U N I V E R S I T E L I B R E D E B R U X E L L E S

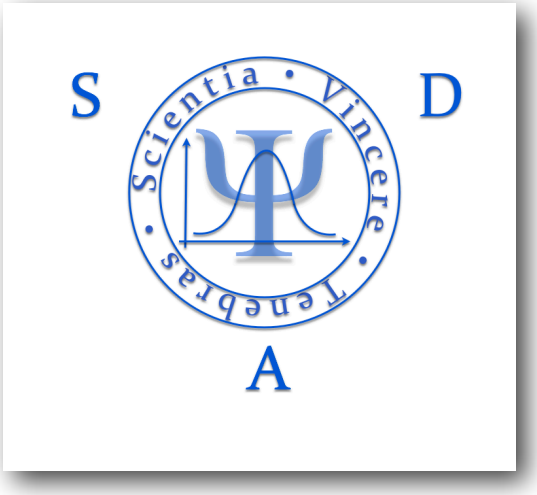
F A C U L T E D E S S C I E N C E S P S Y C H O L O G I Q U E S
E T D E L ' E D U C A T I O N

Syllabus d'Analyse de Données

- Première partie -

Notes de cours et de travaux pratiques

"Je ne crois aux statistiques que lorsque je les ai moi-même falsifiées"
Sir Winston Leonard Spencer Churchill



Professeur titulaire : Leys Christophe

Année académique 2011-2012

REMERCIEMENTS

Je remercie Monsieur Holender, mon prédécesseur, pour ses conseils avisés et pour son autorisation relative à l'utilisation de ses anciennes notes. Je m'en suis inspiré pour plusieurs chapitres, reprenant textuellement certains paragraphes. Les autres chapitres sont essentiellement inspirés du livre de Judd et McClelland, référencé en bibliographie, dont il existe une traduction française en bibliothèque.

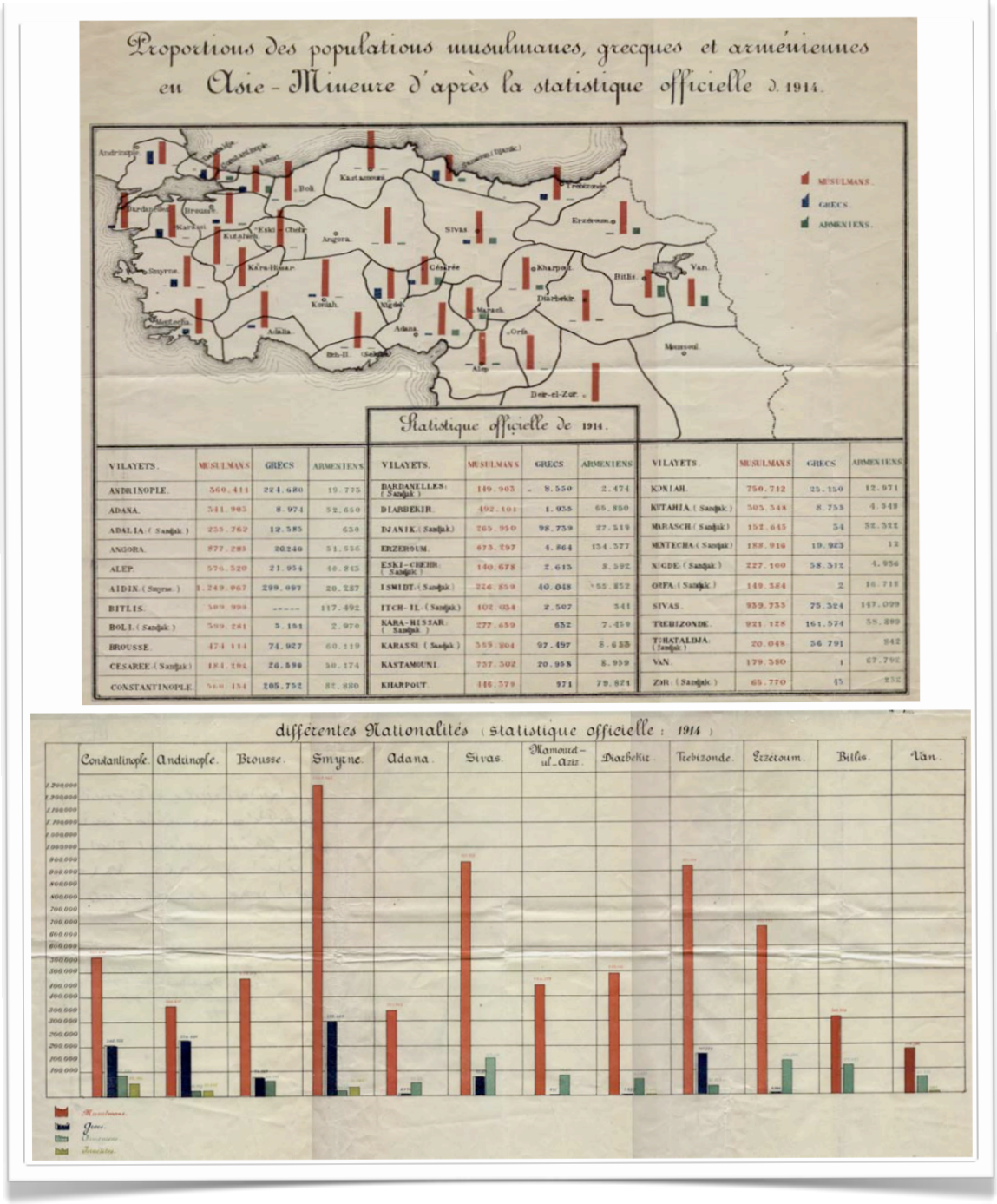
Je remercie également les assistants (Damien Debot, Marie Delacre, Dyna Delle-Vigne, Mohamed El Hamouti, Lola Seyll, Tania Spruch, et Doris Van Cleemput) pour leur travail important dans la conception des exercices de fin de chapitre. J'espère qu'ils éclaireront le lecteur en illustrant de différentes manières la théorie et en le ramenant, tant que possible, aux applications utiles aux psychologues. Je les remercie également pour les remarques qui ont contribué à la clarification de certains concepts.

Je remercie enfin Aude Fenaux pour son travail de relecture et de correction. Je salue son courage pour avoir su lire plusieurs centaines de pages de statistique sans aucune obligation de se pencher sur cette matière.

TABLE DES MATIERES

CHAPITRE 1 : IMPORTANCE DES STATISTIQUES POUR UN PSYCHOLOGUE

1.1. Qu'est-ce que la statistique?



Source : http://commons.wikimedia.org/wiki/File:Proportions_des_populations_en_Asie_Mineure_statistique_officielle_d1914.png?uselang=fr Récupéré le 23/10/11.

La statistique est l'ensemble des instruments de recherches mathématiques permettant de déterminer les caractéristiques d'un ensemble de données (généralement vaste). Les statistiques (au pluriel) sont le produit des analyses reposant sur l'usage de la statistique.

Cette activité regroupe trois principales branches :

a) La collecte des données
b) Le traitement des données collectées, aussi appelé la statistique descriptive
c) L'interprétation des données, aussi appelée l'inférence statistique, qui s'appuie sur la théorie des sondages et la statistique mathématique

Durant la suite de ce chapitre, je vais considérer, je l'espère à tort, que beaucoup d'entre vous n'apprécient absolument pas les statistiques et ne voient pas l'intérêt de ce cours. Cette méthode me permettra de vous expliquer pourquoi vous devez néanmoins vous y atteler.

1.2. En m'inscrivant dans cette faculté, je savais déjà que ce cours n'aurait aucun intérêt pour moi, je me demande bien à quoi ça sert?!

Etudier les méthodes d'analyse de données en Psychologie et en Sciences de l'Education est une activité importante qui s'étale sur trois ans. Il est pourtant évident que la plupart des étudiants entamant ces études ne le fait pas pour l'opportunité d'étudier les statistiques. Plusieurs raisons justifient néanmoins la présence de ce cours au sein de la faculté.

L'un des piliers fondateurs de l'Université est de développer les activités de recherche. La recherche vise essentiellement à décrire la réalité de manière modélisée, c'est-à-dire simplifiée (voir chapitre 2). Cette simplification a comme but de permettre à l'esprit humain de gérer la complexité de la réalité et de décrire ou prédire avec le moins d'erreurs possibles les événements. Actuellement la recherche repose principalement sur l'approche de la connaissance par une méthode empirique. Cette méthode impose l'établissement d'hypothèses (voir chapitre 2) que l'expérimentateur tentera (en principe avec acharnement) d'invalider. S'il n'y parvient pas, il considérera que son hypothèse est solide et l'acceptera comme une vérité temporaire (jusqu'au moment ou quelqu'un parviendra à l'invalider au profit d'une autre hypothèse décrivant mieux la réalité). Cette méthode est à la base des principes du Libre Examen prônés par l'Institution. Malheureusement, pour évaluer la

pertinence d'une hypothèse, il existe peu (à mon sens, pas) d'alternatives à l'utilisation d'une approche mathématique.

Imaginons que vous désiriez simplement montrer qu'en Belgique les hommes ont une taille supérieure aux femmes. Vous pourriez choisir n'importe quelle femme adulte dans la population, puis n'importe quel homme et comparer leur deux tailles. Par cette méthode, vous auriez évidemment une chance non nulle de sélectionner une femme plus grande que l'homme, il suffit pour ça d'avoir accidentellement comparé une basketteuse avec un jockey. Pour diminuer ce genre de risque, vous décidez donc de choisir une centaine de femmes et une centaine d'hommes, totalement au hasard dans la population d'adultes belges. Vous remarquez alors que la taille des femmes tourne autour d'une valeur égale à 169cm, alors que celle des hommes tourne autour de 177cm. Vous annoncez donc avec joie et soulagement que cette hypothèse (qui, pour une raison obscure, vous tenait à coeur) est confirmée (nous verrons plus tard qu'en fait vous avez rejeté l'hypothèse d'égalité des tailles plus que confirmé celle d'une différence). Cependant, un de vos collègues, jaloux de cette trouvaille, vous défie en affirmant que vous n'avez rien prouvé du tout : la différence que vous observez n'est due, selon lui, qu'au hasard de votre sélection. Il prétend que si vous aviez pris l'entièreté des hommes et l'entièreté des femmes adultes vous n'auriez observé aucune différence. A partir de là s'en suit une discussion sans fin (et absolument stérile) où chacun tentera de convaincre l'autre du bien fondé de son opinion… Qui reste une opinion.

Objectif principal de la statistique

La statistique offre une approche mathématique permettant d'évaluer les chances d'affirmer que ces deux tailles sont effectivement différentes les unes des autres (ou pas). Elle vous permettrait, par exemple, de dire à votre collègue que vous affirmez que les hommes sont plus grands que les femmes, mais que vous êtes d'accord de reconnaître qu'il existe une possibilité que vous vous trompiez. Cependant, vous êtes maintenant capable de **quantifier** cette possibilité. Par exemple vous pourriez dire : j'affirme que les hommes adultes sont en général plus grands que les femmes adultes mais j'ai trois chances sur cent de me tromper en vous l'affirmant. En revanche, vous m'affirmez l'inverse, mais vos risques d'erreurs sont très nettement supérieurs aux miens, dès lors je préfère me faire confiance.

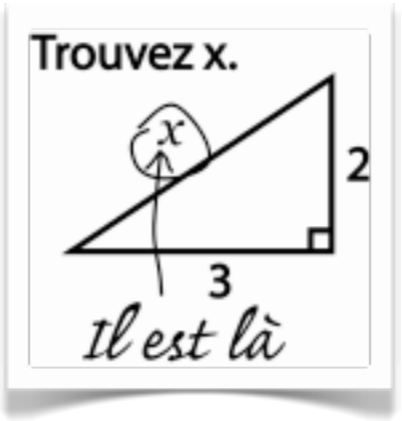
Si vous comprenez ce petit exemple, vous avez d'ores et déjà compris le principe d'une grande partie de nos activités durant ces trois prochaines années. Cet exemple lié à la recherche pourrait vous conduire à me répondre que vous n'avez aucune envie d'entamer une carrière de chercheur, que vous redoutez déjà la perspective de votre mémoire et qu'une fois assis(e) sur votre fauteuil de psychologue thérapeute (objectif de nombreux d'entre vous) vous n'aurez plus jamais rien à voir avec la statistique. Pas si vite! Un thérapeute utilise souvent des tests psychologiques. Ces tests ont été établis à l'aide de nombreuses méthodes statistiques. Si vous ne les comprenez pas, ou que vous ne comprenez pas les implications de ce que signifient ces statistiques, il n'y a environ aucune chance que vous puissiez les utiliser avec pertinence. Et un thérapeute qui utilise un test psychologique comme un moyen de déterminer avec certitude les caractéristiques de son patient est, à mes yeux, nécessairement un thérapeute dangereux.

En outre, l'un des avantages de la formation universitaire est de vous offrir le choix. Quelle que soit votre orientation, vous acquerrez, si tout va bien, une capacité d'analyse d'une situation et une capacité d'assimilation d'informations qui sont recherchées dans le monde du travail. Parallèlement, vous sous-estimez peut-être l'éventail de vos aspirations. Si nombre d'entre vous ont, actuellement, la vocation de devenir "*psychologue thérapeute*", au bout du compte, vous risquez de vous diversifier bien plus que prévu. Or, sur le marché du travail, une formation en analyse de données est précieuse. Il s'agit d'un réel atout à faire valoir pour un engagement, quelle que soit l'institution ou l'organisation que vous désirez rejoindre professionnellement.

Enfin, vous ne sortirez pas d'ici sans vous être adonné à une activité de recherche. Que ce soit au travers de cours, ou lors de votre mémoire, vous allez devoir établir des hypothèses, et tenter de les invalider. Eviter l'utilisation des statistiques est parfois possible, mais rarement (à mon sens jamais) intéressant. D'une part, un jury qui lit dans votre mémoire des phrases de type "*j'ai choisi une méthode qualitative, que je trouve plus riche et plus nuancée qu'une méthode quantitative dans le cas qui nous occupe*" comprendra le plus souvent "*je ne comprends rien du tout aux stats, je vais plutôt essayer de m'en sortir par une dissertation qui, je l'espère, vous convaincra que j'ai correctement réfléchi à la problématique que j'approche*", ce qui n'est pas une bonne manière d'aborder la lecture d'un travail qui couronne votre maîtrise. Les cas où une approche qualitative est à privilégier sont rares et, lorsqu'ils existent, sont beaucoup plus difficiles à gérer avec rigueur qu'une approche quantitative

simple telle qu'elle est demandée lors d'un travail de mémoire. Ne fuyez donc pas l'analyse de données et appropriez-vous cette discipline.

1.3. Je suis nul(le) en math, je ne comprendrai jamais rien!



Ce serait mentir que de dire qu'aucune compétence mathématique n'est nécessaire pour ce cours. Cependant, si vous êtes ici, c'est que vous avez, au moins, acquis les compétences minimales pour en affronter la complexité. L'essentiel de la difficulté se trouve dans la notation, qui devra être votre premier souci. Une fois maîtrisée, le reste devrait être nettement plus aisé. En outre, si vous êtes capable d'ingérer les quantités de matières que vous aurez à ingérer lors de votre cursus, vous êtes plus que certainement capable de traiter l'information contenue dans les cours d'Analyse de Données. Le seul réel empêchement que vous pourriez rencontrer serait lié à l'éventuelle croyance que vous entretenez à propos de vous-même concernant vos affinités pour les mathématiques. Si telle est votre situation, je vous enjoins à lire les articles concernant l'effet Pygmalion et la confirmation du stéréotype. Un très bon article, en français, sur le sujet est celui de Désert, Croizet et Leyens (2002) que vous trouverez sur le site de l'Université Virtuelle. Vous y verrez qu'une raison fondamentale de sous-performer dans une branche pour laquelle on se considère comme mauvais est précisément cette croyance. Je vous garantis donc que, moyennant un état d'esprit suffisamment confiant, vous êtes capables de réussir ce cours qui n'est pas plus compliqué que les autres cours.

Un second piège est à éviter pour réussir un cours qui n'attire pas l'essentiel de votre intérêt : le temps de travail. Un réflexe très humain est d'investir beaucoup de temps dans les activités que l'on aime et très peu dans celles que l'on n'aime pas. Selon moi, l'une des clefs essentielles de la réussite universitaire est de faire exactement l'inverse. Les enseignements qui vous intéressent moins, ou pour lesquels vous éprouvez plus de difficultés, doivent être traités plus longuement, si vous voulez parvenir au même niveau de maîtrise que celui d'une matière qui vous passionne. Et, avec un peu de chance, ce faisant, vous pourriez développer un intérêt là où, au départ, il n'y en avait aucun. Flaubert disait (en substance) : tout finit par être intéressant pour peu que nous l'observions suffisamment longtemps. De manière très pratique, si je devais aborder mon cours, je réfléchirais de cette

manière : il contient 6 ECTS (pour l'ensemble de la théorie et de la pratique), habituellement j'investis 10 heures d'étude par ECTS pour réussir très convenablement un examen (à vous de trouver votre nombre d'heures nécessaires). Je dois donc consacrer 60 heures de mon temps sur ce cours. Cependant, je n'aime pas les statistiques, donc je serai moins performant et, pour en tenir compte, je rajoute 20 heures d'étude. Total 80 heures, soit 10 jours de travail particulièrement soutenu (éventuellement modulé par mon taux de présence au cours). Le reste est une question d'agenda.

1.4. La structure du cours

Une difficulté supplémentaire à laquelle il vous faudra être attentif est la structure de ce cours. Il s'étale donc sur trois ans, et il vous faudra attendre la deuxième année pour commencer à comprendre réellement l'organisation de la matière. En effet, la première année nous nous contenterons de poser les bases pour vous permettre de comprendre les fondements théoriques qui sous-tendent l'entièreté des outils statistiques que l'on utilise traditionnellement en psychologie. Dès lors, à l'issue de cette année vous ne serez toujours pas capables de traiter quantitativement la moindre hypothèse, mais vous aurez en main toutes les clefs vous permettant d'aborder cette notion en BA2. L'inconvénient de cette situation est double : d'une part, il va vous falloir me faire confiance lorsque je prétends que tout ce que je vous raconte est essentiel pour la suite ; d'autre part, vous ne pourrez pas oublier ce cours d'une année à l'autre. En effet, nous nous occuperons cette année, pour l'essentiel, d'aborder les concepts de récolte de données et de traitement descriptif, en ne faisant qu'effleurer l'inférence. C'est pourtant cette inférence qui constituera l'essentiel de votre travail d'analyse de données lorsque vous désirerez vérifier vos hypothèses. Mais l'inférence est l'étape qui suit la récolte de données et le traitement descriptif, elle est donc indissociable de ces deux premières étapes.

Cette situation est extrêmement regrettable. En effet, peu de choses sont aussi agréables que le sentiment d'en avoir fini avec un cours que l'on n'apprécie qu'à moitié. C'est avec un intense soulagement que l'on a tendance à l'entreposer dans le coin le plus reculé de notre mémoire voire même à l'en chasser définitivement. C'est cette tendance qui pose les jalons d'un échec l'année qui suit. Ayez donc le réflexe de conserver votre cours de cette année à portée de main et n'hésitez pas à le compulser à nouveau l'année prochaine. Vous pourriez avoir l'impression désagréable de perdre votre temps, mais en vous remémorant rapidement les concepts présents dans cette partie, vous vous faciliterez grandement le travail de

deuxième année. Au bout du compte, je vous promets un gain de temps par rapport à l'effort que vous devriez fournir pour tenter de comprendre un cours de deuxième alors que vous avez oublié l'intégralité de celui de première.

Enfin, vos cours d'Analyse de Données s'arrêtent en BA3. Vous aurez donc deux ans durant lesquels certains n'auront plus aucune notion de statistique, alors que les autres auront encore l'un ou l'autre cours complémentaire mais dont les enseignants postuleront que vous vous rappelez correctement des informations présentées lors de mes cours. C'est donc deux ans plus tard que vous aurez un réel besoin de monopoliser vos connaissances dans le domaine, lors de votre mémoire. C'est à ce moment que vous devrez être capables de retourner dans ces cours et de retrouver les informations pratiques utiles. Ne les égarez donc pas.

1.5. Aspects pratiques

1.5.1. Organisation des cours pratiques et théoriques et cotation

Le cours d'analyse de données de BA1 contient 6 ECTS. Il est subdivisé en une partie théorique et une partie pratique. A l'issue de ces deux parties vous aurez une seule cote sur vingt points. Ces deux parties conduisent à un seul examen, dans lequel les deux parties ne sont pas séparées. La pratique doit donc être envisagée comme un support pour comprendre la théorie, et, inversement, la théorie ne sert qu'à comprendre ce que l'on fait lorsque l'on prend une décision statistique. A la fin de chaque chapitre une série d'exercices vous est proposée. Certains seront corrigés lors des séances de TP, d'autres sont destinés à votre entrainement. Vous pouvez cependant poser les questions à propos d'exercices non-vus aux TP soit à vos assistants soit aux séances supplémentaires animées par les élèves-assistants. Certains exercices sont surlignés en gris. Il s'agit des exercices à faire en premier lieu (ce qui ne vous dispense pas de faire les autres). Il s'agit des exercices fondamentaux sur lesquels des complexifications viennent s'ajouter dans les autres exercices.

1.5.2 Cas particuliers des étudiants doubleurs ou en réussite partielle

Il existe les cas particuliers des étudiants doubleurs en BA1 et des étudiants en réussite partielle en BA2. Ces étudiants qui auraient réussi l'une des parties du cours avec 12 sont en principe dispensés de cette partie du cours. Cependant, comme cette année les deux

examens sont réunis cela n'a plus de sens. De ce fait, quiconque a réussi la partie pratique l'année passée **n'est plus obligé** d'assister aux TP. Ils devront cependant effectuer l'examen complet. Cependant, je leur conseille vivement d'assister aux travaux pratiques, surtout aux derniers, qui envisagent une matière différente de celle de l'année passée. Les étudiants ayant réussi la partie théorique repasseront également l'examen complet.

1.5.3 Horaire des cours pratiques et théoriques

Les horaires du cours théorique sont applicables dès la première semaine du second quadrimestre. Il y aura deux séances par semaine, le mardi de 16h à 18h, l'autre le vendredi de 14h à 16h, à l'auditoire K.1.105. Treize séances sont prévues. Si des congés m'empêchent de les dispenser, une ou deux séances de rattrapage seront organisées et les horaires vous seront alors communiqués. Les travaux pratiques ont lieu à raison d'une séance par semaine. Vous aurez à vous y inscrire au secrétariat de Madame Fenaux (Bâtiment D, 10ème niveau, D10-153). C'est également auprès de Madame Fenaux que vous pouvez vous adresser pour toute question logistique, afenaux@ulb.ac.be.

1.5.4. Evaluation

L'évaluation tiendra en un examen écrit à livre fermé. **Vous êtes autorisés à utiliser une calculatrice scientifique de base non programmable (mais pas de GSM, ni de tablettes)**. L'ensemble de l'évaluation durera environ 4 heures (un peu moins). La date et le lieu de l'examen seront communiqués par la faculté.

1.5.5. Matière

Le syllabus contient l'intégralité des informations, tant théoriques que pratiques, sur lesquelles vous pourrez être interrogés. Il contient de nombreuses pages mais ne doit pas vous effrayer : ce volume est lié à la présence d'exemples et à une ventilation importante du texte (interligne = 1,5) afin d'en améliorer la clarté. Remarquez par exemple que nous sommes à la dix-huitème page et que vous n'avez encore rien à étudier au sens propre du terme. En outre, vous verrez que je répète plusieurs fois la même chose, à différents endroits du cours, conformément à la méthode d'enseignement par spirale de Bruner (1960). L'idée sous-jacente est d'attirer votre attention sur les différents liens qui existent entre les chapitres et d'installer votre compréhension petit à petit (j'espère que ce sera efficace).

Le cours théorique donné lors des séances vise à expliquer, à l'aide de diaporamas, les concepts les plus importants. Bien qu'ils contiennent l'essentiel de l'information, le diaporama ne remplace en rien le syllabus. Il est à voir comme un complément, condensé et illustré. Je vous déconseille donc de vous contenter d'utiliser les diaporamas comme support d'étude et vous enjoins à utiliser les deux médias pour assimiler votre cours.

1.6. Conclusion

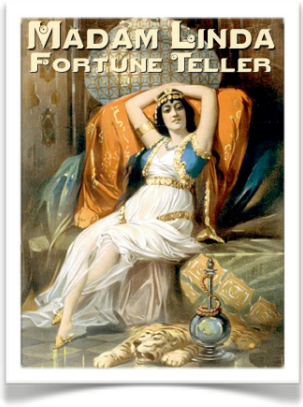
Ce petit chapitre introductif a comme vocation de vous sensibiliser à trois éléments essentiels au bon déroulement de votre réussite à l'Université : (a) l'analyse de données est une branche importante à laquelle vous n'échapperez pas, même si cela vous fâche ; (b) je vous enjoins à considérer ce cours comme un cadeau, presqu'à coup sûr, utile pour votre futur emploi quel qu'il soit ; (c) vous êtes tout à fait capables de surmonter cette difficulté si vous avez les compétences nécessaires pour surmonter les autres cours de la faculté.

P A R T I E I

MODELISATION ET INCERTITUDE

CHAPITRE 2 : LES NOTIONS DE PREDICTION - VARIABLE - HYPOTHESE - LOGIQUE ET REPRESENTATION PAR LES ENSEMBLES

2.1. La prédiction en psychologie



Lorsque nous utilisons le terme prédire dans le langage usuel, nous faisons souvent référence à la prédiction d'un événement futur. Cependant, en science, et plus particulièrement en psychologie, ce terme reflète un concept différent. Comme nous l'avons dit dans le premier chapitre, un enjeu majeur de la recherche scientifique est d'établir un modèle descriptif de la réalité. Notre volonté est de décrire un concept psychologique, de la manière la plus précise possible. Malheureusement, il est impossible, pour l'esprit humain, de prendre en compte l'infinie configuration des informations qui définissent une situation donnée.

2.2. Description de la réalité - concept de variable

Imaginons que je doive évaluer les chances de réussite de l'un d'entre vous. J'éprouverais beaucoup de difficultés à prédire avec certitude votre résultat (vous-même auriez du mal, même juste APRES avoir passé l'examen). Néanmoins, je pourrais, à l'aide de quelques informations, améliorer mes chances de définir une cote. Par exemple, si j'étais capable de mesurer : votre temps de travail ; votre présence au cours ; votre motivation ; la difficulté de mon examen ; votre participation aux guidances ; vos résultats aux autres cours ; votre nom de famille (si par hasard, à votre désavantage, vous portiez le même nom que celui de ma belle-mère, ou, à votre avantage, le même que le mien) ; le temps qu'il fera ; le nombre de bêtises qu'auront fait mes enfants ; la qualité de mes repas ; le nombre d'heures de sommeil que j'ai eu les sept jours avant la correction ; votre nombre d'heures de sommeil ; la qualité de vos repas ; etc. Je pourrais alors éventuellement estimer votre cote avec plus ou moins de précision. Remarquez que je pourrais encore trouver d'autres informations qui influencent de près ou de loin cette cote, potentiellement un grand nombre même (qui sait si la couleur des murs de votre chambre ou de votre bureau n'a pas d'influence sur vos performances?). Remarquez également que certaines informations sont probablement redondantes ou

influencées par d'autres. Par exemple, votre temps d'étude pourrait être influencé par votre motivation.

Au bout du compte, je me retrouve avec une quantité presqu'infinie d'informations, qui s'influencent éventuellement entre elles et avec lesquelles je dois me débrouiller pour prédire votre cote. A partir de maintenant, nous appellerons ces informations des **VARIABLES**.

Cet exemple est très important parce qu'il plante le décor de l'entièreté de votre cours d'analyse de données. Vous devez réaliser, à ce stade, que trois grands problèmes se posent à vous lorsque vous tentez de décrire la réalité :

Contraintes de la modélisation de la réalité

(a) Il est extrêmement difficile, voire impossible, d'**identifier** l'ensemble des variables la définissant.

(b) Même après les avoir identifiées, il est impossible à l'esprit humain d'**évaluer l'influence** de l'ensemble d'entre elles.

(c) Il est souvent difficile et parfois impossible de **mesurer** chacune de ces variables.

2.3. Choix des variables

La solution pour modéliser le mieux possible la réalité est de déterminer les variables qui sont **les plus pertinentes**, c'est-à-dire qui influencent le plus le concept que je veux décrire. Nous pouvons, par exemple, espérer que votre nom ou la qualité de mes repas influencera moins votre cote que le nombre d'heures passées à l'étude de votre syllabus. De plus, il m'est beaucoup plus facile d'acquérir l'information "temps d'étude" qu'une bonne évaluation de l'information "qualité de repas" qui est beaucoup plus subjective. Idéalement, si je pouvais trouver une seule variable qui prédit très bien votre cote, je serais satisfait. Cependant, je peux éventuellement en gérer trois ou quatre (nous irons rarement au-delà en psychologie).

2.4. Hypothèse

2.4.1. Définitions

Définition de l'hypothèse

Pour déterminer ces variables, je dois poser des **HYPOTHESES**. Une hypothèse est une **prédiction** de l'influence d'une variable sur une autre. La variable qui influence l'autre s'appelle la **VARIABLE INDEPENDANTE**. La variable qui est influencée par la variable indépendante s'appelle la **VARIABLE DEPENDANTE**. Il est tout à fait indispensable que vous vous familiarisez avec ces termes (faites les exercices de fin de chapitre).

Dans notre exemple, le temps d'étude est une variable indépendante puisque c'est celle qui va nous permettre de prédire la cote. La cote est donc influencée par le temps d'étude et est donc la variable dépendante.

Remarquez que l'établissement du statut de la variable (dépendante ou indépendante) dépend de l'expérimentateur. Par exemple, j'aurais très bien pu m'intéresser au temps d'étude et tenter de l'évaluer à partir de la cote en partant de l'hypothèse qu'un étudiant qui a bien réussi mon cours a probablement passé beaucoup de temps à l'étudier. Dans cette optique, la cote devient ma variable indépendante à partir de laquelle je vais estimer le temps de travail qui devient ma variable dépendante.

2.4.2. Théorie versus Intuition

Pour générer une hypothèse, j'ai deux outils : **la théorie et l'intuition**. Cette phrase, qui n'a l'air de rien à ce stade-ci, est pourtant fondamentale dans l'approche statistique, comme nous le verrons dans les prochaines années. Ce qu'on appelle "la théorie" est en fait l'ensemble des recherches qui ont été réalisées sur la problématique qui nous intéresse ou sur une problématique très proche. Elle permet bien souvent d'appliquer un modèle déjà existant à notre cas particulier. Elle m'autorise donc à prédire quelque chose ***a priori***. Par exemple, Coulter (1979) rapporte les principaux résultats relatifs à l'influence du temps d'étude sur le résultat de l'examen et me permet d'établir mes hypothèses.

L'intuition est synonyme de "bon sens". Lorsque j'approche la réalité, comme je l'ai fait avec l'exemple ci-dessus, je me suis dit que mon temps sommeil pouvait être un indicateur (en estimant par exemple qu'il influence mon humeur et que je sois suffisamment mauvais enseignant pour que mon humeur ait quoi que ce soit à voir avec vos résultats). Mais rien n'est moins sûr. En fait, je n'en sais rien, et aucune étude (à ma connaissance) ne le montre. Il est même possible que le manque de sommeil ait un effet positif sur votre cote. Par exemple, je pourrais avoir envie de vite en finir avec mes corrections pour pouvoir me reposer et omettre certaines de vos fautes par inattention. Enfin, il est également possible que cette variable n'ait, en fait, aucun effet sur la correction de votre examen. L'intuition me permet donc d'**explorer** avec plus ou moins de succès ma problématique. Ce n'est que suite à cette exploration que je pourrai réellement établir une hypothèse.

Vous constaterez rapidement dans vos autres cours et lors de votre mémoire qu'il est extrêmement mal vu de ne pas justifier théoriquement vos hypothèses. Lorsque vous en établissez une, la seule raison valable de ne pas citer les travaux qui ont été réalisés auparavant est que vous abordiez un domaine complètement nouveau dans lequel il n'y a aucun moyen de prédire quoi que ce soit. C'est évidemment très rare, et, dans ce cas, vous êtes tenus d'être rigoureux dans votre approche exploratoire.

2.4.3. Propriétés d'une hypothèse

La première vertu d'une hypothèse est qu'elle doit être falsifiable, c'est-à-dire qu'il doit être possible de la réfuter. Par exemple, si je vous affirme que la réalité n'existe pas, qu'il ne s'agit que d'une projection de votre esprit, je ne fais pas une hypothèse. En effet, il est totalement impossible de vérifier d'une quelconque manière que je me trompe. Une hypothèse infalsifiable de cette sorte est une croyance. Ce domaine échappe complètement aux statistiques et, plus généralement, à la science (bien qu'elle puisse, malheureusement, être influencée par les croyances des expérimentateurs).

Croyance en une hypothèse

Attention : La certitude que vous ressentez à propos de votre hypothèse ne peut que définir l'acharnement avec lequel vous voudrez la prouver, mais, en soi, elle ne constitue aucune preuve. Une des qualités essentielles que l'on exige de vous est d'être capable d'abandonner cette certitude lorsque l'expérience vous montre que votre hypothèse est incorrecte.

La deuxième caractéristique d'une hypothèse est son aspect prédictif. **Il ne s'agit pas d'une question, mais bien d'une prédiction**. Lorsque l'on pose une hypothèse, on prédit l'influence d'une variable sur une autre.

Il existe deux types d'hypothèses : les hypothèses théoriques et les hypothèses opérationnelles. Les premières envisagent l'influence générale d'une variable sur une autre. Les secondes prédisent très concrètement le résultat d'une expérience. Par exemple, lorsque je dis : "le temps d'étude améliore la performance aux examens", je fais une hypothèse générale. En revanche, lorsque je dis : "les étudiants qui auront étudié cinq heures le syllabus d'analyse de données en BA1 auront une cote plus faible que ceux qui auront étudié cinquante heures", j'énonce une hypothèse opérationnelle.

Enfin, une hypothèse ne peut contenir qu'une seule proposition. Par exemple, lorsque je dis : "les étudiants qui n'ont pas étudié le syllabus d'analyse de données en BA1 auront une cote plus faible que ceux qui l'auront étudié cinq heures, eux-mêmes auront une cote plus faible que ceux qui l'auront étudié cinquante heures", je fais, en fait, trois hypothèses opérationnelles : (a) ceux qui n'ont pas étudié réussiront moins bien que ceux qui ont étudié cinq heures ; (b) ceux qui ont étudié cinq heures réussiront moins bien que ceux qui ont étudié cinquante heures ; (c) ceux qui n'ont pas étudié réussiront moins bien que ceux qui ont étudié cinquante heures. Cependant, ces trois hypothèses opérationnelles testent la même hypothèse théorique.

2.5. Modélisation

Finalement, imaginons qu'après mûre réflexion et nombreuses lectures théoriques, j'en vienne à la conclusion que le temps d'étude, la présence au cours et la motivation soient les mamelles de votre réussite. J'envisage donc de simplifier la réalité au point d'ignorer complètement l'infinité des autres variables et de leurs interactions. Ne pas tenir compte de ces variables suppose nécessairement que je m'apprête, volontairement, à commettre une certaine erreur de prédiction. Cet élément est essentiel, il signifie que je suis conscient de ne pas être déterministe dans mon estimation. Je me situe dans une optique probabiliste : lorsque j'utilise mon modèle, j'ai une **plus grande probabilité de décrire correctement la réalité** que si je n'utilise pas mon modèle, mais je sais que je fais quand même une erreur. Mon espoir est que cette **erreur soit la plus petite possible** et que j'aie réussi à identifier les quelques variables qui prédisent le mieux votre cote. Représenté mathématiquement, je peux m'exprimer de la manière suivante (retenez-la, on l'utilisera souvent) :

Réalité = Modèle + Erreur

En conclusion, une prédiction désignera, dans ce cours, un modèle prédictif et/ou descriptif de la réalité (dans mon équation, le modèle = ma prédiction). Ce modèle sera une simplification de la réalité à laquelle sera donc associée une erreur.

2.6. Historique

Les conceptions probabilistes n'ont pas toujours été utilisées par l'Homme. En Europe, nous pourrions retracer en quelques étapes clefs l'émergence de cette pensée. A ce sujet, je vous conseille de lire le livre de Ian Hacking (1975), si cela vous intéresse.

Les premières collectes de données, à la base des statistiques, semblent assez anciennes. On les situe il y a 4000 ans en Chine en l'an 2 de la dynastie des Han. Ils visaient à recenser la population, les revenus et le nombre de soldats mobilisables[1]. En réponse aux besoins de gestion d'un état, ces recensements sont devenus de plus en plus fréquents et universels. Par exemple, à Rome, les censeurs avaient comme fonction de collecter les données nécessaires

[1] http://factsanddetails.com/china.php?itemid=39&catid=2&subcatid=2

à établir le taux d'imposition et la répartition de ces revenus dans les différents domaines de l'administration. Le Tanakh (l'Ancien Testament) nous offre un éclatant exemple de recensement : le Livre des Nombres (quatrième tome du Pentateuque). Il contient de nombreux dénombrements établissant les effectifs de la population, le nombre de décès, de naissances, etc. Puis, plus on avance dans l'Histoire, plus ces recensements sont fréquents et précis, pour arriver, de nos jours, au recensement extrême que nous connaissons dans nos sociétés modernes (où les instituts de statistique ne manquent pas de travail). Cependant, la notion d'incertitude et de probabilité n'est apparue que très tard. Hacking (1975) situe son apparition au niveau des années 1650.

Auparavant, la certitude a une longue histoire. Le prophète Isaïe (ci-contre à droite) est une figure biblique du Tanakh qui aurait vécu au 8ème siècle avant J.-C.. Outre ses prophéties qui nous intéressent assez peu (mais qui rejoignent la notion usuelle de prédiction assez éloignée de celle que nous envisageons), il dira : "*Si vous ne croyez pas, vous ne comprendrez pas*" (Isaïe, 7, 9). Cette phrase suggère que la certitude est accessible à l'Homme par sa croyance en Dieu. Pour comprendre le monde, il faut écouter les paroles de Dieu car lui seul détient la connaissance. Toute autre recherche de vérité est considérée comme inutile, voire hérétique, pendant de nombreuses années.

Arrêtons-nous à Nicolas Copernic (1473-1543). Ce chanoine catholique, et homme de sciences, établit la théorie de l'héliocentrisme, en profond désaccord avec la théorie généralement admise du géocentrisme (plaçant la Terre au centre de l'univers). Considérer que le Soleil est au centre du système planétaire et que la Terre ne serait qu'un objet tournant autour du Soleil est perçu comme une hérésie. Copernic sera d'ailleurs condamné, à titre posthume, par l'Eglise, défendant la théorie du géocentrisme. A sa suite, Galilée (1564-1642) défenseur acharné de la théorie copernicienne s'est également vu attaqué par l'Eglise pour hérésie en 1616 (représentation ci-dessus) : défendre l'héliocentrisme correspond à questionner la Bible dont le Psaume 93

rapporte la phrase "*Tu as fixé la Terre ferme et immobile*", jugée incompatible avec l'héliocentrisme.

Un contemporain de Galilée, René Descartes, se bat également pour sortir de l'argument d'autorité consécutive à la lecture de la Bible comme celle de la parole de Dieu nécessairement vraie. Pour lui, la vérité est accessible à l'esprit de l'Homme. Pour peu que son raisonnement soit rigoureux et méthodique, c'est-à-dire appuyé par la démonstration mathématique, l'Homme peut atteindre la vérité. Sa conception est importante dans la mesure où elle permet à la pensée humaine de sortir de l'obscurantisme religieux d'alors. Mais elle n'introduit pas encore de pensée probabiliste, la certitude étant encore le but ultime de la quête.

Ce sont Pascal (ci-contre) et Fermat qui ouvrent le champ d'étude en s'attaquant au jeu de dés vers le milieu du 17ème siècle. Ils remarquent qu'en lançant deux dés, le résultat de la somme obtenu n'est pas déterminable à l'avance et que les résultats possibles ne sont pas équiprobables. En effet, les chances d'obtenir 12 (double six) sont moins élevées que les chances d'obtenir, par exemple, 7 (six et un ; un et six ; cinq et deux ; deux et cinq ; trois et quatre ; ou quatre et trois). A la suite de leurs études, ils énoncent le concept de probabilité sous la forme de "*degré d'incertitude*". Il en émergera la notion "*d'espérance*" lorsqu'il s'agit d'une situation d'incertitude. Le "*Pari de Pascal*" est la plus célèbre application de cette nouvelle notion d'incertitude. Pour Pascal, la croyance en Dieu est la seule attitude rationnelle. Le raisonnement (simplifié) est le suivant : selon la Bible, le séjour de l'Homme sur Terre est éphémère, mais l'âme est éternelle. Dès lors, vivre une vie vertueuse en accord avec les commandements est un investissement minime (une soixantaine d'années à l'époque) comparé à la récompense d'une vie éternellement heureuse dans l'au-delà. Dans la mesure où le gain potentiel est infini, aucun investissement temporaire ne peut être trop cher payé.

La conclusion de Pascal sur la vérité est la suivante : pour la découvrir il est nécessaire de s'appuyer sur des prémisses dont la vérité est déjà établie. Cette méthode est cependant impossible dans la mesure où, pour établir ces premières vérités, il faudrait s'appuyer sur

d'autres vérités, et, ultimement, aucune vérité primordiale ne peut être établie. Il établi ce raisonnement sur ses travaux en géométrie où il montre que les démonstrations, aussi parfaites soient-elles, reposent ultimement sur quelques principes (les axiomes) qui sont impossibles à démontrer. Il en résulte une incertitude résiduelle qui, si petite soit-elle, n'est pas nulle.

Nous pourrions ramener cette pensée aux deux catégories de logique : la logique déductive et la logique inductive. La première est une logique certaine, établie *a priori*, au bout du compte assez rarement possible si ce n'est dans le cadre de démonstrations mathématiques (et encore, comme on vient de le voir, jusqu'à une certaine limite). La seconde est un raisonnement risqué, établi *a posteriori*, basé sur l'observation du monde, dans laquelle l'incertitude est toujours présente. C'est aussi la logique inductive qui est à la base de la recherche en sciences humaines et, plus généralement, des raisonnements que nous tenons au quotidien. Les points suivants ont pour but de vous expliquer ces deux approches et leurs conséquences respectives.

2.7. La logique

La logique concerne le raisonnement, ou plus précisément, l'argumentation. Une argumentation consiste à présenter un certain nombre de raisons, que l'on appelle les prémisses, qui étayent une conclusion. Un raisonnement correct consiste à ne pas inférer une conclusion fausse à partir de prémisses vraies. A la fois les prémisses et les conclusions s'expriment linguistiquement sous forme de **propositions**. Une proposition est un énoncé qui peut être vrai au faux (une hypothèse est une forme de proposition).

2.7.1. La logique déductive

Selon la logique de déduction, si les prémisses sont vraies et que l'argumentation (le processus qui agence les prémisses de manière à en faire émerger les conclusions) est valide, alors, les conclusions ne peuvent être que vraies. Il n'y a pas le moindre risque associé à des arguments déductifs valides appliqués à des prémisses vraies.

Un des prototypes du raisonnement déductif est le **syllogisme d'Aristote**. La logique d'Aristote concerne des énoncés qu'on appelle des propositions. Une proposition est une assertion comprenant un sujet et un prédicat (= un attribut). Par exemple, lorsque je dis

"*Tous les Hommes sont mortels*", le sujet est "*Tous les Hommes*" et l'attribut est "*mortels*". D'un point de vue qualitatif, une proposition peut être vraie ou fausse. D'un point de vue quantitatif, une proposition peut concerner tous les cas (comme dans notre exemple), certains cas ou un seul cas. En combinant les deux qualités et les trois quantités on obtient les six types de propositions représentés au Tableau 2.1.

Tableau 2.1. Différents types de propositions étudiées en logique. En excluant les propositions singulières, on obtient les propositions étudiées par Aristote. Les logiciens du Moyen Age les ont organisées sous forme d'un ***carré logique*** et identifiées par les lettres *a, i, e, o*.

	Affirmative	Négative
Universelle	Tous les Hommes[a] sont mortels (**a**)[b]	Aucun Homme n'est mortel (**e**)
Particulière	Quelques (certains) Hommes sont mortels (**i**)	Quelques (certains) Hommes ne sont pas mortels (**o**)
Singulière	Socrate est mortel	Socrate n'est pas mortel

[a] La majuscule à « *Hommes* » signifie que l'on parle de l'espèce humaine.

[b]Au Moyen Age, les logiciens scolastiques ont identifié les propositions affirmatives universelles et particulières par les deux premières voyelles du mot ***affi**rmo* et les propositions négatives universelles et particulières par les voyelles du mot *n**ego***.

La logique d'Aristote ne traite pas des propositions singulières (qui ont été rajoutées ultérieurement), mais seulement des propositions universelles affirmatives et négatives et de propositions particulières affirmatives et négatives de sorte que l'on parle du **carré logique** d'Aristote. Le carré logique sert à effectuer des inférences immédiates entre une prémisse et une conclusion. Par exemple, de la vérité de la proposition « *tous les Hommes sont mortels* » on peut inférer la fausseté de la proposition « *aucun Homme n'est mortel* » ou la vérité de la proposition « *certains Hommes sont mortels* ». Ce type de logique repose donc sur les inclusions d'ensembles qui seront essentiels dans la suite de ce cours. Dans notre cas, inférer que certains Hommes sont mortels vient du fait que l'ensemble "*certains Hommes*" est inclus dans l'ensemble d'ordre supérieur "*Tous les Hommes*".

En plus des inférences immédiates, l'apport d'Aristote à la logique tient des inférences **médiates**. Dans ce cas, la conclusion n'est plus tirée à partir d'une seule prémisse, mais bien de deux ou plus. Lorsque tel est le cas, on parle de **syllogisme**. Un exemple classique est :

Tous les Hommes sont mortels prémisse 1
Tous les Grecs sont des Hommes prémisse 2

Tous les Grecs sont mortels conclusion

Deux types d'erreurs peuvent survenir qui rendraient la conclusion erronée. La première tient de la fausseté d'une prémisse (par exemple, les Hommes ne sont en fait pas mortels). La seconde tient de la fausseté de l'argumentation en elle-même (que l'on appelle la **validité**). Par exemple, bien que les deux prémisses "*Tous les mammifères sont des animaux*" et "*tous les mammifères sont visibles à l'oeil nu*" soient vraies, la conclusion "*donc les microbes n'existent pas*" est fausse.

Remarquez que la prémisse "*Tous les Hommes sont mortels*" est supposément vraie. Cependant, nous ne pouvons en être sûrs que parce que, jusqu'à présent, face à une multitude d'informations sur le sujet, nous avons pu observer que cette proposition s'est toujours avérée vraie et jamais fausse. La deuxième prémisse est également vraie pour la même raison. C'est la limite de la logique déductive dont nous rediscuterons lorsque nous envisagerons la logique inductive.

Le Tableau 2.2. vous montre les deux cas : celui où les prémisses sont fausses (au moins une des deux) et celui où l'argumentation est valide ou pas. Les colonnes montrent l'état des prémisses, les lignes celui de l'argumentation. Vous remarquerez que lorsqu'une prémisse est fausse (par exemple, "*tous les Hommes sont intelligents*") l'argument peut rester valide compte tenu des prémisses. A l'inverse, même lorsque les deux prémisses sont vraies et que la conclusion est vraie (Certains Hommes sont mortels, tous les Grecs sont des Hommes, donc tous les Grecs sont mortels), l'argumentation peut être fausse (le "donc" n'est pas permis car, à l'aide des prémisses données, on ne peut déduire cette conclusion).

Tableau 2.2. Prémisses vraies ou fausses et argumentations valides ou non valides

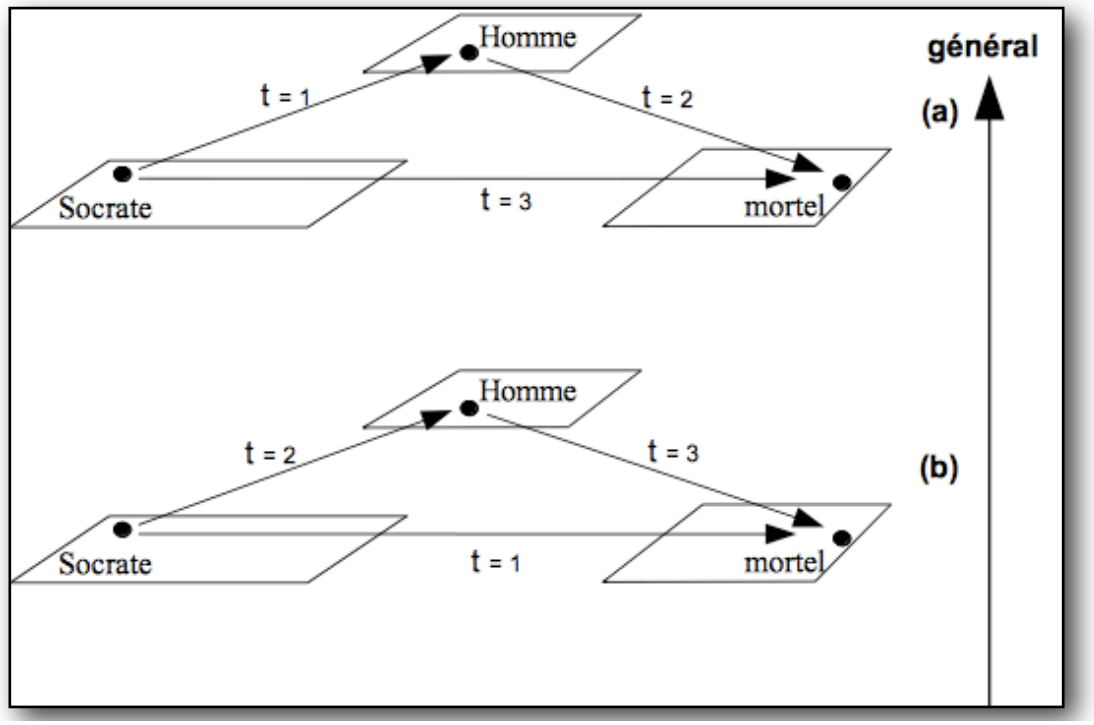
	Prémisses	
Argumentation	Les deux vraies	Une fausse et une vraie
Valide	Tous les Hommes sont mortels Tous les Grecs sont des Hommes donc, tous les Grecs sont mortels	Tous les Hommes sont intelligents Tous les Grecs sont des Hommes donc, tous les Grecs sont intelligents
Non valide	Certains Hommes sont mortels Tous les Grecs sont des Hommes donc, tous les Grecs sont mortels	Certains Hommes sont immortels Tous les Grecs sont des Hommes Donc, tous les Grecs sont immortels

En conclusion, nous voyons que la logique déductive ne peut pas conduire à une erreur pour peu que les prémisses soient vraies et que l'argumentation soit valide.

2.7.2. La logique inductive

La logique inductive repose sur un raisonnement nécessairement incertain dont le principe est de chercher à découvrir des lois générales à partir d'observations de faits. Cette approche est basée sur des inférences probabilistes. L'idée de base est que, plus un phénomène donné est observé, plus il y a de chances qu'il se produise à nouveau. Parallèlement, moins il y a de contre-exemples plus les chances que ce phénomène correspondent à une loi naturelle est grande. Par exemple, la prémisse "*tous les Hommes sont mortels*" que nous utilisions au point précédent correspond, comme je l'ai mentionné auparavant, au cas où de très nombreuses observations montrent que les Hommes finissent par mourir. A cela s'ajoute le fait qu'aucun Homme n'a vécu plus longtemps qu'une 120aine d'années (dans les cas exceptionnels). Le fait est tellement établi que l'on finit par le considérer comme certain et par cesser de chercher un contre-exemple. La Figure 2.1 représente les différences entre un raisonnement inductif et déductif.

Figure 2.1. : (a) raisonnement déductif, (b) raisonnement inductif.

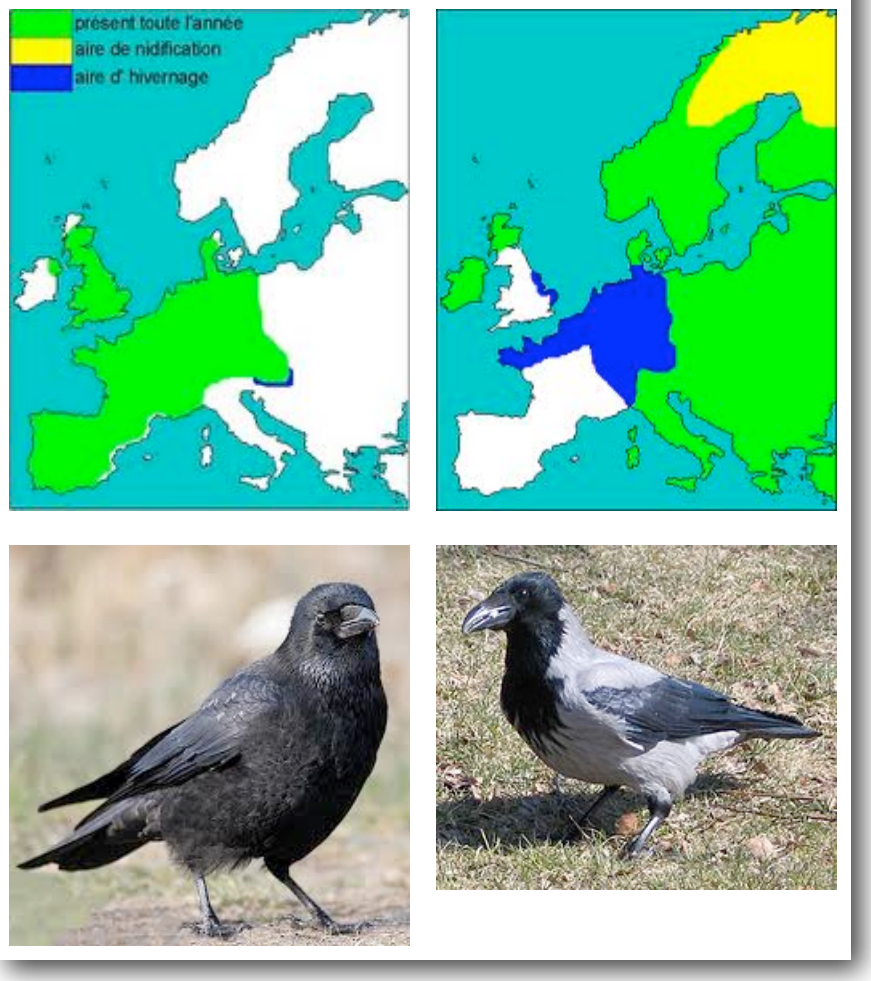


Il existe un lien de réciprocité entre la logique déductive et la logique inductive. Comme nous l'avons vu, la logique déductive dira : tous les Hommes sont mortels, les Grecs sont des Hommes, donc ils sont mortels. A l'inverse, la logique inductive dira : Tous les Grecs finissent par mourir un jour, donc les Hommes sont mortels. Vous remarquerez par cet exemple que le premier raisonnement est certain, pour peu que mes prémisses soient vraies et que mon argument soit valide, alors que le second est probabiliste (le fait que tous les Grecs soient mortels n'implique pas forcément que tous les Hommes doivent l'être, alors que l'inverse bien), dans ce cas-ci il est très probablement vrai également.

Certaines conclusions que l'on a coutume de tirer par un raisonnement inductif n'arrivent pas à un degré de certitude aussi bon que l'exemple du paragraphe précédent (et certainement pas dans le domaine de la psychologie). Par exemple, si je regarde par la fenêtre les corneilles, ces oiseaux noirs au bec noir également, très fréquentes à Bruxelles, je pourrais me dire que les corneilles sont des oiseaux noirs très communs en Europe et nettement plus communes que leur cousines mantelées (grises et noires). En revanche, si je regarde la répartition géographique de cet oiseau en Europe, je m'aperçois que la corneille noire est nettement moins fréquente que la corneille mantelée, qui occupe la plus grande partie du territoire européen (voir Tableau 2.3), mais que je ne vois jamais vu qu'elle ne

fréquente pour ainsi dire pas la Belgique (où je passe le plus clair de mon temps), sauf en hiver à de rares endroits.

Tableau 2.3 : Répartition géographique de la corneille noire (à gauche) et de la corneille mantelée (à droite)



Par essence notre cerveau fonctionne sur base d'une logique inductive. Nous apprenons en nous basant sur la réalité et les faits que nous observons. Une mauvaise utilisation de la logique inductive est, par exemple, souvent à la base des stéréotypes que nous entretenons à propos de groupes sociaux : lorsqu'Eric Zemmour déclare lors d'une émission télévisée que « *les Français d'origine immigrée sont plus contrôlés que les autres parce que la plupart des trafiquants sont Noirs et Arabes. C'est un fait* » (Le Monde, 24 Mars 2010), il fait référence à la sur-représentation des Français d'origine immigrée dans les faits de drogue. Cette sur-représentation est effectivement établie par des chiffres très concrets. Cependant, entretenir la croyance que les Français d'origine immigrée sont des trafiquants serait potentiellement une erreur. C'est pourtant une erreur très fréquente lorsqu'on énonce un stéréotype. En Belgique, le stéréotype d'énoncé "*les Arabes sont des voleurs*" est assez (trop) fréquent. Mais les chiffres montrent en effet que les Arabes ou Belges d'origine arabe sont sur-représentés

pour les faits de petite délinquance (dont le vol). Cependant, le nombre de prisonniers en Belgique tourne autour de 10.000 individus (pas tous Arabes ni d'origine arabe, loin s'en faut) alors que la communauté marocaine (pour ne citer qu'elle) comprend 300.000 personnes. Imaginez le chaos qui règnerait si l'entièreté de cette communauté était réellement délinquante. On décrirait bien mieux la réalité en considérant que 99% des délinquants sont des hommes. Pourtant nous n'inférons pas que tous les hommes sont des délinquants.

Ces exemples vous montrent comment, par l'utilisation d'une logique inductive, l'esprit peut rapidement se fourvoyer dans ses conclusions mais également pourquoi cette même logique est obligatoirement à la base de tout raisonnement scientifique. En effet, dans la mesure où la logique déductive demande des prémisses vraies, et que ces prémisses doivent être établies, il est nécessaire de recourir à la logique inductive pour y parvenir (rappelons-nous, une fois de plus, que si je tiens pour vrai que les Hommes sont mortels ce n'est que parce que l'expérience ne m'a jamais rien montré d'autre). En ce sens, l'essentiel de notre connaissance se base sur l'observation du monde et sur des inférences, des généralisations, à partir de ces observations, c'est-à-dire une logique inductive. Remarquez également que cela correspond à regarder certains événements et à faire des inférences pour tous les événements. Les notions quantitatives de "*tous*" et "*certains*" sont donc primordiales dans la suite des événements (comme nous le verrons dans notre approche des représentations par les ensembles).

Il est important de comprendre qu'une proposition est toujours intrinsèquement vraie ou fausse, elle n'est jamais, en soi, probablement vraie ou probablement fausse. En revanche, nous (en tant qu'êtres humains) n'avons que très rarement accès à cette information avec certitude. Un dernier exemple qui terminera, je l'espère, de vous convaincre et vous permettra d'y réfléchir, est relatif au droit[2]. La détermination de la culpabilité factuelle d'un prévenu est avant tout une question de probabilité. Lorsqu'un crime est commis et qu'un individu est soupçonné, il n'y a aucun moyen, ni pour les magistrats ni pour les jurés, de revenir en arrière et de revivre le moment. On ne peut qu'accumuler des indicateurs (film, empreintes, ADN, témoignages, etc.) qui minimisent la probabilité de déclarer coupable un innocent, mais sans jamais parvenir à la certitude absolue. Cependant, cela ne signifie pas

[2] J'utiliserai cet exemple pour deux raisons : d'une part parce qu'il illustre bien mes propos, et d'autre part, parce que cela vous montre que des considérations statistiques sont très importantes même dans une faculté très peu axée sur les mathématiques.

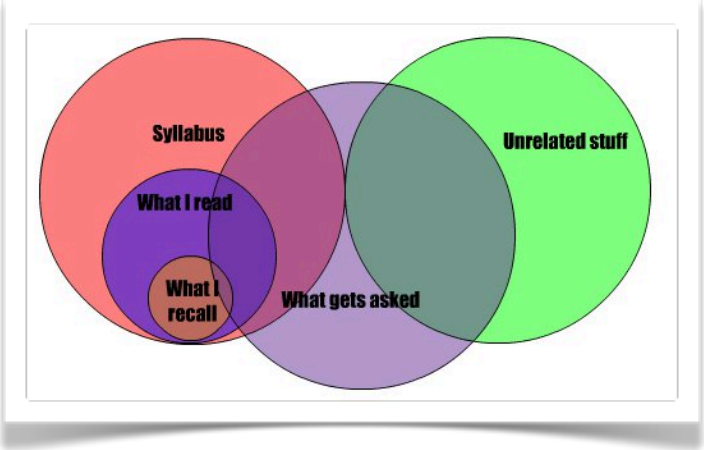
que la logique inductive soit irrationnelle, même si comme nous l'avons vu concernant les stéréotypes, les conclusions peuvent être erronées, ce n'est pas automatiquement le cas. L'enjeu est de trouver des règles permettant d'estimer et de quantifier l'incertitude que l'on a en affirmant une proposition.

D'un point de vue mathématique, l'enjeu revient à admettre que la relation "*implique vraisemblablement*" qui lierait une proposition "*r*" à une proposition "*p*" soit une relation mathématique à part entière. Dès lors, l'intérêt revient à estimer la meilleure relation possible entre les propositions, c'est-à-dire la plus probable.

2.7.3. La logique inductive et la modélisation

Je terminerai ce sujet par faire le lien entre la logique inductive que nous venons de passer en revue et la modélisation, au travers d'un exemple pratique. Imaginons (je n'irai pas jusqu'à dire "*le cas improbable*") que vous désiriez vous rendre à l'Université à temps pour assister à mon cours. Pour y parvenir, vous aurez pris soin d'évaluer le temps que vous prenez entre chez vous et l'Université, fort probablement par simulation abstraite (estimation du temps de transport en commun, évaluation du temps de marche) et par expérience (en venant à l'Université). C'est donc la pratique qui vous permet, par induction, d'évaluer un temps de trajet adéquat. Cependant, cette estimation ne vous donne qu'une probabilité d'arriver à temps. Il suffit d'une grève du personnel de la société de transports en commun, d'un réveil qui ne sonne pas, de clefs introuvables, et vous voilà en retard. Dès lors, vous établissez par une logique inductive un temps plausible de trajet (votre modèle), mais ce temps n'est jamais qu'une estimation à laquelle est associée une erreur. Tout l'enjeu consistera donc à trouver l'heure idéale qui vous permettra d'arriver le plus souvent à l'heure sans toute fois vous faire arriver un temps trop long à l'avance.

2.8. Représentation des raisonnements à l'aide des ensembles



Dans la mesure où les probabilités concernent les chances d'obtention d'un événement donné, il est nécessaire de pouvoir situer cet événement parmi l'ensemble des événements possibles. Il est également indispensable de pouvoir séparer plusieurs ensembles ayant une propriété spécifique ainsi que des ensembles pouvant regrouper plusieurs propriétés. Cette façon, relativement complexe, de considérer les événements peuvent se décrire graphiquement, jusqu'à un certain point (trois ensembles et leurs intersections) ou algébriquement. Ces représentations sont essentielles pour aborder les notions de probabilités (sujet du chapitre 3).

2.8.1. Représentations graphiques des ensembles

Les diagrammes que Venn (1824-1923) a développés à partir de ceux proposés par Euler (1707-1783) sont utilisés en logique déductive notamment pour représenter des raisonnements qui se fondent sur les quantificateurs ***tous***, ***certains*** et ***aucun***. Les quatre cas du Tableau 2.2 peuvent être illustrés comme aux Figures 2.2a démontrant la validité de deux des arguments et 2.2b démontrant la non validité des deux autres arguments.

La Figure 2.2a montre, par la transitivité de la notion d'inclusion d'ensembles, que les deux arguments de la première ligne du Tableau 2.2 incluant le quantificateur "*tous*" sont des arguments valides, même si dans le cas de droite la première prémisse – "*tous les Hommes sont intelligents*" – est fausse. La Figure 2.2b montre, grâce aux relations d'inclusion et d'intersection d'ensembles, que les deux arguments de la deuxième ligne du Tableau 2.2 incluant le quantificateur « *certains* » sont non valides, même si dans le cas de gauche les deux prémisses sont vraies.

Figure 2.2a. Diagrammes de Venn montrant les arguments valides du Tableau 2.2 incluant le quantificateur "tous" dans les deux prémisses même si dans un cas une des prémisses est fausse.

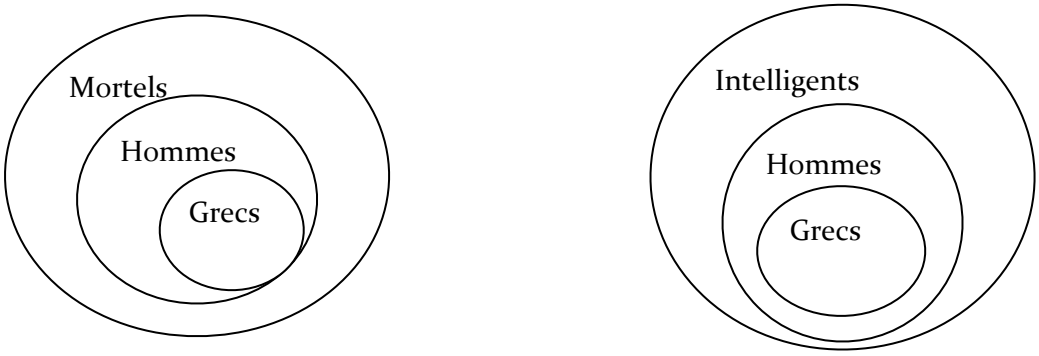
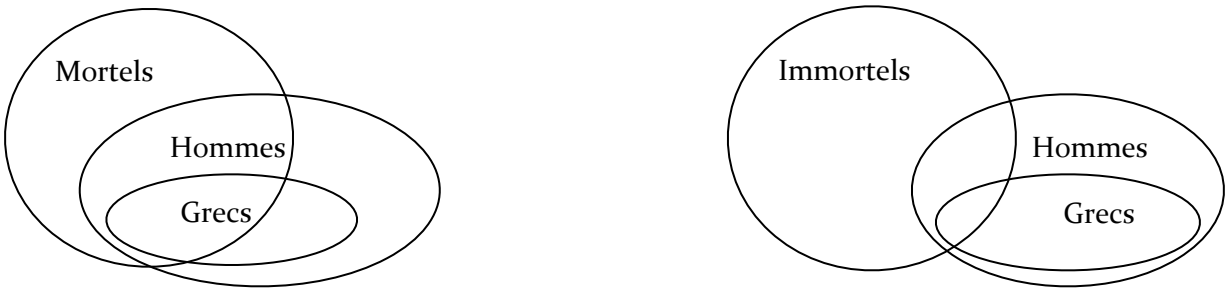


Figure 2.2b. Diagrammes de Venn montrant que les arguments non valides du Tableau 2.2 incluant le quantificateur « certains » même si dans un cas les deux prémisses sont vraies.



Remarques importantes : On se limite à discuter les cas où tous les ensembles dont les extensions sont représentées par les ellipses sont des ensembles contenant un ***nombre fini d'éléments dénombrables***. Tous les ensembles dont il sera question dans ce chapitre et le suivant ont cette propriété. Ceci implique que la notion de probabilité dont il sera question dans ce qui suit porte exclusivement sur des ***événements discrets*** (c'est-à-dire finis, par opposition à continus où chaque valeur est possible rendant ainsi le dénombrement impossible puisqu'infini) ***dénombrables***. En outre, les diagrammes de Venn sont des représentations de propriétés (de prédicats) qui sont supposées vraies de manière absolue pour tous les éléments constituant l'***extension*** des ensembles. En effet, en logique déductive, la proposition « *tous les Hommes sont mortels* » veut dire qu'il n'y a pas le moindre doute sur ce que le prédicat « *mortels* » exprime à propos du concept « *Hommes* ». Pour empiéter sur le chapitre suivant, en termes probabilistes, cela veut dire que l'on attribue une probabilité de 1 à l'assertion « *tous les Hommes sont mortels* ». En effet, lorsque nous parlerons de probabilité nous aurons coutume d'attribuer le chiffre 0 lorsque la

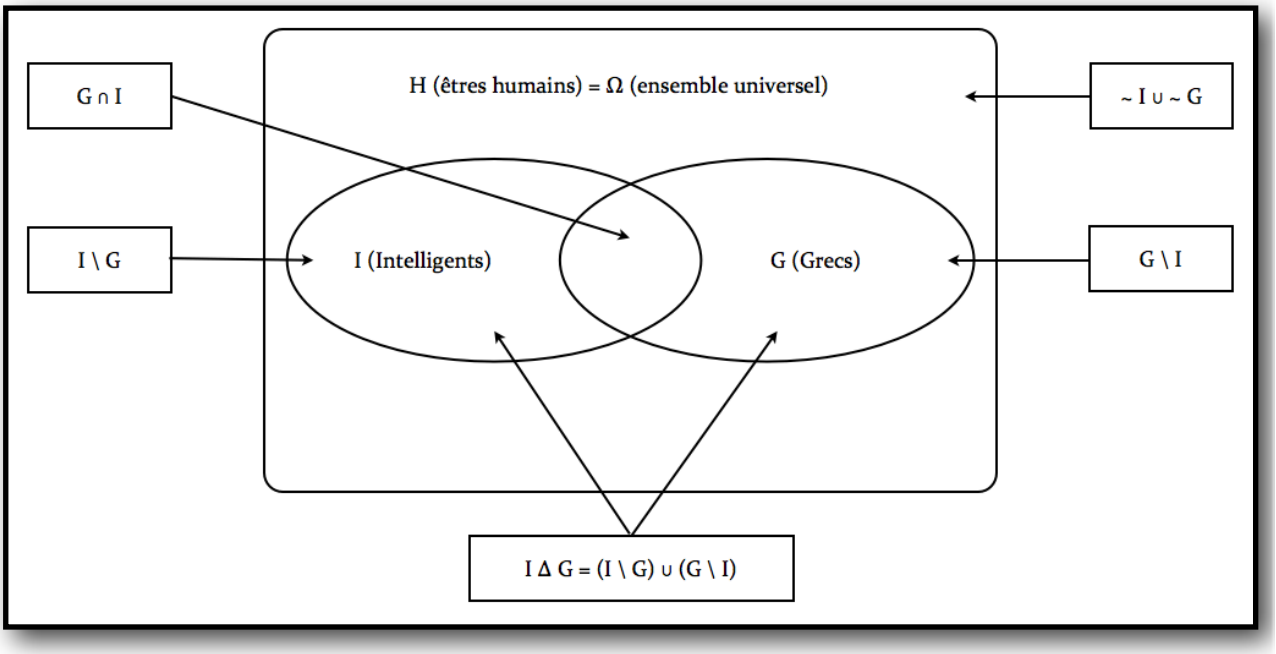
probabilité est nulle (il n'y a aucune chance de voir l'événement apparaître) et 1 lorsque l'événement est certain. **Une probabilité sera donc toujours exprimée par une valeur comprise entre ces deux bornes**.

2.8.2. Représentations algébriques des ensembles

La représentation graphique des ensembles est très rapidement limitée par les deux dimensions de la feuille. Dès lors que nous devons représenter plus de trois ensembles et leurs intersections (si elles existent), le recours au graphique devient impossible. Nous ne pouvons alors plus que nous reposer sur l'algèbre. Pour le comprendre, il est nécessaire de se familiariser avec les notations et avec les concepts qu'elles représentent. C'est l'objectif de ce point.

La Figure 2.3 sert de support aux définitions qui suivent. On considère cette fois l'ensemble des êtres humains noté H comme ensemble universel représenté par un rectangle et deux ensembles représentés par des ellipses en recouvrement partiel (ayant une intersection non vide) : l'ensemble des individus intelligents noté I et l'ensemble des individus grecs noté G (notez que la surface de l'intersection n'est pas forcément proportionnelle au nombre d'individus qui l'occupent, je ne suis pas en train de suggérer qu'il y a très peu de Grecs intelligents).

Figure 2.3. : Ensemble des individus possédant la propriété d'être grec et ensemble des individus possédant la propriété d'être intelligent considérés comme des sous-ensembles de l'ensemble universel constitués par les êtres humains.



2.8.2.1. L'inclusion et l'appartenance

La relation d'inclusion d'ensembles représentée à la Figure 2.2a correspond à la notion logique d'implication. L'inclusion de l'ensemble G des Grecs dans l'ensemble H des êtres humains (« *Hommes* ») s'écrit G ⊂ H (personnellement je me représente ce signe comme un "*c*" que je nomme "*contient*"), celle de l'ensemble des êtres humains dans l'ensemble des êtres mortels s'écrit H ⊂ M et, par transitivité de la relation d'inclusion, l'ensemble des Grecs est inclus dans l'ensemble des êtres mortels, ce qui s'écrit G ⊂ M. On peut traduire la relation d'inclusion d'ensembles en termes d'implication logique comme suit :

- Etre grec implique d'être humain, ou, si x est grec, alors x est humain.
- Etre humain implique d'être mortel, ou, si x est humain, alors x est mortel.
- Etre grec implique d'être mortel, ou, si x est grec, alors x est mortel.

Lorsqu'un ensemble appartient à un ensemble plus grand, on dit qu'il est inclus et on le note "⊂". En revanche, lorsqu'un élément appartient à un ensemble, on utilise le signe "∈". Par

exemple, Socrate est un élément qui appartient à l'ensemble des être humains H (Socrate ∈ H).

2.8.2.2. L'ensemble vide

La notion d'ensembles s'étend même aux cas d'ensembles ne contenant aucun élément. Cet ensemble unique s'appelle l'ensemble vide et est noté "ø". La notion d'ensemble vide joue dans la théorie des ensembles un rôle similaire à la notion de zéro en arithmétique. Cet ensemble qui ne contient aucun élément est considéré comme un sous-ensemble qui est inclus dans n'importe quel ensemble. En effet, si on admet que l'ensemble vide est bien un ensemble, il n'y a rien de contradictoire à le considérer comme inclus dans n'importe quel ensemble puisque la notion d'inclusion d'ensemble a une portée tout à fait générale. Le concept d'ensemble vide est utile pour définir la notion d'ensembles disjoints (c'est-à-dire dont l'intersection est l'ensemble vide).

2.8.2.3. Le singleton

Jusqu'ici, on a traité d'ensembles ou de sous-ensembles contenant un grand nombre d'éléments. Mais la notion d'ensembles s'étend aussi aux cas d'ensembles contenant un seul élément (un singleton). Cette notion sera importante à la section suivante où les événements élémentaires sur lesquels portent la définition de la notion de probabilité seront considérés comme les membres de sous-ensembles contenant chacun un seul élément (un seul événement élémentaire).

2.8.2.4. L'ensemble universel

La notion d'ensemble s'étend aussi à l'ensemble de tous les éléments auxquels on s'intéresse dans une situation particulière, ce qu'on appelle l'ensemble universel noté "Ω" (lettre grecque oméga majuscule). Cet ensemble joue également un rôle important. Dans le cas des deux raisonnements valides illustrés à la Figure 2.2a, on peut penser que c'est l'ensemble le plus inclusif qui sert de référence au discours, donc l'ensemble des organismes mortels dans le premier cas et l'ensemble des êtres intelligents dans le second, mais ceci n'a été ni dit, ni représenté explicitement. La notion d'ensemble universel est indispensable à la définition de la notion de complément d'un ensemble que nous verrons. Elle joue également un rôle

crucial dans la définition de la notion de probabilité (cf. la section suivante) où l'ensemble universel portera le nom d'espace-échantillon (c'est-à-dire le lieu duquel on peut prélever tous les échantillons).

L'ensemble considéré comme universel dépend du contexte du discours. Le choix de l'ensemble des êtres humains est en continuité avec les exemples précédents. On aurait pu choisir comme ensemble universel l'ensemble des habitants de l'Europe, ou du pourtour méditerranéen, ou de la Grèce continentale ou envisager seulement les individus de sexe masculin, etc.

2.8.2.5. Complément d'un ensemble

La notion de complément d'un ensemble correspond à la négation logique de la possession d'une propriété. Par exemple, l'ensemble des individus qui ne possèdent pas la propriété d'être grec est noté ~G. Cet ensemble ~G complémentaire de l'ensemble G ne peut pas être défini dans les contextes où l'ensemble universel demeure non spécifié. Dans notre exemple, toute la surface du rectangle délimitant l'ensemble universel H qui n'est pas recouverte par la surface de l'ellipse délimitant l'ensemble des individus grecs représente les individus non grecs, donc tous les êtres humains qui n'ont pas la propriété d'être grec.

Remarque

Une notation alternative est de représenter la lettre avec une barre au-dessus. Par exemple, l'auteur de science fiction Alfred Elton Van Vogt a écrit le monde des A et le monde des Ā (non-A). mais depuis l'existence de logiciels informatiques qui rendent difficile d'accès ce type de représentation, on préfèrera celle que je viens de présenter. Par ailleurs, ce livre est une belle illustration de la **sémantique générale**, qui se trouve être une logique non-aristotélicienne plus adaptée aux limites de l'esprit humain. Elle fut théorisée essentiellement par Korzybski dans son ouvrage majeur "*Science and Sanity, an Introduction to Non-Aristotelian Systems and General Sémantiques*", dont la première édition paraît en 1933 (la même année que l'axiomatique des probabilités comme nous le verrons). Certains chapitres importants de cet ouvrage sont disponibles gratuitement sur internet à l'adresse http://semantiquegenerale.free.fr/doc.htm. Ce type de pensée logique est à la source de la pensée symbolique et fait intervenir la notion ***percept*** qui tient compte du fait que l'homme n'a pas un accès direct à la réalité mais n'en a que des perceptions qui mènent à des représentations approximatives. Ce point dépasse le présent cours de statistique, mais me semble une approche fondamentale pour tout psychologue. Il rejoint cependant mon discours en appuyant la nécessité de modéliser la réalité et souligne l'émergence obligatoire d'une erreur associée au modèle. Le premier ouvrage de Korzybski qui pose déjà les jalons de la sémantique générale (bien que moins complète que son ouvrage plus complet cité plus haut) est en accès libre (et anglophone) sur internet à l'adresse http://esgs.free.fr/uk/art/manhood.htm.

2.8.2.6. L'intersection d'ensembles

La notion d'intersection d'ensembles correspond au "*et*" logique (conjonction). L'intersection notée I ∩ G (I inter G) de l'ensemble des individus intelligents et de l'ensemble des Grecs contient seulement des individus à la fois intelligents et grecs. La proposition « *il est intelligent et il est Grec* » n'est vraie que si elle s'applique à un individu qui est à la fois intelligent et Grec, elle est fausse pour un individu intelligent mais pas grec, pas intelligent mais Grec et pas intelligent et pas Grec.

2.8.2.7. Différence d'ensembles

Notion de différence entre ensembles. Notez qu'à la Figure 2.3, vous n'avez aucune peine à repérer la portion de la surface de l'ensemble I qui correspond à des individus intelligents qui ne sont pas grecs : il s'agit de l'ensemble noté I \ G constitué par la différence entre les ensembles I et G (dans le sens I moins G). C'est pareil pour repérer la portion de l'ensemble des individus grecs qui ne sont pas intelligents : il s'agit de l'ensemble G \ I. Mais, où se trouvent les individus qui ne sont ni intelligents, ni grecs ? La réponse est qu'ils occupent toute la surface du rectangle délimitant l'ensemble universel H qui est extérieure aux deux ellipses en recouvrement partiel.

2.8.2.8. Union d'ensembles

a) Union inclusive

La notion d'union d'ensembles correspond au "*ou*" logique qu'on appelle "*ou inclusif*". L'union des ensembles des individus intelligents et de l'ensemble des Grecs s'écrit I ∪ G (I union G). Cette union correspond à un nouvel ensemble qui contient trois sous-ensembles : le sous-ensemble des individus intelligents qui ne sont pas Grecs (I \ G), le sous-ensemble des individus Grecs qui ne sont pas intelligents (G \ I) et le sous-ensemble des individus Grecs et intelligents (I ∩ G). Donc la proposition logique « *il est intelligent ou il est Grec* » est vraie si l'individu considéré est intelligent mais pas Grec ; s'il est Grec mais pas intelligent ; et s'il est Grec et intelligent. Elle est fausse dans le cas où cet individu n'est ni intelligent, ni Grec.

b) Union exclusive

La notion de "*ou exclusif*" consiste à exclure l'intersection du cas précédent. Donc, la proposition « *il est intelligent ou il est grec, mais pas les deux* » est vraie pour un individu intelligent mais pas grec et pour un individu grec mais pas intelligent ; elle est fausse pour un individu qui est intelligent et grec et pour un individu non intelligent et non grec. La relation entre ensembles s'écrit I Δ G ; elle définit un ensemble constitué de deux sous-ensembles, excluant l'intersection entre ensembles.

2.8.2.9. Ensembles disjoints (= mutuellement exclusifs)

Notion d'ensembles disjoints ou mutuellement exclusifs. Dans le cas d'ensembles comme l'ensemble des individus intelligents et l'ensemble des Grecs, il est évident qu'il existe un ensemble $I \cap G \neq \varnothing$, c'est-à-dire un ensemble contenant des individus qui ont à la fois la propriété d'être intelligent et la propriété d'être Grec. Ces deux propriétés sont compatibles, c'est-à-dire qu'elles peuvent coexister en même temps chez un même individu. En revanche, la propriété d'être immortel (noté Im) et la propriété d'être Grec sont (jusqu'à preuve du contraire) incompatibles. L'intersection entre ces deux ensembles est vide : $Im \cap G = \varnothing$. Si le diagramme de Venn de la partie droite de la figure 2.2b représentait correctement la réalité, il ne devrait pas y avoir d'intersection non vide entre l'ellipse contenant les êtres immortels et les ellipses qui contiennent respectivement les êtres humains et les Grecs.

2.8.2.10. Tableau de synthèse

Apellation française	Notation	Nom de la notation
Et	$\cap$	Inter
Ou (inclusif)	$\cup$	Union
Ou (exclusif)	Δ	Delta
Non	$\sim$	Non
Si (sachant que)	$\mid$	Barre
(Ensemble) inclus dans (un autre ensemble)	$\subset$	Contient
(Ensemble) n'est pas inclus dans (un autre ensemble)	$\not\subset$	Ne contient pas
(Elément) appartient à (un ensemble)	$\in$	Appartient
(Elément) n'appartient pas à (un ensemble)	$\notin$	N'appartient pas
Ensemble vide	$\varnothing$	Phi
Ensemble universel ou espace-échantillon	Ω	Omega

2.9. Exercices de fin de chapitre

T.P. 1 : CHAPITRE 2

Exercice 1 : Hypothèses théoriques et opérationnelles Variables dépendantes et indépendantes

1. Identifiez les hypothèses parmi les propositions suivantes. Lorsqu'il s'agit d'une hypothèse identifiez la variable dépendante et indépendante. Lorsqu'il ne s'agit pas d'une hypothèse expliquez pourquoi.

   a. Le son du piano énerve les enfants.

   b. Est-ce que le nombre de roues d'un camion améliore sa tenue de route ?

   c. Le monde a été créé il y a une demi-heure.

   d. Les femmes sont moins bonnes en statistique que les hommes.

   e. Les clous rouillent plus rapidement lorsqu'ils sont en présence d'adultes dépressifs.

2. Donnez une hypothèse opérationnelle correspondant aux hypothèses suivantes et identifiez les variables dépendantes et indépendantes :

   a. Aider les gens rend heureux celui qui aide.

   b. Aller aux guidances permet d'améliorer la réussite à l'examen concerné.

   c. Le climat influence l'état dépressif des individus.

   d. Les enfants victimes de traumatismes sont plus résilients lorsqu'ils parviennent à l'exprimer graphiquement.

e. Les victimes de parents abuseurs deviendront souvent abuseurs avec leurs propres enfants.

3. Donnez l'hypothèse générale correspondant aux hypothèses suivantes et identifiez les variables dépendantes et indépendantes :

a. Lorsque le scientifique demande aux sujets d'infliger des chocs électriques à un individu, les sujets vont obéir jusqu'à infliger des chocs d'une intensité provoquant la mort de l'individu.

b. Les sujets qui ont bu trois verres de bières rateront plus de manoeuvres que les sujets sobres.

c. Après s'être entraînés une journée à jouer aux fléchettes, les sujets qui passent leur journée du lendemain à se remémorer l'entraînement seront plus performants que les sujets qui n'y pensent plus.

d. Les sujets qui ont été témoins d'une scène au cours de laquelle un individu d'origine africaine subit une injustice auront moins de préjugés sur les Africains.

e. Les sujets qui ont été touchés physiquement par l'expérimentateur ont davantage accepté de laisser leurs coordonnées à cette personne que les sujets que l'expérimentateur n'a pas touchés.

T.P. 1 : CHAPITRE 2

Exercice 2 : Mise en situation - Article

Auto-régulation des immunoglobulines salivaires A par les enfants

Karen Olness, MD, Timothy Culbert, MD, and Donald Uden

Des observations et études cliniques ont montré que les enfants présentent l'habileté d'utiliser une variété de techniques d'imagerie mentale comme traitement de plusieurs problèmes aigus et chroniques. Citons l'hémophilie, l'arthrite, l'énurésie, les migraines, l'incontinence fécale.
Quelques études réalisées chez la personne adulte suggèrent que certains aspects de la fonction immunitaire puissent être soumis à un contrôle volontaire par le biais de l'hypnose.

La présente étude s'intéresse à la possibilité d'une modulation volontaire du système immunitaire chez les enfants suite à une séance d'auto-hypnose.

Pour ce faire, les auteurs ont montré à des enfants (âgés entre 6 et 12 ans) une vidéo présentant des marionnettes. Une marionnette représentait un virus; l'autre, qui ressemblait à un policier, représentait le système immunitaire. Cette vidéo constituait ainsi une illustration simplifiée du fonctionnement interne du corps aisément compréhensible par les enfants. Une fois la vidéo visionnée les enfants étaient soumis à une séance d'auto-hypnose. Pour un premier groupe d'enfants, cette séance n'était associée à aucune suggestion spécifique qui puisse augmenter le taux d'immunoglobuline (groupe A). Pour un second groupe d'enfants, cette séance était associée à des suggestions spécifiques visant à augmenter le niveau d'immunoglobulines : on demandait aux enfants de fermer les yeux, de se relaxer et d'imaginer de nombreuses marionnettes policiers parcourant leur corps (groupe B). L'hypothèse étant que les enfants du second groupe montreront une plus grande augmentation du niveau d'immunoglobuline que ceux du premier groupe.

L'analyse des échantillons de salive relevés chez chacun des enfants a révélé une augmentation substantielle du niveau d'immunoglobuline chez les enfants ayant expérimenté la séance d'auto-hypnose associée à des suggestions spécifiques (groupe B). C'est-à-dire que le système immunitaire de ces enfants s'est mis à fonctionner comme s'il combattait de vraies infections. Cette augmentation significative n'a pas été relevée dans l'autre groupe d'enfants (groupe A).

1. Quelles sont les hypothèses théorique et opérationnelle des auteurs ?

2. Pour pouvoir modéliser le mieux possible la réalité, il convient de déterminer les informations les plus pertinentes, c'est-à-dire les plus susceptibles d'influencer la question de recherche.

    a. Dans la présente étude, quelle est la question de recherche que les auteurs souhaitent décrire ?

    b. Quelle(s) est(sont) l(les)'information(s) choisie(s) par les auteurs comme étant les plus pertinentes, c'est-à-dire qui influence(nt) le plus leur question de recherche ?

3. Comment appelle-t-on ce type d'information ?

4. Déterminez les variables dépendante et indépendante. Expliquez.

5. L'hypothèse des auteurs s'est-elle vue validée ou invalidée ? Expliquez.

6. Que constitue la théorie ? Que permet-elle ? En quoi est-elle importante ?

7. Quel est le contexte théorique sur lequel se basent les auteurs pour générer leur hypothèse ?

8. Les auteurs observent les résultats (les « faits »), d'une étude et en tirent des conclusions générales. Ce type d'approche correspond à la logique .....

T.P. 1 : CHAPITRE 2

Exercice 3 : Diagrammes de Venn et notation ensembliste

Dans un groupe de 100 étudiants en psycho (données fictives) :

- 10 sont inscrits à un cours de statistiques par obligation
- 85 parce qu'ils trouvent ce cours très intéressant (eh oui… ;-))
- 80 parce que c'est un cours facile (mais si…)
- 10 parce que c'est une obligation et parce que c'est un cours facile.
- 10 avancent les trois raisons à la fois.

1. Tracez le diagramme de Venn correspondant à ces différentes propositions en indiquant dans chaque partie le nombre de personnes qui s'y trouvent.

2. Combien d'étudiants n'ont choisi le cours de statistiques que par obligation ?

3. Utilisez les notations adéquates pour indiquer :

   a. Les parties comprenant les personnes qui n'ont pris ce cours que par obligation ou parce qu'il s'agit d'un cours facile à l'exclusion des personnes qui ont pris le cours pour ces deux raisons réunies. De combien de personnes s'agit-il ?

   b. Qu'une certaine personne que nous appellerions « a » a pris le cours par intérêt :

   c. La(les) partie(s) dans lesquelles se trouvent les personnes qui ont pris le cours par obligation mais pas par intérêt :

   d. La(les) partie(s) dans lesquelles se trouvent les personnes qui ont pris le cours par obligation ou par intérêt (ou inclusif) :

   e. La(les) partie(s) dans lesquelles se trouvent les personnes qui ont pris le cours par intérêt et parce qu'il s'agit d'un cours facile :

f. La(les) partie(s) dans lesquelles se trouvent les personnes qui ont pris le cours pour les trois raisons à la fois :

4. Le chercheur se rend compte que 5 étudiants ont été oubliés dans sa liste. Il s'agit de 5 étudiants qui n'ont évoqué aucune des trois raisons pour lesquelles ils ont pris ce cours. Représentez le diagramme de Venn complété par cette information.

5. Utilisez la notation adéquate pour indiquer la partie qui se trouve autour des ensembles tracés là, c'est-à-dire la partie qui comprend l'entièreté du schéma (y compris la partie qui n'appartient à aucun des ensembles). À quoi correspond-elle dans le cas présent ?

6. Donnez les probabilités suivantes dans ce nouveau contexte :

T.P. 1 : CHAPITRE 2

Exercice 4 : Logique déductive et Syllogisme d'Aristote

Selon la logique déductive, pour aboutir à une conclusion vraie, il faut d'une part que les prémisses soient vraies et, d'autre part, que l'argumentation soit valide. Donc, il suffit qu'une prémisse soit fausse ou que l'argumentation soit non valide pour que la conclusion soit erronée. Le tableau ci-dessous montre les différents cas de figure possibles.

	Prémisses		
Argumentation	Les deux vraies	Une fausse et une vraie	Les deux fausses
Valide			
Non valide			

Placez les exemples suivants dans le tableau ci-dessus. Représentez les diagrammes de Venn correspondant à ces syllogismes.

1. Certaines personnes qui suivent une thérapie ont de graves problèmes.
   Certaines psychologues ont suivi une thérapie.
   Donc, certains psychologues ont de graves problèmes.

4.

2. Toutes les personnes qui suivent une thérapie sont folles.
   Certains psychologues ont suivi une thérapie.
   Donc, tous les psychologues sont aussi fous que leurs patients.

5.

3. Tous les adolescents ont touché à la drogue.
   Tous les gens qui ont touché à la drogue vont mourir jeunes.
   Donc, tous les adolescents vont mourir jeunes.

4. Tous les rectangles ont quatre côtés.
   Tous les carrés sont des rectangles.
   Donc, tous les carrés ont quatre côtés.

   a. À quoi correspond l'ensemble universel ?

   b. En français, à quoi correspond Q \ R (sachant que Q = ensemble des formes géométiques à quatres côtés, R = l'ensemble des recangles et C = ensemble des carrés) ?

   c. Que vaut C \ R ?

   d. En français, à quoi correspond ~ R ?

5. Certains félins ont des griffes.
   Tous les chats sont des félins.
   Donc, tous les chats ont des griffes.

   a. En français, à quoi correspond F ∩ G ?

   b. En français, à quoi correspond F Δ C ?

6. Certains travailleurs productifs sont heureux au travail.
   Tous les patrons sont des travailleurs productifs.
   Donc, certains patrons sont heureux au travail.

7. Tous les êtres humains ont des stéréotypes.
   Tous les psychologues sont des êtres humains.
   Donc, tous les psychologues ont des stéréotypes.

8. Toutes les fleurs sont oranges.
   Toutes les carottes sont des fleurs.
   Donc, toutes les carottes sont oranges.

   a. En français, à quoi correspond ~ F ∩ ~ C ?

   b. En français, que signifie C ⊂ F ?

   c. Si F ⊂ O, par transitivité, que pouvez-vous conclure ?

T.P. 1 : CHAPITRE 2

Exercice 5 : Les variables

6. Une hypothèse peut être définie comme une prédiction à propos des effets d'une variable (a) _____ sur une variable (b) _____.

7. Des chercheurs désirent comparer les capacités d'interactions sociales d'un groupe d'enfants de 3 ans qui sont allés à la crèche avec celles d'un groupe d'enfants de 3 ans qui n'y sont pas allés. Quelle est la variable dépendante dans cette étude ? Quelle est la variable indépendante ?

8. Dans une étude sur les capacités mnésiques, les participants doivent mémoriser en soirée une liste de 20 mots qu'on leur demande de rappeler après un intervalle de 8 heures. Un groupe de participants dort pendant cet intervalle, l'autre groupe

est maintenu éveillé. L'expérimentateur désire investiguer si le type d'activité (sommeil vs veille) pendant l'intervalle influence le nombre de mots correctement rappelés. Dans cette étude, quelles sont les variables dépendante et indépendante ?

9. Déterminez les variables dépendantes et indépendantes pour chacune des hypothèses suivantes :

   a. Les individus présentant un score élevé au test de Q.I ont de meilleurs résultats aux tests de mathématiques que ceux qui présentent un faible score.

   b. Les femmes ont une meilleure perception du langage non verbal que les hommes.

   c. Le taux de naissance d'un pays dépend de sa situation socio-économique.

   d. Le taux d'absentéisme chez les opérateurs est influencé positivement par la présence de conditions de travail difficiles.

10. Des chercheurs s'intéressent à l'obésité chez les jeunes.

   a. Donnez deux exemples de variables (et leur codage) quantitatives[3] et qualitatives[4] qui pourraient être utilisées dans le cadre de cette étude.

   b. Donnez deux exemples de variables indépendantes et dépendantes qui pourraient être utilisées dans le cadre de cette étude.

11. Lors de l'élaboration d'une hypothèse, pourquoi est-il utile de se limiter à un nombre restreint de variables qui semblent particulièrement pertinentes ? (2 éléments de réponse)

[3] Une variable qualitative est une variable catégorielle.

[4]Une variable quantitative est une variable dont les valeurs sont des valeurs numériques issues d'un processus de mesure ou d'un processus de dénombrement (comptage) (ex : taille, poids, nombre d'enfants, …).

T.P. 1 : CHAPITRE 2

Exercice 6 : Hypothèses - Théorie

Les exercices suivants sont à faire en exercice supplémentaire, avec à côté de vous votre cours. Il s'agit de trouver les réponses telles quelles dans le syllabus.

1. Donnez quatre propriétés d'une hypothèse ?

2. A quoi correspond une hypothèse infalsifiable ?

3. Une hypothèse n'est pas une question mais bien .....

4. Quelle est la différence entre une hypothèse théorique et une hypothèse opérationnelle ?

5. Est-il possible, en psychologie, de créer un modèle théorique qui correspondrait exactement à la réalité (c'est-à-dire qu'en l'utilisant, on ne commettrait aucune erreur) ? Justifiez votre réponse.

T.P. 1 : CHAPITRE 2

Exercice 7 : Hypothèses - Exercices

1. Donnez une hypothèse théorique et trois hypothèses opérationnelles qui reflètent cette même hypothèse théorique.

2. Formulez des hypothèses de recherche à partir des questions de recherches suivantes.

    a. Y a-t-il une plus grande prévalence de maladies dépressives chez les personnes subissant quotidiennement des agressions verbales ?
    b. Les personnes optimistes sont-elles plus susceptibles d'obtenir une promotion que les personnes pessimistes ?

c. Les hommes possèdent-ils une plus grande capacité à s'orienter que les femmes ?

d. Les capacités cognitives des personnes diminuent-elles avec l'âge ?

3. Soit, l'affirmation suivante :

Dans la continuité des études qui ont démontré que le partage social des émotions a des effets bénéfiques sur le bien-être physique du locuteur, Véronique Christophe et Jean-Pierre Di Giacomo (2003)[5] ont tenté de tester l'impact des réactions de l'auditeur sur le locuteur. Il en est notamment ressorti que les réactions de l'auditeur centrées sur lui-même sont perçues négativement, alors que celles centrées sur le locuteur sont perçus positivement par le locuteur.

a. Sur base de cette affirmation veuillez formuler une hypothèse valide.

b. Sur base de cette affirmation, veuillez déterminer la (ou les) variables dépendante(s) et indépendante(s).

c. Sur base de quel outil cette affirmation a-t-elle été générée ? Théorie ou intuition ?

[5] Véronique Christophe et Jean-Pierre Di Giacomo (2003), « Est-il toujours bénéfique de partager ses expériences émotionnelles ? Rôle du partenaire dans les situations de partage social des émotions », *Revue internationale de psychologie sociale , t. xvi, n° 2*.

P A R T I E I I

PROBABILITE

CHAPITRE 3 : PROBABILITES ET ANALYSE COMBINATOIRE

3.1. Objectifs

Ce chapitre vise à établir les quelques bases de calcul de probabilités dont vous aurez besoin ultérieurement pour comprendre de nombreux tests statistiques. Il s'agit essentiellement de l'indépendance des événements et des probabilités conditionnelles. Je vous enjoins donc à être particulièrement attentifs à ces points. Bien que, en dehors de ce cours, vous n'aurez sans doute jamais à calculer de probabilités, ce chapitre est à la base de la logique statistique. C'est ce qui lui donne toute son importance. Par ailleurs, l'analyse combinatoire est également abordée. Elle permet de dénombrer les événements par un ensemble de règles mathématiques reposant sur deux grands critères que nous envisagerons au point 3.5.

3.2. Bref historique

A la suite de Pascal et Fermat, en 1657, Huygens a écrit le premier manuel de probabilités. En 1662, les derniers chapitres de la logique de Port Royal de Arnauld et Nicole traitent aussi des probabilités et de la manière dont elles peuvent être utilisées pour justifier des croyances et des décisions prises dans l'incertitude. Avec la publication posthume en 1713 de l'*Ars conjectandi* (l'art de conjecturer) de Jacques Bernoulli (1654-1705) et la publication en 1718 de la première édition de "*The doctrine of chance, or a method for calculating the probability of events*" de Abraham De Moivre (1667-1754), la théorie des probabilités devient une branche des mathématiques. Il a donc fallu seulement une soixantaine d'années pour que les principaux concepts de la théorie des probabilités soient développés. Toutefois, l'axiomatisation de la théorie des probabilités ne sera accomplie que deux siècles plus tard par Kolmogorov en 1933. Ces probabilités, dont les axiomes de Kolmogorov, feront l'objet de ce chapitre.

3.3. Définition de la notion de probabilité

On a vu à la section précédente qu'il est possible d'établir une relation entre la théorie des ensembles et la logique. Dans cette section, on exploite le fait qu'il est possible aussi d'établir une relation entre la théorie des probabilités et la théorie des ensembles.

3.3.1. Dualité de la notion de probabilité

La notion de probabilité a toujours eu deux sens différents. Au **sens épistémique** (c'est-à-dire relatif à la connaissance), la probabilité reflète le degré de croyance en des propositions qui n'ont rien de statistiques. C'est par exemple le cas d'un prévenu que l'on va déclarer coupable ou acquitter. Bien que, lorsqu'une personne soit prise en flagrant délit, que l'on retrouve ses empreintes sur l'arme du crime et son ADN un peu partout, on ait tendance à estimer la probabilité qu'il ait commis le crime comme étant égale à 1, nous n'avons en fait aucun moyen mathématique de l'estimer réellement.

Au **sens fréquentiste**, la probabilité reflète la tendance de certains dispositifs aléatoires à produire des événements avec des fréquences relatives[6] qui tendent à se stabiliser au fur et à mesure qu'on augmente le nombre de répétitions de l'expérience aléatoire. C'est le cas d'une pièce de monnaie dont la fréquence du nombre de "*face*" se stabilisera autour des 50% des occurrences au fur et à mesure du nombre de jets (pour peu qu'elle soit bien équilibrée).

Nécessité de l'incertitude dans l'approche probabiliste

Par contre, pour que l'on puisse parler de probabilité, que ce soit au sens fréquentiste ou au sens épistémique, vous devriez maintenant avoir compris qu'il est nécessaire d'avoir un degré d'incertitude possible. Si tous les événements étaient certains, il ne serait plus nécessaire d'envisager les probabilités.

Prenons trois exemples de propositions et interrogeons-nous sur le degré avec lequel nous croyons à leur véracité :

[6] Nous verrons qu'une fréquence relative représente un nombre d'observations intéressantes divisé par le nombre total d'observations (point 3.3.8.1).

Proposition 1. « *Tous les Hommes sont mortels*[7] ».

La plupart d'entre nous attribuerions sans doute une probabilité subjective de 1 à cette proposition, exprimant ainsi la certitude que cette proposition est vraie. Pourtant, comme nous l'avons déjà souvent soulevé, il s'agit d'une inférence inductive présentée sous la forme d'une proposition universelle affirmative valable en tout lieu et en tout temps. Or, seul un nombre limité (mais très grand) de cas ont été observés. Toutefois, même s'il existe effectivement un problème logique de l'induction, psychologiquement notre degré de certitude dans la véracité de cette proposition est absolu.

Proposition 2. "*Le réchauffement climatique actuel est dû à l'activité humaine*"

Admettons qu'il n'y ait plus de doute sur le fait que nous vivons actuellement dans une période de réchauffement climatique ; il reste des doutes sur le degré avec lequel l'activité humaine est responsable de ce phénomène. Rappelons qu'il n'est fait aucune mention de la notion de probabilité dans la proposition 2. Il s'agit d'une assertion vraie (le réchauffement est dû à l'activité humaine) ou fausse (le réchauffement est dû à d'autres causes). Mais il est possible de reformuler la proposition 2 sous la forme du jugement incluant la notion de probabilité et même une estimation chiffrée de celle-ci.

Proposition 2' (jugement probabiliste). A la lumière de ce que l'on sait, la probabilité que « *le réchauffement climatique actuel soit dû à l'activité humaine* » est de .90.

Il reste à se demander comment une personne serait capable d'établir un tel niveau de probabilité. Ces différentes propositions sont des probabilités au sens épistémique du terme.

Si maintenant nous contrastons la proposition 2' à la proposition 3 qui comprend aussi une mention chiffrée d'une probabilité :

[7] Si vous en avez assez de cette proposition, imaginez-en une autre, par exemple "*il y a autant de particules de matière que de particules d'anti-matière*" (théorie décrite, entre autres, par Jean-Pierre Petit, un physicien très amusant qui, après avoir réalisé une carrière scientifique brillante, a prétendu avoir été inspiré par des représentants extra-terrestres venus de la planète Ummo).

Proposition 3 (jugement probabiliste). *"Cette pièce privilégie l'événement face dont la probabilité d'apparition est de .6".*

Nous voyons que la proposition 2 et la proposition 3 sont des jugements qui ont quelque chose en commun : ils disent quelque chose de vrai ou de faux à propos du monde indépendamment de ce que l'on sait de la pièce ou du climat. Mais la proposition 3 est un jugement probabiliste qui fait implicitement référence à une expérience aléatoire consistant à lancer un grand nombre de fois la pièce. Si cette proposition est vraie, c'est que la pièce est n'est pas équilibrée ou que le dispositif de lancer induit un biais. On peut tester cette hypothèse, par exemple, en vérifiant l'homogénéité de la pièce ou en lançant, par exemple, 100 fois la pièce. Si la fréquence relative de l'événement face est de .63, on aura tendance à croire que la proposition 3 est vraie. Il s'agit d'un jugement factuel qui est indépendant d'un degré quelconque de croyance. Ce jugement peut être erroné. Il pourrait provenir d'une estimation fondée sur un nombre trop petit de lancers de la pièce. Ce jugement peut aussi être correct. Il pourrait se confirmer suite à un très grand nombre de lancers de la pièce. La probabilité au sens fréquentiste semble exprimer quelque chose non pas concernant le degré de croyance en une proposition mais concernant une propriété physique objective de la pièce. L'usage du mot probabilité au sens fréquentiste est lié à des notions telles que : fréquence, disposition, tendance, propension. On dira que, soumise à un grand nombre de lancers, la pièce produit l'événement face avec une fréquence relative proche de .60 ou qu'elle a une disposition, une tendance ou une propension à produire l'événement face dans 60% des cas.

En revanche, la proposition 2' ne peut pas être vérifiée en répétant plusieurs fois une expérience aléatoire puisque le réchauffement climatique dont il est question a lieu une seule fois (maintenant, du moins pour les causes invoquées). La proposition 2 est une proposition disant quelque chose sur les causes du réchauffement climatique actuel. La proposition 2' est un jugement qui indique le degré avec lequel quelqu'un croit que cette proposition est correcte, compte tenu des informations disponibles. Cette notion de la probabilité est donc associée à des notions telles que : croyance, confiance, crédibilité. La probabilité de .90 attribuée à la croyance, la confiance ou la crédibilité de la cause humaine du réchauffement climatique porte donc sur la proposition 2', pas sur le fait du monde exprimé dans la proposition 2. La probabilité au sens épistémique a un côté subjectif et personnel. Par exemple, un juré pourrait croire que les éléments de preuve dont il dispose font qu'il y a une probabilité de .70 que l'accusé soit coupable. Un autre juré pourrait évaluer

cette probabilité à .65, un autre encore à .80. Le premier juré pourrait revoir son jugement et réévaluer cette probabilité à .75 suite aux nouvelles informations fournies.

La probabilité au sens fréquentiste a un côté plus objectif. C'est de cette probabilité dont il sera surtout question dans ce qui suit. Toutefois, je tenterai de vous montrer qu'on oscille souvent entre une interprétation épistémique et une interprétation fréquentiste de la probabilité.

3.3.2. Expérience aléatoire et événement aléatoire

Comprendre les notions d'événement et d'expérience aléatoires sont fondamentales pour la suite du cours. Nous utiliserons souvent ces termes et il est important de les comprendre pour s'approprier la logique des tests statistiques usuels et, de là, mettre au point les méthodologies que vous utiliserez dans vos expériences (si vous ne parvenez pas à réaliser des expériences aléatoires, vous ne pourrez analyser vos résultats).

On qualifie d'**expérience aléatoire une action ou un processus qui lors de chaque répétition engendre un et un seul événement élémentaire parmi un ensemble d'événements élémentaires possibles**. L'expérience est qualifiée d'aléatoire parce qu'il existe un certain degré d'incertitude quant à l'occurrence de chaque événement élémentaire possible lors de chaque essai. Chaque événement possible est qualifié d'aléatoire parce qu'il peut se réaliser ou ne pas se réaliser lors de chaque essai (= chaque répétition de l'expérience aléatoire). Un **événement aléatoire est donc un événement qui peut se réaliser lors d'une expérience aléatoire**. L'ensemble des événements élémentaires possibles auxquels on s'intéresse s'appelle espace-échantillon. Par exemple, lorsqu'on lance une pièce de monnaie, on effectue une expérience aléatoire à l'issue de laquelle la pièce tombera soit sur le côté "*pile*" soit sur le côté "*face*". L'événement "*côté face*"[8] est un événement aléatoire parce qu'il peut se réaliser ("*c'est face!*") ou pas ("*c'est pile!*"). L'espace-échantillon contient l'ensemble des événements possibles (c'est-à-dire pile et face, si l'on néglige la tranche), c'est-à-dire l'ensemble universel, en termes ensemblistes.

[8] Lorsque l'on s'intéresse à un événement particulier, dans un contexte donné, par exemple ici la probabilité que la pièce tombe du côté face, on qualifie cet événement de "critique", c'est-à-dire celui qui nous intéresse

Un événement peut être élémentaire ou composé, selon le cas. Un événement élémentaire est un événement indécomposable. L'espace-échantillon contient l'inventaire exhaustif[9] de tous les événements élémentaires observables lors d'une expérience aléatoire. Or, on peut considérer chacun de ces événements différents comme l'élément unique d'un sous-ensemble particulier de l'espace-échantillon. Chaque sous-ensemble ne contient qu'un élément et les différents sous-ensembles sont mutuellement exclusifs. Comme l'inventaire des événements élémentaires est exhaustif, on peut dire que les sous-ensembles d'événements élémentaires constituent une partition de l'espace-échantillon.

Le Tableau 3.1 montre quatre exemples d'expériences aléatoires. Dans chaque expérience, on s'intéresse à un événement élémentaire défini comme critique (par exemple, obtenir pile lors d'un lancer d'une pièce ou obtenir le 6 lors d'un jet d'un dé). L'espace-échantillon est chaque fois défini comme l'extension d'un ensemble ; donc : {pile, face} et {1, 2, 3, 4, 5, 6}.

Tableau 3.1. : Quatre exemples d'expériences aléatoires basées sur des événements élémentaires mutuellement exclusifs et exhaustifs.

Expérience aléatoire	**Evénement critique**	**Espace-échantillon**
Lancer une pièce	Obtenir *pile*	{pile, face}
Lancer un dé	Obtenir le 6	{1, 2, 3, 4, 5, 6}
Lancer un astragale à faces colorées	Obtenir la *face rouge*	{rouge, bleu, noir, non colorée}
Prélever un étudiant au hasard dans l'auditoire	Obtenir un étudiant de *groupe sanguin O*	{A, B, AB, O}

Un événement composé est un événement critique plus complexe défini à partir de plusieurs événements simples. Par exemple, au lieu de s'intéresser à l'événement critique « *obtenir le 6 lors d'un jet d'un dé* », on peut s'intéresser à l'événement critique composé « *obtenir un nombre pair lors d'un jet d'un dé* ». Cet événement composé se produit chaque fois qu'on observe un résultat appartenant au sous-ensemble de trois éléments {2, 4, 6} qui constitue un sous-ensemble de l'espace-échantillon {1, 2, 3, 4, 5, 6}.

[9] On entend par "exhaustif" que toutes les possibilités sont énoncées. Nous en reparlerons.

La question qu'on se pose est : "*comment peut-on déterminer la probabilité d'un événement simple ou composé considéré comme critique dans chaque cas*?" Voyons le cas des événements simples représentés au Tableau 3.1 :

Dans les deux premiers cas du Tableau 3.1 – obtenir l'événement pile ou obtenir l'événement 6 - la réponse à la question peut être fournie de deux manières : soit on procède analytiquement (on se dit qu'il y a une chance sur six de tirer le chiffre 6 à partir d'un dé usuel bien équilibré) et on détermine *a priori*, la probabilité de l'événement sans même se livrer une seule fois à l'expérience aléatoire. Soit on procède empiriquement (on lance un certain nombre de fois le dé) et on détermine *a posteriori* la probabilité, de l'événement en enregistrant sa fréquence relative d'apparition après un grand nombre d'essais (plus ce nombre est grand, plus on aura confiance en la probabilité obtenue).

Dans les deux derniers cas du Tableau 3.1 – obtenir l'événement face rouge ou obtenir un étudiant de groupe O – la seule possibilité est de procéder empiriquement. En effet, il n'existe pas de standard concernant la population concernée : Même s'il existe 4 faces à un astragale (l'os du talon), l'apparition de chaque face n'est pas équiprobable et dépend, dans une certaine mesure, de l'astragale concerné. De la même manière, concernant le groupe sanguin, la proportion de chaque groupe n'est pas fixe au sein de toutes les populations et nous ne connaissons pas, *a priori*, les chances d'être de l'un ou l'autre groupe sanguin dans la population d'étudiants de BA1 en psychologie.

3.3.3. Définition classique (= analytique = a priori) de la probabilité

La probabilité d'obtenir un événement critique A, au sens classique

Est définie comme le rapport entre le nombre de cas favorables et le nombre de cas possibles.

$$P(A) = \frac{Nombre\_de\_cas\_favorables}{Nombre\_de\_cas\_possibles} = \frac{n(A)}{N}$$

Notation : Remarquez que j'ai utilisé un "*n*" minuscule pour le nombre de cas favorables mais un "*N*" majuscule pour le nombre de cas possibles. Par convention, nous aurons l'habitude d'utiliser *n* pour désigner l'effectif d'une partie de l'espace-échantillon et *N* pour l'ensemble des événements de l'espace-échantillon. Par extension, plus tard, nous utiliserons *n* pour l'effectif d'un échantillon et *N* pour l'effectif d'une population, nous y reviendrons en temps voulu.

Cette définition ne peut s'appliquer que lorsque les événements sont équiprobables, mutuellement exclusifs et exhaustifs. Dans les cas particuliers des lancers de la pièce et du dé, il est possible de déterminer les probabilités d'occurrence de chaque événement élémentaire en se livrant à une analyse de la situation *a priori*. Moyennant l'hypothèse que la pièce et le dé sont chacun faits d'un matériau homogène et qu'ils sont chacun parfaitement symétriques, les différents événements élémentaires devraient tous avoir la même probabilité de se produire. Donc, la probabilité d'obtenir l'événement pile lors d'un lancer de la pièce est de 1/2 et celle d'obtenir l'événement 6 lors d'un lancer du dé est de 1/6. De même, le fait d'obtenir un 6 rend tout autre résultat impossible (on ne peut avoir à la fois un 6 et un 3 en lançant une fois le dé), ce qui signifie que deux événements sont mutuellement exclusifs. Enfin, nous connaissons tous les événements possibles d'un jet de dé c'est-à-dire : {1, 2, 3, 4, 5, 6}, ce qui correspond à l'exigence d'exhaustivité.

Il existe des cas où ces exigences ne sont pas remplies. Imaginons par exemple un meurtre commis dans un train décrit par Aghata Christie ("*le crime de l'Orient-Express*"). La victime a été percée de nombreux coups de couteaux. Un enquêteur (Hercule Poirot) considère les suspects potentiels. D'une part, tous les passagers ne sont pas équiprobablement coupables.

Par exemple, le directeur de la compagnie des wagons-lits (Mr Bouc) n'était pas dans le même wagon-lit que la victime ce qui rend le crime difficile à commettre. De plus, si la femme de la victime (La comtesse Andrenyi) s'avérait coupable, cela n'empêcherait pas, par exemple, le valet de la victime (Edward Masterman) d'être également coupable. Enfin, il n'est pas impossible qu'il y ait un certain nombre de passagers clandestins dont Hercule Poirot n'aurait pas connaissance et qui empêcherait l'exhaustivité de l'espace-échantillon (bien qu'il soit, en fait, exhaustif, mais que l'information soit inconnue).

Ces trois conditions, rendant utilisable la définition proposée, posent un réel problème. En effet, exiger l'équiprobabilité équivaut à introduire la notion de probabilité dans la définition de ce même concept. Cela en fait une définition circulaire impropre à la consommation. Elle a cependant l'avantage d'être facile à comprendre et très imagée. Utilisez-là donc pour vous représenter les concepts, mais ne vous en contentez pas et soyez conscients de ses limites.

Si l'on revient à l'un de nos exemples (les dés) compatibles avec la définition proposée, la probabilité d'obtenir des événements composés de plusieurs événements élémentaires peut également s'envisager. Par exemple, avec un dé, la probabilité d'obtenir un nombre pair est de 3/6 = 1/2 puisque ceci peut se réaliser dans le sous-ensemble {2, 4, 6} d'événements élémentaires. La probabilité d'obtenir un nombre impair est aussi de 3/6 = 1/2 puisque ceci peut se réaliser dans le sous-ensemble {1, 3, 5} d'événements élémentaires. La probabilité d'obtenir un nombre divisible par 3 est 2/6 = 1/3 puisqu'il y a deux éléments dans le sous-ensemble {3, 6} qui correspondent à cette définition.

Rappelons le commentaire du chapitre précédent : le rapport entre le nombre d'événements favorables à la réalisation de l'événement considéré comme critique et le nombre total d'événements élémentaires possibles est une fréquence relative, c'est-à-dire un nombre variant entre 0 (aucun cas possible n'est favorable) et 1 (tous les cas possibles sont favorables).

3.3.4. Définition empirique de la probabilité (a posteriori)

Voyons le cas de l'astragale de mouton (ci-contre). Ian Hacking s'est livré à 300 lancers d'un tel astragale trouvé sur un site archéologique turc et datant d'environ 6000 ans. Trois des faces de l'astragale étaient colorées. Sur 300 lancers, Hacking a obtenu 50 fois la face rouge, 88 fois la bleue, 52 fois la face noire et 110 fois la face non colorée. Ceci donne une fréquence relative de 0,17 pour la face rouge (50/300). Cette fréquence relative observée constitue la meilleure estimation de la probabilité d'obtenir la face rouge avec cet astragale si on se limite aux 300 lancers effectués par Hacking. On obtiendrait une meilleure estimation de cette probabilité en se livrant à 1.000 ou à 10.000 lancers (avis aux amateurs).

Cet exemple montre que ce processus conduit à une définition de la probabilité comme étant la limite de la fréquence relative pour un nombre infini de lancers. Cette idéalisation de la notion de fréquence relative observée s'appelle **fréquence relative théorique**.

<u>Probabilité au sens fréquentiste</u>

Limite de la fréquence relative au nombre d'expériences aléatoires égal à l'infini.

Comme évoqué plus haut, on peut se livrer à l'étude empirique de la probabilité d'obtenir l'événement pile après un grand nombre de lancers d'une même pièce de monnaie. Après 10.000 lancers d'une même pièce, Kerrich[10] a obtenu 5067 fois l'événement élémentaire pile, soit une fréquence relative de 0,5067. Buffon[11] a observé 2048 fois pile sur 4040 lancers d'une pièce, soit une fréquence relative de 0,5069. Avec 24.000 lancers, Karl Pearson (qui a eu une vie très intéressante par ailleurs) a obtenu une fréquence relative de 12.012/24.000 = 0,5005 (ce n'étaient cependant pas les mêmes pièces).

[10] John Kerrich est un mathématicien sud-africain qui a réalisé ses 10000 lancers lors de son incarcération au Danemark durant la seconde guerre mondiale.

[11] Georges-Louis Leclerc de Buffon est l'un des grands noms parmi les scientifiques du 18ème siècle, il s'est intéressé à de nombreux domaines (comme beaucoup de scientifiques de l'époque, la science n'étant qu'à l'aube de son expansion). Il était naturaliste, mathématicien, biologiste, astronome, et écrivain (dont le célèbre ouvrage en trois tomes, "*Histoire naturelle, générale et particulière, avec la description du cabinet du Roy*", est sa pièce maîtresse)

L'avantage de la définition au sens fréquentiste est multiple. D'une part, elle n'est plus circulaire. D'autre part, elle ne demande pas la condition d'équiprobabilité ni celle de l'exhaustivité. En effet, si l'on devait répéter un grand nombre de fois une expérience aléatoire pour observer la fréquence d'occurrence d'un événement critique, peu importe s'il existe un certain nombre d'événements inconnus qui pourraient survenir durant l'expérience aléatoire. Le désavantage de cette définition est qu'elle nous contraint à réaliser des expériences aléatoires pour déterminer la probabilité. On perd donc la possibilité d'évaluer une probabilité *a priori*.

3.3.5. Probabilité au sens empirique et loi des grands nombres

Une manière un peu différente d'exprimer la définition de la probabilité au sens fréquentiste du terme nous est donnée par Bernouilli, sous la désignation de Loi des grands nombres :

Loi des grands nombres

Si la probabilité d'un événement X est $P(X)$ et que l'expérience aléatoire est répétée N fois, chaque essai étant indépendant des autres, la probabilité que la fréquence relative d'apparition de X diffère de $P(X)$ d'une quantité aussi faible que l'on veut tend vers o quand le nombre d'essais N tend vers l'infini.

Cette loi des grands nombres est souvent mal comprise parce que la notion d'indépendance est mal comprise. Cette incompréhension se manifeste notamment par ce qu'on qualifie "*d'illusion du joueur*" (ou de "*sophisme du joueur*") Par exemple, à la roulette, beaucoup de joueurs pensent qu'après un certain nombre de répétitions de la même couleur, la probabilité d'apparition de la couleur opposée augmente. Une autre manière d'envisager cette illusion est de croire que la loi des grands nombres implique une sorte de compensation de la part de la nature : tout se passerait comme si la nature s'arrangeait pour que les fréquences absolues des différents événements possibles s'équilibrent après un grand nombre d'essais.

Pour comprendre la notion d'essais indépendants, nous envisagerons le modèle de l'urne : Imaginons que nous voulions estimer la probabilité de prélever un étudiant de groupe

sanguin O dans un auditoire contenant *N* étudiants. Vous me direz que c'est assez simple, il me suffit de demander aux étudiants de groupe sanguin O de lever le doigt et de diviser ce nombre par le nombre d'étudiants présents dans l'auditoire. C'est exact, et cela correspond à la manière de faire conformément à la définition *a priori* de la probabilité, mais si je cherchais à avoir une vie facile je n'enseignerais pas les statistiques. Nous allons plutôt recourir au modèle (infiniment plus drôle) du tirage aléatoire avec replacement dans l'urne : chaque étudiant est représenté par un petit carton sur lequel son groupe sanguin est indiqué, on peut déterminer la fréquence relative de l'événement groupe O après un grand nombre de prélèvements aléatoires d'un carton, chaque carton prélevé étant replacé dans l'urne avant le prélèvement suivant. Au bout d'un nombre infini d'essais, j'aurai la probabilité exacte de l'événement critique "*choisir un étudiant de groupe sanguin O*". Mais ce qui m'intéresse dans cet exemple est que la probabilité de prélever une personne du groupe O reste la même, d'essai en essai, exactement comme la probabilité d'obtenir le 6 reste la même, d'essai en essai puisque le dé garde ses six faces après chaque essai. On dira qu'il y a **indépendance** entre les événements observés lors de chaque essai.

La formulation de la loi des grands nombres qui figure dans l'encadré ci-dessus n'est pas très claire parce que j'essaie de traduire en français une expression mathématique du théorème. En voici une explication plus analytique tirée de la troisième édition du livre *Statistics* (1997, pp. 273-276) de Freedman, Pisani et Purves.

Freedman et al. (1997) imaginent le dialogue que Kerrich aurait pu avoir avec un assistant pour préparer sa visite au roi du Danemark après sa libération à l'issue de la 2ème Guerre Mondiale. Vous vous en doutez déjà : Kerrich a bien compris la loi des grands nombres, l'assistant n'a rien compris du tout, ne vous efforcez donc pas de retenir son discours, comprenez juste les implications. Ici, on s'intéresse à l'analyse des résultats des 10000 jets d'une même pièce de monnaie, dont vous savez déjà qu'elle a fourni un nombre absolu de 5067 fois l'événement face, soit cet événement avec une fréquence relative de 0,5067, donc une estimation de la probabilité de .5067

Voici le résumé du scénario imaginé par Freedman et al. : Kerrich informe l'assistant qu'il a l'intention d'expliquer au roi la signification de la loi des grands nombres. Le dialogue commence avec l'étonnement de l'assistant qui voit mal l'intérêt d'expliquer au roi du Danemark une loi que tout le monde comprend. Invité par Kerrich à expliquer ce que veut dire cette loi des grands nombres, l'assistant fournit la réponse suivante.

Loi des grands nombres selon l'assistant obtus : supposons qu'on lance une pièce. Si on obtient beaucoup de fois l'événement face successivement, alors l'événement pile se met à apparaître. Ou, si on obtient trop de fois l'événement pile, les chances d'obtenir l'événement face augmentent. A la longue, les nombres d'événements pile et d'événements face s'équilibrent.

Dans la suite du dialogue, Kerrich met en évidence les deux erreurs commises par l'assistant dans sa conception de la loi des grands nombres :

a) Erreur 1 : Confusion à propos de la notion d'indépendance entre les essais. Il s'agit de la croyance erronée que la probabilité d'obtenir l'événement face est plus grande après, par exemple, une séquence de 10 événements pile qu'après une séquence de 2 événements pile. Pour éviter de faire cette erreur je vous enjoins à retenir une règle d'or en probabilité :

<u>Les probabilités n'ont pas de mémoire!</u>

Si, avant de lancer votre pièce une première fois, vous vous interrogez sur la probabilité d'obtenir 20 fois "pile" d'affilée, cette probabilité sera évidemment très faible. En revanche, si vous venez d'obtenir 19 fois pile, la probabilité d'avoir pile au prochain lancement de la pièce est de 0,5 et ce indépendamment de ce qui s'est déjà passé.

Kerrich essaie de convaincre l'assistant de l'inanité de sa croyance en lui montrant un exemple tiré de ses données. Au cours des 2000 premiers essais, il a observé 130 fois une séquence de quatre événements face consécutifs. A l'essai suivant, il a obtenu l'événement face dans 69 cas et l'événement pile dans 61 cas. Ce résultat va légèrement dans le sens opposé à celui prédit par l'assistant.

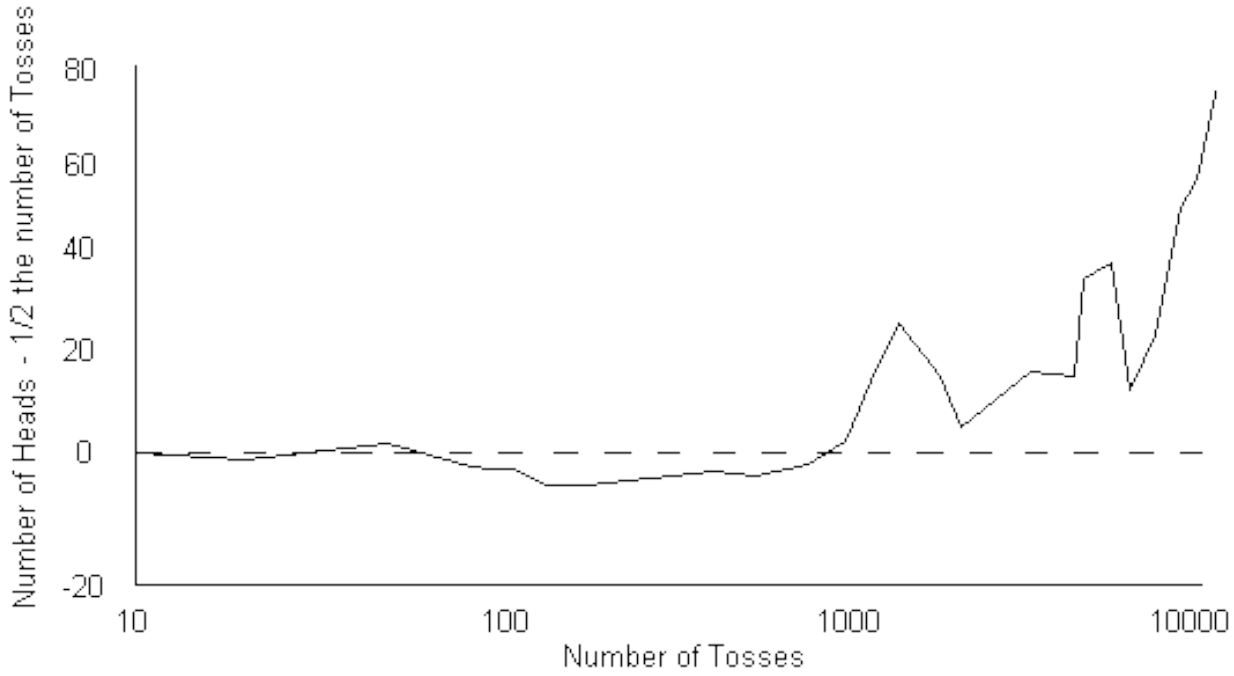
L'assistant reste sceptique, surtout que Kerrich a ajouté qu'il avait obtenu 5067 fois l'événement face sur les 10000 jets, soit un excès de 67 par rapport aux 5000 attendus. Kerrich précise que la différence absolue de 67 représente une différence de moins de 1% par rapport au .50 attendu (5067/10000 = .5067 et .5067 - .5000 = .0067). L'assistant trouve cependant que cet excès est trop grand et que le roi du Danemark ne sera pas impressionné si tout ce que la loi des grands nombres peut faire, c'est d'aboutir à une erreur absolue aussi

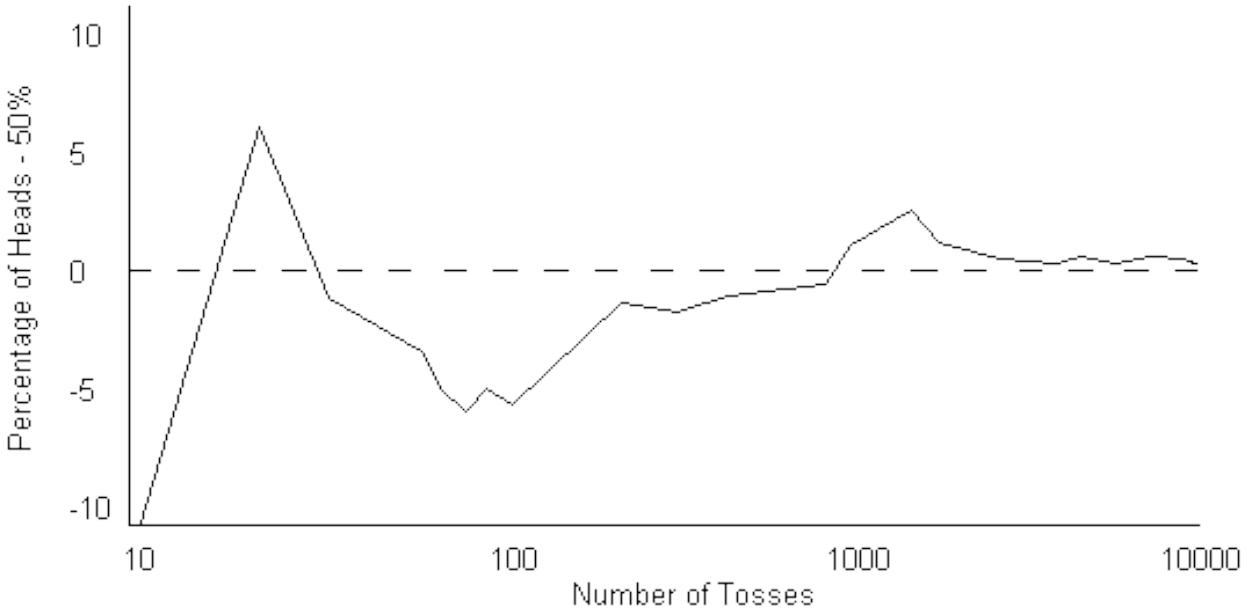
grande (67) par rapport à l'idéal de 5000. Il conclut que les 10.000 essais sont sans doute insuffisants pour la démonstration et il suggère que si on en ajoutait 10.000, le nombre d'événements pile et d'événements face s'équilibreraient beaucoup mieux (donc, que l'excès de l'un ou de l'autre par rapport à 20000/2 = 10000 serait plus faible que l'excès de 67 obtenu par Kerrich par rapport au 10000/2 = 5.000).

b) Erreur 2 : Cette confusion n'est pas indépendante de la première. Elle consiste à croire que la nature fonctionne selon un principe de compensation qui tend à rééquilibrer les fréquences absolues des deux événements au fur et à mesure que les essais progressent. Selon cette interprétation de l'assistant, le processus aléatoire fait que localement, il peut y avoir une dérive qui rend trop fréquent un des deux événements supposés équiprobables, mais la loi des grands nombres fonctionne comme un processus de compensation qui ramène l'équilibre. Donc, après un excès d'événements pile, la loi fait qu'un plus grand nombre d'événements face sera observé dans la suite de la série.

Kerrich utilise les Figures 3.1a et 3.1b pour montrer que cet argument est fallacieux. La Figure 3.1a montre que la tendance générale est celle d'une augmentation de la différence entre le nombre de fois que l'événement "face" apparaît et la moitié du nombre de jets. On remarque qu'avec les essais ce nombre augmente, même si localement, la différence peut être relativement petite ou relativement grande quel que soit le nombre de jets. En revanche, la Figure 3.1b montre que la différence entre pourcentage d'occurrence de l'événement critique ("*face*"), c'est-à-dire le nombre de fois que face est sorti divisé par le nombre de lancers multiplié par cent, et 50% (le pourcentage que l'on obtiendrait sur un nombre infini de lancers si la pièce était parfaitement équilibrée) tend vers zéro lorsque le nombre de lancers augmente.

Figures 3.1a (au-dessus) et **3.1b** (en-dessous) : Loi des grands nombres. Source : http://iimk.ac.in/gsdl/cgi-bin/library?e=d-000-00---ostatis--00-0-0--oprompt-10---4------0-1l--1-en-50---20-about---00031-001-1-outfZz-8-00&a=d&cl=CL1&d=HASHe00909ac46143070d8f732.4 récupéré le 23/10/11.





Arrivé à ce point, toutes les certitudes de l'assistant s'écroulent et il demande ce qu'est exactement la loi des grands nombres. Voici en substance la réponse de Kerrich : Après un grand nombre de jets, la taille de la différence entre le nombre d'événements face et le nombre attendu peut devenir assez grande en fréquences absolues (Figure 3.1a) mais, comparée au nombre de jets, la différence en fréquences relatives tend à devenir petite (Figure 3.1b).

Donc, plus le nombre de jets augmente, plus la différence entre la fréquence relative de l'événement critique et sa fréquence relative théorique (50%) devient petite. Or, au sens fréquentiste, la fréquence relative théorique est par définition la probabilité d'un événement. Donc, ce qu'illustre la Figure 3.1b, c'est bien que la fréquence relative observée tend à se stabiliser autour de la valeur de la probabilité. L'amplitude des fluctuations autour de cette fréquence relative théorique diminue avec l'augmentation de *N*. La Figure 3.1b est donc une illustration du théorème de Bernoulli énoncé dans l'encadré ci-dessus[12].

On peut reformuler le théorème de Bernouilli (la loi des grands nombres) de la manière suivante : la différence entre la fréquence relative observée et la fréquence relative théorique (= la probabilité) tend vers 0 quand *N* tend vers l'infini.

Notation : Remarquez ici que j'ai utilisé le nombre d'expériences aléatoires par *N*. En effet, même si, théoriquement, on pourrait effectuer cette expérience un nombre infini de fois, en pratique on ne le fait qu'un certain nombre de fois et lorsque l'on s'arrête, on peut considérer que l'on a notre population complète d'expériences aléatoires (puisqu'on n'en fait pas d'autre).

Lors de chaque répétition d'une expérience à la Kerrich, on obtiendrait une estimation différente de la probabilité, donc un écart différent par rapport à la probabilité de l'événement face qui est un nombre fixe. Si on admet que la pièce est parfaite, la probabilité de l'événement face est de .5000. Kerrich trouve .5067 après 10000 essais, soit un écart entre la fréquence relative observée et la probabilité qui est de .0067 (= .5067 - .5000). S'il recommençait son expérience de 10000 jets plusieurs fois, il trouverait chaque fois une

[12] Pour ceux qui s'inquiètent du théorème à étudier, libre à vous de retenir la version que vous préférez, du moment que vous êtes capables de m'expliquer ce que cela veut dire et de l'appliquer.

estimation différente de la probabilité, donc un écart différent entre la fréquence relative observée et la probabilité de .5000.

Avec 10000 essais, les écarts entre les valeurs estimées de la probabilité de l'événement critique "*face*" sont donc distribués autour de la moyenne de 5000. A chaque valeur d'écart est associée une probabilité. Par exemple, l'écart de .0067 trouvé par Kerrich a une certaine probabilité que je ne connais pas mais que je peux noter *P*(fréquence relative observée – fréquence relative théorique = .0067) = ??? et lire "*la probabilité que la différence entre la fréquence relative observée et la fréquence relative théorique soit de .0067 vaut ???*".

Supposons que Kerrich ait ajouté 10.000 essais à son expérience. La probabilité de trouver un écart de .0067 aurait diminué parce que la distribution des écarts possibles par rapport à .5000 est plus faible avec 20000 essais qu'avec 10000. Elle serait encore plus faible avec 30000 essais, etc. Donc, quand *N* tend vers l'infini, la probabilité d'observer un écart de .0067 tend vers 0. Il en va évidemment de même pour les probabilités de n'importe quelle valeur d'écart par rapport à .5000.

Reprenons la formulation de la première version de la loi des grands nombres à la lumière de ce qui vient d'être expliqué.

<u>Loi des grands nombres (reprise)</u>

Si la probabilité d'un événement *X* est *P(X)* [*X* = événement face et *P(X)* = .5 chez Kerrich] et que l'expérience aléatoire est répétée *N* fois [*N* = 10000 chez Kerrich], chaque essai étant indépendant des autres [ce qui est le cas, la probabilité de l'événement face = .5 est constante d'essai en essai], la probabilité que la fréquence relative d'apparition de *X* diffère de *P(X)* d'une quantité aussi faible que l'on veut [disons une quantité ε = .0067 pour prendre comme valeur de l'écart celle trouvée par Kerrich] tend vers 0 quand le nombre d'essais *N* tend vers l'infini.

Notation

Remarquez que j'ai utilisé la notation "ε" pour l'erreur. Cette notation reviendra régulièrement lorsque nous aborderons plus en détails ce concept tout à fait central d'erreurs (nous l'avons déjà utilisé au point 2.5 sans l'écrire sous forme d'équation mathématique).

3.3.6. Définition axiomatique des probabilités

Il est temps maintenant d'établir un certain nombre de lois qui régissent et définissent les probabilités autrement que de manière intuitive. Par "*axiome*", j'entends "*vérité indémontrable qui doit être admise*". Pour pouvoir travailler avec les probabilités, il est nécessaire, comme pour tout cadre de travail, fusse-t-il mathématique, de disposer d'une armature constituée d'un minimum d'affirmations que l'on tient pour vraies et qui supportent l'entièreté des raisonnements qui en découlent. Le principe d'une axiomatisation est que tous les théorèmes de la théorie axiomatisée peuvent être dérivés à partir des axiomes. En général, on procède de proche en proche, avec la démonstration du théorème $N + 1$ qui peut être dérivée, soit en se basant sur les axiomes seulement, soit sur des théorèmes parmi les N déjà démontrés, soit sur un mélange d'axiomes et de théorèmes déjà démontrés. Concernant les probabilités, c'est donc Kolmogorov qui a établi trois affirmations, que nous allons maintenant passer en revue. J'attire votre attention sur le fait que les axiomes d'un cadre théorique sont, presque toujours, évidents en soi. Dès lors, lorsque vous les lirez vous vous direz probablement : "*Evidemment, ça ne pourrait pas être autrement, je me demande pourquoi il le dit!*". La difficulté ne vient pas de ce qui est dit, mais bien des implications qui en découlent. Au fur et à mesure qu'on progresse dans les raisonnements qui s'insèrent sur le cadre axiomatique, la complexité s'installe et n'est possible que parce que les quelques axiomes de base existent. Il est donc nécessaire que vous gardiez ces axiomes en tête de manière à pouvoir identifier en quoi les raisonnements ultérieurs sont possibles. Ne faites pas l'erreur de passer ces quelques pages en vous disant que ce qui est dit est tout à fait évident.

Axiomes de Kolmogorov (1933)

Soit un espace-échantillon Ω associé à une expérience aléatoire. La probabilité $P(E)$ d'un événement critique E consiste en un nombre réel qui satisfait aux axiomes suivants :

Axiome 1 : Pour tout événement E, $P(E) \geq 0$

Axiome 2 : $P(\Omega) = 1$

Axiome 3 : Si E_1, E_2....E_m sont m événements mutuellement exclusifs, alors :

$$P(E_1 \cup E_2 ... \cup E_m) = P(E)_1 + P(E_2)....+ P(E_m)$$

L'axiome 1 implique que la probabilité minimum est de 0. Il s'agit de la probabilité d'un événement impossible, donc de la probabilité de l'ensemble vide. Donc, $P(\varnothing) = 0$. L'axiome 2 dit que la probabilité d'un événement certain est de 1. Lors de chaque essai, on a la certitude qu'un des événements de l'espace-échantillon doit se produire. En combinant les axiomes 1 et 2, on peut dire que la probabilité est un nombre réel qui varie entre 0 et 1. Donc, tout événement E est soit impossible [$P(E) = 0$], soit certain [$P(E) = 1$] soit probable [$P(E)$ est plus grand que 0 et plus petit que 1] (je vous avais prévenus que ça aurait l'air évident).

L'axiome 3 porte sur des **événements mutuellement exclusifs** aussi appelés **événements disjoints** ou **événements incompatibles**. Comme nous l'avons vu, de tels événements ne peuvent se produire simultanément lors d'un essai. Cela équivaut à dire que $P(E_x \cap E_y) = \varnothing$ où x et y sont différents et compris entre 1 et m (où m est le nombre d'événements différents). En pratique ça veut juste dire que je ne sais pas obtenir à la fois 3 et 6 (par exemple) lors d'un jet de dé. Dans la mesure où les événements sont mutuellement exclusifs, l'axiome 3 décrit la **principe d'additivité** des probabilités. Par exemple, les chances de tirer un "1" ou un "2" ou un "3", c'est-à-dire $P(1 \cup 2 \cup 3)$, sur un jet de dé sont de 1/6 + 1/6 +1/6 = 3/6 = 0.5, c'est-à-dire $P(1) + P(2) + P(3)$. Plus loin, nous envisagerons une extension du principe d'additivité aux cas d'événements qui ne sont pas mutuellement exclusifs.

3.3.7. Quelques propriétés des probabilités dérivées des axiomes

3.3.7.1. Propriétés des probabilités pour des événements disjoints et exhaustifs

Dans tous les cas envisagés au Tableau 3.1 (que je présente à nouveau ci-dessous pour vous faciliter la tâche, ne reculant devant aucun sacrifice), les événements élémentaires sont non seulement mutuellement exclusifs mais aussi exhaustifs, c'est-à-dire qu'ils constituent une partition de l'espace-échantillon en un certain nombre de sous-ensembles qui sont tels qu'il n'existe aucun événement élémentaire appartenant à Ω qui n'appartienne pas aussi à un des sous-ensembles. J'illustre ce point avec le cas de la pièce de monnaie en appelant A la réalisation de l'événement pile et B la réalisation de l'événement face.

Tableau 3.1. : Quatre exemples d'expériences aléatoires basées sur des événements élémentaires mutuellement exclusifs et exhaustifs.

Expérience aléatoire	**Evénement critique**	**Espace-échantillon**
Lancer une pièce	Obtenir *pile*	{pile, face}
Lancer un dé	Obtenir le 6	{1, 2, 3, 4, 5, 6}
Lancer un astragale à faces colorées	Obtenir la *face rouge*	{rouge, bleu, noir, non colorée}
Prélever un étudiant au hasard dans l'auditoire	Obtenir un étudiant de *groupe sanguin O*	{A, B, AB, O}

a) Propriété d'une partition (c'est-à-dire du cas où on envisage tous les événements de l'espace-échantillon) :

- On peut dire que $\boldsymbol{P(A \cup B) = P(A) + P(B)}$ par l'axiome 3
- mais aussi que $\boldsymbol{P(A \cup B) = P(A) + P(B) = P(\Omega) = 1}$

b) Propriétés de multiplication

Deux ou plusieurs événements indépendants peuvent néanmoins se produire en même temps. Par exemple, si je jette trois fois un dé, je pourrais obtenir trois fois un 6. Je peux également me poser la question de savoir, avant de jeter trois fois mes dés, quelle est la

probabilité d'obtenir trois fois un 6? La règle de multiplicativité dit qu'il suffit, pour répondre à cette question, de multiplier les probabilités de chaque événement entre elles. Rappelons, en effet, que nous sommes en train de nous poser la question de savoir quelle est la probabilité qu'un 6 soit tiré au premier jet ET qu'un autre soit tiré au deuxième jet ET qu'un autre encore soit tiré lors du dernier jet. Cette liaison "*ET*" correspond à la multiplication (alors que le "*OU*" correspond à la somme). Dès lors, selon cette loi multiplicative (uniquement valable, rappelons-le, lorsque les événements sont indépendants), la probabilité est de 1/6*1/6*1/6 = 1/216.

Loi multiplicative des probabilités

La probabilité d'occurrence conjointe de deux ou plusieurs événements indépendants est égale au produit de leurs probabilités individuelles :

$$P(A \cap B \cap C) = P(A) .P(B) .P(C)$$

c) Propriétés d'événements complémentaires

Soit A un événement dans Ω et ~A, l'événement complémentaire de A (remarquez que, si on néglige l'événement "*la pièce tombe sur la tranche*", B = ~A dans notre exemple, cependant, en pratique, je vous conseille d'utiliser la notation ~A dès que vous parlez du complémentaire, cette notation permet d'ôter tout doute de type "*et la tranche?*").

Donc, si A représente l'événement élémentaire pile, ~A représente le fait que A ne se réalise pas, donc le fait que l'événement face se produise au cours d'un essai. Pour reprendre l'exemple du dé, si A représente l'événement élémentaire 6, ~A représente le fait de ne pas avoir obtenu 6, donc, le fait qu'un des événements élémentaires de l'ensemble {1, 2, 3, 4, 5} se produise, c'est-à-dire, dans ce cas, un événement composé (remarquez que nulle part je n'ai imposé que les éléments soient des éléments simples). Donc, l'événement A et son complément ~A étant mutuellement exclusifs et exhaustifs :

- $\boldsymbol{P(A \cup \sim A) = P(A) + P(\sim A) = 1} \Leftrightarrow P(A) = 1 - P(\sim A) \Leftrightarrow P(\sim A) = 1 - P(A)$.
- En outre, $\boldsymbol{P(A \cap \sim A) = 0}$

Cette propriété est fondamentale : l'utilisation du complémentaire permet de réfléchir à de nombreux problèmes, notamment à tous ceux qui contiennent des expressions du type "*au moins un*". Par exemple, si je vous parle d'une expérience aléatoire de type "*tirer trois cartes d'un paquet de 52 cartes*[13]". Je vous demande ensuite quelle est la probabilité d'obtenir l'événement critique "*au moins une figure*". Cette probabilité est assez difficile à établir parce qu'il faut calculer la probabilité d'avoir une figure, puis celle d'avoir deux figures, puis celle d'avoir trois figures, mais qu'il faut également tenir compte du fait que lorsqu'on a deux figures on en a nécessairement une également (et donc on a déjà envisagé partiellement cette probabilité lors du calcul de la probabilité d'obtenir une figure). Cela conduit à un casse-tête dans lequel on doit veiller à ne compter qu'une seule fois chaque cas sans en oublier un seul. Une manière beaucoup plus simple de résoudre ce problème est de l'envisager à l'aide de l'événement complémentaire. Il est en effet très simple de calculer la probabilité de n'avoir aucune figure durant les trois tirages : il y a 12 figures, parmi les 52 cartes, donc 40 cartes qui ne contiennent pas de figure. Lors du premier tirage il y a donc 40 cas favorables sur 52 cas possibles de ne pas avoir de figure. Lors du deuxième tirage, puisque vous venez de tirer une carte qui n'est pas une figure, il n'y a plus que 39 cartes favorables sur 51 cartes possibles. De même lors du troisième tirage, il ne vous restera que 38 cartes favorables sur 50 cartes possibles. Nous avons vu que la probabilité que ces trois événements indépendants se réalisent consécutivement est égale au produit de ces événements. Cependant, dans ce cas, nous ne sommes pas dans une situation d'indépendance puisqu'après chaque tirage, une carte est ôtée du paquet et change les probabilités d'obtenir une image. Dans ce cas, le calcul devient 40/52 * 39/51 * 38/50 = 38/85 (je ne vous démontre pas ce calcul, mais intuitivement vous devriez pouvoir le reproduire, faites l'exercice). Cependant, la probabilité qui était demandée dans l'énoncé était celle d'avoir au moins une figure. C'est donc le complémentaire de la probabilité que nous venons de calculer. Les propriétés des événements complémentaires nous permettent alors de dire que :

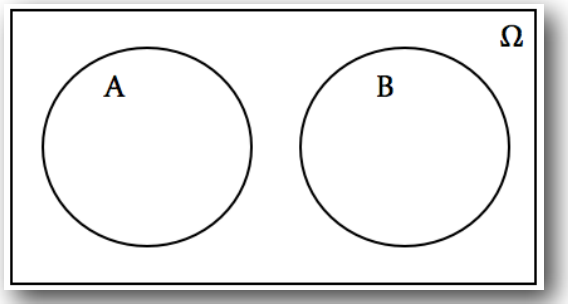
[13] Je me suis rendu compte que certains étudiants ne jouaient pas aux cartes (ce n'est pas une tare). Pour ceux-là, je précise qu'un jeu de cartes habituel (en Belgique) contient 52 cartes : 13 coeurs, 13 carreaux, 13 trèfles et 13 piques. Ces cartes vont du 2 à l'as. Au-dessus du 10, on trouve le valet, la dame et le roi qui sont des figures, il y a donc 12 figures. Les Honneurs sont le valet, la dame, le roi et l'as (qui est souvent une carte privilégiée et très forte dans de nombreux jeux), il y a donc 16 Honneurs. Il y a également deux jokers en supplément des 52 cartes mais qui sont, le plus souvent, écartés du jeu. Ils sont cependant intéressants dans certains exercices de probabilité parce qu'ils n'appartiennent à aucune couleur.

- *P(avoir au moins une figure sur trois tirages) = 1 - P(n'avoir aucune figure sur trois tirages)*
- *P(avoir au moins une figure sur trois tirages) = 1 - 38/85 = 47/85 = .553*

3.3.7.2. Propriétés des probabilités pour des événements non mutuellement exclusifs

Considérons d'abord l'espace-échantillon constitué par les quatre groupes sanguins : {A, B, AB, O} et les événements mutuellement exclusifs (mais non exhaustifs) : A = groupe sanguin A et B = groupe sanguin B (Figure 3.2).

Figure 3.2. : Ω = groupes sanguins ={A, B, AB, O}



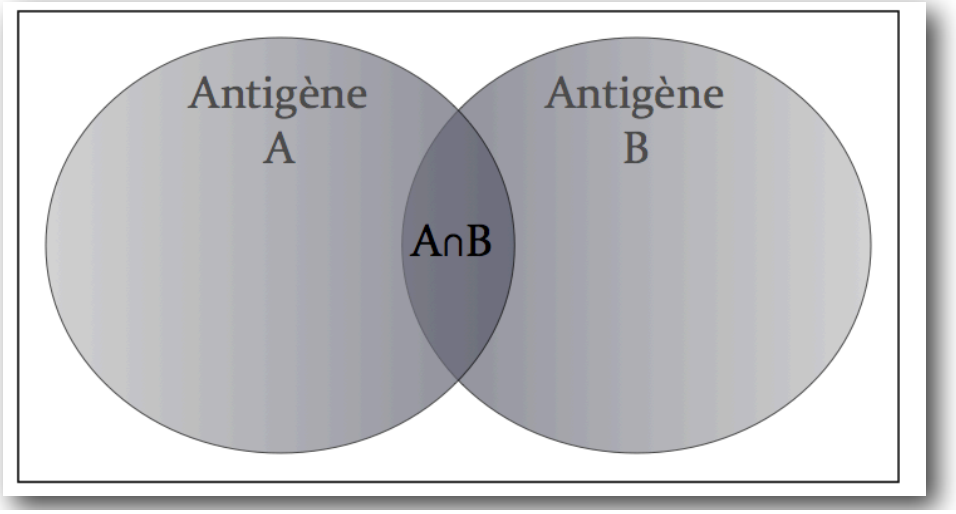
Dans ce cas P(A U B) = P(A) + P(B) par l'axiome 3. Par exemple, si la probabilité du groupe A est de .40 et la probabilité du groupe B est de .20 dans une population, la probabilité de trouver une personne qui appartient soit au groupe A soit au groupe B est de .60. Supposons que la probabilité du groupe AB soit de .30, alors la probabilité de trouver une personne qui appartient soit au groupe A, soit au groupe B, soit au groupe AB est de .90. Les 10% restants des personnes sont nécessairement du groupe O.

En fait, les quatre groupes sanguins résultent du croisement de deux attributs : posséder ou ne pas posséder l'antigène A et posséder ou ne pas posséder l'antigène B (chez l'Homme, l'antigène A, ou B, est une protéine membranaire des globules rouges, lorsqu'un individu n'a ni le A ni le B il est O, qui se prononce d'ailleurs "*zéro*" et pas "*O*" contrairement à la croyance fréquente). Considérons donc ce nouvel espace-échantillon {antigène A, antigène B} constitué de deux attributs qui ne sont pas mutuellement exclusifs (on peut être de groupe sanguin AB). De tels attributs sont aussi qualifiés de compatibles (contrairement à

des attributs disjoints, donc incompatibles, comme synthétiser à la fois des plumes et des poils, par exemple).

En termes de probabilités de possession de l'antigène A, ceci donne donc .70 [car *P(antA)* + *P(antA ∩ antB)* = .40 + .30], celle de posséder l'antigène B est de .50 [car *P(antB* + *P(antA ∩ antB)* = .20 + .30]. Toutefois, la probabilité de trouver une personne qui possède soit l'antigène A, soit l'antigène B ne peut pas être de 1.20. Ceci est dû au fait que *P(antA ∩ antB)* = .30 a été considéré deux fois. Il faut donc soustraire une fois .30 pour obtenir le bon résultat qui est de .90. Cet exemple se comprend très facilement une fois que l'on envisage la situation sous la forme d'un diagramme de Venn (Figure 3.3). Vous voyez qu'en comptant la probabilité A (tout le cercle A) et en ajoutant la probabilité B (tout le cercle B), je compte deux fois l'intersection A∩B (c'est pour ça qu'il est gris foncé). Je dois donc l'enlever une fois pour corriger cette erreur.

Figure 3.3. : Diagramme de Venn représentant des événements non mutuellement exclusifs.



Mathématiquement, dans le cas des événements non mutuellement exclusifs (non disjoints) cela devient :

P (antA U antB) = P(antA) + P(antB) – P(antA ∩ antB) = .70 + .50 – .30 = .90

Cette manière de calculer fonctionne en fait très bien avec des événements mutuellement exclusifs également si ce n'est que dans ce cas la probabilité de l'intersection est nulle. Par extension, on peut établir la règle générale.

Règle générale d'addition des probabilités de deux événements

Si A et B sont deux événements quelconques (mutuellement exclusifs ou pas)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Dans le cas particulier où A et B sont des événements disjoints (Axiome 3)

$$P(A \cup B) = P(A) + P(B)$$

parce que $P(A \cap B) = 0$

La notation ensembliste permet de se faire une idée concrète de la notion de probabilité et des relations entre probabilités de différents événements simples ou composés qui peuvent être définis dans des espaces-échantillons (= ensembles universels) spécifiques. En revanche, cette notation est nettement moins pratique quand on veut indiquer aussi les différentes probabilités associées à ces différents événements. Le recours à un tableau de contingence est beaucoup plus pratique. Ci-dessous, un tableau de contingence partitionné horizontalement en fonction de la possession ou de la non possession de l'antigène A et verticalement en fonction de la possession ou de la non possession de l'antigène B est présenté (Tableau 3.2). Cette représentation est très importante car elle permettra de représenter tous vos plans expérimentaux lorsque vous devrez tester vos hypothèses.

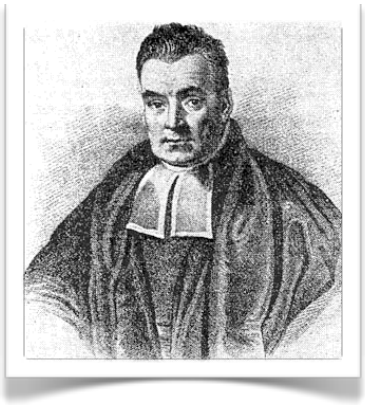
Tableau 3.2. : Tableau de contingence pour deux événements (simples ou composés)

	antB	~antB	Probabilité
antA	P(antA ∩ antB) = 0,30 (groupe AB)	P(antA ∩ ~antB) = 0,40 (groupe A)	$P(antA)$ = 0,70
~antA	P(~antA ∩ antB) = 0,20 (groupe B)	P(~antA ∩ ~antB) = 0,10 (groupe O)	$P(\sim antA)$ = 0,30
Probabilité	$P(antB)$ = 0,50	$P(\sim antB)$ = 0,50	$P(\Omega)$ = 1

3.3.8. Probabilités conditionnelles et indépendance

3.3.8.1. Probabilités conditionnelles

Les probabilités conditionnelles répondent à déterminer la probabilité d'un événement SACHANT qu'un autre événement est présent. Ce point est extrêmement important pour deux raisons :



La première concerne une branche entière des statistiques basée sur le théorème de Bayes, et appelée statistique bayésienne. Thomas Bayes (1702-1761) était un pasteur mathématicien britannique (illustré ci-contre). Ses découvertes en probabilités ont été résumées dans son Essais sur la manière de résoudre un problème dans la doctrine des risques (*Essay towards solving a problem in the doctrine of chances* - 1763) publié à titre posthume. En ce qui nous concerne, l'idée principale de Bayes était qu'il nous faut tenir compte de ce que l'on connait déjà avant d'inférer la probabilité d'événements inconnus. Par exemple, si je demande la probabilité qu'un être humain soit mort pendu, vous me répondrez sans doute qu'elle est très faible, dans la mesure où la plupart des gens meurent d'autre chose que de pendaison. En revanche, si je vous demande quelle est la probabilité qu'un être humain meure SACHANT qu'il s'est pendu, là vous me répondrez que la probabilité est proche de 1 (je parle évidemment d'une pendaison par le cou). Ce type de raisonnement conditionne la pensée bayésienne que nous n'aborderons pas durant ces prochains cours parce que ce domaine est encore fort peu appliqué en sciences humaines. Cependant, je vous enjoins à ne pas perdre cette information de vue, surtout si vous vous intéressez à la recherche, parce que je ne serais pas étonné qu'elle se développe dans les prochaines années. En outre, vous aborderez fort probablement cette problématique dans le cours "*d'Histoire, Concepts et Méthodes*" en BA3.

La deuxième raison concerne la notion d'indépendance entre deux événements qui est liée aux probabilités conditionnelles (comme nous le verrons). Or, comme nous l'avons envisagé dans le chapitre 2, cette notion est fondamentale. Rappelez-vous que nous fonctionnons par modèle. Nous voulons simplifier la réalité et décrire un événement à partir de quelques variables, le moins possible, qui influencent beaucoup le concept décrit. Nous établirons des

hypothèses, chaque hypothèse décrivant le lien qui peut exister entre une variable indépendante et une variable dépendante. Dès lors, montrer que deux variables sont complètement indépendantes l'une de l'autre, c'est-à-dire qu'il n'est pas nécessaire de tenir compte de l'une pour prédire l'autre, est évidemment tout à fait essentiel puisque cela permet de rejeter l'hypothèse qui postule ce lien.

Voyons maintenant comme cela s'opérationnalise mathématiquement. Supposons que nous envisagions le lien entre deux variables : le fait de réussir (ou pas) un examen (d'Analyse de Données par exemple) et le fait d'avoir lu (ou pas) les lectures conseillées pour ce cours. Nous sommes donc devant un espace-échantillon partitionné selon deux propriétés : le fait d'avoir effectué les lectures d'accompagnement d'un cours (L et ~L) et le fait d'avoir réussi l'examen (R ou ~R). Il y a 120 étudiants qui se répartissent dans les quatre catégories comme indiqué au Tableau 3.3. Chaque cellule représente un événement conjoint qui provient de la combinaison de deux événements non exclusifs (= compatibles) : il est possible de réussir ou de rater en ayant ou en n'ayant pas lu, et réciproquement, le fait d'avoir lu ou pas lu n'empêche pas qu'on puisse avoir réussi ou raté.

Le Tableau 3.3 représente cette situation. Par exemple, la première case contient l'information du nombre de sujets qui ont à la fois réussi l'examen et lu les lectures conseillées (n = 70). Dans la marge de la colonne de droite on trouve la proportion des sujets qui ont réussi l'examen, indépendamment du fait qu'ils aient lu ou non les lectures conseillées (n = 75). Remarquez que les "n" représentent les fréquences absolues, c'est-à-dire le nombre de sujets qui appartiennent à la case, indépendamment du nombre total de sujets (lorsque je dis : 70 sujets ont réussi, je ne parle pas du nombre total de sujets qui ont essayé de réussir, donc qui ont passé l'examen). La somme des "n" de chaque case donne le nombre total "N" de sujets. Les fréquences relatives : $f = n/N$ représentent le nombre de sujets d'une case, divisé par le nombre total de sujets, d'où le terme "relatif" puisqu'il s'agit du nombre de sujets **par rapport** au nombre total de sujets. Ces fréquences sont considérées comme des estimations des probabilités correspondantes qu'on notera P (somme des P = 1). Ce faisant, on envisage en fait la définition épistémique des probabilités puisque nous sommes en train de calculer le nombre "n" de cas favorables de l'événement critique concerné divisé par le nombre "N" de cas possibles. Par exemple, la probabilité d'avoir réussi et d'avoir lu est de 70/120 = .583.

Tableau 3.3. : Tableau de contingence indiquant les fréquences absolues et les probabilités estimées par les fréquences relatives de réussite en fonction du fait d'avoir lu les ouvrages conseillés pour 120 étudiants.

	L (Lecture)	~L (Non Lecture)	Total marginal Probabilité marginale
R (Réussite)	$n(R \cap L) = 70$	$n(R \cap \sim L) = 5$	$n(R) = 75$
	$P(R \cap L) = .583$	$P(R \cap \sim L) = .042$	$P(R) = .625$
~R (Non Réussite)	$n(\sim R \cap L) = 20$	$n(\sim R \cap \sim L) = 25$	$n(\sim R) = 45$
	$P(\sim R \cap L) = .167$	$P(\sim R \cap \sim L) = .208$	$P(\sim R) = .375$
Total marginal	$n(L) = 90$	$n(\sim L) = 30$	$N = 120$
Probabilité marginale	$P(L) = .750$	$P(\sim L) = .250$	$P = 1.00$

Revenons à notre considération des totaux marginaux. Chaque marge représente une partition en deux groupes de l'espace-échantillon constitué par les 120 étudiants. Dans la marge inférieure, on voit que la **probabilité inconditionnelle** d'avoir effectué les lectures est de .750, donc celle de ne pas les avoir effectuées est de .250. Dans la marge de droite, on voit que la probabilité inconditionnelle d'avoir réussi est de .625, donc celle de ne pas avoir réussi est de .375. On dit, donc, que ces probabilités sont inconditionnelles parce qu'elles concernent la probabilité d'une variable (la lecture ou la réussite) **sans envisager la probabilité de l'autre variable**. En effet, lorsque par exemple, je dis que 75 sujets ont réussi l'examen, je ne donne aucune information sur le fait qu'ils ont ou non lu les lectures conseillées.

En revanche, si je me pose la question de savoir quelle est la probabilité de réussir SACHANT que le sujet a lu les lectures conseillées, je suis en train de m'intéresser aux **probabilités conditionnelles**. Je note cette question $P(R \mid L)$ qui se lit : "*probabilité de R si L*" ou "*probabilité de réussir si on a lu*". Intuitivement on peut se représenter cette probabilité de nombreuses manières. Les Figures 3.4a et 3.4b représentent la situation par les ensembles. Vous constaterez que, à intersection égale, le lien entre la lecture et la réussite est beaucoup plus fort dans la Figure 3.4b que dans la Figure 3.4a.

Figure 3.4a : La proportion de sujets qui ont lu et réussi est très petite par rapport à la proportion des sujets qui ont lu (R∩L prend peu de place dans L).

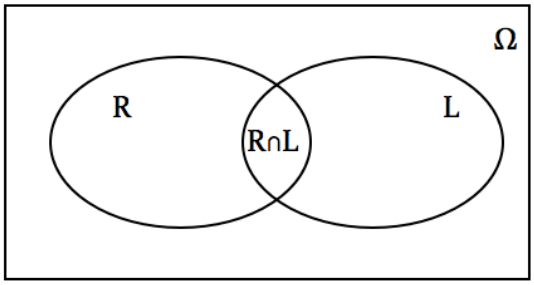
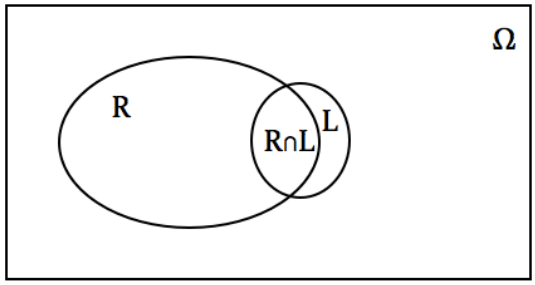


Figure 3.4b : La proportion de sujets qui ont lu et réussi est très grande par rapport à la proportion des sujets qui ont lu (R∩L prend beaucoup de place dans L).



Notes :

- L'intersection R∩L se veut identique dans les deux figures
- La partie L sans R∩L vaut ~R∩L (et inversément la partie de R sans R∩L vaut R∩~L)

Du point de vue algébrique, dans le Tableau 3.3, nous connaissons la probabilité qu'un sujet réussisse et ait lu : $P(R \cap L)$ =.583. Mais cette probabilité ne nous dit rien, en elle-même, sur les chances d'avoir réussi SI on a lu. En effet, cette probabilité-là doit tenir compte du nombre de sujets qui ont lu (c'est aussi ce qui est représenté par les Figures 3.4a et 3.4b)! Si je dis, par exemple, que 118 personnes ont lu mais que seuls 70 de ces sujets ont réussi, je ne peux pas conclure à un lien aussi fort entre le fait d'avoir lu et celui d'avoir réussi que si 70 sujets ont lu et que tous ont réussi. Pour tenir compte de cette information, je vais calculer ma probabilité conditionnelle $P(R \mid L)$ en considérant le rapport entre la proportion de sujets qui ont lu et réussi et la somme de la proportion de sujets qui ont lu et réussi et de la proportion de sujets qui ont lu et raté (donc la proportion de tous les sujets qui ont lu[14]) : $P(R \mid L) = P(R \cap L)/(P(R \cap L) \cup P(\sim R \cap L)) = P(R \cap L)/P(L) = .583/.750 = .778$. Cette probabilité ainsi obtenue est à comparer avec la probabilité inconditionnelle $P(R) = .625$. Vous voyez que la probabilité de réussir SI on a lu est plus élevée que la probabilité de réussir sans tenir compte de l'information de la lecture.

[14] Je vous mets au défi de comprendre cette phrase! Alors que la représentation algébrique qui suit dit exactement la même chose et me semble beaucoup plus limpide. Surtout si vous regardez les Figures 3.4a et b pour visualiser chaque terme. J'espère par ce commentaire vous sensibiliser à l'intérêt de la notation mathématique.

Voyons ensuite quelle est la probabilité de réussite étant donné que les lectures n'ont pas été faites, que l'on note $P(R \mid \sim L)$. Ceci revient à calculer la probabilité de réussite dans la colonne droite du Tableau 3.3. Par le même raisonnement on obtient : $P(R \mid \sim L) = P(R \cap \sim L)/P(\sim L) = .042/.250 = .168$. Cette probabilité est plus petite que la probabilité inconditionnelle $P(R) = .625$. On peut donc en déduire que la probabilité de réussir si l'on a pas lu est nettement inférieure à la probabilité de réussir sans tenir compte de la variable "*lecture*".

Toujours selon le même raisonnement, on peut aussi examiner les choses dans l'autre sens, c'est-à-dire la probabilité d'avoir lu étant donné la réussite (première ligne du tableau) et la probabilité d'avoir lu étant donné l'échec (deuxième ligne du tableau). Il faut ensuite comparer ces probabilités avec la probabilité inconditionnelle $P(L) = .750$.

$$P(L \mid R) = P(R \cap L)/P(R) = .583/.625 = .933$$
$$P(L \mid \sim R) = P(\sim R \cap L)/P(\sim R) = .167/.375 = .445$$

Donc, sachant que la probabilité inconditionnelle d'avoir lu est de .750 et connaissant le sens de la relation entre lecture et réussite comme indiqué ci-dessus, on peut prédire qu'un étudiant qui a réussi a une probabilité supérieure à .750 d'avoir lu et qu'un étudiant qui a échoué a une probabilité plus faible que .750 d'avoir lu.

En fait, cette information est redondante. Si on obtient un lien de dépendance entre une variable A et une variable B, on obtiendra forcément un lien de dépendance entre la variable B et la variable A. Dès lors, en pratique, on envisagera le problème dans un des deux sens seulement, et ce sens sera défini par la théorie. Dans notre exemple, il me semble plus pertinent de déterminer si les gens ont réussi, sachant qu'ils ont lu, que de déterminer s'ils ont lu sachant qu'ils ont réussi…

A partir des exemples que nous venons de traiter ci-dessus, nous pouvons généraliser les relations algébriques pour deux événements A et B quelconques. Nous allons le faire en isolant deux termes différents : soit la probabilité conditionnelle soit la probabilité conjointe.

Probabilités conditionnelles et probabilités conjointes pour événements dépendants

Probabilités conditionnelles		Probabilités conjointes
$P(A\|B) = \frac{P(A \cap B)}{P(B)}$	$\Leftrightarrow$	$P(A \cap B) = P(A\|B)P(B)$
	SSi $P(B) > 0$	
$P(B\|A) = \frac{P(A \cap B)}{P(A)}$	$\Leftrightarrow$	$P(A \cap B) = P(B\|A)P(A)$
	SSi $P(A) > 0$	

Remarquez, pour conclure ce point, qu'à chacune des quatre probabilités conditionnelles calculées ci-dessus correspond une probabilité complémentaire. Par exemple, le complément de $P(R \mid L) = .778$ est $P(\sim R \mid L) = 1 - .778 = .222$. Il s'agit de la probabilité d'avoir raté étant donné que les lectures ont été faites. Bien entendu, si $P(R \mid L)$ n'avait pas été calculée, $P(\sim R \mid L)$ pouvait être calculée en appliquant la formule de définition de la probabilité conditionnelle, soit $P(\sim R \cap L)/P(L) = .167/.750 = .222$.

3.3.8.2. Indépendance entre événements

Imaginons maintenant une autre situation dans laquelle on collecte (bizarrement) deux types d'informations : la réussite à l'examen d'Analyse de Données et la couleur, claire (Light) ou pas, des murs du salon des grands-parents des sujets. Vous vous direz sans doute, intuitivement, qu'il n'y a aucune raison pour qu'il existe un lien de dépendance entre ces deux événements. C'est précisément ce que je vais montrer, d'autant plus facilement que le Tableau 3.4 qui présente ces informations est totalement fictif et adapté à mon propos. Donc, sachant que la probabilité inconditionnelle de réussite est de .625, cette probabilité resterait la même pour un étudiant dont les grands-parents ont un salon aux murs clairs et pour un étudiant qui a des grands-parents qui vivent dans un salon foncé.

Tableau 3.4. : Notion d'indépendance entre événements non exclusifs

	L (Light)	~L (Non Light)	Total marginal Probabilité marginale
R (Réussite)	$n(R \cap L) = 60$	$n(R \cap \sim L) = 15$	$n(R) = 75$
	$P(R \cap L) = .500$	$P(R \cap \sim L) = .125$	$P(R) = .625$
~R (Non Réussite)	$n(\sim R \cap L) = 36$	$n(\sim R \cap \sim L) = 9$	$n(\sim R) = 45$
	$P(\sim R \cap L) = .300$	$P(\sim R \cap \sim L) = .075$	$P(\sim R) = .375$
Total marginal	$n(L) = 96$	$n(\sim L) = 24$	$N = 120$
Probabilité marginale	$P(L) = .800$	$P(\sim L) = .200$	$P = 1.00$

En appliquant les relations algébriques que nous avons établies au point précédent, nous pouvons évaluer les probabilités conditionnelles (remarquez que je ne pousse pas le vice jusqu'à envisager la probabilité d'avoir des murs clairs en sachant qu'on a réussi, mais bien l'inverse). Comme on s'y attend, la probabilité de réussir si les murs sont clairs est égale à la probabilité marginale de réussir qui ne tient pas compte de la couleur des murs, de la même manière, la probabilité de réussir quand les murs ne sont pas clairs est également égale à la probabilité marginale de réussir :

$$P(R \mid L) = P(R \cap L)/P(L) = .500/.800 = .625 = P(R)$$
$$P(R \mid \sim L) = P(R \cap \sim L)/P(\sim L) = .125/.200 = .625 = P(R)$$

Vu sous l'angle des probabilités d'avoir un salon clair conditionnellement à la réussite, la conclusion est la même :

$$P(L \mid R) = P(R \cap L)/P(R) = .500/.625 = 0{,}800 = P(L)$$
$$P(L \mid \sim R) = P(\sim R \cap L)/P(\sim R) = .300/.375 = .800 = P(L)$$

A partir de cet exemple, nous pouvons établir les formules générales pour l'indépendance :

- En cas d'indépendance entre deux événements A et B, $P(A \mid B) = P(A)$ *(équation 1)*
- On sait, par ailleurs que : $P(A \cap B) = P(A \mid B)\, P(B)$ *(équation 2)*
- A partir des équations 1 et 2, on obtient, par remplacement : $P(A \cap B) = P(A)\, P(B)$
- On peut tenir le même raisonnement pour $P(B \mid A)$ qui vaudrait la même chose.

On aboutit à une règle des produits des probabilités qui est **applicable au cas des événements compatibles et indépendants**. Quand les événements sont non exclusifs, donc compatibles, et indépendants, la probabilité conjointe est obtenue en faisant le produit des probabilités marginales, c'est-à-dire des probabilités non conditionnelles.

Probabilités conditionnelles et probabilités conjointes pour événements indépendants

Probabilités conditionnelles		**Probabilités conjointes**
$P(A\|B) = P(A)$ $P(B\|A) = P(B)$	$\Leftrightarrow$	$P(A \cap B) = P(A)P(B)$

Cette information fait partie des informations très importantes dans le contexte des probabilités. En effet, dans la suite du programme nous aurons souvent à évaluer si les variables A et B sont indépendantes ou non. Partir du principe que si elles le sont alors la probabilité de leur intersection A ∩ B est égale à la probabilité de l'événement A multiplié par la probabilité de l'événement B et comparer nos résultats à cette hypothèse sera un bon moyen de procéder.

Voici un exemple issu de l'utilisation du Tableau 3.4 : La probabilité conjointe de l'événement « *avoir réussi et avoir un salon clair* » notée $P(R \cap L) = .500$ est obtenue en faisant le produit des deux probabilités marginales des événements « *avoir réussi* » notée $P(R) = .625$ et de la probabilité marginale « avoir un salon clair » notée $P(L) = .800$. Donc, $P(R \cap L) = P(R).P(L)$, soit .500 = .625 x .800, ce qui implique l'indépendance. Ceci n'est vrai que parce qu'au Tableau 3.4, les deux événements sont compatibles (on peut avoir un salon clair et avoir réussi alors qu'un animal ne peut pas avoir à la fois des poils et des plumes) et

indépendants (la réussite est complètement indépendante du fait d'avoir un salon clair ou pas).

Comparez avec ce qui se passe au Tableau 3.3 où les événements conjoints sont aussi compatibles mais pas indépendants. Ici $P(R \cap L) \neq P(R).P(L)$, donc .583 ≠.625x.750 = .469, ce qui implique la non indépendance. Je vais ici faire une remarque que nous ne comprendrons pleinement que bien plus tard mais qui devrait vous sensibiliser à l'inférence statistique : il devrait vous sembler évident que lorsque je mesure mes variables "*Réussite*" et "*Couleur de Salon*" chez 120 sujets pris au hasard, je n'aurai pas forcément une probabilité aussi clairement indépendante que dans mon Tableau 3.4. En effet, je suis aussi soumis au hasard de l'occurrence des événements et il se peut qu'il y ait, par exemple, un peu plus de sujets qui ont réussi et dont les grands-parents ont un salon clair, de la même manière que si je lance 50 fois une pièce de monnaie je n'aurai pas systématiquement 25 "*pile*" et 25 "*face*". Dès lors, si je veux réellement savoir si deux événements sont indépendants, à partir d'un échantillon de 120 sujets comme c'est le cas dans mon tableau, je dois me poser la question de savoir si la probabilité que j'obtiens diffère suffisamment du produit des probabilités des deux événements pour que je puisse dire que les événements sont dépendants l'un de l'autre. En d'autres termes, dans notre exemple, je me poserai un jour (mais pas aujourd'hui) la question de savoir si .583 est vraiment fort différent de .469 ou bien si je dois considérer que ces deux probabilités sont trop proches pour pouvoir conclure, avec une chance raisonnable de ne pas me tromper, qu'elles sont différentes.

3.3.9 Résumé de la théorie des notions vues

Evénement : fait qui peut se produire ou non.

Evénements indépendants : Deux événements sont indépendants si l'occurrence de l'un n'a aucune influence sur l'occurrence de l'autre.

Evénements mutuellement exclusifs (= incompatibles) : Deux événements sont mutuellement exclusifs si l'occurrence de l'un empêche l'occurrence de l'autre.

Ensemble d'événements exhaustif : Un ensemble d'événements est dit exhaustif si l'un de ces événements se réalise nécessairement.

Règle de multiplicativité : Si A et B sont deux événements indépendant, alors :

$$P(A \cap B) = P(A)\,P(B).$$

Règle de l'additivité : Si A et B sont deux événements mutuellement exclusifs, alors :

$$P(A \cup B) = P(A) + P(B)$$

Probabilité conditionnelle :

$$P(A \mid B) = P(A \cap B)/P(B)$$

Si deux événements sont indépendants, alors :

$$P(A \mid B) = P(A)$$

et

$$P(A \cap B) = P(A)\,P(B)$$

3.3.10. Synthèse des sections 3.3.7 et 3.3.8

Freedman et al. (3ème éd., 1998, pp. 243-244) fournissent une excellente synthèse sur les liens entre les notions suivantes : événements mutuellement exclusifs et règle d'addition des probabilités; événements indépendants et règle de multiplication des probabilités. Cette synthèse apparaît sous la rubrique « *Two FAQs (frequently asked questions)* ». En voici une traduction quasi littérale augmentée de quelques commentaires. On s'intéresse à deux événements A et B dont les probabilités sont *P(A)* et *P(B)*.

Les deux questions fréquemment posées sont :

- **Quelle est la différence entre événements mutuellement exclusifs et indépendants?**

Le fait d'être « *mutuellement exclusifs* » constitue une idée, la notion d'indépendance en constitue une autre. Cependant, les deux idées s'appliquent à des paires d'événements et les deux idées disent quelque chose à propos de la manière dont ces événements peuvent être en relation. Cependant, il faut distinguer deux types de relation :

Deux événements sont mutuellement exclusifs si la survenue de l'un empêche la survenue de l'autre (on dit aussi de tels événements qu'ils sont disjoints ou incompatibles).

Deux événements sont indépendants si la survenue de l'un ne change pas la probabilité de survenue de l'autre (ceci suppose que les événements puissent se produire conjointement, donc qu'ils soient compatibles).

- **Quand peut-on additionner et quand peut-on multiplier les probabilités ?**

La règle d'addition et la règle de multiplication concernent des manières de combiner des probabilités. Cependant, les deux règles permettent de résoudre des problèmes différents. La règle d'addition permet de trouver la probabilité qu'au moins un des deux événements se produise. La règle de multiplication permet de trouver la probabilité que deux événements se produisent conjointement.

Donc, la première question à se poser pour décider s'il faut additionner ou multiplier les probabilités est : veut-on connaître *P(A ou B)*, *P(A et B)*, ou quelque chose d'autre. Une fois qu'on est sûr que la question posée nécessite de connaître *P(A ou B)* ou *P(A et B)*, il faut se demander si la relation entre les événements est propice à l'utilisation de la règle d'addition qui dit que $P(A \text{ ou } B) = P(A) + P(B)$ ou de la règle de multiplication qui dit que $P(A \text{ et } B) = P(A).P(B)$, ou bien s'il faut recourir à des règles plus générales. L'addition des probabilités de deux événements requiert que ces événements soient mutuellement exclusifs. La multiplication des probabilités non conditionnelles de deux événements nécessite que ces événements soient indépendants. Ce texte établit clairement la distinction entre des événements mutuellement exclusifs et des événements indépendants et les liens qui existent

entre ces deux notions et les règles d'addition et de multiplication des probabilités prises au sens restreint, donc au sens non général.

Remarque : par défaut, quand on parle de la règle d'addition des probabilités, il s'agit de la règle au sens restreint qui s'applique seulement aux cas des événements disjoints. C'est la règle exprimée dans l'axiome 3 de Kolmogorov. On a vu qu'il existe aussi une règle générale qui s'applique aux cas des événements quelconques (donc aussi aux événements non mutuellement exclusifs). De même, par défaut, quand on parle de la règle de multiplication des probabilités, il s'agit de la règle au sens restreint qui s'applique seulement aux cas des événements indépendants. Il existe aussi une règle générale qui s'applique aux événements non indépendants.

Dans le livre de Freedman et al. (pp. 244-245), ce point est suivi par un exemple commenté que je résume. Dans chaque cas on se demande si la solution peut être trouvée en appliquant la règle de l'addition ou du produit des probabilités (au sens restreint) ou si la question demande de recourir aux règles générales plus complexes.

Question 1. Un dé est lancé six fois. Quelle est la probabilité d'obtenir le six lors du 1er lancer ou lors du 6ème lancer ?

Réponse 1 : la question porte sur la survenue de l'un ou l'autre de deux événements. Donc, la règle d'addition devrait s'appliquer. Mais, attention, les événements ne sont pas mutuellement exclusifs (on peut obtenir le six au 1er lancer et aussi au 6ème) ; donc, la règle d'addition ne s'applique pas. A-t-on plus de chances de répondre à la question en se tournant vers la règle de multiplication? Celle-ci s'applique aux cas des événements indépendants. Or, il y a bien indépendance entre obtenir le six au 1er lancer et obtenir le six au 6ème lancer (il n'y a aucune raison que le fait d'obtenir 6 au premier lancer influence la probabilité d'obtenir 6 au sixième lancer). Mais la question ne porte pas sur la probabilité d'obtenir le six lors du 1er lancer et lors du 6ème. Donc, la règle de multiplication ne s'applique pas non plus. En conclusion : le problème est insoluble si on dispose seulement de la règle d'addition des probabilités au sens restreint et de la règle de multiplication des probabilités au sens restreint.

Question 2. Un dé est lancé six fois. Quelle est la probabilité d'obtenir le six lors du 1er lancer et lors du 6ème lancer ?

Réponse 2 : La question porte sur la survenue de l'un et l'autre événements et ces événements sont indépendants. Donc, la règle de multiplication des probabilités s'applique et la réponse est : 1/6 x 1/6 = 1/36.

Question 3. Dans un paquet de 52 cartes bien mélangées, quelle est la probabilité que la carte du dessus soit l'as de pique ou que la carte du dessous soit l'as de pique ?

Réponse 3 : La probabilité que la carte du dessus soit l'as de pique est de 1/52. Avant tout tirage d'une carte, la probabilité que celle du dessous soit l'as de pique est aussi de 1/52. Les deux événements sont mutuellement exclusifs (si la 1ère carte est l'as de pique, la dernière ne peut pas l'être et réciproquement). Or, la question porte bien sur la survenue de l'un ou de l'autre des deux événements exclusifs. Donc, la règle d'addition des probabilités s'applique et la réponse est : 1/52 + 1/52 = 2/52.

Question 4. Dans un paquet de 52 cartes bien mélangées, quelle est la probabilité que la carte du dessus soit l'as de pique et que la carte du dessous soit l'as de pique ?

Réponse 4 : Vous savez que cette probabilité est nulle. En effet, si la 1ère carte est l'as de pique, la dernière ne peut pas l'être ; si la dernière carte est l'as de pique, la première ne peut pas l'être. Donc, les deux événements sont mutuellement exclusifs. Mais on ne vous demande pas la probabilité de l'un ou l'autre, mais la probabilité de l'un et l'autre des événements. Or, la règle de multiplication des probabilités ne s'applique pas puisque les événements ne sont pas indépendants. Donc, la question ne trouve pas de réponse dans le cadre de la règle des produits au sens restreint.

Réponse à la question 1 en utilisant la règle générale d'addition des probabilités. En fait, la question 1 porte sur la probabilité d'observer l'un ou l'autre de deux événements non exclusifs. Comme il s'agit d'un "*ou*" il faut recourir à la règle d'addition des probabilités. Mais la règle d'addition au sens restreint ne marche pas, comme on l'a vu ci-dessus. En revanche la règle générale permet d'obtenir la réponse. Donc :

P(6 au 1er lancer ou 6 au 6ème lancer) = P(6 au 1er lancer) + P(6 au 6ème lancer) – P(6 au 1er lancer et au 2ème lancer) = 1/6 + 1/6 – 1/36 = 11/36.

Réponse à la question 4 en utilisant la règle générale du produit des probabilités. La question 4 porte sur la probabilité d'observer l'un et l'autre de deux événements non indépendants. Comme il s'agit d'un "*et*" logique, il faut utiliser la règle des produits. Mais, la règle des produits au sens restreint n'est pas valide. En revanche, la règle générale des produits permet d'obtenir la réponse.

P(carte du dessus soit l'as de pique et la carte du dessous soit l'as de pique) = P(carte du dessus est l'as de pique) x P(carte du dessous est l'as de pique | carte du dessus est l'as de pique) = 1/52 x 0 = 0 puisque la probabilité conditionnelle est nulle.

3.4. Principe fréquentiste

La dualité du concept de probabilité a été mise en évidence à la section 3.3.1. Au sens épistémique, la probabilité indique le degré de croyance, de confiance ou de crédibilité que l'on peut accorder à des propositions du type « *le réchauffement climatique actuel est en grande partie causé par l'activité humaine* ». Au sens fréquentiste, la probabilité désigne la fréquence avec laquelle un dispositif aléatoire comme le jet répété d'une pièce de monnaie tend à fournir, par exemple, le résultat face. La probabilité est ici conçue comme une idéalisation de la notion de fréquence relative. Si la pièce était parfaitement équilibrée, la fréquence relative de l'événement face se stabiliserait progressivement autour de .50 au fur et à mesure de l'augmentation du nombre de jets.

La probabilité au sens fréquentiste correspond à une caractéristique physique de la pièce qui est vraie ou fausse indépendamment de nos croyances. Cette propriété physique un peu particulière peut être qualifiée de disposition, de tendance ou de propension parce que, contrairement à d'autres propriétés physiques comme la masse ou la forme, elle ne se manifeste pas en permanence. Elle ne se manifeste que quand la pièce est placée dans les circonstances appropriées, c'est-à-dire soumise à des jets successifs dont les résultats sont aléatoires.

Le point important souligné par Hacking et Dufour (2004) est que nous effectuons souvent un va-et-vient entre l'acception fréquentiste et l'acception épistémique de la probabilité, sans nécessairement toujours en être conscient. Ce faisant, nous obéissons à une sorte de règle que les auteurs proposent d'appeler principe fréquentiste. Pour comprendre le principe fréquentiste, il faut contraster trois manières de justifier le crédit que l'on peut accorder à une proposition, donc trois types de raisons qui peuvent être avancés pour justifier nos croyances.

Situation fréquentiste pure : Dans une situation fréquentiste pure, une proposition du type « *cette pièce particulière semble bien équilibrée* » peut être vérifiée empiriquement. En effet, on aurait des raisons de croire que cette proposition est vraie si, à l'issue d'une série de quelques dizaines ou de quelques centaines de lancers successifs, les fréquences relatives des deux événements possibles étaient proches de 0,50. Grâce à cette possibilité de vérification empirique, la probabilité au sens fréquentiste a souvent été qualifiée d'objective.

Situation épistémique pure : Dans une situation épistémique pure, une proposition du type « *il y a au moins huit chances sur dix pour que le réchauffement climatique actuel soit causé par l'activité humaine* » exprime le degré de confiance élevé qu'une personne accorde aux différents arguments qui appuient cette thèse. Les notions de fréquence, de tendance et de propension ne s'appliquent pas puisque le réchauffement climatique actuel est un événement unique. Comme le degré de confiance accordé à une proposition varie d'une personne à l'autre, la probabilité au sens épistémique a souvent été qualifiée de subjective.

Situation de va-et-vient : Un va-et-vient entre les deux acceptions de la notion de probabilité se produit, par exemple, dans la situation où une pièce serait lancée une seule fois et où un observateur prédirait que l'événement face a une chance sur deux de se produire pendant que la pièce est encore en l'air ou alors qu'elle a déjà atterri, mais à un endroit dissimulé à son regard. Le jugement de l'observateur sur le résultat de ce lancer unique de la pièce est épistémique au même titre que le jugement sur le réchauffement climatique actuel. Cependant, la justification est différente. Elle repose sur la connaissance de la probabilité de l'événement face au sens fréquentiste. L'observateur attribue donc la valeur de la probabilité fréquentiste de l'événement face à la probabilité épistémique de l'événement face dans le cas du lancer unique d'une pièce particulière. C'est l'utilisation de ce lien entre probabilité fréquentiste et probabilité épistémique pour justifier une croyance que Hacking et Dufour (2004) qualifient de principe fréquentiste.

Ce principe fréquentiste est continuellement à l'œuvre dans les situations de décision dans l'incertitude portant sur des individus particuliers. Un patient dépressif à qui on permet d'arrêter son traitement commettra ou ne commettra pas un suicide. Il n'y a pas de probabilité fréquentiste associée à cet événement unique qui peut se produire ou ne pas se produire. Toutefois, si on sait sur une base fréquentiste que dans 90% des cas un tel patient ne commet pas un suicide, on peut justifier, par le principe fréquentiste, d'arrêter le traitement. Cette décision repose sur une probabilité épistémique de 90% de chances de succès, elle-même basée sur une probabilité de type fréquentiste : dans 90% des cas comparables observés, cette décision n'a pas donné lieu à une issue négative.

3.5. Notions d'analyse combinatoire : "*counting rules*"

Cette partie fait écho au besoin, en probabilité, d'évaluer le nombre de cas possibles et de dénombrer des événements ainsi qu'à l'utilisation des combinaisons dans l'équation de la distribution binomiale (voir chapitre 7). Bien que vous aurez, en pratique, peu d'occasions d'utiliser ces notions dans les tests inférentiels que vous vous apprêtez à réaliser dans la suite des événements, elles sont importantes pour une bonne compréhension du cadre théorique.

Ces règles concernent le dénombrement du nombre de dispositions d'un certain nombre d'objets pris parmi un ensemble d'objets avec ou sans remise. Deux critères sont importants lorsque l'on cherche à dénombrer des séquences, nous envisagerons les justifications théoriques par après : **il est essentiel de s'intéresser à la remise ou à la non remise des éléments avant chaque tirage (c'est-à-dire à l'indépendance ou à la dépendance des tirages successifs) et à l'importance ou non de l'ordre de tirage**. En vous basant sur ces deux critères de sélection, vous aurez facile à identifier la méthode correcte de dénombrement. Vous pouvez déjà, en principe, faire le lien théorique avec la dépendance ou l'indépendance des tirages vu dans ce chapitre.

Lorsque l'on désire déterminer la probabilité d'un événement, la première question à se poser est : de combien de manières différentes cet événement peut-il se produire? Pour parvenir à cette conclusion, une démarche en quatre étapes est nécessaire.

1. L'établissement de l'inventaire exhaustif des événements élémentaires constituant l'espace-échantillon.
2. Le comptage des événements élémentaires. La probabilité de n'importe lequel d'entre eux est : 1/nombre total d'événements élémentaires. Par exemple, l'inventaire des événements élémentaires à prendre en compte pour établir la distribution des probabilités de la somme des points lors du jet de deux dés montre qu'il y a 36 séquences à envisager (rappelez-vous : il y a 6 possibilités pour le premier dé, 6 possibilités pour le second, les deux événements sont indépendants, il y a donc 36 combinaisons possibles). Chaque séquence a une probabilité de 1/36 de se produire lors d'un essai.
3. La délimitation des classes d'événements élémentaires correspondant à des événements composés dont on cherche la probabilité et comptage des événements dans ces classes. Par exemple, il existe cinq manières différentes d'obtenir une somme de 6 lors du jet de deux dés (1-5 ; 2-4 ; 3-3 ; 4-2 ; 5-1), ce qui donne une probabilité de 5/36 = .138.
4. Même quand les événements élémentaires ne sont pas équiprobables, on peut se livrer à l'inventaire décrit en 3. Supposons qu'on numérote les faces de l'astragale dont il est question au chapitre 3 (voir point 3.3.4) de la manière suivante 1 pour la face rouge ($\hat{p}$ = 50/300 = .167), 2 pour la face bleue ($\hat{p}$ = 88/300 = .293), 3 pour la face noire ($\hat{p}$ = 52/300 = .173) et 4 pour la face incolore ($\hat{p}$ = 110/300 = .367) [15]. En lançant deux fois cet astragale, quelle est la probabilité d'obtenir une somme de 5? En dressant l'inventaire des 16 événements élémentaires possibles, on s'aperçoit qu'il y a quatre manières d'obtenir ce résultat : { 1-4, 4-1, 2-3, 3-2 } . Les événements étant indépendants au sein de chaque couple, la loi du produit des probabilités s'applique et donne *P* = .167 x .367 = .061 pour les séquences 1,4 et 4,1 et *P* = .293 x .173 = .051 pour les séquences 2,3 et 3,2. Les quatre événements composés donnant une somme de 5 étant mutuellement exclusifs, la loi de l'addition des probabilités (au sens restreint) s'applique. Donc, la probabilité d'obtenir une somme de 5 est donnée par la probabilité d'avoir obtenu 1 et 4 ou 4 et 1 ou 2 et 3 ou 3 et 2, soit .061 + .061 + .051 + .051 = .224.

[15] L'accent circonflexe au-dessus du p suggère qu'on parle de l'estimation de la probabilité.

La procédure décrite ci-dessus est générale. Elle permet en principe de déterminer la probabilité de n'importe quel événement. En pratique, il serait souvent fastidieux de se livrer à l'inventaire des événements élémentaires. Heureusement, on peut s'aider d'un certain nombre de règles qui permettent d'obtenir les résultats voulus par calcul.

3.5.1. Règle des produits

Cette règle sert à dénombrer les événements élémentaires distincts (indépendants) de l'espace-échantillon d'une expérience aléatoire comprenant *N* essais indépendants [16] donc on peut considérer qu'il s'agit de tirages **avec remise** (puisqu'indépendants) et dans lequel **l'ordre est défini**. Elle dit que :

<u>Règle des produits</u>

Si $K_1, K_2, \ldots K_N$ sont les nombres d'événements distincts et indépendants qui peuvent se produire au cours des essais 1, 2,…*N* dans une série, le nombre de séquences différentes de *N* événements est donné par le produit $K_1 \times K_2, \ldots \times K_N$.

Admettons que *N* = 3, c'est-à-dire que l'on réalise trois essais d'un événement aléatoire "*jet de dé*" mais qu'on utilise plusieurs dés différents : au premier jet un dé deux faces (donc une pièce de monnaie), au deuxième essai on jette un dé dodécaédrique (12 faces), au troisième essai on jette un dé tétraédrique (4 faces, illustration ci-contre). Le nombre de séquences possibles devient : 2 x 12 x 4 = 96. Remarquez qu'il s'agit bien d'événements indépendants puisque quel que soit le résultat du jet concernant la pièce de monnaie, cela n'a aucune influence sur le résultat du jet des autres dés (pareillement pour les autres jets qui ne s'entre-influencent pas).

Une application très courante de cette règle est la détermination d'un plan factoriel expérimental : lorsqu'un psychologue crée une expérience, il voudra mesurer l'influence

[16] Remarquez que j'utilise un *N* majuscule puisqu'ici nous parlons de l'ensemble de l'espace-échantillon et non d'un échantillon de l'espace-échantillon.

d'une ou plusieurs variables sur une autre (rappelez-vous du chapitre 2 sur les hypothèses). Supposons le cas d'un psychologue qui envisage d'étudier l'effet de la culpabilité montrée par un enfant sur la punition qu'on lui attribue. Supposons également que le psychologue ait des raisons de penser que cette influence dépende également du genre, c'est-à-dire qu'il suppose que les femmes ne seront pas influencées de la même manière que les hommes par cette démonstration (ou absence de démonstration) de culpabilité. Ce psychologue devra donc établir deux variables indépendantes (le sentiment de culpabilité de l'enfant et le genre) et mesurer une variable dépendante (la punition). De nombreux choix sont possibles pour mettre au point une telle expérience. Imaginons qu'il envisage trois conditions pour la variable Démonstration de Culpabilité : (a) dans un cas, l'enfant se sent coupable ; (b) dans un autre le sujet n'a aucune information sur la culpabilité ressentie par l'enfant ; (c) dans un dernier cas, l'enfant ne se sent explicitement pas coupable. La variable Genre est assez évidente : soit le sujet est féminin, soit il est masculin. La punition est mesurée par une échelle de sévérité en 7 points (ce qu'on appelle une échelle de Likert).

Pour décrire un tel plan expérimental, on ne prend en compte que les conditions expérimentales, c'est-à-dire les variables indépendantes. Donc le Sentiment de Culpabilité et le Genre. Le nombre de conditions possibles est donné, selon la règle des produits, par le nombre de possibilités pour la première variable indépendante (trois) et le nombre de possibilités pour la seconde variable indépendante (deux). On écrira qu'on est dans un plan expérimental 3 x 2. Il y a donc 6 conditions possibles. En effet, si on les dénombre, nous obtenons : Homme-Coupable ; Femme-Coupable ; Homme-pas d'info ; Femme-pas d'info ; Homme-non coupable ; Femme-non coupable. Nous pourrons donc mesurer la punition attribuée dans chacune des conditions et comparer les moyennes entre les conditions. Ce genre de procédure est très courante. Elle permet par exemple de comparer la sévérité entre deux conditions spécifiques, comme la punition attribuée par une femme lorsqu'un enfant se sent coupable et la punition attribuée par un homme lorsque l'enfant se sent coupable. Mais il est également possible de comparer la punition lorsque l'enfant se sent coupable et lorsqu'il ne se sent pas coupable (indépendamment du genre) ou la punition moyenne attribuée par les hommes et par les femmes (indépendamment du sentiment de culpabilité). Nous verrons ces cas de figure dans les cours des années qui viennent.

Il existe un cas particulier de la règle des produits qui simplifie un petit peu le calcul : lorsque les K essais ont le même nombre de conditions. Imaginons qu'on lance 5 dés cubiques (à 6 faces). Dans ce cas, $K = 6$ et $N = 5$. Le nombre de possibilités, selon la règle des

produits, est de 6x6x6x6x6 = 6^5 = 7776 possibilités. Dès lors, dans ce cas, le nombre de possibilités vaut K^N.

Encore une fois, nous sommes dans un cas où les jets sont indépendants (avec remise) et l'ordre compte. L'idée de la remise est peut-être difficile à visualiser dans le cas de jets de dés. Ce que j'entends par là c'est que le fait d'obtenir le score "*1*" au premier jet n'implique pas que cette valeur ne soit plus possible lors du second jet (on n'a pas enlevé le "*1*" du deuxième dé). La situation serait différente si le fait de faire un "*1*" au premier jet abolissait la valeur pour tous les autres jets et pareillement pour les autres valeurs. Dans ce dernier cas, sitôt un score obtenu, il ne pourrait plus être répété puisqu'il serait ôté des possibilités. Admettons que le premier jet soit un "*4*". Le deuxième peut valoir n'importe quoi sauf "*4*". Mettons que ce soit un "*2*". Le troisième jet ne peut plus valoir que 1-3-5 ou 6. Mettons qu'ensuite un "*3*" sorte, puis un "*5*", puis un "*1*". Le dernier jet ne pourrait donc plus être autre chose que le "*6*" qui n'est toujours pas sorti. La séquence est donc 4-2-3-5-1-6. Ici, le problème reviendrait à calculer les possibilités d'arranger les chiffres 1-2-3-4-5-6 entre eux, et nous envisagerons cette procédure plus tard.

Deux exemples d'application de la règle des produits où la probabilité des *K* est constante pourraient être : (a) le nombre de séquences possibles de trois dés (cubiques) de couleurs différentes est de 6^3 = 216 ; (b) le nombre de séquences de trois lettres latines est de 26^3 = 17576. On peut multiplier ces exercices à l'envi, vous en verrez lors des travaux pratiques.

3.5.2. Règles combinatoires

Lorsqu'il s'agit de combiner des éléments entre eux, il est nécessaire de distinguer trois cas : les permutations, les arrangements et les combinaisons. Nous verrons les propriétés et l'utilisation de chacun d'entre eux et un tableau de synthèse vous permettra, je l'espère, d'y voir clair en un clin d'oeil.

3.5.2.1. Permutations

Les permutations concernent le cas où on dispose de tous les éléments d'un ensemble et que l'on doit les arranger **sans remise et où l'ordre est important**. Par exemple, si j'ai un ensemble de deux objets (un couteau et une fourchette) et que je cherche toutes les manières de les arranger, j'en obtiens deux : le couteau d'abord, la fourchette ensuite, ou

l'inverse. Remarquez que je suis en train de considérer que les tirages ne sont **pas indépendants**, puisqu'il n'y a pas de remise. En effet, lors de mon premier choix, je peux commencer par la fourchette ou par le couteau, mais, dès lors que j'ai choisi, l'autre s'impose. Si je place la fourchette en premier, je n'ai d'autre choix que de mettre le couteau ensuite. Dans la règle des produits, ce n'était pas le cas. A ce moment, si je considérais un pile ou face (qui est déterminé par deux possibilités, comme les couverts que nous venons d'envisager) le premier lancer pouvait être soit pile soit face et le second lancer pouvait également être soit pile soit face. Il n'était alors pas question de dire que si je faisais pile lors du premier jet je devais faire face lors du second. Dans ce cas, j'avais le choix entre pile-pile ; pile-face ; face-pile ; face-face. J'avais donc quatre possibilités.

Si j'ai trois objets parmi trois, par exemple, un couteau, une fourchette et une cuillère, j'obtiens 6 séquences possibles : couteau-fourchette-cuillère ; couteau-cuillère-fourchette ; fourchette-couteau-cuillère ; fourchette-cuillère-couteau ; cuillère-couteau-fourchette ; et cuillère-fourchette-couteau. Remarquez la logique de la méthode que j'utilise pour parvenir à ce résultat : je fixe le premier couvert de manière à permuter les deux derniers de toutes les manières possibles. Puis je change le premier couvert et permute à nouveau les deux derniers, etc. Si j'avais plus de types de couverts, j'étendrais la méthode : je fixerais tous les couverts sauf les deux derniers que je permuterais, puis je changerais l'antépénultième et permuterais les deux derniers, et remonterais ainsi jusqu'au premier. On peut définir que :

Nombre de Permutations possibles

$$P = N!$$

Cette formule peut être comprise intuitivement : lors du premier tirage, j'ai le choix entre un élément parmi N ; lors du second tirage, je n'ai plus le choix que parmi N-1 ; lors du troisième tirage, je n'ai le choix qu'entre N-2 ; etc. jusque 1. Dès lors, le nombre de séquences correspond à Nx(N-1)x(N-2)x…x1, c'est-à-dire N!. Dans nos exemples, 2! = 2 et 3!= 3x2 = 6. Si je vous demandais maintenant de faire l'exercice avec la fourchette et le couteau destinés au plat d'entrée, la fourchette et le couteau destinés au plat principal, la cuillère à soupe, la fourchette à dessert et la cuillère à café, vous devriez prévoir 7! = 5040 séquences. Si vous

cherchez une solution facile pour rompre votre couple, invitez votre conjoint au restaurant, effectuez cet exercice sans lui parler et surtout pas pour lui dire que l'idée vient de moi.

3.5.2.2. Arrangements

Les arrangements obéissent au même principe que les permutations. La différence est qu'ici on ne s'occupe pas de l'entièreté des objets mais bien d'une partie. On dit que l'on arrange *r* objets parmi *N*. L'**ordre compte et il n'y a pas de remise**, comme dans le cas des permutations. Par exemple, si je vous demandais de combien de manières il est possible d'arranger trois lettres parmi vingt-six (mais cette fois sans répétition) puisqu'il n'y a pas de remise, le résultat serait différent de l'exemple précédent où je demandais combien de séquences de trois lettres pouvaient être créées. Dans l'exemple précédent, il était possible d'obtenir "*aaa*". Ici, le fait de tirer un "*a*" en première position empêche d'avoir un autre "*a*" en deuxième position. Chaque lettre n'a donc plus une chance de 1/26 d'être tirée mais bien une chance de 1/25, puisque le "*a*" est retiré des possibilités. Le nombre de manières de sélectionner et d'arranger *r* objets pris parmi *N* objets distincts est donné par :

Nombre d'Arrangements possibles

$$A_N^r = \frac{N!}{(N-r)!}$$

donc, dans le cas qui nous occupe, A = 26!/23! = 15600.

Prenons un exemple plus simple afin de pouvoir dénombrer facilement le résultat et vérifier la pertinence du calcul : supposons que je sois canadien passionné par le Bénélux (ne souriez pas : il y a des gens qui sont passionnés par les emballages de cigarettes ou le curling et moi j'aime les statistiques). Je veux établir un itinéraire aléatoire pour le visiter. Etre canadien n'a aucune importance, sauf qu'un Belge a nécessairement commencé par la Belgique. Il se trouve que je n'ai le temps de visiter que deux des trois contrées. J'ai donc un choix de deux parmi trois options (la Belgique, B, la Hollande, NL et le Luxembourg, L). Mes possibilités sont de visiter : B-NL ; B-L ; NL-B ; NL-L ; L-NL ; L-B, soit 6 possibilités. La formule donnerait 3!/1! = 3x2/1 = 6 possibilités également.

La logique de cette formule n'est qu'une extension de celle des permutations. En fait, un arrangement n'est rien d'autre qu'une permutation qui s'arrête après une sélection de *r* éléments. Nous avons vu qu'une permutation se calcule en prenant *N*!. Ici, le processus s'arrête après *r* sélections, et donc il resterait (*N*-*r*) éléments à prélever pour terminer la permutation. Ces éléments n'étant pas pris en compte, il est nécessaire de les ôter du dénombrement en simplifiant les termes concernés, c'est-à-dire en divisant la permutation par les éléments résiduels. Dès lors, nous obtenons N!/(N-r)!.

3.5.2.3. Les combinaisons

Les combinaisons consistent à sélectionner une partie d'un échantillon **sans remise** (tirages dépendants) mais dans une situation où **l'ordre ne compte pas**. Prenons comme illustration l'échantillonnage. Supposons que je définisse les BA1 comme ma population, et que vous êtes 500. J'aimerais connaître la taille moyenne des étudiants qui la compose. Mais je n'ai aucune envie de mesurer tout le monde, et je me dis qu'en mesurer 50 me permettrait déjà de me faire une idée de la taille moyenne de la population. Question : combien de possibilités d'échantillons différents me sont offertes?

Dans ce cas, que je choisisse d'abord l'étudiant numéro 1 et puis le numéro 7 (par exemple) ou bien d'abord le 7 et puis le 1 importe peu (je vous désignerais bien par autre chose qu'un numéro, mais je devrais adapter mon syllabus tous les ans, ne vous offensez donc pas). Au bout du compte, j'ai mesuré mes 50 étudiants et je ne me soucie pas de savoir dans quel ordre. Si l'ordre ne compte plus, cela signifie que tous les ordres possibles correspondant à une configuration donnée ne comptent plus que comme une seule séquence. Le tout est donc d'être capable de calculer cet effet d'ordre et de l'ôter du dénombrement.

Avant de calculer la réponse à mon problème d'échantillonnage, choisissons un cas plus simple que nous pouvons facilement visualiser. Supposons une classe plus petite de 3 personnes (A, B et C, ce n'est pas beaucoup mieux qu'un numéro, mais je progresse), et choisissons-en deux, indépendamment de l'ordre mais sans remise. Les possibilités sont : AB ; AC ; BA ; BC ; CA ; CB. Cependant, AB et BA ne compte que pour une séquence, AC et CA également, de même pour BC et CB. Il nous reste donc trois séquences possibles au lieu des 6 que nous avions lorsque l'ordre comptait. J'ai donc divisé mon nombre de séquences possibles par deux. En exercice, je vous enjoins à réaliser ce calcul pour un échantillonnage de trois personnes parmi quatre. Si vous calculez bien, vous vous rendrez compte que, cette

fois, vous avez réduit le nombre de possibilités par 6 en ne tenant pas compte de l'ordre. En fait, vous réduirez toujours vos possibilités d'un coefficient égal à *r*! c'est-à-dire le nombre de permutation des *r* objets (une autre manière d'énoncer ce raisonnement est que : puisque l'ordre importe peu, cela signifie que la manière dont on permute les *r* objets importe peu et donc qu'il faut ôter toutes les permutations possibles de ces *r* objets, c'est-à-dire *r*!). Nous obtenons donc la formule :

Nombre de Combinaisons possibles

$$C_N^r = \frac{N!}{(N-r)!\,r!}$$

Egalement noté $\binom{N}{r}$, au lieu de C_N^r

Cela correspond donc bien à un arrangement dont on a ôté les permutations possibles des *r* objets. Nous pouvons maintenant calculer le nombre d'échantillons de 50 étudiants possibles dans votre classe de 500 étudiants : 500!/[(500-50)! x 50!] = 2,42 10^{68}. Vous comprendrez facilement que si lors d'un TP deux étudiants ont recueilli des données identiques et prétendent être tombés par hasard sur le même échantillon, l'enseignant (pourvu qu'il aime les statistiques) vous répondra qu'il y a 1 chance sur 2,42 10^{68} que vous n'ayez pas triché. Il peut donc annuler votre épreuve avec la conscience tranquille : le risque d'injustice est sans doute plus faible que le risque que le plafond de l'auditoire lui tombe sur la figure.

Remarquez qu'il n'y a aucun sens à se demander ce qu'il se passerait si on cherchait à combiner l'entièreté de la population, donc une combinaison de *N* parmi *N*. En effet, dans ce cas, puisque les tirages sont dépendants et que l'ordre ne serait pas important, cela conduirait nécessairement à une seule possibilité. Imaginons la combinaison de 3 étudiants (A, B et C), j'obtiendrais : ABC ; ACB ; BAC ; BCA ; CAB ; CBA. Mais si l'ordre n'a pas d'importance, ces 6 séquences seraient équivalentes puisqu'elles contiennent toutes A, B et C et que seul l'ordre varie. Nous parviendrions à la même conclusion en utilisant la formule et en n'oubliant pas que, par convention 0! = 1 : 3!/[(3-3)! 3!] = 3!/(0! 3!) = 3!/3! = 1

3.5.2.4. Synthèse

Les permutations, arrangements et combinaisons concernent toujours des événements dépendants (tirage sans remise). L'ordre peut être important (permutation et arrangement) ou pas (combinaison).

Permutations	**Arrangements**	**Combinaisons**
de N objets	de r objets pris parmi N	de r objets pris parmi N
L'ordre des objets est pertinent	**L'ordre des objets est pertinent**	**L'ordre des objets est non pertinent**
$N!$	$A_N^r = \frac{N!}{(N-r)!}$	$C_N^r = \frac{N!}{(N-r)! \times r!}$
Si $N = 7$,	Si $N = 7$ et $r = 4$,	Si $N = 7$ et $r = 4$,
$N! = 5040$	$\frac{N!}{(N-r)!} = \frac{5040}{6} = 840$	$\frac{N!}{(N-r)! \times r!} = \frac{5040}{6 \times 24} = 35$

3.6. Exercices de fin de chapitre

T.P. 2 - 3 : CHAPITRE 3

A. Probabilités et Ensembles

Exercice 1 : Tirage d'une carte

On tire une carte au hasard dans un jeu ordinaire de 52 cartes.

1. Comment caractérise-t-on l'action de tirer au hasard une carte de ce jeu. Donnez-en la définition.

2. Comment s'appelle l'ensemble des événements élémentaires possibles auxquels on s'intéresse ?

3. Quel est ici l'espace-échantillon ?

4. Par quelle lettre grecque désigne-t-on l'espace-échantillon ?

5. Que vaut $P(\Omega)$?

6. Quelle est la probabilité de prélever une carte rouge. S'agit-il d'un événement élémentaire ou d'un événement composé.

7. Quelle est la probabilité de prélever un trèfle. S'agit-il d'un événement élémentaire ou d'un événement composé.

8. Quelle est la probabilité de prélever un as ? S'agit-il d'un événement élémentaire ou d'un événement composé.

9. Quelle est la probabilité de prélever un roi noir ? S'agit-il d'un événement élémentaire ou d'un événement composé.

10. Quelle est la probabilité de prélever une dame de cœur ? S'agit-il d'un événement élémentaire ou d'un événement composé.

11. Quelle définition de la probabilité (formule) avez-vous utilisé pour répondre aux questions précédentes ?

12. Quelles en sont les conditions d'utilisation ?

13. Ces conditions sont-elles vérifiées dans notre exemple ?

14. Quand parle-t-on d'évenements complémentaires ? Quelles sont les propriétés de deux évenements complémentaires ?

T.P. 2 – 3 : Partie A

Exercice 2 : Familles de trois enfants

Supposons que, dans les familles de trois enfants, les 8 cas suivants soient également possibles : GGG, GGF, FGG, GFG, GFF, FGF, FFG, FFF (où, par exemple, GGF représente l'événement « avoir un garçon pour aîné, un garçon comme deuxième enfant et une fille comme benjamine »). Trouvez les probabilités des événements suivants :

1. avoir exactement un garçon

2. avoir un garçon comme aîné

3. avoir un garçon comme aîné et une fille comme benjamine

4. avoir exactement deux garçons

5. avoir au moins un garçon

6. avoir au moins deux garçons

7. avoir au plus un garçon

8. avoir plus de garçons que de filles

9. avoir au moins une fille et au moins un garçon

10. n'avoir aucune fille plus jeune qu'un garçon

T.P. 2 – 3 : Partie A
Exercice 3 : Probabilité d'un événement composé d'alternatives inclusives

1. Quand peut-on dire que P(AUB) = P(A)+P(B) ? Est-ce le cas ici ?

2. Représentez A ∪ B à l'aide d'un diagramme de Venn et donnez la formule permettant de calculer la probabilité de A ∪ B.

3. Quel est le nom donné à cette règle dans le cours ?

4. Etendez cette règle à A ∪ B ∪ C (indice : A ∪ B ∪ C = A ∪ (B ∪ C)). Confirmez votre raisonnement en représentant cette nouvelle situation sur un diagramme de Venn.

5. Montrez que le résultat obtenu en 2 peut être déduit par calcul du résultat établi en 1.

T.P. 2 - 3 : Partie A
Exercice 4 : Combinaison d'événements aléatoires

On considère deux événements aléatoires A et B. On connaît les probabilités suivantes :

$$P(\sim A) = 2/3 \; ; \; P(B) = 2/3 \; ; \; P(A \cap B) = 1/4$$

On demande de calculer les probabilités des événements composés suivantes en indiquant la formule théorique utilisée. La réponse à la première question est donnée à titre d'exemple.

Remarque : vous pouvez utiliser, si nécessaire, les lois de Morgan :

$$\sim(A \cup B) = \sim A \cap \sim B$$

$$\sim(A \cap B) = \sim A \cup \sim B$$

1. Dressez pour les événements A et B un tableau de contingence semblable à celui de la page 24 pour mieux vous aider à visualiser ces événements.

2. Représentez ces mêmes données sous la forme d'un diagramme de Venn en indiquant dans chaque zone la probabilité correspondante.

3. Calculez les probabilités suivantes à l'aide de formules.

Probabilité demandée	Formule théorique utilisée	Réponse numérique
$P(\sim B)$		
$P(A)$		
$P(A \cup B)$		
$P(A-B)$		
$P(A \Delta B)$		
$P(\sim A \cap \sim B)$		

T.P. 2 - 3 : Partie A

Exercice 5 : Probabilités conditionnelles et indépendance

Un athénée offre aux étudiants du dernier degré les options suivantes : a) math – sciences (**MS**), b) sciences – langues modernes (**SL**), c) langues classiques et modernes (**LL**), d) sciences humaines (**SH**). Les élèves se repartissent donc en fonction de quatre sections et en fonction du sexe de la manière suivante :

	MS	SL	LL	SH	
Filles	2	11	6	14	
Garçons	5	8	4	7	

Supposons que l'on choisisse un candidat au hasard.

On définit les événements suivants :

- F : l'élève choisi est une fille ;
- MS : l'élève choisi suit l'option MS ;
- SL : l'élève choisi suit l'option SL ;
- SH : l'élève choisi suit l'option SH.

1. 1. Calculez les sommes marginales afin de compléter les cases vides du tableau ci-dessus.

2. 2. Calculez les probabilités des événements suivants (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées).

P(F) =
P(MS) =

P(SL) =
P(SH) =
P(G) =
P(LL) =

3. Peut-on appeler les probabilités ci-dessus des probabilités : (a) indépendantes? ; (b) marginales? ; (c) conditionnelles? Pourquoi?

4. Montrez que P(F) et P(G) sont complémentaires.

5. Exprimez en français les événements suivants et calculez leur probabilité (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées).

MS ∩ F : P(MS ∩ F) =
SL ⎪ F : P(SL ⎪ F) =
F ⎪ SH : P(F ⎪ SH) =

6. Les paires d'événements suivants sont-ils indépendants ou non ? Justifiez vos réponses à l'aide d'une formule en utilisant uniquement les probabilités calculées aux points 1 et 2.

Evénements	Indépendants ou non ?	Justification
MS et F		
SL et F		
SH et F		

Déterminez, à partir des résultats des points 1, 2 et 6, les probabilités des événements suivants (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées et justifiez en indiquant la formule utilisée).

Probabilité	Résultat numérique	Formule
$P(SL \cap F)$		
$P(SL \mid \sim F)$		

T.P. 2 - 3 : Partie A

Exercice 6 : Probabilités conditionnelles et indépendance

On tire au hasard une carte d'un jeu de 52 cartes. On considère les événements suivants :

R : tirer un ROI C : tirer un COEUR

1. Quel concept de la théorie des probabilités permet de dire que la proposition : $P(C) + P(R) = 1$ est fausse ? Expliquer pourquoi cettre proposition est fausse.

2. Calculez les probabilités des événements suivants (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées) :

$P(R)=$	$P(C)=$	$P(R \cap C)=$	$P(R \cup C)=$

3. Les événements R et C sont-ils compatibles ($P(Z \geq 1)$) ou incompatibles ($P(Z = 1) = 0$). Ces 2 événements sont-ils indépendants? Justifiez vos réponses et dressez le tableau correspondant.

On répète la même expérience en tirant la carte d'un jeu auquel on a ajouté deux jokers. (Ce jeu comporte 54 cartes).

4. Calculez les probabilités des événements suivants (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées) :

P(R) =	P(C) =	P(R∩C) =	P(R∪C) =

5. Les événements R et C sont-ils compatibles ? Sont-ils indépendants? Justifiez vos réponses et dressez le tableau correspondant.

On répète la même expérience en tirant la carte d'un jeu incomplet, dont on a retiré le roi de coeur. (Ce jeu comporte donc 51 cartes).

6. Calculez les probabilités des événements suivants (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées) :

P(R) =	P(C) =	P(R∩C) =	P(R∪C) =

7. Les événements R et C sont-ils compatibles ? Sont-ils indépendants? Justifiez vos réponses

T.P. 2 - 3 : Partie A

Exercice 7 : Propositions concernant des probabilités d'événements aléatoires

Pour chacune des propositions suivantes, mettez une croix dans la case correspondant à la bonne réponse.

Proposition	Toujours vraie	Parfois vraie	Jamais vraie
La probabilité de la réalisation simultanée de deux événements est supérieure aux probabilités de réalisation de chacun de ces événements			
La probabilité de la non réalisation d'un événement est inférieure à la probabilité de réalisation de celui-ci.			
La somme des probabilités de deux événements incompatibles est inférieure ou égale à 1.			
La somme des probabilités associées à n événements est inférieure à 1.			
La somme de la probabilité de la réalisation d'un événement et de la probabilité de la non réalisation du même événement est égale à 1.			
Deux événements compatibles sont indépendants			
Deux événements indépendants sont incompatibles			
La somme des probabilités associées aux événements élémentaires de l'ensemble fondamental est égale à 1.			
La probabilité de la réunion de n événements mutuellement exclusifs est supérieure à 1.			
La probabilité de la réalisation simultanée de n événements est supérieure aux probabilités de réalisation de chacun de ces événements.			
Deux événements indépendants et réalisables sont compatibles.			

La somme des probabilités de deux événements indépendants est égale à la somme de la probabilité de la réunion de ces 2 événements et du produit des probabilités de chacun de ces événements.			

T.P. 2 - 3 : Partie A

Exercice 8 : Hommes et femmes, européens et non européens

Un groupe de personnes est composé de 20 hommes (dont 10 européens) et de 30 femmes (dont 20 européennes). Si l'on choisit au hasard une personne dans ce groupe, déterminez la probabilité pour qu'elle soit :

1. de sexe féminin
2. de sexe masculin
3. de nationalité européenne
4. un homme de nationalité non européenne

T.P. 2 - 3 : Partie A

Exercice 9

1626 personnes diplômées au cours de ces quelques dernières années ont été classées suivant le sexe d'une part, le niveau du diplôme obtenu le plus élevé d'autre part (il s'agit de données recueillies aux Etats-Unis). Les résultats sont présentés dans le tableau de contingence ci-dessous.

	Diplôme de bachelor	Diplôme de master	Formation professionnelle	Doctorat	Total
Homme	529	171	44	26	
Femme	616	194	30	16	
Total					

On tire au hasard une personne parmi les 1626 sujets. Soient les événements :

H = "La personne sélectionnée est un homme";
F = "La personne sélectionnée est une femme";
BA = "La personne sélectionnée a obtenu un diplôme de bachelor";
MA = "La personne sélectionnée a obtenu un diplôme de master";
PR = "La personne sélectionnée est diplômée d'une formation professionnelle";
DO = "La personne sélectionnée a obtenu un diplôme de doctorat".

1. Retraduisez à l'aide des opérations ensemblistes usuelles appliquées aux différents événements définis ci-dessus, les événements ci-dessous et calculez-en la probabilité. Donnez vos réponses avec une précision de trois décimales.

a. "La personne sélectionnée est une femme."

Formulation mathématique :

Probabilité :

b. "La personne sélectionnée est une femme, sachant qu'elle a obtenu un diplôme de master."

Formulation mathématique :

Probabilité :

c. "La personne sélectionnée a un diplôme de bachelor ou est diplômée d'une formation professionnelle, sachant qu'il s'agit d'un homme."

Formulation mathématique :

Probabilité :

d. "La personne sélectionnée n'est pas diplômée d'une formation professionnelle."

Formulation mathématique :

Probabilité :

e. "La personne sélectionnée est une femme, sachant qu'elle a obtenu un diplôme de master ou de doctorat."

Formulation mathématique :

Probabilité :

2. Les variables « sexe » et « niveau de diplôme » sont-elles indépendantes l'une de l'autre ? En d'autres termes, la variable « niveau de diplôme » se distribue-t-elle de la même façon chez les hommes et les femmes ? Justifiez avec soin votre réponse.

T.P. 2-3 : Partie A

Exercice 10

Dans un magasin employant 10 caissières, l'une d'elles est malhonnête, les autres sont honnêtes.

Pour une caissière honnête, le risque d'obtenir un contenu de caisse supérieur au montant enregistré (suite à des erreurs fortuites) est de 0,02. Ce risque atteint 0,05 pour la caissière malhonnête

A la fin de la journée de travail, on prélève une caisse au hasard.

On définit les événements suivants :

M : la caisse provient de la caissière malhonnête ;
S : le montant de la caisse est supérieur au montant enregistré.

1. Exprimez en français les deux événements suivants :

~M	
~S	

2. Exprimez en français les probabilités suivantes et donnez-en leur valeur numérique :

		Expression	**Résultat**
	P(~M)		
	P(S ⎮ M)		
	P(~S|~M)		

3. Exprimez en français les probabilités suivantes, indiquez la formule permettant de les calculer et donnez-en leur valeur numérique :

		Expression	Formule	Résultat
	$P(M \cap S)$			
	$P(\sim M \cap S)$			
	$P(S)$			
	$P(M\|S)$			

4. Constatant que le montant de la caisse est supérieur au montant enregistré, le gérant du magasin en conclut que la caissière est malhonnête Que pensez-vous de cette conclusion ?

T.P. 2-3 : Partie A

Exercice 11

Une école professionnelle et technique offre les options suivantes : puériculture, carrosserie et travaux de bureau. Les propositions d'étudiants fréquentant les diverses sections sont reprises dans le tableau suivant :

Puériculture	Carrosserie	Travaux de bureau
35%	23%	42%

Dans la section puériculture, on observe 95% de filles;
dans la section carrosserie, on observe 90% de garçons;
dans la section travaux de bureau, on observe 60% de filles;

Un élève est choisi au hasard. On définit les événements suivants :

P : élève fréquente la section puériculture ;

C : élève fréquente la section carrosserie ;

B : élève fréquente la section travaux de bureau ;

F : élève est une fille;

G : élève est un garçon

1. Déterminez les probabilités des événements suivants :

P(P) =
P(C) =
P(B) =

2. Exprimez en français les probabilités suivantes et donnez-en leur valeur numérique:

		Expression	**Résultat**
	$P(F\|P)$		
	$P(G\|P)$		
	$P(F\|C)$		
	$P(G\|C)$		
	$P(F\|B)$		
	$P(G\|B)$		

3. Exprimez en français les probabilités suivantes, indiquez la formule permettant de les calculer et donnez-en leur valeur numérique:

	Expression	Formule	Résultat
$P(F \cap P)$			
$P(G \cap P)$			
$P(F \cap C)$			
$P(G \cap C)$			
$P(F \cap B)$			
$P(G \cap B)$			

4. Exprimez en français les probabilités suivantes, indiquez la formule permettant de les calculer et donnez-en leur valeur numérique:

	Expression	Formule	Résultat
P(F)			
P(F)			

5. Exprimez en français les probabilités suivantes, indiquez la formule permettant de les calculer et donnez-en leur valeur numérique:

	Expression	Formule	Résultat
$P(P \mid F)$			
$P(P \mid G)$			
$P(C \mid F)$			
$P(C \mid G)$			

P(B | F)			
P(B | G)			

T.P. 2-3 : Partie A

Exercice 12

Certains exercices sont inventés, d'autre tirés d'études existant, d'autres encore tirés du livre de Gauvrit (2005).

1. Dans le cadre d'une étude de Goldberg & al. (2003) sur la perception de l'infidélité dans les couples, les auteurs étudient l'amour propre dans un échantillon contenant 50% d'hommes et 50% de femmes. Au moyen d'un questionnaire, ils ont classé les personnes en trois groupes selon le type d'amour propre : certaines personnes placent leur amour propre essentiellement dans la fonction sexuelle (cat. 1), d'autres dans leur capacité à attirer les sentiments de l'autre (cat. 2) et une dernière catégorie (cat. 3) rassemblent les individus qui n'appartiennent à aucune des deux premières. Les effectifs sont les suivants :

	Cat. 1	Cat. 2	Cat. 3
Hommes	45	20	35
Femmes	23	47	30

   a. Les événements critiques « être un homme » et « appartenir à la première catégorie » sont-ils indépendants ?

   b. Plus généralement, la variable "genre" et la variable "catégorie" sont-elles indépendantes ?

   c. Au vu des résultats, si je vous dis que les hommes sont plus choqués que les femmes par une infidélité sexuelle et moins par une infidélité sentimentale, êtes-vous d'accord avec moi? Justifiez.

2. D'après une étude de Rozin & al. (2003), 22% de la population des Etats-Unis sont obèses contre 7% de la population française. Pourtant les Français mangent plus gras et ont plus de cholestérol. On choisit une personne au hasard dans une population comprenant une proportion de x Américains et de (1-x) Français. Quelle est la probabilité de tomber sur une personne obèse?

3. Gadeau et Billon-Galland (2003) ont distingué de nombreux motifs qui poussent les enseignants à signaler des enfants en difficulté à un psychologue scolaire. Dans 25% des cas, il s'agit d'un trouble de l'apprentissage, dans 11% des cas d'un trouble des relations avec les adultes et dans 8% des cas de facteurs familiaux (le reste étant divers). Considérez les tirages comme indépendants.

   a. Si l'on choisit deux élèves au hasard parmi ceux qui ont été signalés au psychologue scolaire, quelle est la probabilité qu'ils l'aient été tous deux pour des facteurs familiaux?

   b. On choisit 10 élèves, au hasard, parmi ceux qui ont été signalés. Les 10 élèves ont été envoyés au psychologue pour un trouble relationnel avec les adultes. Quelles étaient les chances que cela arrive?

4. Schmid Mast et Hall (2003) ont classé des volontaires en deux groupes selon qu'ils préfèrent commander (groupe 1) ou être commandés (groupe 2). Chacun des groupes est ensuite séparé en deux de manière aléatoire. La moitié des joueurs de chaque groupe jouent alors le rôle de subordonné dans un jeu de rôle. L'autre moitié des joueurs jouent le rôle du patron. On note ensuite si les personnes ont bien joué leur rôle (réussite) ou pas (échec) selon des critères décidés à l'avance. On note G la variable groupe et X la variable réussite. X vaut 0 en cas d'échec et 1 en cas de réussite. Dans le groupe des subordonnés, on observe que la probabilité de faire partie du groupe 1 et d'avoir réussi est plus petite que la probabilité d'appartenir au groupe 1 multipliée par la probabilité d'avoir réussi. En revanche, dans le groupe des patrons, la probabilité de faire partie du groupe 1 et d'avoir réussi est égale à la probabilité d'appartenir au groupe 1 multiplié par la probabilité d'avoir réussi. Traduisez ces phrases en langage algébrique et commentez la signification de ces résultats.

5. Soit le tableau de contingence suivant reprenant deux variables (par comportement prosocial on entend un comportement d'aide à autrui) :

	Comportements prosociaux fréquents	Comportement prosociaux rares	Comportements non prosociaux	Total marginal
Heureux				17
Pas heureux				27
	25	10	9	44

a. Quels effectifs absolus retrouverions-nous dans les différentes cellules en cas d'indépendance des variables ?

b. A quelles probabilités correspondent chacune des cellules (ainsi que les cellules marginales) ?

c. Les effectifs correspondant à l'indépendance ne sont pas possibles physiquement (étant donné qu'il serait nécessaire d'avoir des morceaux d'individus pour y parvenir précisément). Est-ce que cela signifie, en pratique, que les deux variables dépendent obligatoirement l'une de l'autre? Justifiez votre réponse de manière intuitive.

T.P. 2-3 : Partie A

Exercice 13

Entre 1950 et 1963, le psychosociologue Stanley Milgram a effectué plusieurs séries d'expériences en laboratoire autour du thème de la soumission à l'autorité. Ci-dessous, nous vous proposons quelques extraits de son livre sorti en français, en 1974, aux Editions Calmann-Lévy sous le titre « Soumission à l'autorité ».

Nous n'évoquerons pas ici les questions éthiques relatives à ces expérimentations ni les conclusions qui pourraient en être tirées.

« L'obéissance à l'autorité, longtemps prônée comme une vertu, revêt un aspect différent quand elle est au service d'une cause néfaste ; la vertu se mue alors en vice odieux. […]

Afin d'analyser avec précision l'acte d'obéissance, j'ai réalisé à l'Université de Yale une expérience simple. […] Une personne vient dans un laboratoire de psychologie où on la prie d'exécuter une série d'actions qui vont entrer progressivement en conflit avec sa conscience. La question est de savoir jusqu'à quel point précis elle suivra les instructions de l'expérimentateur avant de se refuser à exécuter les actions prescrites. […]

L'expérimentateur [lui] explique qu'il s'agit d'étudier les effets de la punition sur le processus d'apprentissage. Il emmène l'*élève* dans une pièce, l'installe sur une chaise munie de sangles qui permettent de lui immobiliser les bras pour empêcher tout mouvement désordonné et lui fixe une électrode au poignet. Il lui dit alors qu'il va avoir à apprendre une liste de couples de mots ; toutes les erreurs qu'il commettra seront sanctionnées par des décharges électriques d'intensité croissante. **Le véritable sujet de l'expérience, c'est le *moniteur***.

Après avoir assisté à l'installation de l'*élève*, il [le sujet naïf, le *moniteur*] est introduit dans une salle principale du laboratoire où il prend place devant un impressionnant simulateur de chocs. Celui-ci comporte une rangée horizontale de trente manettes qui s'échelonnent de quinze à quatre cent cinquante volts par tranches d'augmentation de 15 volts et sont assorties de mentions allant de CHOC LÉGER à ATTENTION : CHOC DANGEREUX. On invite alors le *moniteur* à faire passer le test d'apprentissage à l'*élève* […]. Quand celui-ci répondra correctement, le *moniteur* passera au couple de mots suivant. Dans le cas contraire, il devra lui administrer une décharge électrique en commençant par le voltage le plus faible (quinze volts) et en augmentant progressivement d'un niveau à chaque erreur (trente volts, quarante-cinq volts, ainsi de suite).

Le *moniteur* est un sujet absolument « naïf », venu au laboratoire pour participer à une expérience. Par contre, l'*élève*, ou victime, est un acteur qui ne reçoit en réalité aucune décharge électrique. L'expérience a pour objet de découvrir jusqu'à quel point un individu peut pousser la docilité dans une situation concrète et mesurable où il reçoit l'ordre d'infliger un châtiment de plus en plus sévère à une victime qui proteste énergiquement. A quel instant précis le sujet refusera-t-il d'obéir à l'expérimentateur ?

Le conflit surgit quand l'*élève* commence à donner des signes de malaise. À soixante-quinze volts, il gémit. À cent vingt volts, il formule ses plaintes en phrases distinctes. À cent cinquante volts, il supplie qu'on le libère. À mesure que croît l'intensité des décharges électriques, ses protestations deviennent plus véhémentes et pathétiques. À deux cent quatre-vingt-cinq volts, sa seule réaction est un véritable cri d'agonie. »

Cette expérience a été reproduite sous différentes conditions en fonction de la proximité ou de la distance entre le *moniteur* et l'*élève*.

Expérience 1 : Aucune plainte vocale de l'*élève* n'est perceptible par le *moniteur*.
Expérience 2 (Feedback vocal) : même situation, mais le *moniteur* entend les protestations vocales de l'*élève*.
Expérience 3 (Proximité) : Même expérience, mais le *moniteur* et l'*élève* sont dans la même pièce, à quelques dizaines de centimètres l'un de l'autre.
Expérience 4 (Contact) : Pour recevoir la décharge l'*élève* doit placer sa main sur une plaque. S'il refuse de poser sa main sur la plaque, il faut un contact physique entre le *moniteur* et lui pour l'y contraindre.

Quarante adultes ont participé à chacune des expériences, soit un total de 160 personnes.

Voici le tableau des chocs maxima administrés par les *moniteurs* dans chaque expérience :

Niveau de choc	Mention verbale et niveau de voltage	Expérience 1 Feedback à distance (n=40)	Expérience 2 Feedback vocal (n=40)	Expérience 3 Proximité (n=40)	Expérience 4 Contact (n=40)
	Choc léger				
1	15				
2	30				
3	45				
4	60				
	Choc modéré				
5	75				
6	90				
7	105			1	
8	120				
	Choc fort				
9	135		1		1
10	150		5	10	16
11	165		1		
12	180		1	2	3

	Choc très fort				
13	195				
14	210				1
15	225			1	1
16	240				
	Choc intense				
17	255				1
18	270			1	
19	285		1		1
20	300	5	1	5	1
Choc extrêmement intense					
21	315	4	3	3	2
22	330	2			
23	345	1	1		1
24	360	1	1		
Attention : choc dangereux					
25	375	1		1	
26	390				
27	405				
28	420				
	XXX				
29	435				
30	450	26	25	16	12

À partir des données réelles de l'expérience rapportées dans le tableau ci-dessus, imaginons que nous tirions au sort un sujet pour l'interroger sur l'effet qu'a pu avoir la participation à une telle expérience sur sa vie dans les années qui ont suivi.

Soient les événements :

- V = « Le sujet sélectionné a refusé catégoriquement à un moment ou à un autre d'obéir à l'expérimentateur » ;
- A = « Le sujet sélectionné a participé à l'expérience 1 » ;
- B = « Le sujet sélectionné a participé à l'expérience 2 » ;
- C = « Le sujet sélectionné a participé à l'expérience 3 » ;
- D = « Le sujet sélectionné a participé à l'expérience 4 » ;

- I = « Le sujet sélectionné a été jusqu'à administrer un choc intense à potentiellement mortel. ».

1. Retraduisez à l'aide des opérations ensemblistes usuelles appliquées aux différents événements les événements ci-dessous et calculez-en la probabilité. Donnez vos réponses avec une précision de deux décimales. Pour les questions b, d et e, nous vous demandons de donner, outre la réponse, la formule que vous avez utilisée (ou que vous pourriez utiliser si vous avez procédé autrement).

a.	« Le sujet sélectionné a administré un choc de 450 volts à l'*élève*. » Formulation mathématique : Probabilité :
b.	« Sachant que le *moniteur* et l'*élève* étaient dans la même pièce, le sujet sélectionné a infligé à l'*élève* un choc de 450 volts.» Formulation mathématique : Probabilité :
c.	« Le sujet sélectionné a participé à l'expérience 1 et à l'expérience 4. » Formulation mathématique : Probabilité :
d.	« Sachant que le sujet sélectionné a participé à l'expérience 2, il s'est arrêté avant que le choc ne puisse être qualifié d'intense. »

	Formulation mathématique : Probabilité :
e.	« Le sujet sélectionné a dû avoir des contacts physiques avec l'*élève* et est allé jusqu'au bout de l'expérience. » Formulation mathématique : Probabilité :

2. Les événements A et V sont-ils indépendants ? Justifiez votre réponse à l'aide d'une formule et commentez votre constatation.

3. Commentez le résultat obtenu au point 1c.

T.P. 2-3 : Partie A

Exercice 14 : Probabilités - vrai ou faux ?

1. Vrai ou faux ? Si on obtient 6 fois pile sur 6 lancers, la probabilité d'obtenir face au lancer suivant devient très élevée. Pourquoi ?

2. Vrai ou faux ? Après un excès d'événements pile, la loi des grands nombres fait qu'un plus grand nombre d'événements face sera observé dans la suite de la série. Pourquoi ?

3. Vrai ou faux ? Après un grand nombre de jets, la taille de la différence entre le nombre d'événements face et le nombre attendu en fréquence absolue diminue.

4. Vrai ou faux ? Comparée au nombre de jets, la différence en fréquence relative entre le nombre de piles et de faces tend à devenir petite.

5. À quelle loi fait référence cette dernière affirmation ?

6. Vrai ou faux ? La différence entre la fréquence relative observée et la fréquence relative théorique tend vers o quand N tend vers l'infini.

7. Quel autre nom peut-on donner à la fréquence relative théorique ?

T.P. 2 - 3 : CHAPITRE 3

B. Notions d'analyse combinatoire : « counting rules »

Exercice 1 : Arrangements et permutations

1. On organise un tirage au sort pour déterminer l'ordre de passage d'un examen pour 3 étudiants. Nous appellerons ces 3 étudiants A, B et C pour plus de facilité.

2. Combien y a-t-il de possibilités pour la première position ?

3. Combien y a-t-il de possibilités pour les deux premières positions ? Donnez ces différentes possibilités. Quelle formule permet de calculer cela ?

4. Combien y a-t-il de possibilités pour l'ensemble des trois étudiants ? Donnez ces différentes possibilités. Quelle formule permet de calculer cela ?

6. Combien y a-t-il de possibilités pour le tirage au sort si le nombre d'élèves s'élève à 10 ?

T.P. 2 – 3 : Partie B

Exercice 2 : Principe de dénombrement et probabilités

Arrangements et combinaisons

Quatorze boules identiques, numérotées de 1 à 14, sont placées dans une urne.

1. Calculez le nombre de tierces possibles dans l'ordre et dans le désordre.

2. Quelle est la probabilité que les boules 1, 2 et 3 (dans l'ordre) soient tirées ? Et que la séquence 1, 2, 3 dans un ordre indifférent soit tirée ?

3. Calculez le nombre de quartes possibles dans l'ordre et dans le désordre.

4. Quelle est la probabilité que les boules 1, 2, 3 et 4 (dans l'ordre) soient tirées ?

5. Calculez le nombre de quintes possibles dans l'ordre et dans le désordre.

6. Calculez le nombre de sixtes possibles dans l'ordre et dans le désordre.

T.P. 2-3 : Partie B

Exercice 13 : Répartition hebdomadaire d'activités

Principe multiplicatif – combinaison – permutation

Un étudiant se propose d'établir un programme hebdomadaire en fixant ses activités pour chacune des 7 soirées de la semaine. Le but de l'exercice est de dénombrer ces programmes sous différentes contraintes.

1. De combien de manières différentes peut-il répartir trois soirées d'étude sur la semaine ?

2. Sachant que l'étudiant choisit de déterminer aléatoirement la répartition de ses trois jours d'étude quelle est la probabilité qu'il étudie lundi, mardi et mercredi.

3. Sachant que l'étudiant choisit de déterminer aléatoirement la répartition de ses trois jours d'étude, sans tenir compte de ses autres activités, quelle est la probabilité qu'il étudie lundi, mardi et mercredi ou lundi, mardi et jeudi ?

T.P. 2-3 : Partie B

Exercice 14

Stanley Milgram a réalisé une expérience étudiant la soumission à l'autorité en essayer de voir jusqu'où un sujet naïf pouvait obéir à un expérimentateur qui lui demandait d'infliger un choc électrique à un élève chaque fois qu'il commet une erreur dans une séquence d'apprentissage. L'élève est bien sûr un comédien-complice qui simule la douleur liée aux chocs).

NB : la liste des conditions ci-dessous n'est pas exhaustive.

Concernant l'effet d'influence de la personnalité de l'expérimentateur, Milgram a utilisé 2 conditions :

1° Expérimentateur = technicien assez sec et cassant

2° Expérimentateur = quelqu'un qui paraissait doux et pacifique

Concernant l'effet d'influence de la présence ou de l'absence de l'expérimentateur :

1° Expérimentateur assis à quelques dizaines de centimètres du sujet

2° Après avoir donné ses premières instructions, il quittait le laboratoire et donnait ses ordres par téléphone.

Concernant le lieu d'expérimentation, Milgram a étudié 3 conditions :

1° Un luxueux laboratoire de l'Université de Yale

2° Un laboratoire plus modeste dans le sous-sol du même immeuble ; laboratoire fonctionnel, mais assez simple.

3° Immeuble de bureaux à Bridge-Port

Concernant le sexe du sujet, Milgram a étudié 2 conditions :

1° Homme

2° Femme

Concernant l'engagement de la victime, il a étudié deux conditions :

1° engagement sans restriction énoncée

2° Engagement limité de la victime

Concernant le choix du niveau de choc électrique, deux conditions :

1° imposé

2° libre

Concernant la distance entre le sujet naïf et son « élève », Milgram a étudié 4 conditions :

1 : Aucune plainte vocale de l'*élève* n'est perceptible par le *moniteur*.

2 (Feedback vocal) : même situation, mais le *moniteur* entend les protestations vocales de l'*élève*.

3 (Proximité) : Même expérience, mais le *moniteur* et l'*élève* sont dans la même pièce, à quelques dizaines de centimètres l'un de l'autre.

4 (Contact) : Pour recevoir la décharge l'*élève* doit placer sa main sur une plaque. S'il refuse de poser sa main sur la plaque, il faut un contact physique entre le *moniteur* et lui pour l'y contraindre.

1. Combien de condition aurait potentiellement pu prendre son plan expérimental si on se limite aux conditions énoncées ci-dessus ? Qu'en pensez-vous ?

P A R T I E I I I

STATISTIQUES DESCRIPTIVES ET PREAMBULES A L'INFERENCE

CHAPITRE 4 : LES ECHELLES DE MESURE

4.1. Introduction

Jusqu'à présent, j'ai tenté de vous faire percevoir l'intérêt et les enjeux des statistiques au travers de ces trois premiers chapitres. Actuellement vous devriez avoir bien compris que :

a) Nous voulons décrire la réalité à partir de modèles qui la simplifient.
b) A un modèle correspond une erreur, c'est-à-dire une incertitude qui rend tout modèle probabiliste (par opposition à déterministe).
c) Vos deux objectifs sont que votre modèle soit le plus simple possible et que l'erreur soit la plus petite possible.
d) Pour décrire une situation, nous devons utiliser de l'information, que l'on appelle des variables. Différentes variables peuvent s'influencer (être dépendantes) ou pas (être indépendantes) les unes des autres.
e) L'être humain n'a accès à ces variables qu'en regardant autour de lui, c'est-à-dire en recueillant l'information dans le milieu. Cela implique qu'il fonctionne par induction et non par déduction pour établir ses modèles. Cette méthode inductive est risquée et l'enjeu est de pouvoir contrôler ce risque.
f) L'établissement de modèles prédictifs se fait étape par étape en vérifiant, petit à petit des hypothèses, c'est-à-dire des liens entre variables indépendantes et variables dépendantes. Ces hypothèses sont de plus en plus sûres au fur et à mesure de leur résistance aux tentatives de les invalider.

A ce stade, vous êtes donc capables de décortiquer une hypothèse et d'en générer par vous-même. Vous savez également distinguer une hypothèse d'une question et d'une croyance. Vous comprenez la distinction entre la logique inductive et déductive et savez que la logique inductive est incontournable mais doit être traitée avec prudence. De plus, vous avez en tête une petite ligne du temps qui vous permet de comprendre l'émergence de la pensée probabiliste. D'un point de vue mathématique, vous êtes à l'aise avec l'utilisation des calculs de probabilités dans les conditions que nous avons définies : les axiomes, les règles d'additivité et de multiplication, les probabilités conditionnelles et les critères d'indépendance. Vous savez, enfin, déterminer et appliquer la bonne méthode pour dénombrer les événements. Si vous n'en êtes pas là, vous n'avez pas suffisamment travaillé les chapitres précédents.

Nous allons maintenant nous intéresser à la collecte d'informations nécessaires à la modélisation. Le premier problème auquel il faut faire face est la mesure des variables. Imaginez que vous deviez mesurer une distance, par exemple, entre l'ULB et votre domicile. Vous pouvez estimer qu'ils sont à quelques kilomètres de distance comme vous pouvez me dire qu'ils sont séparés de 8627,3324 mètres l'un de l'autre. Vous pourriez également vous contenter de dire qu'ils sont dans la même commune ou dans la même ville. Décider de la manière dont vous allez mesurer vos variables est une décision importante et riche en implications.

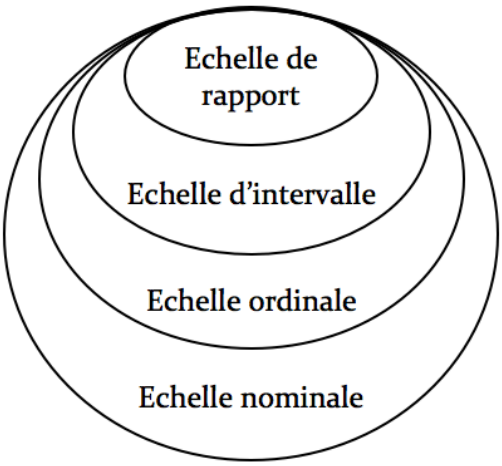
Règles d'Or

Lorsque vous désirez vérifier une hypothèse, vous aurez **PREALABLEMENT** réfléchi à la manière dont vous allez mesurer vos variables **et** au traitement statistique que vous allez appliquer ensuite. Ne faites **JAMAIS** l'erreur de réaliser une étude et de vous demander après coup comment traiter vos résultats, sans quoi je vous promets des heures d'ennui. Dites-vous qu'un chercheur qui utilise correctement les méthodes d'analyse de données trouvera vraisemblablement plus de résultats que celui qui bricole au petit bonheur la chance.

4.2. Mesure des variables

Mesurer une variable correspond à attribuer un code, le plus souvent chiffré, en veillant à ce que les propriétés de la variable soient conservées le mieux possible grâce aux propriétés du code utilisé. Nous distinguerons 4 échelles de mesure fondamentales : les échelles nominales, les échelles ordinales, les échelles d'intervalles et les échelles de rapport (Howell, 1999). **Les échelles nominales regroupent les mesures qualitatives, les autres les mesures quantitatives**. Toutes ces échelles ont une relation d'inclusion les unes par rapport aux autres, à savoir que : (a) elles peuvent toutes être considérées comme nominales ; (b) les échelles d'intervalle et de rapport peuvent être considérées comme des échelles ordinales ; (c) les échelles de rapport sont également des échelles d'intervalle (Figure 4.1.). Plus on s'approche de l'échelle de rapport, plus l'information est précise, mais il n'est pas possible de tout mesurer sur ce type d'échelle.

Figure 4.1 : Inclusion des propriétés des échelles de mesure



4.2.1. Les échelles nominales

Le code correspondant à ce type d'échelle ne tient compte que d'un lien d'appartenance d'un objet à une classe, ce qu'on appelle la **distinctivité**. On peut distinguer les éléments en fonction des catégories. Par exemple, la couleur des cheveux est souvent mesurée sous la forme : blond, brun, noir, blanc, roux, sans. Une attribution numérique n'impliquerait que l'attribution d'un symbole chiffré plutôt qu'une chaîne alphabétique à une classe donnée. Par exemple, si je conviens que le blond s'appelle "*1*", le brun "*2*" et le noir "*3*" je ne fais qu'établir une convention sans aucun sens mathématique. Je pourrais tout aussi bien décider que le noir s'appelle "*1*" et que c'est le blond qui s'appelle "*3*". Je pourrais également décider d'autres valeurs et dire que le blond s'appellera "*879*". Il n'y a aucun lien arithmétique entre ces classes. La valeur 2 n'a par exemple pas le statut de "*double de la valeur 1*", elle n'est même pas "*plus grande que la valeur 1*", elle est juste différente qualitativement.

Cette échelle n'est cependant pas inutile. D'une part, elle est souvent incontournable : comment mesurer, facilement, le genre d'un individu, l'appartenance à une religion, ou quelconque variable purement qualitative autrement? En outre, elle permet de classer les sujets en fonction de caractéristiques qualitatives mais néanmoins de développer des indices de tendance centrale ou de dispersion (voir plus bas). Par exemple, on peut dénombrer les sujets de chaque classe et établir la classe la plus représentée. Ou bien, envisageons une classification nominale binaire, comme : homme = -1 ; femme = 1. Si je somme tous les

individus pour lesquels j'ai l'information je peux avoir une idée grossière de la proportion d'hommes et de femmes dans le groupe de sujets (la valeur 0 indiquerait une parité, une valeur positive traduirait un surplus de femmes et une valeur négative indiquerait un surplus d'hommes). Mais il n'en reste pas moins que les informations contenues dans ce type d'échelle sont peu nombreuses.

4.2.2. Les échelles ordinales

Ces échelles sont légèrement plus informatives que les échelles nominales : l'**ordination** des chiffres importe. Supposons, par exemple, que je vous demande si vous aimez ou non mon cours, sur une échelle allant de 1 à 7. Si vous répondez 1 parce que vous n'aimez vraiment pas, cela signifie que vous aimez moins que si vous aviez répondu 2, mais cela ne signifie pas pour autant que vous aimez **à moitié** moins que si vous aviez répondu 2. En fait, la quantité qui sépare le 1 et le 2 n'est pas forcément la même que celle entre le 2 et le 3, elle est impossible à estimer si ce n'est son signe (positive ou négative).

Ces échelles sont parfois bien dissimulées et peuvent nous illusionner. Par exemple, une note sur 20 à mon examen pourrait donner l'impression que j'évalue vos compétences en statistiques et que j'estime qu'un 12 est deux fois mieux qu'un 6. Ce n'est évidemment pas le cas, au mieux je peux conclure qu'un étudiant en sait suffisamment pour continuer son parcours alors que son voisin ayant réalisé un 6 n'en sait pas assez par rapport à mon niveau d'exigence. Mais je ne peux pas réellement calculer la quantité de matière intégrée. En fait, pour mieux y parvenir je devrais être capable d'estimer le niveau de difficulté de chaque question. Une question facile recevrait un certain nombre de points, puis une question deux fois plus difficile devrait recevoir le double de points. Cependant, d'une part l'information relative à la difficulté d'une question est très difficile à mesurer, et d'autre part, il peut y avoir d'autres manières d'envisager la cotation. Par exemple, vaut-il mieux attribuer des points à une réponse qui montre que l'étudiant a intégré ce qui est important ou qu'il a intégré ce qui est difficile? Un autre indice que peut m'apporter l'étude de votre cote est votre position par rapport à l'ensemble des étudiants qui ont réalisé l'examen. Nous reviendrons sur ces notions ultérieurement.

4.2.3. Les échelles d'intervalle

Cette échelle acquiert un niveau supplémentaire d'information. Ici la distance entre deux unités est constante. De nombreuses variables quantifiables sont mesurées sur ce type d'échelles. C'est le cas, par exemple de la température exprimée en degrés Celsius. Cette échelle est basée sur la température à laquelle l'eau gèle et à laquelle l'eau bout. On pourrait tout à fait changer les conventions de ces échelles. C'est d'ailleurs le cas des systèmes de mesure anglo-saxons qui travaillent, par exemple, en degrés Fahrenheit pour la température. Cela ne les empêche pas de déterminer la température extérieure. Ils peuvent également utiliser ces valeurs dans des calculs complexes de problèmes physiques de la même manière que nous, sans qu'aucun problème ne survienne (en tout cas pas plus que si on utilise les degrés Celsius). C'est là un énorme avantage, à tel point que la majorité des calculs statistiques que nous effectuerons demandent que les variables soient mesurées sur une échelle d'intervalle au minimum. Nous verrons qu'en psychologie, on y parvient rarement, mais que dans certains cas on se permet de traiter des échelles ordinales comme s'il s'agissait d'échelles d'intervalle (et que ça marche très bien). L'une des manières de s'en sortir est l'étalonnage. Je vous renvoie à votre cours de psychologie différentielle pour traiter ce problème.

En revanche, les échelles d'intervalle ne donnent pas certaines informations : par exemple, on ne peut pas dire qu'il fait deux fois plus chaud lorsqu'il y 30°C que lorsqu'il y a 15°C. En effet, lorsqu'il y a 0°C je ne suis pas au zéro absolu, il y a tout à fait moyen d'obtenir des températures négatives. C'est l'information supplémentaire que donnent les échelles de rapport, comme nous allons le voir.

4.2.4. Les échelles de rapport

Ces échelles apportent un dernier niveau d'information : le zéro absolu. Dans les échelles précédentes le zéro ne signifiait rien de particulier, si ce n'est par convention. Par exemple, lorsque je dis qu'il fait 0°C dehors, je ne suis pas en train de dire qu'il n'y a pas de température, je suis juste en train de dire que si vous avez oublié un verre d'eau sur votre appui de fenêtre vous pouvez vous attendre à ce qu'il soit gelé. En revanche, si je mesure la température en Kelvin et que je vous annonce qu'il fait 0 degré dehors, soyez plus alarmés : je suis en train de vous dire qu'il n'y a plus de température, donc plus de pression, donc plus de volume, donc que vous êtes dans le vide intersidéral. Le zéro en Kelvin représente ce

qu'on appelle le zéro absolu, il n'y a pas de température négative sur cette échelle, alors qu'il y en a sur l'échelle en degrés Celsius. En outre, dans une échelle de rapport, si je dis qu'il fait 300K de température, je suis réellement en train de vous dire que la température est trois fois supérieure à celle de l'endroit où il ne fait que 100K. Ce n'était pas le cas pour la température en Celsius. De nombreuses variables sont mesurées sur de telles échelles, par exemples, le poids (ou la masse), la taille, le nombre d'enfants, etc.

Le temps répond également à cette définition : la seconde a été définie de la manière suivante :

> *"La seconde est la durée de 9 192 631 770 périodes de la radiation correspondant à la transition entre les deux niveaux hyperfins de l'état fondamental de l'atome de césium 133."* (Bureau International des Poids et Mesures, http://www.bipm.org/fr/si/si_brochure/chapter2/2-1/2-1-1/second.html retrouvé le 25/9/2011).

Dès lors, le temps écoulé entre, par exemple, midi et midi une seconde est le même que le temps écoulé entre midi une seconde et midi deux secondes. De même, la période écoulée en une seconde est exactement le cinquième de la période écoulée en cinq seconde. Pour peu que nous décidions d'attribuer un zéro au commencement d'un événement, nous nous trouvons sur une échelle de rapport. Par exemple, si deux individus naissent et que l'on décide que le moment de leur naissance est le zéro, puis que l'un meurt à 40 ans et l'autre à 80 ans, on peut dire que l'un a vécu deux fois plus longtemps que l'autre.

Remarquez qu'au sein de ces échelles il reste un niveau de complexité différent selon le cas. Par exemple, exprimer la taille en mètres n'est qu'un choix conventionnel. Les Anglo-Saxons utilisent les pieds et les pouces (feet and inches) au lieu des mètres et des centimètres et s'en sortent très bien : un pieds vaut 30,48 cm et un pouce vaut 2,54 cm. En revanche, changer l'échelle mesurant le nombre d'enfants ou le nombre d'étages d'un immeuble est absurde (même les Anglo-Saxons disent qu'ils ont UN enfant lorsque c'est le cas). Ce sous-type particulier d'échelle se nomme "***échelle absolue***". Elle a la caractéristique de n'admettre aucune transformation linéaire : il est en effet inutile de vouloir changer la convention en rajoutant 2 à chaque étage par exemple, ça n'aurait aucun sens d'avoir un rez-de-chaussée qui se nomme "*deuxième étage*", un premier étage qui se nomme "*troisième étage*" etc.. Sauf , pour une raison curieuse, à l'ULB!

4.2.5. Résumé des caractéristiques

Propriétés	Nominale	Ordinale	Intervalle	Rapport
Distinctivité	oui	oui	oui	oui
Ordination	non	oui	oui	oui
Intervalles égaux	non	non	oui	oui
Zéro absolu	non	non	non	oui

4.3. Exercices de fin de chapitre

T.P. 1 : CHAPITRE 4

Exercice 1 : Les échelles de mesure

1. Un instituteur note l'ordre dans lequel ses élèves terminent leur interrogation. Le premier à finir, le deuxième… Quelle échelle de mesure est utilisée ?

2. Dans une étude sur la perception des expressions faciales, les participants doivent classer les photographies qu'on leur montre dans une des 3 catégories d'expression émotionnelle suivantes : triste, joyeux ou neutre. Sur quelle échelle est mesurée l'expression émotionnelle ?

3. Un questionnaire porte sur (a) l'âge, (b) le genre, (c) la profession, (d) la taille, (e) le nombre d'enfants d'une personne. Quelles sont les échelles utilisées pour ces différentes mesures ?

4. Imaginez une variable mesurée au minimum sur une échelle ordinale mesurant une caractéristique psychologique d'un individu (par exemple son QI, ou son courage). Nommez la caractéristique que vous mesurez et déterminez le type d'échelle que vous utilisez.

T.P. 1 : CHAPITRE 4

Exercice 2 : Échelles de mesures et Ensembles

1. Représentez les échelles de mesure dans un diagramme de Venn.

2. Comment appelle-t-on en termes ensemblistes le lien qui unit ces différents types d'échelles de mesures ? Expliquez ce que cela signifie et donnez la notation correspondante.

3. Laquelle de ces échelles nous donne l'information le plus d'information ?

T.P. 1 : CHAPITRE 4

Exercice 3 : Les échelles de mesure - Exemples

1. Un chercheur a une liste avec la taille en cm d'un groupe d'enfants de 8 ans ; par exemple 123 cm, 146 cm, 130 cm… Quelle échelle de mesure est utilisée ?

2. Ce chercheur convertit ces tailles sur une nouvelle échelle en calculant la différence pour chaque enfant entre sa taille et la taille moyenne du groupe. Un enfant qui a exactement la taille moyenne du groupe a un score de 0 ; un enfant qui a 1 cm de plus que la taille moyenne du groupe obtient un score de +1 ; un enfant qui a 2 cm de moins obtient un score de -2… Quelle échelle est utilisée ?

3. Une étude porte sur l'influence de la couleur de la peinture des toilettes du bâtiment D. Des résultats ont montré que le temps d'occupation des toilettes, calculé en secondes, est supérieur lorsque les parois sont peintes en mauve que lorsqu'elles sont peintes en jaune

   a. Dans cet énoncé, veuillez identifier la variable indépendante et la variable dépendante.

   b. Pour chacune des variables, déterminez si elles sont quantitatives ou qualitatives.

   c. Sur quelle échelle va-t-on mesurer ces variables ?

4. Sur quelle échelle de mesure ces variables sont-elles habituellement mesurées ?

   N.B. : les échelles de mesure étant hiérarchisées, indiquez le niveau le plus élevé.

	Echelle nominale	Echelle ordinale	Echelle d'intervalles	Echelle de rapports	Echelle absolue
Groupe sanguin					
Nationalité					
Température en Kelvin					
Température en degrés Celsius					
Connaissance en informatique évaluée sur une échelle à quatre points.					
Couleur des yeux					
Heure à laquelle une personne va dormir					
Nombre d'heures de sommeil					
Latéralité					
Ancienneté					
Intelligence (Q.I.)					
Indice de masse corporelle (IMC)					
Degré de satisfaction sur une échelle de Lickert à 7 point					
Classement sportif					
Décours du temps selon un calendrier grégorien					

T.P. 1 : CHAPITRE 4
Exercice 4 : Hypothèses – variables dépendantes et indépendantes – échelles de mesures – modélisation

Un chercheur souhaite évaluer les effets de l'allaitement sur les performances intellectuelles, « pratiques » et sur les performances au travail des mères primipares après la reprise du travail par ces dernières quand leurs bébés sont gardés par des tiers et qu'elles tirent leur lait. Il constitue pour ce faire deux groupe de mères ayant un travail de bureau : celles qui a priori ont choisi d'allaiter à temps plein malgré la reprise du travail et celles qui ont a priori choisi de ne pas (plus) allaiter. Il sépare ensuite chacun de ces groupes en deux, la moitié de chacun des groupes bénéficiant d'une sieste sur leur lieu de travail (en plus de la pause allaitement) accordée par l'employeur et rémunérée (par un budget spécial attribué à des fins de recherche). Les femmes allaitantes peuvent prendre ce temps de sieste en une fois ou prolonger leurs pauses allaitement de deux fois une demi-heure afin de dormir un peu après avoir tiré leur lait.

Le chercheur en question pense en effet qu'il serait judicieux que les mères aient droit à des siestes dans un lieu aménagé sur leur lieu de travail, sans réduction de salaire. Il pense en effet que les performances durant le temps de travail restant seraient multipliées et que la quantité de travail ne serait pas moins importante malgré un temps de travail réduit par ces siestes.

Il mesure les performances intellectuelles de ces personnes par un test de logique et par un test de mémoire à court terme.

Le chercheur a pris contact avec des employeurs d'accord de participer à la recherche.

Les performances « pratiques » sont évaluées par le temps mis à réaliser une tâche ordinaire du quotidien : préparer une quiche selon une recette donnée par le chercheur, la mettre au four, donner le bain au bébé (y compris la préparation qui va avec) et faire tourner une

machine à laver et sortir des poubelles parce que, chez certaines d'entres elles, Bruxelles-Propreté passe le soir.
Le chercheur enregistre le temps total mis pour effectuer ces tâches ainsi que les temps partiels (éventuellement en plusieurs fois, qu'il additionnera à la fin pour une tâche donnée). (La réponse éventuelle au coup de téléphone de la belle-mère qui téléphone toujours au pire moment pour savoir si tout va bien et si le bébé est rentré vivant de la crèche où chaque jour il risque sa vie, n'étant pas comptabilisée dans le temps total).

La tétée d'un enfant qui surviendrait durant cette période serait aussi réduite. Ainsi que les moments où bébé pleurant, les mamans doivent s'en occuper autrement, également. Cependant ces temps sont enregistrés séparément afin de voir si les bébés ne sont pas également plus calmes le soir au retour de la crèche quand leur maman a eu la possibilité de se reposer dans la journée.

Les performances au travail sont évaluées par une tâche « générale » établie par le chercheur et commune à toutes les femmes, et une autre établie par l'employeur. Ces tâches sont évaluées sur base du temps mis pour les réaliser et de la qualité du travail évaluée par le nombre d'erreurs commises.

Accessoirement, la quantité de lait tirée est notée systématiquement. En effet, le chercheur se demande si les mères plus reposées n'auraient pas des quantités de lait augmentées.

Ces différentes épreuves seront passées :

- avant la grossesse quand la mère a été recrutée suffisamment tôt ;
- au deuxième, au cinquième et au huitième mois de grossesse ;
- deux semaines après la reprise du travail ;
- deux mois après la reprise du travail ;
- un mois après le début de la diversification ;
- deux mois après la fin de l'allaitement (différent pour chaque mère).

1. Quelles sont les deux grandes familles d'hypothèses ?

2. Quelles sont la (ou les) hypothèse(s) de ce chercheur (classez-les en fonction de la « famille » dont ils font partie (cf. supra) ?

3. Quelles sont les variables en présence ? Sur quelle échelle se mesurent-elles ?

4. Quelles sont les variables dépendantes et indépendantes données clairement dans l'énoncé ?

5. Quelles variables autres variables dépendantes ou indépendantes serait-il intéressant d'inclure dans la recherche ? Énoncez les nouvelles hypothèses éventuelles.

6. Dans ce contexte, modélisez, de la même manière que dans le syllabus, la réalité des performances d'une maman donnée au test de logique à un moment T en utilisant la moyenne de groupe comme modèle.

CHAPITRE 5 : EXPLORATION GRAPHIQUE DES DONNEES A UNE DIMENSION ET TERMINOLOGIE

5.1. Introduction

Imaginons que vous récoltiez de l'information (peu importe laquelle) auprès de 200 sujets, vous aurez devant vous un tableau de chiffres très lourd et indigeste (même pour vous). Si vous avez un message à faire passer à l'aide de ce tableau, il n'y a aucune chance que vous y parveniez en l'état. Vous allez devoir simplifier cette information pour la rendre compréhensible. Si nous reprenons notre exemple du début, à propos de la taille des hommes et des femmes adultes, vous avez fini par sélectionner 100 hommes et 100 femmes dont vous avez mesuré la taille. Ce qui signifie que vous avez récolté deux cents valeurs peu parlantes en elles-mêmes. L'une des manières de présenter avantageusement vos données est de se centrer sur une approche visuelle. Nous allons envisager différentes méthodes graphiques de représentation des données, leurs avantages et leurs inconvénients.

En revanche, lorsque vous dites que la taille moyenne[17] des femmes est de 169 cm et celle des hommes est de 177cm, vous simplifiez vos données de telle manière à rendre l'information plus claire et plus parlante. Cette transformation est un exemple de **transnumérisation**, c'est-à-dire de manipulation des données chiffrées pour leur donner un sens facilement compréhensible.

[17] Je considère ici que vous avez déjà une notion, au moins usuelle, du concept de moyenne. Dans l'exemple original du chapitre 1 je me contente de dire que les tailles "*tournent autour d'une valeur*". Nous aborderons plus précisément ce concept ultérieurement.

Règles d'Or

Lorsque vous rédigez un travail, un mémoire, un rapport professionnel, un article scientifique, ou un quelconque document, n'oubliez pas que des gens sont supposés le lire (sinon, ne l'écrivez pas). Ces gens sont humains comme vous et moi (disons plutôt comme vous), et n'aiment probablement pas plus que vous les statistiques. Ils savent qu'il est indispensable de mesurer le risque de se tromper dans leurs prédictions, comprennent l'importance de l'outil, mais ne l'apprécient pas pour autant. Dès lors, il est important que vous réfléchissiez au message que vous voulez transmettre et à la meilleure méthode pour y parvenir. Simplifiez-les donc un maximum, en prenant soin d'y incorporer l'information importante de la manière la plus visuelle et la plus claire possible.

5.2. Problématique : Population et échantillon

Il est très important de réaliser ce que vous faites lorsque vous mesurez une variable (donc lorsque vous prélevez de l'information dans le milieu). Le Tableau 4.1 montre une série fictive de 45 étudiants qui ont passé leurs examens d'Analyse des Données théorique et pratique. Admettons que ces 45 étudiants appartiennent à une classe (toute aussi fictive) de 600 étudiants. Le plus souvent, les cotes de ces 45 étudiants ne m'importent absolument pas en tant que telles. Ce qui m'intéresse, c'est, par exemple, d'avoir une idée du taux de réussite des 600 étudiants sans devoir recueillir l'information pour l'entièreté de ce groupe (imaginons que cette information soit difficilement accessible). Dès lors, en tant que chercheur, je vais définir ma **population**, l'ensemble des 600 étudiants, et prélever aléatoirement une partie de cette population, ce qu'on appelle l'**échantillon** (les 45 étudiants). A partir de cet échantillon je vais ensuite tenter d'inférer les caractéristiques de ma population et espérer faire la plus petite erreur possible. Par convention, j'appellerai ***N*** la taille de ma population (souvent inconnue) et ***n*** la taille de mon échantillon. La logique devient donc : (a) je m'intéresse à un problème concernant une population énorme ; (b) je prélève un échantillon gérable auprès duquel ; (c) je vais mesurer (comme je peux, en utilisant une échelle de mesure adéquate) les informations (variables) qui m'intéressent ; (d) je vais ensuite inférer les caractéristiques de l'entièreté de la population à propos de cette variable sur base de cet échantillon.

Si je ne m'intéresse qu'aux caractéristiques de mon échantillon, sans me soucier de la population dont il est issu et sans vouloir inférer les caractéristiques de cette population, je parle de **statistiques descriptives**. En revanche, si j'ai la vocation de déterminer certaines caractéristiques de ma population à partir d'un échantillon comme je viens d'en discuter – points (a) à (c) – je parle alors d'**inférence statistique**.

Remarquez que la définition de ma population est essentiellement un choix de chercheur. Imaginons par exemple que j'aie de bonnes raisons de penser que la distribution des résultats de l'examen théorique des étudiants ne change jamais, quelle que soit l'année. Dans ce cas, je pourrais considérer que ma population est l'ensemble des étudiants de BA1 de Sciences Psychologiques et de l'Education. Je pourrais également envisager que n'importe quels étudiants réussissent de manière semblable les examens de statistique de base, quelle que soit leur nationalité ou leur spécialité. Je considérerais alors que ma population est l'entièreté des étudiants de BA1 dans le monde, indépendamment de leur faculté. Lorsque je vous dis cela, vous devez certainement être en train de penser que je commets une lourde erreur en croyant à cette hypothèse. Mais soyez conscients qu'environ 40% des études en Psychologie se font sur des étudiants de BA1 (selon Schultz, 1969), que 40 autres pour-cent se font sur des étudiants d'autres facultés et que le reste est réparti entre des enfants ou des adultes (le plus souvent, pour moitié, avec des caractéristiques spéciales que l'on désire étudier). Les conclusions que l'on tire sont souvent sensées s'appliquer à l'entièreté de l'humanité. Rien n'est donc moins sûr. Mais rassurez-vous, elles s'appliquent quand même à vous. Cette réflexion devrait vous renvoyer à la problématique de la logique inductive, de la répartition aléatoire des sujets et de la représentativité.

Remarquez également une dualité intrinsèque à l'échantillonnage : il y a deux critères opposés qui le conditionnent. D'une part, le prélèvement d'un sujet doit être indépendant du prélèvement d'un autre sujet. D'autre part, plus mon échantillon est grand, plus je serai précis dans mes estimations. Or pour que l'indépendance soit respectée le mieux possible il est nécessaire que l'échantillon représente un prélèvement d'une toute petite partie de la population au point que prélever un sujet de la population n'influence que de manière totalement négligeable les chances de prélever un autre sujet. En effet, rappelez-vous : si je prélève un étudiant au hasard dans une population de 600 individus, j'ai une chance sur 600 de prélever celui-ci. Si j'en prélève un second, je n'ai plus qu'une chance sur 599 que ce soit le suivant et les prélèvements ne sont en fait pas indépendants. Si je prélève 40 humains

parmi les 7 10^9 qui existent, cette violation de l'indépendance est totalement négligeable. Par contre, si je prélève 200 étudiants parmi 600, la dépendance devient sensible.

Mais d'autre part, vous pouvez sans peine admettre que, si je prélève 599 étudiants sur mes 600, les erreurs que je pourrais faire concernant l'inférence des caractéristiques de ma population de 600 sont sans doute très faibles. A l'inverse, si j'utilise 5 sujets sur 600 pour acquérir une information pertinente sur les 600 étudiants, j'ai de forte chance d'être éloigné de la bonne information, beaucoup plus que si je prenais 599 étudiants. Donc, d'un côté l'échantillon doit être suffisamment petit pour que la violation de l'indépendance ne soit pas problématique, mais suffisamment grand pour permettre d'inférer le plus précisément possible les paramètres de la population. En pratique, on essaie souvent d'avoir le plus possible de sujets, et les populations sont le plus souvent suffisamment grandes pour que l'indépendance ne soit pas réellement menacée.

5.3. Présentation des données sous forme de distribution de fréquences

Lorsque vous récoltez de l'information à propos d'une variable, je vous conseille, avant toute chose, de représenter vos données graphiquement. Il existe plusieurs moyens d'y parvenir, certains sont plus utilisés que d'autres. L'avantage principal est d'avoir un accès visuel facile et immédiat sur vos données, ce qui vous permet de repérer facilement les tendances principales et les éventuelles anomalies (par exemple les erreurs de codage).

Une représentation graphique doit être sobre, simple, claire et représentative de vos données. Oubliez tout ce qui vous écarte de ces objectifs. Nous allons voir deux manières de représenter graphiquement une variable (plus tard, nous envisagerons la représentation de deux variables ou plus) : l'histogramme (et le diagramme en bâtons) et le diagramme en tiges et feuilles (Howell, 1999). Je déconseille toute autre alternative (dans le cadre de la Psychologie) à moins d'être bien renseigné sur ce que vous faites. Evitez par exemple les représentations en trois dimensions, qui perdent nécessairement en précision (sauf si vous voulez épater l'audience par l'esthétique plutôt que par le contenu). Evitez également les couleurs inhabituelles, les textures de simili pierre, ou bois, ou autres que vous proposent la plupart des logiciels habituels.

5.3.2. Données brutes

La première distinction utile à réaliser est celle qui sépare les variables discontinues des variables continues. Les **valeurs discontinues** sont des variables qui ne peuvent prendre que certaines valeurs et pas d'autres. On les appelle également **variables discrètes**. Par exemple, le nombre d'enfants est une variable discontinue dans la mesure où il est impossible d'avoir des valeurs qui n'appartiennent pas aux entiers positifs ou nuls (il est difficile d'avoir 2,6 enfants). Une **variable continue**, en revanche, peut, théoriquement au moins, prendre n'importe quelle valeur sur un intervalle donné. Par exemple, la taille d'un être humain adulte peut valoir un nombre infini de valeurs comprises entre 51 cm (le Népalais Khagendra Thapa Magar) et 272 cm (Robert Wadlow, mort en 1940, jamais surpassé à ce jour) même si la précision de nos instruments de mesure nous imposent une limite finie.

Comme premier exemple, prenons les points d'ANAD (théorie et pratique) de 45 étudiants fictifs (Tableau 5.1). Remarquez, encore une fois, que cette échelle est évidemment particulière : on pourrait la considérer comme une variable continue, si les enseignants étaient réellement capables de mesurer la "connaissance" d'un cours sur une telle échelle. En pratique, non seulement il s'agit d'une échelle ordinale, mais en plus, son niveau de précision est très mauvais. Il suffit pour vous en convaincre de demander à 10 personnes de corriger la même épreuve d'une évaluation pour vous rendre compte de la disparité des scores attribués. Donc, nous considérerons cette variable comme discontinue et ne pouvant prendre que les valeurs entières.

Tableau 5.1. : Séries statistiques de 45 étudiants de BA1.

Num	Théorie	Prat.	Num	Théorie	Prat.	Num	Théorie	Prat.
1	18	15	**16**	12	16	**31**	15	12
2	13	15	**17**	17	12	**32**	12	15
3	17	14	**18**	13	12	**33**	12	12
4	16	14	**19**	14	15	**34**	17	14
5	17	15	**20**	14	15	**35**	14	10
6	15	13	**21**	17	14	**36**	13	11

Tableau 5.1. : Séries statistiques de 45 étudiants de BA1.

Num	Théorie	Prat.	Num	Théorie	Prat.	Num	Théorie	Prat.
7	16	14	**22**	13	14	**37**	11	13
8	14	15	**23**	14	9	**38**	13	14
9	14	14	**24**	13	11	**39**	14	12
10	12	15	**25**	15	13	**40**	13	12
11	16	12	**26**	18	14	**41**	12	14
12	12	18	**27**	11	10	**42**	13	12
13	16	14	**28**	15	15	**43**	12	10
14	14	14	**29**	16	13	**44**	11	14
15	12	14	**30**	12	14	**45**	14	15

5.3.3. Transnumérisation en tableau de fréquences

La première étape de simplification du Tableau 5.1 est d'établir un tableau de **distribution de fréquences**. Le Tableau 5.2 correspond à cette étape. Comme vous pouvez l'observer, les cotes possibles ont été envisagées et le nombre d'étudiants l'ayant obtenue constitue ce que l'on appelle la **fréquence absolue**.

La colonne suivante contient la même information mais par rapport à l'ensemble des sujets, c'est la **fréquence relative**. Par exemple, 3 sujets sur 45 ont obtenu un 11/20, 3/45 = 0,067. Remarquez que la fréquence relative correspond à la probabilité de trouver une note donnée dans l'échantillon. Par exemple, on peut dire qu'il y a une probabilité de 0,20 (ou 20%, ou 1/5) d'avoir un 12, donc que, si je prélève au hasard un étudiant parmi mes 45, j'ai une chance sur cinq d'en sélectionner un qui a obtenu un score de 12/20 en théorie. De même, si l'on se trouve à la troisième colonne, la somme des fréquences relatives représente la probabilité d'avoir 13 ou moins. Donc : 0,067+0,20+0,18 = 0,447, il y a 44,7% de l'échantillon qui a 13 ou moins à sa cote d'Analyse de Donnée.

La colonne suivante contient la somme des fréquences absolues, ce qu'on appelle les **fréquences cumulées**. Cette information peut s'exprimer en pourcentage pour plus de facilité, c'est-à-dire en le divisant par l'effectif total de l'échantillon (le nombre de sujets,

donc 45) et en le multipliant par 100. C'est ce qu'on retrouve à la dernière colonne du tableau. Par exemple, 8 étudiants ont obtenus un 13/20, la fréquence cumulée est de 20, ce qui en pourcentage donne : 20/45*100 = 44,4%.

Enfin, remarquez que j'ai commencé mon tableau à la cote de 11 et que je l'ai terminé à la cote de 18. C'est un choix. J'aurais pu commencer à 0 et finir à 20 et indiquer des fréquences de 0 aux scores qu'aucun étudiant n'a obtenu. Cela m'aurait permis de représenter l'entièreté des scores possibles. Cependant, si je n'avais, par exemple, eu aucun étudiant qui obtient 13, qui est en milieu de série, j'aurais néanmoins été obligé d'indiquer cette cote pour respecter la taille des intervalles.

Tableau 5.2. : Transnumérisation du Tableau 5.1.

Cotes	Fréquences absolues	Fréquences relatives	Fréquences cumulées	Fréquences cumulées en %
11	3	0,067	3	6,7
12	9	0,20	12	26,7
13	8	0,18	20	44,4
14	9	0,20	29	64,4
15	4	0,089	33	73,3
16	5	0,11	38	84,4
17	5	0,11	43	95,6
18	2	0,044	45	100

5.4. Représentations graphiques

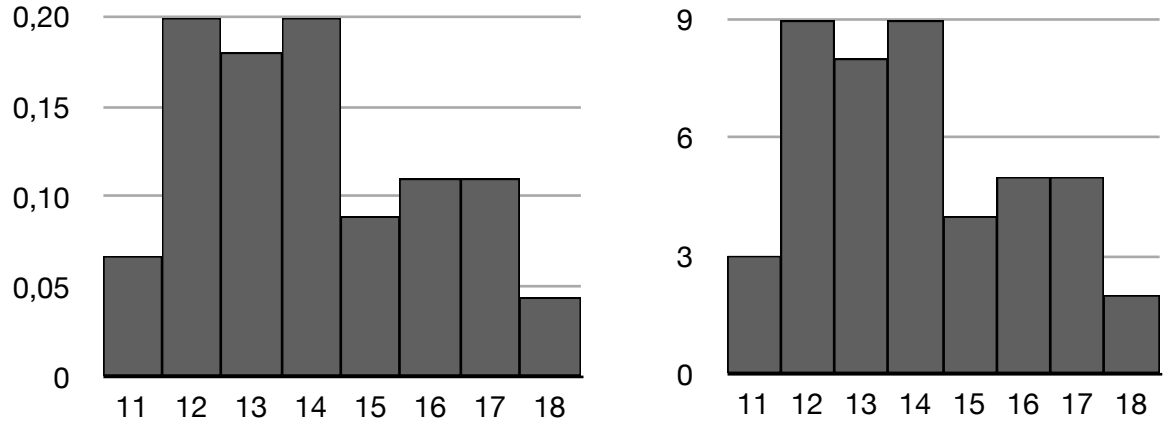
5.4.1. Représentation sous forme d'histogramme

Les Figures 5.1 montrent les graphes en barres de notre distribution. Que ce soient les fréquences absolues ou les fréquences relatives importe peu en termes de forme de graphe. En revanche, l'information donnée est un petit peu différente, dans le premier cas, on peut dénombrer le nombre d'étudiants qui ont reçu une note donnée, dans le second cas, on peut déterminer la proportion des étudiants qui ont reçu cette note. En fait, j'aurais tendance à

conseiller la première représentation si vous voulez discuter de votre échantillon en particulier (statistique descriptive) et la seconde si vous désirez faire des inférences par rapport à votre population.

Figure 5.1. : Diagramme en barre de la distribution du Tableau 5.2

(a) Distribution en fréquences relatives (b) Distribution en fréquences absolues



Malgré l'apparente simplicité de ces diagrammes, certaines caractéristiques valent la peine d'être développées :

1. Selon le type d'échelle, l'axe horizontal aura différentes conventions. Par exemple, pour une échelle nominale, cet axe ne sera pas normé (une unité n'aura pas obligatoirement une dimension constante) et n'aura pas de direction. Il servira juste de support visuel aux barres (dont la largeur importe peu, mais qu'on uniformise pour le confort des yeux) qui peuvent apparaître dans n'importe quel ordre sans altérer le contenu informatif de la présentation. Pour les échelles ordinales, cet axe a une direction et l'ordre des classes est important, mais la largeur des barres toujours pas (même si on les garde toujours uniformes). Pour les échelles d'ordre supérieur, l'axe a une direction et est normé. La largeur des barres apporte une information valide.
2. Dans certains cas, on utilise des classes. Par exemple, pour représenter l'âge, on pourrait travailler par 5 ans (de 15-20 ; 20-25 ; 25-30 ; etc.). Dans ce cas, le premier problème qui se pose est la limite : une personne de 20 ans, où se situe-t-elle? C'est en réalité un faux problème pour plusieurs raisons : d'une part, on prendre la décision soi-même sur des critères subjectifs. On peut, par exemple considérer qu'une personne qui dit avoir 20 ans a nécessairement plus que cet âge (ne fusse que les secondes qui s'égrènent durant l'expérimentation) et donc

va dans la classe 20-25. Mais on pourrait également décider au hasard de l'attribution. En fait, ce cas est très rare et ne perturbera vraisemblablement pas vos résultats quelle que soit la décision.

3. Toujours en cas d'utilisation de classe, une valeur de classe importante est son milieu (ou son centre). En effet, une manière efficace de se représenter la classe 20-25 ans est de considérer que tous les sujets qui la peuplent ont 22,5 ans. Ces valeurs vous permettront éventuellement de tracer une courbe qui passe par tous les centres des classes (ce qu'on appelle un polygone des fréquences).
4. Il reste à déterminer le nombre de classes et leur limites. Il n'y a pas vraiment de règles concernant leur nombre, mais Howell (1999) suggère qu'une dizaine de classes constitue un nombre facilement gérable, permettant de distinguer les informations utiles. Si vous en avez besoin de plus ou de moins, laissez votre bon sens vous guider. En revanche, concernant les limites de classe, il est utile de se rendre compte des enjeux. Par exemple, supposons que vous vouliez distinguer des caractéristiques des enfants, des adolescents, des adultes et des personnes âgées (et adultes). Vous feriez donc quatre classes : une de 0 à 12, une de 12-18, une de 18-60 et une au-delà de 60 ans. Se faisant, les deux premières classes ont une étendue de 12 et de 6 ans, alors que la troisième est de 42 ans et que la dernière est indéterminée. Bien que ce choix puisse être pertinent dans certains cas, il faut se rendre compte que vous n'êtes plus dans une échelle de rapport si vous faites cela, mais bien dans une échelle ordinale. Dès lors, si c'est pertinent par rapport à votre question de recherche, je vous conseille, tant que faire se peut, de garder des étendues identiques pour chacune de vos classes.

5.4.2. Représentation sous forme de tiges et feuilles

Cette représentation est assez ingénieuse. Elle représente l'entièreté des données tout en donnant la forme de la distribution, tel que le ferait un histogramme. Elle a été proposée par Tukey (1977). Le Tableau 5.3 présente ce diagramme à partir de données brutes fictives représentant, par exemple, le temps de trajet entre votre domicile et l'Université. Supposons que vous mettez entre 2 minutes et 60 minutes pour y parvenir. Je peux, comme dans la première colonne du tableau, noter tous les temps de trajet en les rassemblant par dizaine de minutes. Dans un diagramme en tiges et feuilles, les tiges (deuxième colonne) représenteront les dizaines (donc la classe) et les feuilles (troisième colonne) seront

représentées par l'unité en nombre correspondant aux nombres de sujets qui prennent cette valeur précise. De cette manière, la forme de la distribution apparaît horizontalement en considérant la ligne de séparation des tiges et des feuilles comme l'équivalent de l'axe horizontal de l'histogramme. Vous pouvez constater que la plupart des étudiants de cette série habitent à 10-20 minutes de l'Université, alors que seuls trois d'entre eux habitent à 50-60 minutes.

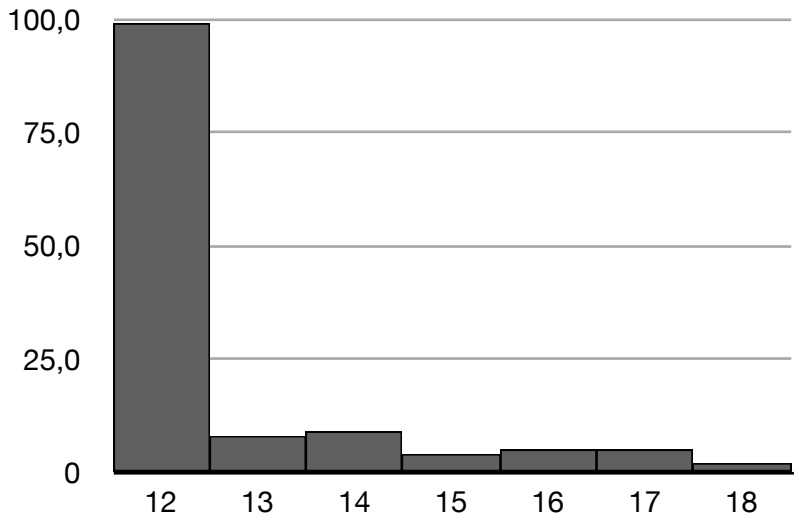
Tableau 5.3. : Diagramme en tiges et feuilles

Données brutes	Tiges	Feuilles
2,2,2,3,4,5,6,8,8,9,9, 10,10,10,11,11,12,12,13, 13,13,13,13,14,15,16,17, 17,17,18,18,19,19,19,20, 20,21,21,22,22,23,24,25, 25,26,26,26,27,27,28,29, 30,31,32,33,33,35,35,36, 37,38,39,39,40,41,42,45, 45,45,46,47,48,48,49,50, 50,59.	0	22234568899
	1	112233333456778000
	2	11223455666778900
	3	12335567899
	4	1255567889
	5	009

5.4.3. Les valeurs aberrantes

Un des avantages des représentations graphiques est la détection de valeurs dites "aberrantes", c'est-à-dire tout à fait anormales par rapport aux autres valeurs de votre série. Ces valeurs, comme nous le verrons plus tard, sont importantes à détecter parce qu'elles peuvent perturber sérieusement les analyses. Il y a plusieurs raisons pour lesquelles une valeur peut être anormale. La plus évidente (et probablement la plus fréquente) est l'erreur d'encodage. Si, dans le Tableau 5.3, j'avais par mégarde écrit 599 au lieu de 59 pour la dernière valeur, j'aurais eu une quantité de tiges énormes (60) dont 53 auraient été vides. Graphiquement, que ce soit avec un histogramme ou un diagramme en tiges et feuilles, cela m'aurait sauté aux yeux. La Figure 5.2 montre cette situation, la seule différence entre cet histogramme et l'histogramme de la Figure 5.1.a est la première valeur : au lieu de n'avoir que 9 personnes qui ont eu 12, j'ai appuyé deux fois sur le 9 et ai donc 99 personnes qui ont eu 12. Vous percevez immédiatement le changement drastique qui s'ensuit.

Figure 5.2. : Même histogramme que la figure 4.2. mais avec la première valeur aberrante suite à une erreur d'encodage.



D'autres causes peuvent expliquer des valeurs aberrantes. Par exemple, si je mesure des temps de réaction (admettons que les sujets doivent appuyer sur un bouton dès qu'ils perçoivent la couleur jaune), il se peut qu'un sujet soit distrait, et n'appuie sur le bouton que 4 secondes après avoir vu la couleur. Cette valeur ne représentant pas du tout l'aptitude que je cherche à mesurer, il est tout à fait légitime de l'ôter de l'analyse.

En revanche, il existe des valeurs exceptionnellement élevées (ou faibles) que je ne peux pas traiter comme une simple anomalie et dont je dois tenir compte. C'est le cas d'erreurs systématiques. Imaginons, pour reprendre l'exemple des temps de réaction, qu'une des photos, toujours la même, conduise, pour un cinquième des sujets, à un temps de réponse de 4 secondes, alors que toutes les autres donnent un temps d'environ 0,2 seconde. Dans ce cas, je ne peux pas considérer que ces sujets sont simplement distraits et ôter leur score. Je dois me dire que cette photo contient quelque chose de particulier qui génère, chez une partie importante des sujets, une perturbation. A ce moment, j'ai deux choix. Soit le problème théorique sous-jacent m'indiffère complètement et j'enlève simplement la photo (y compris pour les 80% des sujets qui ont un temps de réponse habituel). Soit je m'intéresse à ce résultat et je tente de lui trouver une explication plausible, par exemple en réalisant une nouvelle expérience adaptée au problème. Dans le cadre d'une recherche, vous aurez

l'honnêteté intellectuelle de rapporter ce résultat dans votre compte rendu et, si vous optez pour la première solution, vous justifierez votre choix.

Plus loin, nous traiterons en détails des méthodes pour détecter ces valeurs aberrantes (voir points 5.6 et 7.5.3). Mais pour l'heure, il est d'abord utile d'envisager les distributions possibles.

5.5. Les distributions

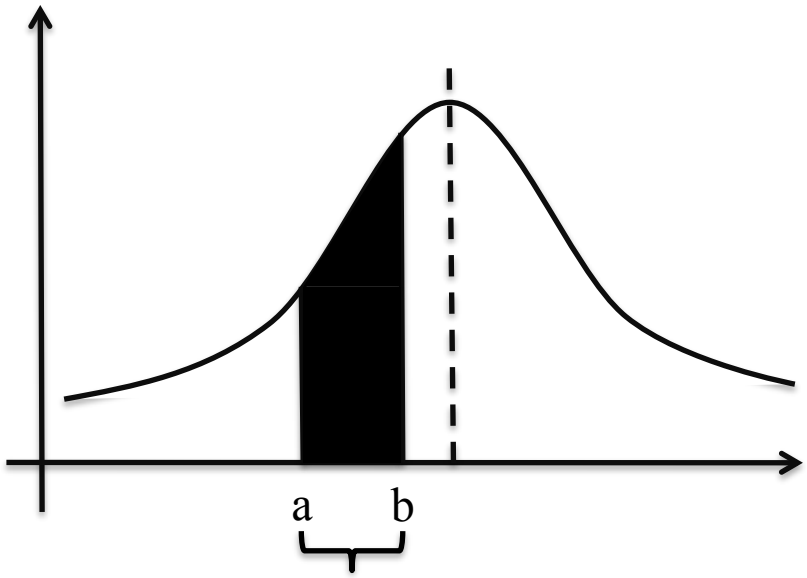
5.5.1. Caractéristiques d'une distribution

Une distribution statistique peut être vue comme une courbe représentant l'infinité des résultats possibles d'une expérience aléatoire avec leur probabilité respective. En effet, si vous reprenez la Figure 5.1 et que vous imaginez que les rectangles se multiplient (c'est-à-dire que l'on ait un nombre infini de classes) et que leur largeur devienne de plus en plus petite jusqu'à devenir ponctuelle (donc d'avoir la largeur d'un point, c'est-à-dire tendant vers zéro), vous finirez par obtenir une courbe qui peut se définir par une équation de type $y = f(x)$. En fait, je ne fais rien d'autre qu'envisager qu'une variable discrète devienne continue, pour vous sensibiliser à une vision sous forme de courbe. Pour certaines distributions, l'équation de la fonction vous sera communiquée, mais même dans ce cas, vous n'aurez pas à l'utiliser, il existe des abaques (sous forme de tables) qui vous fournissent les informations nécessaires sans vous obliger à calculer ces fonctions compliquées. Nous traiterons plus en détails de ces considérations au chapitre 7 qui s'attaque à deux distributions importantes : la binomiale et la normale.

Pour un intervalle donné, l'aire sous cette courbe représente la probabilité de cet intervalle (voir Figure 5.3). Si l'on calcule l'aire sous l'entièreté de la courbe, on obtient une probabilité de 1. Par exemple, si vous considérez la Figure 5.4.d, vous pouvez voir graphiquement que le milieu de la courbe représente exactement la moitié de l'aire (puisque la courbe est symétrique). Dès lors, vous pouvez être sûr(e)s que la probabilité qu'une observation (un sujet de votre échantillon) correspondant à cette courbe ait un score contenu entre $-\infty$ et le milieu de la courbe est de 0,5 (une chance sur deux, puisque l'autre chance sur deux correspond à être sur l'autre moitié de la courbe allant du milieu à $+\infty$). En fait, à chaque segment de l'axe horizontal (appelé l'axe "x") correspond une probabilité d'appartenir à ce segment et cette probabilité se calcule en mesurant l'aire sous la courbe sur ce segment,

comme l'illustre la Figure 5.3. D'un point de vue mathématique, il suffit d'intégrer la fonction de la courbe sur l'intervalle concerné, mais dans ce cours, nous n'aurons jamais à le faire (je lis déjà la déception sur vos visages), il vous suffit de comprendre le principe.

Figure 5.3. : Représentation de la probabilité comme l'aire sous la courbe

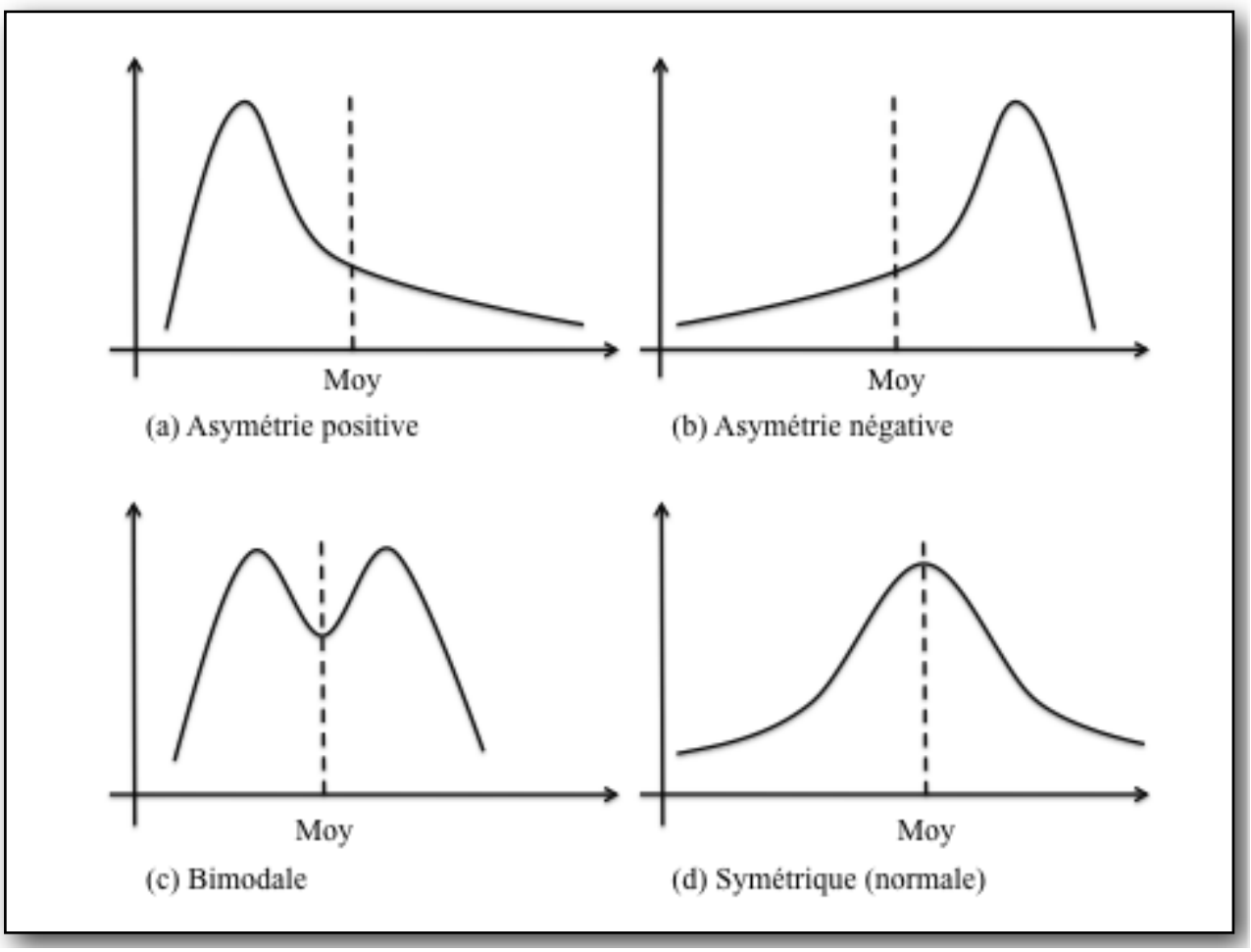


5.5.2. Formes des distributions

Les distributions statistiques peuvent avoir de nombreuses formes, comme le montre la Figure 5.4. Bien qu'il y ait une infinité de distributions possibles, on en distingue habituellement quatre grands types. La distribution peut être asymétrique et ce, de manière positive (si la queue de la distribution tend vers les valeurs positives de l'axe x) ou négative (si la queue de la distribution tend vers la gauche). Elle peut également être bimodale (ou multimodale), c'est-à-dire avoir deux (ou plus) modes (nous verrons ce que cela signifie plus bas). Ce type de distribution peut, par exemple, être obtenu en mesurant la taille des individus belges adultes. En effet, si l'on ne sépare pas les hommes des femmes, il y aura un premier pic à la taille la plus fréquente des femmes et un second à la taille la plus fréquente des hommes qui est une valeur supérieure à celle des femmes. Enfin, la forme peut être symétrique, comme peut l'être une distribution dite **normale**. Cette distribution particulière est une distribution qui occupe une position prépondérante dans le domaine

des statistiques et à laquelle nous allons nous attacher tout particulièrement dans les chapitres qui viennent. J'attire d'ores et déjà votre attention sur le fait qu'il ne suffit pas qu'elle soit symétrique pour pouvoir être qualifiée de normale. Mais avant d'en arriver à cette distribution, continuons à nous interroger sur la manière dont nous pouvons décrire nos données quelle que soit la distribution. Le point suivant termine l'exploration graphique. Ensuite, nous passerons à quelques caractéristiques algébriques, pour revenir enfin sur les distributions particulières.

Figure 5.4. : Formes des distributions



5.6. Les quantiles et les boîtes à moustaches

5.6.1. Définition des quantiles et quantiles particuliers

Une manière utile de décrire une distribution est de la diviser en un certain nombre d'intervalles ayant comme caractéristique de contenir des proportions identiques d'observations. Ces intervalles sont les **quantiles** (ou fractiles). On cherche alors à

déterminer les valeurs observées ou les valeurs proches des valeurs observées de la variable qui correspondent aux limites de ces intervalles.

Certains quantiles sont plus utilisés que d'autres et portent des noms particuliers. On distinguera :

- La **médiane (θ)**, qui sépare la distribution en deux portions contenant chacune 50% des observations. Ce quantile est extrêmement important car il donne une mesure de ce que l'on nomme la **tendance centrale**. Elle correspond souvent (lorsque la distribution est unimodale, c'est-à-dire n'a qu'un seul pic) à déterminer l'endroit de la distribution qui regroupe un maximum de sujets.
- Les **quartiles** qui séparent la distribution en quatre portions qui contiennent chacune 25% des observations. On spécifie leur ordre en parlant de *premier quartile* (noté Q_1), *deuxième quartile* (noté Q_2), *troisième quartile* (noté Q_3) et de *quatrième quartile* (noté Q_4).
- Les **déciles** qui séparent la distribution en 10 parties contenant chacune 10% des observations.
- Les **percentiles** qui séparent la distribution en 100 parties contenant chacune 1% des observations. Remarquez que le 25ème percentile est le premier quartile ; le 50ème percentile est le 5ème décile ou le 2ème quartile ou encore la médiane ; etc.

5.6.2. Distinction entre les notions de percentiles et les rangs percentiles

- Un percentile est un score, donc une valeur de la variable envisagée.
- Le rang percentile est le rang exprimé en pourcent occupé par un sujet possédant ce score. C'est donc le rang médian qui représente la mesure de la tendance centrale.

Dans le Tableau 5.2, la ligne grisée montre que la cote 15 correspond à une fréquence cumulée (la somme des fréquences relatives) de 73,3%. On prend l'habitude d'arrondir le percentile ce qui nous donne le 73ème[18]. On dira d'un étudiant qui a la cote de 15 qu'il se situe au rang percentile 73 pour ce cours. Remarquez qu'exprimer de la sorte, cela

[18] Remarque sur les arrondis : deux stratégies sont possibles. Soit on considère qu'une fois un percentile entamé on se retrouve déjà dans le suivant, auquel cas 73,3 deviendrait déjà le 74ème percentile. Soit on arrondit de manière traditionnelle, auquel cas toute valeur strictement inférieure à 73,5 deviendrait 73 et toute valeur supérieure ou égale à 73,5 (et strictement inférieure à 74,5) s'arrondi à 74.

correspond à établir la position de l'étudiant en imaginant que le groupe contient 100 classes (même si en réalité il en contient moins).

5.6.3. Distinction entre séries statistiques ordonnées et distributions de fréquences

Lorsque l'on détermine les quantiles, on peut utiliser une présentation des données sous deux formes : les distributions de fréquences (comme nous l'avons fait jusqu'à présent) ou la série statistique ordonnée. Selon ce choix, la manière de traiter les quantiles est un petit peu différente. Je vais envisager ces différences sur base d'exemples de petites séries de quelques sujets dont la cote à l'examen d'ANAD a à nouveau été inventée.

5.6.3.1 Cas d'une série statistique ordonnée

Envisageons la médiane par cinq exemples : dans le premier nous aurons un nombre impair de sujets ; dans le deuxième un nombre pair ; dans le troisième des valeurs répétées et un nombre pair de sujets ; dans le quatrième des valeurs répétées et un nombre impair de sujets ; enfin, dans le cinquième la même situation que dans l'exemple trois mais avec une série qui conduit à une situation un peu particulière. Une fois le principe compris pour la médiane, vous n'aurez aucune difficulté à reproduire le processus pour n'importe quel quantile (la médiane n'étant rien d'autre que le 50ème percentile, c'est-à-dire un parmi d'autres).

Pour déterminer la valeur de la médiane, il faut ranger les données par ordre numérique croissant de la variable et déterminer la position médiane (ou rang médian) par la formule suivante : **position médiane (ou rang médian) = (n + 1)/2**.

Supposons la série statistique de cinq sujets (n = 5) : 5, 8, 3, 7, 15 (exemple 1)
On peut l'ordonner de manière croissante et obtenir : 3, 5, 7, 8, 15.

Dans ce cas, la médiane occupe le rang (5+1)/2 = 3 en comptant à partir de la gauche, ce qui correspond à la valeur 7. Il y a un nombre égal (2) de valeurs de la variable de part et d'autre de la médiane de 7, mais ces deux portions égales ne représentent pas exactement 50% des données puisque la valeur de 7 correspondant à la médiane est exclue.

Supposons maintenant la série statistique de n = 6 : 5, 11, 3, 6, 15, 14 (exemple 2)

L'ordre numérique croissant est : 3, 5, 6, 11, 14, 15.

La médiane occupe la position (6 +1)/2 = 3,5, ce qui correspond à une valeur comprise entre 6 et 11. Par convention, on prend la moyenne des deux valeurs, ce qui donne 8,5 pour la médiane. **La médiane ne correspond donc pas nécessairement à une valeur observée**. Dans ce cas, il y a non seulement le même nombre de valeurs de la variable de part et d'autre de la médiane, mais ces deux portions représentent chacune exactement 50% des données (puisque la médiane ne correspond pas à une valeur observée ; donc, aucune valeur n'est exclue par le partage en deux).

Lorsque des valeurs sont répétées dans les données, elles doivent évidemment être répétées dans la série de données rangée par ordre croissant.

Supposons la série statistique de *n* = 10 : 5, 8, 14, 3, 15, 14, 8, 5, 14, 14 **(exemple 3)**

L'ordre numérique croissant est : 3, 5, 5, 8, 8, 14, 14, 14, 14, 15.

La médiane occupe le rang (10 + 1)/2 = 5,5. En comptant 5 et 6 positions depuis la gauche on arrive aux valeurs 8 et 14, ce qui donne 11 [= (8 + 14)/2] pour la médiane (donc pas une valeur observée parce que *n* est pair).

Si on supprimait un des "*8*" dans cette série, on se retrouverait avec les neuf valeurs suivantes déjà rangées par ordre croissant (*n* = 9) :

3, 5, 5, 8, 14, 14, 14, 14, 15 **(exemple 4)**

La médiane serait de 14 [rang 5 = (9 +1)/2].

Si on supprimait le 2ème 8, on se retrouverait avec la série ordonnée (*n* = 8) :

3, 5, 5, 14, 14, 14, 14, 15 **(exemple 5)**

La médiane occuperait le rang (8 + 1)/2 = 4,5 ; elle aurait donc la valeur 14 parce que la moyenne de la valeur de 14 occupant le rang 4 et de la valeur 14 occupant le rang 5 est aussi de 14.

En résumé, avec un nombre impair d'observations, la valeur de la médiane correspond nécessairement à une valeur observée de la variable (exemples 1 et 4) ; avec un nombre pair d'observations, la médiane ne correspond pas nécessairement à une valeur observée de la variable. « Pas nécessairement » veut dire que si les valeurs de la variable de part et d'autre du rang médian [défini comme étant le rang $(n+1)/2$] sont différentes, la médiane prend une valeur qui n'est pas observée (exemples 2 et 3). En revanche, si les valeurs de part et d'autre du rang médian sont égales, la médiane prend cette valeur observée (exemple 5).

5.6.3.2 Cas d'une série présentée sous forme de fréquences relatives cumulées

a) Variable discontinue (= discrète)

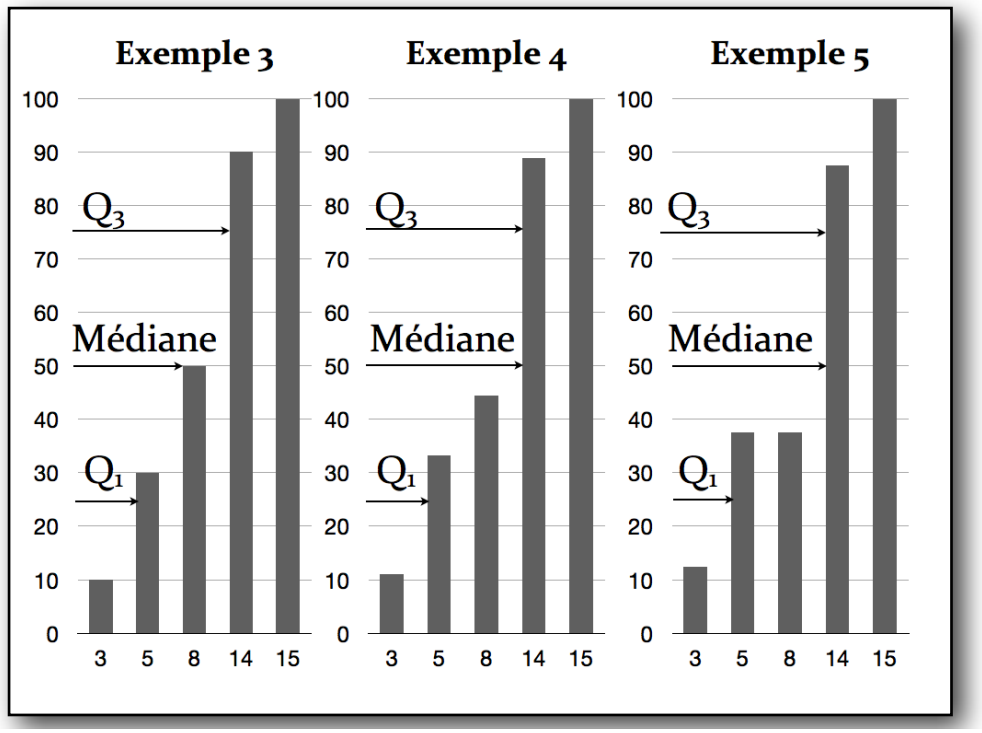
Idéalement, la médiane devrait correspondre à la valeur de la variable dont la fréquence relative cumulée est d'exactement 50%. De la même manière, les quartiles 1 et 3 devraient correspondre aux valeurs de la variable dont les fréquences relatives cumulées sont exactement de 25 et 75%. Cependant, cet idéal ne peut généralement pas être atteint quand on détermine les valeurs des quartiles à partir de la distribution des fréquences relatives cumulées d'une variable discontinue ou considérée comme telle.

Le Tableau 5.4 reprend les exemples de 3 à 5 présentés sous forme de tableau de fréquences. La Figure 5.5 montre les diagrammes en escalier des fréquences relatives cumulées (exprimées en %) pour ces mêmes données. Les flèches horizontales correspondent aux fréquences cumulées de 25, 50 et 75%.

Tableau 5.4. : Tableaux de fréquences des exemples 3 - 4 - 5

	Exemple 3		Exemple 4		Exemple 5	
Cotes	Fréq. abs	Fréq. rel. cumul.	Fréq. abs	Fréq. rel. cumul.	Fréq. abs	Fréq. rel. cumul.
3	1	10,00%	1	11,11%	1	12,50%
5	2	30,00%	2	33,33%	2	37,50%
8	2	50,00%	1	44,44%	0	37,50%
14	4	90,00%	4	88,89%	4	87,50%
15	1	100,00%	1	100,00%	1	100,00%

Figure 5.5. : Diagramme des fréquences cumulées pour les exemples 3 - 4 - 5 et leurs quartiles (flèches noires)



Deux cas doivent être envisagés :

- La flèche bute sur une contremarche. La valeur de l'abscisse à la verticale de cette contremarche est la valeur du quartile recherché (Figure 5.5. exemples 4 et 5).
- La flèche est exactement au niveau d'une marche. Par convention, on prend la moyenne des valeurs correspondant aux deux limites de cette marche comme la valeur du quartile recherché (Figure 5.5. exemple 3, médiane).

Les résultats obtenus à la Figure 5.5. en se basant sur les fréquences cumulées sont les mêmes que ceux obtenus à la section précédente en se fondant sur les séries statistiques ordonnées. Les voici :

- Exemple 3 : Q_1= 5, Q_2 = 11 [= (8 + 14) / 2] et Q_3 = 14
- Exemple 4 : Q_1 = 5, Q_2 = 14 et Q_3 = 14
- Exemple 5 : Q_1 = 5, Q_2 = 14 et Q_3 = 14

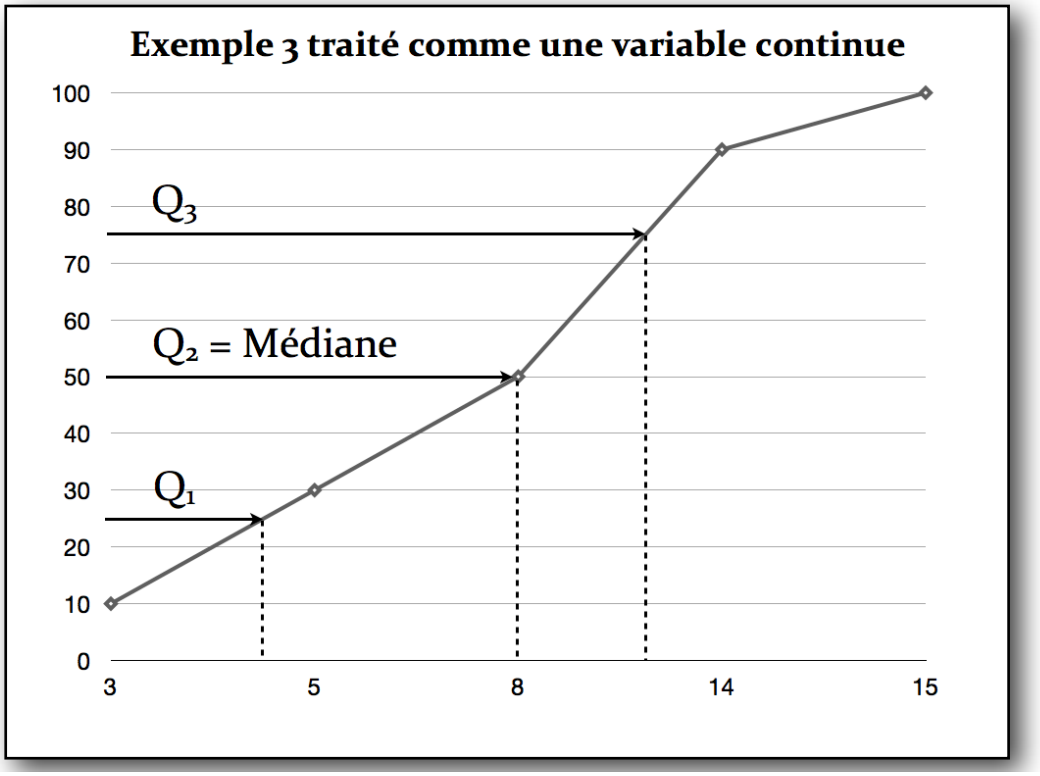
Notez qu'il est rare que les quartiles ainsi définis correspondent exactement aux 25, 50 et 75% de la distribution. En effet, au Tableau 5.4, on voit que les fréquences cumulées pour les

cotes de 8, 11 et 13 correspondent respectivement à 28,75%, 53,75 % et 76,875% des observations.

b) Variable continue

Admettons maintenant que je considère les valeurs de l'exemple 3 comme des valeurs continues. Cela signifie que toutes les valeurs de cette variable sont considérées comme possibles (je considère donc qu'il est théoriquement possible d'observer une cote de, par exemple, 5,6879879). La conséquence est que je n'ai plus de classe, mais que chaque point infinitésimal est représenté sur la fonction. Je n'ai donc plus de diagramme en barre mais bien une fonction qui relie les points observés du diagramme de fréquences cumulées par un segment de droite, c'est ce qu'on appelle un **polygone de fréquences cumulées** (comme le montre la Figure 5.6 qui n'est donc rien d'autre que la représentation des données du Tableau 5.4. exemple 3 et qui est la même que la partie gauche de la Figure 5.5 mais représentée sous forme de variable continue).

Figure 5.6. : Polygone de fréquences cumulées correspondant aux données de l'exemple 3



Dans ce cas, nous pouvons trouver les valeurs exactes des quartiles par **interpolation linéaire**. Il y a plusieurs manières de réaliser une telle interpolation, voici deux méthodes :

- **<u>Méthode 1</u>** : Qui se rappelle de ses cours de math sait qu'avec deux points a, de coordonnées (x_a, y_a), et b, de coordonnées (x_b, y_b), on peut construire une droite d'équation $\mathbf{(Y - y_a) = (y_b-y_a)/(x_b-x_a)*(X-x_a)}$. Les autres le savent maintenant. Le premier quartile se trouve entre le point (3, 10) et le point (5, 30). A partir de ces deux points nous obtenons la droite (Y - 10) = (30 - 10)/(5 - 3) * (X - 3) qui se simplifie en Y = 10X - 20. Je cherche le premier quartile, qui correspond à un Y = 25. En remplaçant cette valeur dans l'équation de la droite, nous obtenons 25 = 10X - 20, donc X = 4,5. Le premier quartile se situe donc à une cote de 4,5. Le deuxième quartile est immédiat : X = 8. Le troisième quartile peut se calculer comme le premier. Je vous laisse faire l'exercice, vous devriez obtenir la cote de 11,75. Pour ceux qui s'en souviennent il est également possible de trouver l'équation de la droite par un système de deux équations de droites à deux inconnues. Il suffit de remplacer x et y par chacun des points et de trouver a et b.

- **<u>Méthode 2 :</u>** La fréquence de la cote 3 se situe à 10%. Celle de la cote 5 à 30%. Je cherche la cote correspondant à la fréquence de 25%. Il y a une distance de 20 unités entre 30% et 10%. La fréquence de 25% se situe à 15 unités des 10% (et à 5 unités de 30%) donc à 15/20*100 = 75% de la distance entre 10% et 30%. Reportons cette proportion sur la distance entre les coordonnées de l'axe x (les cotes). On voit que 3 et 5 sont séparés de 2 unités. Les 75% de deux unités correspondent à 2*0,75 = 1,50 point. La cote correspondant à une fréquence de 25% est donc de 3 + 1,5 = 4,5 comme on l'obtenait par la méthode 1. Le deuxième quartile est également immédiat, étant sur le point 8. Le troisième quartile devrait également vous conduire à la cote de 11,75 si vous faites l'exercice par vous-même (il ne s'agit de rien d'autre qu'une règle de trois).

5.6.4. Les boîtes à moustaches

Une information intéressante, lorsque l'on désire caractériser notre distribution, est l'**écart interquartile**. Il est défini par l'écart entre le troisième quartile et le premier quartile (Q_3 - Q_1). Si nous reprenons l'exemple 3 de la Figure 5.5, on obtient un écart interquartile de 14 - 5 = 9. L'intérêt de cet intervalle tient du fait qu'il représente les 50% de l'échantillon regroupé au centre de la distribution. C'est donc une mesure de ce que nous appellerons désormais la **dispersion** des données. En effet, si l'écart interquartile est grand, cela signifie que les sujets

ont obtenu des scores très étalés sur l'ensemble des cotes possibles. A l'inverse, si cet écart est très faible, cela signifie que la plupart des sujets ont obtenu des scores regroupés autour d'une valeur centrale.

Mesurer la dispersion est essentiel pour plusieurs raisons, sur lesquelles nous reviendrons lorsque nous aborderons la description des données par des méthodes algébriques. Une des raisons principales est la considération des valeurs aberrantes (ou anormales). En effet, imaginez que 95% des étudiants réalisent un 12/20 à l'examen d'ANAD et que les 5% restant réalisent soit un 11, soit un 13. Dans ce cas, un 10 ou un 14 sera assimilé à des valeurs surprenantes. En revanche, imaginez que 50% des étudiants obtiennent une cote comprise entre 5 et 15 ; 40 autres % ont une cote entre 1 et 4 ou 16 et 19 ; et les 10% restant ont soit un 0 soit un 20. Dans ce cas, aucune cote ne pourrait être considérée comme particulièrement surprenante. Remarquez le lien entre ce que j'explique ici et la notion de modèle et d'erreur que nous avons envisagée plus tôt :

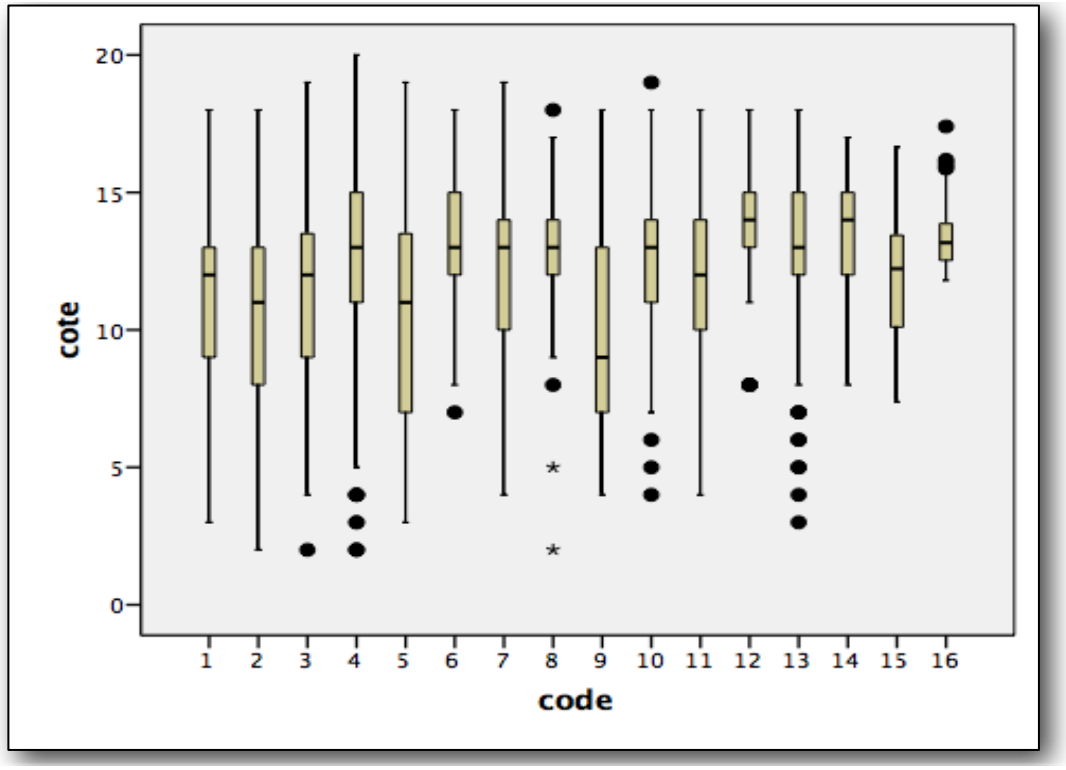
1. Dans le premier cas, lorsque la dispersion est très faible, si je prédis la cote de 12 à un étudiant pris au hasard, je n'ai que peu de chance de me tromper. De plus, si je me trompe, je ne fais qu'une erreur d'un point sur vingt. Dans ce cas, mon modèle prédictif est très bon et ma marge d'erreur est très faible.

2. Dans le second cas, lorsque la dispersion est très importante, prédire une cote de 12 à un étudiant ne me permet que rarement d'identifier correctement sa cote. L'erreur associée à ce modèle est nettement plus importante. Si je tombe sur un étudiant qui à 0 je fais une erreur de 12 points. La dispersion aura donc une importance capitale dans l'établissement de mon modèle prédictif.

Une manière particulièrement explicite de représenter graphiquement ce type de raisonnement (et la conclusion) est d'établir ce que l'on appelle des boîtes à moustaches. La Figure 5.7 donne une telle représentation. Elle a été créée sur base d'un fichier de points recueillis, il y a quelques années, par Monsieur Holender (mon éminent prédécesseur) alors que les Bacheliers s'appelaient encore Candidatures. Les points sont établis sur 20 et concernent les 14 cours de la 2ème candidature d'il y a quelques années. Les graduations 15 et 16 de l'axe horizontal correspondent à la moyenne totale (sur 20) de la 2ème et de la 1ère candidature pour ces étudiants. Seuls les étudiants qui ont au moins une note de 7,4/20 au total pour la 2ème candidature sont pris en compte. Tous ces étudiants ont donc réussi leur

première candidature. En conséquence, alors que les cotes pour les 14 cours ont une étendue possible de 2 à 20, l'étendue possible de la note moyenne en 2ème candidature est de 7,4 à 20 et celle de 1ère candidature, de 12 à 20.

Figure 5.7. : Cotes pour les 14 examens d'une 2ème candidature en psychologie d'il y a quelques années (étendue possible de 2 à 20) ainsi que les résultats cette 2ème candidature (étendue possible de 7,4 à 20) et pour la 1ère candidature (étendue possible de 12 à 20). Basé sur 160 étudiants ayant plus de 7,4/20 pour la moyenne de la 2ème candidature.



Vous remarquerez plusieurs parties sur les boîtes de cette figure. D'une part, la boîte proprement dite ou boîte centrale. D'autre part, les lignes qui en sortent, et que l'on nomme les moustaches. Et enfin, les points qui sont hors des moustaches et qui représentent les valeurs extrêmes. Nous allons passer chacun de ces éléments en revue.

5.6.4.1. La boîte centrale

La boîte centrale s'étend de Q1 à Q3 (donc l'écart interquartile). Elle correspond donc aux 50% des données centrales de la distribution. La barre à l'intérieur de la boîte représente la

médiane. La position de la médiane à l'intérieur de la boîte indique le degré de symétrie ou d'asymétrie de la portion centrale de la distribution.

Si l'on compare, par exemple, les cours 1, 4 et 6 de la Figure 5.7, on constate que les médianes indiquent que le cours 1 est moins bien réussi que les cours 4 et 6. La position de la médiane pour le cours 1 indique une asymétrie négative dans la portion centrale des données puisque la distance entre Q2 et Q1 est plus grande que la distance entre Q3 et Q2. La portion centrale de la distribution pour le cours 4 est plutôt symétrique puisque la médiane est à peu près au milieu de la boîte. Enfin, la portion centrale de la distribution du cours 6 présente une asymétrie positive puisque la distance entre Q2 – Q1 est plus petite que la distance entre Q3 et Q2. Les écarts interquartiles sont à peu près égaux pour les cours 1 et 4 et plus grands que pour le cours 6.

5.6.4.2. Les moustaches

Ce sont les lignes continues qui s'étendent de part et d'autre de Q1 et Q3. Leurs limites dépendent des **barrières** qui sont situées à une distance de 1,5 fois la taille de la boîte de part et d'autre de la boîte. Or, la taille de la boîte est représentée par l'écart interquartile. Par exemple, pour le cours 2, Q1 = 8 et Q3 = 13, donc l'écart interquartile est de 13 – 8 = 5 points. En multipliant 5 par 1,5 on trouve 7,5 points. La barrière inférieure vaut Q1 – 7,5, donc 8 – 7,5 = 0,5. La barrière supérieure vaut Q3 + 7,5, donc 13 + 7,5 = 20,5.

Cependant, une fois les barrières déterminées, il se peut qu'elles ne correspondent à aucune valeur existante de ma distribution. Nous allons donc observer les valeurs adjacentes à l'intérieur des barrières. Les cotes du cours 2 varient entre 2 et 18 (comme on pourrait le voir sur les données brutes que je ne vous ai pas fournies). Donc, la valeur adjacente inférieure est de 2. C'est la valeur la plus proche de la barrière inférieure de 0,5. La valeur adjacente supérieure est de 18. C'est à ces valeurs adjacentes que correspondent les extrémités visibles des moustaches.

Lorsqu'une valeur de la variable est confondue avec une barrière, la valeur adjacente est confondue avec cette barrière. Par exemple, si la barrière supérieure avait été de 18 au lieu de 20,5, la valeur 18 aurait été considérée comme valeur adjacente supérieure et la limite de la moustache aurait été confondue avec la barrière supérieure. De plus, il peut arriver qu'une valeur adjacente soit très loin d'une barrière. Supposons que la cote la plus basse ait été de 7

au lieu de 2, alors cette cote de 7 aurait été considérée comme la valeur adjacente inférieure, même si elle est assez éloignée de la barrière inférieure de 0,5. Dès lors, les barrières inférieures et supérieures sont toutes les deux à une distance de 1,5 fois l'écart interquartile par rapport à Q1 et Q3, cela n'empêche pas que les deux moustaches puissent avoir des longueurs très différentes l'une de l'autre (puisque les barrières n'apparaissent en général pas sur le diagramme ; ce sont les valeurs adjacentes qui apparaissent).

5.6.4.3. Les valeurs extrêmes

A la Figure 5.7, les points et les étoiles, de part et d'autre des moustaches, sont des valeurs extrêmes. Les valeurs extrêmes supérieures sont > à la barrière supérieure et les valeurs extrêmes inférieures sont < à la barrière inférieure. En outre, on différencie entre deux types de valeurs extrêmes, celles qui sont plus éloignées que 1,5 fois la largeur de la boîte et jusqu'à 3 fois la largeur de la boîte (représentées par des points) et celles qui sont plus éloignées que 3 fois la largeur de la boîte (représentées par des astérisques).

Il arrive qu'un point (ou une étoile) représente plus d'une donnée de même valeur. Après vérification, il se fait qu'un point ou une étoile représente souvent une seule donnée et au maximum deux données. Remarquons que l'impact des valeurs extrêmes se pose d'autant plus que les effectifs des échantillons sont plus faibles. La pertinence de leur élimination se pose aussi. Supposons qu'on ait disposé d'échantillons de 20 étudiants au lieu de 160. L'élimination d'une, deux ou trois valeurs extrêmes reviendrait à éliminer 5%, 10% ou 15% des données. Est-ce acceptable? Peut-on être sûr qu'avec des échantillons plus grands, ces mêmes données seraient encore considérées comme extrêmes?

5.7. Exercices de fin de chapitre

T.P. 4 : CHAPITRE 5

A. Problématique : Population et échantillon

Exercice 1 : Population et échantillon

Lorsque nous faisons des statistiques, notamment en psychologie, nous sommes amenés à vérifier des hypothèses à propos d'une population sur base d'un échantillon dit représentatif.

1. Des chercheurs s'intéressent à la notion de dépression chez les adolescents résidents en Belgique. Ces chercheurs font leur étude sur base d'un échantillon de 200 adolescents de 15 ans qu'ils ont choisi au hasard dans un établissement d'enseignement secondaire général de Virton. Peut-on considérer que cet échantillon est représentatif des adolescents scolarisés en Belgique ?

2. Dans quelles conditions l'ensemble des étudiants de l'ULB serait-il considéré comme une population ?

3. Dans quelles conditions l'ensemble des étudiants de l'ULB serait-il considéré comme un échantillon ?

4. Si l'ensemble des étudiants de l'ULB était considéré comme un échantillon, s'agirait-il d'un échantillon aléatoire ? Expliquez votre réponse.

T.P. 4 : Exercice 2

Échantillons et population - Indépendance du tirage

1. Quelle est la différence entre n et N ?

2. Je tire un échantillon de 4 personnes dans une population de 10, de manière aléatoire. Le tirage est-il indépendant ? Que faut-il faire pour qu'il le soit ?

3. Je tire un échantillon de 4 personnes dans une population de 6000, de manière aléatoire. Le tirage est-il indépendant ? Que faut-il faire pour qu'il le soit.

4. Un chercheur décide de mener une étude relative aux convictions religieuses des Congolais vivant en Belgique. Il mène à cette fin une série d'entretiens au sein de plusieurs familles, toutes vivant à proximité du quartier Matongué[19]. Un échantillon récolté de cette manière est-il aléatoire ? Expliquez pourquoi.

T.P. 4 : CHAPITRE 5

B. Présentation des données : Distributions & Représentations graphiques

Exercice 1 : Types de variables et Représentation graphique

1. Comment appelle-t-on une variable dont les valeurs sont discontinues ? Donnez-en un exemple.

2. Donnez des exemples de variables continues.

3. Concrètement comment traite-t-on ces variables continues?

T.P. 4 : Exercice 2

Vocabulaire statistique descriptive

Observons le tableau de données ci-joint (Annexe 2). Utilisons notre bon sens pour l'analyser. Une première approche des statistiques peut se faire de manière très intuitive.

1. Chaque ligne du tableau (sauf la première) représente :

[19] Matongé est un quartier situé à Ixelles, à la Porte de Namur. Il est essentiellement fréquenté par des Africains Congolais, mais aussi des Rwandais, Burundais, Maliens et Sénégalais.

2. D'une ligne à l'autre, les valeurs attribuées à un sujet différent peuvent **varier**. Chaque colonne représente donc :

Le tableau que nous sommes en train d'analyser est intéressant, mais peu économique et peu lisible en tant que tel.

3. Comment pourrions-nous représenter la variable « cheveux » de manière plus « économique » ? Placez ces données dans un tableau au nombre de lignes le plus limité possible.

4. Donnez les fréquences relatives et les fréquences relatives exprimées en pourcentages pour cette variable.

5. Quelles valeurs représentent les notations n_2 et f_3 et quelle est la signification de ces valeurs ?

6. Représentez les données que vous avez regroupées à l'exercice 4 sous la forme de deux diagrammes en barres, l'un pour les fréquences absolues (nombre absolu de personnes concernées), l'autre pour les fréquences relatives (exprimées en %). Comparez les deux graphiques. Notez sur chaque graphe la légende, le titre, le sens et l'échelle de l'axe Y.

7. Repérez dans le tableau (Annexe 2) les variables qualitatives et quantitatives. À quelles échelles de mesure renvoient-elles ? Donnez les 2 méthodes utilisées pour coder les variables qualitatives.

8. Dressez le tableau des fréquences des notes en complétant le tableau suivant (cf. annexe 2) :

j	Notes	Fréq. abs.	Fréq. abs. cum.	Fréq. rel. en %	Fréq. rel. cum.
1	11				
2	12				
3	14				

4	15				
5	16				
6	18				
	total				

9. Représentez graphiquement les fréquences absolues de la variable « note ». N'oubliez pas d'indiquer un titre, le sens des axes et une légende. Comment s'appelle un tel graphe ? Comment s'appelle l'axe horizontal et que représente-t-il ?

10. Que peut-on dire par rapport à la surface de chaque rectangle dans un histogramme et par rapport à la surface totale des rectangles ? Comparez à ce qu'il en est pour un diagramme en barres.

11. Représentez sous la forme d'une distribution de fréquences groupées en classes de 3 points la variable *note*. Indiquez les centres des classes, les fréquences absolues, relatives, relatives cumulées et relatives cumulées exprimées en %.

T.P. 4 : Exercice 3

Distributions de fréquences

1. Comment appelle-t-on le tableau ci-dessous ?

j	Notes sur 20 X_j	Fréquences absolues n_j	Fréquences absolues cumulées	Fréquences relatives f_j	Fréquences relatives cumulées
1	**X_1=12**	**n_1=3**	**3**	**f_1=0.3**	**0.3**
2	**X_2=13**	**n_2=1**	**4**	**f_2=0.1**	**0.4**
3	**X_3=14**	**n_3=2**	**6**	**f_3=0.2**	**0.6**
4	**X_4=16**	**n_4=1**	**7**	**f_4=0.1**	**0.7**
5	**X_5=17**	**n_5=1**	**8**	**f_5=0.1**	**0.8**
6	**X_6=18**	**n_6=2**	**10**	**f_6=0.2**	**1.00**
	TOTAL	**n=10**		**1**	

2. Combien de personnes constituent notre échantillon ? .....

3. En regardant ce tableau, estimez, si on tirait au hasard une personne, quelle serait la probabilité qu'elle ait eu 14/20 à son examen. .....

4. En regardant ce tableau, estimez, si on tirait au hasard une personne quelle serait la probabilité qu'elle ait obtenu un note inférieure ou égale à 14. .....

5. Regardez le tableau suivant.

j	Notes sur 20 X_i	Fréquences absolues n_j	Fréquences absolues cumulées	Fréquences relatives f_j	Fréquences relatives cumulées
1	12	3	3	0.3	0.3
2	13	3	4	0.1	0.4
4	14	2	7	0.2	0.6
4	16	1	7	0.1	0.7
5	17	1	8	0.1	0.8
6	18	1	9	0.1	0.9
7	18	1	10	0.1	1.4
	TOTAL	n = 10		1,4	

a. Combien de sujets constituent l'échantillon ? .....

b. Il contient des erreurs ou des imprécisions (une erreur par colonne). Corrigez-les.

T.P. 4 : Exercice 4

Représentation sous forme de tiges et feuilles

Voici un ensemble de données brutes, correspondant au nombre de kilomètres que parcourent 40 employés d'une entreprise pour rejoindre leur lieu de travail à partir de leur domicile.

1,5,5,8,10,10,11,12,14,15,15,16,17,20,20,20,22,23,24,24,25,26,27,27,28,28,29,30,32,34,37,38,39,39,39,39,42,43,44,48.

1. Veuillez présenter ces données sous forme de tiges et feuilles, comme proposé par Tukey (1977).

2. Que pouvez-vous dire au sujet de la forme de la distribution ?

3. Est-ce correct de dire que seul le diagramme en tiges et feuilles permet de détecter les valeurs aberrantes et que les histogrammes ne présentent pas cet avantage ?

4. Est-ce correct de dire que lorque certaines valeurs semblent exceptionnellement élevées (ou faibles) par rapport à l'ensemble de la distribution, celles-ci doivent systématiquement être exclues de l'échantillon ?

T.P. 4 : Exercice 5
Distribution observée des fréquences absolues et relatives
Diagramme en barres

La clinique de santé mentale d'une université utilise les lettres suivantes pour coder les principaux types de problèmes poussant les patients à demander une assistance :

A : Anxiété générale
B : Dépression générale
C : Problèmes liés à la sexualité
D : Problèmes liés à l'alcool et aux stupéfiants
E : Problème de comportement social
F : Problèmes familiaux
G : Autres problèmes

Cinquante-quatre patients se sont rendus à la clinique un jour donné. On a attribué à chacun d'eux une lettre en fonction du problème dont ils souffraient.

A	B	B	E	B	D	B	G	F	G	B
B	C	E	B	B	A	B	B	B	G	B
C	F	D	G	G	D	D	G	G	B	F
G	A	G	A	F	G	G	G	C	D	E
B	B	B	G	G	G	A	B	C	B	

. Sur quel type d'échelle nous situons-nous dans cet exercice ?

2. Construisez le tableau de la distribution des fréquences absolues et relatives associé à ces observations.

3. Dessinez le diagramme en barres des fréquences relatives correspondant.

4. Commentez les résultats des deux points précédents.

T.P. 4 : Exercice 6

Histogramme et polygone des effectifs

Diagramme en tiges et feuilles

Considérons la série statistique ordonnée suivante relative aux âges des membres d'un club sportif :

17	18	19	20	22	22	23	23	24	25
25	26	26	27	27	27	27	27	28	28
28	28	28	29	29	29	29	30	30	30
30	30	31	31	31	31	31	32	32	33
33	34	35	35	36	36	38	39	40	41

1. Groupez cette série en classes et construisez le tableau de la distribution observée des fréquences absolues (effectifs) et relatives correspondant. Prenez, pour ce faire, 5 classes de même largeur. Calculez les valeurs centrales des classes.

2. Construisez un diagramme en tiges et feuilles pour cette série.

3. Dessinez l'histogramme et le polygone des effectifs de cette distribution observée.

4. Commentez vos résultats.

T.P. 4 : Exercice 7

Mise en situation : Article

Auto-régulation des immunoglobulines salivaires A par les enfants

Karen Olness, MD, Timothy Culbert, MD, and Donald Uden

Des observations et études cliniques ont montré que les enfants présentent l'habileté d'utiliser une variété de techniques d'imagerie mentale comme traitement de plusieurs problèmes aigus et chroniques. Citons l'hémophilie, l'arthrite, l'énurésie, les migraines, l'incontinence fécale.
Quelques études réalisées chez la personne adulte suggèrent que certains aspects de la fonction immunitaire puissent être soumis à un contrôle volontaire par le biais de l'hypnose.

La présente étude s'intéresse à la possibilité d'une modulation volontaire du système immunitaire chez les enfants suite à une séance d'auto-hypnose.

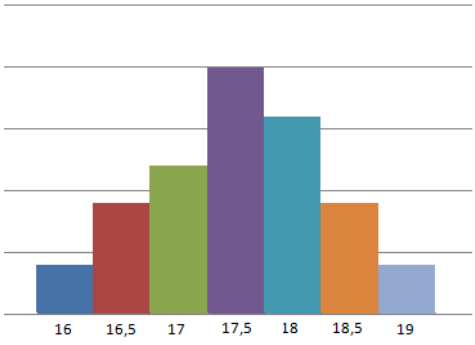
Pour ce faire, les auteurs ont montré à des enfants (âgés entre 6 et 12 ans) une vidéo présentant des marionnettes. Une marionnette représentait un virus; l'autre, qui ressemblait à un policier, représentait le système immunitaire. Cette vidéo constituait ainsi une illustration simplifiée du fonctionnement interne du corps aisément compréhensible par les enfants. Une fois la vidéo visionnée les enfants étaient soumis à une séance d'auto-hypnose. Pour un premier groupe d'enfants, cette séance n'était associée à aucune suggestion spécifique qui puisse augmenter le taux d'immunoglobuline (groupe A). Pour un second groupe d'enfants, cette séance était associée à des suggestions spécifiques visant à augmenter le niveau d'immunoglobulines : on demandait aux enfants de fermer les yeux, de se relaxer et d'imaginer de nombreuses marionnettes policiers parcourant leur corps (groupe B). L'hypothèse étant que les enfants du second groupe montreront une plus grande augmentation du niveau d'immunoglobuline que ceux du premier groupe.

L'analyse des échantillons de salive relevés chez chacun des enfants a révélé une augmentation substantielle du niveau d'immunoglobuline chez les enfants ayant expérimenté la séance d'auto-hypnose associée à des suggestions spécifiques (groupe B). C'est-à-dire que le système immunitaire de ces enfants s'est mis à fonctionner comme s'il combattait de vraies infections. Cette augmentation significative n'a pas été relevée dans l'autre groupe d'enfants (groupe A).

1. Quelle est la population à laquelle s'intéressent les auteurs ?

2. Les auteurs rapportent les résultats obtenus auprès d'un échantillon de 16 participants. S'ils s'étaient contentés d'analyser les résultats de leur échantillon sans tenir compte de la population dont il est issu, nous nous situerions dans une optique ..... Cependant, le but des auteurs est d'inférer des caractéristiques de l'entièreté de la population d'intérêt à partir des résultats de leur échantillon. Nous nous situons donc dans une optique .....

Voici les données récoltées auprès du groupe B :



3. Comment s'appelle ce type de présentation de données ? Pourquoi a-t-on utilisé un tel graphique pour représenter les données ?

4. A partir de ce graphique, construisez le tableau de données correspondant (dans sa version économique). Incluez-y les indices j, les effectifs, les fréquences relatives, les effectifs cumulés et les fréquences relatives cumulées.

T.P. 4 : CHAPITRE 5

B. Les quantiles et les boîtes à moustaches

Partie 1 : Quantiles basés sur des séries statistiques

1. Quelle est la différence entre le rang médian et la médiane ?

2. Comment détermine-t-on la valeur du rang médian pour une série statistique ?

3. Voici les données d'un test de QI de trois enfants de 8 ans : 105, 115, 95.

   a. Représentez-les graphiquement avec l'effectif en ordonnée et le QI en abscisse.

   b. Ordonnez ces données par ordre croissant et déterminez la médiane intuitivement, puis utilisez la formule pour la trouver.

4. Soit la série statistique suivante : 99, 100, 113, 145

   a. Représentez-les graphiquement.

   b. Ordonnez ces données par ordre croissant et déterminez la médiane intuitivement, puis utilisez la formule pour la trouver.

   c. Déterminez les quartiles.

T.P. 4 : Les quantiles et les boîtes à moustaches

Partie 2 : Détermination des quantiles dans des distributions de fréquences non groupées dont les données sont considérées comme discrètes

La valeur de la médiane et des quartiles est facile à déterminer sur la seule base des rangs quand on a de petites séries, mais pour les autres quantiles et pour des séries ou des distributions plus grandes, il est beaucoup plus simple de se baser sur les distributions de fréquences relatives cumulées.

Idéalement, la médiane devrait correspondre à la valeur de la variable dont la fréquence cumulée est d'exactement 50%, mais cet idéal ne peut généralement être atteint quand on détermine la médiane à partir de la distribution des fréquences relatives cumulées d'une variable discontinue ou considérée comme telle. Il en est de même pour les autres quantiles.

A. Ci-dessous les données relatives aux résultats d'étudiants de BA2 à un cours (cours6) noté sur 20. Considérez ces valeurs comme discrètes (discontinues).

Note sur 20 (Cours6)	Effectif	Effectif cum.	Fréq. Rel. Cum.
7	1	1	0,01
8	1	2	0,01
9	4	6	0,04
10	8	14	0,09
11	6	20	0,13
12	38	58	0,36
13	25	83	0,52
14	23	106	0,66
15	23	129	0,81
16	18	147	0,92
17	11	158	0,99
18	2	160	1,00
n=	160		

a. Pour quelle valeur de la variable cours6, ordonnée de manière croissante, les effectifs cumulés dépassent-ils la moitié de l'effectif total ? Justifiez de deux manières votre réponse, sur base des fréquences absolues cumulées et relatives cumulées.

b. Déterminez les 1er, 2ème et 3ème quartiles ainsi que le 36ème percentile.

Premier quartile :	
Deuxième quartile :	

Troisième quartile	
36ème percentile	

c. La deuxième valeur de la variable cours 6 déterminée au point précédent est le quantile ½, appelé aussi Q_2 ou médiane. Que signifie ce paramètre ? Que pouvez-vous dire de cette valeur en la resituant dans le contexte d'une série statistique ?

d. Tracez un diagramme en barres des fréquences absolues cumulées et indiquez graphiquement où se situent la médiane et le 36ème percentile.

Rappel : Dans le cas particulier où la proportion correspondant au quantile est exactement atteinte pour une valeur de la variable, le quantile est défini par convention comme la moyenne entre cette valeur et la valeur suivante.

e. Que vaut l'écart interquartile ?

f. Tracez la boîte à moustaches (version rudimentaire) pour le cours 6.

g. Calculez les valeurs pivots pour les données du cours 6. À quoi correspondent-elles ?

h. Déterminez les valeurs adjacentes gauche et droite pour nos données :

i. Tracez la boîte à moustaches (version sophistiquée) pour le cours 6. Indiquez-y les valeurs pivots, les valeurs adjacentes et les valeurs extrêmes.

B. Une étude a été planifiée pour déterminer si la réactivité émotionnelle des enfants de familles monoparentales est différente de celle d'enfants de familles traditionnelles avec deux parents. Un échantillon est prélevé pour chaque type de famille. Ces enfants sont soumis à un test de réactivité émotionnelle. Pour ce test, plus le résultat du score est élevé, plus l'enfant présente une réactivité émotionnelle prononcée.

groupe	score
1	6
1	9
1	4
1	13
1	14
1	9
1	8
1	12
1	11
1	9
2	12
2	18
2	14
2	10
2	19
2	8
2	15
2	11
2	10
2	13
2	15
2	16

Groupe 1 : Famille monoparentale

		Frequency	Cumulative Percent
Valid	4	1	10,0
	6	1	20,0
	8	1	30,0
	9	3	60,0
	11	1	70,0
	12	1	80,0
	13	1	90,0
	14	1	100,0
	Total	10	

Groupe 2 : Famille traditionnelle à deux parents

		Frequency	Cumulative Percent
Valid	8	1	8,3
	10	2	25,0
	11	1	33,3
	12	1	41,7
	13	1	50,0
	14	1	58,3
	15	2	75,0
	16	1	83,3
	18	1	91,7
	19	1	100,0
	Total	12	

a. Déterminez les quantiles et l'écart interquartile, et détaillez vos calculs.

b. Tracez les boîtes à moustaches correspondantes, et commentez-les. Déterminez VD et VI.

T.P. 4 : Les quantiles et les boîtes à moustaches

Partie 3 : Détermination des quantiles dans le cas de distributions de fréquences pour des données considérées comme continues

Ci-dessous les données relatives au poids de 181 étudiants de psycho de l'ULB.

Poids	Fr. abs.	Fr. abs. Cum.	Fr. rel.	Fr. rel. Cum.
40	1	1	0,006	0,006
41	1	2	0,006	0,011
45	4	6	0,022	0,033
46	3	9	0,017	0,050
47	1	10	0,006	0,055
48	3	13	0,017	0,072
49	1	14	0,006	0,077
50	4	18	0,022	0,099
51	4	22	0,022	0,122
52	13	35	0,072	0,193

Poids	Fr. abs.	Fr. abs. Cum.	Fr. rel.	Fr. rel. Cum.
65	4	133	0,022	0,735
66	1	134	0,006	0,740
67	4	138	0,022	0,762
68	5	143	0,028	0,790
69	5	148	0,028	0,818
70	2	150	0,011	0,829
71	1	151	0,006	0,834
72	3	154	0,017	0,851
73	2	156	0,011	0,862
74	2	158	0,011	0,873

53	9	44	0,050	0,243
54	6	50	0,033	0,276
55	11	61	0,061	0,337
56	4	65	0,022	0,359
57	7	72	0,039	0,398
58	15	87	0,083	0,481
59	3	90	0,017	0,497
60	16	106	0,088	0,586
61	4	110	0,022	0,608
62	7	117	0,039	0,646
63	6	123	0,033	0,680
64	6	129	0,033	0,713

75	7	165	0,039	0,912
77	1	166	0,006	0,917
78	4	170	0,022	0,939
79	1	171	0,006	0,945
80	3	174	0,017	0,961
82	1	175	0,006	0,967
83	1	176	0,006	0,972
84	1	177	0,006	0,978
85	1	178	0,006	0,983
86	1	179	0,006	0,989
88	1	180	0,006	0,994
91	1	181	0,006	1,000

1. Déterminez à partir du tableau ci-dessus les quantiles suivants (à l'aide des deux méthodes mentionnées dans le cours théorique), ainsi que l'écart interquartile :

Réponses :

1er quartile

Médiane
3ème quartile
Écart interquartile

2. Tracez la boîte à moustaches (version rudimentaire) correspondant aux données ci-dessus.

3. Calculez les valeurs pivots correspondant à nos données :

4. Déterminez les valeurs adjacentes pour nos données :

5. Tracez la boîte à moustaches (version sophistiquée).

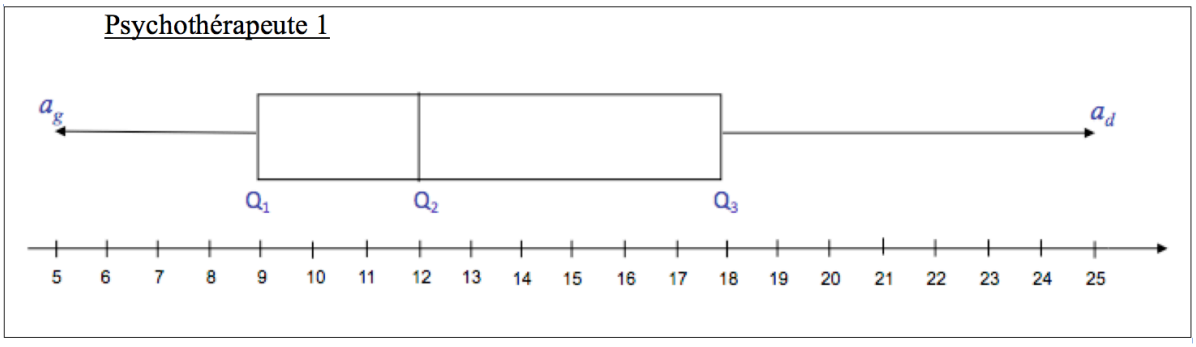
T.P. 4 : Les quantiles et les boîtes à moustaches

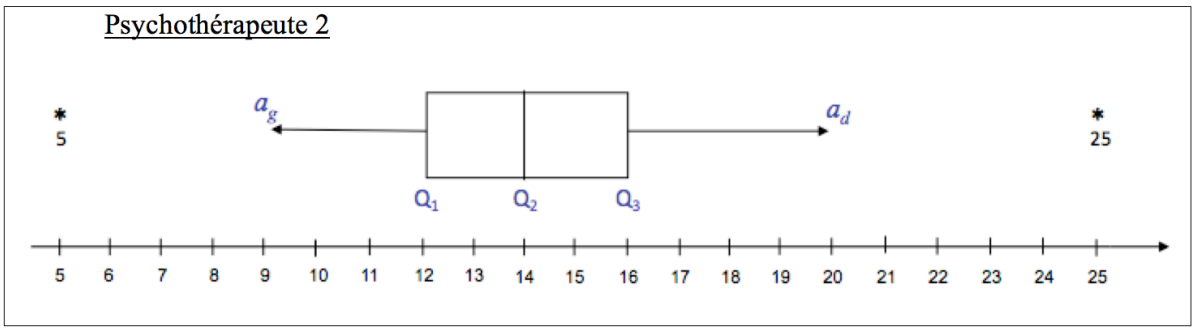
Partie 4 : Les boîtes à moustaches – Observation

A. Deux jeunes psychothérapeutes s'interrogent sur la durée de leur psychothérapie. Pour tenter de répondre à cette question. Ils rapportent le nombre de séances qui ont été nécessaires à leur psychothérapie. Le premier psychothérapeute répertorie le nombre de séances de 30 patients alors que le second, ayant moins d'expérience ne peut les répertorier que pour 20 patients.

1. Quelle est la variable que l'on va mesurer ? Comment s'appelle-t-elle ? Sur quelle type d'échelle de mesure est-elle être mesurée ?

Voici les deux boîtes à moustaches correspondant au nombre de séances répertoriées par les deux psychothérapeutes :

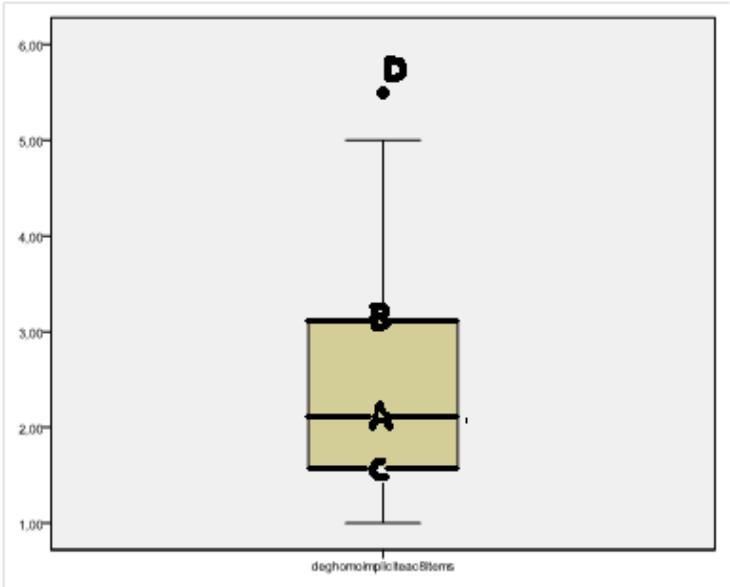




2. Que représente l'écart interquartile ? Quel est son intérêt ?

3. Commentez les deux boîtes à moustaches ci-dessus (dispersion, symétrie, valeurs extrêmes, …). Que pouvez-vous conclure ? Quelle est l'influence de l'écart interquartile sur la qualité du modèle de prédiction et la marge d'erreur ?

B. Voici une boite à moustaches créée sur base des scores de 131 sujets obtenus sur une échelle de mesure relative au dégoût implicite de l'homosexualité. Cette échelle regroupait une série d'items que les participants devaient évaluer sur base d'une échelle de Likert à 7 points (1 = pas du tout d'accord ; 7 = tout à fait d'accord).

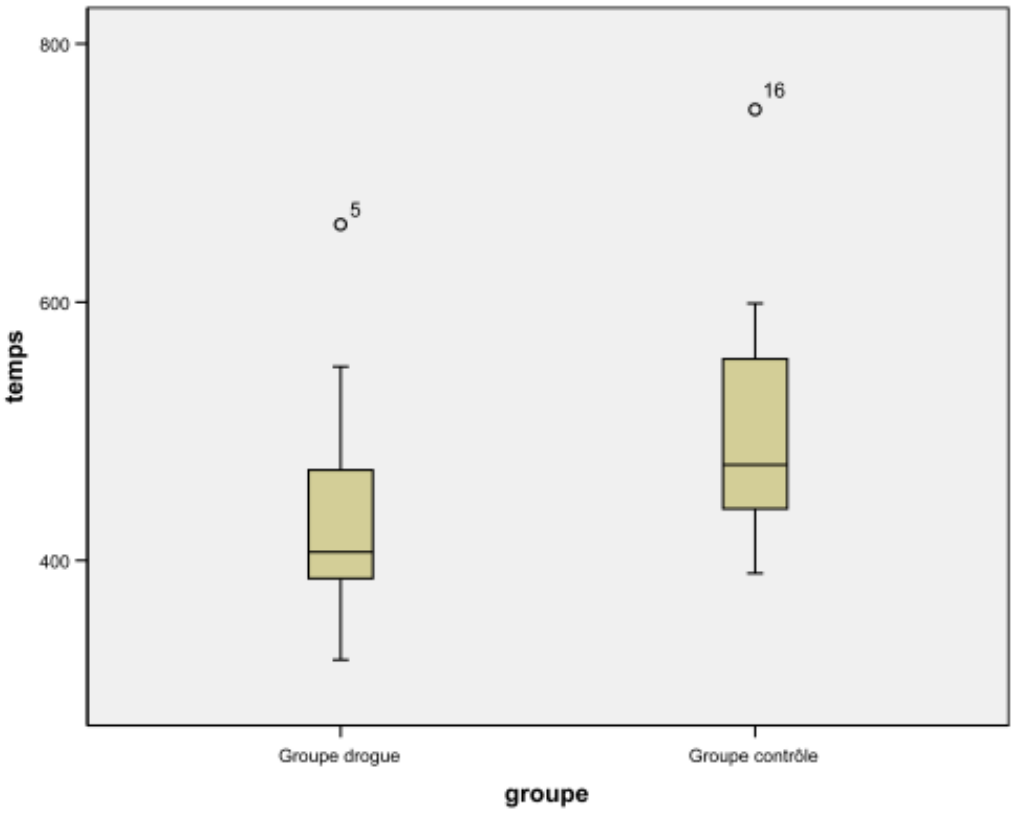


1. Veuillez localiser, sur base de ce graphique, les différents éléments repris dans le tableau suivant :

	Lettre :
Q_1	
Q_3	
Médiane	
Valeur extrême	

2. Que pouvez-vous déduire de la position de la médiane ?

C. Une équipe de chercheurs a de bonnes raisons de croire qu'un faible dosage d'une certaine drogue augmente la vitesse avec laquelle les gens peuvent prendre des décisions. Ils décident d'essayer de confirmer ceci en menant une expérience dans laquelle les temps de décision (en millisecondes) de 14 personnes ayant ingéré de la drogue sont comparés avec ceux d'un groupe contrôle de 14 autres personnes qui ont réalisé la tâche sous une condition placebo. Les résultats des mesures sur les deux groupes sont repris ci-après :



a. Que représentent les deux boules du dessin ? Quelle conclusion peut-on tirer au vu de ces deux boîtes ?

b. Quelle est la variable dépendante ? indépendante ?

T.P. 4 : CHAPITRE 5

EXERCICE RECAPITULATIF

Soit le **tableau 5.1**. du cours théorique : Séries statistiques de 45 étudiants de BA1

.

Num	Théorie	Prat.	Num	Théorie	Prat.	Num	Théorie	Prat.
1	18	15	**16**	12	16	**31**	15	12
2	13	15	**17**	17	12	**32**	12	15
3	17	14	**18**	13	12	**33**	12	12
4	16	14	**19**	14	15	**34**	17	14
5	17	15	**20**	14	15	**35**	14	10
6	15	13	**21**	17	14	**36**	13	11
7	16	14	**22**	13	14	**37**	11	13
8	14	15	**23**	14	9	**38**	13	14
9	14	14	**24**	13	11	**39**	14	12
10	12	15	**25**	15	13	**40**	13	12
11	16	12	**26**	18	14	**41**	12	14
12	12	18	**27**	11	10	**42**	13	12
13	16	14	**28**	15	15	**43**	12	10
14	14	14	**29**	16	13	**44**	11	14
15	12	14	**30**	12	14	**45**	14	15

Et, le **tableau 5.2**, transnumérisation du Tableau 5.1. pour les notes obtenues en théorie :

Cotes	Fréquences absolues	Fréquences relatives	Fréquences cumulées	Fréquences cumulées en %
11	3	0,067	3	6,7
12	9	0,20	12	26,7
13	8	0,18	20	44,4
14	9	0,20	29	64,4
15	4	0,089	33	73,3

16	5	0,11	38	84,4
17	5	0,11	43	95,6
18	2	0,044	45	100

1. Réalisez un tableau similaire au tableau 5.2 pour les notes obtenues au cours pratique.

2. Représentez graphiquement les données associées aux cotes à l'examen théorique d'une part, et à l'examen pratique d'autre part pour ces 45 étudiants de BAC1.

3. Déterminez les différents quartiles et réalisez les boîtes à moustaches en considérant les points du cours théoriques comme une variable discontinue et les points du cours pratique comme une variable continue.

T.P. 4 : Annexe

Exemple de tableau de données

Num	Groupe	Âge	genre	Taille (cm)	Poids (kg)	Yeux	Chev	Sport	Fumer	Lunette	note / 20	nombre de frères
1	1	22	F	166	55	bleus	blonds	1	1	1	18	1
2	1	19	M	165	59	bleus	blonds	2	4	1	12	0
3	1	22	F	173	86	marron	bruns	2	1	0	11	2
4	1	20	M	178	55	bruns	bruns	2	2	0	14	1
5	2	20	M	178	61	bleus	bruns	1	1	1	15	1
6	2	19	F	155	58	bruns	châtains	2	1	1	16	2
7	2	19	F	171	70	verts	bruns	2	3	0	15	2
8	2	20	F	165	58	verts	blonds	3	1	0	18	0
9	3	21	F	170	58	bruns	noirs	2	3	1	12	3
10	3	19	M	181	74	bruns	bruns	4	3	0	12	0
11	3	19	F	159	52	verts	châtains	2	1	0	11	0
12	3	19	F	162	45	verts	châtains	3	1	1	16	2

CHAPITRE 6 : EXPLORATION ALGEBRIQUE DES DONNEES A UNE DIMENSION

6.1. Introduction

Jusqu'à présent nous n'avons envisagé la présentation d'une distribution statistique que sous forme graphique. Cependant, plusieurs caractéristiques sont importantes à déterminer algébriquement. On peut distinguer trois grandes catégories d'indicateurs algébriques essentiels :

1. **Les mesures de la tendance centrale**, qui indiquent les valeurs de la distribution qui sont les plus représentées par les sujets qui la composent (nous avons déjà envisagé la médiane). Elles nous donnent une valeur prédictive pouvant servir de modèle simple.
2. **Les mesures de la dispersion**, qui indiquent jusqu'à quel point les sujets s'éloignent des valeurs centrales de la distribution (nous avons déjà envisagé l'écart interquartile). C'est donc lié à la mesure de l'erreur résiduelle.
3. **La détermination algébrique de la symétrie et de l'aplatissement** d'une distribution (deux indicateurs importants qui caractérisent une distribution donnée).

C'est l'objectif de ce chapitre. De plus, nous allons faire le lien entre la mesure algébrique de ces mesures et l'établissement du modèle prédictif que nous envisageons depuis le deuxième chapitre.

6.2. Les mesures de la tendance centrale

Parmi elles, on retrouve la moyenne, le mode et la médiane. Nous avons déjà envisagé la médiane, nous reverrons son importance dans la présentation qui suit mais nous n'allons pas consacrer de point spécifique à cette notion. Nous envisagerons en revanche le mode et la moyenne.

Pour chacun de ces indices, nous imaginerons la cote sur 10 à un examen théorique d'ANAD1 (je pense que vous êtes maintenant habitués à imaginer ce genre de situation, bien qu'ici l'échelle soit sur 10 au lieu de 20 pour vous montrer à quel point ça n'a aucune importance) pour 9 étudiants fictifs (Tableau 6.1.). Nous prendrons également l'habitude d'utiliser

certaines notations[20]. On symbolise généralement les variables par des lettres majuscules choisies à la fin de l'alphabet. On tend à utiliser X quand il n'y a qu'une seule variable.

Dans une série statistique, on utilise l'indice *I* pour indicer cette variable et *i* pour se référer à un étudiant quelconque. L'indice *I* varie de 1 à *n* (donc, il peut prendre les valeurs 1, 2, 3, …, *i*, …, *n*). On symbolise par X_i (X indicé *i*) la cote d'un étudiant quelconque. La cote de l'étudiant n° 2 est symbolisée par $X_2 = 5$, celle de l'étudiant n° 8, par $X_8 = 6$, etc.

Dans une distribution de fréquences, chaque valeur possible de la variable apparaît dans la 1ère colonne. Comme ces valeurs sont numériques, elles sont rangées par ordre croissant. On symbolise les valeurs observées (ou les valeurs possibles) de la variable par une lettre minuscule. Ici, les valeurs observées de la variable X sont représentées par x_j avec *J* pouvant prendre les valeurs 1 à 5. Donc, x_1 a pour valeur 3, x_2 a pour valeur 4, … x_5 a pour valeur 7.

Tableau 6.1. : (a) Série statistique de n = 9 représentant les cotes sur 10 à un examen. (b) même série représentée sous forme de distribution de fréquences.

(a) Série

Num	Cotes (X_i)
1	3
2	5
3	7
4	4
5	5
6	6
7	5
8	6
9	4

(b) Distribution

Valeurs observées (notées x_j)	Fréquences absolues (notées n_j)	Fréquences relatives (notées f_j)	Fréquences relatives exprimées en %
3	1	0,111	11,1
4	2	0,222	22,2
5	3	0,333	33,3
6	2	0,222	22,2
7	1	0,111	11,1
Total	9	1,00	100

6.2.1. Le mode

Le mode se définit très simplement comme la classe la plus représentée. Dans le Tableau 6.1 (a ou b) vous constaterez aisément que la cote 5 est la plus représentée puisque trois sujets

[20] Je rappelle qu'une des difficultés majeures de la compréhension des statistiques est la notation. Soyez donc bien attentifs aux conventions d'écriture et tentez de vous en imprégner dès que possible.

l'ont obtenue. C'est donc le mode. Remarquez qu'une distribution peut être multimodale. Par exemple, si un seul étudiant avait eu la cote 5 (et que ma série statistique ne comportait plus que 7 sujets), les cotes 4 et 6 auraient toutes deux été les plus représentées (deux sujets par classe) et il y aurait eu deux modes (distribution bimodale).

Le mode est donc une manière de modéliser une distribution. En effet, si j'admets, comme modèle, que les étudiants ont une cote en ANAD de 5/10, j'aurai raison dans 33% des cas. Si je choisis, n'importe quelle autre valeur, j'aurai une chance plus faible de ne pas me tromper. Par exemple, si je choisis 4/10 plutôt que 5, je n'ai que 22% de chances de ne pas me tromper.

Le grand avantage du mode est d'être insensible aux valeurs aberrantes. Supposons que j'aie mal encodé mes données et que j'ai écrit 77, ou même 777, au lieu de 7 pour le troisième sujet. Le mode resterait identique, ce serait 5/10. Ce ne sera pas du tout le cas pour la moyenne, comme nous allons le voir. En revanche, le grand désavantage du mode est qu'il ne dépend finalement que de la ou des quelques valeurs les plus représentées et est totalement insensible au reste de la distribution (même des valeurs qui n'ont rien d'aberrantes).

6.2.2. La moyenne

6.2.2.1. Procédure de calcul de la moyenne

Lorsque l'on parle de moyenne en statistique, on entend toujours (à notre niveau) la moyenne arithmétique (l'alternative étant la moyenne géométrique [21]). La moyenne arithmétique consiste à prendre la somme des valeurs et à la diviser par le nombre de valeurs qui constituent cette somme. Donc, en termes mathématiques, la moyenne devient :

Formule de définition de la moyenne calculée à partir d'une série statistique

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

[21] Pour mémoire, la moyenne géométrique consiste à prendre le produit des valeurs et à extraire la racine de degré égal au nombre de valeurs qui constitue ce produit, mais vous n'aurez jamais à utiliser une telle procédure.

Le problème qui se pose avec cette définition de la moyenne survient lorsqu'il y a un grand nombre de données, rendant l'encodage relativement ardu. Dans ce cas, on peut utiliser les données présentées sous forme de distribution de fréquences, et la formule devient :

Formule de définition de la moyenne calculée à partir d'une distribution non groupée de fréquences absolues

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j$$

Formule de définition de la moyenne calculée à partir d'une distribution non groupée de fréquences relatives

$$\overline{X} = \sum_{j=1}^{J} f_j x_j$$

Cependant, en pratique et à terme, vous n'aurez jamais à calculer cette moyenne dans l'avenir puisque, dès l'année prochaine, vous utiliserez des logiciels informatiques capables de réaliser ces simples calculs en une fraction de seconde. Il n'en reste pas moins que, cette année-ci, vous devez non seulement être capable de les réaliser, mais en plus de bien comprendre les enjeux de ces indicateurs, tellement ils sont essentiels pour la suite des événements.

Dès lors, il est important de connaître les propriétés du signe de sommation (∑) résumées dans la Figure 6.1. Remarquez que je n'utilise que la lettre sigma majuscule. Pourtant, si je devais être complet, je devrais y mettre l'indice *i*, qui peut aller de 1 jusque *n*, comme c'est le cas dans la Figure 6.1. En pratique, on considère cette information comme sous-entendue et on ne l'indique plus (sauf dans les rares cas où il y a deux signes de sommation et deux indices différents comme c'est le cas pour la Figure 6.1.3). Les deux premières règles sont essentielles à comprendre : une constante multipliant la somme (dans la Figure 6.1, elle se nomme *a* sans indice) peut être indiquée après ou avant le signe de sommation (propriété 1) ; La somme de *n* sommes de termes *a* et *b* est égale à la somme de la somme tous les *a* et

de la somme de tous les *b* (propriété 2). La dernière propriété est indiquée par souci d'exhaustivité mais n'est pas essentielle pour les problèmes qui nous occuperont.

Figure 6.1. : Propriétés du signe de sommation

$$\sum_{i=1}^{n} ab_i = a\sum_{i=1}^{n} b_i; \qquad (1)$$

$$\sum_{i=1}^{n}(a_i + b_i) = \left(\sum_{i=1}^{n} a_i\right) + \left(\sum_{i=1}^{n} b_i\right); \qquad (2)$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n}\sum_{i=1}^{m} a_{ij}; \qquad (3)$$

Sachant ceci, la moyenne du Tableau 6.1.a peut s'exprimer des manières suivantes :

$$\overline{X} = \frac{1}{9}\sum_{i=1}^{9} X_i = \frac{1}{9}(3+5+7+4+5+6+5+4+6) = \frac{1}{9}(45) = 5 \qquad (1)$$

$$\overline{X} = \sum_{i=1}^{9}\frac{1}{9}X_i = \frac{1}{9}3 + \frac{1}{9}5 + \frac{1}{9}7 + \frac{1}{9}4 + \frac{1}{9}5 + \frac{1}{9}6 + \frac{1}{9}5 + \frac{1}{9}4 + \frac{1}{9}6 = 5 \qquad (2)$$

Pour 9 valeurs, la formule alternative n'est pas nécessaire, mais à titre d'exercice, nous allons néanmoins l'appliquer en se basant cette fois sur le Tableau 6.1.b :

$$\overline{X} = \frac{1}{9}\sum_{j=1}^{5} n_j x_j = \frac{1}{9}\left[(1\times 3) + (2\times 4) + (3\times 5) + (2\times 6) + (1\times 7)\right] = \frac{1}{9}(45) = 5 \qquad (3)$$

Nous n'allons pas réaliser ce calcul pour la dernière formule, étant donné que la fréquence f_j n'est rien d'autre que n_j/n, c'est-à-dire 1/9, 2/9, …, 1/9 ce qui n'est rien d'autre que la distribution du 1/9 dans l'expression (3).

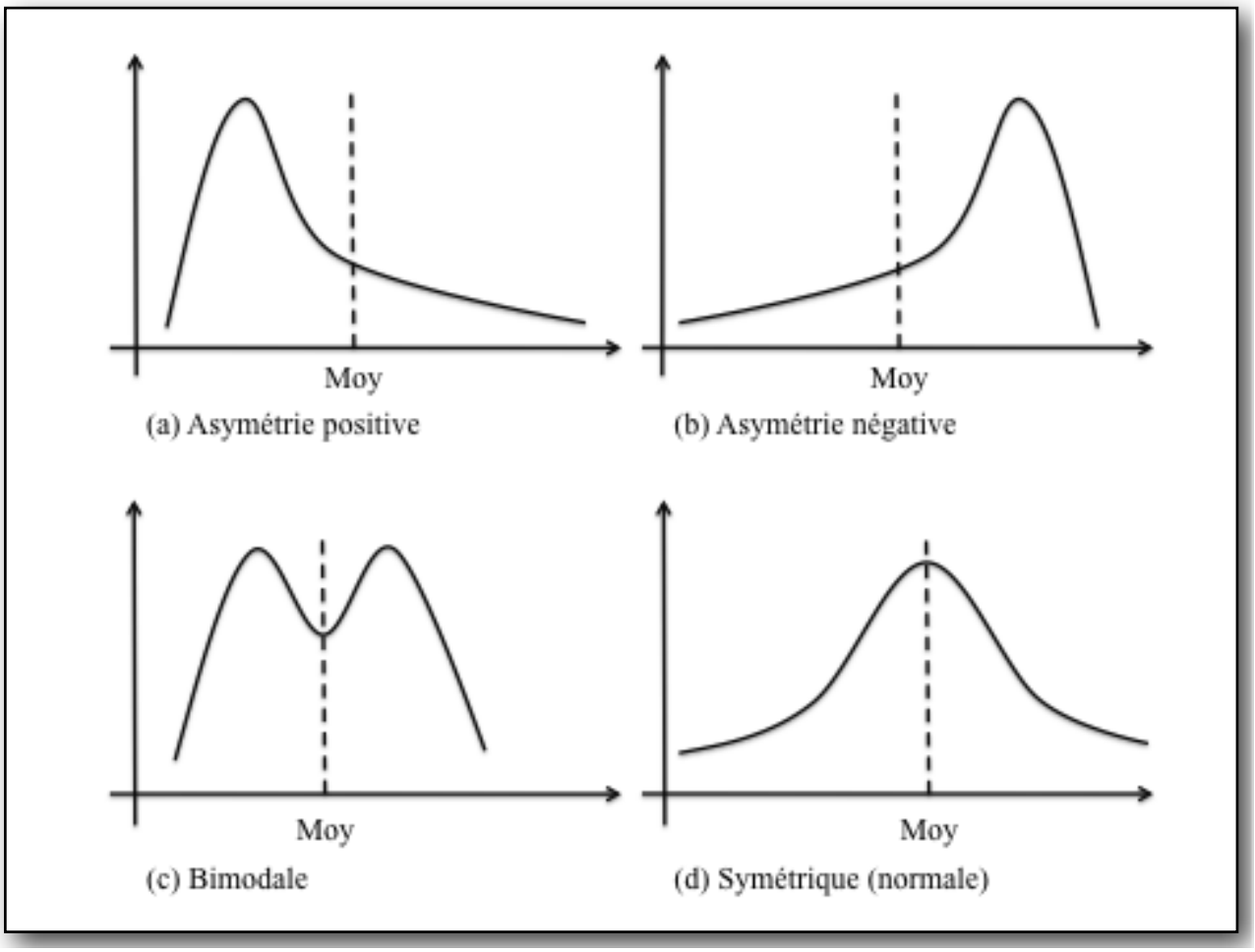
La moyenne est probablement la manière la plus habituelle de représenter la tendance centrale d'une distribution. Elle permet également de modéliser les données de manière efficace, dans la plupart des cas. Envisageons cependant plusieurs cas où la moyenne pose problème.

6.2.2.2. Inconvénients et avantages de la moyenne

Le premier inconvénient est sa sensibilité aux valeurs aberrantes. Rappelez-vous que, lorsque nous avons envisagé le mode, nous nous réjouissions de le voir insensible à un 77 ou 777 qui remplacerait le 7 de la série. Dans le cas de la moyenne, il n'y a pas de quoi se réjouir : le remplacement du 7 par un 77 ferait passer la moyenne de 5 à 12,78 et le remplacement par 777 conduirait à une moyenne de 90,56 (vérifiez ce calcul par vous-même)! Vous conviendrez qu'estimer la distribution à l'aide d'un 90,56 ne permet pas du tout de représenter correctement le résultat habituel des étudiants (surtout à une évaluation cotée sur 10). C'est donc un désavantage sérieux qui nous oblige à être attentifs aux valeurs aberrantes.

Le deuxième inconvénient, lorsque l'on s'exprime en terme de moyenne, est qu'elle peut, si la distribution n'est pas symétrique ou n'est pas unimodale, ne pas la représenter correctement. En effet, rappelez-vous des formes possibles de distribution du chapitre 5 (Figure 5.2 que je remets ci-dessous). Vous remarquerez que la moyenne, lorsque les distributions sont asymétriques (Figure 5.2. a et b), est loin du regroupement le plus important des valeurs observées (qui se trouve au niveau du pic). De même, lorsque la distribution est bimodale (Figure 5.2.c) la moyenne représente une valeur qui n'est, finalement, pas très représentative des données puisque différente des deux pics représentatifs. En revanche, lorsque la distribution est unimodale et symétrique (Figure 5.2.d), vous conviendrez que la moyenne représente parfaitement bien la distribution puisqu'elle correspond à sa valeur la plus probable.

Figure 5.2. : Formes des distributions



Nous revenons donc, pour la deuxième fois, sur l'importance d'avoir une distribution symétrique et unimodale. Heureusement, c'est le plus souvent le cas. Il semble qu'il existe une loi naturelle qui fasse correspondre la plupart des phénomènes observés à une telle distribution. Par exemple, si je me demande quelle est la taille moyenne des hommes au sein de la population belge, j'obtiens 1m77 (comme mentionné auparavant). Il existe évidemment des hommes plus grands ou plus petits. Mais plus on s'éloigne de cette valeur centrale (1m77), moins on trouvera de spécimens de taille correspondante. Il y a un peu moins de gens qui mesurent 1m80, encore un peu moins qui mesurent 1m90, très très peu qui mesurent 2m et identiquement pour les valeurs inférieures à cette moyenne. Ce type de distribution, dite **normale** fera l'objet du chapitre 7. En outre, vous remarquerez donc qu'il est nécessaire de déterminer la symétrie d'une distribution. C'est pourquoi, nous passerons également en revue les indicateurs de symétrie dans ce chapitre.

6.2.2.3. Modèle de la moyenne

Revenons maintenant à notre conceptualisation par modèle. Nous supposerons donc à partir de maintenant que les distributions sont plus ou moins symétriques et unimodales. Je vous ai dit, dès le deuxième chapitre, que nous désirions modéliser la réalité pour pouvoir prédire le mieux possible un événement. Je vous ai également dit que les chercheurs n'étaient pas (toujours) fous et qu'ils étaient tout à fait conscients qu'une prédiction contient une part d'erreur. Leur volonté étant que cette erreur soit la plus petite possible. De sorte que :

Réalité = Modèle + Erreur

Ecrite sous forme mathématique, cette équation prendra la forme suivante, par convention :

$$Y_i = \beta_0 + \varepsilon_i \quad (4)$$

Cette convention d'écriture correspond à celle adoptée dans l'ouvrage de Judd, McClelland, Ryan, Muller, & Yzerbyt (2010) dont s'inspire une bonne partie des cours de BA1, 2 et 3. **L'utilisation de lettres grecques se justifie lorsque l'on fait référence à la population. L'utilisation de lettres latines correspondantes se fera lorsque l'on fait référence à l'échantillon.** Nous entrons donc, dès à présent, dans une logique de statistique inférentielle. En effet, le paramètre β_0 d'une population n'est presque jamais connu. Le plus souvent, nous allons l'estimer à l'aide d'un échantillon. Il en va de même pour l'erreur. Donc, le modèle que nous définirons à partir de l'échantillon peut s'écrire de la manière suivante :

$$Y_i = b_0 + e_i \quad (5)$$

Où b_0 et e_i représentent les paramètres de l'échantillon. Dans l'exemple de la taille des adultes masculins belges, nous avons, au début du cours, imaginé que l'on prélève 100 hommes belges et les mesurions, c'est notre **échantillon**. A partir de cet échantillon nous voulions savoir quelle était la taille moyenne pour tous les Belges, ce qui représente notre **population**, et notre estimation était de 177cm (= b_0). Donc, à partir de notre échantillon nous avons calculé la moyenne qui nous permet de modéliser non seulement notre série statistique des 100 sujets (et nous avons fait des statistiques descriptives) mais également le

paramètre correspondant pour la population que ces 100 sujets sont sensés représenter (et là nous avons fait de l'inférence statistique). On dit que la moyenne de l'échantillon est un **estimateur** de la moyenne de la population. De même, nous allons savoir calculer les erreurs de notre échantillon, en évaluant la différence entre notre prédiction et la taille réelle de chaque sujet, et en tirer des conclusions concernant les caractéristiques des erreurs au sein de la population. Là encore il s'agira d'inférence. Voyons de plus près comment cela se passe pour nos 9 sujets du Tableau 6.1.

Rappelez-vous que la moyenne était un modèle pertinent pour décrire une distribution symétrique (comme l'est celle du Tableau 6.1, représentez-la sous forme d'histogramme si vous n'êtes pas convaincus). Dès lors, nous allons envisager la moyenne comme notre modèle, et considérer l'erreur. Nous pouvons donc écrire :

$$Y_i = 5 + e_i$$

Où 5 est le modèle (b_0), c'est-à-dire notre prédiction (également notée $\hat{Y}_i$), Y_i est la cote à l'examen d'ANAD, et e_i est l'erreur que je commets en utilisant ma moyenne. Cela signifie que, du point de vue des statistiques descriptives, je considérerai qu'un individu appartenant à cet échantillon a obtenu 5/10 à mon examen. D'un point de vue inférentiel, je me dirai qu'à chaque fois que je croise un étudiant de BA1, il a une note de 5/10 à mon examen. En effet, en l'état actuel des choses, l'information concernant les 9 personnes sont les seules informations que je possède concernant cet examen. Dès lors, je peux considérer que je suis plus précis en tenant compte de ces 9 personnes pour inférer quelque chose sur l'entièreté de ma population qu'en décidant moi-même sans aucune information.

Cela peut vous paraître peu crédible, mais imaginez une situation semblable (qui vous semblera sans doute nettement plus crédible…) : l'extraterrestre de Roswell. Nous n'avons absolument jamais vu d'extraterrestres. Imaginons qu'ils existent néanmoins mais que nous n'avons aucune idée de leur apparence. Imaginons également que ce soit bien un vaisseau spatial contenant un extraterrestre qui s'est écrasé près de Roswell en juillet 1947 et pas un ballon sonde espion classé par la défense américaine (tant qu'à choisir une hypothèse improbable, autant que ce soit la plus amusante). Nous sommes donc devant le seul exemplaire d'une population d'extraterrestres d'effectif inconnu et de caractéristiques inconnues également. Admettons que cet extraterrestre mesure 1m60 ait la peau grise et une forme humanoïde. Si vous deviez émettre une prédiction concernant la taille moyenne de la

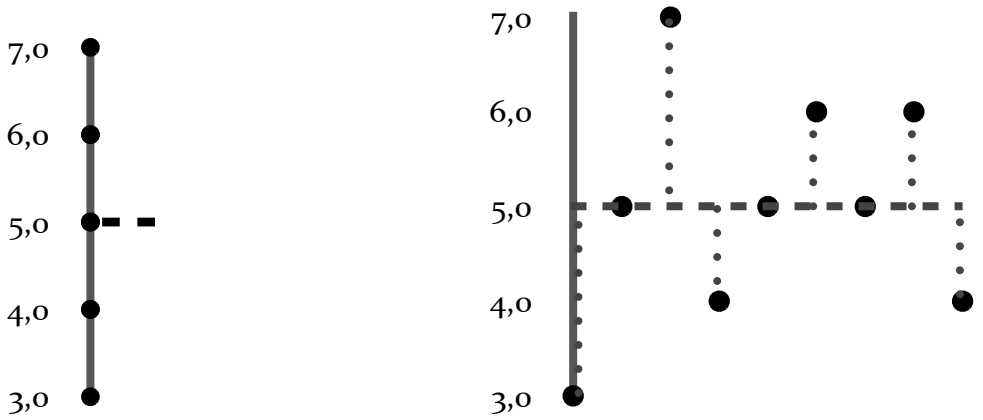
population entière d'extraterrestres de même espèce, votre meilleure estimation serait de dire 1m60, alors que vous n'avez qu'un seul sujet cette fois (et que les extraterrestres sont peut-être des milliards!). Bien entendu, vous vous attendez à ce que certains soient plus grands et d'autres plus petits, mais vous ne vous attendez pas à ce que certains fassent 180m et que d'autres soient microscopiques. En revanche, vous pourriez évidemment être plus précis si ce vaisseau avait contenu un millier d'individus, que vous aviez pu les mesurer tous et en calculer la taille moyenne.

Si vous n'aviez pas vu cet extraterrestre de vos propres yeux, vous n'auriez pas pu tirer ce genre de conclusion. Par exemple, moi qui ai raté cet exemplaire, je continue de penser qu'il vaut mieux chercher des indices de vie de type bactérien là où il y a eu (ou où il y a encore) de l'eau. Je m'attends donc plus à voir des organismes microscopiques fossilisés (sur un débris de planète quelconque qui parviendrait jusqu'à nous) que des organismes d'1m60 parce que j'ai une bonne idée de la taille d'une bactérie.

Revenons maintenant à nos 9 étudiants à qui j'attribue par défaut la note de 5/10. Je peux sans problème calculer mon erreur d'estimation. Il suffit de modifier un petit peu mon équation (5) et d'isoler le terme erreur : $e_i = Y_i - b_0$. Donc, pour mon premier sujet, je fais une erreur de 3 - 5 = -2. Selon le même raisonnement, je fais une erreur pour les autres sujets de, dans l'ordre à partir du deuxième sujet : 0, 2, -1, 0, 1, 0, 1, -1. La Figure 6.2 vous montre comment, à partir de cette série, on peut représenter l'information que nous avons (la cote sur 10) pour nos 9 sujets, soit sur un seul axe (Figure 6.2.a), soit étalé et avec une représentation de l'erreur (Figure 6.2.b).

Figure 6.2. : Représentation des données du Tableau 6.1 sur un seul axe (a) et étendu avec une représentation graphique de l'erreur (b).

(a) Axe vertical, moyenne = 5 (b) Etalement, moyenne = 5



Une information intéressante serait de pouvoir estimer l'entièreté de l'erreur que je commets sur mon échantillon, puisque c'est justement ce que j'essaie de minimiser en utilisant une prédiction. Donc, je désire connaître la somme des erreurs. Cependant, je suis alors confronté à un ennui : mes erreurs sont tantôt supérieures à la moyenne et tantôt inférieures à celle-ci. Dès lors, lorsque j'en calcule la somme, je me retrouve avec une valeur nulle. Vous le voyez très bien graphiquement, mais même algébriquement : $-2 + 0 + 2 + (-1) + 0 + 1 + 0 + 1 + (-1) = 0$.

Il existe deux moyens de contourner ce problème. Soit prendre la valeur absolue de chaque erreur, soit élever toutes les erreurs au carré (et donc ne plus se soucier du signe, puisqu'un chiffre négatif élevé au carré devient positif). Nous allons opter pour la seconde solution (et nous ne serons pas les seuls, l'écrasante majorité des méthodes statistiques sont basées sur cette approche). Ce faisant nous obtiendrons : $(-2)^2 + 0^2 + 2^2 + (-1)^2 + 0^2 + 1^2 + 0^2 + 1^2 + (-1)^2 = 12$. Nous appellerons cette valeur la **somme des carrés de l'erreur** et la symboliseront **SCE**.

Il est intéressant de constater que la moyenne est en fait la valeur qui conduit à la plus petite SCE possible. Démontrons cela en calculant la SCE que nous aurions eu si nous avions utilisé n'importe quelle autre valeur que la moyenne comme modèle. C'est ce que le Tableau 6.3 rapporte comme information (je vous suggère de refaire les calculs par vous-même). Vous pouvez remarquer qu'à l'instant où l'on s'éloigne de la valeur moyenne (5/10) la SCE augmente, passant à 21 lorsque l'on utilise 4/10 comme modèle et augmentant à 51 lorsque

l'on utilise 3/10. J'aurais également pu prendre d'autres valeurs qui ne correspondent à aucune valeur observée comme 10/10 ou 5,1/10, j'en serais arrivé à la même conclusion, vous pouvez réaliser le calcul par vous-même.

Tableau 6.3. : Evolution de la SCE en fonction du choix du modèle.

Valeur modèle	SCE
3	51
4	21
5	12
6	21
7	51

En conclusion, nous savons maintenant qu'à l'instant où nous recueillons de l'information, nous pouvons en calculer la moyenne. Cette moyenne nous permet d'établir un modèle qui a l'avantage de nous permettre d'effectuer une prédiction la plus précise possible, c'est-à-dire conduisant à l'erreur minimale d'estimation, à condition d'avoir une distribution symétrique et unimodale.

Plus tard, nous nous intéresserons à l'inférence que je n'ai, jusqu'ici, fait que suggérer. L'enjeu est le suivant (**ne passez certainement pas ce paragraphe sans l'avoir compris**) : avec mes 9 sujets j'ai estimé une moyenne. Cette moyenne est l'estimation la plus précise que j'aie de la moyenne de la population dont sont issus mes 9 sujets (donc de l'ensemble des BA1). **Cependant, si chacun d'entre vous avait également choisi au hasard 9 étudiants de BA1 (et donc probablement pas les mêmes que moi), nombreux d'entre vous auraient obtenu une moyenne différente de la mienne**. Sans doute pas très éloignée, mais néanmoins différente. Vous auriez tous considéré que votre moyenne représentait également la moyenne de la population et il n'y a aucune raison de penser que mon estimation est moins bonne ni meilleure que la vôtre. Dès lors, l'estimation que j'ai obtenue, en faisant la moyenne de mes 9 sujets, me donne une estimation approximative de la moyenne de la population parmi un ensemble d'estimations possibles, mais probablement pas la moyenne exacte (sauf coup de chance). Ce que j'aimerais savoir c'est **l'intervalle dans lequel la moyenne de ma population se trouve presque certainement**. Par exemple, si je trouve une moyenne de 5/10 avec mes 9 sujets, j'estime donc que la moyenne

de la population correspondante doit tourner autour de ce 5/10 mais peut-être pas valoir exactement ce score. Puis-je me dire que la moyenne de la population est comprise entre 4 et 6/10? Ou bien dois-je me dire qu'elle est dans un intervalle compris entre 3 et 7/10? Et si je décide de prendre ce dernier intervalle, quel risque y a-t-il pour que la moyenne soit finalement de 8/10 et que je me sois complètement trompé dans mon estimation? Enfin, très importante question également, si je dois estimer que ma moyenne est comprise entre 3 et 7/10 et que je ne suis pas du tout satisfait de ce niveau de précision, comment puis-je faire pour être plus précis dans mon estimation? Telles sont les questions que nous devrons encore traiter avant la fin de l'année.

6.3. Les mesures de la dispersion ou description de l'erreur

L'intérêt de cette mesure est intrinsèquement lié à la discussion que nous venons d'avoir sur le modèle. Je vous ai déjà sensibilisé à l'importance de la dispersion lorsque nous avons envisagé les boîtes à moustaches et l'écart interquartile, qui constituent une manière de visualiser la dispersion des données autour de la valeur centrale. Je vous avais demandé d'imaginer une distribution de cotes d'examen où la plupart des étudiants obtiennent un 12/20 et les quelques individus qui s'en écartent obtiennent un 11 ou un 13. Dans ce cas, la moyenne de 12 représente une excellente estimation puisque mon erreur serait très faible. Ensuite je vous avais demandé d'imaginer une dispersion beaucoup plus importante des cotes et j'avais conclu qu'alors l'estimation serait nettement moins bonne. De fait, l'erreur associée au modèle serait nettement plus élevée. Nous voyons donc que l'étude de la dispersion est intimement associée à l'évaluation de l'erreur de notre modèle. C'est même une autre manière de dire la même chose puisque l'erreur n'est rien d'autre que la mesure de l'écart entre les valeurs observées et une valeur théorique, donc la dispersion. Il est donc nécessaire de bien la décrire. Pour y parvenir, il existe quatre méthodes : l'écart interquartile que nous avons déjà envisagé ; l'étendue ; l'écart moyen absolu ; et enfin, la variance (et l'écart-type).

6.3.1. L'étendue

C'est la manière la plus rudimentaire de mesurer la dispersion. Elle consiste simplement à mesurer la différence entre la valeur maximale observée et la valeur minimale observée de la distribution. Donc, dans notre exemple de la Figure 6.1, il s'agit de 7 - 3 = 4.

Il devrait maintenant vous sembler évident que cette mesure pose problème dès lors qu'une valeur un petit peu atypique est présente. Un 77 au lieu du 7 conduirait immédiatement à une étendue de 77 - 3 = 74 qui ne représente pas bien du tout la réalité. En outre, cette mesure ne dépend que de deux valeurs, elle ne tient absolument pas compte de toutes les autres. Vous pouvez donc regarder cet indicateur lorsque vous abordez en un coup d'oeil la distribution de vos résultats, mais n'accordez aucun crédit à cet indicateur dès lors que vous entrez dans une analyse plus fine.

6.3.2. Ecart moyen absolu

Une fois que nous avons une estimation ponctuelle de la moyenne, c'est-à-dire de la tendance centrale, un moyen évident d'envisager la dispersion est de mesurer les écarts par rapport à cette valeur centrale. Cela correspond en fait à estimer l'erreur comme nous l'avons fait au point 6.2.2.3. Nous avons donc vu que la somme des écarts par rapport à la moyenne (qui ne sont rien d'autres que les erreurs) est nulle. Nous avions donc résolu ce problème en élevant au carré chacune des déviations (des erreurs), mais nous aurions pu en faire la somme des valeurs absolues (de telle sorte que les valeurs négatives deviennent positives). Cette stratégie correspond à la première étape de l'établissement de l'écart moyen absolu.

Cependant, nous nous trouvons maintenant devant un second problème (que nous rencontrions déjà en calculant la SCE mais dont je n'ai pas encore parlé) : la valeur que nous obtenons est très difficilement interprétable en l'état. Une manière de la rendre plus parlante est de calculer la moyenne de ces écarts (ce qui nous donnerait une estimation de l'écart à la moyenne en moyenne par sujet[22]). C'est de cette manière que nous définirons l'**écart moyen absolu (EMA)**. La formule générale devient donc :

$$EMA = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right|$$

[22] Ne confondez pas les deux termes de "moyenne". L'écart à la moyenne correspond à l'écart par rapport à la moyenne de la distribution concernée. La moyenne des écarts à la moyenne est donc la moyenne des erreurs.

Appliquée aux données du Tableau 6.1 nous obtenons :

$$EMA = \frac{1}{9}\sum_{i=1}^{9}\left|X_i - \overline{X}\right| = \frac{1}{9}(2+0+2+1+0+1+0+1+1) = 0{,}89$$

Cela revient à dire que les sujets s'écartent en moyenne de 0,89 point par rapport à l'estimation moyenne de 5/10. Cette mesure est en fait une excellente façon de se représenter la dispersion. Cependant, elle est totalement supplantée par l'écart-type. Cela provient du fait que l'écart-type est dérivé de la variance qui jouit de propriétés mathématiques qui lui font jouer un rôle central en statistique théorique.

6.3.3. La variance et l'écart-type

6.3.3.1. Etablissement des concepts dans une optique descriptive

Nous avons déjà envisagé de résoudre le problème de la somme nulle des erreurs (des écarts à la moyenne) en élevant chaque écart au carré. Cela nous donnait ce que nous avons appelé la somme des carrés de l'erreur (SCE). Nous avions vu qu'appliquée à la série statistique du Tableau 6.1, la SCE avait une valeur de 12. Mais, de la même manière que pour l'écart moyen absolu, ce chiffre ne nous informe que très peu. Dès lors, nous allons adapter la même stratégie que pour l'EMA et considérer la moyenne de la SCE. Cela nous conduit à la diviser par le nombre de sujets de l'échantillon. Il s'agit, par définition, de la **variance**.

La formule est indiquée plus bas. Mais avant de l'écrire et de l'appliquer à notre exemple, attachons-nous à un dernier problème : la moyenne de la somme des carrés de l'erreur nous donne donc le **carré moyen de l'erreur (CME) par sujet**. C'est-à-dire un chiffre dont l'unité est le carré de l'unité de la distribution. En pratique donc, la CME est synonyme de variance, mais retenez les deux termes, parce que nous aurons besoin de les utiliser séparément plus tard. Dans mon exemple issu du Tableau 6.1, la SCE valait 12. La moyenne pour mes 9 sujets est donc 12/9 = 1,33 par sujet, ce qui constitue la variance de mon échantillon. Cela signifie donc que les sujets s'écartent d'1,33 "points au carré" de la moyenne de 5/10. Cette unité est évidemment difficilement interprétable. Pour résoudre ce problème, on va ramener cet indicateur à la même unité que l'unité de la moyenne, c'est-à-dire les "points". Pour ce faire, il suffit d'extraire la racine carré de la variance : √1,33 = 1,15. Cette

mesure constitue l'**écart-type**. Nous pouvons maintenant dire que les sujets s'écartent en moyenne de 1,15 point autour de la moyenne de 5/10. Remarquez que cette évaluation de la dispersion est un petit peu différente de celle que nous obtenions en utilisant EMA (0,89), mais on reste dans un même ordre de grandeur : l'écart-type est un petit peu plus conservateur, c'est-à-dire qu'il surestime un petit peu l'erreur par rapport à l'EMA. L'encadré ci-dessous résume le concept de variance et d'écart-type sous forme de formules mathématiques. Par convention, nous noterons la variance d'un échantillon S^2 et l'écart-type S. Remarquez qu'il s'agit, dans les deux cas, de paramètres de l'échantillon et que, donc, nous utilisons des lettres latines.

Formules de définition de la variance et de l'écart-type calculés à partir d'une série statistique

Variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Ecart-type

$$S = \sqrt{S^2}$$

avec

$$\sqrt{S^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Nous pouvons aussi exprimer ces formules en nous basant sur l'utilisation des distributions de fréquences, absolues ou relatives. Il suffit pour ça d'adapter les équations de la manière suivante :

Formule de définition de la variance calculée à partir d'une distribution non groupée des fréquences absolues

$$S^2 = \frac{1}{n}\sum_{j=1}^{J} n_j (x_j - \overline{X})^2$$

Formule de définition de la variance calculée à partir d'une distribution non groupée des fréquences relatives

$$S^2 = \sum_{j=1}^{J} f_j (x_j - \overline{X})^2$$

6.3.3.2. Etablissement des concepts dans une optique inférentielle

De la même manière que j'utilise la moyenne de mon échantillon pour inférer la moyenne de la population, il m'est possible d'utiliser l'estimation de l'erreur (ou de la dispersion, vous aurez maintenant compris le lien entre les deux), c'est-à-dire ma variance ou mon écart-type, de mon échantillon pour inférer le paramètre correspondant à la population. Cependant, il existe une différence entre la moyenne et la variance (je ne parlerai plus de l'écart-type, mais considérez que tout ce que je dis à propos de la variance s'applique à l'écart-type jusqu'à la fin de ce chapitre). En effet, je peux utiliser la moyenne de mon échantillon comme estimateur de la moyenne de la population sans qu'aucune correction ne soit nécessaire. On parle dans ce cas d'**estimateur non biaisé**. En revanche, il se trouve, pour des raisons mathématiques que je ne développerai pas ici, que l'utilisation de la variance de l'échantillon comme estimateur de la variance de la population conduit en moyenne à une sous-estimation de cette variance (c'est-à-dire que la vraie variance de la population est souvent un peu supérieure à l'estimation que constitue la variance de l'échantillon). On dit que la variance de l'échantillon est un **estimateur biaisé** de la variance de la population. Pour tenir compte de cette sous-estimation (= de ce biais) il est nécessaire de corriger la variance de l'échantillon pour rendre l'estimateur un petit peu plus grand qu'il ne l'est lorsqu'on envisage la variance non corrigée. La méthode consiste à diviser la SCE non pas par le nombre de sujets, mais bien par le nombre de ce qu'on appelle les **degrés de liberté** et qui est nécessairement inférieur au nombre de sujets (de sorte qu'en

divisant la SCE par un nombre plus petit, on obtient une variance plus grande). Les degrés de liberté tiennent compte du nombre de paramètres estimés que l'on utilise dans une expression. Dans la variance, nous utilisons la moyenne, qui est un paramètre de la distribution. Nous devons donc l'ôter du nombre de sujet et le nombre de degrés de liberté qui nous reste est de n - 1. Ce qui implique que, pour estimer la variance de la population, nous utiliserons la SCE divisée par le nombre de degré de liberté n - 1 et non par n.

Les degrés de liberté

Il n'est pas aisé de développer l'approche théorique des degrés de liberté. En revanche, une approche intuitive peut suffire et est nettement plus accessible. Les degrés de liberté peuvent se conceptualiser comme l'ensemble de valeurs aléatoires qui ne peuvent être déterminées par une équation. Par exemple, une équation de k inconnues, où k est un nombre entier supérieur à 1, est indéterminée. En effet, l'équation x + y = 1 (où k = 2 : x et y) est indéterminée car il est nécessaire de connaître soit x, soit y pour trouver l'inconnue manquante (il y a donc une infinité de solutions). Dans cette équation il y a donc un seul degré de liberté (soit x, soit y). En revanche, une fois une valeur attribuée à l'une des inconnues, l'autre inconnue est nécessairement déterminée. Par exemple, si j'attribue la valeur 6 à y (y = 6), alors x est déterminé et ne peut valoir que -5. En effet, x + 6 = 1 ⇔ x = 1 - 6 ⇔ x = -5. On peut généraliser le concept : pour déterminer complètement k inconnues, il faut k équations.

De la même manière dans une série statistique, il y a n - 1 degrés de liberté dans la combinaison linéaire conduisant à la moyenne. Si j'ai comme information que n = 9 sujets, comme dans le Tableau 6.1, et que ma moyenne est de 5, je me trouve devant une équation indéterminée. Il y a une infinité de valeurs attribuables à mes 9 sujets qui permettent d'obtenir une moyenne de 5. Néanmoins, si je connais la valeur des scores de 8 de mes sujets, alors le neuvième sujet ne peut prendre qu'une seule valeur pour que ma moyenne soit égale à 5. Par exemple, si mes 8 sujets ont un score de 5, alors, le neuvième sujet ne peut avoir que la valeur 5 pour que ma moyenne soit égale à 5. Dès lors, dans cette équation, j'ai 8 degrés de liberté, c'est-à-dire n - 1.

Il reste à expliquer le lien entre les degrés de liberté et leur utilisation comme diviseur de la variance. Là encore, l'intuition sera plus facile à mobiliser que les démonstrations mathématiques. Posons-nous trois questions :

1. Pourquoi faut-il utiliser le nombre de degrés de liberté comme diviseur lorsque l'on estime la variance d'une population?

En fait, il faut distinguer deux situations. La première est celle où l'on ne connaît pas l'ensemble des individus (optique inférentielle). Dans ce cas, nous estimons d'abord la moyenne, puis la variance (dans cet ordre puisqu'il est nécessaire de connaître la moyenne pour calculer la variance). Dans la mesure où la moyenne n'est pas connue mais estimée, elle est associée à une incertitude (rappelez-vous que si nous estimions tous la moyenne d'une population à partir d'un échantillon, nous obtiendrions tous une estimation sans doute proche mais différente). Or nous devons tenir compte de cette incertitude dans l'estimation de la variance et son existence implique que la variance doit être revue un petit peu à la hausse. Donc, dans la mesure où la variance s'estime en tenant compte de la moyenne qui est un paramètre lui aussi estimé, on perd ce degré de liberté. Dès lors, la variance dépend de n - 1 degrés de liberté. La deuxième situation est en rapport avec la deuxième question.

2. Pourquoi ne pas l'utiliser lorsque l'on calcule la variance d'une population dont tous les individus sont connus?

Dans ce cas (optique descriptive), nous n'estimerons pas la moyenne, nous la décrirons. En effet, si nous connaissons l'entièreté des individus, il n'est pas question d'utiliser un estimateur de la moyenne, la moyenne obtenue à l'aide des scores de chaque individu pour la variable concernée nous donne la vraie moyenne de la population. Si chacun d'entre nous la calculait, nous obtiendrions tous exactement la même valeur. Par exemple, si je veux connaître la taille moyenne des étudiants de votre auditoire de 500 personnes (que je définis donc comme la population), et que je demande à chacun de mesurer 50 personnes au hasard, tout le monde aura une moyenne proche mais un peu différente. En revanche, si tout le monde mesure les 500 personnes au complet, nous aurons tous exactement la même moyenne (en considérant que nous ne fassions pas d'erreur de mesure). Si l'on revient à l'exemple de l'équation $x + y = 1$, la première situation revient à trouver y, en connaissant la valeur de x et en ayant estimé la valeur 1. La deuxième situation revient à se rendre compte

que x + y valent 1 parce qu'on connaît x et y. Dès lors, dans ce cas-ci, on n'utilise pas la correction de la variance par le nombre de degrés liberté et on utilise l'ensemble des *N* individus parce que la moyenne utilisée pour calculer la variance n'est pas une estimation, c'est la vraie moyenne de la population.

3. Pourquoi ne pas utiliser les degrés de liberté lorsque l'on estime la moyenne de la population à partir d'un échantillon?

En y réfléchissant, vous devriez être capable maintenant de répondre à cette question par vous-même. En effet, si vous avez compris que ce qui génère la perte d'un degré de liberté est le fait que l'on estime un paramètre, vous vous rendrez compte que l'on n'en estime aucun dans le calcul de la moyenne. L'estimation de la moyenne s'effectue en tenant compte des scores de chaque individu de l'échantillon et ce score est mesuré, mais pas du tout estimé.

En conclusion, chaque fois que nous aurons besoin de calculer les degrés de liberté dans le futur (et nous en aurons encore besoin), vous aurez à vous poser la question de savoir combien de paramètres vous avez estimés dans le problème qui vous occupe. Le nombre de degrés de liberté sera le nombre d'observations ôté du nombre de paramètres estimés.

La plupart des logiciels informatiques, et des manuels de statistiques, donnent la variance corrigée. Par exemple, sur SPSS (le logiciel que vous utiliserez en BA2) et sur Excel, la variance rapportée est toujours calculée à l'aide des degrés de liberté. Cependant, sur Excel il existe la possibilité d'obtenir la variance non-corrigée en utilisant la commande VAR.P au lieu de VAR. De la même manière, on peut trouver l'écart-type non-corrigé (la racine carrée de la variance non-corrigée) en utilisant la commande ECARTTYPEP au lieu d'ECARTYPE. Le P signifie que le logiciel estime que l'on connait l'entièreté de la population et que donc nous n'avons pas besoin d'estimer les paramètres à partir de l'échantillon. En effet, si nous disposons de l'information pour tous les individus d'une population, l'inférence statistique devient inutile : nous connaissons dès lors la moyenne vraie de la population ainsi que sa variance et son écart-type. Dans ce cas, seules les statistiques descriptives ont du sens.

Dans l'éventualité où vous ignorez si la valeur proposée par un logiciel (par exemple une calculatrice) tient compte ou non des degrés de liberté, il vous suffit de demander la

variance de la série statistique {-1, 1}. La variance non-corrigée et l'écart-type correspondant sont nécessairement de 1 alors que la correction conduit à une valeur de respectivement 2 et √2 (1,41). De même, si vous désirez obtenir la variance non-corrigée à partir de la variance corrigée, il vous suffit de la multiplier par $(n - 1)/n$.

Remarquez que la différence entre variance corrigée et non-corrigée s'estompe au fur et à mesure que n est grand. En effet, il y a une grande différence entre diviser par 4 ou par 3, mais une très petite différence entre diviser par 4000 ou par 3999.

Formules de définition de la variance et de l'écart-type corrigés

Variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Ecart-type

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

6.3.3.3. Désavantage de la variance et de l'écart-type

Nous retrouvons également un inconvénient semblable à celui de la moyenne (ce qui n'est pas étonnant puisque l'on calcule la moyenne des écarts à la moyenne) : la sensibilité aux valeurs aberrantes. Cette sensibilité est exacerbée par le fait que l'on élève les erreurs au carré! Pour reprendre la valeur 77 qui remplace le 7, cela conduirait à une variance de 576,89 au lieu de 1,33!! Un simple écart de 2 points pourrait doubler la variance (dans ce cas) : si le 7 n'était remplacé que par un 9, la variance passerait déjà à 2,67.

6.3.3.4. Approfondissement du lien entre la variance et l'estimation de l'erreur : méthode des moindres carrés

Comme nous l'avons vu auparavant, la volonté générale de quiconque modélise la réalité est d'avoir une erreur minimale. Nous avons également vu que la moyenne possédait la caractéristique de minimiser l'erreur. Cette démonstration était purement empirique, j'ai présenté un tableau (Tableau 6.3) qui montrait que toute valeur différente de la moyenne conduisait à une somme des carrés supérieure à la somme des carrés résiduelle obtenue en utilisant la moyenne. Mais on peut se demander pourquoi la moyenne offre cette propriété. Cette démonstration est assez célèbre car elle a découlé sur un conflit entre deux grands mathématiciens ayant vécu à cheval sur la fin du 18ème siècle et début du 19ème : Adrien-Marie Legendre (1752 -1833), ci-à droite, et Johan Carl Friedrich Gauss (1777-1855), ci-à gauche, également célèbre pour sa courbe décrivant la distribution normale, sujet du prochain chapitre. Ces deux personnages semblent être parvenus aux mêmes conclusions de manière indépendante et dans un intervalle assez court : Legendre en 1805, Gauss en 1809 (les bienfaits d'Internet n'étant pas encore connus, ce n'est pas impossible).

La démonstration se trouve actuellement dans de nombreux ouvrages, mais le plus ancien est sans doute celui de Legendre dans son traité intitulé "*Nouvelles méthodes pour la détermination de l'orbite des comètes*". Nous n'allons pas considérer cette démonstration qui se base sur un principe assez simple, mais rendu complexe dans sa version universelle qui s'applique à un nombre important de variables. En revanche, nous allons brièvement nous intéresser au cas particulier de la moyenne qui ne concerne qu'une seule variable. L'idée générale est donc de trouver la valeur qui minimise la somme des carrés de l'erreur. En mathématique, il existe une manière très classique de minimiser une fonction, si vous vous souvenez de vos cours de secondaires et surtout du chapitre sur les études de fonctions, c'est d'annuler la dérivée première de la fonction. Dès lors, nous essayons de faire quelque chose de très simple :

La somme des carrés de l'erreur :

$$SCE = (x_1 - M)^2+(x_2 - M)^2+...+(x_n - M)^2 \text{ où M est la moyenne de la variable X.}$$

Cette SCE doit être minimale. La dérivée selon X de la SCE doit donc être nulle :

$$d(SCE)/dx = [(x_1 - M)^2+(x_2 - M)^2+...+(x_n - M)^2]' = 0$$

Ce qui donne, en utilisant la propriété de la dérivée d'une variable avec exposant $(x^a)' = ax^{a-1}$:

$$2(x_1 - M)+2(x_2 - M)+...+2(x_n - M) = 0 \Leftrightarrow$$

$$2x_1+2x_2+...+2x_n = 2nM \Leftrightarrow$$

$$M = 2(x_1+x_2+...+x_n)/2n \Leftrightarrow$$

$$\mathbf{M = (x_1+x_2+...+x_n)/n}$$

Cette dernière formule est celle de la moyenne puisque $(x_1+x_2+...+x_n)$ n'est rien d'autre que Σx_i.

En conclusion : cette démonstration vous prouve que la moyenne est bien la valeur qui minimise la SCE puisqu'elle représente le résultat de la dérivée de la SCE lorsqu'elle est nulle. Néanmoins, la SCE n'est pas nulle, elle n'est que minimale. Dès lors, on en revient à trouver l'erreur minimale résiduelle par sujet en divisant la SCE par le nombre de sujets, ce qui nous donne la variance de l'échantillon.

Par ailleurs, si on veut **estimer** cette variance pour la population, il nous reste à remplacer le nombre de sujets par le nombre de degrés de liberté et diviser la SCE par *n*-1, conformément à notre argumentation précédente. Vous comprendrez maintenant que la variance est bien une mesure de dispersion et qu'elle est liée à l'erreur, puisque l'erreur n'est rien d'autre que l'écart par rapport à la moyenne, c'est-à-dire le degré auquel les valeurs se dispersent autour de la valeur moyenne. Enfin, l'écart-type ramène l'information de la SCE à la bonne unité, puisque la somme des **carrés** de l'erreur est précisément au carré, qui se trouve être une échelle moins confortable pour nos esprits. A ce stade je m'attends à ce que vous ressentiez une passion infinie pour les mathématiques et les statistiques.

6.4. Détermination algébrique de la symétrie et de l'aplatissement d'une distribution

6.4.1. Introduction

Nous avons déjà envisagé les différentes formes possibles des distributions au point 5.5.2. Parmi les formes possibles, nous avons isolé plus particulièrement les distributions asymétriques (positives ou négatives), les distributions multimodales ou les distributions symétriques. Jusqu'ici notre approche était essentiellement graphique. Maintenant nous avons les outils nécessaires pour traiter algébriquement le problème de la forme d'une distribution puisque nous pouvons caractériser la tendance centrale et la dispersion d'une courbe lorsqu'il s'agit d'une distribution symétrique et unimodale.

Approchons d'abord une distribution totalement symétrique de manière intuitive. Si vous reprenez sous les yeux la Figure 5.2.d vous verrez maintenant que la moyenne se trouve exactement au centre, au plus haut point du pic puisque la dispersion des valeurs est exactement la même des deux côtés de ce pic. C'est également à cette valeur que vous trouverez le mode et la médiane puisque le sommet du pic est la valeur la plus probable (par définition) et que la symétrie implique qu'il y a exactement le même nombre de sujets à sa gauche qu'à sa droite (c'est donc bien la médiane). Dès lors, **pour une distribution parfaitement symétrique, moyenne, mode et médiane se confondent**.

Si vous portez maintenant attention aux distributions asymétriques (Figure 5.2.a et b) vous constaterez que le mode (sommet du pic) est différent de la moyenne. **Le mode est plus petit que la moyenne en cas d'asymétrie positive et plus grand en cas d'asymétrie négative**. Cette situation n'est pas anormale puisque le mode ne dépend que d'une seule valeur (la plus probable, donc le sommet du pic) alors que la moyenne est sensible aux valeurs extrêmes et donc tirée vers la queue de la courbe.

Enfin, si l'on s'attache à la distribution bimodale (Figure 5.2.c), on remarque que, même symétrique, les modes ne correspondent (forcément) pas à la moyenne. Les raisonnements qui suivent peuvent s'appliquer à n'importe quelle distribution multimodale. Dans ce cas, la difficulté résultante est évidente : pour décrire correctement les valeurs les plus fréquentes de la courbe, il est nécessaire de prendre plusieurs valeurs en compte (autant qu'il y a de modes, ici, deux). Cette situation génère une chaîne de conséquences : (a) la moyenne ne signifie plus grand-chose ; (b) et donc, pour évaluer la dispersion des données, la situation se

complexifie, puisqu'on doit le faire par rapport à deux valeurs (ou plus) au lieu d'une. La variance n'est donc plus utilisable, ni l'EMA. Seule l'étendue garde une information intacte avec les limites dont on a parlées. En fait, il existe des alternatives que nous verrons au chapitre 10.

Cette discussion devrait à nouveau vous convaincre, cette fois à l'aide de l'information sur la manière dont on évalue la tendance centrale et la dispersion, qu'idéalement, on a tout intérêt à avoir une distribution la plus unimodale et la plus symétrique possible. Ce n'est que dans ce cas que nous pouvons utiliser sereinement les **paramètres** habituels qui décrivent la distribution c'est-à-dire la moyenne et la variance. C'est donc dans ce cas uniquement que nous ferons ce que l'on appelle des **statistiques paramétriques**. Dans les cas où la moyenne et la variance ne caractérisent pas correctement la distribution, on est contraint de faire appel à des méthodes d'analyses différentes dites **non-paramétriques**.

Accordons maintenant notre attention à la description de ce qu'on appelle les moments d'une distribution. Cette approche nous permettra de comprendre comment créer des indices algébriques de mesure de la symétrie et de l'aplatissement d'une courbe.

6.4.2. Notions de moments d'une distribution

En mathématique, il existe un concept intéressant, que nous n'allons pas démontrer mais dont nous allons voir les implications : le moment. Un moment est une quantité calculée par la moyenne des valeurs de la distribution (à une constante près) élevées à un certain degré (appelé l'ordre). Lorsque je dis "à une constante près", j'entends qu'à chaque valeur on peut ôter ou rajouter une constante avant de l'élever à un degré quelconque. Pour l'instant, cette valeur doit vous paraître un petit peu obscure, mais vous y verrez plus clair lorsque nous envisagerons les applications.

Moments d'ordre k

$$A_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^k$$

Moments d'ordre k centré par rapport à l'origine (a = 0)

$$A_k = \frac{1}{n}\sum_{i=1}^{n}X_i^k$$

En considérant ces formules, vous pouvez retrouver les notions de moyenne et de variance que nous venons de voir. En effet, la moyenne n'est rien d'autre que le moment centré sur l'origine (sur zéro) d'ordre 1 (k = 1). De même, la variance n'est rien d'autre que le moment centré sur la moyenne (a = la moyenne) d'ordre 2 (k = 2). Vous ne devriez pas rencontrer de problèmes en effectuant les remplacements dans les formules par vous-même pour vous rendre compte de la situation.

De manière plus générale, on distinguera les propriétés des moments d'ordres pairs (k = 2, 4, 6, …) et les moments d'ordres impairs de degrés supérieurs à 1 (k = 3, 5, 7, …). La raison qui nous conduit à devoir distinguer les deux est qu'un exposant pair implique que les termes négatifs deviennent positifs une fois élevés à l'ordre concerné. En revanche, un exposant impair ne change jamais le signe de la valeur exposée.

Les moments centrés d'ordres pairs donnent des informations sur la dispersion des données, alors que les moments centrés d'ordres impairs supérieurs à 1 donnent des informations sur l'asymétrie. Cependant, en pratique, les moments d'ordres élevés ne sont pas envisagés. Seuls les moments d'ordre 1 (associé à la moyenne), 2 (associé à la variance), 3 (associé à l'asymétrie, voir point 6.4.3) et 4 (associé à l'aplatissement, voir point 6.4.4) seront utilisés.

6.4.3. Indice d'asymétrie (Skewness) : coefficient G1 de Fisher

Nous avons déjà dit que la différence entre la moyenne et le mode pouvait constituer une manière de mettre l'asymétrie d'une distribution en évidence. C'est d'ailleurs ce que certains mathématiciens ont proposé comme indicateur d'asymétrie. Par exemple, Pearson proposait

de mesurer cet écart par la formule suivante : (moyenne - mode) / écart-type. Cet indicateur est très rarement utilisé.

Cependant, Fisher a établi une mesure, plus facile et plus immédiate à interpréter, basée sur le moment d'ordre 3. Cette mesure s'est imposée comme un standard dans l'analyse descriptive des distributions.

Définition du coefficient G_1 de Fisher

$$G_1 = \frac{M_3}{S^3}$$

avec $M_3 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^3$ et $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Dès lors :

- Lorsque la distribution est parfaitement symétrique, G_1 est égal à zéro.
- Plus G_1 est positif plus la distribution est en asymétrie positive (donc que la queue de la distribution s'étend vers la droite).
- Plus G_1 est négatif plus la distribution est en asymétrie négative (donc que la queue de la distribution s'étend vers la gauche).

Vous pouvez vérifier que l'asymétrie de la distribution correspondant au Tableau 6.1 est de $G_1 = 0$. Ce résultat est tout à fait logique puisque cette distribution est parfaitement symétrique.

6.4.4. Indice d'aplatissement (Kurtosis) : Coefficient G2 de Fisher

Nous avons vu que les moments d'ordres pairs sont en rapport avec la dispersion des valeurs, dû au fait que les termes négatifs deviennent positifs lorsqu'ils sont exposés par un nombre pair. Cependant, alors que le moment d'ordre 2 nous informe sur la dispersion des valeurs par rapport à la moyenne, le moment d'ordre 4 peut nous informer sur l'aplatissement de la distribution. Le coefficient d'aplatissement se définit comme le moment d'ordre 4 divisé par la variance au carré (ou l'écart-type exposant quatre). Cette

mesure est nommée le coefficient d'aplatissement de Pearson. Cependant, elle n'est pas entièrement satisfaisante. En effet, voilà plusieurs fois que je vous parle de notre fameuse distribution normale, Graal sacré des distributions. Or, pour que le coefficient de Pearson corresponde à la valeur zéro lorsque l'aplatissement est compatible avec une telle distribution, il est nécessaire de l'ajuster en lui retirant la valeur 3. C'est ce que Fisher propose, et qui donne le coefficient G_2 de Fisher.

Définition du coefficient G_2 de Fisher

$$G_2 = \frac{M_4}{S^4} - 3$$

avec $M_4 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^4$ et $S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$

Dès lors, on peut voir que :

- G_2 vaut o quand la distribution n'est ni plus pointue ni plus aplatie qu'une distribution normale de même moyenne et de même écart-type.
- G_2 prend des valeurs positives d'autant plus grandes que la classe modale de la distribution s'étire plus fort vers le haut. Ceci veut dire que la distribution devient plus pointue que la distribution normale correspondante. En conséquence, par rapport à une distribution normale, il y a une plus grande fréquence de valeurs de la variable proches de la moyenne et, corrélativement, une plus faible fréquence de valeurs de la variable éloignées de la moyenne.
- G_2 prend des valeurs négatives d'autant plus grandes que le centre de la distribution s'enfonce plus fort, élargissant ainsi la classe modale jusqu'à des valeurs éloignées de la moyenne. Ceci veut dire que la distribution est moins pointue que la distribution normale correspondante. En conséquence, par rapport à une distribution normale, il y a une plus grande fréquence de valeurs éloignées de la moyenne et, corrélativement, une plus faible fréquence de valeurs proches de la moyenne (une telle distribution est qualifiée de **platykurtique**).

Je vous enjoins à réaliser le calcul de G_2 sur base de la distribution du Tableau 6.1. Vous devriez obtenir $G_2 = -0{,}29$. La distribution est presqu'équivalente à une distribution normale, mais est un tout petit peu plus plate que l'idéal.

6.5. Synthèse

Nous avons vu les principaux moyens de caractériser algébriquement une distribution. Trois grandes classes de mesures ont été envisagées : la tendance centrale, la dispersion et la forme des distributions (aplatissement et asymétrie). En résumé, la moyenne et la variance sont les deux paramètres les plus utilisés et les plus importants pour caractériser une distribution. Cependant, ils n'ont de sens que si la distribution a une forme adéquate pour permettre leur utilisation, c'est-à-dire unimodale et symétrique. Lorsqu'une distribution est caractérisable par ces deux paramètres uniquement, on dit qu'elle est normale. Deux mesures supplémentaires permettent de décrire correctement une distribution : le coefficient G_1 de Fisher qui mesure l'asymétrie et le coefficient G_2 de Fisher qui mesure l'aplatissement. Plus les valeurs de ces coefficients s'éloignent de zéro, plus l'écart entre la distribution étudiée et la distribution normale idéale est important.

Tout ce que nous avons vu depuis le chapitre 5 devrait vous sensibiliser à l'importance de vous approprier vos données par une approche descriptive avant d'envisager tout traitement plus complexe de l'information récoltée. Mon conseil est donc, dans un premier temps, de représenter graphiquement vos distributions. Cela vous permettra d'avoir une idée visuelle des caractéristiques de votre distribution ainsi que de débusquer d'éventuelles erreurs d'encodage. Ensuite, vous pouvez (à mon sens, vous devez) calculer les principaux indicateurs algébriques qui vous informerons plus précisément sur la distribution de vos données. A ce stade, nous en sommes déjà à l'établissement d'un modèle sommaire vous permettant de simplifier la réalité en associant à votre distribution une mesure de la tendance centrale (le plus souvent la moyenne) qui vous servira de modèle prédictif. De même, vous serez capables d'évaluer l'importance de l'erreur résiduelle en mesurant les écarts au carré par rapport à votre modèle (la SCE). Avant d'envisager des modèles plus complexes vous permettant de diminuer plus encore votre erreur, nous allons envisager deux distributions importantes : la distribution binomiale et la, maintenant fameuse, distribution normale.

6.6. Exercices

T.P. 5 - 6 : CHAPITRE 6

Partie 1 : Mesures de la tendance centrale

Exercice 1 : Le Mode

Voici un tableau sous forme de distribution de fréquences qui correspond au nombre de fois qu'un échantillon d'adolescents a vu un film donné :

x_j	n_j	f_j	Fréquences relatives (en %)
1	3	0,061	6,1
2	5	0,102	10,2
3	9	0,184	18,4
4	8	0,163	16,3
5	2	0,041	4,1
6	7	0,143	14,3
7	9	0,184	18,4
8	6	0,122	12,2
Total :	49	1	100

1. Que pouvez-vous dire de la forme de la distribution ? Quel est le mode ?

2. Dans quel pourcentage de cas aurai-je raison si j'affirme que les adolescents de ce groupe ont vu ce film : a) 3 fois ? b) 6 fois ?

3. Dans quel pourcentage de cas aurai-je raison si j'affirme que les adolescents de ce groupe ont vu ce film : a) - de 3 fois ? b) - de 6 fois ? c) au - 4 fois ? d) au - 7 fois ?

T.P. 5 - 6 : Partie 1

Exercice 2 : La Moyenne

A. Voici une série statistique de 5 notes sur 10 obtenues à l'examen d'histoire par les 5 élèves d'une classe.

i	X_i Notes histoire (/10)
1	$X_1=5$
2	$X_2=8$
3	$X_3=8$
4	$X_4=6$
5	$X_5=9$

1. Calculez la moyenne de cette série statistique. Notez explicitement le calcul que vous avez effectué.

2. Recalculez la moyenne de la série statistique ci-dessus en utilisant la formule adéquate vue au cours, en détaillant à l'aide des notations mathématiques le processus de calcul.

Nous allons maintenant voir ce qu'il se passe au niveau de la moyenne si on modifie une valeur de la variable (ou qu'on ajoute ou enlève une donnée), comme par exemple dans le tableau suivant :

Num i	Notes histoire / 10 X_i	Notes histoire modifiées X_i'
1	5	**8**
2	8	8
3	8	8
4	6	6
5	9	9

3. Calculez la moyenne de la nouvelle variable « notes histoire modifiées». Que se passe-t-il ?

4. Remplacez maintenant le premier 8 par un 1 (c'est-à-dire une valeur plus petite que le 5 initial). Calculez la nouvelle moyenne. Que constatez-vous ?

5. Ajoutez à la nouvelle série statistique une valeur égale à sa nouvelle moyenne, c'est-à-dire un sujet qui a eu une note égale à la moyenne. Calculez la nouvelle moyenne. Que se passe-t-il ?

B. Voici les notes sur 20 obtenues par un groupe de 20 élèves, respectivement à une interrogation d'histoire et de français.

(a) ***Série[23]***

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
$X_{histoire}$	7	9	11	13	16	16	19	17	16	9	12	17	15	13	18	12	18	15	11	12
$Y_{français}$	12	17	10	15	12	14	17	18	18	8	11	14	13	15	16	1	14	15	15	13

1. Quelle est la valeur de X_5 et de Y_5 ?

(b) ***Distribution de fréquences correspondante***

Valeurs observées		**Fréquences absolues (n_j)**		**Fréquences relatives (f_j)**	
Histoire (x_j)	**Français (y_j)**	**histoire**	**français**	**histoire**	**Français**
7	1	1	1	0,05	0,05
9	8	2	1	0,1	0,05
11	10	2	1	0,1	0,05
12	11	3	1	0,15	0,05
13	12	2	2	0,1	0,1
15	13	2	2	0,1	0,1
16	14	3	3	0,15	0,15
17	15	2	4	0,1	0,2
18	16	2	1	0,1	0,05

[23] N.B. : Notons que la variable a été présentée en ligne pour un souci d'économie de place. Prenez cependant l'habitude de représenter les variables en colonnes comme nous l'avons vu à la première séance de travaux pratiques. En effet, c'est la présentation qui est exigée par la plupart des programmes informatiques, notamment SPSS que vous découvrirez l'année prochaine.

19	17	1	2	0,05	0,1
	18		2		0,1

2. Calculez la moyenne de la classe au **cours d'histoire** de 3 manières différentes. Et, donnez les formules associées à vos calculs.

3. Quelle est la moyenne de la classe pour le cours **de français** ?

4. La moyenne de classe pour le cours de **français** est particulièrement influencée par la note obtenue par l'un des élèves. De quel élève s'agit-il ? Justifiez votre réponse et expliquez l'inconvénient que cela entraîne.

5. Le calcul de la moyenne est-il une bonne manière de représenter une distribution multimodale ?

T.P. 5 - 6 : Partie 1

Exercice 3 : Sensibilité aux valeurs extrêmes

1. Un psychologue passioné d'éthologie désire examiner le nombre d'erreurs que font des rats avant de connaître parfaitement le labyrinthe dans lequel ils sont placés. L'échantillon compte 10 rats (n =10). Le nombre d'erreurs des 9 premiers rats est le suivant : 6, 2, 6, 4, 5, 4, 7, 6, 5. Calculez la moyenne et le mode de ces données.

2. Le dixième rat a commis 100 erreurs avant de connaître parfaitement le labyrinthe. Quand ce rat est inclus dans l'échantillon, que se passe-t-il au niveau de la moyenne et du mode ? Quelles conclusions générales peut-on conclure par rapport aux informations que fournissent ces 2 indices de tendance centrale ?

3. Un onzième rat est ajouté. Il a commis 5 erreurs avant de connaître parfaitement le labyrinthe. Que deviennent le mode et la moyenne ?

4. Quels sont les avantages et les désavantages des différents indicateurs algébriques de tendance centrale ? Complétez le tableau ci-dessous par oui ou non.

	Sensible aux valeurs aberrantes ?	**Proche du regroupement le plus important des valeurs ?**
Mode		
Moyenne		
Médiane		

T.P. 5 - 6 : Partie 1

Exercice 4 : Sommes simples

Le tableau ci-dessous reprend les valeurs correspondant aux résultats de 5 sujets à un test de math
noté sur 10 (X) et à un test d'évaluation du stress (Y) où les sujets doivent évaluer leur niveau de stress sur une échelle de 1 à 10. Complétez le tableau (sauf la dernière ligne).

i	X	Y	X²	Y²	X-Y	XY
1	$X_1 = 3$	$Y_1 = 9$				
2	$X_2 = 8$	$Y_2 = 3$				
3	$X_3 = 4$	$Y_3 = 5$				
4	$X_4 = 5$	$Y_4 = 2$				
5	$X_5 = 5$	$Y_5 = 1$				
$\sum$	$\sum_{i=1}^{N} X_i =$	$\sum_{i=1}^{N} Y_i$	$\sum_{i=1}^{N} X_i^2$	$\sum_{i=1}^{N} Y_i^2$	$\sum_{i=1}^{N} (X_i - Y_i)$	$\sum_{i=1}^{N} X_i Y_i$

1. Calculez les sommes suivantes et placez les formules et les réponses dans le tableau ci-dessus.

$\sum_{i=1}^{5} X_i$	$\sum_{i=1}^{5} Y_i$
$\sum_{i=1}^{5} X_i^2$	$\left(\sum_{i=1}^{5} X_i\right)^2$
$\sum_{i=1}^{5} (X_i - Y_i)$	
$\sum_{i=1}^{5} X_i - \sum_{i=1}^{5} Y_i$	
$\sum_{i=1}^{5} X_i Y_i$	
$\sum_{i=1}^{5} X_i - \sum_{i=1}^{5} Y_i$	

2. Calculez les moyennes suivantes et écrivez la réponse dans la colonne prévue pour les résultats (colonne RES). Exprimez vos calculs en utilisant le signe $\sum$ (colonne sigma) et sous une forme algébrique.

		RES.	FORME SIGMA	FORME DEVELOPPEE
1	Au test de math. pour tous les sujets			
2	Au test de stress, pour tous les sujets			
3	Au test de math, pour les sujets 1 à 3			
4	Au test de stress, pour les sujets 1 à 3			

5	Au test de math, pour les sujets 2 à 4			
6	Au test de stress, pour les quatre premiers sujets			

3. Que pensez-vous du fait d'avoir calculé des moyennes pour ces variables ? Pensez les choses en termes d'échelles de mesure.

T.P. 5 - 6 : Partie 1

Exercice 5 : La Modélisation

1. Modélisez une situation sur base du mode, de la médiane et de la moyenne ?

2. Donnez toutes les autres valeurs qui peuvent être utilisées pour modéliser la réalité ?

3. Quel est le risque que vous courrez en faisant cela ?

4. Donnez 6 manières différentes d'écrire la formule pour calculer la moyenne (en fonction du mode de présentation des données ou du mode de présentation de la formule elle-même).

T.P. 5 - 6 : Partie 1

Exercice 6 : Sommes simples

1. Développez les sommes suivantes:

	FORME SIGMA	FORME DEVELOPPEE
1	$\sum_{i=0}^{5} X_i =$	

2	$\sum_{i=1}^{4}(Y_i - 3)^2 =$	
3	$\sum_{i=1}^{4}\left(X_i - \frac{1}{4}\sum_{i=1}^{4}X_i\right) =$	

2. Soient les variables X et Y suivantes

i	X_i	Y_i
1	3	3
2	4	5
3	1	2
4	10	6
5	7	4

n = 5 et a = 3

		Développement	Résultat
1.	$\sum_{i=1}^{5}X_i =$		
2.	$\sum_{i=3}^{4}X_i =$		
3.	$\frac{1}{3}\sum_{i=1}^{3}X_i =$		
4.	$\frac{1}{3}\sum_{i=2}^{4}X_i =$		
5.	$\frac{1}{3}\sum_{i=1}^{3}(X_i - 3) =$		

6.	$\frac{1}{3}\sum_{i=1}^{3}(X_i - 2)^2 =$		
7.	$\sum_{i=1}^{n} \mathbf{X_i} =$		
8.	$\sum_{i=1}^{n} \mathbf{a^2} =$		
9.	$\sum_{i=1}^{4} \frac{X_i - 3}{4} =$		
10.	$\sum_{i=1}^{n} \mathbf{(X_i - a)^2} =$		

3. Écrivez les expressions suivantes sous une forme sigma :

1.	$X_4^2 + X_5^2 + X_6^2 + X_7^2 + X_8^2 + X_9^2 =$	
2.	$X_0 + X_1 + X_2 + Y_0 + Y_1 + Y_2 =$	

T.P. 5 - 6 : CHAPITRE 6

Partie 2 : Mesures de la dispersion

Exercice 1 : Le modèle de la moyenne - Mesure de l'erreur

1. Voici deux séries statistiques rangées par ordre croissant, avec les X_i représentant les points obtenus par des étudiants à un examen noté sur 20.

Série A :

i	X_i
1	0
2	2
3	10
4	14
5	20
6	20

Série B :

i	X_i
1	10
2	10
3	10
4	10
5	12
6	14

a. Quel est le mode de ces deux séries ?

b. Calculez la moyenne de chacune de ces séries. Indiquez la formule utilisée.

c. Quelle est la place (le rang) de l'étudiant qui a 14/20 dans le classement par ordre décroissant dans chacune des séries ?

d. Que vaut l'étendue ? Indiquez la formule utilisée pour les données rangées par ordre croissant.

e. Quelle que soit la variable, dans quelle unité est donnée l'étendue ?

f. Admettons que l'on ait une série de cent sujets dont le moins bon a obtenu 1/20 et le meilleur 19/20 mais que les 98 autres aient des résultats variant entre 8 et 14/20. Pensez-vous que l'étendue vous donne une bonne idée de la dispersion des résultats ?

Un autre moyen d'envisager la dispersion est de calculer les écarts de chaque note par rapport à un modèle, la moyenne de la série et de prendre la moyenne de ces écarts.

g. Comment, dans le cours, sont appelés les écarts de chaque valeur de la variable par rapport à sa moyenne ? Quelle est la notation utilisée ?

h. Calculez les erreurs (= écarts à la moyenne) pour chacune des données de la série A et notez-les dans la colonne appropriée, puis calculez la somme des erreurs et finalement, la moyenne de ces erreurs. Indiquez la formule utilisée.

i. Faites la même chose pour la série B. Pas de corrigé ici, puisque vous pouvez vous auto-corriger. Grâce à quel indice ?

Ce résultat sera donc toujours le même quelle que soit la série, parce que les différences positives sont annulées par les différences négatives. Dès lors, on ne peut pas prendre conscience des différences qui existent réellement par rapport à la moyenne (c'est-à-dire au modèle). Un moyen de contourner ce problème est d'utiliser les valeurs absolues mais il est très peu utilisé. On préfèrera élever toutes les valeurs des erreurs au carré. Ainsi, tous les termes deviendront positifs. La somme de ces valeurs élevés au carré se nomme « **Somme des Carrés de l'Erreur** » (SCE). Nous prendrons ensuite la moyenne des carrés de l'erreur, appelée « **variance** » (S_X^2).

j. Calculez l'écart moyen absolu pour la série A.

k. Calculez les carrés de l'erreur pour les deux séries, puis la somme de ces erreur et enfin la $\overline{X}$ de ces erreurs. Indiquez la formule utilisée.

l. Dans quelle série cette valeur est-elle la plus basse ? Cela est-il en accord avec la dispersion que vous constatez en comparant les séries ? Que pensez-vous de ces chiffres ?

Les unités de la variance sont élevées au carré. Si on avait une moyenne en centimètres, on aurait des cm au carré, si on avait des kg, on aurait des kg au carré, si on avait des degrés, on aurait des degrés au carré…
Donc, pour revenir à des valeurs du même ordre d'unité que la moyenne de départ, on peut prendre la racine carrée de la moyenne des carrés des écarts à la moyenne, c'est-à-dire la racine carrée de la variance, et on obtient ainsi ce que l'on appelle l'écart-type (S_X).

m. Calculez l'écart type pour les deux séries. Notez-en la formule et commentez les résultats.

Conclusion : La dispersion d'une série statistique se calcule en prenant la moyenne des carrés des écarts par rapport à la moyenne. Plutôt que de donner ce terme à rallonge, on l'appellera VARIANCE, notée S^2 (calculée au point k). Pour retrouver des unités semblables à celles des données de base, et les rendre interprétables, on prend la racine carrée de cette variance ; c'est ce qu'on dénommera l'ÉCART TYPE (calculé au point m), noté « S », plutôt que de l'appeler « racine carrée de la variance». Ces deux valeurs sont les indices de dispersion les plus utilisés en statistique.

Vous avez probablement pu réaliser les différents calculs sans trop de problème. Cependant il n'y avait que 6 sujets. Certaines études statistiques en comptent plusieurs centaines voire plusieurs milliers. Les calculs sont exactement identiques à ceux que vous avez réalisés tant pour la moyenne que pour les indices de dispersion. Cependant, il existe une notation simple qui permet d'écrire en quelques signes un calcul portant sur des milliers de sujets. Vous devez IMPERATIVEMENT être parfaitement familiarisés avec cette notation et en connaître les propriétés.

n. Voici six formules. À quoi correspondent-t-elles ? En quoi les trois premières sont-elles différentes des trois suivantes ?

Moyenne :	Variance :	Écart type :
$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$	$S_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sqrt{S_X^2}$

Moyenne :	Variance :	Écart type :
$\overline{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j$	$S_X^2 = \frac{1}{n}\sum_{j=1}^{J} n_j (x_j - \overline{X})^2$	$S_X = \sqrt{\frac{1}{n}\sum_{j=1}^{J} n_j (x_j - \overline{X})^2} = \sqrt{S_X^2}$

T.P. 5 - 6 : Partie 2

Exercice 2 : Le modèle de la moyenne

1. Voici une série statistique : 1, 2, 3, 4, 5, 6 ,7, 8, 9, 10. Calculez-en la moyenne.

2. Prenez la valeur 10 pour modéliser la réalité. Que pouvez-vous dire de la moyenne de l'erreur ?

3. Prenez la moyenne pour modéliser la réalité. Que pouvez-vous dire de la moyenne de l'erreur ?

4. Donnez pour cette série, le mode, la médiane, l'étendue, l'écart moyen absolu, la variance et l'écart type, la variance et l'écart type estimés de la population. Indiquez la formule utilisée et le détail de votre calcul.

5. a+b+c+d+e=15. En admettant qu'il s'agisse d'une série statistique ayant pour somme 15, combien vaut n ? Que signifie n ? Combien cette équation contient-elle de degrés de liberté ? Expliquez votre réponse.

T.P. 5 - 6 : Partie 2

Exercice 3 : Moyenne et variance - Effet de l'ajout et de la suppression de données

Attention : Répondez aux différentes questions posées sans utiliser votre calculatrice. Faites plutôt appel à votre bon sens.

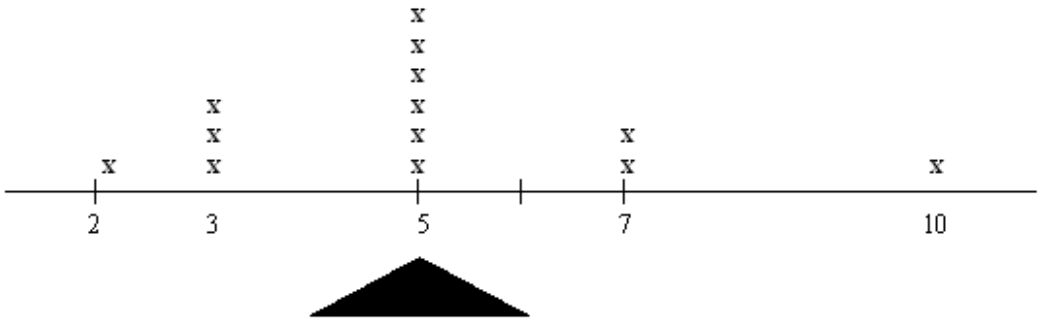
Les auteurs d'un manuel de statistique ont représenté les situations suivantes, comme si elles étaient placées sur une balance. La moyenne se trouve au point d'équilibre de la balance et en constitue donc le centre de gravité. C'est d'ailleurs pour cela que la somme des écarts par rapport à la moyenne est toujours nulle.

N.B. : Nous considérerons dans cet exercice que le poids du plateau de la balance est nul.

SITUATION INITIALE

Scores : 2; 3; 3; 3; 5; 5; 5; 5; 5; 5; 7; 7; 10

Moyenne des scores : 5



SITUATION # 1

On considère la situation nouvelle #1 obtenue en supprimant le score 10.

1. Représentez la balance et son inclinaison après cette modification.

2. Indiquez par une croix l'effet de cette modification sur la moyenne.

La moyenne se déplace vers la gauche (elle diminue)	
La moyenne reste inchangée	
La moyenne se déplace vers la droite (elle augmente)	

3. Indiquez par une croix l'effet de cette modification sur la variance.

La variance des scores augmente	
La variance des scores ne change pas	
La variance des scores diminue	

4. Où faudrait-il mettre la base de la balance pour rétablir son équilibre ? Faites les calculs nécessaires pour la rééquilibrer.

SITUATION # 2

Partant de la situation initiale, on considère maintenant une situation nouvelle #2 obtenue en supprimant le score 2.

5. Représentez la balance et son inclinaison après cette modification.

6. Indiquez par une croix l'effet de cette modification sur la moyenne.

La moyenne se déplace vers la gauche (elle diminue)	
La moyenne reste inchangée	
La moyenne se déplace vers la droite (elle augmente)	

7. Indiquez par une croix l'effet de cette modification sur la variance.

La variance des scores augmente	
La variance des scores ne change pas	
La variance des scores diminue	

8. Où faudrait-il mettre la base de la balance pour rétablir son équilibre ? Faites les calculs nécessaires pour la rééquilibrer.

SITUATION # 3

On considère enfin une troisième modification de la situation initiale obtenue en supprimant cinq des six scores 5. On a donc les figures suivantes.

Scores : 2; 3; 3; 3; 5; 7; 7; 10

9. Représentez la balance et son inclinaison après cette modification.

10. Indiquez par une croix l'effet de cette modification sur la moyenne.

La moyenne se déplace vers la gauche (elle diminue)	
La moyenne reste inchangée	
La moyenne se déplace vers la droite (elle augmente)	

11. Indiquez par une croix l'effet de cette modification sur la variance.

La variance des scores augmente	
La variance des scores ne change pas	
La variance des scores diminue	

12. Le cas échéant, où faudrait-il mettre la base de la balance pour rétablir son équilibre ? Faites les calculs nécessaires pour la rééquilibrer.

T.P. 5 - 6 : Partie 2

Exercice 4 : Sommes simples

Introduction au calcul de la variance

Considérons les couples de valeurs ci-après :

$$X_1 = 3, \quad Y_1 = 9$$

$$X_2 = 8, \quad Y_2 = 3$$

$$X_3 = 4, \quad Y_3 = 5$$

$$X_4 = 5, \quad Y_4 = 2$$

$$X_5 = 5, \quad Y_5 = 1$$

Evaluez les sommes suivantes :

	FORME SIGMA	FORME DEVELOPPEE	RES.	Que vient-on de calculer ?
1	$\frac{1}{5}\sum_{i=1}^{5}(X_i - 5)^2 =$			
2	$\frac{1}{5}\sum_{i=1}^{5}(Y_i - 4)^2 =$			

T.P. 5 - 6 : Partie 2

Exercice 5 : Variance et Ecart-type

Voici l'âge de 10 étudiants en psycho :

	X_i	$X_i - \overline{X}$	$\left(X_i - \overline{X}\right)^2$	$\left(X_i - \overline{X}\right)^3$	$\left(X_i - \overline{X}\right)^4$
1	19				
2	19				
3	22				
4	20				
5	20				
6	19				
7	19				
8	20				
9	21				
10	19				
$\sum_{i=1}^{n}$					

1. Complétez la colonne $X_i - \overline{X}$. N'arrondissez pas.

2. A quoi correspond la variable que vous venez de calculer ? Donnez-en la formule.

3. Calculez la moyenne de la variable $X_i - \overline{X}$. Que remarquez-vous ?

4. Complétez les troisième, quatrième et cinquième colonnes du tableau des données. A quoi correspondent ces colonnes ?

5. Calculez la moyenne, la variance et l'écart-type de X sur base des sommes calculées sur la dernière ligne du tableau. Arrondissez la réponse finale à deux décimales.

6. Calculez l'écart-type de la variable $X_i - X$. Que remarquez-vous ?

T.P. 5 - 6 : Partie 2

Exercice 6 : Vrai – Faux

1. Pour chacune des propositions suivantes, mettez une croix dans la case correspondant à la bonne réponse.

	Proposition	Toujours vraie	Parfois vraie	Jamais vraie
a	La moyenne et l'écart-type sont mesurés dans les mêmes unités.			
b	La moyenne d'une distribution est égale à une valeur observée de cette distribution.			
c	La moyenne d'une distribution est comprise entre la plus grande et la plus petite valeur observée.			
d	La médiane correspond à une valeur observée			

2. Les différents indicateurs algébriques de dispersion sont-ils sensibles aux valeurs aberrantes ? Complétez le tableau ci-dessous.

	Sensible aux valeurs aberrantes	
	OUI	**NON**
Etendue		
Ecart-type		
Variance		

3. Pour que la variance S_x^2 d'une série statistique simple $\{X_1, X_2, ... X_N\}$ soit nulle, il suffit que la condition suivante soit remplie. Expliquez.

	Condition	Vrai	Faux
a	$\overline{X} = 0$		
b	Toutes les valeurs de la série sont égales entre elles		
c	La série statistique est symétrique par rapport à la moyenne		
d	L'écart type est nul		

4. Les mesures suivantes constituent-elles des mesures de tendance centrale ou de dispersion ?

	Tendance centrale	Dispersion
Etendue		
Ecart-type		
Mode		
Médiane		
Variance		
EMA		
Moyenne		
Coefficient G2		

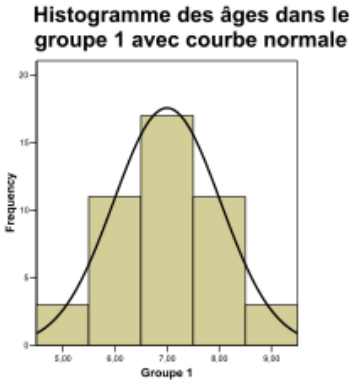
T.P. 5 - 6 : CHAPITRE 6

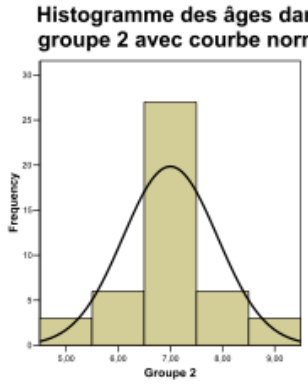
Partie 3 : Les Coefficients de Fisher

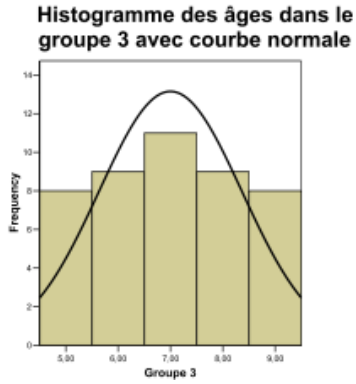
Exercice 1 : Observation graphique

1. Donnez les formules détaillées des coefficients de symétrie et d'aplatissement.

2. Voici trois graphiques représentant chacun la distribution des âges dans des groupes d'enfants.

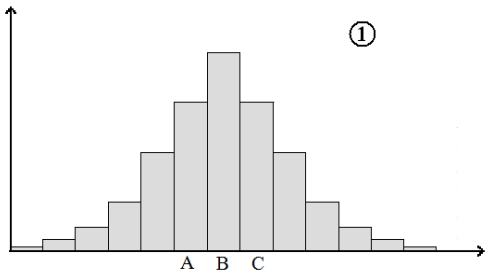


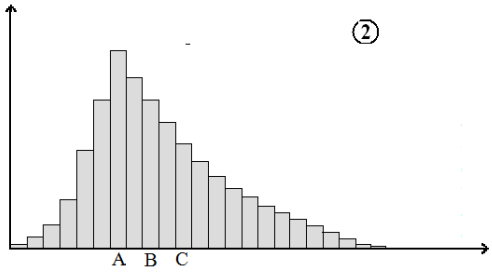


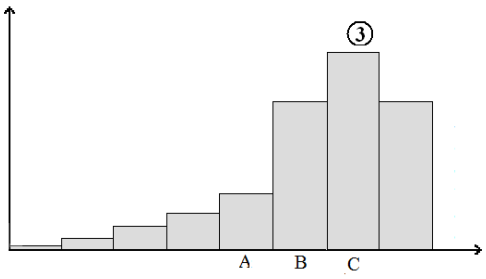


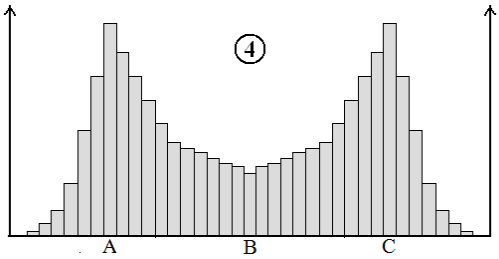
Que pensez-vous du signe du coefficient G_2 de Fischer pour ces trois distributions ? Justifiez votre réponse.

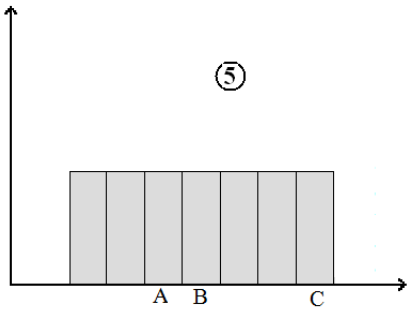
3. Voici cinq graphiques :











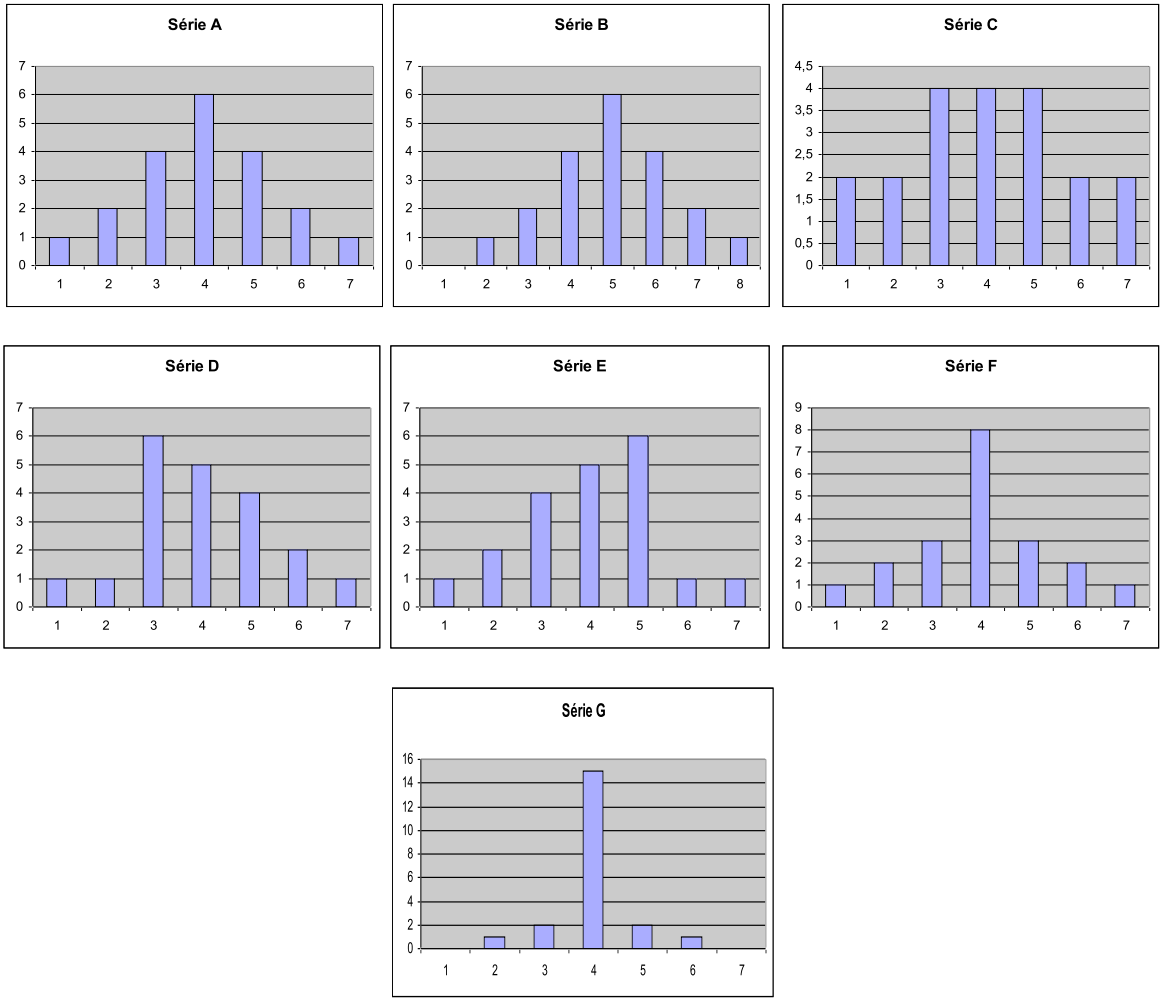
Pour chacun de ces graphiques, complétez le tableau ci-dessous en cochant les cases appropriées :

	1	2	3	4	5
Symétrique					
Asymétrique positive					
Asymétrique négative					
Unimodale					
Multimodale					

Pour chacun des graphiques, veuillez identifier la localisation du (ou des) mode(s), de la moyenne et de la médiane (associé au point A, B ou C ?)

	Mode	Moyenne	Médiane
1			
2			
3			
4			
5			

4. Les cinq diagrammes en barres suivants représentent six distributions d'un score variant de 0 à 10 mesuré pour 20 sujets.



Indiquez, par une lettre, à quelle série correspond chaque ensemble de statistiques.

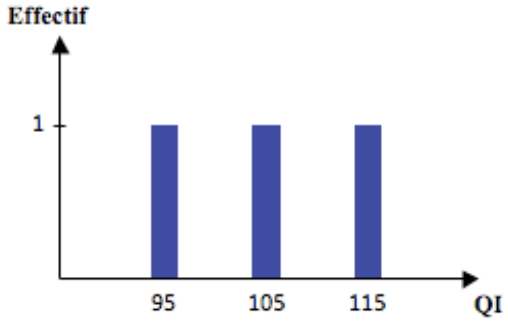
N.B. : Il n'est pas nécessaire d'effectuer de calculs pour répondre à cette question.

Moyenne	Ecart type	G1	G2	Série
4	1,41	-0,11	-0,25	
4	1,45	0	-0,35	
4	1,73	0	-0,8	
5	1,45	0	-0,35	
4	0,76	0	2.25	
4	1,41	0	-0,1	
4	1,41	0,11	-0,25	

T.P. 5 - 6 : Partie 3

Exercice 2 : Mode – Médiane – Moyenne & Symétrie

A. Voici les données d'un test de QI de trois enfants de 8 ans : 95, 105, 115. La médiane de cette distribution vaut 105.



1. Calculez la moyenne de cette série statistique. Notez la formule utilisée et sa décomposition. Comparez la moyenne et la médiane.

2. Intuitivement, que pouvez-vous dire du coefficient G_1 pour cette série ? Justifiez votre réponse.

3. Calculez le coefficient G_1 pour cette série.

4. Transformez cette série de manière à ce que la médiane ne change pas, mais soit plus petite que la moyenne, sans ajouter de données. Vérifiez votre réponse en calculant la moyenne.

5. Représentez cette série graphiquement.

6. Intuitivement, que pouvez-vous dire de G_1 pour cette série ? Justifiez votre réponse.

7. Calculez la valeur du coefficient G_1 pour cette série.

8. Ajoutons à la série de base 4 données mais uniquement d'un seul côté (à droite ou à gauche). Déterminez la médiane et commentez le résultat en la comparant avec la médiane de la série de base.

9. Intuitivement, que pouvez-vous dire de G_1 pour cette série ? Justifiez votre réponse.

9. Calculez la valeur du coefficient G_1 pour cette série.

10. Dans quel cas le mode, la moyenne et la médiane sont confondues ?

T.P. 5 - 6 : Partie 3

Exercice 3 : Mode – Médiane – Moyenne

Statistiques paramétriques et non paramétriques

A. Donnez des séries ou distributions statistiques répondant aux conditions suivantes (aidez-vous de représentations graphiques si nécessaire) :

a) Mode = moyenne = médiane
b) Mode <moyenne<médiane
c) Mode = médiane < moyenne
d) Mode = médiane > moyenne
e) Symétrique mais médiane=moyenne ≠ mode
f) Moyenne < mode = médiane
g) Moyenne > mode = médiane
h) Bimodale avec moyenne = médiane
i) Pas unimodale avec moyenne = mode = médiane

B. Donnez un exemple de données correspondant à de la statistique paramétrique d'une part, et non paramétrique d'autre part.

T.P. 5 - 6 : Partie 3

Exercice 4 : Calcul de l'asymétrie

Soit, les trois séries suivantes :

	1	2	3	4	5	6	7	8	9	10	11
x_a	1	2	3	3	5	2	4	3	3	4	3
x_b	3	2	2	3	2	4	2	5	3	1	4
x_c	1	3	3	4	3	4	5	2	4	4	2

1. Sur base du coefficient G_1 de Fisher, déterminez la forme de la distribution de ces trois séries, en cochant la case appropriée dans le tableau ci-dessous.

	Symétrique	Asymétrique positive	Asymétrique négative
x_a			
x_b			
x_c			

2. Sur base des valeurs obtenues en G_b et G_c, comparez et commentez les distributions. Aidez-vous d'une représentation graphique.

T.P. 5 - 6 : Partie 3

Exercice 5 : Coefficients de Fisher : G_1 et G_2

A. Soit les 3 séries statistiques suivantes :

A :

j	x_j	n_j
1	1	3
2	2	7
3	3	10
4	4	5
5	5	20

B :

j	x_j	n_j
1	1	10
2	2	10
3	3	12
4	4	10
5	5	10

Distrib. C :

j	x_j	n_j
1	1	2
2	2	7
3	3	12
4	4	7
5	5	2

1. Calculez les coefficients G_1 et G_2 pour les distributions statistiques A, B et C en vous rappelant que S_x^m est l'écart type de la série exposant *m*. Arrondissez à deux décimales. Commentez vos résultats.

2. Tracez le graphe de ces distributions et observez la relation entre les coefficients et la forme de votre courbe tracée. N'oubliez pas de donner une légende aux axes, de leurs donner un sens et une échelle. Complétez votre commentaire ci-dessus par le résultat de vos observations.

B. Voici les données de 50 personnes concernant leur taille (en cm) :

j	Valeurs de la variable : x_j	Fréquences absolues : n_j
1	156	1
2	157	2
3	158	1
4	159	1
5	160	1
6	161	2
7	162	3
8	163	2
9	165	1
10	166	10
11	167	2
12	168	8
13	169	1
14	170	1
15	171	2
16	172	2
17	173	2
18	174	1
19	175	2
20	177	3
21	178	2
	TOTAL	50

a. Calculez la moyenne des données suivantes. Donnez le meilleur modèle prédictif que vous connaissez pour ces données.

b. Sur quel type d'échelle vous trouvez-vous ?

c. Le coefficient G1 vaut 0.055 et G2 vaut -0.468. Commentez ces valeurs.

d. Quel graphique pouvez-vous tracer ? Pourquoi ?

T.P. 5 - 6 : CHAPITRE 6

EXERCICE RECAPITULATIF

1. Inventez deux séries statistiques de 10 sujets pour la variable QI en tenant compte des critères suivants :

1°) La première série est bimodale, la seconde est unimodale et symétrique.

2°) La première série doit avoir une moyenne de 58 et un de ses modes doit être de 45.

Exemple de réponse[24] :

i	X_i
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

i	X_i
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	

Une fois ces séries réalisées :

a. Calculez algébriquement la moyenne, indiquez la formule utilisée et identifiez les modes.

[24] Tout l'exercice est basé sur les données que vous aurez inventées, fort probablement différentes de celles proposées dans le corrigé. Il est donc évident que vous n'obtiendrez pas les mêmes réponses pour les questions suivantes.

b. Calculez l'étendue des deux distributions.

c. Dans quelle unité est exprimée cette étendue ?

d. Calculez l'écart moyen absolu de ces deux séries.

e. Donnez le meilleur modèle prédictif que vous connaissez pour chaque série.

f. Lorsque c'est adéquat, calculez une mesure de l'erreur résiduelle (si ce n'est pas possible indiquez pourquoi).

g. Etablissez la moyenne de l'erreur résiduelle par sujet. Comment s'appelle cette valeur?

h. Dans quelle unité avez-vous exprimé cette mesure? Si cette unité est différente de celle de la variable, comment pouvez-vous remédier à cette situation? Quel nom porte cette nouvelle valeur?

i. Vérifiez algébriquement à l'aide de la méthode de Fisher que la série 2 est symétrique.

j. Calculez le coefficient d'asymétrie pour la série 1. Comment se nomme ce coefficient? Quelle est sa traduction anglaise?

k. Quels sont les signes des coefficients d'aplatissement? Quel coefficient avez-vous utilisé ? Pourquoi? Quelle est sa traduction anglaise?

l. Pensez-vous que les coefficients de Fisher (G_1 et G_2) soient pertinents pour la série 1 et la série 2? Justifiez votre réponse.

m. Vérifiez graphiquement l'exactitude de vos résultats. Commentez.

n. Estimez la moyenne, la variance et l'écart-type de la population pour laquelle l'estimation pourrait être pertinente.

CHAPITRE 7 : LES DISTRIBUTIONS BINOMIALES ET NORMALES

7.1. Introduction

Dans ce chapitre, nous allons considérer deux distributions importantes, la distribution binomiale et la distribution normale. Pour être précis, nous devrions toujours parler de distributions binomiales et de distributions normales au pluriel, dans la mesure où il en existe une infinité possible. En effet, ces distributions dépendent de paramètres qui peuvent prendre une infinité de valeurs. Nous verrons qu'une distribution binomiale dépend d'une **variable aléatoire discrète** (discontinue) caractérisée par les paramètres *n* (nombre d'événements aléatoires) et *p* (probabilité d'occurrence d'un événement aléatoire donné). Une distribution normale sera définie par une **variable aléatoire continue** caractérisée par deux paramètres : la moyenne et la variance.

Cette distinction entre variables discrète et continue est très importante. En effet, comme énoncé très simplement par Sanders, Murphy et Eng (1984), une distribution de probabilité n'est rien d'autre qu'une "*énumération complète de tous les résultats possibles d'une expérience avec leur probabilité respective*". Or, dans le cas d'une variable discrète les événements possibles sont dénombrables, alors que dans le cas d'une variable continue, les événements possibles sont infinis.

Les points suivants développeront les notions de variables aléatoires, mettront en évidence l'importance des paramètres qui définissent ces variables dans les deux distributions et dégageront le lien qui existe entre ces deux types de distributions.

7.2. La distribution binomiale

7.2.1. Introduction

Imaginons le cas classique du jet d'une pièce de monnaie. Nous sommes dans un cas très particulier de distribution qui apparaît lorsque les événements qui la composent, les expériences aléatoires, n'ont que deux issues possibles. Soit l'expérience aléatoire donne un résultat, soit l'autre. On parlera souvent de réussite ou d'échec, même si c'est à

l'expérimentateur de définir quel événement correspond à quelle détermination. Lorsque l'on jette une pièce de monnaie en l'air, elle retombera soit sur "*pile*", soit sur "*face*" et ce, en principe, de manière équiprobable. Dès lors, appeler l'événement "*face*" une réussite a autant de sens que d'appeler l'événement "*pile*" de la même manière. On décidera donc en fonction du contexte. D'autres variables peuvent s'établir sur un processus semblable, cependant, la probabilité *p* n'est pas obligatoirement égale à .5 comme on le considère dans le jet d'une pièce de monnaie. Par exemple, admettons que l'on évalue la probabilité qu'une machine à laver sorte de l'usine, et fonctionne, et que cet événement arrive dans 95% des cas. On appellera vraisemblablement *p* la probabilité d'être fonctionnel qui sera la probabilité de réussite égale à .95 et *q* la probabilité d'échec, c'est-à-dire d'avoir une machine à laver défectueuse à la sortie de l'usine, égale à .05. Cependant, le patron d'une société concurrente pourrait considérer qu'une machine à laver défectueuse chez son concurrent est une victoire et appeler cet événement "*réussite*" sans changer les conclusions mathématiques qui vont suivre.

La distribution d'une variable aléatoire dont les événements n'ont que deux issues possibles se nomme binomiale. Vous aurez compris (sinon relisez votre chapitre sur les probabilités) que la probabilité d'une des issues de l'événement est *p* et que la probabilité de l'autre est ce qu'il reste, c'est-à-dire 1-*p* que l'on appelle *q*. Donc nous retrouvons la relation $\boldsymbol{p + q = 1}$.

7.2.2. Equation de la variable aléatoire discrète

Si l'on considère le Tableau 7.1, on retrouve les caractéristiques d'une telle variable. Ici, nous pouvons considérer qu'il s'agit d'une pièce de monnaie bien équilibrée ou, en tout cas, d'une situation similaire : on réalise dix fois un événement aléatoire (admettons un tirage à pile ou face) et la probabilité de réussite est de $p = .5$. Admettons que *p* désigne la probabilité d'obtenir l'événement "*pile*", ce qui implique que *q* désigne la probabilité d'obtenir "*face*" (on néglige la très faible probabilité d'obtenir la tranche). Les deux premières colonnes de ce tableau montrent qu'obtenir 0 fois "*pile*" implique nécessairement d'obtenir 10 fois "*face*", obtenir une fois "*pile*" conduit à l'obtention de 9 "*face*", etc.. En d'autres termes, pile et face sont mutuellement exclusifs (si on obtient l'un, on n'obtient pas l'autre) et exhaustifs (il n'y a donc pas d'autre possibilité que pile ou face). Nous sommes également dans le cas d'une indépendance entre les événements : admettons que j'aie déjà lancé 9 fois ma pièce et obtenu 9 fois "*pile*", j'ai néanmoins une chance sur deux d'obtenir "*pile*" la dixième fois dans

la mesure où tout ce qui s'est passé avant n'influence absolument pas le jet que je considère : je vous rappelle que les statistiques n'ont pas de mémoire.

Tableau 7.1. : Distribution binomiale (n = 10, p = .5)

Nombre de fois *pile*	Nombre de fois *face*	Probabilités	Probabilités cumulées
0	10	0,0010	0,0010
1	9	0,0098	0,0107
2	8	0,0439	0,0547
3	7	0,1172	0,1719
4	6	0,2051	0,3770
5	5	0,2461	0,6230
6	4	0,2051	0,8281
7	3	0,1172	0,9453
8	2	0,0439	0,9893
9	1	0,0098	0,9990
10	0	0,0010	1,0000

L'obtention de la probabilité des événements décrits par les colonnes 1 et 2 du Tableau 7.1, c'est-à-dire les colonnes 3 et 4 de ce tableau, a été réalisée par l'application de l'équation définissant la distribution de probabilité. Cette distribution est liée à un schéma d'expérience aléatoire appelé ***schéma de Bernouilli***, dont un nombre n de lancers d'une pièce de monnaie en est l'illustration la plus courante. La variable aléatoire qui produira la distribution binomiale a pour équation (non démontrée ici) :

Equation de la distribution binomiale

$$p(k) = P(X = k) = \binom{n}{k} p^k q^{n-k}$$

où n = nombre de fois que l'on réalise l'expérience aléatoire (par exemple le nombre de fois qu'on lance une pièce de monnaie) ; k = le nombre de fois que cet événement est une "réussite" ; p est la probabilité de réussite ; q est la probabilité d'échec. La notation particulière mettant entre parenthèses le n et le k correspond en fait à la combinaison de k

éléments parmi *n*. Elle se calcule comme suit (je vous renvoie au chapitre 3.5 sur l'analyse combinatoire) :

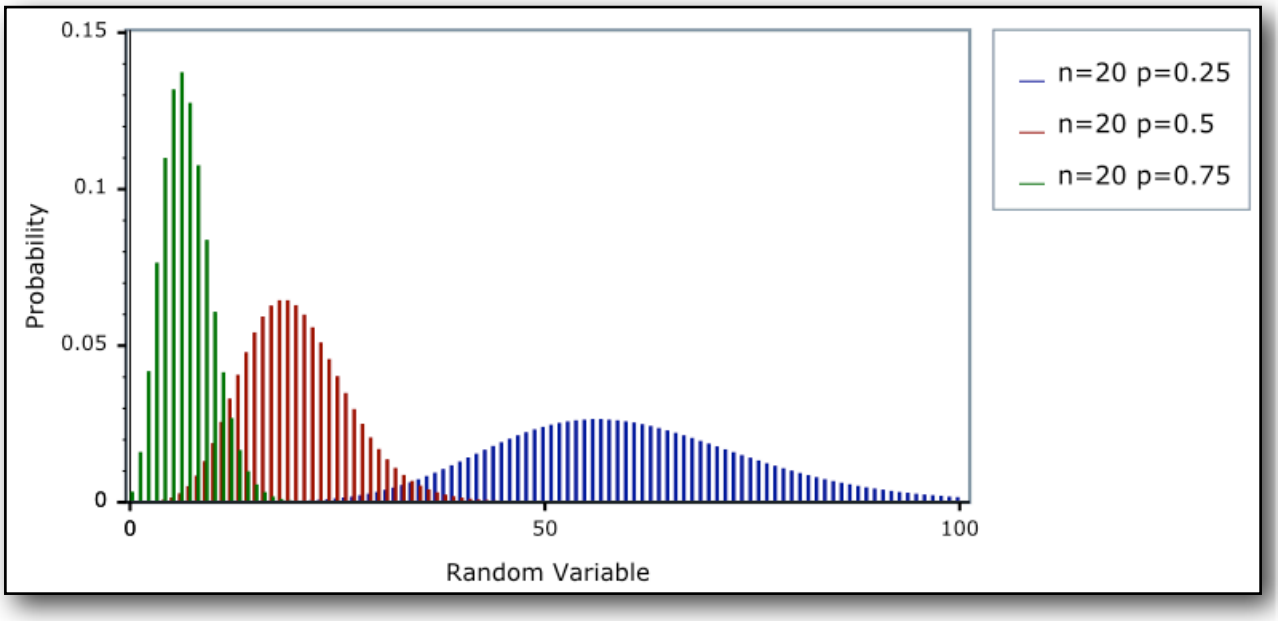
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Je rappelle ici que les points d'exclamation symbolisent l'opération mathématique "*factorielle*" qui consiste à multiplier le nombre entier avant le point d'exclamation par tous les entiers qui lui sont inférieurs. Par exemple, 4! = 4*3*2(*1) = 24. Reprenons un exemple facilement compréhensible de l'utilisation des combinaisons : le *Lotto* (avec deux "t" parce que je l'illustre à partir du Lotto belge et non du Loto français). Il existe 42 numéros différents possibles, et on tire six numéros sans remise (ce qui signifie que les tirages ne sont pas indépendants). On a donc une combinaison de 6 numéros parmi 42. Peu importe l'ordre d'apparition du chiffre, puisqu'il n'est pas imposé aux joueurs d'avoir les 6 chiffres dans un ordre d'apparition précis (à l'inverse du Tiercé où le classement compte). Il y a donc 42!/[6!(42-6)!] combinaisons possibles, c'est-à-dire 5245786 combinaisons, dont une seule est gagnante. La probabilité de gain est donc de 1/5245786, c'est-à-dire environ cinq fois moindre que celle de faire un accident d'avion, qui est déjà ridiculement faible. En d'autres termes, si les probabilités influençaient les émotions des individus, ceux qui pensent sérieusement avoir une chance de gagner au Lotto devraient être terrorisés en avion, être totalement hystériques en voiture, et résignés à la mort en enfourchant un vélo (seul moyen de transport plus dangereux encore que la moto, mais plus accessible aux enfants…).

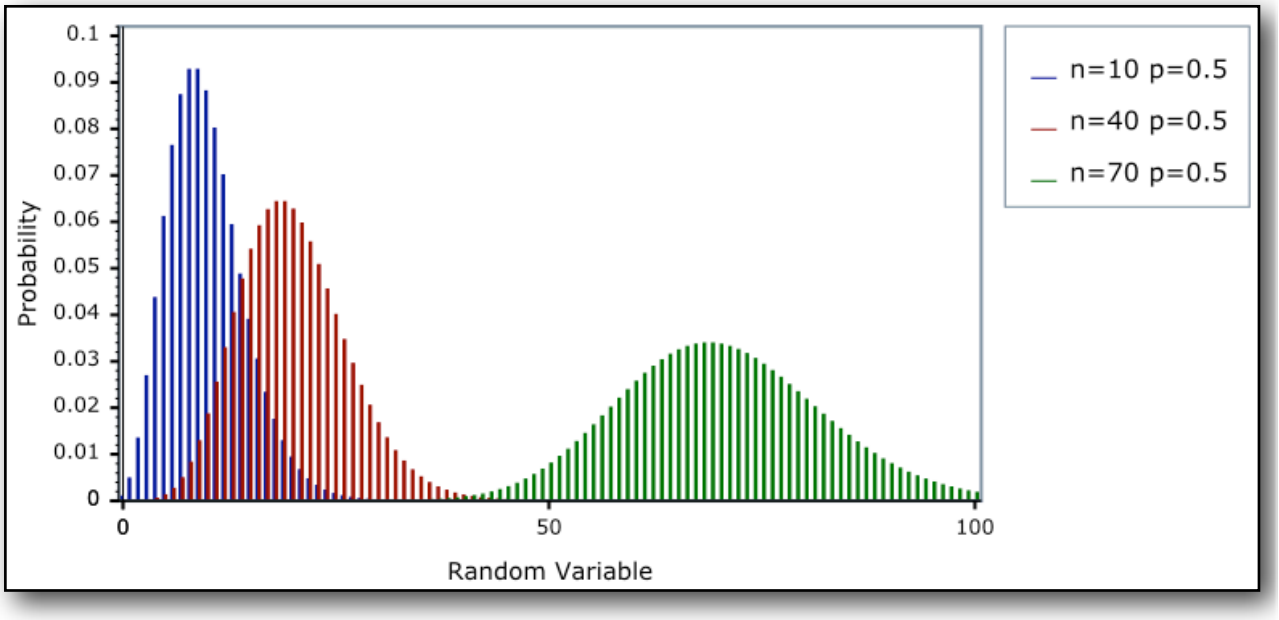
En conclusion, la distribution binomiale est une distribution entièrement caractérisée par deux paramètres : *n* et *p* (dans la mesure où *k* est votre inconnue et que *q* est fonction de *p*). La Figure 7.1 représente une telle distribution. Vous constatez que (Figure 7.1.a) lorsque *n* est constant, plus *p* est grand plus la distribution est pointue et ramassée vers l'axe y (des probabilités) alors que (Figure 7.1.b), lorsque *p* est constant, plus *n* est grand, plus la courbe est plate et décalée vers la droite. De même, les distributions sont toujours symétriques pour *p* = .5 et toujours asymétrique autrement.

Figure 7.1. : Distributions binomiales de paramètres *p* différents (a) et *n* différents (b)

(a) Formes des distributions en fonction de *p*



(b) Formes des distributions en fonction de *n*



Voyons quelles sont les probabilités d'obtenir, 1, 2, 3 ou 4 succès pour *n* valant respectivement 1, 2, 3 et 4 lancers successifs d'une pièce parfaitement équilibrée. Comme d'habitude, ceci implique de faire l'inventaire des événements possibles dans chaque situation et de compter le nombre de cas favorables.

- Pour n = 1, il y a 2^1 = 2 événements élémentaires possibles et la variable aléatoire "*nombre de succès*" peut prendre deux valeurs : 0 ou 1, chaque valeur se produisant avec une probabilité de .50.

- Pour n = 2, il y a 2^2 = 4 événements possibles. L'inventaire des événements élémentaires possibles est constitué par les séquences suivantes : PP, FP, PF et FF, chacune de ces séquences se produisant avec une probabilité .25 (= .50 x .50 puisque les deux événements constituants sont indépendants). La variable aléatoire "*nombre de succès*" peut prendre trois valeurs : 0, 1 et 2 dont les probabilités sont respectivement .25, .50 et .25 (1/4 de cas avec 0 fois pile – FF, 2/4 de cas avec 1 fois pile – PF et FP – et 1/4 de cas avec 2 fois pile – PP).

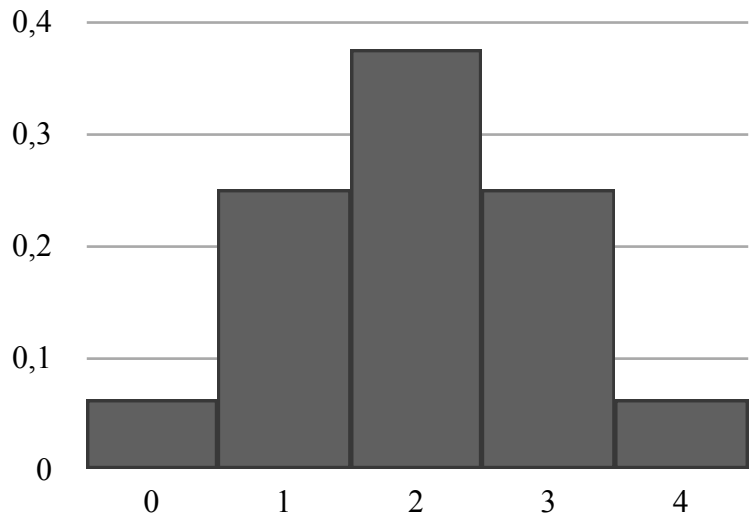
- Pour n = 3, il y a 2^3 = 8 événements possibles. L'inventaire des événements élémentaires possibles est constitué par les séquences suivantes : PPP, FPP, PFP, PPF, FFP, FPF, PFF et FFF, chacune de ces séquences se produisant avec une probabilité de .125 (= .50 x .50 x .50 puisque les trois événements composants sont indépendants). La variable aléatoire "*nombre de succès*" peut prendre quatre valeurs : 0, 1, 2 et 3 dont les probabilités sont .125, .375, .375 et .125 (1/8 de cas avec 0 fois pile – FFF, 3/8 de cas avec 1 fois pile – PFF, PFP, FFP – 3/8 de cas avec 2 fois pile – PPF, PFP, FPP – et 1/8 de cas avec 3 fois pile – PPP).

Ces quatre exemples vous informent sur le mode de calcul de ces probabilités, et vous montrent que cela devient vite fastidieux lorsque n augmente. Utilisons maintenant l'équation de la binomiale. Pour n = 2, il nous faut envisager la possibilité de 0 ou de 1 succès. Zéro succès correspond à effectuer le calcul suivant : $2!/(1!*1!)*.5^0*.5^2 = .5$ ce qui correspond bien à ce que nous suspections. Rappelons que n'importe quelle valeur exposant 0 est égale à 1. De même, factorielle de zéro est aussi égale à 1, par convention. Je vous laisse trouver l'alternative pour un succès.

Imaginons-nous maintenant une situation où n = 4 et cherchons la probabilité d'obtenir trois réussites. Nous avons vu, en réfléchissant de manière probabiliste, que la probabilité attendue est de .25. En remplaçant correctement les valeurs dans l'équation, nous obtenons : $4!/(3!*1!).5^3*.5^1 = .25$. Je vous laisse à nouveau trouver les autres résultats et faire le même exercice pour n = 2 et 3.

Vous pouvez également réaliser facilement l'histogramme de fréquences correspondant à ces résultats. La Figure 7.2 représente le cas de $n = 4$, je vous laisse tracer ceux des n précédents. Si vous comprenez ce principe, vous pourriez, avec un peu de patience, retrouver les distributions de la Figure 7.1 (ne perdez pas votre temps à le faire, comprenez juste le principe).

Figure 7.2. : Distribution binomiale $n = 4$, $p = .5$



Remarquez enfin que toute variable aléatoire discrète peut être considérée comme une binomiale. Par exemple, un jet de dé pourrait être exprimé de la sorte. Bien sûr, il existe 6 expressions possibles de l'événement. Mais nous pouvons envisager la probabilité de tirer un 6 versus celle de ne pas tirer un 6, ou bien la probabilité de tirer un nombre pair, versus la probabilité de ne pas tirer un nombre pair, etc. ce qui constitue une dichotomisation de la probabilité.

Dans le cas où il est impossible de dichotomiser le problème, il devient nécessaire de faire appel à la généralisation de la distribution binomiale, appelée **distribution multinomiale**. Cette distribution ressemble assez fort, mais dépend des paramètres n et p_1, …, p_k où k est le nombre de résultats possibles (6 pour un dé cubique). Nous ne nous attacherons pas à décrire cette distribution. Cependant, si vous désirez en savoir plus, je vous renvoie à l'ouvrage de Forbes, Evans, Hasting, & Peacock (2011).

Par ailleurs, lorsque *p* est très faible, il existe une distribution plus adéquate que la distribution binomiale. Elle se nomme "**distribution de Poisson**", du nom d'un mathématicien célèbre (Siméon-Denis Poisson, 1781-1840) et sans aucun rapport avec l'ichtyologie. Cette distribution concerne donc essentiellement les événements rares, bien qu'elle soit également utilisée dans d'autres cas. Parmi les utilisations, citons, le suicide des enfants, l'apparition d'une maladie rare comme le SIDA (en Belgique), la désintégration radioactive, les mutations, etc.. Cette distribution est entièrement caractérisée par un paramètre λ (lambda) qui est un nombre réel strictement positif traduisant la moyenne d'occurrence d'un événement dans un laps de temps donné. Nous ne l'envisagerons pas, mais retenez son existence.

7.2.3. Utilisation des tables de la binomiale

En pratique, on utilise très peu l'équation de la binomiale (bien que, comme nous venons de le voir, ce ne soit pas inaccessible). On a plutôt tendance à utiliser les tables, plus facilement accessibles dans les manuels de statistiques, et encore plus tendance à utiliser les logiciels statistiques. Le Tableau 7.2 présente une telle table pour quelques valeurs des paramètres *p* et *n*. Vous retrouverez les valeurs que nous avons envisagées au point précédent. Par exemple, pour $n = 4$ et $p = .5$, les probabilités d'avoir 0, 1, 2, 3 ou 4 succès que nous avons calculées sont à la dernière colonne.

Tableau 7.2. : Probabilités de la distribution binomiale pour certaines valeurs de ses paramètres. Source : http://jeancharles.canonne.pagesperso-orange.fr/table_loi_binomiale.htm retrouvé le 23/10/11.

		p=									
n	r	0,0500	0,1	0,15	0,2	0,25	0,3	0,35	0,4	0,45	0,5
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010

A titre d'exercices, répondez aux trois questions suivantes :

a) Quelle est la probabilité d'obtenir quatre fers à repasser défectueux parmi dix fers produits sachant que la probabilité d'être défectueux est de 5%?
b) Quelle est la probabilité d'avoir au moins trois chatons femelles dans trois portées de trois chatons par portée?
c) Quelle est la probabilité dans un groupe de 8 étudiants de trouver entre 3 et 5 étudiants qui n'ont jamais échoué en secondaire, sachant que seuls 35% des étudiants réussissent le secondaire sans redoubler une seule année?

7.2.4. Statistiques descriptives d'une distribution binomiale

La Figure 7.1 vous a montré que ces distributions étaient plus ou moins symétriques et toujours unimodales. En outre, nous verrons par la suite qu'une fois que le *n* est suffisamment grand (au-delà de 30), les distributions sont très proches de la symétrie quelle que soit la probabilité de réussite. Il y a donc un sens descriptif à calculer la moyenne et la variance d'une telle distribution. Il se trouve que ces paramètres sont très faciles à évaluer à partir des paramètres *n* et *p* décrivant la distribution binomiale. Si le résultat est évident, la démonstration mathématique n'apporte pas grand-chose pour nous. Vous pouvez donc admettre sans démonstration que :

La moyenne, la variance et l'écart-type d'une distribution binomiale (*n*, *p*) valent

$$\mu_X = np$$

$$\sigma^2_X = np(1\text{-}p) = npq$$

$$\sigma_X = \sqrt{np(1\text{-}p)} = \sqrt{npq}$$

De même, on peut calculer les coefficients d'asymétrie et d'aplatissement de cette distribution en appliquant des formules relativement simples :

Coefficient d'asymétrie et d'aplatissement d'une distribution binomiale (*n*, *p*) valent

$$Asymétrie = (q-p)/(\sqrt{npq})$$

$$Aplatissement = (1-6p\ q)/(npq)$$

A nouveau, les passionnés pourront trouver les démonstrations de ces paramètres dans l'ouvrage de Forbes, Evans, Hasting et Peacock (2011). Personnellement je serais incapable de faire la démonstration sans leur aide.

7.2.5. Utilisation de la table en termes de proportions de succès

Jusqu'à présent, nous avons considéré le nombre de succès et d'échecs. Cependant, il est souvent nécessaire de s'exprimer en termes de "*proportion de succès*". En effet, il est plus intéressant de savoir combien de succès on a obtenus, par exemple combien de machines à laver fonctionnelles sont sorties de l'usine, on préférera souvent donner l'information du nombre de machines à laver fonctionnelles qui sont sorties par rapport au nombre de machines à laver qui ont été produites. Ce type d'informations est particulièrement utile lorsque l'on cherche à inférer la proportion de succès pour la population à partir de l'échantillon. La proportion de succès se définit donc comme le nombre de succès/n. Pour n = 10, la proportion de succès varie entre 0 et 1 par tranche de 0,1. Cette situation est représentée par le Tableau 7.3.

Tableau 7.3. : Extrait d'une table de distribution binomiale exprimée en proportions

Proportion de succès	*Nbr. de succès*	$P(X = r)$	$P(X \leq r)$
0,00	0	.3487	.3487
0,10	1	.3874	.7361
0,20	2	.1937	.9298
0,30	3	.0574	.9872
0,40	4	.0112	.9984
0,50	5	.0015	.9999
0,60	6	.0001	1.0000
0,70	7	.0000	1.0000
0,80	8	.0000	1.0000
0,90	9	.0000	1.0000
1,00	10	.0000	1.0000

Les conséquences sur les statistiques descriptives correspondantes sont assez évidentes : la moyenne et l'écart-type sont divisés par n alors que la variance est divisée par n^2 (cela devrait vous sembler relativement évident). De sorte que :

Moyenne

$$\mu_p = \frac{Np}{N} = p$$

Variance

$$\sigma_p^2 = \frac{Np(1-p)}{N^2} = \frac{p(1-p)}{N}$$

7.4. La distribution normale

7.4.1. Etablissement d'une fonction de densité de probabilité

Admettons maintenant que notre variable aléatoire soit continue. Cela revient à considérer que les rectangles que je représentais en traçant l'histogramme (par exemple à la Figure 7.2) deviennent infiniment petits et infiniment nombreux. Pour ceux qui se souviennent de leur cours de mathématique, on dira que l'intervalle sur l'axe des x tend vers zéro ($dx \rightarrow 0$). Dès lors, l'équation de la variable aléatoire binomiale ne s'applique plus puisque N est infini, ce qui signifie qu'un des paramètres de la binomiale est nécessairement indéterminé. De fait, imaginons-nous une girouette qui tourne au gré du vent (peu importe si c'est grâce à un coq ou à un lion). Imaginons ensuite une bourrasque soudaine qui provoque un affolement de la girouette, suivit d'un arrêt complet du vent et donc de la girouette. Nous pouvons déterminer qu'il y a une chance sur quatre que la girouette s'arrête dans le premier quadrant (de 0 à 90°), une chance sur quatre dans le deuxième, et identiquement pour le troisième et le quatrième. Nous avons donc envisagé cette variable comme discrète. En revanche, si nous nous interrogeons sur la probabilité que la girouette s'arrête précisément à la valeur de 30° (virgule zéro à l'infini), nous devons considérer que cette probabilité est nulle. En effet, puisque le nombre de degrés est théoriquement infini, le nombre de cas favorables est de 1

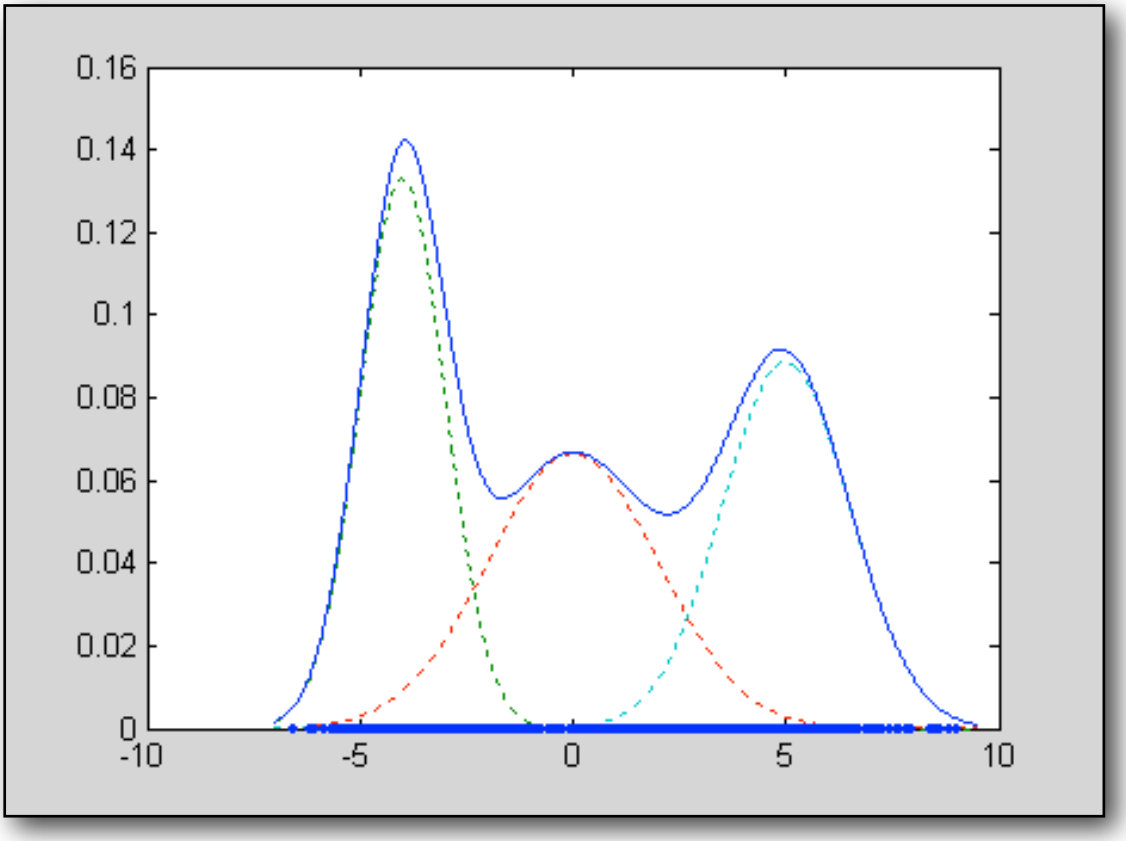
(30° virgule zéro à l'infini) et le nombre de cas possibles est infini. Un divisé par l'infini étant nul, il n'y a en théorie aucune chance que la girouette s'arrête précisément sur cet angle (ni sur un autre d'ailleurs). Toutes les valeurs ont une probabilité nulle d'occurrence. En pratique, la matière nous limite et rend la probabilité non nulle tout simplement parce que les atomes et les molécules ont une taille finie alors qu'un point mathématique n'en a pas. De plus, nos instruments de mesure ne permettent de mesurer l'angle qu'à un certain degré de précision de sorte que la variable n'est pas réellement continue. Elle est discrète mais contient un nombre très très important de valeurs. Dès lors, cela revient à dire que nous ne pouvons mesurer qu'un intervalle d'angles, très petit, mais pas infiniment petit. Or, s'il est impossible de définir une probabilité ponctuelle pour une valeur précise d'une variable aléatoire continue, il est en revanche possible de trouver la probabilité sur un intervalle, si petit soit-il. Par exemple, nous pourrions trouver la probabilité que la girouette s'arrête sur un angle compris entre 30,000000000000000000000 et 30,000000000000000000001 degrés. Il suffit pour cela d'intégrer la courbe qui décrit la distribution sur l'intervalle concerné (l'intégrale d'une courbe donnant l'aire sous la courbe, comme nous l'avons déjà vu à la Figure 5.3). Une telle courbe peut se décrire par une équation. Cette équation est un idéal théorique qui tente de décrire le mieux possible une distribution observée. On nomme ce type de courbe la fonction de **densité de probabilité** d'une variable aléatoire continue. La densité de probabilité de l'entièreté de la courbe est égale à 1. Une fois cette équation établie, sous la forme de $y = f(x)$, l'impossibilité que nous rencontrions ci-avant est surmontée : en effet, si l'on reprend notre exemple de la girouette, il est tout à fait certain qu'elle s'arrêtera quelque part entre 0° et 360°. A l'aide de l'équation, il devient maintenant possible d'estimer la probabilité ponctuelle d'une valeur, puisqu'à chaque valeur de l'axe x correspond une valeur sur l'axe y.

Modélisation d'une variable continue

En résumé (Figure 7.3), une variable aléatoire continue peut être modélisée par une distribution théorique présentée sous la forme d'une courbe correspondant à une équation de type $y = f(x)$, qui se lit y est une fonction de x où y est représenté par l'axe vertical du graphe qui informe sur les fréquences et x par l'axe horizontal qui informe sur la valeur de la variable aléatoire.

1. L'aire sous l'entièreté de la courbe est égale à 1.
2. L'aire sous tout intervalle (x_b-x_a) est positive et inférieure à 1.

Figure 7.3. : Exemple de fonction de densité de probabilité *y=f(x)*. La ligne pleine montre une fonction de densité multimodale formée par la réunion de trois distributions unimodales, en pointillés.



7.4.2. Caractéristiques d'une distribution normale

En pratique, la binomiale devient rapidement inutilisable. Dès que le *N* est trop important, même les ordinateurs les plus sophistiqués deviennent incapables de gérer une telle distribution. Faites par exemple l'exercice de calculer, la combinaison de 3 parmi 500, même à l'aide d'un tableur qui utilisera toute la puissance de calcul de votre processeur, et vous constaterez qu'il ne parviendra pas à gérer de tels nombres. Heureusement, ce n'est pas nécessaire. En effet, une fois le *N* suffisamment grand (en pratique on considère $N > 30$) la distribution binomiale tend vers une distribution normale, c'est-à-dire que l'on peut considérer cette variable aléatoire discrète comme une variable aléatoire continue.

Rappelons qu'une distribution normale est une distribution théorique qui semble représenter correctement la distribution de nombreuses variables aléatoires naturelles. Cela peut sembler contre-intuitif que de nombreuses variables se distribuent de la même manière alors que l'on peut envisager une infinité de distributions différentes, allant d'un mode à une infinité de modes en passant par toutes les asymétries et les aplatissements possibles (voir chapitre 5.5). Cependant, si on réfléchit ce n'est pas si inimaginable. Reprenons, par exemple, la taille des femmes adultes belges que nous avions envisagée au premier chapitre. Il ne vous semblera pas étonnant qu'il existe une taille habituelle (on peut maintenant dire une taille moyenne) qui représente correctement la taille de nombreuses femmes (c'est-à-dire qui minimise la SCE). Nous avions considéré que cette taille est de 169cm. Vous admettrez sans peine que, plus je m'éloigne de cette taille, que ce soit en positif (vers les femmes très grandes) ou en négatif (vers les femmes très petites), moins je trouverai d'individus. Il en va de même pour beaucoup de caractéristiques dans de nombreux domaines. Citons : le bonheur, l'indicateur BMI, le taux de cellules sanguines, la pureté d'un diamant, la pluviométrie d'une zone géographique, la taille d'un banc de thons, la vitesse des voitures en ville à une heure donnée, etc..

C'est Abraham De Moivre (1667-1754), illustré ci-contre, qui a proposé la première équation décrivant une distribution normale lorsque *N* est suffisamment grand, en 1733. Nous sommes 20 ans après la publication de l'*Ars conjectandi* de Jacob Bernouilli (1654-1705). Cette trouvaille sera largement diffusée dans la 2ème édition (datant de 1738) de son livre *The doctrine of chance, or a method for calculating the probability of events* (1ère édition en 1718). Cependant, c'est bien plus tard, au 19ème siècle, que Gauss généralisera le concept et précisera l'équation de la loi normale. Voyez l'encadré ci-dessous pour les formules que vous ne devez pas connaître.

Fonction de densité de probabilité pour la distribution normale

Pour une distribution normale standard (donc basée sur la variable Z qui, comme nous le verrons consiste à ôter à chaque score la moyenne de la distribution et diviser par son écart-type), l'approximation proposée par De Moivre était la suivante :

$$f(x) = \frac{1}{\sqrt{\pi}} e^{-z^2}$$

Actuellement, on utilise plutôt la forme modifiée attribuée à Gauss

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-z^2}{2}} \quad (1)$$

et, pour la variable non standard de moyenne μ et d'écart-type σ

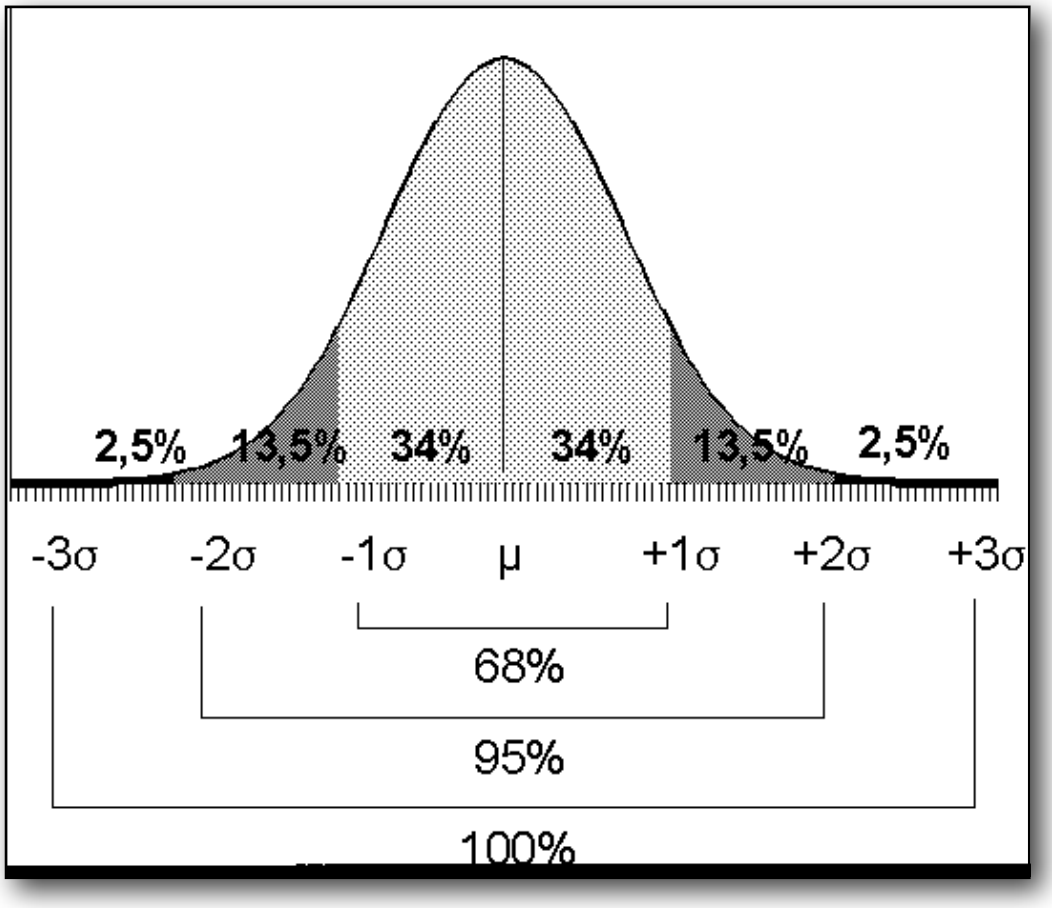
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Cette formule contient cinq constantes : le nombre 1 au numérateur du coefficient qui multiplie *e*, deux fois le nombre 2, une fois le nombre pi (π = 3,14159…) et une fois le nombre de Neper (e = 2,71…). Elle contient aussi deux paramètres : μ et σ. Les paramètres peuvent prendre une infinité de valeurs mais leurs valeurs sont complètement spécifiées pour chaque distribution normale particulière. Donc, si on cherche la probabilité d'obtenir 233 fois "*pile*" en lançant 500 fois une pièce parfaite, il faut remplacer μ et σ dans la formule ci-dessus par les valeurs de la moyenne et de l'écart-type d'une distribution binomiale (N = 500, p =.50). Ceci donne une densité de probabilité de .1257. Cette probabilité de .1257 est une estimation de la probabilité binomiale qui, compte tenu du grand nombre d'essais (N = 500) doit différer très peu de la probabilité binomiale exacte.

Vous retiendrez donc que la distribution normale est entièrement définie par les paramètres de moyenne et de variance de la population concernée. Elle a une forme tout à fait caractéristique et des propriétés particulièrement intéressantes. Tout d'abord, elle est symétrique. Par facilité, on la représente toujours par une variable centrée et réduite, c'est-à-dire de moyenne égale à zéro et d'écart-type égal à un. Deuxièmement, plus l'intervalle autour de la moyenne est élevé, plus la densité de probabilité est grande (nécessairement) et la valeur de cette densité est déterminée par l'équation. La Figure 7.4 montre que lorsqu'on s'éloigne d'un écart-type de la moyenne, 68% des valeurs de la variable sont comprises dans l'intervalle. En s'éloignant de deux écart-types, la probabilité monte à 95% pour atteindre les 100% à trois écart-types (à peu près, en fait les 100% sont atteints à l'infini mais, dès trois écarts-types, on est proche de cette limite : 99,72%). Appliqué à notre exemple des femmes belges, 68% des femmes se trouvent à un écart-type (dont je n'ai pas encore parlé, mais

supposons qu'il vaille 9,5 cm). Donc, si la taille des femmes en Belgique suit bien une distribution normale, 68% des femmes belges mesurent entre 159,5 cm (= 169-9,5) et 178,5 cm (=169+9,5) ; 95% des femmes mesurent entre 150 cm et 188 cm ; et 99,72% des femmes mesurent entre 140,5 cm et 197,5 cm.

Figure 7.4. : Distribution normale et densité de probabilité à 1, 2 ou 3 écarts-types autour de la moyenne. Sur le schéma, la courbe s'arrête à trois écarts-types. En théorie elle s'étend de chaque côté à l'infini. Source : http://www.er.uqam.ca/nobel/r30574/PSY1282/C3P6.html retrouvé le 23/10/11



Admettons qu'un collègue anversois ait estimé la taille moyenne d'une population à 170 cm avec un écart-type de 10 cm (supposons qu'il ait utilisé un échantillon de 100 femmes belges un petit peu différent de celui que j'ai décrit au chapitre 1, donc son estimation est différente de la mienne, mais il était plus près de la frontière hollandaise que moi, peut-être n'avons-nous pas sélectionné nos sujets de manière parfaitement aléatoire). Une taille de 180 cm correspond alors à une densité de probabilité de .0242. Cette valeur est obtenue en remplaçant μ, σ et *X* par les valeurs 170, 10 et 180 dans la formule (2). Il s'agit de la valeur de

l'ordonnée de la courbe de densité de probabilité pour la valeur d'abscisse 180 cm. Vous pouvez faire l'exercice avec notre échantillon de moyenne valant 169 cm et d'écart-type égal à 9,5. Vous devriez obtenir 0,0215 pour une taille de 180cm. Pour calculer la probabilité que la taille soit comprise entre 175 et 185 cm, il faut déterminer la fraction de la surface totale de la courbe qui est interceptée par ces deux limites[25].

La plupart des traités de statistique mathématique spécifient les distributions normales à partir des valeurs de la moyenne et de la variance. On écrit par convention $N(\mu,\sigma^2)$. On dirait, par exemple, de la variable taille, dans l'exemple ci-dessus, qu'elle est distribuée approximativement *N(170, 100)*, où 170 est l'estimation de μ et 100 l'estimation de σ^2. Toutefois, dans ce cours, les paramètres des distributions normales seront la moyenne et l'écart-type. On écrira donc $\boldsymbol{N(\mu,\sigma)}$. Par exemple, $N(170, 10)$ avec 170 l'estimation de μ et 10 l'estimation de σ. Ce choix est de loin préférable puisque la moyenne et l'écart-type sont exprimés dans la même unité que la variable (le cm dans l'exemple). En outre, un logiciel comme Excel demande de spécifier μ et σ pour calculer les valeurs *z* de la courbe de densité de probabilité.

De la même manière que lors de la distribution binomiale, nous n'aurons jamais besoin d'utiliser l'équation de la distribution normale (raison pour laquelle je ne vous impose pas de la retenir). Les logiciels informatiques sont beaucoup plus performants que nous pour calculer avec précision les valeurs qui nous intéressent. Par ailleurs, il est indispensable que vous compreniez la manipulation de cette distribution, et il existe une table appelée "*table normale standard*" dont nous allons bientôt apprendre l'utilisation. Mais avant, il est nécessaire de comprendre le principe de standardisation d'une variable aléatoire.

7.4.3. Standardisation d'une variable aléatoire et spécification formelle de la distribution normale

Comme nous l'avons vu, en abordant l'équation de la distribution normale, deux paramètres la caractérisent : la moyenne et l'écart-type (ou la variance). Ce qui implique qu'il existe une

[25]Mathématiquement, ceci revient à calculer l'intégrale de la courbe dans les limites 175, 185 selon la formule 2 de l'encadré :

$$\int_{x_j=175}^{x_{j'}=185} \frac{1}{10\sqrt{2\pi}} e^{\frac{-(x_j-170)^2}{2.10^2}} dx$$

infinité de distributions normales envisageables. Dès lors, déterminer la probabilité pour la valeur d'une variable à l'aide d'une seule table devient difficile. Nous verrons en BA2 que d'autres ennuis surviennent lorsque l'on désire analyser des liens entre plusieurs variables et qu'elles suivent toutes une distribution normale différente.

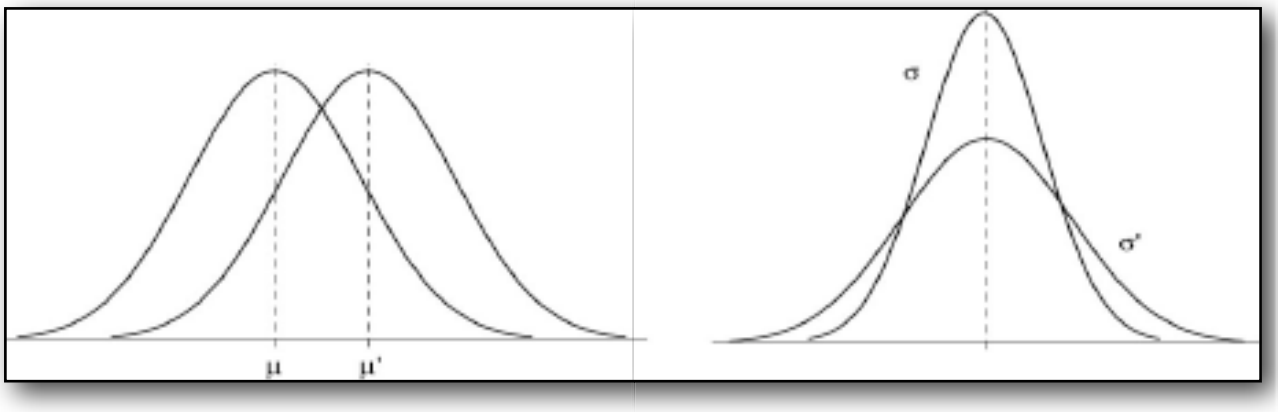
Pour surmonter ce problème, il est possible de faire subir des transformations linéaires aux variables aléatoires de telle sorte qu'elles deviennent comparables les unes aux autres. Pour mieux comprendre le problème, imaginons simplement que l'on s'intéresse non seulement à la taille des individus, mais également à leur poids. Le problème qui émerge est que la taille est exprimée en centimètres et le poids en kilos. Comment, dès lors, comparer des distributions si différentes?

En fait, deux problèmes surviennent lorsque différentes variables sont normalement distribuées mais que les paramètres de la distribution diffèrent : le décalage par rapport à la moyenne (Figure 7.5.a) et le décalage entre les variances (Figure 7.5.b). Pour y remédier, il est courant d'utiliser une distribution de moyenne $\mu = 0$ et de variance $\sigma^2 = 1$. C'est cette distribution que l'on nomme "*standard*", c'est-à-dire $N(0,1)$. Une variable ainsi standardisée prend le nom de "*score Z de la variable*".

Figure 7.5. : Courbes de probabilités normales

(a) Variances identiques, moyennes différentes

(b) Variances différentes, moyennes identiques



Pour standardiser une distribution donnée, il est donc nécessaire, dans un premier temps, de corriger la moyenne de la série statistique. Il suffit pour se faire d'ôter la moyenne à chacune des valeurs de la série. En effet, reprenons la série statistique du Tableau 6.1 illustré dans le Tableau 7.6 ci-dessous, il s'agissait des notes fictives à l'examen d'ANAD1 de 9 sujets.

La moyenne était de 5/10 et l'écart-type de 1,22. La partie (b) du tableau montre comment, en ôtant la moyenne à chaque sujet, la moyenne finale devient nulle. On dit que l'on a **centré la variable** après avoir effectué cette transformation. Remarquez qu'elle ne change rien à la forme de la courbe. Par défaut, lorsqu'on dit qu'on centre une variable, on sous-entend que c'est sur une moyenne nulle. Cependant, on pourrait avoir besoin de la centrer sur une quelconque autre valeur. En fait, centrer une variable veut juste dire qu'on ajuste la moyenne à une valeur qui nous intéresse. Par exemple, imaginons que les notes de travaux pratiques de deux groupes soient distribuées de manière semblable. Cependant, certains ont eu un assistant plus sévère et d'autres un assistant plus clément de sorte que la moyenne du premier groupe soit de 10/20 et la moyenne de l'autre soit de 14/20. Je pourrais décider de corriger cette injustice en décidant de translater la distribution des deux groupes, un vers une moyenne de 12. Je rajouterais deux points aux étudiants du groupe 1 et j'enlèverais deux points aux étudiants du groupe 2 (en espérant que personne n'ait plus que 18 dans le groupe 1 ni moins de 2 dans le groupe 2).

Maintenant que j'ai résolu le problème de ma moyenne, il me reste à résoudre le problème des écart-types différents (Figure 7.5.b). Pour régler ce problème, je peux diviser chacune des valeurs de ma série par l'écart-type de la distribution. La conséquence est que l'écart-type final (et la variance) sera égal à 1. C'est ce que j'ai fait dans la partie (c) du Tableau 7.6. En faisant cette opération on dit qu'on **réduit une variable**. Encore une fois, par défaut réduire signifie ramener l'écart-type à 1, mais on pourrait le réduire à une autre valeur en fonction de nos besoins.

Une variable standardisée se nomme *z* et correspond donc à une variable centrée à une moyenne nulle et réduite à un écart-type de un. L'avantage de cette double transformation linéaire[26] est qu'elle permet d'obtenir une distribution centrée sur zéro et dont l'unité de mesure est l'écart-type. Donc une valeur sur l'axe des *x* de 2,5 correspond à une distance de 2,5 écarts-types à droite de la moyenne. Pareillement pour une distance négative, par exemple -2,5, mais dans ce cas, on se trouve à gauche de la moyenne. Nous allons voir les applications qui en découlent.

[26] Les transformations linéaires sont de deux types : (a) la multiplication (ou la division) par une constante ; (b) l'addition (ou la soustraction) d'une constante. Les transformations linéaires ne changent pas la forme d'une courbe (ce sera toujours une normale si on part d'une normale), mais en modifient les paramètres.

Variable centrée-réduite ou Score *z*

$$z = \frac{(X_i - \overline{X})}{S}$$

Tableau 7.6. : (a) Série statistique de n = 9 représentant les cotes sur 10 à un examen ; (b) même série centrée et ; (c) centrée-réduite (score *z*).

(a) Série (b) Série centrée (c) Série *Z* centrée-réduite

Num	Cotes (X_i)	Variable centrée $(X_i - \overline{X})$	Variable centrée-réduite $z = (X_i - \overline{X})/S$
1	3	3-5 = -2	(3-5)/1,22
2	5	5-5 = 0	(5-5)/1,22
3	7	7-5 = 2	(7-5)/1,22
4	4	4-5 = -1	(4-5)/1,22
5	5	5-5 = 0	(5-5)/1,22
6	6	6-5 = 1	(6-5)/1,22
7	5	5-5 = 0	(5-5)/1,22
8	6	6-5 = 1	(6-5)/1,22
9	4	4-5 = -1	(4-5)/1,22
$\overline{X}$ =	**5**	**0**	**0**
S =	**1,22**	**1,22**	**1**

7.4.4. Utilisations de la table normale standard

7.4.4.1. Lecture simple de la table

Le Tableau 7.7 montre une table normale standard. Vous remarquerez que seules les valeurs comprises entre 0 et 3,90 sont envisagées. Les valeurs négatives ne sont pas reprises et les valeurs dépassant 3,90 non plus. Regardons d'abord les informations qui sont présentes. Si vous vous référez à la représentation graphique au-dessus de la table, vous constaterez que l'information fournie concerne la densité de probabilité entre le début de la courbe (-∞) et la valeur *z* considérée. C'est ce qu'indique l'intégrale de l'équation en tête de table : l'équation de la loi normale est intégrée sur l'intervalle allant de -∞ à *z*, ce qui donne l'aire sous la courbe normale, donc la densité de probabilité. Par exemple, si on regarde la

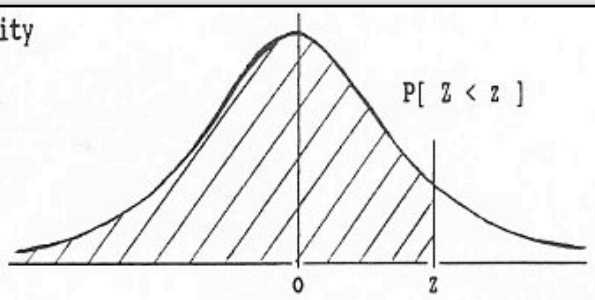
première valeur (z = 0,0), on obtient une densité de probabilité de 0,5000. Le fait qu'il y ait trois zéros après le 5 signale que les probabilités sont données avec une précision de 4 chiffres après la virgule. La valeur de 0,5 est évidente puisque lorsque la moyenne est nulle nous sommes exactement au milieu de la distribution et que cette distribution est parfaitement symétrique. La probabilité qu'un score se trouve entre -∞ et zéro est donc bien d'une chance sur deux, donc de 0,5. Si l'on progresse de un écart-type à droite, on obtient une probabilité de 84,13% (pour *z* = 1,0 la densité de probabilité vaut 0,8413).

Il est inutile de recueillir l'information au-delà de trois écarts-types. En effet, la courbe tend vers zéro et la densité de probabilité n'évolue plus de manière sensible au-delà de cette valeur (elle n'évolue que de manière infinitésimale).

Tableau 7.7. : Table de la distribution normale standardisée. Récupérée sur http://willw9.blogspot.com/2011/06/solace-in-mathematics-how-basic.html le 16-10-2011.

The table gives the cumulative probability up to the standardised normal value z

i.e. $P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}Z^2)\ dZ$



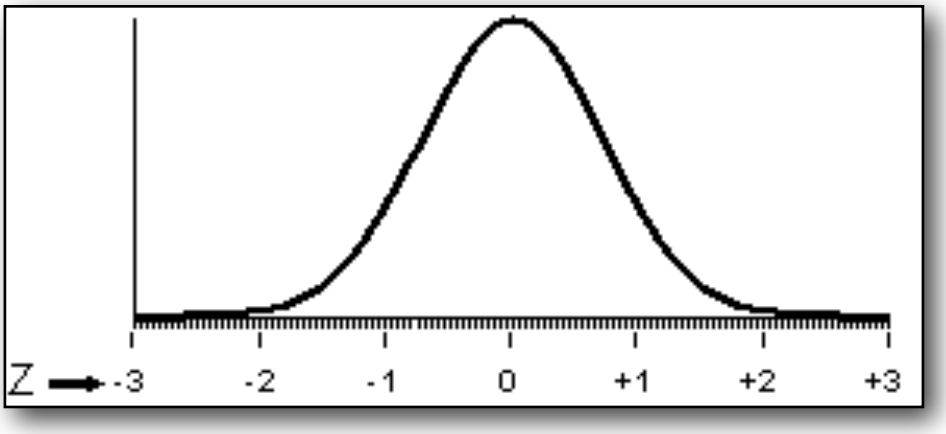
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

7.4.4.2. Utilisation de la table pour les valeurs négatives

Si l'on se réfère à la partie gauche de la courbe, c'est-à-dire aux valeurs négatives de z, aucune information concernant la densité de probabilité n'est donnée par la table. Par exemple, chercher la densité de probabilité à -0,5 écart-type de la moyenne n'est pas directement accessible. Cependant, profitons à nouveau du fait que la distribution soit parfaitement symétrique et que la probabilité totale est égale à 1. Dès lors, il nous suffit de regarder la densité de la valeur positive et d'en prendre la portion résiduelle : la table nous informe que pour z = 0,5 la densité entre -∞ et 0,5 est de 0,6915. Cela signifie que la densité de probabilité entre 0,5 et +∞ est de 1-0,6915 = 0,3085. Par symétrie, cette densité de probabilité correspond à celle de l'intervalle allant de -∞ à -0,5 que nous cherchons.

Comme exercice, représentez cette situation graphiquement. Sur la Figure 7.6.a, Représentez z = 0,5 et z = -0,5. Grisez la zone comprise entre 0,5 et +∞. Grisez la zone comprise entre -∞ et -0,5. Vous devriez vous rendre compte que ces zones sont de surfaces identiques.

Figure 7.6.a. : Silhouette loi normale



7.4.4.3. Utilisation de la table pour déterminer la densité de probabilité d'un intervalle

a) Détermination de la densité de probabilité pour un intervalle donné

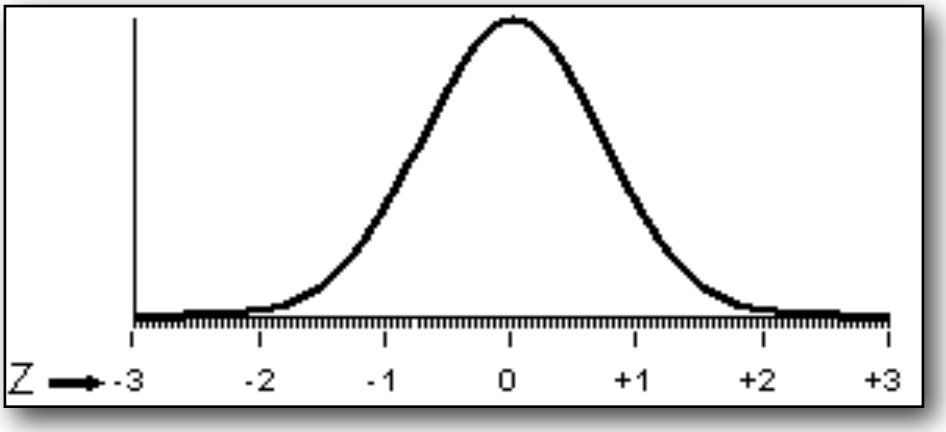
Imaginons que nous désirions, comme ce sera souvent le cas ultérieurement, déterminer la densité de probabilité d'un intervalle, mettons de un écart-type à gauche et à droite de la

moyenne. Cela revient à chercher la densité de probabilité de l'intervalle -1 ≤ *z* ≤ 1. Nous avons vu que cette densité est de 0,8413 pour l'intervalle -∞ ≤ *z* ≤ 1. Il nous reste donc à ôter de cette densité la densité correspondant à l'intervalle -∞ ≤ *z* ≤ -1. Comme nous l'avons vu au point précédent, cette densité correspondant à une valeur négative de *z* se calcule en calculant la densité résiduelle de la valeur positive, soit 1-8413 = 0,1587. Cette valeur doit maintenant être enlevée de la densité de probabilité de -∞ ≤ *z* ≤ 1 : 0,8413-0,1587 = 0,6826. Nous pouvons donc dire que 68,26% des scores *z* sont compris entre -1 et 1 écart-type autour de la moyenne dans une distribution normale. C'est ce qu'illustrait la Figure 7.4.

A titre d'exercice, je vous propose de confirmer les autres informations de la Figure 7.4. A savoir la probabilité de se trouver à ± 2 écarts-types de la moyenne est de 95% (plus exactement 0,9546) et à ± 3 écarts-types de la moyenne est de 99% (0,9972).

Vous pouvez à nouveau effectuer une représentation graphique. Représentez sur la Figure 7.6.b z = 1 et z = -1. Grisez la zone allant de -∞ à 1. Grisez la zone allant de -∞ à -1, qui chevauchera partiellement la zone précédente. Vous vous rendrez compte de deux choses : (a) la zone grisée deux fois correspond à la partie que vous devez ôter de la première zone grisée ; (b) la zone grisée deux fois a une surface identique à la surface de la courbe non grisée.

Figure 7.6.b. : Silhouette loi normale



b) Détermination d'un intervalle pour une densité de probabilité donnée

Vous pouvez également fonctionner dans l'autre sens. Par exemple, demandons-nous quelle est la valeur de *z* qui nous permette d'obtenir exactement 95% des scores de la variable aléatoire distribuée normalement (et non 95,46) de part et d'autre de la moyenne. Cela revient à désirer connaître la valeur de *z* qui contient une probabilité résiduelle de 2,5% à droite de la courbe. De cette manière, en prenant la valeur symétrique, j'aurai également 2,5% de probabilité résiduelle de l'autre côté de la courbe et mon intervalle contiendra bien 95% des scores. La valeur $z = 1{,}96$ correspond à une densité de probabilité de 0,9750. Cela signifie qu'il reste bien une probabilité de 0,0250 (= 1-0,9750) soit 2,5%. La valeur symétrique est donc $z = -1{,}96$ qui laisse 0,0250 entre $-\infty$ et -1,96. Dès lors l'intervalle compris entre $-1{,}96 \leq z \leq 1{,}96$ contient bien exactement 95% des scores.

Je vous suggère de réaliser les exercices interactifs proposé par le site http://homeomath.imingo.net/interactifs144.htm. Aidez-vous du Tableau 7.7 pour répondre aux questions.

7.5. Application de la loi normale aux erreurs aléatoires du modèle

7.5.1. Introduction

Une application tout à fait fondamentale de la distribution normale se retrouve dans la gestion de l'erreur due à la modélisation. Vous aurez maintenant bien compris que lorsqu'on modélise la réalité, on fait une erreur. Cette erreur dépend de plusieurs choses : les différences individuelles, la mesure, l'abandon de diverses variables potentiellement influentes, etc. Par exemple, lorsque je prétends que les femmes adultes belges mesurent 169 cm, nous avons vu que certaines peuvent être plus grandes, d'autres plus petites. Les raisons sont diverses : j'ai pu me tromper en mesurant, j'ai pu mal encoder les valeurs, je n'ai pas tenu compte de leur poids ni de leur origine (et j'aurais peut-être dû, par exemple si certaines sont de descendance norvégienne alors que d'autres sont de descendance pygmée), certaines ont été nourries différemment des autres lors de leur enfance, le système endocrinien a ses spécificités individuelles que je ne connais pas, etc..

Supposons maintenant que j'envisage de construire une chaise de taille spécialement adaptée aux femmes belges à un prix raisonnable. Mais j'hésite entre un rembourrage en copeaux de latex ou en écales de sarrasin. Je suis justement en train de mesurer la taille de mes 100 sujets féminins et j'en profite pour les faire patienter aléatoirement sur des chaises en copeaux ou en écales. Si la répartition est aléatoire, il y a autant de chances que de grandes femmes s'assoient sur des copeaux que sur les écales et pareillement pour chacune des tailles. La distribution de mon erreur est donc identique dans les deux conditions et suit une loi normale. Ce n'est que dans cette condition que je suis en droit de comparer la préférence des femmes pour l'un ou l'autre rembourrage.

En effet, si pour une raison ou une autre je perds ce caractère aléatoire, mes conclusions perdent leur validité. Par exemple en décidant que les femmes flamandes s'assoient sur les chaises en latex, les femmes wallonnes sur les chaises en écales et les femmes bruxelloises restent debout, je pourrais biaiser les résultats : imaginons que les femmes flamandes soient plus grandes que les femmes wallonnes et mon expérience est ratée. Bien sûr, il se peut que, même en distribuant aléatoirement les sujets au sein des conditions, le hasard fasse qu'un groupe soit de taille moyenne supérieure à l'autre. Mais d'une part il est peu probable que cette différence soit forte, d'autre part, la probabilité que cela arrive est identique quelque soit la condition et n'est donc pas systématique (d'où le terme d'erreur aléatoire).

Ce problème est encore plus prégnant en sciences, notamment en physique, où les mesures sont sensées être extrêmement précises. Au XVIIIe et XIXe siècles, il était déjà fréquent qu'un même observatoire astronomique ou des observatoires différents se livrent à des mesures répétées, par exemple, sur la position de la Lune, afin de déterminer le plus précisément possible l'équation de son orbite et de ses mouvements. Par exemple, l'astronome Johann Tobias Mayer (1723-1762) avait effectué 27 relevés de la position du cratère *Manilius* au cours d'une période allant du 11 avril 1748 au 4 mars 1749. Toutes ces observations avaient été faites avec le même télescope, dont Mayer pouvait penser que la précision était restée constante au cours de la période d'observation. Ceci lui donnait 27 équations différentes à 3 inconnues. Or, sans erreur de mesure, il suffit de trois équations pour déterminer les valeurs de trois inconnues. Comment les choisir parmi les 27 ? Ou alors, faut-il calculer les moyennes sur trois groupes de neuf équations et utiliser les trois équations résultantes pour dériver les valeurs des trois inconnues ? Mais alors comment répartir les équations entre les trois groupes ? Sans entrer dans les détails de sa procédure,

disons que Mayer a effectivement réparti les équations en trois groupes avant de déterminer les valeurs des paramètres.

Un problème analogue s'est posé Leonhard Euler (1707-1783) en 1749 dans son analyse des anomalies dans les trajectoires de Jupiter et Saturne. Euler disposait de 75 observations réparties sur une période allant de 1582 à 1745. Ceci lui donnait 75 équations à 8 inconnues. Or, seulement huit équations permettent de dériver la valeur des huit inconnues. Cependant, Euler a renoncé à traiter complètement le problème parce qu'il ne savait pas comment les choisir ou les regrouper. Il était préoccupé par le fait que ces observations réparties sur plus de 160 ans variaient sans doute très fort dans leur degré de précision et il craignait que les grandes erreurs de mesure s'additionnent.

La solution au problème de la variabilité des mesures qui s'est imposée entre 1805 et 1809 est due à la combinaison des travaux de Adrien Marie Legendre (1752-1833), Pierre Simon Laplace (1749-1827) et Carl Friedrich Gauss (1777-1855). Nous avons déjà évoqué la solution de Legendre quand on a parlé de l'analogie entre la moyenne et le centre de gravité au chapitre 6 et de l'invention de la méthode des moindres carrés. Cependant, il n'y a aucune notion probabiliste chez Legendre. Or, plusieurs auteurs, dont Laplace, avaient déjà eu l'intuition que les erreurs aléatoires de taille importante étaient sans doute moins fréquentes que les erreurs aléatoires de petite taille et que la notion d'erreurs aléatoires implique une distribution symétrique de leurs valeurs. Tout ceci a conduit Gauss a postuler que les erreurs aléatoires de mesure sont distribuées normalement parce qu'elles résultent de l'action d'un grand nombre de causes impossibles à identifier, chaque cause d'erreur contribuant pour une petite part indépendante dans la détermination de l'erreur aléatoire résultante lors de chaque mesure. La solution que Gauss a proposé en 1809 pour gérer le problème des erreurs aléatoires de mesure reste d'actualité.

7.5.2. Modèle de l'erreur aléatoire de mesure

Comme nous l'avons vu, lorsque l'on modélise la réalité, nous pouvons établir la relation suivante : Réalité (mesure) = Modèle (estimation) + erreur. L'erreur est donc une erreur que nous devons considérer comme aléatoire (parce qu'elle dépend d'un tel nombre de variables qu'il nous est, probablement à jamais, impossible de la déterminer) et nous devons veiller à ce qu'elle reste aléatoire (c'est-à-dire veiller à ce que chaque erreur soit indépendante des

autres et qu'il n'y ait pas de biais systématique). Lors de chaque mesure, le modèle et l'erreur sont inobservables. Cependant, lors de la répétition de la mesure, la valeur vraie du modèle (par exemple la moyenne de la population) reste inchangée. Seule l'erreur varie, et elle seule induit une variation de la mesure.

Les erreurs aléatoires de mesure se distribuent selon la loi normale avec une moyenne de 0 et un écart-type à déterminer que l'on nomme l'**erreur standard de la mesure**. Retenez bien ce terme, et visualisez bien ce qu'il représente, il sera essentiel dans toute l'inférence statistique. L'erreur standard de la mesure constitue en quelque sorte l'unité qui indique l'ampleur de l'erreur aléatoire. Rappelez-vous que la forme normale de la distribution implique que 68,26% d'entre elles se situent à ± 1 écart-type de zéro, 95,46% à ± 2 écarts-type de zéro et 99,72% à ± 3 écarts-types.

7.5.3. Application à l'élimination des données jugées aberrantes

Rappelez-vous que, lorsqu'une valeur est très éloignée du reste des valeurs d'une distribution, on a tendance, légitimement, à l'ôter de la série. L'étude des boîtes à moustaches nous a montré comment prendre une telle décision graphiquement en fonction des quantiles. La répartition normale de l'erreur peut maintenant justifier un peu mieux ce comportement et permettre de prendre une décision dans le cas d'une variable aléatoire continue. Voyons ce que l'Histoire peut nous apprendre sur ce sujet.

Le kilo étalon actuel date de 1889. C'est la seule unité de mesure qui soit encore définie par un étalon physique déposé au pavillon de Sèvres. Une masse de 1 kg est une masse qui serait en équilibre avec le kilo étalon, à condition que la balance soit parfaite. Donc, pour savoir si le poids de 1 kg utilisé par votre maraîcher est conforme, il faudrait le lui subtiliser et aller le comparer au prototype de Sèvres (ou dérober celui de Sèvres et aller chez votre maraîcher, mais c'est encore un petit peu plus compliqué). En pratique, des copies du prototype sont diffusées un peu partout dans le monde et c'est aussi le cas pour des copies de fractions du kilo. Dans l'exemple qui suit, il est question de la copie du poids de 10 grammes acquise par le National Bureau of Standards de Washington vers 1940. Depuis lors, ce poids de 10 grammes appelé NB10 est pesé environ une fois par semaine sur la même balance et dans des conditions approximativement identiques de pression atmosphérique et de température (pendant que des enfants meurent de faim, je sais, mais ce n'est pas si inutile que ça peut en

avoir l'air). La raison de cette pesée hebdomadaire est d'obtenir une estimation de l'ampleur de l'erreur aléatoire de mesure qui entache chaque pesée particulière.

Tableau 7.8. : Erreurs en microgrammes sous 10 grammes pour 100 pesées successives du poids NB 10 (environ une pesée par semaine en 1962-1963).

Num	Erreur	Num	Erreur	Num	Erreur	Num	Erreur
1	409	26	397	51	404	75	408
2	400	27	407	52	406	76	404
3	406	28	401	53	407	77	401
4	399	29	399	54	405	78	404
5	402	30	401	55	411	79	408
6	406	31	403	56	410	80	406
7	401	32	400	57	410	81	408
8	403	33	410	58	410	82	406
9	401	34	401	59	401	83	401
10	403	35	407	60	402	84	412
11	398	36	423	61	404	85	393
12	403	37	406	62	405	86	437
13	407	38	406	63	392	87	418
14	402	39	402	64	407	88	415
15	401	40	405	65	406	89	404
16	399	41	405	66	404	90	401
17	400	42	409	67	403	91	401
18	401	43	399	68	408	92	407
19	405	44	402	69	404	93	412
20	402	45	407	70	407	94	375
21	408	46	406	71	412	95	409
22	399	47	413	72	406	96	406
23	399	48	409	73	409	97	398
24	402	49	404	74	400	98	406
25	399	50	402	75	408	99	403

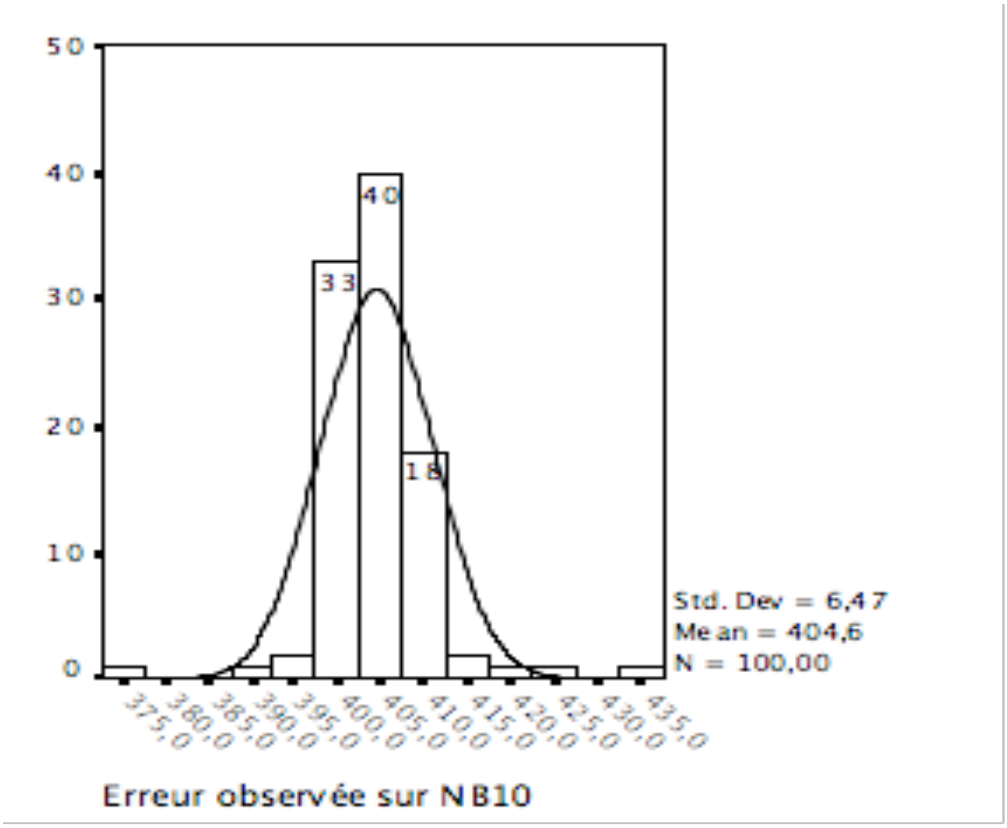
Les données traitées ici concernent 99 pesées de NB10 effectuées en 1962-1963[27]. Toutes les mesures exprimées en grammes sont inférieures à dix grammes et elles sont toutes identiques jusqu'à la 3ème décimale (donc chaque pesée donne 9,999 grammes). Pour les trois décimales suivantes, les mesures varient. Par exemple, les deux premières mesures sont de 9,999591 et 9,999600 grammes. Il manque donc 0,000409 et 0,000400 gramme pour faire

[27] Source: Freedman, D., Pisani, R, & Purves, R. (1998, 3ème édition). *Statistics* (chapitre 6, pp. 97-109).

10 grammes. Les données du Tableau 7.8 constituent des écarts par rapport à 10 grammes exprimés en microgrammes (millionième de gramme) plutôt qu'en grammes de manière à supprimer les décimales. Dans ce tableau, les deux premières données apparaissent donc comme 409 et 400 microgrammes.

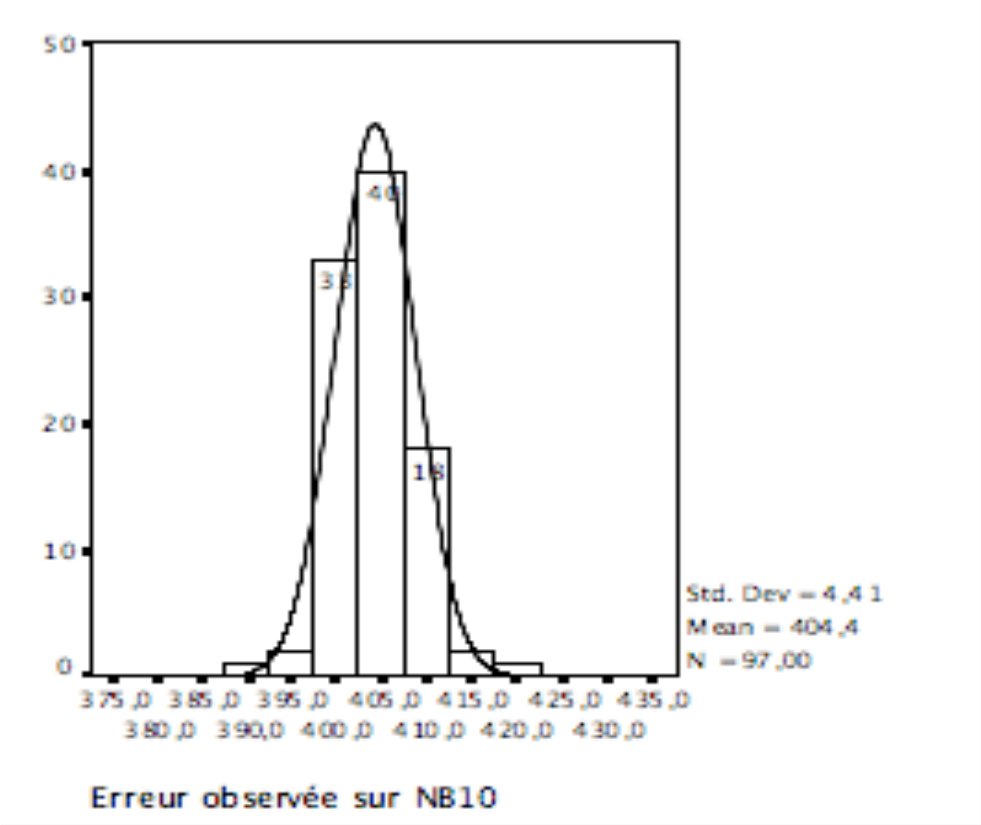
La Figure 7.7 est un histogramme représentant les mêmes données que le Tableau 7.8 (et d'une valeur supplémentaire parce que la base de données originale comprenait 100 mesures mais qu'une a été escamotée pour présenter le Tableau 7.8 de manière économique). Il s'agit de la figure fournie par le logiciel *SPSS* à laquelle aucune modification n'a été apportée si ce n'est de faire inscrire l'effectif des classes au sommet des rectangles. Notez que comme il s'agit d'un histogramme, toutes les classes possibles, depuis celle qui est centrée sur 375 microgrammes jusqu'à celle qui est centrée sur 435 microgrammes sont et doivent être représentées, même quand elles ne contiennent aucune données.

Figure 7.7. : Histogramme des données du Tableau 7.8 tel qu'il est proposé par *SPSS* avec 13 classes de 5 microgrammes de largeur.



A l'extérieur, à droite de la Figure 7.7, on trouve trois indications : la moyenne de 404,6 microgrammes, l'écart-type (corrigé) de 6,47 microgrammes et l'effectif *n* de 100 pesées. La courbe normale superposée à l'histogramme est de même moyenne et de même écart-type. Si on retranche et qu'on ajoute trois fois l'écart-type (3 x 6,47) à la moyenne de 404,6 microgrammes, on obtient, en arrondissant, 385 et 424 microgrammes. Si les données épousaient exactement la distribution normale, les valeurs de 375, 437 et même 424 microgrammes apparaîtraient comme extrêmement improbables, puisqu'à trois écarts-types on retrouve pour ainsi dire tous les scores de la variable. Cela justifie qu'on puisse les considérer comme aberrantes par rapport aux autres et les éliminer. C'est ce qu'on a fait à la Figure 7.8. En conséquence, vous pouvez remarquer que la courbe normale épouse beaucoup mieux nos données. Cela revient à dire que notre modèle est plus précis, et donc qu'on a intérêt à ôter ces valeurs.

Figure 7.8. : Histogramme des données du Tableau 7.8 après retrait de trois valeurs extrêmes (jugées aberrantes) avec 13 classes de 5 microgrammes de largeur comme à la Figure 7.7.



7.5.4. Conséquences du modèle de l'erreur aléatoire de mesure

En calculant la moyenne d'un grand nombre de mesures effectuées sur un même objet, les erreurs aléatoires tendent à s'annuler et la moyenne des mesures observées tend vers la vraie valeur de la mesure (et ce d'autant plus que le nombre de mesures est plus grand). Dans notre exemple basé sur 100 pesées de NB10, on peut dire que la moyenne de 404,6 microgrammes est sans doute très proche de la vraie valeur de ***l'erreur systématique*** de la mesure de NB10 effectuée avec la balance en question dans les années 1962-1963. Cette estimation de l'erreur systématique est entachée d'erreurs aléatoires de mesure dont l'ampleur – l'erreur standard de la mesure - est estimée à 6,47 microgrammes.

En idéalisant, c'est-à-dire en se référant à la distribution normale de moyenne 0 et d'écart-type 6,47, on peut affirmer que pour chaque pesée individuelle de NB10, il y a 68,26 % de chances pour que l'erreur aléatoire de mesure se situe entre –6,47 et +6,47 microgrammes (± 1 écart-type), ou 95,46 % de chances pour qu'elle se situe entre –12,94 et +12,94 microgrammes (± 2 écarts-types), ou encore qu'il y ait 99,72 % de chances qu'elle se situe entre -19,41 et +19,41 microgrammes (± 3 écarts-types). Une manière alternative de présenter les choses est de se référer à l'estimation de 404,6 microgrammes de l'erreur systématique de mesure. On peut alors affirmer avec respectivement 68,26 ; 95,46 et 99,72 % de chances d'avoir raison que les intervalles 398,13-411,07 ; 391,66-417,54 et 385,19-424,01 contiennent la vraie valeur de l'erreur systématique de la pesée de NB10 dans les conditions spécifiées plus haut.

Il est rare qu'on ne désire pas se débarrasser d'un petit nombre de données que l'on juge aberrantes de manière à obtenir des estimations encore plus précises des paramètres auxquels on s'intéresse. L'utilisation de la distribution normale à cette fin permet de définir un critère objectif d'élimination de données. Dans l'exemple de la pesée de NB10, le critère d'élimination a été très strict puisque seules trois données de probabilités extrêmement faibles ont été considérées comme aberrantes. Notez que la décision de qualifier ces données d'« aberrantes » est elle-même une décision prise dans l'incertitude. Des données aussi éloignées de la moyenne pourraient consister en de véritables erreurs aléatoires de mesure de probabilités extrêmement faibles. On court donc un petit risque chiffrable en les considérant d'une autre nature (donc, comme non aléatoires). La prise de ce petit risque valait la peine puisqu'elle provoque une réduction de l'erreur standard de la mesure d'environ 1/3 (4,41 au lieu de 6,47 microgrammes).

Si nous analysons d'un peu plus près cette erreur de mesure systématique (les 404 microgrammes qui manquent), cinq possibilités peuvent expliquer cet état : (a) soit le poids de cuivre est trop faible et on devrait rajouter les 404 microgrammes manquant ; (b) soit la balance est imprécise et sous-estime le poids de 404 microgrammes ; (c) soit la masse de cuivre est insuffisante et la balance sous-estime le poids ; (d) soit la masse de cuivre est excessive et la balance sous-estime le poids de manière importante ; (e) soit la masse de cuivre est nettement insuffisante et la balance surestime le poids. Il n'y a aucun moyen de déterminer la possibilité correcte (bien que les deux dernières soient moins probables que les trois premières). On peut néanmoins s'aider d'approches différentes. Par exemple, on pourrait mesurer le volume exact de cuivre et, connaissant la masse volumique, en établir la masse exacte. Malheureusement, les appareils de mesure d'un tel volume risquent fort d'être encore plus imprécis que la balance. On pourrait également titrer le cuivre à l'aide d'une pile en utilisant la masse comme électrode, mais on ne pourra alors que savoir après coup combien pesait la masse de cuivre. En somme, on risque fort de rester dans l'incertitude jusqu'à la fin de nos jours.

7.6. Applications de la loi normale à des mesures physiques et psychologiques

C'est avec un certain étonnement qu'on s'est progressivement aperçu que la distribution de mesures physiques comme la taille ou le poids adopte des formes approximativement normales. L'étonnement provenait du fait que la courbe de Gauss était universellement connue sous le nom de *courbe de l'erreur de mesure* et son utilisation exclusivement réservée à l'élimination des données jugées aberrantes, comme expliqué à la section précédente. Pourquoi une mesure comme la taille est-elle distribuée comme une erreur de mesure? Cette question a surtout hanté Quételet (1796-1874), un mathématicien belge d'abord connu pour ses travaux en astronomie et en météorologie (il est le fondateur de l'Observatoire Royal de Bruxelles à Uccle) et ensuite pour ses travaux en statistique.

Comme principal instigateur d'une discipline appelée ***statistique sociale***, Quételet était surtout intéressé par la stabilité et la prévisibilité des phénomènes sociaux de masse, comme le taux de suicide, le taux de mortalité ou de mariage, le taux de criminalité. Par exemple, le taux de mortalité dans une ville comme Paris diffère en fonction des tranches d'âge. Mais, pour chaque tranche d'âge, il tend à rester très stable d'une année à l'autre. La prévisibilité de ces phénomènes de masse contraste avec l'imprévisibilité des comportements individuels.

Concernant des mesures comme la taille, Quételet a forgé le concept ***d'homme moyen*** en se calquant sur le modèle de l'erreur de mesure. Chaque personne est censée réaliser un certain ***type*** sous-jacent propre à la population à laquelle elle appartient. Les raisons pour lesquelles les personnes diffèrent en taille sont dues aux aléas de la vie dont les multiples influences sont comparables aux erreurs aléatoires de mesure. Nous en avons déjà parlé.

La théorie de l'homme moyen n'a cependant jamais été très convaincante ; elle n'a joué qu'un rôle anecdotique dans l'histoire de la discipline. En revanche, l'importance de l'utilisation de la courbe de Gauss dans l'étude des phénomènes sociaux et psychologiques n'a cessé de croître. Les Figures 7.9.a,b,c en sont des illustrations flagrantes. On y voit trois distributions empiriques qui épousent des formes approximativement normales. La Figure 7.9.a montre la distribution de la taille en pouces d'adultes masculins, la Figure 7.9.b montre la distribution de QI au Stanford-Binet et la Figure 7.9.c montre les scores à un test mesurant un trait de personnalité (sur une échelle de Likert à 5 points) du Big Five[28] (en l'occurrence, l'agréabilité).

[28] Le Big Five est un modèle à cinq facteurs (Neuroticisme, agréabilité, extraversion, conscience et ouverture à l'expérience) très utilisé actuellement, établi par Fiske (1949) et validé par McCrae et Costa (1987).

Figure 7.9.a. : Distribution de la taille des hommes adultes dans le monde et comparaison à une distribution normale. Source : http://www.askamathematician.com/2010/02/q-whats-so-special-about-the-gaussian-distribution-a-k-a-a-normal-distribution-or-bell-curve/ retrouvé le 23/10/11.

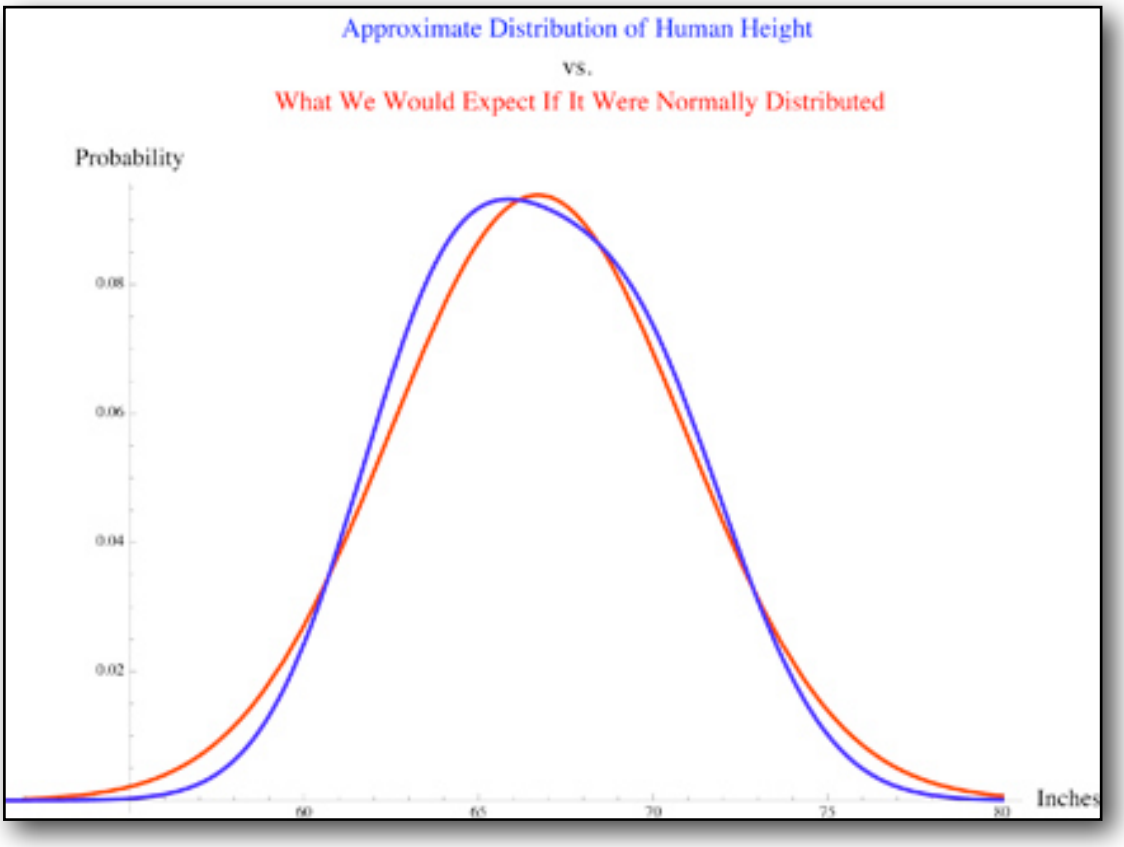


Figure 7.9.b. : Distribution du test de quotient intellectuel de Binet correspondant à une distribution normale. Source : http://www.mindsparke.com/iq.php?id=2 retrouvé le 23/10/11.

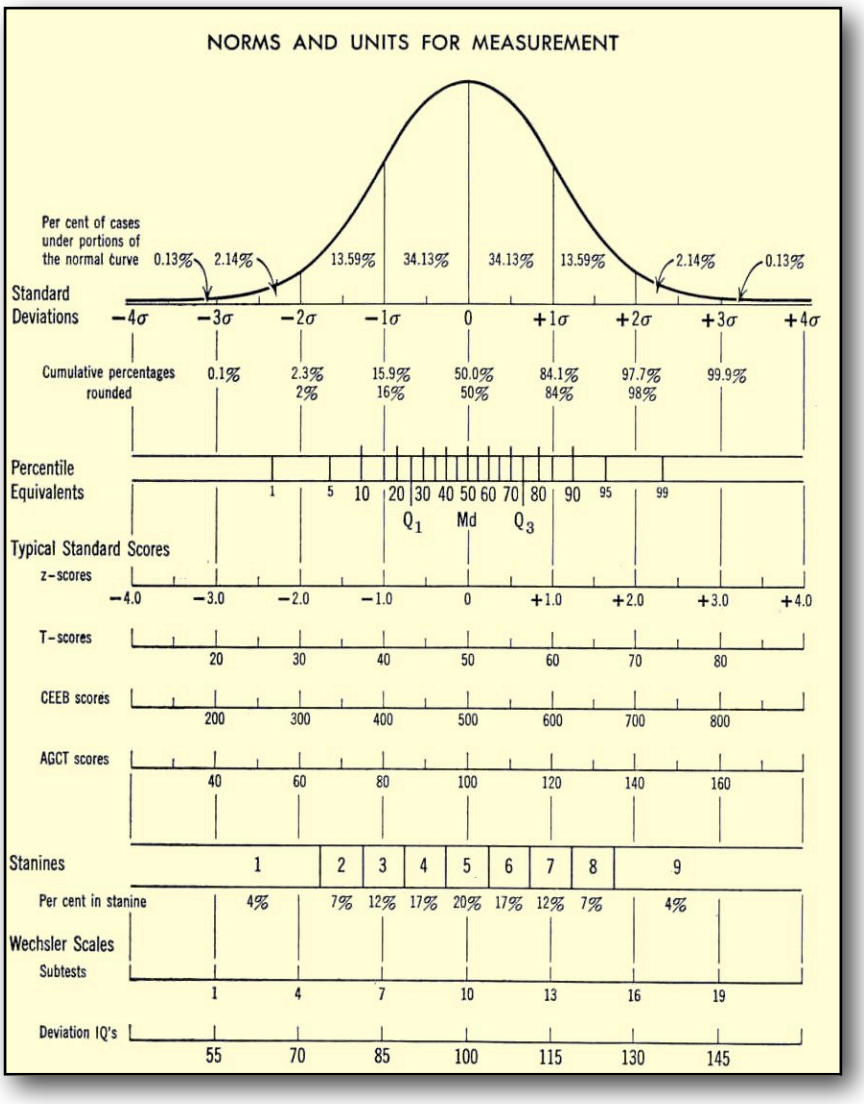
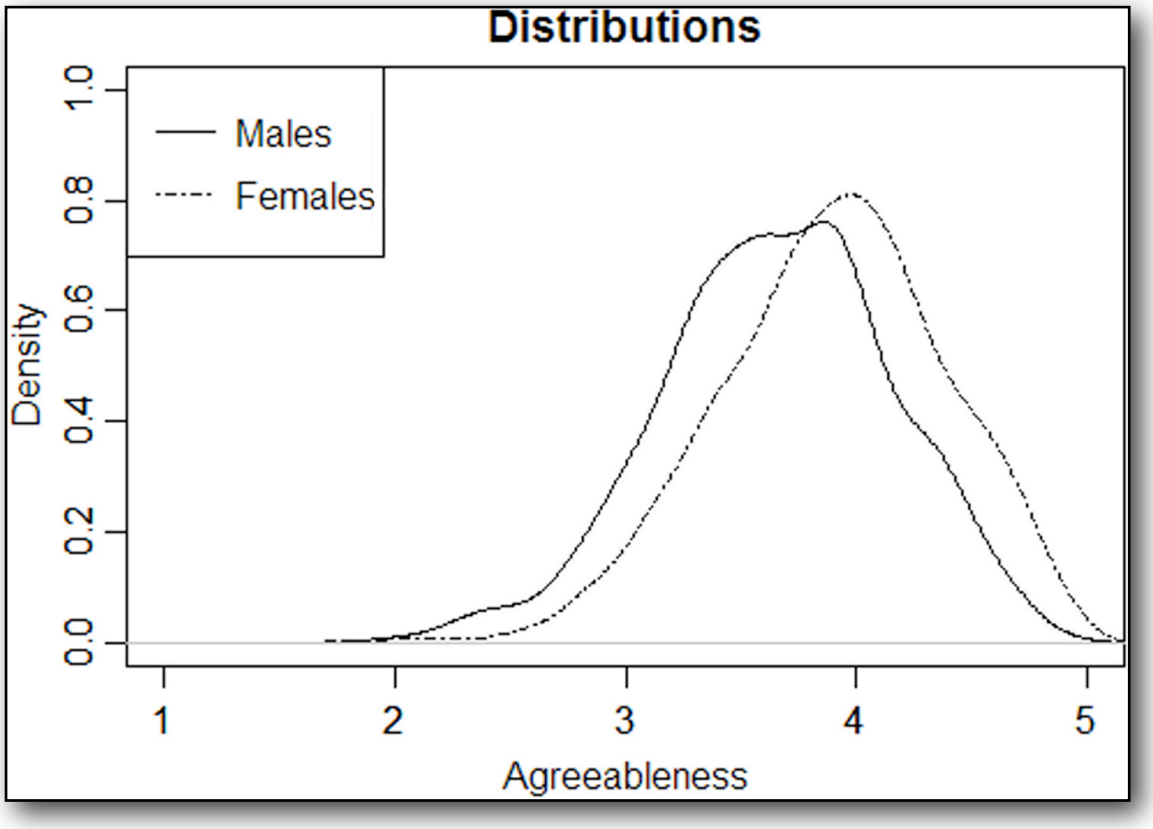


Figure 7.9.c. : Distribution du trait d'agréabilité du Big Five pour un échantillon masculin et un échantillon féminin. Source : http://www.frontiersin.org/personality_science_and_individual_differences/10.3389/fpsyg.2011.00178/full retrouvé le 23/10/11.



Un individu peut se situer, par exemple, 1 écart-type au-dessus de la moyenne en taille, ce qui le situe au rang percentile 84 (en arrondissant 84,13) pour la taille. Ce même individu pourrait avoir un QI de 130, ce qui le situerait 2 écarts-types au-dessus de la moyenne, donc au rang percentile 98 (en arrondissant 97,72) pour le QI.

Une conclusion intéressante est que, alors que cela n'a pas de sens de demander si un individu est plus intelligent que grand, la comparaison de la position ordinale relative qu'il occupe dans une population en termes d'intelligence et de taille ne nous paraît pas saugrenue du tout. La première question n'a pas de sens parce que l'intelligence et la taille ne sont pas mesurées par des unités comparables et que les contenus empiriques de ces deux mesures concernent des propriétés qui n'ont aucun rapport entre elles. En revanche, en envisageant les mesures du QI et de la taille dans leur aspect ordinal plutôt que numérique, la comparaison des positions relatives occupées par un individu dans une population est parfaitement légitime.

7.7. Conclusions

Au vu de ces chapitres, vous aurez compris la différence entre la distribution binomiale, réservée aux variables aléatoires discrètes et la distribution normale applicables aux variables aléatoires continues. Vous savez qu'il existe d'autres distributions que ces deux-là, mais que ce sont elles qui jouent un rôle majeur dans le domaine des statistiques qui nous occupe. Vous connaissez le lien qui permet d'estimer une binomiale par une normale lorsque le nombre de sujets devient trop important et que l'analyse combinatoire produit des nombres ingérables. Vous êtes également capables d'utiliser les tables des deux distributions pour déterminer la probabilité d'événements ou la densité de probabilité (et vous êtes conscients de la différence entre ces deux terminologies). Enfin, vous conceptualisez l'importante application de la loi normale aux erreurs de mesure propres au problème de modélisation dont nous parlons depuis le premier chapitre.

Le point suivant sera consacré aux exercices sur ces distributions. Le chapitre suivant envisagera la distribution d'échantillonnage à la lumière de ce que vous connaissez maintenant.

7.8. Exercices de fin de chapitre

T.P. 7 – Exercice 1
Distribution Binomiale

1. On lance une pièce 2 fois de suite. Quelle est la probabilité d'obtenir :
   - **1.1.** 0 face
   - **1.2.** 1 face
   - **1.3.** 2 faces

N.B. : Aidez-vous d'un dessin en « arbre » pour trouver la réponse.

2. Même exercice avec les formules adéquates.

3. Calculez de trois manières différentes, la moyenne, la variance et l'écart type de la distribution de probabilités des deux lancers de notre pièce équilibrée.Basez-vous :

   1/ sur la formule habituelle,

   2/ sur la formule simplifiée pour la binomiale basée sur les nombres de succès

   3/ sur la formule simplifiée pour la binomiale basée sur les proportions de succès.

T.P. 7 – Exercice 2
Approximation normale de la loi binomiale
et correction pour la continuité

Correction pour la continuité

Pour simplifier les calculs, il est possible de calculer une probabilité qui approche celle de la binomiale en se fondant sur une fonction continue qui donne une bonne approximation de la distribution binomiale discontinue. Cette distribution continue est connue sous le nom de distribution normale.

Il existe un critère, un peu rudimentaire, nous le verrons, qui consiste à accepter l'approximation normale de la binomiale à partir de N = 30.

Un autre critère, plus subtil, tenant compte de la forme de la distribution est proposé par Moore. Selon lui, il est nécessaire, pour pouvoir utiliser l'approximation de la binomiale par la normale, de se baser sur le critère suivant :

$$Np \geq 10 \text{ } \underline{\textbf{et}} \text{ } N(1-p) \geq 10.$$

N.B. : 1-p = q

1. Soit la distribution binomiale suivante :

		p
N	*x*	0.10
10	0	.3487
	1	.3874
	2	.1937
	3	.0574
	4	.0112
	5	.0015
	6	.0001
	7	.0000
	8	.0000
	9	.0000
	10	.0000

5. Peut-on utiliser l'approximation de la binomiale par la normale ?

Les formules de correction sont les suivantes :

Limite inférieure	Limite supérieure
$Z_{j\,\text{inf}} = \frac{(x_j - 0{,}5) - Np}{\sqrt{Np(1-p)}}$	$Z_{j\,\text{sup}} = \frac{(x_j + 0{,}5) - Np}{\sqrt{Np(1-p)}}$

N.B. :

Les formules de correction portent sur les valeurs des nombres (pas des proportions) de succès.

Les formules ci-dessus sont présentées avec une disposition des termes un peu différente par rapport au cours, en dehors de cela elles sont parfaitement identiques.

6. Qu'on soit arrivé à la conclusion que l'on peut utiliser l'approximation par la normale ou pas, calculez de deux manières différentes (en utilisant respectivement la binomiale et son approximation par la normale), la probabilité que le nombre de réussites soit compris entre 2 et 4 (2 et 4 inclus). Comparez les résultats et expliquez-en la différence.

T.P. 7 – Exercice 3

Considérations sur la binomiale

Les variables suivantes sont-elles binomiales ? Si « non », peuvent-elles êtres considérées comme telles ? Sous quelles conditions ? Transformez-les si possible et donnez-en les caractéristiques.

	Binom ?	Conditions et transformations éventuelles	Caractéristiques
1. Le lancer d'une pièce de monnaie non pipée			
2. Le lancer d'une pièce pipée			
3. Le lancer d'un dé non pipé			

4. La couleur des cheveux dans la population belge.			

T.P. 7 – Exercice 4

Binomiale

1. Dans une usine, on prélève un échantillon de 10 appareils pour un contrôle de qualité. Quelle est la probabilité qu'un ou moins de ces 10 appareils soit défectueux sachant que 10% des appareils produits dans cette usine sont défectueux ? Aidez-vous de la table 1 (ci-dessous) qui est la distribution (non cumulée) des probabilités pour $B(10,0.1)$. Tracez un histogramme de cette distribution de probabilités avec en abscisse le nombre de machines défectueuses possibles dans l'échantillon et en ordonnée les probabilités associées. Calculez la moyenne et la variance de cette distribution.

Table 1

		p
N	*x*	0.10
10	0	.3487
	1	.3874
	2	.1937
	3	.0574
	4	.0112
	5	.0015
	6	.0001
	7	.0000
	8	.0000
	9	.0000
	10	.0000

Réponse :

La probabilité qu'un appareil ou moins soit défectueux est de :

$\mu =$

$\sigma^2 =$

2. D'après certaines données, 25% des femmes au chômage n'ont jamais été mariées. On prélève un échantillon de 10 femmes chômeuses au hasard :
   2.1. Quelle est la probabilité qu'exactement 2 femmes n'aient jamais été mariées ? Aidez-vous de la table 2 (ci-dessous).
   2.2. Quelle est la probabilité que 2 ou moins n'aient jamais été mariées ?
   2.3. Quelle est la probabilité qu'au moins 8 ne soient pas mariées ?
   2.4. Quelle sont la moyenne et la variance de cette distribution ?

Table 2

		p
N	*x*	0.25
10	0	.05631
	1	.18771
	2	.28157
	3	.25028
	4	.14600
	5	.05840
	6	.01622
	7	.00309
	8	.00039
	9	.00003
	10	.00000

Réponse :

T.P. 7 – Exercice 5

Considérations sur la binomiale : approche intuitive pour mieux comprendre la formule

Pour cet exercice, nous nous situons dans le cadre d'une distribution binomiale, donnant la probabilité de tirer la boule 1 d'une urne qui contient 10 boules (des sphères de même taille et de même poids) numérotées de 1 à 10.

1. Dans ce contexte, combien d'événements élémentaires prend-on en considération ?

2. Dans ce contexte, que signifie N=1 ?

3. Pour N=1, combien y a-t-il d'événements possibles ? Quelles en sont les probabilités ?

4. Pour N=2, combien y a-t-il d'événements possibles ? Quelles en sont les probabilités ?

N.B. : Pour cet exercice, nous noterons E l'échec et R la réussite.

5. Que doit valoir la somme des probabilités calculées au point précédent ? Pourquoi ?

6. En termes ensemblistes, que pouvez-vous dire de ce que nous avons observé au point précédent ?

7. Pour N = 3, combien y a-t-il d'événements possibles ? Calculez-en les probabilités.

8. Comparez, sans faire de calculs, les résultats obtenus au point précédent à ceux que vous obtiendriez dans le cadre de trois lancers successifs d'une pièce de monnaie équilibrée.

9. Déterminez la distribution binomiale pour les deux exemples (celui des 10 boules et celui de la pièce) en complétant le tableau suivant :

Pour p=.1

Nombre X de réussites dans la combinaison	Combinaisons prises en compte	Nombre de combinaisons prises en compte	Calcul de la probabilité pour chaque combinaison	P(X)
0				
1				
2				
3				

Pour p=.5

Nombre X de réussites dans la combinaison	Combinaisons prises en compte	Nombre de combinaisons prises en compte	Calcul de la probabilité pour chaque combinaison	P(X)
0				
1				

2				
3				

Pour N=3, il nous a été aisé de déterminer le nombre de combinaisons à prendre en compte pour chaque valeur de X. Mais ce n'est pas aussi évident pour des valeurs de N plus élevées. Nous pouvons cependant les calculer à l'aide de la formule suivante :

$$\binom{n}{k}$$

Et c'est encore plus facile en allant directement chercher la valeur de ce coefficient binomial dans le tableau que vous trouverez en annexe.

10. Quelle est la formule à utiliser pour calculer la dernière colonne des deux tableaux ci-dessus.

11. Appliquez cette formule pour vérifier l'exactitude de vos réponses.

T.P. 7 – Exercice 6

Distribution binomiale, distribution normale, un peu de théorie

1. Complétez le texte suivant en vous aidant de votre syllabus :

Une distribution binomiale dépend d'une ……..……. caractérisée par les paramètres …… (…… ……) et …. (……..). Les événements qui la composent n'ont que ……..…. issues possibles. On dit aussi que cette expérience aléatoire est …… …… (ou du moins dichotomisée). On parle souvent d' …….. et de ……... (si on obtient l'un, on n'obtient pas l'autre) ; il n'y a pas d'autre possibilité que ces deux événements : on dit que ces événements sont ……..……. ……. et …….....…..…….

Nous sommes également dans le cas d'une ……..…… entre les événements : si j'ai lancé une pièce bien équilibrée neuf fois, la probabilité d'obtenir à nouveau pile est de …… .

En pratique, la binomiale devient inutilisable quand ……..…….. ; la distribution binomiale tend vers……..…….. c'est-à-dire que l'on peut considérer cette variable discrète comme une …….……

2. Quelle probabilité représente la surface totale sous la courbe normale ?

3. . Quelle est la notation que nous utilisons pour définir une distribution normale ? Donnez la formulation théorique sur base de l'estimation de la moyenne et de l'écart type et donnez un exemple chiffré.

4. Est-ce que modifier les paramètres d'une variable normale, en centrant la distribution autour de o et en réduisant risque, dans certains cas, de modifier la forme de la courbe ?

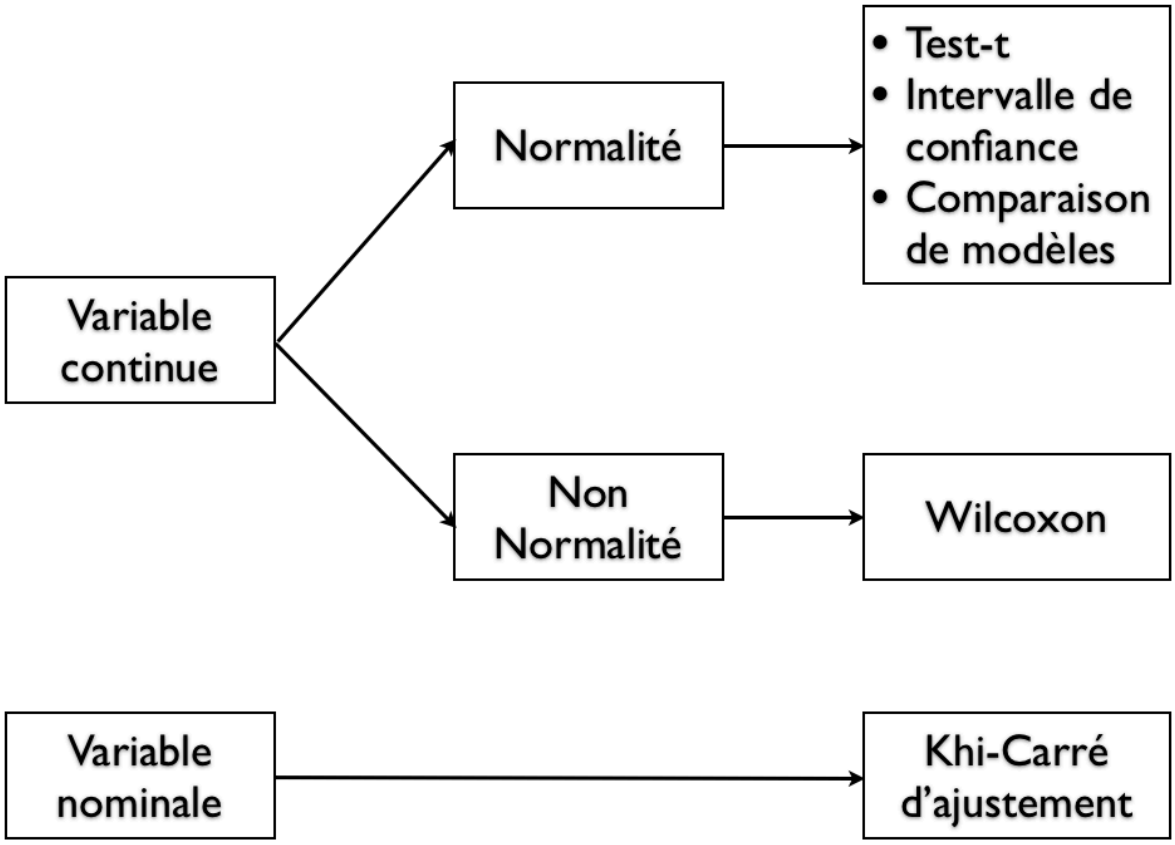
P A R T I E I V

INFERENCE STATISTIQUE ET MODELISATION

Introduction

Cette section s'intéresse au début de l'inférence, c'est-à-dire à l'estimation de paramètre d'une population à partir d'un échantillon. Le principe de base est d'estimer le paramètre d'une population à partir d'un échantillon et de comparer la compatibilité de cette estimation avec une valeur théorique connue. Par exemple, imaginez que vous mesuriez le nombre de voitures par ménage dans une commune wallonne dont vous soupçonnez un appauvrissement. Vous connaissez la moyenne nationale belge du nombres de voitures par ménage parce que les statistiques annuelles existent depuis plusieurs générations. Vous sélectionnez aléatoirement 30 ménages de la commune ciblée. Vous estimez la moyenne de cet échantillon et en déduisez la moyenne de la population dont est supposée être issu cet échantillon. Puis vous comparez cette valeur avec la moyenne nationale pour savoir si, oui ou non, votre estimation basée sur les habitants de la commune est compatible avec la valeur théorique correspondant à la population belge ou si vous devez considérer que cette commune est défavorisée. En d'autres termes vous vous demandez si les ménages issus de votre commune appartiennent bien à la population belge ou s'il faut considérer qu'il s'agit d'une autre population, défavorisée, pour la variable "nombre de voitures par ménage". Ce principe guidera toute cette section. Mais nous aborderons plusieurs approches du même problème et envisagerons des conditions différentes. Ci-dessous, la structure sous forme d'un arbre de décision. Les critères de décisions sont liés à l'échelle de mesure de la variable et aux conditions d'application de la méthode.

Arbre de décision



CHAPITRE 8 : INFERENCE STATISTIQUE A PROPOS DES VALEURS DE PARAMETRES

8.1. Introduction

Dès que j'en ai eu l'occasion dans les chapitres précédents, je vous ai parlé de l'erreur d'estimation. Cependant, jusqu'ici nous nous sommes attachés à décrire les données recueillies en mesurant une variable un certain nombre de fois. Même si nous avons abordé de temps à autre le rapport qui existait entre un échantillon et une population, nous n'avons pas encore abordé l'inférence statistique. Pourtant, l'écrasante majorité des outils statistiques utilisés par les chercheurs psychologues est directement liée à l'inférence. Au point 6.3.3.1 je vous ai parlé de cet aspect. Je vous ai dit que les psychologues mesuraient une variable sur un certain nombre de sujets, que l'on appelait l'échantillon et qu'à partir de ces informations, ils déduisaient les caractéristiques de la population (qui parfois est définie comme l'humanité toute entière). Je vous ai expliqué que, lorsqu'on estime un paramètre de la population (par exemple la moyenne) on ne tombe vraisemblablement pas sur la vraie valeur, c'est-à-dire celle qu'on obtiendrait si on mesurait la variable auprès de tous les individus qui constituent la population concernée. Par ailleurs, si plusieurs personnes utilisent un échantillon de même taille que moi (ou d'une taille différente d'ailleurs), plusieurs estimations de la même moyenne seront établies et elles seront relativement différentes les unes des autres. Un enjeux important était donc de déterminer une fourchette de valeurs dans laquelle la moyenne de la population se trouve probablement. C'est ce point qui nous préoccupera tout particulièrement dans ce chapitre et dans les deux suivants.

Structure des chapitres VIII, IX et X

Il est nécessaire que vous compreniez bien ce que nous sommes en train de faire : le chapitre 8 pose les bases de l'inférence statistique et vous montre les enjeux de trois manières différentes : par **intervalle de confiance**, par **comparaison de modèles** et par **test-*t***. Distinguez bien ces trois approches. En fait elles sont identiques, donc je répète trois fois la même chose, mais chacune a ses avantages et ses inconvénients, de sorte que vous devez être familier avec les trois. Nous verrons que, quelle que soit la manière, il y a des conditions d'application, certaines sur l'erreur, d'autres sur le type d'échelle de mesure utilisée (au moins une échelle d'intervalle). Le chapitre 9 vous montre ce que l'on doit faire pour traiter le même problème qu'au chapitre 8 mais lorsque nous utilisons une échelle nominale. Enfin, le chapitre 10 vous montre ce qu'il convient de faire lorsque les postulats concernant l'erreur (voir point 8.2) ne sont pas respectés. Les chapitres 9 et 10 traitent donc, eux aussi, du même problème, mais dans des conditions différentes.

Les tests statistiques se séparent en deux grands groupes : les **tests paramétriques et les tests non-paramétriques**. Les premiers sont basés sur l'utilisation de deux paramètres habituels que vous connaissez maintenant : la moyenne et l'écart-type (ou la variance). Les seconds s'appliquent lorsque nous ne sommes pas dans les conditions pour appliquer les premiers. Pour utiliser les approches paramétriques, nous allons devoir faire un certain nombre de postulats sur les caractéristiques de nos erreurs. C'est ce que les points suivants développent.

8.2. Postulats concernant l'erreur

8.2.1. Distribution normale de l'erreur

Au chapitre 7, nous avons vu que la distribution normale a été établie pour décrire l'erreur de mesure. **Cette erreur est donc considérée comme étant distribuée normalement autour d'une moyenne nulle**. En fait c'est habituellement le cas (c'est *normalement* le cas, d'où le nom de la distribution normale), mais pas toujours. Lorsque ce n'est pas le cas, il sera nécessaire d'avoir l'information sur la distribution concernée et de prendre les mesures qui s'imposent pour tenir compte de la situation (voir chapitre X, point 10.2).

Le postulat de normalité de la distribution de l'erreur est une condition essentielle pour utiliser la moyenne comme modèle prédictif et la SCE comme mesure de l'erreur. Si l'erreur suit une autre distribution, nous ne pourrons sans doute pas utiliser ces estimateurs. Cette conclusion devrait vous apparaître clairement, le chapitre 4 sur les distributions et le chapitre 6 sur les estimateurs de la tendance centrale et sur la dispersion devraient vous avoir sensibilisé à cette condition.

Ce postulat est une des raisons majeures pour lesquelles il est important de se soucier des valeurs aberrantes et de les ôter de l'analyse. En effet, rappelez-vous que la moyenne et la SCE sont très sensibles à ces valeurs et drastiquement corrompues si on ne gère pas le problème. Je vous renvoie au chapitre 10 pour savoir comment définir si l'erreur est ou non distribuée normalement.

8.2.2. Indépendance de l'erreur

Les erreurs sont considérées comme étant indépendantes. Cela signifie que l'erreur d'une mesure n'a absolument aucun effet sur l'erreur d'une autre mesure. Ce n'est pas toujours le cas. L'exemple le plus évident est représenté par les séries temporelles. Imaginez que l'on cherche à mettre en évidence l'effet de l'entraînement sur le lancer de fléchettes. On mesure les scores de 100 sujets au jet de 12 fléchettes (temps zéro). On entraîne ces mêmes 100 sujets pendant une semaine, 6 heures par jour. On les fait relancer 12 fléchettes à la fin de cet entraînement et on observe l'amélioration (temps un).

Cette manière de procéder est courante, cependant, les erreurs ne sont pas indépendantes (et la procédure d'analyse de ces résultats tiendra compte de cette non-indépendance, nous verrons cela en BA2). En effet, les aptitudes individuelles de chaque sujet sont constantes d'un essai à l'autre. Un sujet atteint de strabisme qui vise sans entraînement sera toujours atteint de strabisme après une semaine et l'erreur due à cet inconvénient physique sera présent de la même manière. En revanche, un sujet étant particulièrement doué pour viser et coordonner les mouvements du bras en fonction de la visée gardera également son aptitude avant et après. Les erreurs entre le temps zéro et le temps un sont donc bien liées.

Un autre cas de dépendance de l'erreur concerne par exemple les couples ou les enfants et les parents. Si je mesure la taille des pères, et la taille des fils, je peux me dire qu'un père de

grande taille aura sans doute un fils de taille supérieure à la moyenne également. C'est d'ailleurs ce qu'à montré Galton, le cousin de Darwin, dont nous reparlerons plus tard. De même, si je m'informe sur les valeurs morales (par exemple) de 50 couples, il y a de fortes chances pour qu'au sein de chaque couple les avis soient relativement partagés.

Remarquez également que, comme nous en avons déjà discuté, le fait même de prélever un sujet rompt le principe de l'indépendance, puisque cela influence la probabilité d'obtenir le sujet suivant. Cependant, je rappelle que, souvent, le prélèvement est tellement infime par rapport à l'effectif de la population que l'influence est négligeable.

8.2.3. Les erreurs sont identiquement distribuées

Ce point diffère de l'exigence de distribution normale de l'erreur. Dans ce cas, il s'agit de considérer que toutes les erreurs ont la même forme de distribution normale, c'est-à-dire la même dispersion.

Retournons à notre exemple de joueurs de fléchettes. Imaginons cette fois que 50 des 100 sujets soient en pleine santé, mais que les 50 autres soient pratiquement aveugles. Dans ce cas, même si tant les 50 premiers que les 50 suivants ont une distribution normale de l'erreur, la variance du second groupe sera fort probablement plus grande, vu qu'ils sont susceptibles de lancer la fléchette à peu près n'importe où. Cette différence de variabilité peut également poser problème. On dit que, lorsque les valeurs diffèrent de la sorte et que les variances ne sont pas les mêmes, selon les sous-échantillons concernés, il y a ***hétérosédasticité*** (= les variances diffèrent). Or il est postulé que les variances de toutes les erreurs soient équivalentes, c'est-à-dire qu'on postule l'***homosédasticité***.

8.2.4. Les erreurs sont dénuées de biais

Ce postulat est une autre manière de dire que les erreurs, distribuées normalement, sont centrées sur zéro (donc ont une moyenne de zéro). Cela revient à dire qu'il y a autant de chances d'observer des erreurs plus grandes que la prédiction que des erreurs plus petites que la prédiction.

Imaginons, par exemple, le cas décrit par Taleb (2001). Cet auteur, analyste financier de profession, décrit dans son ouvrage la manière dont on s'intéresse aux analystes financiers compétents. Il montre comment, en se focalisant sur ceux qui ont réussi, on commet un biais énorme au niveau de l'erreur. En effet, supposons que 1000 analystes placent une certaine quantité d'argent en bourse. Le pur hasard fera probablement qu'une certaine partie d'entre eux, 500, fassent de meilleurs résultats que la moyenne (c'est-à-dire que les indicateurs boursiers génériques comme le NASDAQ). Les 500 autres ont échoué et on n'entendra plus parler d'eux. L'année suivante, une partie des analystes chanceux, 250, réussissent à nouveau une bonne performance. La troisième année, il en reste 125 (toujours en considérant le même taux de réussite) et l'année d'après 62. En quatre ans, ces 62 analystes ont amassé une fortune confortable et sont réputés pour leurs placements judicieux. La cinquième année, de nombreuses enquêtes rapporteront les méthodes que les 31 analystes brillants restant ont utilisées et tenteront de trouver la solution miracle pour faire un bon analyste. Si ces enquêteurs avaient pris la peine de demander aux 969 analystes qui ont échoué la méthode qu'ils ont utilisée pour s'abîmer de la sorte, ils recueilleraient probablement les mêmes informations que celles recueillies auprès des analystes fructueux. Seulement, personne ne s'intéresse à ces 969 analystes ruinés de sorte que les réponses recueillies auprès de 31 analystes deviendront la base des nouvelles stratégies d'investissement des petits porteurs en quête de bons conseils. Cela revient à dire que l'erreur est toujours largement sur-estimée et n'est absolument plus centrée sur zéro. Je vous conseille vivement la lecture de ce livre. Non seulement il est écrit tel un roman sans une seule formule mathématique, mais en plus il est drôle et intéressant. Une fois lu, vous aurez une meilleure compréhension du hasard et de la manière dont il peut nous berner si nous n'y prenons pas garde.

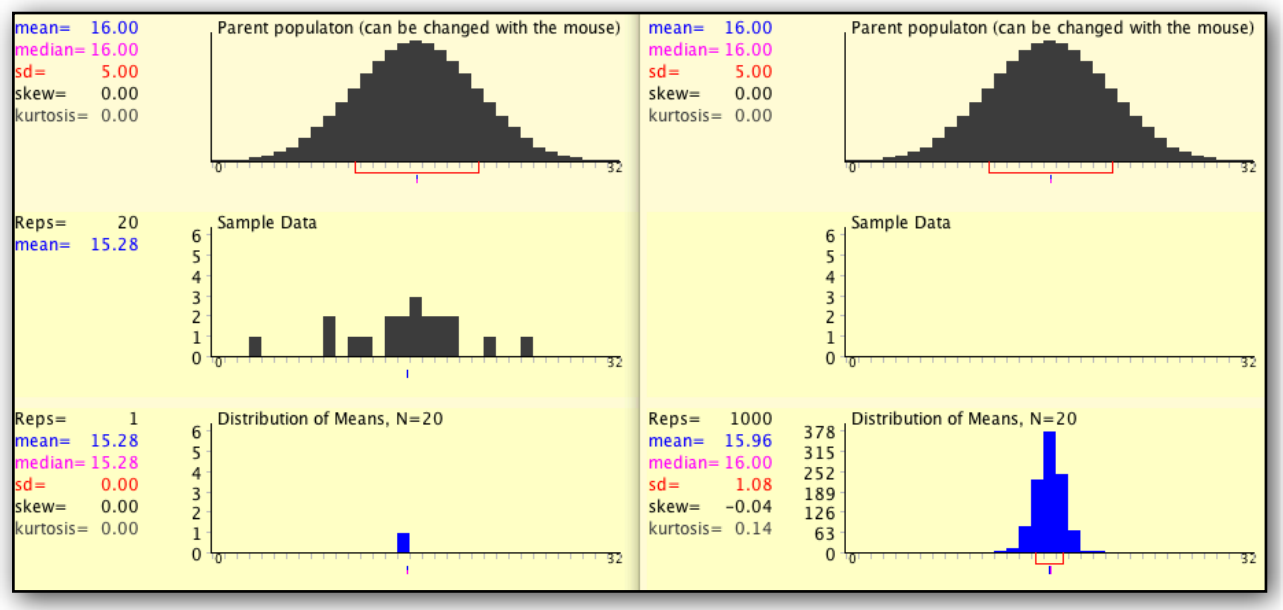
8.3. Les trois concepts essentiels à distinguer en inférence

Une habitude traditionnelle pour les étudiants est de ne pas parvenir à distinguer clairement trois distributions, et leurs paramètres respectifs : la distribution de l'échantillon, la distribution de la population et la distribution d'échantillonnage. Dans l'état actuel des choses, vous devriez être parfaitement à l'aise avec les deux premières, mais pas du tout avec la troisième. Ce point-ci est **LA** notion la plus importante à comprendre, à visualiser et à intégrer comme une seconde nature. Si vous ne comprenez pas, retravaillez ce chapitre, puis encore, et encore jusqu'à ce que ce soit intégré, c'est totalement indispensable. **Ne vous présentez pas à l'examen sans avoir intégré la notion de distribution**

d'échantillonnage et la différence qui existe entre distribution de l'échantillon, de la population et d'échantillonnage[29].

Le plus souvent, vous serez dans une situation où vous ne connaissez rien de votre population ni de votre distribution d'échantillonnage (que j'explique plus bas). Vous aurez recruté un certain nombre de sujets pour une expérience, mettons 40 personnes. Vous aurez mesuré une série de variables auprès de ces sujets, et voilà l'étendue des informations auxquelles vous aurez accès. Vos 40 sujets représentent votre échantillon, mais vous ne vous intéressez à eux que parce qu'ils constituent, selon vous, une proportion représentative d'un nombre bien plus important d'individus (par exemple toute l'humanité) qui est la population qui vous intéresse. C'est ce que représente la Figure 8.1. sauf que l'échantillon ne contient que 20 personnes.

Figure 8.1. : Echantillonnage de *n* = 20. **A gauche**, prélèvement d'un échantillon de 20 sujets (graphe du milieu) appartenant à la population (graphe du dessus) de moyenne 16 et d'écart-type de 5. Le graphe du dessous montre l'estimation de la moyenne de la population par l'échantillon (en bleu). **A droite**, même exercice réalisé 1000 fois, conduisant à la distribution d'échantillonnage (en bleu, graphe du dessous) dont l'erreur standard vaut 1,08 (représenté en rouge dans le graphique d'en dessous). Source : http://onlinestatbook.com/stat_sim/sampling_dist/index.html retrouvé le 25/10/11.



[29] J'espère avoir assez insisté pour vous faire comprendre que c'est important.

A partir des données de votre échantillon, et pour peu que les postulats, vus aux points précédents, concernant l'erreur soient respectés, vous estimez deux paramètres : la moyenne et la variance (corrigée, oublions l'autre). Ce que vous obtenez là sont les paramètres de votre échantillon. La moyenne de votre **échantillon** représente votre **estimation non biaisée** de la moyenne de la **population** correspondante. La **variance corrigée** représente votre **estimation** (non biaisée puisqu'elle est corrigée) de votre **population** (Tableau 8.2).

Cependant, vous êtes conscients que, si vous aviez pris 40 autres sujets, ou que vos collègues avaient fait la même expérience que vous mais sur un autre échantillon, vous auriez obtenu une estimation de vos paramètres légèrement différente. Par exemple, imaginons que vous ayez mesuré la taille moyenne des 40 sujets de votre échantillon (constitué de femmes belges de taille adulte), et que vous obteniez 1m68. Nous avons vu au début du cours qu'une autre expérience basée sur 100 personnes avait conduit à une estimation de la taille moyenne d'1m69. Ces deux valeurs ne sont pas très éloignées, mais ne sont pas identiques. Elles représentent toutes les deux une estimation de la taille moyenne des femmes adultes belges.

La vraie moyenne de la population des femmes adultes belges est inconnue. Supposons, pour l'exercice, que j'aie pris mon courage à deux mains (il faut au moins ça) et que je les aie mesurées toutes! J'ai obtenu une moyenne d'1m68. Je connais donc maintenant la vraie moyenne de ma population (il n'y a là aucune inférence, c'est purement descriptif). Vous vous rendez bien compte qu'il est en pratique impossible de réaliser cette mesure et que donc, **fictivement** je considère que connais la moyenne vraie de ma population, mais que dans la réalité ce n'est presque jamais réalisable (sauf dans de très rares cas, par exemple la taille des bébés en Belgique, parce qu'on les mesure systématiquement à la naissance depuis des années ; le nombre de voiture par habitant, parce qu'on tient une comptabilité précise des ventes et des déclassements ; les revenus moyens déclarés de la population, etc.).

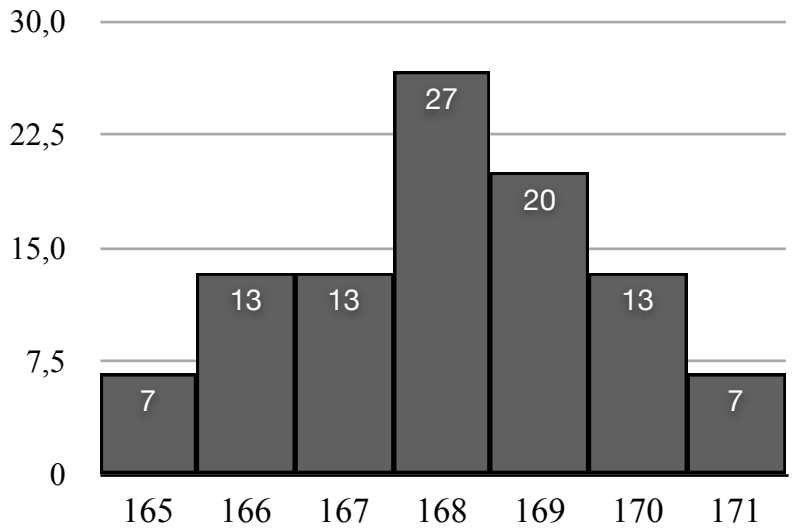
Imaginons, par ailleurs, que 14 autres personnes aient une mesure de la taille moyenne de 40 autres femmes adultes belges. C'est ce que le Tableau 8.1 rapporte comme données. La ligne grisée représente votre échantillon. Ce tableau ne prend pas en compte les 100 personnes qui donnaient une moyenne de 169 cm, ne nous en soucions plus pour l'instant. Comprenez bien qu'il s'agit ici de 15 moyennes hypothétiquement recueillies auprès de 15 fois 40 personnes. La Figure 8.2 représente l'histogramme de cette distribution de 15 moyennes et

c'est donc une **distribution d'échantillonnage de moyennes**. Remarquez également que nulle part, je vous informe sur l'écart-type associé à ces moyennes. Si je l'avais fait, j'aurais pu établir une moyenne des écarts-types et obtenir une **distribution d'échantillonnage d'écarts-types (ou de variance)**, ce sera l'objectif du point 9.2.4, mais ne complexifions pas les choses pour l'instant.

Tableau 8.1. : Distribution d'échantillonnage de 15 expérimentateurs (que j'appelle sujets) ayant mesuré la taille moyenne de 40 femmes adultes belges

n° sujet	Taille (cm)	n° sujet	Taille (cm)	n° sujet	Taille (cm)
1	168	6	167	11	170
2	167	7	168	12	166
3	169	8	168	13	166
4	168	9	169	14	170
5	165	10	171	15	169

Figure 8.2. : Histogramme des données du Tableau 8.1. (les nombres inscrits dans les barres représentent le pourcentage de l'échantillon qui ont obtenu la valeur concernée).



Vous pouvez constater que, le plus souvent, la moyenne des 40 personnes donne la bonne estimation de la moyenne de la population (1m68). Mais, dans certains cas on s'en éloigne, selon une distribution plus ou moins normale (elle serait parfaitement normale, si j'avais

utilisé un très grand nombre d'échantillons, par exemple 5000 au lieu de 15). Cela signifie que la plupart des expérimentateurs (ici 27%), dont vous, estimeront correctement la moyenne de la population. Une partie relativement importante (ici 20+13=33%) des expérimentateurs tomberont sur une évaluation pas trop éloignée mais quand même différente de la bonne moyenne, c'est-à-dire 167 ou 169 cm au lieu de 168. Une proportion encore plus petite (ici 13+13 = 26%) aura une moins bonne précision (166 ou 170 cm au lieu de 168). Une très petite proportion fera une erreur encore plus importante (ici 14% obtiendront 165 ou 171 au lieu de 168).

Dans l'exemple de la Figure 8.2, on peut se dire que 86% des expérimentateurs se trouvent à plus ou moins deux centimètres de la moyenne de la population et que 100% des expérimentateurs se trouvent à plus ou moins 3 centimètres de la vraie moyenne. Si j'avais une infinité d'expérimentateurs qui avaient mesuré la taille de 40 femmes adultes belges, j'aurais peut-être eu une très petite proportion des expérimentateurs qui auraient eu des tailles encore plus éloignées de la vraie moyenne, mais proportionnellement la grande majorité des expérimentateurs seraient tombés sur la bonne estimation ou une estimation très proche. La distribution serait une distribution normale qui pourrait être décrite par sa moyenne et son écart-type et dont la moyenne serait la vraie moyenne de la population c'est-à-dire 168 cm (puisque la distribution de l'erreur de mon modèle est centré sur zéro, comme on le postule au point 8.2.1).

Comprenez bien que j'ai imaginé ce qui se passerait **si** 15 personnes avaient mesuré la taille des 40 femmes (et j'ai inventé des valeurs plausibles) et j'ai imaginé également ce qui se passerait s'il s'agissait d'une infinité d'expérimentateurs. Mais en réalité, vous êtes tout(e) seul(e) à avoir mesuré la taille de 40 femmes et vous n'avez que cet échantillon comme information. Seulement, sachant ce qui se serait passé si vous n'étiez pas tout(e) seul(e), vous voudriez maintenant estimer les paramètres de la distribution d'échantillonnage théorique à laquelle votre échantillon appartient. De cette manière, vous pourriez vous dire que la vraie moyenne de la population doit se trouver autour de votre estimation (1m68) et vous devriez pouvoir trouver une fourchette de valeurs dans laquelle vous êtes raisonnablement sûr qu'elle se trouve sachant que plus on s'éloigne de votre estimation, moins la moyenne de la population a de chances de se trouver : si votre estimation est de 1m68, il y a peu de chances que la vraie moyenne de la population soit de 1m20, mais elle n'est pas nécessairement exactement égale à 1m68.

L'estimation de la moyenne de votre distribution d'échantillonnage est la moyenne de votre échantillon (Tableau 8.2). Vous n'avez, en effet, pas de raison d'utiliser autre chose, vous n'avez aucune information plus pertinente que celle-là. En revanche, l'écart-type de votre distribution d'échantillonnage, que l'on appelle l'**erreur standard** (termes que vous devez vous graver dans le cerveau au petit burin) et que nous avons déjà envisagé aux points 7.5.2 et 7.5.4, doit tenir compte de la taille de votre échantillon. En effet, plus il est grand, plus l'erreur standard sera petite (et mieux c'est, puisque moins vous aurez d'incertitude sur l'intervalle de valeurs dans lequel se situe la vraie moyenne de la population). Pourquoi?

Imaginons une situation un peu différente. Cette fois-ci, la population est représentée par les 200 étudiants de BA2. Je voudrais connaître leur taille moyenne (je parle un peu en "*je*" pour changer). Admettons qu'il n'y ait que des femmes adultes et que la taille moyenne de la population, que je ne connais pas, soit également de 168cm, pour faire simple. Donc, si je mesurais tout le monde, j'obtiendrais une taille moyenne de 168cm. Mais j'aime l'inférence statistique (c'est pour ça que j'ai changé de style et que je parle en "*je*" au lieu d'en "*vous*") et je suis extrêmement fainéant. Je n'ai donc aucune envie de mesurer tout le monde. J'hésite entre mesurer 2 personnes, 10 personnes ou 100 personnes. Donc, j'hésite entre prélever un échantillon **aléatoire** (parce que je m'apprête à sélectionner les sujets par une procédure complètement aléatoire) **simple** (appelé simple parce que le fait de choisir un étudiant ne change rien à la probabilité de choisir le suivant[30]) de $n = 2$, $n = 10$ ou $n = 100$. Admettons que dans cet auditoire, il y ait une étudiante géante de 200 cm.

Puisque la distribution de la population suit une loi normale, j'aurai beaucoup d'étudiantes qui mesurent 168 cm, et de moins en moins d'étudiantes qui s'écartent de cette valeur. Je n'ai d'ailleurs qu'une seule étudiante qui mesure 200 cm (et qu'en pratique il doit y en avoir une tous les dix ans).

Si je prends deux sujets pour estimer la moyenne de la classe, j'ai finalement très peu de chances de tomber sur l'étudiante de 200 cm (je peux calculer cette chance : 1/200 + 1/199 = 0,01 soit environ une chance sur 100)[31]. J'ai beaucoup plus de chances de tomber sur des

[30] On a déjà vu que ce n'était pas tout à fait vrai, mais c'est suffisamment vrai pour qu'on le considère comme tel

[31] Si j'avais réellement prélevé un échantillon aléatoire SIMPLE, la probabilité serait de 1/200 + 1/200 soit exactement 1%.

étudiantes d'une taille proche des 168cm que je cherche. En revanche, SI je tombe sur l'étudiante de 200 cm, je fais obligatoirement une très grosse erreur (à moins, par miracle, de tomber sur une étudiante toute petite pour contrebalancer mon étudiante géante). Imaginons que l'autre étudiante ait la taille moyenne de 168 cm, mon estimation devient : (200+168)/2 = 184 cm. Donc, avec ces deux étudiantes, je me dis que la population entière mesure 184 cm, alors qu'en réalité la vraie moyenne est de 168 cm : s'en suit une erreur de 16 cm (184-168) si mon étudiante de 200cm fait, malheureusement, partie de l'échantillon.

Admettons maintenant que j'utilise 10 sujets au lieu de 2. J'ai donc un peu plus de risques d'inclure dans mon échantillon l'étudiante de 200 cm (1/200 + 1/199 + 1/198 + … + 1/191 = 0,05), environ 5% de risques[32]. Mais, si j'inclus cette étudiante, l'effet est déjà moins catastrophique que dans le premier cas. En effet, imaginons que les 9 autres étudiantes aient exactement la taille moyenne de 168 cm, la moyenne devient (200+168*9)/10 = 171,2 cm. Je me trompe encore, mais de 3,2 cm au lieu de 16 cm, donc nettement moins fort que lorsque *n* était égal à 2 et que mon étudiante de 200cm était inclue dans l'échantillon.

Imaginons enfin que je prélève 100 étudiantes dans l'auditoire ce qui représente la moitié de ma population. Dans ce cas, le risque d'inclure mon étudiante de 200 cm est beaucoup plus élevé (69%)[33]. Cependant, l'impact de cette étudiante sera nettement moins grand. Admettons, à nouveau, que, par ailleurs, toutes les autres étudiantes mesurent 168 cm, l'estimation de la moyenne de ma population devient (200+168*99)/100 = 168,32 cm, donc je ne me trompe plus que de 0,32 cm (admettez que c'est mieux que les 16cm obtenus avec *n*=2).

Vous aurez donc compris que, plus la taille de mon échantillon augmente, plus mes estimations sont, en moyenne, proches de la vraie moyenne de ma population. En d'autres termes, plus l'erreur standard (l'écart-type de ma distribution d'échantillonnage) est petite. Cette situation est représentée à la Figure 8.3. Il ne s'agit en fait de rien d'autre qu'une illustration de la loi des grands nombres discutée au point 3.3.5. Sans démonstration, nous accepterons que, pour tenir compte de la taille de l'échantillon, l'estimation de l'erreur

[32] Ce qui est toujours une estimation proche du cas où mon échantillon était SIMPLE, puisque 5*1/200 = 5% exactement.

[33] Remarquez que lorsque je prends une grande partie de ma population (ici la moitié), la dépendance de la sélection commence à se faire cruellement sentir : si les prélèvements étaient indépendants, j'aurais eu une probabilité de 50% d'inclure le sujet de 200cm, ici j'ai une probabilité de 69%! Cela illustre notre discussion du point 5.2.

standard à partir de l'écart-type de mon échantillon se calcule comme suit : $\boldsymbol{\sigma_M = \sigma/\sqrt{n}}$ où σ_M est l'erreur standard, σ est l'écart-type de ma population et n la taille de l'échantillon (Tableau 8.2). Cependant, je ne connais pas σ. Je dois l'estimer à partir de s, l'écart-type corrigé de mon échantillon. Dès lors on estimera l'erreur standard de la manière suivante : $\boldsymbol{\sigma_M = s/\sqrt{n}}$

Figure 8.3. : Distribution d'échantillonnage (en bleu) représentant les estimations de la moyenne de la population (en noir) de moyenne égale à 16 et d'écart-type égal à 5. Plus l'effectif de l'échantillon est grand (à gauche n = 5, à droite n = 25) plus l'erreur standard (en rouge en-dessous de la distribution d'échantillonnage en bleue) est petite (à gauche σ_M = 2,22 ; à droite σ_M = 1,00) et donc moins les estimations de la moyenne s'écartent de la vraie moyenne de la population. Source : http://onlinestatbook.com/stat_sim/sampling_dist/index.html retrouvé le 25/10/11.

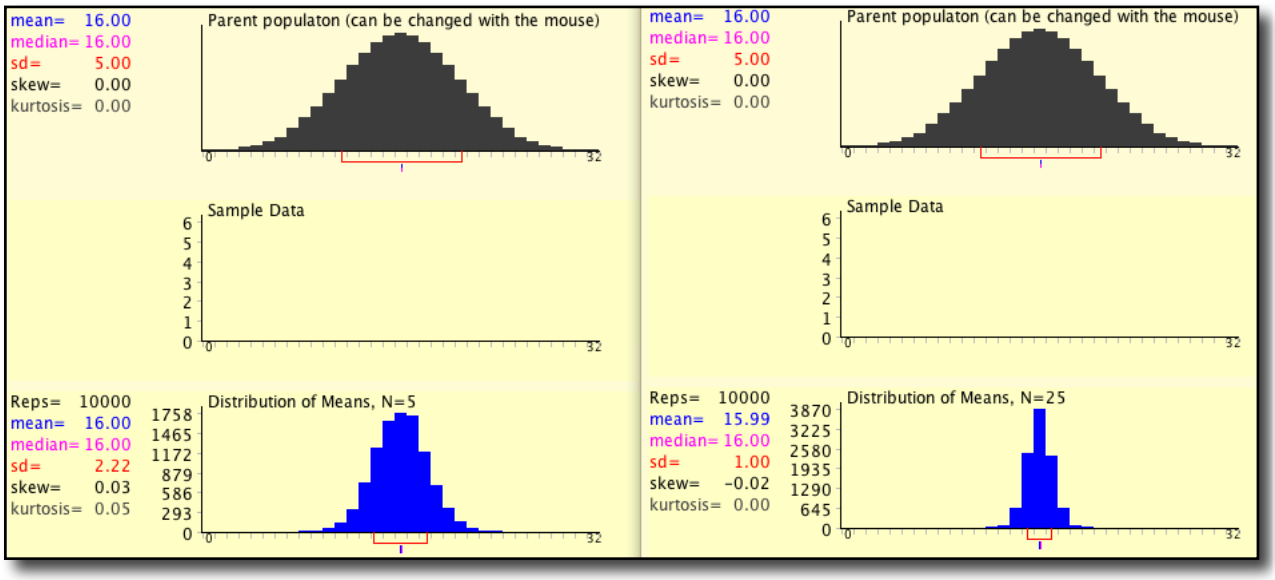


Tableau 8.2. : Détermination et estimation des paramètres des distributions de l'échantillon, de la population et de la distribution d'échantillonnage (j'utilise le signe ≈ pour "*est estimé par*").

	Echantillon	Population	Distribution d'échantillonnage
Moyenne	$\overline{X} = \frac{1}{n}\sum_{j=1}^{J} n_j x_j$	$\mu \approx \overline{X}$	$\mu_M = \mu \approx \overline{X}$
Ecart-type	$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$	$\sigma \approx S$	$\sigma_M = \frac{\sigma}{\sqrt{n}} \approx \frac{S}{\sqrt{n}}$ (= **Erreur standard**)

L'estimation des paramètres de la distribution d'échantillonnage est à la base de l'inférence statistique. C'est ce qui permet d'établir la fourchette de valeurs autour de l'estimation des paramètres de mon échantillon dans laquelle doit se trouver le paramètre correspondant de ma population. Cette fourchette de valeurs s'appelle l'**intervalle de confiance**. Le point suivant explique comment établir un tel intervalle de confiance, à l'aide de distribution théorique comme la distribution normale, et quelles conclusions tirer à partir de cet intervalle. Le point d'après appliquera ces notions au modèle de la moyenne que nous avons établi dans les chapitres précédents.

8.4. Etablissement de l'intervalle de confiance

Vous devriez maintenant avoir compris que l'intervalle de confiance dépend intrinsèquement de l'erreur standard. Et que l'erreur standard diminue au fur[34] et à mesure que l'effectif augmente. Cette erreur standard est estimée à partir de l'écart-type de l'échantillon, sauf si on connait la vraie variance de la population (auquel cas on l'utilise), ce qui n'arrive pour ainsi dire jamais[35] mais qui vaut la peine d'être considéré théoriquement.

[34] Vous êtes vous jamais demandé ce que voulait dire "fur"? Et bien il semble que cela veuille dire "mesure" ou "taux" et qu'en conséquence la locution "au fur et à mesure" soit en fait un pléonasme (un peu comme "aujourd'hui", "hui" voulant dire "jour"). Fur viendrait du latin "forum" voulant dire "marché" et est apparu comme définition de "mesure" au XVème siècle. (source : Dictionnaire de l'Académie française, huitième édition, 1932-1935).

[35] Je ne trouve aucun exemple de population dont on connait la variance mais pas la moyenne, et si on connait la variance et la moyenne, alors ce n'est pas la peine de calculer l'erreur standard puisqu'il n'est plus nécessaire de faire de l'inférence statistique.

En effet, si on connaissait la variance de la population, nous ne devrions pas l'estimer et donc nous ne ferions aucune erreur d'estimation. En revanche, le fait que l'on doive estimer la variance de la population par la variance de l'échantillon implique que l'on fasse une certaine erreur d'estimation. Nous allons envisager les deux cas et examiner les conséquences de cette erreur.

8.4.1. Cas où la variance de la population est connue

Dans ce cas, la distribution de l'erreur standard est considérée comme suivant strictement une distribution normale. Comme nous l'avons suggéré dans le point précédent. Trouver l'intervalle de confiance devient alors un jeu d'enfant pour qui sait utiliser la table de la loi normale, c'est-à-dire vous. Reprenons l'exemple du chapitre 7 concernant les notes sur 10 de 9 sujets (voir Tableau 7.6. mis en rappel plus bas). Imaginons cette fois que la population soit l'ensemble des étudiants de BA1 en psychologie et que j'ai choisi au hasard ces 9 sujets pour estimer la moyenne des étudiants de cette année sans ouvrir mon fichier contenant les cotes de tout le monde.

La moyenne des 9 étudiants est de 5/10. C'est donc ce que j'estime être la moyenne des étudiants de BA1. L'écart-type corrigé est de 1,22 ce qui représente l'estimation non biaisée de l'écart-type de la population. Seulement, imaginons que, par un moyen magique inexplicable, je sache en toute certitude que 1,22 est en fait la vraie valeur de l'écart-type de ma population (donc je ne l'estime pas, je la connais).

Je suppose que la note moyenne de toute la classe de BA1 est de 5/10 mais je me rends bien compte qu'il peut y avoir une petite différence entre cette estimation et la vraie moyenne des étudiants de BA1. J'aimerais donc bien savoir dans quel intervalle je dois m'attendre à trouver ma vraie moyenne et je vais calculer l'erreur standard. J'ai 9 étudiants, ce qui signifie que je divise mon estimation de l'écart-type par 3 (racine de 9) : $\sigma_M = 1{,}22/\sqrt{9} = 1{,}22/3 = 0{,}41$.

Le problème est maintenant de définir le risque de se tromper. Rappelez-vous qu'une distribution normale a comme caractéristique d'avoir une densité de probabilité de 68% lorsqu'on s'écarte de un écart-type à gauche et à droite de la moyenne (voir Figure 7.4). Dès lors, si l'on décidait de considérer que la vraie valeur de la moyenne de la population se trouvait à un écart-type à gauche ou à droite de 5 (c'est-à-dire entre 4,59 et 5,41) on aurait

32% de risques de se tromper, puisque l'intervalle ne comprend que 68% des valeurs possibles.

En psychologie la convention veut qu'un risque de 5% soit acceptable (tandis qu'un risque de 32% est totalement inacceptable vu que cela signifie qu'on prendrait une mauvaise décision une fois sur trois et que donc nos articles scientifiques contiendraient une erreur sur trois inférences). Nous devons, dès lors, trouver l'intervalle qui regroupe 95% des valeurs possibles et accepter qu'il y ait 5% de risques que la vraie valeur de la moyenne de la population soit en dehors de cet intervalle. Le risque résiduel de se tromper (donc, ici 5%) est désigné par le risque d'erreur de première espèce, ou **risque α**. Cela suggère donc qu'il existe un second risque d'erreur, nous y reviendrons (voir encadré plus bas). La table de la distribution normale (voir Tableau 7.7) nous informe que la valeur de 1,96 donne une densité de probabilité de 0,975 entre -∞ et 1,96. Soit 2,5% d'erreur. Par symétrie, on aurait 5% d'erreur si on considérait l'intervalle allant de -1,96 à 1,96. C'est donc la valeur qu'on retiendra.

Risque α et *p*-valeur

Lorsque l'on parle du **risque α**, on parle donc de la probabilité de faire l'erreur de première espèce. C'est-à-dire de la probabilité de décider qu'une valeur n'est issue de la même population que celle autour de laquelle j'ai construit ma distribution d'échantillonnage alors qu'en fait elle l'est. Ce risque est déterminé à l'avance. Je décide, *a priori*, qu'il sera de 5% et c'est sur cette base que je calcule mon intervalle de confiance. Puis, lorsque j'observe où se trouve une éventuelle valeur alternative, je peux attribuer une probabilité, correspondant à cette valeur précise, d'être issue de la même population que la moyenne observée de mon échantillon. Cette probabilité est la ***p*-valeur**, qui peut être plus petite ou plus grande que mon risque α déterminé *a priori*. Nous illustrerons cette différence au point 8.6.3.

L'intervalle de confiance autour de la moyenne qui regroupera 95% des valeurs possibles est donc de 5 ± 1,96*0,41 = 5 ± 0,8. Cela revient à dire que, $4{,}2 \leq \mu \leq 5{,}8$. On se dira que nous estimons la moyenne de la population de BA1 à 5/10 et que nous sommes sûrs à 95% que, si cette moyenne n'est sans doute pas exactement égale à cette valeur, elle est à tout le moins

comprise entre 4,2/10 et 5,8/10. L'intervalle entre 4,2 et 5,8 représente notre intervalle de confiance et est centré sur l'estimation de la moyenne de la population de 5.

Imaginons maintenant que la moyenne et l'écart-type de l'échantillon soient les mêmes mais que cette fois l'échantillon contient un nombre de sujets plus important, mettons 100. L'estimation de l'erreur standard devient : σ_M = 1,22/√100 = 1,22/10 = 0,12. Et l'intervalle de confiance à 5% de risque d'erreur devient : 5 ± 1,96*0,12 = 5 ± 0,24. Avec 100 sujets, j'estimerais toujours ma moyenne à 5 dans ce cas-ci, mais je considérerais que ma vraie moyenne de la population se trouve entre $4{,}76 \leq \mu \leq 5{,}24$. J'ai donc pu réduire assez fort mon incertitude en augmentant le nombre de sujets de mon échantillon.

Tableau 7.6. : (a) Série statistique de n = 9 représentant les cotes sur 10 à un examen ; (b) même série centrée et ; (c) centrée-réduite.

(a) Série		(b) Série centrée	(c) Série centrée-réduite
Num	Cotes (X_i)	Variable centrée $(X_i - \overline{X})$	Variable centrée-réduite $z = (X_i - \overline{X})/S$
1	3	3-5 = -2	(3-5)/1,22
2	5	5-5 = 0	(5-5)/1,22
3	7	7-5 = 2	(7-5)/1,22
4	4	4-5 = -1	(4-5)/1,22
5	5	5-5 = 0	(5-5)/1,22
6	6	6-5 = 1	(6-5)/1,22
7	5	5-5 = 0	(5-5)/1,22
8	6	6-5 = 1	(6-5)/1,22
9	4	4-5 = -1	(4-5)/1,22
$\overline{X}$ =	**5**	**0**	**0**
S =	**1,22**	**1,22**	**1**

Remarquez que nous aurions pu être plus ou moins exigeants au niveau du risque d'erreur α que nous prenons. Par exemple, j'aurais pu construire un intervalle de confiance ayant pour exigence de n'avoir qu'un seul pourcent d'erreur ou, au contraire, de tolérer 10% d'erreur. Je vous suggère, à titre d'exercices, de construire de tels intervalles.

Remarque sur le Risque α, notion de Risque β et de Puissance

En psychologie, comme je l'ai dit, l'usage veut qu'un niveau de confiance de 95% soit de mise. Mais dans certains cas, il y a lieu d'être beaucoup plus exigeant. Imaginons par exemple que vous désiriez savoir si vous êtes ou non séropositif. Vous faites donc une prise de sang dont le principe est de trouver des anticorps anti-HIV, donc repérant les antigènes (molécule qui est détectée par l'anticorps) de ce virus. En effet, si vous possédez de tels anticorps, c'est que, manifestement, vous avez été en contact avec le virus HIV (responsable du SIDA). Donc, vous êtes vraisemblablement toujours contaminés par ce virus qui résiste habituellement bien à vos défenses. Un tel anticorps a une configuration bien particulière, mais qui tolère de petits changements. La molécule utilisée par les biologistes pour détecter cet anticorps tolère donc elle aussi un certain niveau de variabilité de la part de cet anticorps. Ce qui signifie que la molécule s'associera à coup sûr à l'anticorps anti-HIV, mais éventuellement aussi à un anticorps dirigé vers un autre virus dont les antigènes ont une configuration proche de celle reconnue par l'anticorps anti-HIV. De cette manière, la présence de l'anticorps anti-HIV est détectée dans 99,9% des cas et mon risque α est très faible. C'est évidemment essentiel : il ne s'agit pas de rater la présence du HIV dans 5% des cas, sans quoi beaucoup d'individus mourront, convaincus de ne pas avoir la maladie. Malheureusement il y a un coût : cette stratégie conduit à détecter la présence du HIV **même dans des cas où il n'est en fait pas présent**. Ce risque de fausse détection correspond au risque de seconde espèce ou **risque *β***. C'est ce qui explique qu'en cas de détection positive du HIV, votre médecin demandera en principe toujours une deuxième prise de sang pour confirmer le diagnostic. On retrouve ce raisonnement pour à peu près toutes les détections médicales (test de grossesse, test de l'hépatite, etc.). Vous aurez compris que ces risques sont complémentaires : diminuer le risque α signifie augmenter le risque β et inversement, à vous de décider de leur importance relative.

Cette situation est tout à fait normale. D'un point de vue mathématique, diminuer le risque α correspond à augmenter l'intervalle de confiance. Pour reprendre ma série statistique du tableau 7.6, cela revient à considérer que ma vraie moyenne n'est pas comprise entre 4,2 et 5,8 mais, par exemple, entre 3 et 7 (comme exercice, vous devriez calculer le risque α qui correspond à cet intervalle, je trouverais ça une question d'examen intéressante). Dès lors, je ne suis pas prêt à exclure la valeur 3 des valeurs possibles de ma moyenne de la population. Or, il y a très peu de chances que cette moyenne soit la vraie moyenne de ma population (puisqu'elle est fort éloignée de mon estimation de 5/10). Donc en l'estimant possible, je prends un gros risque de me tromper et de l'accepter à tort (risque β). Je ferais sans doute mieux de l'exclure des valeurs plausibles de la moyenne. N'oublions pas que le but est d'être le plus précis possible dans notre estimation, sans quoi il suffirait de considérer que la moyenne de la population est comprise entre $-\infty$ et $+\infty$. Nous serions certains d'avoir raison (le risque α serait nul), mais ce serait totalement inutile comme conclusion (nous ne prédirions rien).

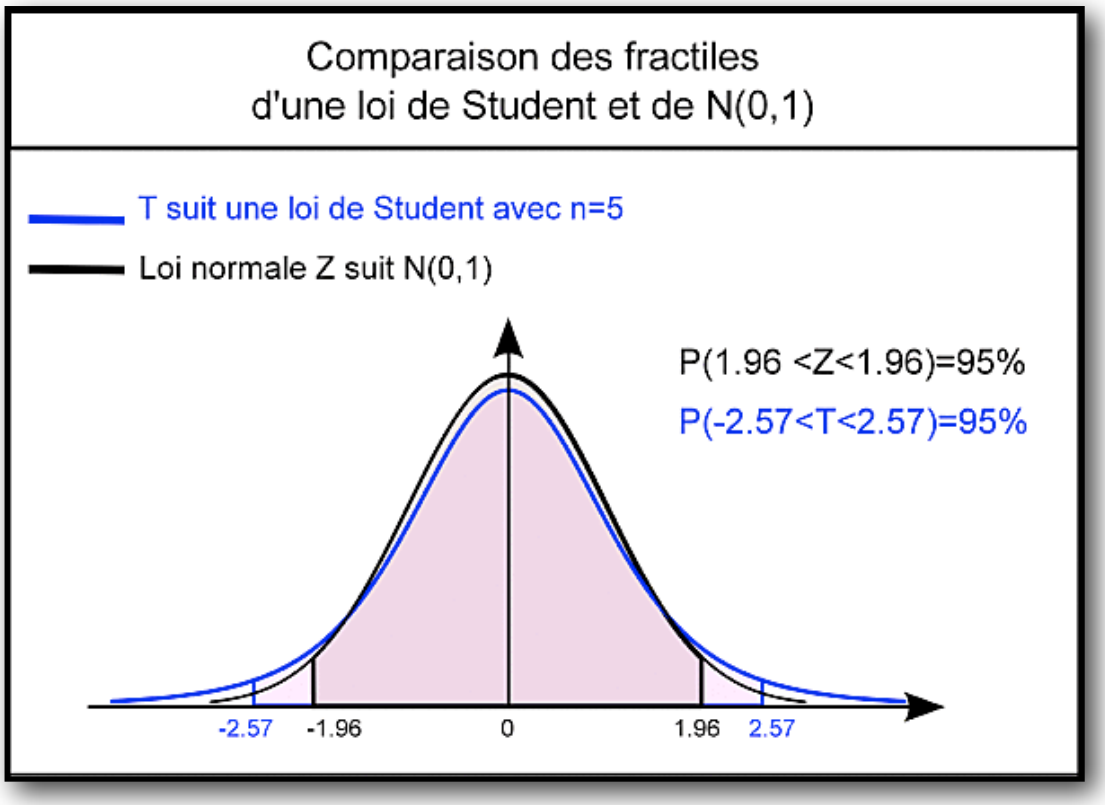
Ces considérations suggèrent également que nous puissions, dans certains cas, être moins exigeants sur le risque α. En effet, imaginons que j'aie une hypothèse à laquelle je crois très fort. J'aimerais m'assurer rapidement qu'elle soit plausible avant de pousser plus loin mes investigations. Je construis une petite étude sur une dizaine de sujets. Nous avons vu que, avec si peu de sujets, l'erreur standard sera très grande. Donc, je peux accepter un risque α de 10% en me disant que cela suffit pour continuer à tenir compte de mon hypothèse et aller plus avant dans mes recherches. Cette conclusion ne sera pas publiable au niveau d'exigence de la communauté scientifique, mais me permettra, à faible coût, d'orienter mes recherches dans une direction ou une autre.

Nous verrons ultérieurement comment calculer ce risque β et les notions annexes à ces considérations. D'ores et déjà, comprenez que nous désirons minimiser les deux risques, qu'ils sont tout deux liés à l'erreur standard et que donc plus on a de sujets, mieux c'est. Enfin, sachez déjà qu'un indicateur traditionnel de ce risque est en fait la considération inverse, c'est-à-dire la probabilité de ne pas faire d'erreur de seconde espèce, qui est la probabilité complémentaire du risque β c'est-à-dire $\boldsymbol{1-\beta}$. Cette probabilité d'éviter le risque β s'appelle la **puissance**.

8.4.2. Cas où la variance de la population est inconnue (estimée)

Imaginons-nous maintenant dans un cas nettement plus probable où 1,22 représente bien l'estimation de l'écart-type de la population (et non sa vraie valeur comme c'était le cas au point précédent). Le fait que nous ne soyons pas certains de la valeur vraie de l'écart-type de la population implique que nous sommes susceptibles de commettre une erreur d'estimation (c'est même fort probablement le cas). Nous pourrions d'ailleurs, selon le même principe que pour la moyenne trouver un intervalle de confiance dans lequel l'écart-type de la population devrait se trouver. Nous y reviendrons. Pour tenir compte de cette incertitude, William Gosset (ci-contre) a établi une distribution, assez proche de la distribution normale, plus conservatrice qui propose une correction de l'intervalle de confiance. Il s'agit de la distribution *t* de Student. Cette distribution ne dépend que d'un seul paramètre : le degré de liberté, *k*. La Figure 8.4 montre la différence entre une distribution *t* de Student pour $k = 4 = n-1$ et une distribution normale $N(0,1)$. Vous pourrez constater que, l'intervalle qui regroupe 95% des scores de la variable aléatoire est de ± 1,96 pour une loi normale alors qu'elle est sensiblement plus grande pour la distribution de Student, étant égale à ± 2,57.

Figure 8.4. : Comparaison des fractiles entre une distribution *t* de Student (k = 4) et une distribution Normale $N(0,1)$. Source : http://probastat.over-blog.com/article-student-versus-loi-normale-centree-reduite-55417756.html retrouvé le 23 octobre 2011.



William Gosset (1876-1937), originaire du comté d'Oxford, a réalisé ses études de Mathématiques et de Chimie à Oxford. Il a obtenu un prix dans chaque matière, d'abord en Mathématiques en 1897, puis en Chimie deux ans plus tard. Il fut ensuite engagé dans la brasserie *Guinness et fils* à Dublin, en 1899, comme chimiste. Durant cette période, il a effectué de nombreux travaux en statistiques, tout en étant en contact avec d'autres mathématiciens passionnés par ce sujet, tel Karl Pearson. Ses travaux empiriques sur le sujet l'ont conduit à établir la distribution dont il est question ici ainsi que les tests de comparaison de deux moyennes appelés test-*t* de Student. Tant la distribution que les tests de comparaison de moyennes ont été utilisés par Gosset pour contrôler la qualité de la stout (bière brune irlandaise produite, entre autres, par Guinness). Ce procédé de contrôle étant considéré comme secret industriel par l'employeur de Gosset, ce dernier publia ses travaux sous le pseudonyme de Student.

La distribution de Student n'est, en fait, utile que lorsque les degrés de liberté sont suffisamment peu nombreux. En effet, à l'infini, la distribution de Student tend à rejoindre

les valeurs de la distribution normale, de sorte qu'au-delà de k = 30 nous considérerons que la distribution est normale. La Figure 8.5 illustre cette situation et le Tableau 8.3 rapporte la table de Student. Si vous reprenez la table de la loi Normale, vous constaterez effectivement que les valeurs de la table de Student correspondent pour k > 30 (même s'il reste de petites différences).

Figure 8.5. : Distributions t de Student pour différents paramètres. Source : http://fr.wikipedia.org/wiki/Fichier:Student_densite_best.JPG retrouvé le 23/10/11.

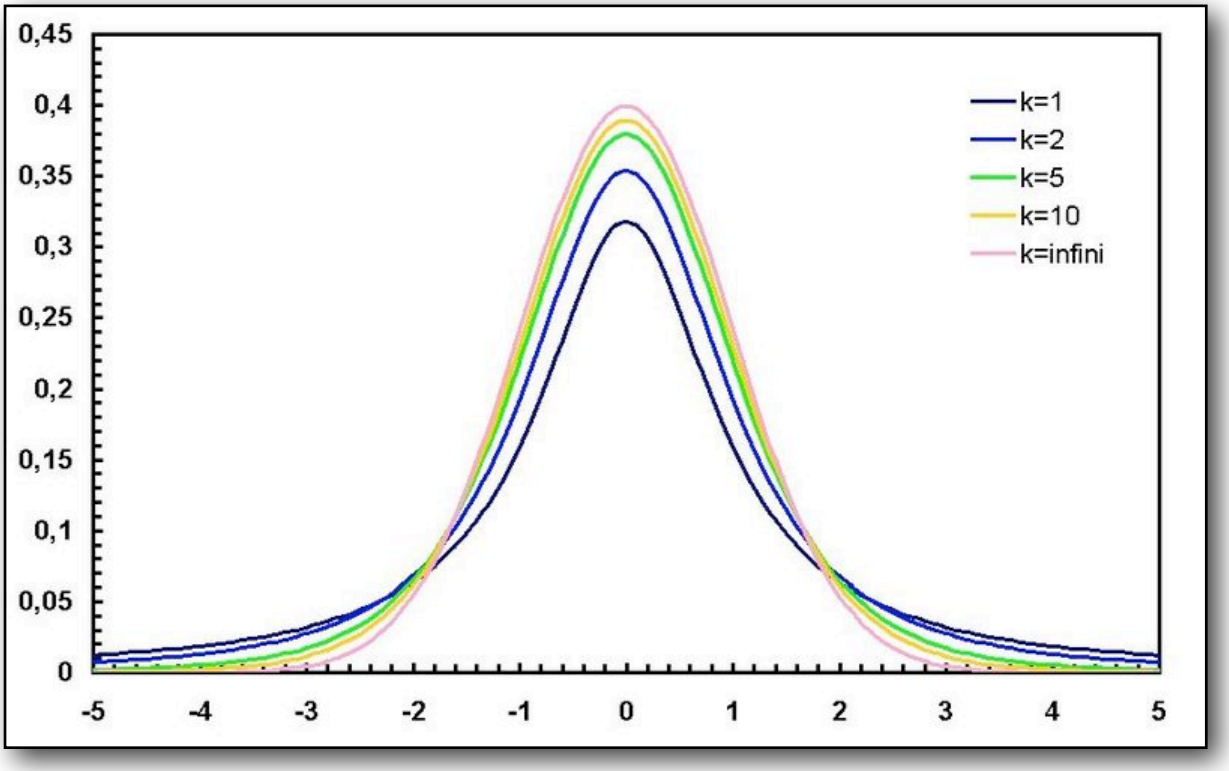
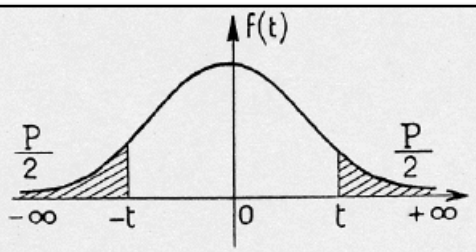


Tableau 8.3. : Table de la distribution de Student. Source : http://rfv.insa-lyon.fr/~jolion/STAT/node146.html retrouvé le 23/10/11.



ν \ P	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,929
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
80	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,254	0,386	0,526	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,253	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Reprenons maintenant notre exemple. La moyenne des cotes est donc de 5/10. Nous avons 9 personnes et une estimation de l'écart-type égale à 1,22. L'erreur standard est donc de 1,22/3 = 0,41 (ce calcul n'a pas changé). Nous voulons trouver l'intervalle de confiance autour de la moyenne en tenant compte du fait que nous avons estimé un écart-type de la population mais qu'il y a une imprécision puisque nous ne connaissons pas cet écart-type. La table de Student nous informe que, pour un nombre de degrés de liberté égal à 8 (parce que nous avons 9 sujets et que $k = n-1$), la valeur de la distribution qui autorise 5% d'erreur est 2,306. L'intervalle de confiance devient donc : IC = 5 ± 2,306*0,41 = 5 ± 0,95. Donc, nous estimons

cette fois que la vraie moyenne de la population se situe entre 4,05 ≤ μ ≤ 5,95. En résumé nous obtenons :

Formule de l'intervalle de confiance en utilisant la distribution t de Student

$$\overline{X} - t_{(ddl,\alpha)} \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{(ddl,\alpha)} \frac{S}{\sqrt{n}}$$

Si le nombre de sujets avait été de 100, nous aurions estimé l'intervalle de confiance égal à : IC = 5 ± 1,99*0,12 = 5 ± 0,24. L'estimation est donc bien la même qu'en utilisant la loi normale que nous aurions utilisé puisque n > 30. Remarquez que la valeur 1,99 est obtenue par interpolation linéaire entre k = 80 (valeur correspondante = 2,00) et 120 (valeur correspondante = 1,98) qui sont les valeurs données par la table.

Comme exercice, trouvez l'intervalle de confiance pour un risque α de 1% et pour un risque α de 10%. Quelles sont vos conclusions?

8.5. Application des concepts à la comparaison de modèles

Ici, commence réellement la comparaison de modèles dans le cadre de l'inférence statistique. Si vous comprenez bien l'entièreté de ce point, vous aurez les bases nécessaires à la compréhension de tout le cours de BA2 qui n'est qu'une complexification (que j'estime légère) de ce que nous verrons ici. Je rappelle que je vais vous dire exactement la même chose que ce que je viens de dire à propos des intervalles de confiance. Je ne fais qu'aborder le même problème sous un autre angle.

8.5.1. Etablissement des modèles à comparer

Imaginons une situation dans laquelle nous avons une prédiction possible autre que la moyenne, et plus économique, mais que nous nous demandons si la moyenne est une meilleure prédiction. Comprenez bien cet énoncé, il est très important pour la suite des événements.

Une telle situation peut par exemple se trouver avec un examen d'analyse de données (il sert vraiment à tout, c'est une aubaine que vous l'ayez!). Supposons que je teste vos connaissances à l'aide d'un questionnaire contenant 100 "*vrai ou faux*" et que je sois suffisamment peu versé dans les probabilités pour ne pas mettre de point négatif en cas de mauvaise réponse (rassurez-vous ça n'arrivera pas). Dès lors, quelqu'un qui n'y connait rien du tout à la matière aurait, en répondant totalement au hasard, en moyenne 50 points (on dit que sont espérance mathématique est de 50). Evidemment, j'espère que lorsque vous vous présentez à un examen, vous ne répondez pas au hasard. En étudiant, je suppose que vous ferez mieux. Je suis donc face à deux modèles : un modèle qui prédit une note obtenue complètement au hasard et un modèle qui me laisse suggérer que vous avez efficacement étudié quelque chose avant de venir. Cependant, je n'ai pas de valeur alternative à proposer. Je sais que 50 est le nombre que j'obtiendrais par réponse au hasard, mais j'ignore si en étudiant vous obtiendrez en moyenne 60, 70 ou 80% de bonnes réponses (ou n'importe quelle autre valeur, je pourrais même envisager avoir truffé mes questions de pièges de sorte que l'espérance est en fait inférieure à 50% parce que vous seriez induits à choisir la mauvaise réponse). Tout ce que je veux savoir, c'est si votre moyenne sera aussi bonne, moins bonne ou meilleure que le hasard et, comme je suis un enseignant bienveillant (si si, je vous assure), j'ai l'espoir que ce soit mieux. Cet espoir n'a aucune influence sur les mathématiques de ce qui va suivre.

Voici deux séries statistiques fictives obtenues par 14 étudiants de BA1, supposons dans deux cours différents, qui reprend le score sur 100 obtenu. La différence est au niveau de l'écart-type de la série. La première est très dispersée, la seconde beaucoup moins. Attachons-nous d'abord à la série concernant le premier cours : Y_i.

Tableau 8.4. : Séries statistiques fictives de deux cours de moyennes égales mais d'écarts-types différents pour une classe d'étudiants de BA1.

Numéro de sujet	Y_i	Y'_i
1	87	60
2	55	60
3	43	61
4	57	59
5	87	58
6	32	62
7	88	60
8	74	60
9	65	61
10	44	59
11	9	60
12	45	60
13	92	59
14	62	61
Moyenne	60	60
Ecart-Type	**24,31**	**1,04**

Nos précédentes discussions sur l'établissement de modèle de prédiction (voir point 6.2.2.3) nous autorise à écrire les équations suivantes :

$$\text{Modèle compact : } Y_i = B_0 + \varepsilon_{ci} \Leftrightarrow Y_i = 50 + \varepsilon_{ci}$$

$$\text{Modèle augmenté : } Y_i = \beta_0 + \varepsilon_{ai} \Leftrightarrow Y_i = 60 + \varepsilon_{ai}$$

Hypothèse nulle, h_0

L'hypothèse initiale à tester est que le modèle compact est identique au modèle augmenté. Puis, on espère pouvoir rejeter cette hypothèse. Cette logique peut sembler saugrenue en premier abord, mais tout ce que nous avons vu, lors des chapitres 2 et 3, à propos de la logique inductive et déductive sous-tend cette décision. En effet, il est impossible de confirmer une hypothèse, on ne peut que l'infirmer. Imaginez que votre hypothèse soit : "*toutes les voitures sont bleues*". Confirmer votre hypothèse reviendrait à chercher une multitude de voitures bleues et de vouloir utiliser ce recensement comme preuve. C'est peu efficace. En revanche, il suffit de trouver une voiture d'une autre couleur pour que l'hypothèse soit fausse. En conséquence, nous parlerons de h_0 comme de l'hypothèse nulle d'un test qui correspond toujours à une égalité entre deux modèles. Et nous espérerons que cette hypothèse puisse être rejetée avec un risque maximum de 5% de se tromper en l'affirmant. Dans notre exemple, on dit que :

h_0 : modèle compact = modèle augmenté

Voyons si nous pouvons rejeter cette hypothèse…

Je me trouve donc devant un modèle qualifié par Judd & al. (2010) de compact et un modèle augmenté. Dans le modèle compact, je considère simplement que les étudiants ont répondu au hasard à l'examen. A cette estimation est liée une erreur notée ε_{ci}. Dans le modèle augmenté, je considère la moyenne obtenue par mon échantillon et l'erreur associée est notée ε_{ai}. Je m'apprête à calculer si le fait de prendre en compte cette nouvelle information me permet de diminuer mon erreur de prédiction (donc si $\varepsilon_{ai} < \varepsilon_{ci}$) et si cette diminution est suffisamment grande pour être prise en compte. En d'autres termes, je me demande si la prédiction issue de mes données s'écarte suffisamment de ma prédiction qui ne dépend pas de ces données pour estimer la cote d'un étudiant de BA1 à partir de la moyenne de mon échantillon plutôt qu'à partir du hasard.

Une autre manière de voir les choses est la suivante : si je compare la distribution de l'erreur de mon modèle compact (ε_{ci}) à celle de mon modèle augmenté (ε_{ai}), suis-je tenté de conclure qu'elle appartient à deux distributions d'échantillonnage différentes (celle de ε_{ai} centrée sur

une moyenne plus basse que celle de ε_{ci}, puisque mon erreur doit être aussi faible que possible). Ou bien dois-je considérer qu'elles proviennent toutes les deux de la même distribution d'échantillonnage. Dans ce dernier cas, mon modèle augmenté ne sert à rien du tout et je peux me contenter de mon modèle compact.

8.5.2. Proportion de réduction de l'erreur entre les deux modèles

Dans les points précédents, nous avons envisagé la construction d'un intervalle de confiance autour d'une valeur. Nous pourrions, ici également, construire l'IC autour de 60 et regarder si la valeur 50 est comprise dans l'intervalle à un risque α de 5%. Si pas, je considérerais qu'il vaut mieux tenir compte de ma moyenne parce que les étudiants ont étudié et réussissent mieux que s'ils répondaient au hasard. Une autre manière de procéder, plus en accord avec le principe d'une comparaison de modèle, est de mesurer la proportion de réduction de l'erreur (PRE). Cet indicateur se mesure comme suit :

$$PRE = \frac{SCE(C) - SCE(A)}{SCE(C)} = \frac{SCR}{SCE(C)} \qquad \text{(équation 1)}$$

Rappelons que la SCE est la somme des carrés de l'erreur (voir chapitre 6, point 6.2.2.3). SCE (C) est la SCE du modèle compact et SCE (A) celle du modèle augmenté. Vous constatez donc que si SCE (A) = SCE (C) alors, PRE = 0 et il n'y a pas lieu de prendre le modèle augmenté en considération. En revanche, plus la SCE (A) diminue, plus la PRE se rapproche de 1 et plus il est intéressant d'utiliser le modèle augmenté. Le cas le plus propice serait que la SCE (A) = 0. Dans ce cas, nous n'aurions plus aucune erreur d'estimation en utilisant le modèle augmenté, PRE serait égal à 1, c'est-à-dire qu'on diminuerait l'erreur de 100%[36]. La SCR se définit comme la réduction de la somme des carrés de l'erreur que l'on obtient en soustrayant l'erreur due au modèle A de l'erreur initiale due au modèle C. Une autre manière de calculer la SCR est d'utiliser les prédictions de chacun des modèles C et A. En effet, plus la différence des prédictions est élevée, plus l'erreur est réduite (on ne peut jamais

[36] Au risque de générer de l'incompréhension, je suis en train de me dire que j'adore les stat au moment où j'écris ces lignes! Il y a de la magie là-dedans, ne la ressentez-vous pas? Tout névrosé en quête de contrôle ne peut qu'aimer cette matière, quoi de plus magique que de maîtriser le hasard, de se complaire dans l'incertitude? C'est complètement sécurisant : fini la dépendance à la mère, fini le Doudou (sauf éventuellement pour les montois), à nous l'autonomie et la libre pensée je vous dis!

augmenter l'erreur en tenant compte de nouvelles informations). Nous admettrons sans démonstration que :

$$SCR = \sum (\hat{Y}_{ic} - \hat{Y}_{ia})^2 = \sum (50 - 60)^2 = 14 * 100 = 1400$$

Le Tableau 8.5 calcule la SCE de chaque série pour le modèle augmenté et pour le modèle compact, c'est-à-dire par rapport à 60 et par rapport à 50. Comme vous le constatez, l'erreur est très faible, lorsque l'écart-type est très faible également. Cela devrait être une évidence absolue pour vous si vous avez compris que l'écart-type n'est qu'une autre manière d'exprimer l'erreur mais reste basée sur la SCE : la variance se calcule en divisant cette SCE par le nombre de degrés de liberté et l'écart-type est la racine carrée de la variance.

Tableau 8.5. : SCE(C) et SCE(A) des deux séries du Tableau 8.4.

Numéro de sujet	Y_i	SCE Compact $(Y_i - 50)^2$	SCE Augmenté $(Y_i - 60)^2$	Y'_i	SCE Compact $(Y'_i - 50)^2$	SCE Augmenté $(Y'_i - 60)^2$
1	87	1369	729	60	100	0
2	55	25	25	60	100	0
3	43	49	289	61	121	1
4	57	49	9	59	81	1
5	87	1369	729	58	64	4
6	32	324	784	62	144	4
7	88	1444	784	60	100	0
8	74	576	196	60	100	0
9	65	225	25	61	121	1
10	44	36	256	59	81	1
11	9	1681	2601	60	100	0
12	45	25	225	60	100	0
13	92	1764	1024	59	81	1

Tableau 8.5. : SCE(C) et SCE(A) des deux séries du Tableau 8.4.

14	62	144	4	61	121	1
Moyenne = 60		Σ = 9080	Σ = 7680	Moy = 60	Σ = 1414	Σ = 14
Ecart-Type = **24,31**				ET = 1,04		

A l'aide de l'équation 1, nous pouvons trouver la PRE de la série Y_i (nous envisagerons l'autre série au point 8.5.4) : PRE = (9080-7680)/9080 = 1400/9080 = 0,15. Cela signifie que nous diminuons de 15% la somme des carrés de l'erreur en utilisant le modèle augmenté par rapport au modèle compact.

Souvenons-nous maintenant de ce qui a été dit à propos de la distribution d'échantillonnage. Nous voulons savoir si l'erreur associée au modèle compact appartient à la même distribution d'échantillonnage que l'erreur associée au modèle augmenté. En d'autres termes, nous nous demandons si la moyenne vraie de la population des erreurs estimée par le modèle compact se trouve dans un intervalle de confiance qui inclut la moyenne vraie de la population des erreurs estimée à l'aide du modèle augmenté. Une méthode pour répondre à cette même question est de se demander si la proportion de la réduction de l'erreur (PRE) est comprise dans l'intervalle de confiance de la valeur zéro. Si oui, cela signifie que la PRE n'est pas distinguable de o et donc que le modèle A n'apporte rien. Si non cela signifie que la PRE est différente de zéro et vaut donc la peine d'être prise en compte. Il est donc nécessaire de connaître la distribution de la PRE. C'est, d'une manière un petit peu différente, ce que donne la statistique *F*.

8.5.3. La distribution F et ANOVA

Dans les livres de statistique usuels (et dans les logiciels informatiques) vous ne trouverez, en général, pas de table de distribution d'une PRE. En revanche, il existe une distribution directement liée à la PRE nommée distribution *F* de Fisher-Snedecor. Voyons d'abord les adaptations apportées à la PRE avant d'aborder la logique de cette distribution.

Telle que nous l'avons décrite jusqu'à présent, la PRE ne donne qu'une information partielle. Par exemple, elle n'apporte aucune information sur la proportion de réduction d'erreur **par paramètre estimé**. Or, nous avons déjà vu que c'était un critère fondamental. En effet, ici,

nous avons ajouté un paramètre (la moyenne de notre échantillon) et nous avons réduit l'erreur. Nous aurions pu, comme nous le ferons plus tard, envisager plusieurs paramètres qui améliorent notre prédiction. Rappelez-vous, au point 2.2, nous avons émis l'idée qu'une bonne description de la réalité peut prendre en compte plusieurs variables explicatives (indépendantes) d'une variable dépendante. Nous pouvons, à la lumière de ce que nous avons vu sur la PRE, imaginer qu'à chaque variable prise en compte correspondrait une PRE. Dès lors, il peut être important de considérer la PRE moyenne des différentes variables envisagées. Par exemple, imaginons que nous ayons pris 3 variables en considération et que la PRE soit de 0,63 (donc que nous avons réduit l'erreur de 63%). Nous pourrions nous dire que nous avons réduit l'erreur de, en moyenne 0,21 (= 0,63/3) par variable indépendante.

Un second élément dont nous ne tenons pas compte avec la PRE telle qu'énoncée à présent est la proportion d'erreur résiduelle, donc l'erreur qui demeure. Elle peut s'exprimer sous la forme (1-PRE). Et il peut être intéressant de connaître l'erreur résiduelle moyenne qui reste à expliquer à l'aide des variables explicatives dont l'effet reste à estimer. Nous avons vu, dans l'encadré sur les degrés de liberté au point 6.3.3.2, qu'il pouvait y avoir autant de paramètres estimés que de sujets. Sachant que certains paramètres ont déjà été estimés (dans notre exemple, un paramètre a déjà été estimé, la moyenne), il est question de se demander quelle quantité d'erreur il reste à expliquer et combien de paramètres sont disponibles pour l'expliquer (donc on se demande quelle est l'erreur moyenne par paramètre restant).

En conséquence, nous avons deux valeurs de degrés de liberté importantes à calculer : le nombre de degrés de liberté lié au nombre de paramètres estimés et le nombre de degrés de liberté lié au nombre de paramètres qui restent possibles d'estimer. La première valeur est égale à (PA - PC) où PA est le nombre de paramètres estimés dans le modèle augmenté et PC le nombre de paramètres estimés dans le modèle compact. Dans notre exemple, le modèle compact n'estime aucun paramètre (la valeur 50 correspond à une valeur basée sur le hasard) et le modèle augmenté estime un paramètre (la moyenne de l'échantillon, 60). PA - PC = 1 - 0 = 1 pour cet exemple. Le nombre de paramètres qui restent possibles à estimer est égal au nombre de sujets moins le nombre de paramètres estimés par le modèle augmenté, soit (n - PA), dans notre exemple, n - 1.

A partir de ces valeurs, nous pouvons calculer la statistique *F*. Elle consiste en fait, à calculer le rapport entre la réduction de l'erreur par paramètre estimé et l'erreur résiduelle par

paramètre encore potentiellement estimable. Dans la mesure où une réduction d'erreur associée à un paramètre est intéressante à prendre en compte SI elle diminue mieux l'erreur qu'un paramètre estimé quelconque, plus ce rapport est élevé plus cela signifie que l'information (la variable) prise en compte est pertinente. Mathématiquement cela donne :

$$F = \frac{PRE/(PA - PC)}{(1 - PRE)/(n - PA)} = \frac{0{,}15/(1-0)}{0{,}85/(14-1)} = 2{,}29$$

Habituellement, un tableau de résultat d'une telle comparaison de modèle contient les informations transmises par le Tableau 8.6. Remarquez que la valeur de *F* est un petit peu différente du 2,29 trouvé ci-dessus. Il s'agit simplement d'une erreur d'arrondi, la vraie valeur de PRE étant de 0,154185 et donc 1-PRE étant de 0,845815. Les informations correspondent à : la réduction de l'erreur par le modèle A ; l'erreur qui demeure inhérente au modèle A ; et l'erreur totale initiale, c'est-à-dire celle du modèle C. Les degrés de liberté de chacune de ces valeurs sont également transmises. A l'aide de ces informations, il y a également moyen d'exprimer la valeur de *F* et c'est traditionnellement de cette manière qu'on le fait. Je vous ai donné l'autre pour vous aider à comprendre la logique de ce calcul, mais une simple transformation mathématique vous permettrait de découvrir par vous même que :

$$F = \frac{SCR/(PA - PC)}{SCE(A)/(n - PA)} = \frac{CMR}{CME} = 2{,}37$$

Dans cette équation, le CMR est la réduction du carré moyen (moyen voulant donc dire : par paramètre estimé) et CME est le carré moyen (cette fois par paramètre qui reste à estimer) de l'erreur qui reste après avoir utilisé le modèle augmenté. Une fois la valeur de *F* obtenue, il reste à nous interroger sur les valeurs de *F* qui sont anormales. Si la CMR ne fait pas mieux que n'importe quelle autre variable ne pourrait le faire, le rapport serait de 1 (parce que CMR serait égal à CME). Je me demande donc quelle est la distribution d'échantillonnage de *F* et quelle est la valeur de *F* qui correspond à une densité de probabilité de 95% (parce qu'en psychologie, on accepte un risque d'erreur de première espèce de 5%)? Si je sais répondre à cette question, je peux alors à partir de cette valeur, considérer que le rapport (CMR/CME) est suffisamment plus grand que 1 pour penser que la réduction de l'erreur par paramètre estimé est réellement plus élevée que la réduction d'erreur moyenne par paramètre résiduel.

En d'autres termes, je me dirai, à ce moment là, que j'ai choisi une variable qui réduit significativement mon erreur, donc qui valait la peine d'être considérée. Cette distribution *F* se nomme la distribution de Fisher-Snedecor du nom des mathématiciens qui l'ont établie. Il s'agit d'une distribution d'un rapport entre deux variances (la réduction de l'erreur au numérateur et l'erreur résiduelle au dénominateur). C'est pourquoi lorsque l'on parle de cette méthode d'analyse, on la qualifie d'**analyse de variance**, en Anglais ANalysis Of VAriance ou **ANOVA**.

La distribution *F* a une forme caractéristique et il existe une table de valeurs associée (voir Tableau 8.7) qui fonctionne selon le même principe que la table *t* de Student si ce n'est qu'elle est décrite par deux paramètres au lieu d'un seul : le nombre de degrés de liberté au numérateur et le nombre de degrés de liberté au dénominateur. Cette table nous informe qu'une $F(1, 13)$ doit être plus grande que 4,67 pour considérer que la valeur est significativement supérieure à 1 avec une probabilité de 95% d'avoir raison. La valeur que nous observons étant de 2,37, nous ne pouvons pas considérer que la différence de 10 points entre 50 et 60 soit due au fait que les étudiants ont bien étudié. Nous ne pouvons pas rejeter l'hypothèse qu'ils ont répondu au hasard mais que le hasard a été clément pour cet échantillon particulier qui a une moyenne supérieure à 50. L'information donnée par la moyenne ne diminue pas suffisamment l'erreur par rapport à ce qu'un autre paramètre pourrait, en moyenne, faire.

Remarque TRES IMPORTANTE sur le rejet de l'hypothèse h_0

Lorsque l'on rejette une hypothèse, c'est qu'on a pu observer une différence significative et nous pouvons être confiants dans notre décision de la rejeter (avec un risque de 5% de se tromper, mais on accepte ce risque). En revanche, **si on ne rejette pas l'hypothèse, on ne peut pas pour autant l'accepter**! On doit conclure qu'on ignore la conclusion! Pourquoi?

Vous devriez déjà l'avoir compris suite à l'encadré sur h_0 (point 8.5.1) et suite au cours de logique des premiers chapitres, mais illustrons-le, une fois de plus, d'une autre manière : supposons que vous lanciez 4 fois une pièce de monnaie, et que vous obteniez 3 "*face*" et 1 "*pile*". Pouvez-vous en conclure que votre pièce est mal équilibrée? La réponse est évidemment non. Une pièce parfaitement bien équilibrée pourrait tout à fait produire un tel résultat. Cependant, il n'est pas impossible qu'elle soit mal équilibrée et que sa probabilité de faire "*face*" soit bien de 75%, vous n'en savez rien.

Supposons maintenant que vous lanciez 1000 fois la pièce, et que cette fois vous obteniez 750 fois "*face*" et seulement 250 fois "*pile*". Dans ce cas, vous pouvez avoir un peu plus confiance en vous en rejetant l'idée que la pièce est bien équilibrée.

Donc, soit vous êtes capables de rejeter votre hypothèse et d'adopter votre modèle augmenté avec une probabilité de vous tromper connue (et définie *a priori*, c'est votre risque α). Soit vous n'êtes pas dans les conditions pour pouvoir rejeter votre hypothèse et dans ce cas, vous n'avez aucune conclusion à donner (c'est ce qui explique qu'on ne publie jamais un résultat non significatif dans la littérature, sauf s'il fait partie d'un ensemble de résultats dont certains sont significatifs).

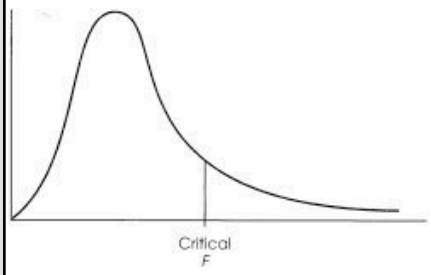
Tableau 8.6. : Tableau d'ANOVA appliqué à la série 1 du Tableau 8.4

Source	*SC*	*ddl*	*CM*	*F*	*p*
Réduction modèle A	SCR	PA - PC	CMR = SCR/(PA - PC)	CMR/CME	
Série 1	*1400*	*1*	*1400/1*	*2,37*	>,05
Erreur modèle A	SCE(A)	n - PA	CME = SCE(A)/(n - PA)		
Série 1	*7680*	*13*	*7680/13 = 590,77*		
Total	SCE(C)	n - PC			
Série 1	*9080*	*14*			

Remarquez la forme de la distribution *F* représentée au-dessus de la table des valeurs. Cette distribution est asymétrique (asymétrie positive). Elle a comme minimum 0 et comme maximum +∞. Tout le risque alpha est reporté sur la queue de la courbe, c'est-à-dire sur les valeurs élevées de la distribution. Nous sommes donc bien en train de nous demander si le rapport est suffisamment plus grand que 1, mais nous n'envisageons pas la possibilité qu'il soit significativement plus petit que 1. Cela n'aurait d'ailleurs pas de sens puisqu'un rapport plus petit que 1 correspondrait à l'idée que l'information que nous prenons en compte RAJOUTERAIT de l'erreur. C'est bien entendu impossible. Toute information, si bénigne soit-elle ne peut que rendre notre estimation plus précise (même si on découvre que deux variables ne sont pas dépendantes l'une de l'autre, c'est une information). Cependant, l'erreur d'échantillonnage, due aux différences individuelles, pourrait dans certains cas conduire à un rapport inférieur à 1, mais certainement pas significativement. Cette valeur inférieure ne peut être due qu'au hasard de l'échantillonnage et à rien d'autre. Nul besoin donc d'envisager un test d'inférence à ce sujet.

Tableau 8.7. : Table des valeurs critiques de la distribution *F* à p = 0,05 (risque α = 5%) et p = 0,01 (risque α = 1%).

*Table entries in lightface type are critical values for the .05 level of significance. Boldface type values are for the .01 level of significance.



DEGREES OF FREEDOM: DENOMINATOR	DEGREES OF FREEDOM: NUMERATOR														
	1	2	3	4	5	6	7	8	9	10	11	12	14	16	20
1	161 **4052**	200 **4999**	216 **5403**	225 **5625**	230 **5764**	234 **5859**	237 **5928**	239 **5981**	241 **6022**	242 **6056**	243 **6082**	244 **6106**	245 **6142**	246 **6169**	248 **6208**
2	18.51 **98.49**	19.00 **99.00**	19.16 **99.17**	19.25 **99.25**	19.30 **99.30**	19.33 **99.33**	19.36 **99.34**	19.37 **99.36**	19.38 **99.38**	19.39 **99.40**	19.40 **99.41**	19.41 **99.42**	19.42 **99.43**	19.43 **99.44**	19.44 **99.45**
3	10.13 **34.12**	9.55 **30.92**	9.28 **29.46**	9.12 **28.71**	9.01 **28.24**	8.94 **27.91**	8.88 **27.67**	8.84 **27.49**	8.81 **27.34**	8.78 **27.23**	8.76 **27.13**	8.74 **27.05**	8.71 **26.92**	8.69 **26.83**	8.66 **26.69**
4	7.71 **21.20**	6.94 **18.00**	6.59 **16.69**	6.39 **15.98**	6.26 **15.52**	6.16 **15.21**	6.09 **14.98**	6.04 **14.80**	6.00 **14.66**	5.96 **14.54**	5.93 **14.45**	5.91 **14.37**	5.87 **14.24**	5.84 **14.15**	5.80 **14.02**
5	6.61 **16.26**	5.79 **13.27**	5.41 **12.06**	5.19 **11.39**	5.05 **10.97**	4.95 **10.67**	4.88 **10.45**	4.82 **10.27**	4.78 **10.15**	4.74 **10.05**	4.70 **9.96**	4.68 **9.89**	4.64 **9.77**	4.60 **9.68**	4.56 **9.55**
6	5.99 **13.74**	5.14 **10.92**	4.76 **9.78**	4.53 **9.15**	4.39 **8.75**	4.28 **8.47**	4.21 **8.26**	4.15 **8.10**	4.10 **7.98**	4.06 **7.87**	4.03 **7.79**	4.00 **7.72**	3.96 **7.60**	3.92 **7.52**	3.87 **7.39**
7	5.59 **12.25**	4.47 **9.55**	4.35 **8.45**	4.12 **7.85**	3.97 **7.46**	3.87 **7.19**	3.79 **7.00**	3.73 **6.84**	3.68 **6.71**	3.63 **6.62**	3.60 **6.54**	3.57 **6.47**	3.52 **6.35**	3.49 **6.27**	3.44 **6.15**
8	5.32 **11.26**	4.46 **8.65**	4.07 **7.59**	3.84 **7.01**	3.69 **6.63**	3.58 **6.37**	3.50 **6.19**	3.44 **6.03**	3.39 **5.91**	3.34 **5.82**	3.31 **5.74**	3.28 **5.67**	3.23 **5.56**	3.20 **5.48**	3.15 **5.36**
9	5.12 **10.56**	4.26 **8.02**	3.86 **6.99**	3.63 **6.42**	3.48 **6.06**	3.37 **5.80**	3.29 **5.62**	3.23 **5.47**	3.18 **5.35**	3.13 **5.26**	3.10 **5.18**	3.07 **5.11**	3.02 **5.00**	2.98 **4.92**	2.93 **4.80**
10	4.96 **10.04**	4.10 **7.56**	3.71 **6.55**	3.48 **5.99**	3.33 **5.64**	3.22 **5.39**	3.14 **5.21**	3.07 **5.06**	3.02 **4.95**	2.97 **4.85**	2.94 **4.78**	2.91 **4.71**	2.86 **4.60**	2.82 **4.52**	2.77 **4.41**
11	4.84 **9.65**	3.98 **7.20**	3.59 **6.22**	3.36 **5.67**	3.20 **5.32**	3.09 **5.07**	3.01 **4.88**	2.95 **4.74**	2.90 **4.63**	2.86 **4.54**	2.82 **4.46**	2.79 **4.40**	2.74 **4.29**	2.70 **4.21**	2.65 **4.10**
12	4.75 **9.33**	3.88 **6.93**	3.49 **5.95**	3.26 **5.41**	3.11 **5.06**	3.00 **4.82**	2.92 **4.65**	2.85 **4.50**	2.80 **4.39**	2.76 **4.30**	2.72 **4.22**	2.69 **4.16**	2.64 **4.05**	2.60 **3.98**	2.54 **3.86**
13	4.67 **9.07**	3.80 **6.70**	3.41 **5.74**	3.18 **5.20**	3.02 **4.86**	2.92 **4.62**	2.84 **4.44**	2.77 **4.30**	2.72 **4.19**	2.67 **4.10**	2.63 **4.02**	2.60 **3.96**	2.55 **3.85**	2.51 **3.78**	2.46 **3.67**
14	4.60 **8.86**	3.74 **6.51**	3.34 **5.56**	3.11 **5.03**	2.96 **4.69**	2.85 **4.46**	2.77 **4.28**	2.70 **4.14**	2.65 **4.03**	2.60 **3.94**	2.56 **3.86**	2.53 **3.80**	2.48 **3.70**	2.44 **3.62**	2.39 **3.51**
15	4.54 **8.68**	3.68 **6.36**	3.29 **5.42**	3.06 **4.89**	2.90 **4.56**	2.79 **4.32**	2.70 **4.14**	2.64 **4.00**	2.59 **3.89**	2.55 **3.80**	2.51 **3.73**	2.48 **3.67**	2.43 **3.56**	2.39 **3.48**	2.33 **3.36**
16	4.49 **8.53**	3.63 **6.23**	3.24 **5.29**	3.01 **4.77**	2.85 **4.44**	2.74 **4.20**	2.66 **4.03**	2.59 **3.89**	2.54 **3.78**	2.49 **3.69**	2.45 **3.61**	2.42 **3.55**	2.37 **3.45**	2.33 **3.37**	2.28 **3.25**

8.5.4. Importance de la dispersion de l'erreur

Voyons maintenant ce que donnerait le même raisonnement mais avec la deuxième série Y'_i du Tableau 8.4. Vous constatez que la différence n'est pas au niveau de la prédiction. L'estimation du modèle augmenté est la même que pour la première série, c'est-à-dire 60, la moyenne. Cependant, dans ce cas-ci, peu de sujets s'écartent de la moyenne et lorsqu'ils le font c'est de très peu de points (deux maximum). C'est donc bien au niveau de l'erreur que les choses changent. Calculons : la SCR ne change pas, 1414-14 = 1400 ; en revanche la SCE(A) est beaucoup plus petite, elle vaut 14! Cela signifie que, cette fois, la réduction de l'erreur est de 1400 mais sur une erreur totale de 1414. En utilisant le modèle augmenté, on ne fait pratiquement plus d'erreur résiduelle. La valeur *F* devient du coup énorme (très largement supérieure à la valeur critique de 4,67), elle vaut 1300. Pour la série 1, la réduction de l'erreur était la même. Cependant, l'erreur totale de départ était beaucoup plus importante et l'erreur résiduelle était encore conséquente. Cette différence est cruciale : lorsque l'on réduit l'erreur, l'impact est différent selon qu'elle soit réduite mais qu'il en reste encore beaucoup (cas de la série 1) ou qu'elle soit réduite et qu'il n'en reste presque plus (cas de la série 2). C'est finalement ce rapport qu'évalue la statistique *F*.

Tableau 8.8. : Tableau d'ANOVA appliqué à la série 2 du Tableau 8.4

Source	*SC*	*ddl*	*CM*	*F*	*p*
Réduction modèle A	SCR	PA - PC	CMR = SCR/(PA - PC)	CMR/CME	
Série 2	*1400*	*1*	*1400/1*	*1300*	<,05
Erreur modèle A	SCE(A)	n - PA	CME = SCE(A)/(n - PA)		
Série 2	*14*	*13*	*14/13 = 1,08*		
Total	SCE(C)	n - PC			
Série 2	*1414*	*14*			

8.5.5. Décision statistique et estimation de la puissance

Nous avons parlé au point 8.4.1 des risques de première espèce, ou risque α. A ce risque correspond ce qu'on appelle l'erreur de type I. Nous avons également évoqué les risques de

seconde espèce, ou risque β. A ce risque correspond l'erreur de type II. Le Tableau 8.9 récapitule les décisions que nous prenons et les risques de se tromper.

Tableau 8.9. : Décision statistique en fonction de la situation réelle et erreur associée

	Situation réelle	
Décision statistique	Modèle C correct	Modèle C incorrect
Rejet Modèle C	*Erreur de type I*	*Décision correcte*
Non rejet Modèle C	*Décision correcte*	*Erreur de type II*

Remarquez que, pour l'instant, nous connaissons bien notre risque d'erreur de type I. Nous l'avons fixé à 5% par convention en psychologie. Nous avons également vu qu'il existe des situations dans lesquelles notre niveau d'exigence doit être plus élevé (le diagnostic d'une maladie par exemple) et d'autres où notre niveau d'exigence peut être plus bas (la recherche d'hypothèses, par exemple). En revanche, nous ne connaissons encore rien concernant la détermination du risque d'erreur de type II. Nous avons dit que l'erreur de seconde espèce se nommait erreur β (qui a donc une probabilité β de survenir) et que la probabilité de ne pas commettre cette erreur ($1-\beta$) était la puissance. Une préoccupation que devrait avoir tout chercheur est de déterminer la puissance. Pourquoi?

L'erreur de type II correspond à ne pas rejeter un modèle compact alors qu'il est mauvais. Donc de ne pas voir que notre modèle augmenté est en fait meilleur que le modèle compact. Si ce risque est trop élevé, mettons qu'il soit égal à 0,5, cela signifie que, même si notre modèle augmenté est meilleur, une fois sur deux nous ne le verrons pas.

Ce point vous sensibilisera d'une part, au concept, et d'autre part, aux éléments qui participent à la détermination de la puissance. En fait, vous avez toutes les clefs en main pour comprendre de quoi dépend la puissance. Il suffit maintenant d'agencer les concepts que vous connaissez.

8.5.5.1. La taille du PRE

Une des principales questions que nous devons nous poser pour déterminer la puissance, est **la taille de la proportion de réduction de l'erreur que l'on veut détecter**. En effet,

supposons qu'une variable diminue significativement l'erreur mais de manière très faible, c'est-à-dire que la variable indépendante ait un effet sur la variable indépendante qui m'occupe, mais un effet assez petit.

Par exemple, imaginons que je me demande ce qui influence l'absentéisme au travail. Il y a de fortes chances qu'un patron particulièrement désagréable ou une charge de travail inhumaine influencent fort cette variable. Mais peut-être que le climat est également une variable à prendre en compte (les travailleurs étant malades quand il fait froid). L'erreur de prédiction sera sans doute réduite de manière plus importante en considérant la charge de travail que le climat. Mais il se peut que le climat diminue l'erreur parce que son effet est réel, même si la PRE est beaucoup plus faible. De ce fait, j'aurai plus de mal à détecter le petit effet lié au climat que le gros effet lié à la charge de travail. Plus la PRE que je veux détecter est petite, plus j'ai de mal à y parvenir (puisqu'elle a plus de risques d'être noyée dans l'erreur d'échantillonnage).

8.5.5.2. L'erreur

Comme nous l'avons vu en comparant la série 1 et la série 2 du Tableau 8.4, **l'erreur peut également nous empêcher de voir les effets d'une variable**. Par exemple, dans la série 1, nous avions une très grande erreur. Mais nous ne contrôlions que très peu de variables chez nos participants. L'influence de toutes ces variables peut générer un "*bruit de fond*" qui rend les effets de l'étude invisibles. Une manière de réduire l'erreur serait par exemple de s'assurer que les étudiants aillent tous dormir tôt, n'ingurgitent aucune substance diminuant l'attention ou l'éveil, aient travaillé un nombre d'heures plus ou moins identique et suffisant, aient eu de bonnes conditions de travail, etc.. Plus on contrôle les variables potentiellement perturbatrices, plus on a de chances de réduire l'erreur. A l'extrême, nous nous trouvons face à la série 2 où l'erreur est minime et permet la détection de taille nettement plus faible. En fait, même si la moyenne de la série 2 avait été de 54, avec le même écart-type, j'aurais considéré la différence avec la valeur 50 (mon modèle compact) comme significative. Vous pouvez vous en assurer en faisant vous-mêmes le calcul. Pourtant une moyenne de 60 peut, comme dans le cas de la série 1, ne pas être significative si l'erreur est trop élevée.

8.5.5.3. Le risque α

Nous avons discuté au point 8.4.1 (encadré) du lien qui unit le risque α au risque β. **Nous avons vu que si nous désirions être sûrs de ne pas commettre d'erreur de type I, cela provoquait nécessairement une plus grande probabilité de commettre une erreur de type II et inversement**. Dès lors, une des manières d'augmenter la puissance (donc de diminuer l'erreur de type II) est d'être moins exigeant sur le risque α. Cette situation explique que ce risque soit fixé à 5% en psychologie et non à moins. On peut se dire que cela signifie qu'une étude sur 20 conduit à de fausses conclusions, mais en fait ce n'est pas le cas. Un élément fondamental de la recherche est que les découvertes importantes sont reproduites par d'autres scientifiques pour s'assurer de la solidité du résultat. Dès lors, une étude qui est reproduite, mettons deux fois, aurait moins d'une chance sur cent de ne pas être valide. En revanche, opter pour un risque α de 5% permet de conserver une puissance correcte, donc d'éviter de refuser trop souvent à tort le rejet du modèle compact.

8.5.5.4. La taille de l'échantillon

Nous avons vu que la forme de la distribution d'échantillonnage se contractait au fur et à mesure que l'échantillon grandissait (Figure 8.3). Cela signifie que **plus l'échantillon est grand, plus l'erreur standard est petite**. C'est d'ailleurs évident si l'on se souvient que l'estimation de l'erreur standard s'obtient en divisant l'écart-type de l'échantillon par la racine carrée de l'effectif de cet échantillon. En conséquence, plus l'échantillon est grand, moins l'intervalle de confiance est grand et plus on va considérer rapidement qu'une valeur différente de la moyenne appartient à une autre distribution d'échantillonnage, donc plus on est puissant.

8.5.5.5. Conclusion

En conclusion, pour détecter de petite PRE, j'ai intérêt à m'arranger pour que l'erreur indésirable soit la plus petite possible et que l'échantillon soit le plus grand possible. J'ai également intérêt à augmenter mon risque de première espèce, mais évidemment, ce faisant, je risque de commettre l'erreur qui y est associée, de sorte que ce n'est pas une réelle solution.

Ce chapitre sur la puissance devrait vous éclairer sur les enjeux principaux de l'inférence statistique. A partir d'un échantillon, nous avons trouvé un modèle augmenté et nous nous

sommes demandé s'il était plus performant que le modèle compact basé sur le hasard. Nous avons ensuite vu que, pour répondre à cette question, nous avions intérêt à ce que la proportion de réduction de l'erreur soit maximale et que l'erreur résiduelle soit minimale. L'approche par intervalle de confiance nous a montré que nous désirions réduire la distribution d'échantillonnage au minimum de manière à avoir l'estimation de la moyenne de la population la plus précise possible (la fourchette de valeurs la plus petite possible). De cette manière, nous pouvions détecter avec une relativement bonne certitude les valeurs qui sont suffisamment éloignées de l'estimation de la moyenne de la population comme appartenant à une autre distribution d'échantillonnage, centrée sur une autre valeur.

Pour ce qui est du calcul de la puissance, la difficulté provient justement du nombre de facteurs qui la définissent. En effet, puisqu'elle dépend de la taille de l'échantillon, du niveau de risque d'erreur de type I, de la taille de la PRE à détecter et de l'erreur, il est extrêmement difficile de créer une table facilement lisible qui prenne en compte tous ces facteurs. Cependant, certaines de ces informations sont conventionnellement constantes ce qui peut rendre la tâche plus facile. Par exemple, il est rare de calculer une puissance pour détecter des effets très grands comme des PRE supérieurs à .30. En effet, bien souvent on désirera être plus précis que cela et détecter des PRE inférieures à cette valeur. Il est également habituel de considérer un risque α de 5% et rare d'en envisager un autre dans le domaine de la psychologie. En revanche, l'effectif et, conjointement, les degrés de liberté sont eux très fluctuants. Le Tableau 8.10 vous montre une table de puissance basée sur un $\alpha = 5\%$, et un degré de liberté du modèle compact nul (PC = 0) et du modèle augmenté égal à un (PA = 1).

Tableau 8.10. : Table de puissance pour α = .05 quand PC = 0 et PA = 1. Source : Judd, McClelland, Ryan, Muller & Yzerbyt (2010), p. 85.

			Prob(PRE > *Valeur critique*)							
	Valeur Critique		*Vrai* PRE, η^2							
n	*F*	PRE	*0*	*.01*	*.03*	*.05*	*.075*	*.1*	*.2*	*.3*
2	161.45	.994	.05	.05	.05	.05	.05	.06	.06	.07
3	18.51	.903	.05	.05	.05	.06	.06	.07	.08	.11
4	10.13	.771	.05	.05	.06	.06	.07	.08	.11	.15
5	7.71	.658	.05	.05	.06	.07	.08	.09	.14	.21
6	6.61	.569	.05	.05	.06	.07	.09	.10	.17	.26
7	5.99	.499	.05	.06	.07	.08	.10	.12	.20	.31
8	5.59	.444	.05	.06	.07	.09	.11	.13	.23	.36
9	5.32	.399	.05	.06	.08	.09	.12	.14	.26	.41
10	5.12	.362	.05	.06	.08	.10	.13	.16	.29	.46
11	4.96	.332	.05	.06	.08	.11	.14	.17	.32	.50
12	4.84	.306	.05	.06	.09	.11	.15	.18	.35	.54
13	4.75	.283	.05	.06	.09	.12	.16	.20	.38	.58
14	4.67	.264	.05	.06	.09	.13	.17	.21	.41	.62
15	4.60	.247	.05	.07	.10	.13	.18	.23	.44	.66
16	4.54	.232	.05	.07	.10	.14	.19	.24	.46	.69
17	4.49	.219	.05	.07	.10	.14	.20	.25	.49	.72
18	4.45	.208	.05	.07	.11	.15	.21	.27	.52	.74
19	4.41	.197	.05	.07	.11	.16	.22	.28	.54	.77
20	4.38	.187	.05	.07	.12	.16	.23	.29	.56	.79
22	4.32	.171	.05	.07	.12	.18	.25	.32	.61	.83
24	4.28	.157	.05	.08	.13	.19	.27	.35	.65	.87
26	4.24	.145	.05	.08	.14	.20	.29	.37	.69	.89
28	4.21	.135	.05	.08	.15	.22	.31	.40	.72	.92
30	4.18	.126	.05	.08	.15	.23	.33	.42	.75	.93
35	4.13	.108	.05	.09	.17	.26	.37	.48	.82	.96
40	4.09	.095	.05	.10	.19	.29	.42	.54	.87	.98
45	4.06	.085	.05	.10	.21	.32	.46	.59	.91	.99
50	4.04	.076	.05	.11	.23	.36	.51	.64	.93	**
55	4.02	.069	.05	.11	.25	.39	.55	.68	.95	**
60	4.00	.064	.05	.12	.27	.42	.58	.72	.97	**
80	3.96	.048	.05	.14	.34	.53	.71	.84	.99	**
100	3.94	.038	.05	.17	.41	.62	.81	.91	**	**
150	3.90	.026	.05	.23	.57	.80	.93	.98	**	**
200	3.89	.019	.05	.29	.70	.90	.98	**	**	**
500	3.86	.008	.06	.61	.98	**	**	**	**	**

** puissance > .995

Prenons quelques exemples pour lire cette table. La première colonne (lorsque la vraie PRE est nulle c'est-à-dire lorsque le modèle augmenté n'apporte aucune information supplémentaire). Nous voyons que dans ce cas, quelle que soit la taille de l'échantillon (sauf 500) nous avons bien 5% de risques de percevoir une PRE significative malgré le fait qu'en réalité il n'y en ait pas (c'est ce qu'on veut). Ensuite, plus on augmente la valeur de PRE, plus on a facile à la détecter. Si on considère 0,80 comme puissance acceptable, on voit qu'une

PRE de 0,01 est pratiquement indétectable (même pour un échantillon de 500 personnes, on est qu'à 0,61 de puissance). En revanche, il est assez aisé de détecter des PRE de 0,30 (une vingtaine de personnes suffisent).

Elaborer de telles tables pour tous les degrés de liberté de la statistique *F* est ardu. Il est également nécessaire de trouver des tables pour les autres distributions comme les *t* de Student. Dès lors, il est souvent plus simple d'utiliser des logiciels informatiques. G*Power[37] est un tel logiciel, gratuit, facile d'utilisation, et très performant, que nous utiliserons l'année prochaine. Nous distinguerons également deux stratégies différentes : (a) la détermination *a priori* de la puissance, qui permet de déterminer les conditions expérimentales nécessaires pour détecter les effets que l'on désire ; et (b) la détermination *a posteriori* de la puissance qui permet de savoir si on avait une chance de percevoir un effet existant avec notre plan expérimental (en général il vaut évidemment mieux le savoir avant).

8.6. Equivalence entre la comparaison par modèles et le test-*t* à un échantillon

La comparaison par modèle nous autorise à définir si le modèle augmenté est plus intéressant que le modèle compact, avec un risque de 5% de se tromper dans notre décision (erreur de type I). Cette conclusion est en fait exactement la même que celle que nous obtenions à l'aide des intervalles de confiance. En effet, on concluait pour la série 1 du Tableau 8.4 que la moyenne de 60 n'était pas plus intéressante à utiliser que la valeur 50 correspondant à une réponse au hasard à l'interrogation. Cela revient à dire que la valeur 50 est inclue dans l'intervalle de confiance construit autour de la valeur 60. De même, on concluait, pour la série 2 du même tableau, que la moyenne de 60 était une prédiction plus intéressante à utiliser que la valeur 50. Cela revenait à estimer que la valeur théorique de 50 était cette fois en dehors des limites de l'intervalle de confiance.

Le fait que ces conclusions soient similaires implique deux choses. Premièrement, il doit y avoir moyen de construire un intervalle de confiance à l'aide de la distribution *F*. Deuxièmement, il doit exister un lien entre la distribution *F* et la distribution *t* de Student qui autorise l'utilisation d'une table ou de l'autre.

[37] http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

8.6.1. Construction de l'intervalle de confiance à l'aide d'une distribution F ou d'une distribution t

Dans ce cadre, nous allons utiliser deux informations : celle que l'on trouve dans la table *F* et le carré moyen de l'erreur. La valeur autour de laquelle nous allons calculer l'intervalle de confiance est la valeur β_0 du modèle augmenté. Cela devrait vous sembler évident puisque le principe est, finalement, de construire la distribution d'échantillonnage correspondant à la moyenne de l'échantillon et d'en prendre la densité de probabilité égale à 95% (avec les 5% de rejet réparti symétriquement sur les extrémités de la courbe normale de la distribution d'échantillonnage).

Le départ est donc, comme nous l'avons vu auparavant, de calculer l'erreur standard. Cette erreur est la racine carré du carré moyen de l'erreur du modèle A. En effet, le CME est la somme des écarts par rapport à la moyenne élevée au carré divisé par le nombre de degrés de liberté. C'est donc la variance corrigée de l'échantillon. Cela se voit bien en regardant la quatrième colonne du Tableau 8.5 (la SCE augmenté) qui est la somme des carrés des écarts par rapport à la moyenne. Cette valeur est divisée par le nombre de degrés de liberté dans le Tableau 8.6 pour obtenir la CME. Il s'agit donc bien de la variance de la distribution du modèle A. Pour obtenir l'erreur standard, on divise l'écart-type corrigé de l'échantillon (qui est une estimation de l'écart-type de la population) par la racine carré de l'effectif de l'échantillon. Cela revient à prendre la racine carré de la variance corrigée de l'échantillon divisée par l'effectif de l'échantillon. Donc pour la série 1 :

$$\sigma_M = \sqrt{\frac{CME}{n}} = \sqrt{\frac{S^2}{n}} = \frac{S}{\sqrt{n}} = \sqrt{\frac{590,77}{14}} = 6,50$$

Vous vous rappelez qu'à un écart-type de part et d'autre de la moyenne, seuls 68% des scores sont englobés (Figure 7.4) et il fallait donc multiplier par 1,96 cet écart-type pour obtenir 95% des valeurs. La valeur correspondante à ce principe serait pour la distribution *t* de Student de 2,160 (voir Tableau 8.3, pour 13 degrés de liberté). Sur la distribution F, il s'agit de la valeur 4,67 (voir Tableau 8.7). Cependant, la table *F* est une table qui concerne les distributions de variances, et non d'écarts-types. Dès lors, c'est la variance de l'échantillonnage qui doit être multipliée par 4,67 et non l'écart-type. Nous obtenons donc :

$$IC = \beta_0 \pm \sqrt{\frac{F_{(1,n-1,\alpha)}CME}{n}} = 60 \pm \sqrt{\frac{4{,}67 * 590{,}77}{14}} = 60 \pm 14{,}03$$

Nous voyons donc bien que l'intervalle de confiance englobe la valeur 50 puisque la moyenne de la population est comprise entre 45,97 ≤ μ ≤ 74,03. Nous ne pouvons donc pas considérer que 50 est une valeur qui appartient à une autre population que la valeur de 60.

Cette même conclusion est obtenue avec la construction de l'intervalle de confiance qui emploi la distribution *t* de Student (et comme nous avons dû estimer la variance de la population c'est bien cette table qui s'appliquait et non la distribution normale). Dans ce cas nous aurions eu :

$$IC = \beta_0 \pm t_{(n-1)}\sigma_M = \beta_0 \pm t_{(n-1)}\sqrt{\frac{CME}{n}} = 60 \pm 2{,}16 * 6{,}50 = 60 \pm 14{,}03$$

Vous devriez être tout à fait capables de réaliser le même exercice pour la série 2 du Tableau 8.4, mais le Tableau 8.11 s'en charge pour vous. On y retrouve les mêmes procédures que celles que nous venons de voir, mais la dispersion, et donc l'erreur, étant nettement moins grande, l'intervalle de confiance est nettement plus petit et la valeur 50 en sort.

Tableau 8.11. : Comparaisons intervalles de confiance par la table *F* ou *t*.

$IC = \beta_0 \pm t_{(n-1)}\sqrt{\frac{CME}{n}}$	$IC = \beta_0 \pm \sqrt{\frac{F_{(1,n-1,\alpha)}CME}{n}}$
$IC = 60 \pm 2{,}16 * \sqrt{\frac{1{,}08}{14}} = 60 \pm 0{,}60$	$IC = 60 \pm \sqrt{\frac{4{,}67 * 1{,}08}{14}} = 60 \pm 0{,}60$
Conclusion : 59,4 ≤ μ ≤ 60,60 et donc 50 est largement hors de l'IC.	

8.6.2. Lien entre la distribution F et la distribution t

Puisque la distribution *F* est applicable à des variances et que la distribution *t* est applicable à des écarts-types, la relation entre les deux distributions devient évidente. Elle dépend des

nombres de degrés de liberté qui doivent être comparables, de sorte que le lien entre une distribution *t(n-1)* de Student et la distribution *F(1, n-1)* devient :

$$t_{(n-1)} = \sqrt{F_{(1,n-1)}}$$

Certains d'entre vous s'en sont probablement doutés en observant les équations de l'intervalle de confiance avec l'utilisation de la table *F* et avec l'utilisation de la table t. La seule chose qui change est que le *F(1,n-1,α)* est sous la racine carré tandis que la valeur de *t(n-1)* est hors de cette racine.

8.6.3. Alternative à la méthode de l'intervalle de confiance : le test-t pour un échantillon

Souvent, dans la littérature scientifique, l'intervalle de confiance n'est pas utilisé (et c'est, à mon sens, dommage). L'alternative plus fréquente est de présenter les résultats sous forme de test-*t*. Il s'agit en fait exactement de la même chose que l'intervalle de confiance, sauf que dans ce cas, nous allons calculer un *t* observé et le comparer à une valeur *t* théorique correspondant au risque α désiré. Si notre valeur *t* observée est supérieure à la valeur du *t* théorique, on conclura que la valeur théorique à laquelle nous comparons la moyenne de notre échantillon est en dehors de l'intervalle de confiance, c'est-à-dire que la moyenne de l'échantillon n'appartient pas à une distribution d'échantillonnage qui inclut cette valeur théorique.

Il reste à comprendre comment trouver la valeur *t* observée. Au point 8.4.2, nous avons vu que l'intervalle de confiance se déclinait comme suit :

$$\overline{X} - t_{(ddl,\alpha)}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{(ddl,\alpha)}\frac{S}{\sqrt{n}}$$

Nous pourrions, en ne considérant que la partie droite de cette relation, la transformer assez facilement, en isolant le terme $t_{(ddl,\alpha)}$ en :

Formule du test-t alternatif à la comparaison par IC

$$\frac{\mu - \overline{X}}{\frac{S}{\sqrt{n}}} \leq t_{(ddl,\alpha)}$$

Dès lors, si le terme de gauche est inférieur (en valeur absolue) à la valeur $t_{(ddl,\alpha)}$ trouvée dans la table, nous en conclurons que la moyenne de l'échantillon et la moyenne théorique sont dans le même intervalle de confiance (cela revient à considérer que leur différence n'est pas significativement éloignée de zéro). En revanche, si le terme de gauche est supérieur à la valeur du $t_{(ddl,\alpha)}$ théorique, nous en rejetterons l'hypothèse que la distribution d'échantillonnage de la moyenne de l'échantillon inclut la moyenne de la population théorique (μ) considérée.

Lorsque vous adoptez cette approche, vous rapporterez la valeur du $t_{(ddl,\alpha)}$ observée et indiquerez si la p-valeur est supérieure ou inférieure à 0,05. En reprenant l'exemple que nous avons utilisé pour la comparaison de modèle cela donne :

Application de l'approche par test-*t* aux séries 1 et 2 du tableau 8.5

Série 1	**Série 2**
$\frac{50-60}{\frac{24,31}{\sqrt{14}}} = -1,54$	$\frac{50-60}{\frac{1,04}{\sqrt{14}}} = -35,98$

Valeur du *t* théorique pour ddl = 13 et α = 0,05 : ***t* = 2,16** (voir table 8.3)

p-valeur = 0,15 (>0,05, n.s.)	p-valeur = 2.10^{-14} (<0,05, s.)
Remarque : $1,54^2 \cong 2,37$ (valeur de la *F*)	Remarque : $35,98^2 \cong 1300$ (valeur de la *F*)

Remarquez que, dans l'encadré ci-dessus, je rapporte la vraie p-valeur. Vous n'aurez pas à le faire, vous vous contenterez de signaler si la valeur observée est supérieure ou inférieure à la valeur théorique. Cependant, les logiciels de statistique, tel SPSS, effectuent le calcul précis en utilisant l'équation de la distribution et rapportent la vraie p-valeur.

8.6.4. Pourquoi présenter les trois méthodes si la distribution t de Student marche très bien?

Plusieurs raisons m'ont motivé à vous présenter ces trois méthodes, mais deux sont essentielles. La première est que depuis le début j'essaie d'utiliser une méthode d'enseignement en spirale. Je vous redis plusieurs fois les mêmes concepts mais je les utilise dans différents contextes de manière à vous imprégner petit à petit de la matière. Donc, présenter trois méthodes pour dire la même chose pouvait vous aider à comprendre.

La seconde raison est que, si le test-*t* est probablement plus aisé à comprendre que l'approche par comparaison de modèles, il n'en ira pas de même lorsque les méthodes se complexifieront par la suite. En effet, pour passer à des modèles plus complexes, l'approche par comparaison de modèle ne demande que peu d'efforts alors que les test-*t* peuvent tout de suite être abandonnés au profit de l'ANOVA. En revanche, il était obligatoire de présenter le test-*t* également parce que, d'une part, ce test se retrouve fréquemment dans la littérature et que, d'autre part, il permet plus facilement d'expliquer les intervalles de confiance.

8.7. Exercices de fin de chapitre

T.P. 8 – Exercice 1

Standardisation - score Z

1. Au début de l'année scolaire, un enseignant fait passer un test de français à ses élèves pour connaître leur niveau en cette matière. Les résultats de ce test sont transformés en scores Z. Un élève obtient un score de Z = + 2,4. Que peut-on dire de sa note?

2. Une distribution de notes à un examen a une moyenne de 50 et un écart type de 8. Pour cette distribution, quels sont les scores Z correspondant à X = 58 et X = 46. Pour cette distribution, quelle note correspond à un score Z de +2 ? Quelle formule générale pouvez-vous tirer de vos calculs ?

3. Déterminer la valeur z correspondant aux résultats obtenus par dix élèves à une interrogation en mathématique (notée sur 20) :

N	1	2	3	4	5	6	16	8	9	10
Note (/20)	14	15.5	13	12	19	17	16	16	11	9
Score Z										

T.P. 8 – exercice 2

Application du score Z à la distribution normale

Notion de probabilité – distribution normale – score Z – Utilisation de la table Z

On considère une variable aléatoire X dont la distribution est normale, continue et symétrique par rapport à l'origine.

On sait – par calcul ou par lecture d'une table – que dans une distribution normale :

$$P(-1 \leq Z \leq +1) = 0{,}6826 = 68{,}26\%$$

$$P(-2 \leq Z \leq +2) = 0{,}9544 = 95{,}44\%$$

$$P(-3 \leq Z \leq +3) = 0{,}9974 = 99{,}74\%$$

En présence d'une distribution normale, quel pourcentage de données est (approximativement) situé…

1. Entre m et m + 3 σ ? Représentez graphiquement la zone concernée (à main levée).

2. Entre m - 1 σ et m + 2 σ ? Représentez graphiquement la zone concernée (à main levée).

3. En-deça de m +1 σ ? Représentez graphiquement la zone concernée (à main levée).

4. Au-delà de m + 1 σ ? Représentez graphiquement la zone concernée (à main levée).

T.P. 8 – Exercice 3
Les distributions normales

Concernant la distribution du quotient intellectuel, nous savons que la moyenne (μ) est égale à 100 et que l'écart type (σ) vaut 15.

1. Veuillez représenter cette distribution et donner les valeurs de la moyenne et des différents écart-types.

2. Sachant cela, quel est le pourcentage d'individus… :

- Ayant un QI compris entre 70 et 115 ?

- Ayant un QI inférieur à 85 ?

- Ayant un QI supérieur à 55 ?

3. Avec l'aide de la table de la distribution normale standardisée, veuillez déterminer la probabilité d'obtenir un score compris …

- Entre -.56 et 1.78

- Entre 0 et - .91

- Entre 0.57 et 1.57

T.P. 8 – exercice 4

Application du score Z à la distribution normale

Notion de probabilité – distribution normale – score Z

1. Un chercheur s'intéresse à la répartition des QI en Belgique. Il sait que la moyenne est de 100 et l'écart-type de 15. D'après le test Wais-R, on considère qu'une personne avec un QI inférieur à 85 est atteinte de débilité légère et s'il est inférieur à 70, de débilité profonde. A l'opposé, une personne avec un QI supérieur à 125 est surdouée.

Intrigué par ces catégories, il veut vérifier quels pourcentages de la population entrent dans ces différentes catégories.

a) Vérifiez quel pourcentage de la population fait partie des catégories « débile léger » ou « débile profond ».

b) Vérifiez quel pourcentage de la population fait partie de la catégorie « surdoué »

c) Vérifiez quel pourcentage de la population est « normale » (c'est-à-dire entre les catégories « débile léger » et « surdoué »)

d) Vérifiez quel pourcentage de la population présente une « débilité légère » mais pas « profonde »

2. D'après une étude inspirée des travaux C. Murray, les gens ont tendance à s'entourer de personnes qui ont un QI de ±15 points par rapport au leur.

a) Suivant ce principe, quel pourcentage de la population satisfait à ce critère pour une personne ayant un QI de 117.

b) En vous basant sur le résultat trouvé au point a) (arrondi à 2 décimales), quelle est la probabilité que cette personne rencontre exactement 3 autres personnes qui satisfont à ce critère parmi un groupe de 12 personnes.

c) En vous basant sur le résultat trouvé au point a) (arrondi à 2 décimales), quelle est la probabilité que cette personne rencontre au moins 2 autres personnes qui satisfont à ce critère dans un groupe de 7 personnes.

d) En suivant ce principe, quelle est la probabilité qu'une personne ayant un QI de 130 rencontrent exactement 1 autre personne qui satisfait à ce critère dans un groupe de 10 personnes.

e) En suivant ce principe, quelle est la probabilité qu'une personne ayant un QI de 145 rencontrent au moins 5 autres personnes qui satisfont à ce critère dans un auditoire de 600 personnes.

f) Quelle autre distribution (citée dans le cours) serait plus adaptée pour répondre à la question précédente et pourquoi ?

T.P. 8 – Exercice 5

Propriétés de la variance et de l'écart type

Complétez le tableau ci-dessous. Complétez les trois dernières colonnes en utilisant ce que vous savez des propriétés de la moyenne, de l'écart type et de la variance, sans effectuer de calculs complexes, puis vérifiez votre calcul en utilisant les formules des statistiques en question.

X_i	5	4	2	3	9	5	$\overline{X} = 4{,}67$	$S_X^2 = 4{,}89$	$S_X = 2{,}21$
$X_i' = 2X_i$							$\overline{X'} =$	$S_{X'}^2 =$	$S_{X'} =$
$X_i'' = X_i + 1$							$\overline{X''} =$	$S_{X''}^2 =$	$S_{X''} =$
$X_i''' = X_i - \overline{X}$							$\overline{X'''} =$	$S_{X'''}^2 =$	$S_{X'''} =$
$X_i'''' = Z_i = \frac{X'''}{S_X} = \frac{X_i - \overline{X}}{S_X}$							$\overline{Z} =$	$S_Z^2 =$	$S_Z =$

1. Sur base des résultats que vous venez de calculer, déduisez la relation liant :

1) $\overline{X'}$ à $\overline{X}$, $S_{X'}^2$ à S_X^2 et $S_{X'}$ à S_X :
2) $\overline{X''}$ à $\overline{X}$, $S_{X''}^2$ à S_X^2 et $S_{X''}$ à S_X :
3) $\overline{X'''}$ à $\overline{X}$, $S_{X'''}^2$ à S_X^2 et $S_{X'''}$ à S_X :

2. Quelle est la caractéristique de la série statistique constituée des observations $\boldsymbol{Z_i, i = 1,2,...6}$?

T.P. 8 – exercice 7
Calcul de probabilités à partir d'une distribution normale de tailles

Dans une population donnée, la taille des sujets se distribue suivant une loi normale de moyenne égale à 172 cm et d'écart- type égal à 22 cm.

On sélectionne au hasard un individu dans cette population.

1. Calculez, en supposant que les tailles soient mesurées au cm près, la probabilité qu'un individu ait une taille entre 179,5 et 180,5 cm. Arrondissez à trois décimales.

2. Calculez, en supposant que les tailles soient mesurées au mm près, la probabilité que l'individu sélectionné ait une taille entre 179,9 et 180,1 cm.

3. Que vaut, pour une variable aléatoire ayant une distribution continue, normale, la probabilité que la variable aléatoire prenne EXACTEMENT une valeur particulière ? Expliquez.

T.P. 8 – Exercice 8
Distribution normale – Utilisation de la table standardisée

1. En sachant que la distribution est normale, continue et symétrique, on peut déduire du graphique ci-dessus, la probabilité liée à différentes surfaces représentées sous la courbe. Quand c'est possible, basez-vous sur ce que vous avez déjà calculé.

	PROBABILITÉ CONSIDÉRÉE	VALEUR NUMÉRIQUE	JUSTIFICATION
1.	$P(0 \le Z \le 1)$		
2.	$P(Z \le 0)$		
3.	$P(Z \le 1)$		
4.	$P(Z > 1)$		
5.	$P(Z = 1)$		
6.	$P(Z \ge 1)$		
7.	$P(\lvert Z \rvert \le 1)$		
8.	$P(Z \ge 0)$		

9.	$P(-2 \le Z \le -1)$		

10.	$P(1 \leq Z \leq 2)$		

2. Sur base de la table de distribution normale des scores Z, déterminez les probabilités suivantes :

	PROBABILITÉ RECHERCHÉE	VALEUR NUMÉRIQUE
1.	$P(Z \leq 1)$	
2.	$P(Z \leq -3)$	
3.	$P(Z \geq 2)$	
4.	$P(Z \leq 2)$	
5.	$P(Z \leq -2)$	
6.	$P(Z \geq -2)$	
7.	$P(Z \leq 1{,}65)$	
8.	$P(0 \leq Z \leq 1{,}53)$	
9.	$P(0 \leq Z \leq 0{,}58)$	
10.	$P(-1{,}25 \leq Z \leq 0)$	
11.	$P(Z \geq 1{,}95)$	
12.	$P(-1{,}61 \leq Z \leq 0{,}73)$	
13.	$P(-1 \leq Z \leq +1)$	
14.	$P(-2 \leq Z \leq +2)$	
15.	$P(-3 \leq Z \leq +3)$	
16.	$P(-4 \leq Z \leq +4)$	
17.	$P(Z \geq -2{,}04)$	
18.	$P(Z \geq -1{,}72)$	
19.	$P(Z \geq 1{,}94)$	
20.	$P(Z \leq -0{,}48)$	

3. Commentez les résultats obtenus pour $P(-1 \leq Z \leq +1)$ et $P(-4 \leq Z \leq +4)$

4. Sur base de la table de distribution Z, exprimez mathématiquement les probabilités suivantes :

	VALEUR NUMÉRIQUE	PROBABILITÉ RECHERCHÉE 1	PROBABILITÉ RECHERCHÉE 2
0.	.5000		
1.	.9975		
2.	.9564		
3.	.2177		
4.	.0000		
5.	.3632		
6.	.6844		
7.	.002		
8.	.0119		
9.	.7422		
10.	.1064		
11.	.4983		

T.P. 8 CHAPITRE 8 - EXERCICE 9
LOI NORMALE

1. Dans un échantillon, la moyenne est de 62 kg pour le poids et de 168 cm pour la taille. L'écart type est respectivement de 8 et de 5. Une personne a un score standard de 0 pour le poids et de 0 pour la taille. Si on la compare aux autres membres de notre échantillon, peut-on dire qu'elle est plus lourde que haute, qu'elle est plus haute que lourde ou qu'elle est aussi lourde que haute ? Expliquez votre réponse.

2. Dans ce même échantillon, une personne a un score standard de 1 pour le poids et de 1 pour la taille. Si on la compare aux autres membres de notre échantillon, peut-on dire qu'elle est plus lourde que haute, qu'elle est plus haute que lourde ou qu'elle est aussi lourde que haute ? Expliquez votre réponse.

3. Dans ce même échantillon, une personne a un score standard de 0,65 pour le poids et de –0,25 pour la taille. Si on la compare aux autres membres de notre échantillon, peut-on dire qu'elle est plus lourde que haute, qu'elle est plus haute que lourde ou qu'elle est aussi lourde que haute ? Expliquez votre réponse.

4. Une personne a un score standard de 0,65 pour le poids et de 0,75 pour la taille. Si on la compare aux autres membres de notre échantillon, peut-on dire qu'elle est plus lourde que haute, qu'elle est plus haute que lourde ou qu'elle est aussi lourde que haute ? Expliquez votre réponse.

5. De quels types de distributions la distribution normale est-elle la forme idéalisée ?

6. Qu'est-ce que la distribution d'échantillonnage de la moyenne ?

T.P. 8 – Exercice 10

Score Z et ses usages

1. Une distribution A se distribue normalement autour d'une moyenne de 20 et d'un écart type de 7. Une distribution B se distribue normalement autour d'une moyenne de 23 et d'un écart type de 2. Dans quelle distribution un score de 27 est-il mieux placé par rapport aux autres ? Expliquez votre réponse.

2. Pourquoi est-il possible de comparer des scores provenant de distributions différentes après que chaque distribution soit transformée en scores Z ?

T.P. 8 – Exercice 11

Utilisation de la table de distribution normale standardisée

1. Combien de tables faudrait-il pour déterminer les probabilités pour toutes les variables ?

2. À l'aide de la table, trouvez la proportion d'une distribution normale qui est localisée strictement dans la partie au-dessus des scores Z suivants :

1.1. Z = + 1	
1.2. Z = + 0,72	
1.3. Z = - 2	
1.4. Z = - 0,33	

3. Trouvez la proportion de la distribution normale située entre la moyenne et les scores Z suivants :

2.1. Z = +0,67	
2.2. Z = -1,5	
2.3. Z = - 0,5	

4. Trouvez la proportion d'une distribution normale située entre les scores Z suivants :

3.1. Entre Z = - 0,5 et Z = + 0,5	
3.2. Entre Z = -1 et Z = +1	

5. Trouvez la proportion d'une distribution normale située en dessous des scores Z suivants:

4.1. Z = + 0,2	
4.2. Z = - 0,71	

T.P. 8 – exercice 12

Distribution normale

On considère une variable aléatoire X dont la distribution est normale, continue et symétrique par rapport à l'origine.

On sait – par calcul ou par lecture d'une table – que dans une distribution normale :

$$P(-1 \leq Z \leq +1) = 0{,}6826 = 68{,}26\%$$

$$P(-2 \leq Z \leq +2) = 0{,}9544 = 95{,}44\%$$

$$P(-3 \leq Z \leq +3) = 0{,}9974 = 99{,}74\%$$

1. Représentez sur une courbe normale standardisée les probabilités données ci-dessus.

2. Placez sur le graphique ci-dessous les proportions liées aux différentes surfaces indiquées.

T.P. 9 – 10 : Chapitre 8

Intervalle de confiance et Modélisation

Exercice 1 : Type de distribution

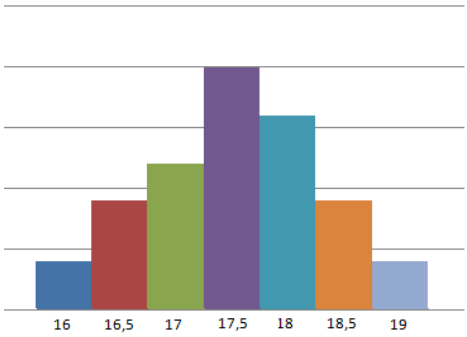
Un professeur souhaite connaitre la moyenne d'âge des 700 élèves de première bac et demande à trois assistants de rechercher cette information pour lui. L'assistant A, décide de calculer la moyenne sur base de l'entièreté de la classe. L'assistant B, quant à lui, prélève un échantillon de 60 personnes et en infère la moyenne de la population totale. L'assistant C, finalement, demande comme exercice pratique à 20 élèves de BAC 2 de chacun, isolément et sans se concerter, calculer la moyenne d'âge de 30 étudiants de bac 1 sélectionnés aléatoirement. Il effectue ensuite un histogramme sur base des vingt moyennes récoltées. À quel type de distribution s'intéresse chacun des assistants ?

	Distribution de l'échantillon	Distribution de la population	Distribution d'échantillonnage
Assistant A			
Assistant B			
Assistant C			

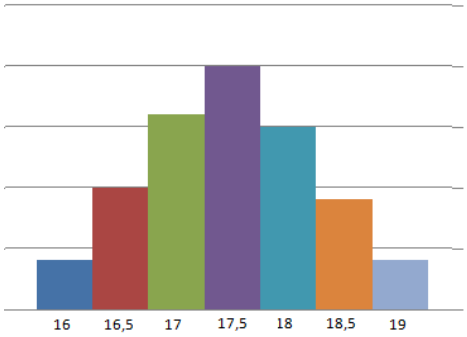
T.P. 9 – 10 : Chapitre 8

Exercice 2 : Propriétés d'une distribution sur base graphique

Voici deux histogrammes qui représentent la distribution d'échantillonnage de l'IMC des mannequins belges. L'un a été constitué sur base des moyennes d'IMC calculées par 156 expérimentateurs différents. L'autre, sur base de seulement 18 expérimentateurs.



A



B

1. Laquelle de ces deux distributions a été constituée sur base des moyennes calculées par 156 expérimentateurs ? Expliquez votre réponse.

2. Laquelle de ces deux distributions sera associé à la plus petite erreur standard ?

3. Quelle forme prendrait cette distribution si nous avions une infinité d'expérimentateurs ayant mesuré l'IMC moyen des mannequins belges ?

T.P. 9 – 10 : Chapitre 8

Exercice 3 : Distributions de Student et Normale.

1. Dans la table de la distribution de Student, relevez les valeurs qui correspondent à l'intervalle qui regroupe 95% des scores de la variables aléatoires pour : k = 30, k = 60, k = 100 et k = ∞. Que remarquez-vous ?

T.P. 9 – 10 : Chapitre 8

Exercice 4 : Type de distribution

1. Vrai ou Faux ? Justifiez en cas de nécessité.

Echantillonnage aléatoire simple : Tous les échantillons possibles de même taille ont la même probabilité´ d'être choisis et tous les ´éléments de la population ont une chance ´égale de faire partie de l''echantillon	V – F
Si nous prélevons un ´échantillon de taille n dans une population donnée, la moyenne de l''echantillon nous donnera une idée approximative de la moyenne de la population.	V – F
Dans le cadre de la distribution d'échantillonnage, plus l'effectif augmente, plus l'erreur standard diminue.	V – F
le cas de grands échantillons (n>30) et lorsque l'on connait la variance, la distribution de l'erreur standard est considérée comme suivant une distribution de student.	V – F
le cas de petits échantillons (n>30) et lorsque l'on ne connait pas la variance, la distribution de l'erreur standard est considérée comme suivant une distribution de normale.	V – F
En psychologie, la convention fixe le risque d'erreur à 10% (α)	V – F

T.P. 9 – 10 : Chapitre 8
Exercice 5 : Intervalles de confiance

1. Pour déterminer l'âge moyen de ses clients, une grande entreprise de confection pour hommes prélève un échantillon aléatoire de 50 clients et trouve 36ans. On suppose que la variance (corrigée) est de 144. Trouvez un intervalle de confiance à 99% pour la moyenne de l'âge μ de l'ensemble de ses clients.

T.P. 9 – 10 : Chapitre 8
Exercice 6 : Intervalles de confiance

Nous voulons estimer la taille moyenne d'un groupe de 2000 filles, et extrayons à cette fin un échantillon de 50 filles. Au sein de cet échantillon, la taille moyenne est de 168 cm et l'écart type de la population est connu et vaut de 5 cm. Déterminez …

1. Un I.C. autour de la moyenne dont le risque α associé est de .05. Concluez cette information par une phrase pour inférer.

2. Un I.C. autour de la moyenne (1-α) à .90. Concluez cette information par une phrase.

3. On choisit ensuite un plus grand échantillon, ayant exactement la même moyenne et le même écart type. On trouve pour ce nouvel échantillon que l'intervalle de confiance autour de la moyenne à .90 est borné à 167,8 et 168,2. Sur base des résultats obtenus ci-dessus, pouvez-vous retrouver de combien de sujets il est composé ?

T.P. 9 – 10 : Chapitre 8
Exercice 7 : Intervalles de confiance

1. Voici les résultats obtenus par 15 élèves à un examen d'histoire (noté sur 10), prélevés au hasard parmi une classe de 100 élèves.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X_i	9	9	6	7	5	6	6	8	8	7	8	5	6	4	7

Sachant que la moyenne de l'échantillon est de 6,73 et que l'écart-type estimé est de 1.49, veuillez trouver l'intervalle de confiance autour de la moyenne ayant 95% de chances de contenir la moyenne exacte de la classe totale.

2. Dans quel pourcentage de cas aurais-je raison si j'affirme que les adolescents de ce groupe ont vu ce film : a) 3 fois ? b) 6 fois ?

T.P. 9 – 10 : Chapitre 8

Exercice 8 : Intervalles de confiance

1. D'après une étude californienne[38], le temps de sommeil recommandé par nuit serait d'environ 7h15 (435 minutes). D'après eux, le manque de sommeil, tout comme le surplus serait lié à un risque de mortalité plus élevé (ils n'ont cependant pas démontré de lien de causalité). Un chercheur convaincu par ces résultats veut vérifier si la population belge dort le nombre approprié d'heures par jour. Il prend un échantillon de 100 personnes et observe que leur temps moyen de sommeil par jour est de 457 minutes (un peu moins de 8 heures) avec un écart-type (corrigé) de 89 minutes.

    a. A quelle valeur allons-nous comparer la moyenne de notre échantillon ? Est-ce que la variance de la population est connue ? La distribution d'échantillonnage suit-elle une distribution normale ou t de student ? (si c'est une student, précisez le nombre de degrés de liberté)

    b. Est-ce que la moyenne obtenue par l'échantillon est une bonne estimation de la moyenne de la population ? Est-ce une bonne estimation de la moyenne de la distribution d'échantillonnage ?

[38] Daniel F. Kripke, MD; Lawrence Garfinkel, MA; Deborah L. Wingard, PhD; Melville R. Klauber, PhD; Matthew R. Marler, PhD ; Mortality Associated With Sleep Duration and Insomnia, *Arch Gen Psychiatry.* 2002;59:131-136.

c. Est-ce que l'écart-type (corrigé) de l'échantillon est une bonne estimation de l'écart-type de la population ? Est-ce une bonne estimation de l'écart-type de la distribution d'échantillonnage ?

d. Calculez l'erreur standard de la distribution d'échantillonnage.

e. Maintenant que vous avez toutes les informations nécessaires, établissez un intervalle de confiance à 95% autour de la moyenne de votre échantillon.

f. Que pouvez-vous conclure des résultats obtenus au point précédent.

g. C'est bien beau tout ça mais qu'est-ce notre chercheur va pouvoir faire de ce résultat et de sa conclusion ?

T.P. 9 – 10 : Chapitre 8

Exercice 9 : Modélisation : Réduction de l'erreur

Voici les résultats obtenus par une classe de 10 élèves à une interrogation d'histoire (noté sur 10) qui consistait en une centaine de questions de type « vrai ou faux » sans cotation négative.

Xi	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Note	4	9	8	7	6	6	7	9	4	7	7,5	8	8	6	6

$\mu = 6{,}833 \quad \sigma = 1{,}531$

1. De quel pourcentage réduirions-nous l'erreur si nous considérons que les élèves n'ont pas répondu au hasard à toutes les questions ?

2. Veuillez compléter le tableau suivant :

Tableau 8.6. : Tableau d'ANOVA appliqué à la série 1 du Tableau 8.4

Source	*SC*	*ddl*	*CM*	*F*	*p*
Réduction modèle A	SCR	PA - PC	CMR = SCR/(PA - PC)	CMR/CME	
Série 1					
Erreur modèle A	SCE(A)	n - PA	CME = SCE(A)/(n - PA)		
Série 1					
Total	SCE(C)	n - PC			
Série 1					

3. Sur base des résultats obtenus et de la table, pouvez-vous rejeter l'hypothèse selon laquelle les résultats des élèves sont **significativement** différents de ce qu'ils auraient obtenus en répondant au hasard, avec une probabilité de 95% d'avoir raison ?

4. Veuillez vérifier cette même hypothèse, mais cette fois en construisant un intervalle de confiance à l'aide d'une distribution F (ou d'une distribution t), qui tolère un risque α de 5%.

T.P. 9 – 10 : Chapitre 8

Exercice 10 : I.C. à l'aide d'une distribution F ou d'une distribution t

Voici deux distributions statistiques, qui correspondent aux notes obtenues par deux groupes de 12 élèves à un examen statistique (noté sur 100).

	1	2	3	4	5	6	7	8	9	10	11	12	βo	S
Yi	75	72	74	73	72	75	78	73	74	76	77	77	74,67	2,015
Y'i	85	68	81	65	82	75	70	69	67	72	81	81	74,67	6,998

1. Pour chacune d'elle, veuillez construire un intervalle de confiance acceptant un risque α de 5%, à l'aide de la distribution F. Sur base de cet intervalle, pouvez-

vous déterminer si la moyenne obtenue est significativement différente de celle qui serait obtenue par le hasard ?

2. Veuillez vérifier les valeurs obtenues pour les bornes des I.C pour ces deux séries, acceptant un risque α de 5%, mais en utilisant cette fois la distribution t

3. Comment pouvez-vous expliquer la différence entre les valeurs de CME obtenues par ces deux groupes d'élèves, qui ont pourtant la même moyenne et le même nombre de sujets ?

T.P. 9 – 10 : Chapitre 8

Exercice 11 : Modélisation : Réduction de l'erreur

1. Un psychologue examine les résultats d'un protocole de dépression coté sur 25. Plus le score est élevé, plus la personne est considérée comme dépressive. Voici les résultats obtenus par les 13 premiers patients.

x	4	5	6	8	9	10	11	12	13	15	16	18	20
xi	1	2	3	4	5	6	7	8	9	10	11	12	13

μ = 10,50	σ = ,523

a. Etablissez le modèle compact et le modèle augmenté.

b. De quel pourcentage réduisons-nous l'erreur si nous considérons que les patients n'ont pas répondu au hasard au questionnaire de dépression ?

c. Vérifiez vos conclusions précédentes à l'aide de la table F. Pouvez-vous rejeter l'hypothèse selon laquelle les résultats obtenus sont dus au hasard, avec une probabilité de 95 ? 99% ?d'avoir raison ? Détaillez tous vos calculs pour arriver aux conclusions

2. Vrai ou Faux

V-F ?	Justification
α est la probabilité de l'hypothèse nulle	
α est la probabilité d'obtenir le résultat qui a été observé si le traitement est en réalité inefficace	
Une diminution du risque alpha diminue le risque bêta pour tout échantillon donné	
La possibilité d'accepter à tort l'hypothèse nulle lorsqu'elle est fausse est représentée par le risque β	
La probabilité de commettre l'erreur de seconde espèce décroît lorsque la taille de l'échantillon augmente.	
La puissance 1- β ne dépend pas de la taille de l'échantillon, ni de la taille de PRE à détecter et de l'erreur, ni du risque de première espèce.	

T.P. 9 – 10 : Chapitre 8

Exercice 12 : Modélisation : Réduction de l'erreur

Pour un examen, le professeur décide de poser des questions à choix multiple ayant toujours 4 propositions dont une seule correcte. Un étudiant qui répond au hasard devrait donc obtenir une note de 5/20. (Il n'y a pas de système de points négatifs)

Comme il espère que ses étudiants ont étudié son cours, il va supposer que la moyenne de la classe sera une meilleure prédiction de la note d'un étudiant pris au hasard que la valeur de 5.

Les résultats sont les suivants :

i	Score Xi	$(Xi-5)^2$	$(Xi-9,76)^2$
1	11		
2	9		
3	15		
4	5		
5	11		
6	5		
7	4		

8	9		
9	8		
10	9		
11	15		
12	12		
13	8		
14	11		
15	7		
16	14		
17	13		
Σ			

Moyenne = 9.76 écart-type = 3.42

1. Que vaut la PRE ? (vous pouvez remplir le tableau ci-dessus pour vous aider)

2. Calculer la SCR de deux façons différentes.

3. Remplissez le tableau suivant en indiquant les valeurs en dessous des initiales.

Tableau 8.6. : Tableau d'ANOVA appliqué à la série 1 du Tableau 8.4

Source	*SC*	*ddl*	*CM*	*F*	*p*
Réduction modèle A	SCR	PA - PC	CMR = SCR/(PA - PC)	CMR/CME	
Série 1					
Erreur modèle A	SCE(A)	n - PA	CME = SCE(A)/(n - PA)		
Série 1					
Total	SCE(C)	n - PC			
Série 1					

.

4. Quel autre nom avons-nous donné jusqu'à présent (dans les TP précédents) à la CME ? :

5. Sur base des résultats du tableau de la question 3, peut-on conclure que le modèle augmenté présente un intérêt ?

T.P. 9 – 10 : Chapitre 8

Exercice 13 : Propriétés des erreurs α et β

En psychologie, on accepte généralement un risque de 5% de commettre une erreur de première espèce (soit un α de .05), cependant, certaines situations nous obligent à considérer un intervalle de confiance supérieur à .95. Quel est le problème dans de telles situations ?

T.P. 9 – 10 : Chapitre 8

Exercice 14 : Lecture de la table de puissance

Considérons une situation dans laquelle l'α est de .05, le degré de liberté du modèle compact vaut 0 et le degré de liberté du modèle augmenté vaut 1 (voir tableau 8.10).

1. Si l'on considère 0,75 comme puissance acceptable, à partir de quelle valeur une PRE sera-t-elle détectable ?

2. Lorsque l'on considère 0,75 comme puissance acceptable, de combien de sujets au moins devra être composé notre échantillon pour pouvoir détecter des PRE de .05, .2 et .3 ?

T.P. 9 – 10 : Chapitre 8

Exercice 15 : Risque α, risque β et Puissance

A. Un neurologue s'inquiète en remarquant chez un de ces patients des signes typiques de la maladie de Parkinson (tremblements, mouvements lents, marche à petits pas, etc). Avant de faire passer un DATScan (examen coûteux et utilisé à des fins de confirmation) et des examens plus spécifiques, il administre à son patient quelques tests cliniques préalables qui mettent en évidence la présence de déficits associés à cette maladie. Il est important de détecter cette maladie pour pouvoir administrer le traitement approprié.

1. A quoi correspond l'erreur de type I que le neurologue pourrait commettre ? Donnez une autre manière d'exprimer « erreur de type I ».

2. A quoi correspond l'erreur de type II que le neurologue pourrait commettre ? Donnez une autre manière d'exprimer « erreur de type II ».

3. A votre avis, le neurologue doit-il fixer le niveau de confiance à 90%, 95% ou 99% pour la passation des tests cliniques? Expliquez pourquoi. Donnez-en les avantages et les inconvénients.

4. Si le patient est atteint de maladie de Parkinson et que le niveau de confiance est fixé à 90%, quelle est la probabilité de faire une erreur de type I ?

5. De manière générale, à quoi correspond la puissance ? Comment la note-t-on ? Comment la traduiriez-vous dans le présent exemple ?

6. Complétez les quatre cases du tableau ci-dessous par une phrase qui correspondrait (certaines réponses ont déjà été proposées aux questions 1, 2 et 5 ; il suffit de les placer dans les cases appropriées) :

	Réalité	
Décision statistique	Présence de la maladie	Absence de la maladie
Rejet de l'hypothèse (présence de la maladie)	Erreur de type I (risque α)	Décision correcte
Non rejet de l'hypothèse (présence de la maladie)	Décision correcte	Erreur de type II (risque β)

T.P. 9 – 10 : Chapitre 8

Exercice 16 : Risque α, risque β, puissance 1- β et Proportion de réduction de l'erreur.

1. Placez les 6 propositions suivantes dans les cases appropriées du tableau ci-dessous :

a. Erreur de type II
b. 1-β
c. risque α
d. Décision correcte
e. Erreur de première espèce
f. Puissance

	Réalité	
Décision statistique	Modèle C correct	Modèle C incorrect
Rejet du modèle C		
Non rejet du modèle C		

2. Un ingénieur construit un pont. Il teste l'amplitude des vibrations sur ce pont en comparaison avec la moyenne habituelle de l'amplitude de vibrations des autres structures qu'il a construites. Cette moyenne de vibrations des autres structures correspond au modèle compact. Si le test ne rejette pas le modèle C, le pont est considéré comme sécuritaire. A l'inverse, si le test rejette le modèle C (s'il y a une différence significative entre les vibrations de son pont et la moyenne des autres structures), le pont est considéré comme dangereux. Selon vous, quel type d'erreur (α ou β) l'ingénieur devrait davantage tenir en compte et tenter de minimiser ? Pourquoi ?

3. Combien faudrait-il de sujets au minimum pour détecter :

   a. Une PRE 0.20, si on considère une puissance de 0.85 comme acceptable ?
   b. Une PRE 0.075, si on considère une puissance de 0.50 comme acceptable ?
   c. Une PRE 0.10, si on considère une puissance de 0.50 comme acceptable ?

Aidez-vous de la table de puissance (Tableau 8.10).

T.P. 9 – 10 : Chapitre 8

Exercice 17 : Récapitulatif

Voici les résultats d'un test (noté sur 10) mesurant les aptitudes sociales de 10 étudiants belges en Psychologie de 2009: 5,5,6,6,7,7,7,8,8,9. La moyenne de ces étudiants vaut 6,8 et son écart-type corrigé vaut 1,32

1. Le chapitre 8 aborde une nouvelle façon de faire des statistiques. En effet, il nous permet de se familiariser avec l'inférence statistique. De quoi s'agit-il exactement ? Illustrez le concept sur base de l'exemple ci-dessus.

2. La base de l'inférence statistique réside dans la distribution d'échantillonnage théorique de la moyenne de ces aptitudes sociales. Comment pouvez-vous expliquer ce que cette distribution d'échantillonnage représente ?

3. Quelles sont les paramètres de cette distribution d'échantillonnage ?

4. Si un chercheur, plus que motivé, a testé TOUS les étudiants de Psycho en 2009 et qu'il connait la moyenne et l'écart-type, qu'en advient-il de notre inférence statistique ?

5. Donnez 3 manières équivalentes de pouvoir faire de l'inférence statistique ?

6. Quelle est la forme de la distribution d'échantillonnage si le chercheur ne possède que la moyenne de l'échantillon et la variance de la population et qu'il souhaiterait construire un intervalle de confiance ? Pourquoi ?

7. Dans la plupart de la littérature, vous entendrez parler d'intervalle de confiance de 95%… Qu'est-ce que ça signifie ?

8. Que se passe-t-il si, tout d'un coup, le chercheur perd la variance de la population… Sur quelle distribution devrions-nous nous baser pour construire un intervalle de confiance ? Quelles en sont les caractéristiques ?

9. Dans quel type d'inférence statistique nous trouvons-nous, si nous avons une autre prédiction que la moyenne ; cette prédiction étant le score théorique de 5. Néanmoins, nous estimons que la moyenne de notre échantillon est plus représentative des étudiants en Psycho.

10. Créez et annotez mathématiquement un modèle compact et un modèle augmenté sur base de l'énoncé 9.

11. Pourquoi n'auriez-vous pas fait l'inverse ? Appuyez votre raisonnement sur l'erreur associée à chacun des modèles.

12. Calculez la proportion de la réduction de l'erreur due au modèle augmenté ? Que cela signifie-t-il ?

13. Pour voir si cette réduction de proportion de l'erreur est significative par rapport au modèle compact, vous allez devoir utiliser une distribution liée directement au PRE. Comment se nomme-t-elle ? Quels sont ses avantages ?

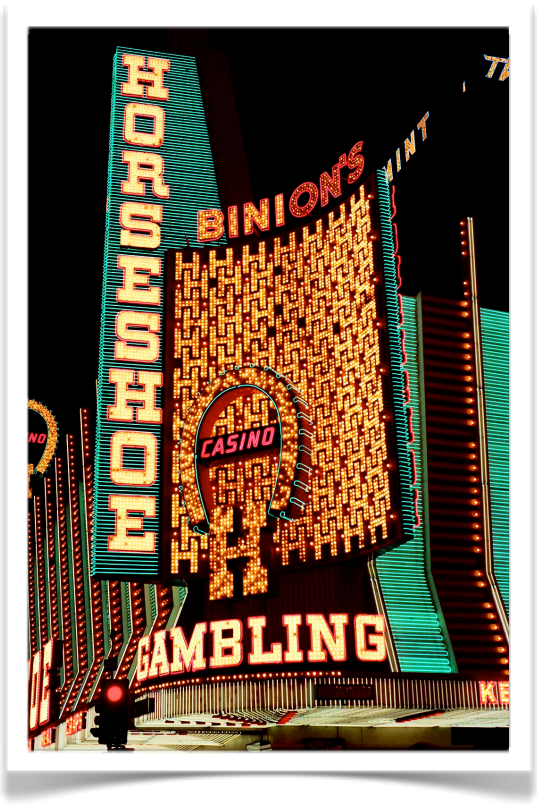
14. Calculez la Statistique de test F pour notre comparaison de modèle.

15. Cette valeur de F est-elle comprise dans la densité de probabilité de 95% ?

CHAPITRE 9 : INFERENCE STATISTIQUE SUR DES VARIABLES NOMINALES - TEST χ^2

9.1. Introduction

Jusqu'à présent, nous avons fait de l'inférence statistique en considérant que les variables étaient sur une échelle d'intervalle au minimum. Une note a été introduite sur la possibilité d'utilisation d'échelle ordinale également, et c'est, quelque part, ce que nous avons fait en exemple en utilisant des notes de cours. Cependant, ces tests ne peuvent être effectués sur des variables nominales. En effet, pour ce type de variables, la moyenne et l'écart-type ne représentent rien. Rappelons-nous que si j'attribue la valeur 1 aux cheveux blonds, 2 aux bruns et 3 aux noirs, ce n'est que conventionnel. Je pourrais tout aussi bien changer cette attribution pour obtenir 1 - Noirs ; 2 - Blonds ; et 3 - Bruns. Avoir une couleur de cheveux moyenne de 1,7 (par exemple) n'aurait aucun sens.



Imaginons une situation d'un jeu de dés : le Craps. William Lee Bergstrom, dit "*l'homme à la malette*", est originaire du Texas et est connu pour avoir été particulièrement chanceux à ce jeu. Nous sommes en 1980. Au départ de sa fortune, s'élevant à 777.000 dollars, il gagne 1,5 million de dollars aux dés au Binion's Horseshoe Hotel à Las Vegas (ci-contre). Durant les quatres années qui suivent, il parvient à gagner un autre million de dollar. Le samedi suivant la fin de la vraie histoire, imaginons que William L. B., maintenant célèbre, entre dans un bar, confiant en sa chance, mais se retrouve, dés en mains, face à des joueurs suspicieux. Ceux-ci le somment de jeter un de ses dés une soixantaine de fois afin de déterminer si les dés sont bien équilibrés. La Tableau 9. 1 donne les résultats des jets effectués. Ces données n'ont absolument rien d'historiques, elles sont tirées du livre de Freedman, Pisani & Purves (1997).

Tableau 9. 1. : Jet de dés

Valeur	Fréquence observée	Fréquence attendue
1	4	10
2	6	10
3	17	10
4	16	10
5	8	10
6	9	10
Somme	60	60

La fin de l'histoire de William Lee Bergstrom appartient à la légende, néanmoins il paraît assez clair que la loi des grands nombres l'ait rattrapé. Il semble que quelques années plus tard, il ait retenté son coup de chance, une troisième fois, mais, cette fois là, a perdu la quasi entièreté de sa fortune. Certains prétendent qu'il aurait à nouveau tenté sa chance en misant le reste de ses avoirs à la roulette russe et aurait perdu, cette fois en y laissant la vie. Selon une source plus crédible (le Los Angeles Times) le corps de William L. B. fut retrouvé sans vie dans sa chambre au Marina Hotel. Le sergent de police Frank Jergovic l'aurait trouvé en état d'overdose, il était âgé de 33 ans[39], nous sommes le 4 février 1985.

Avant de déterminer si les dés de William L. B. sont truqués ou pas, il est nécessaire d'attirer votre attention sur une distinction importante lorsque l'on parle de Khi-carré. Ce terme s'applique en fait à deux concepts, certes liés, mais néanmoins différents : la distribution Khi-carré et les tests Khi-carrés. Nous allons commencer par décrire la distribution et distinguer à la fois ce qui la caractérise et ce qui la rend utile pour les tests qui vont suivre.

[39] Comme d'autres, plus connus : Jésus ; Charles IV le Bel, dernière victime de la malédiction de Jacques de Molay (Chef des Templiers condamné au bucher, par Philippe le Bel, d'où il y maudit ce dernier et sa descendance) ; Alexandre le Grand ; Bon Scott, premier chanteur d'AC-DC ; etc.

9.2. La distribution Khi-carré

9.2.1. L'équation

La distribution Khi-carré est une distribution qui ne dépend que d'un seul paramètre : les degrés de liberté. Elle est définie par une équation compliquée que vous n'aurez pas à retenir, mais que je vous donne pour information :

Equation de la distribution Khi-Carré

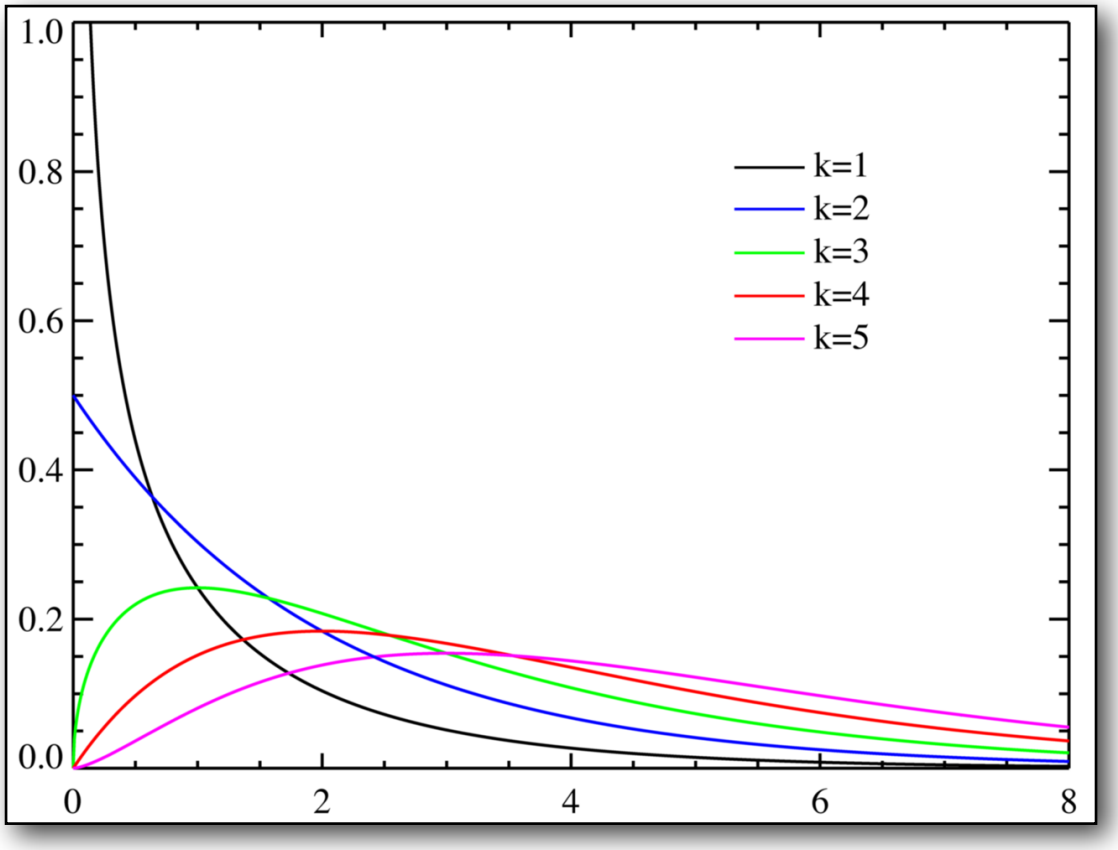
$$f(\chi^2) = \frac{1}{2^{k/2}\Gamma(k/2)} \chi^{2[(k/2)-1]} e^{-\chi^2/2}$$

Où Γ est appelée "fonction gamma". Il s'agit en fait d'une fonction proche de la fonction factorielle que vous connaissez mais appliquée à des nombres qui ne sont pas toujours entiers. En outre, *k* est le nombre de degrés de liberté et χ la valeur de la distribution sur l'axe horizontal, nécessaire pour trouver la densité de probabilité $f(\chi^2)$ associée. J'arrête ici ma description de cette équation qui n'a donc pas beaucoup d'intérêt pour nous.

9.2.2. Lien avec la distribution normale et approche intuitive

La figure 9.1 vous montre la forme caractéristique d'une telle distribution. Elle a tendance à être asymétrique, bien que cette asymétrie diminue au fur et à mesure que le nombre de degrés de liberté augmente. La plus petite valeur possible du χ^2, sur l'axe horizontal, de cette distribution est zéro. Puis les valeurs progressent vers l'infini.

Figure 9. 1. : Distribution Khi-carré en fonction du paramètre k (degrés de liberté)



Cette distribution est utilisée notamment par le test Khi-carré que nous verrons au point suivant et qui est en réalité un test de concordance entre une distribution observée et une distribution théorique. Le principe de base est de regarder, comme le suggère le Tableau 9.1 (où une colonne est dénommée valeurs observées, et l'autre valeurs théoriques), si les données que l'on observe sont compatibles avec des données théoriques. Dans le Tableau 9.1, les scores obtenus lors des lancers de dés (distribution observée) seront comparés aux valeurs que l'on obtiendrait si le dé était parfaitement bien équilibré (distribution théorique). La distribution Khi-carré est donc une distribution basée sur les écarts entre valeurs, élevés au carré pour les raisons habituelles de signe. C'est par définition ce qu'on appelle une variance (puisque la variance n'est qu'une mesure des écarts au carré entre les valeurs d'une distribution et la moyenne de cette distribution). Lorsque l'on prend la somme de tous les écarts au carré entre deux valeurs, le minimum que l'on puisse obtenir est évidemment zéro (lorsqu'il n'y a aucun écart, par exemple, lorsque toutes les valeurs sont égales à la moyenne). En effet, puisque les écarts sont élevés au carré, il est impossible d'obtenir de valeurs négatives. En revanche, il n'y a pas de valeur maximale. Nous devons donc considérer que cette distribution peut, théoriquement, aller jusqu'à l'infini.

Il existe un lien entre cette distribution et la distribution normale. Rappelons-nous que la distribution normale a, comme abscisse, les valeurs z qui sont des scores standardisés. Par définition, $z = (X - \mu)/\sigma$. Supposons maintenant que nous élevions ces valeurs au carré. Nous obtiendrions $z^2 = (X - \mu)^2/\sigma^2$. Nous aurions donc la somme des carrés des écarts à la moyenne divisée par la variance de la population. Par définition une distribution Khi-carré à un degré de liberté est égale à cette valeur : $\chi^2_{(1)} = z^2$. En augmentant le nombre de degrés de liberté nous ne faisons rien d'autre que prendre la somme des scores *z* au carré : $\chi^2_{(N)} = \Sigma z^2$. Remarquez que, présenté de cette manière, il apparaît clairement que la distribution khi-carré ne peut que prendre des valeurs nulles ou positives.

La conséquence de ces liens est que la distribution Khi-carré est soumise aux mêmes conditions que celles imposées aux tests que nous avons réalisés précédemment : les observations doivent être indépendantes et provenir d'une population distribuée normalement.

9.2.3. La table des valeurs critiques

De la même manière que pour la table de la distribution normale, de Student ou de Fisher-Snedecor, il existe, pour la distribution Khi-carré, une table des valeurs critiques. Elle est organisée en fonction de deux critères : la paramètre *k* et la probabilité α de commettre l'erreur de type I. Le Tableau 9.2 rapporte un extrait de cette table.

Tableau 9.2. : Table de valeurs critiques de la distribution Khi-carré

Critical Points of the Chi Square Distribution

D. F.	Cumulative probability												
	0.005	0.010	0.025	0.05	0.10	0.25	0.50	0.75	0.90	0.95	0.975	0.99	0.995
1	0.39E-4	0.00016	0.00098	0.0039	0.0158	0.102	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5
7	0.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.9	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.3	12.3	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9

Imaginons une situation dans laquelle nous avons 20 degrés de liberté. Notre exigence habituelle est d'accepter un risque α de 5% (0,05). La table indique une valeur de 31,4 pour ce risque. Pour trouver cette valeur vous devez utiliser la colonne 0,95 puisque la table n'envisage pas le risque mais bien la probabilité complémentaire, comme le montre la petite représentation graphique en-tête de table. Puis, vous descendez jusqu'à rencontrer la ligne correspondant au nombre de degrés de liberté. Cette valeur est donc la valeur Khi-carré qui comprend 95% des scores de la distribution pour ce nombre de degrés de liberté. Le point

suivant illustre cette utilisation au travers d'une analyse rarement envisagée, mais pourtant importante : la distribution d'échantillonnage de la variance.

9.2.4. Distribution d'échantillonnage de la variance

Outre l'utilisation que nous allons en faire pour effectuer de l'inférence statistique sur des variables nominales, la distribution Khi-carré a une application qui devrait vous sembler évidente compte tenu de ce que nous venons de voir, mais qu'il est nécessaire d'énoncer : elle correspond à la forme de la distribution d'échantillonnage de la variance. En effet, lorsque nous estimons la variance de la population à l'aide de notre échantillon, nous pouvons tenir exactement le même raisonnement que celui que nous avons tenu concernant la moyenne. Supposons que de nombreux expérimentateurs estiment eux-aussi la variance de la population à l'aide de leur échantillon. Nous aurions tous une estimation sans doute proche, mais néanmoins différente de la variance de la population. Cependant, cette estimation ne peut pas être distribuée normalement comme le serait celle de la moyenne puisque les variances ne peuvent pas être négatives. La distribution d'échantillonnage de la variance suit en fait une distribution en lien étroit avec une distribution Khi-carré de degrés de liberté égaux à n-1. Vous admettrez sans démonstration que ce lien consiste en la relation linéaire suivante (vous pouvez le démontrer vous-même en remplaçant S^2 par sa formule et en n'oubliant pas qu'il s'agit de la variance corrigée) :

Equation de la distribution d'échantillonnage de la variance

$$\chi^2_{(n-1)} = \frac{(n-1)S^2}{\sigma^2}$$

Où S est l'écart-type de l'échantillon et σ celui de la population. La variance de la population (l'écart-type au carré) peut être isolée à partir de cette équation en inter-changeant simplement cette variance avec la valeur Khi-carré qui se retrouve donc au dénominateur. Cependant, lorsque l'on cherche à déterminer l'intervalle de confiance, nous faisons face à une petite subtilité mathématique. Voici le développement :

$$IC : \chi^2_{(n-1,\alpha/2)} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{(n-1,1-\alpha/2)}$$

$$IC : \frac{\chi^2_{(n-1,\alpha/2)}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{(n-1,1-\alpha/2)}}{(n-1)S^2}$$

$$IC : \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha/2)}} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha/2)}}$$

$$IC : \frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha/2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha/2)}}$$

Vous remarquez entre la deuxième et la troisième étape le changement de sens du signe, obligatoire pour réaliser l'inversion dans une inéquation de ce type. La transition entre la troisième et la quatrième étapes n'est rien d'autre qu'une réécriture de l'équation en tenant compte du sens logique de l'intervalle (la plus petite valeur à gauche, la plus grande à droite). De sorte que l'intervalle de confiance d'une distribution de variance devient :

Intervalle de confiance bilatéral pour une distribution de variances

$$IC : \frac{(n-1)S^2}{\chi^2_{(n-1,1-\alpha/2)}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha/2)}} \rightarrow bilatéral$$

Et, selon le même raisonnement, il est nécessaire de tenir compte de la probabilité α et non 1-α pour un test unilatéral :

Intervalle de confiance unilatéral pour une distribution de variances

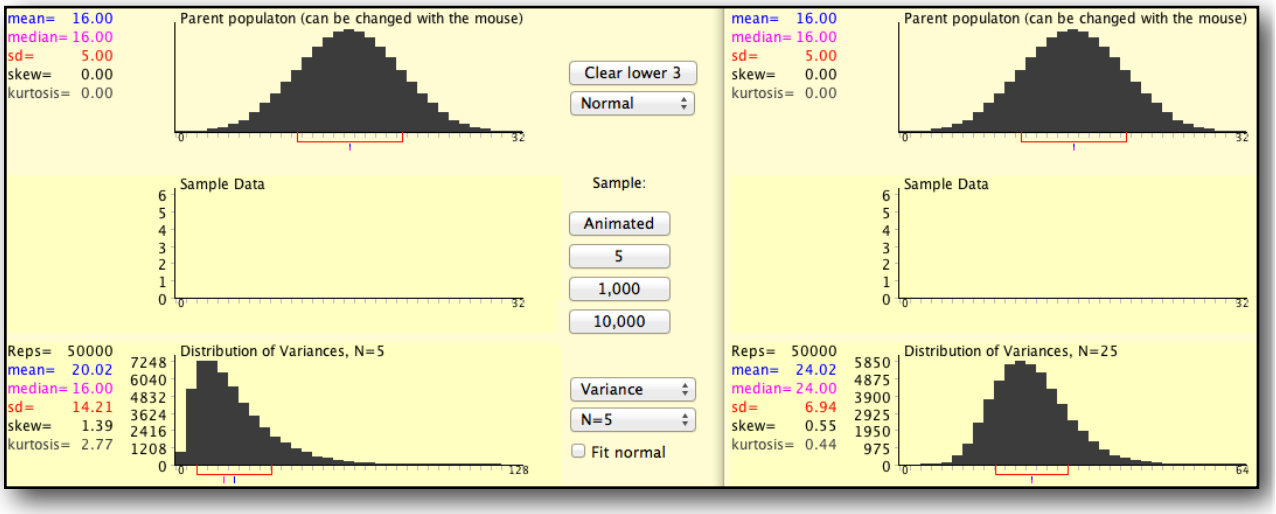
$$IC : \sigma^2 < \frac{(n-1)S^2}{\chi^2_{(n-1,\alpha)}} \rightarrow unilatéral$$

Deux choix sont possibles. Soit nous réalisons le test en bilatéral, en considérant qu'il n'y a pas de raison de conserver les valeurs proches de zéro de cette distribution. Soit nous réalisons le test en unilatéral en considérant que la variance peut être nulle, mais que ce qui

nous intéresse avant tout est la valeur maximale que la variance peut prendre. Nous illustrerons ce propos dans l'exemple que nous envisagerons ci-dessous.

La Figure 9.2 montre une distribution d'échantillonnage de la variance pour des échantillons de taille n = 5 et pour des échantillons de taille n = 25. On voit que, lorsque le nombre de degrés de liberté est plus important, l'asymétrie de la distribution tend à s'estomper. Par ailleurs, l'erreur standard de cette distribution tend également à diminuer en fonction de l'augmentation de n. En revanche, la variance estimée sera de plus en plus grande au fur et à mesure que n augmente. Cette situation peut se comprendre intuitivement et a déjà été discuté auparavant (chapitre 8 et théorie des grands nombres): imaginez que vous mesuriez la variance de la taille des étudiants d'un auditoire en prélevant aléatoirement deux personnes. Vous auriez de grandes chances de prendre deux personnes de taille moyenne puisque nous avons vu que la plupart des individus ont une taille proche de la valeur moyenne. Donc, le plus souvent votre variance est assez faible. S'il y a un individu de deux mètres, vous le prendrez dans de très rares cas et dans ce cas votre variance sera très élevée, mais la plupart du temps elle sera très faible. En revanche, si vous prenez un échantillon plus grand, par exemple 50 étudiants, vous aurez nettement plus de risques d'inclure dans cet échantillon de grands ou de petits individus en plus de ceux qui ont une taille proche de la moyenne. Dès lors, la variance que vous obtiendrez sera nécessairement plus grande que lorsque votre échantillon est de petite taille. Cependant, les fluctuations de votre estimation de la variance (l'erreur standard) sera plus faible en prenant de grands échantillons qu'en en prenant des petits puisque l'impact des grands et des petits individus est moins grand que lorsque l'échantillon contient peu de sujets.

Figure 9. 2. : Distribution d'échantillonnage, basée sur 50000 itérations, de la Variance lorsque *n* = 5 (gauche) ou *n* = 25 (droite). On remarque une atténuation de l'asymétrie lorsque n augmente, une estimation moyenne de la variance plus élevée (24 au lieu de 20) et une erreur standard plus faible (6,94 au lieu de 14,21).



Reprenons maintenant notre Tableau 8.4. Nous voyons que, pour la série 1, l'écart-type est de 24,31. La variance est donc le carré de cette valeur, soit 590,98. Pour la série 2, l'écart-type est de 1,04 et la variance de 1,08. Attachons-nous, dans un premier temps, à la série 1. Le nombre de sujets de cette série était égal à 14, ce qui conduit à 13 degrés de liberté (= 14-1).

En appliquant le raisonnement décrit ci-dessus, nous pouvons calculer l'intervalle de confiance en unilatéral de cette variance : la table nous informe que la valeur du Khi-carré pour 13 degrés de liberté et une probabilité de 1-α vaut 22,4, cependant, nous devons utiliser la probabilité complémentaire pour tenir compte du changement de sens du signe et choisir 1-α = 0,05. Dès lors, nous en déduisons que la variance de la population ne sera pas supérieure à 13*590,98/5,89 = 1304,37.

Imaginons que nous ayons considéré qu'il n'y a aucune raison de travailler en unilatéral et que nous refusons de considérer que la variance de la population pourrait valoir zéro (qui est trop éloignée des 590,98 que nous observons au niveau de notre échantillon). Dans ce cas, nous devons répartir le risque sur les deux côtés de la courbe et considérer les valeurs de la table pour un risque de 0,025 à droite, et 0,975 à gauche. Ces valeurs sont de 5,01 et de 24,7. Reporté à notre exemple, nous aurions une variance comprise entre (13*590,98/24,7) = 311,04 et (13*590,98/5,01) = 1533,48.

De la même manière nous pouvons calculer l'intervalle de confiance dans lequel devrait se trouver la variance de la série 2. Vous pourriez faire l'exercice vous-même pour vérifier votre compréhension, mais je vous donne néanmoins la réponse (n'hésitez pas à refaire les calculs pour d'autres séries vues dans ce syllabus, ça pourrait vous servir…).

IC bilatéral : $13*1{,}08/24{,}7 < \sigma^2 < 13*1{,}08/5{,}01 \Leftrightarrow \mathbf{0{,}57 < \sigma^2 < 2{,}80}$

IC unilatéral : $\sigma^2 < 13*1{,}08/5{,}89 \Leftrightarrow \mathbf{\sigma^2 < 2{,}38}$

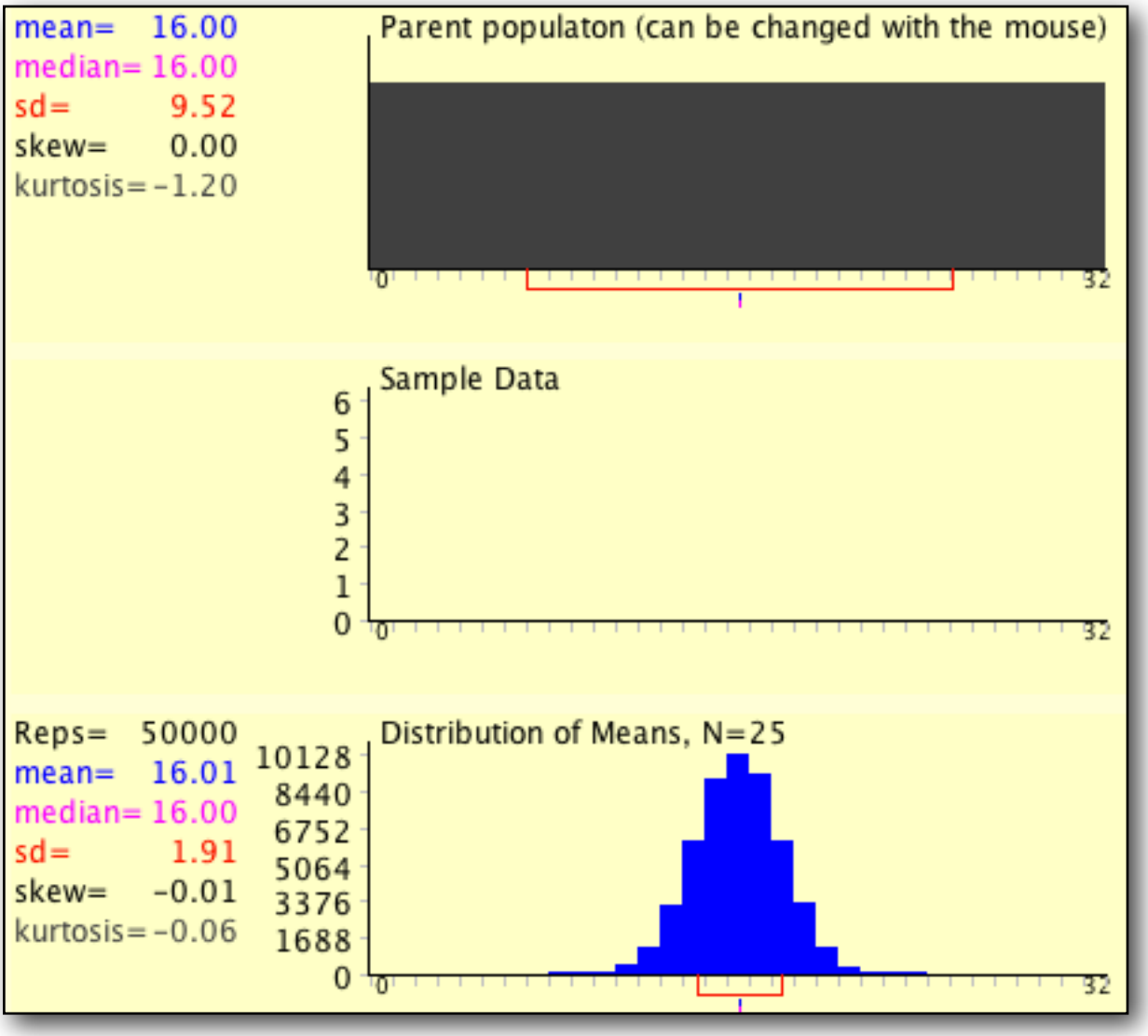
9.3. Test Khi-carré d'ajustement

Abordons à présent le test Khi-carré. Il se distingue de la distribution dans la mesure où il s'agit d'un test de comparaison entre une distribution observée et une distribution théorique. En fait, nous l'utiliserons pour traiter les problèmes liés à l'utilisation de variables dépendantes catégorielles, mais d'autres applications sont possibles. Par exemple, nous pourrions, à l'aide d'un tel test, vérifier si la distribution des valeurs d'une série, ou des erreurs associées à un modèle, sont compatibles avec une distribution normale, bien que nous verrons que ce n'est pas la méthode la plus puissante.

Revenons à l'embarrassante situation de feu William L. B., contraint de lancer 60 fois un de ses dés. Remarquez que nous aurions pu vérifier que les dés étaient mal équilibrés en testant simplement si la moyenne est différente de 3,5. En effet, la moyenne théorique d'un grand nombre de jets de dé est de (6+1)/2=3,5. Cependant, ce n'est pas une si bonne idée. En effet, si vous voulez savoir si le dé fait plus régulièrement un six que les autres valeurs, cela pourrait marcher (la moyenne serait plus élevée). Mais, qui joue au Craps sait que vous pourriez avoir intérêt à ce que votre dé fasse plus souvent un 3 ou un 4 que les autres valeurs. En effet, si vous réalisez un total de 7 au premier coup, vous êtes gagnants. Or si le dé est mal équilibré parce qu'il fait plus régulièrement 3 et 4 que toutes les autres valeurs, la moyenne (3,5) sera compatible avec un bon équilibre. Enfin, la distribution des scores est uniforme, il n'y a pas plus de chance de réaliser une valeur que les autres, la distribution de la population n'est donc pas une normale. Or, si William L. B. faisait plus de 3 et de 4 que d'autres valeurs, la forme de la distribution ne serait plus uniforme. C'est donc bien un test de compatibilité entre une forme de distribution observée et une forme de distribution théorique qui constitue le bon test.

Remarquez cependant que la distribution d'échantillonnage de la moyenne suit bien une distribution normale, même si la distribution de la population est, elle, uniforme. En effet, même si chaque score (de 1 à 6) a la même probabilité d'occurrence, la moyenne n'en serait pas moins 3,5 et, si on prend plusieurs échantillons, c'est assez rare que cette moyenne fluctue (et les fluctuations seront de plus en plus faibles en fonction de l'augmentation du *n*). La Figure 9.3 illustre cette situation particulière (pour une distribution uniforme continue allant de 0 à 32, de moyenne égale à 16). Elle montre que la forme de la distribution de la population ne conditionne pas la distribution d'échantillonnage qui, elle, reste une distribution normale.

Figure 9.3. : Distribution d'échantillonnage de la moyenne dont la distribution de la population est uniforme. L'effectif des 50000 échantillons est de $n = 25$ par échantillon. La distribution d'échantillonnage suit une loi normale.



Le Tableau 9.1 vous montre que la distribution observée contient des différences par rapport à la distribution théorique. Théoriquement, sur 60 jets, chaque chiffre devrait être représenté 10 fois (vu qu'ils ont tous la même probabilité d'occurrence si le dé est bien équilibré). Ces 10 fois attendues représentent en fait la **fréquence attendue**. Il est cependant normal d'observer des écarts par rapport à cette situation, de la même manière que si je jetais 60 fois une pièce de monnaie, je m'attends à avoir 30 faces et 30 piles tout en sachant très bien que ce ne sera probablement pas exactement le cas. **La question devient donc de savoir si les écarts entre la distribution théorique et la distribution observée sont suffisamment importants pour être attribués à autre chose qu'au hasard d'échantillonnage**.

Nous n'allons pas démontrer la formule du test Khi-carré donnée ci-dessous. Notez cependant qu'elle est en lien direct avec le carré des écarts entre la distribution observée et la distribution théorique. C'est ce terme qui nous intéresse évidemment le plus. Le reste n'est qu'un arrangement mathématique pour rendre juste la démonstration[40] basée sur la relation de base $\chi^2 = z^2$.

Test Khi-Carré d'ajustement

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$ Où O = fréquence observée et E = Fréquence attendue (Expected).

Appliqué à notre exemple (Tableau 9.1), cela donne le calcul suivant :

$$\chi^2 = \frac{(4-10)^2}{10} + \frac{(6-10)^2}{10} + \frac{(17-10)^2}{10} + \frac{(16-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} = 14{,}2$$

La valeur observée du χ^2 est donc de 14,2. Il y a 5 degrés de liberté. En effet, si nous avons 60 jets dont chacun peut produire 6 chiffres, nous obtenons la fréquence pour chaque chiffre

[40] Une démonstration convaincante et accessible par le truchement de la loi binomiale étendue au Khi-carré (si, si, il y a un lien : la distribution binomiale est la plus simple et on peut également vérifier qu'il n'y a pas d'écart significatif entre une série d'expériences aléatoires binaires, par exemple, une série de jets de pile ou face, et la distribution binomiale théorique) est trouvable dans le Howell (1999) p. 160.

(tel que présenté au tableau 9.1). Dès lors, si on connait la fréquence pour 5 de ces 6 chiffres, nous pouvons déduire la fréquence pour le sixième et nous perdons donc un degré de liberté. La table Khi-carré nous informe que la valeur limite du $\chi^2_{(5)}$ qui correspond à une densité de probabilité de 95% vaut 11,1. Il y a donc moins de 5% de risque de nous tromper en considérant que la distribution que nous observons est différente de la distribution théorique considérée.

Si nous comptons simplement décliner poliment l'invitation à jouer avec William L. B., nous pouvons nous arrêter là. Si, en revanche, nous voulons l'enduire de goudron et de plumes et lui faire subir les pires humiliations dans les rues de Las Vegas, nous aurons peut-être la courtoisie d'être plus exigeant sur le risque de prendre la mauvaise décision. Nous pourrions par exemple n'accepter que cinq chances sur mille de nous tromper. Dans ce cas, nous regardons la dernière colonne de la table (0,995) pour 5 degrés de liberté. La valeur correspondante est de 16,7. Nous ne pouvons donc pas prendre le risque de lui infliger un traitement aussi extrême (si cela vous fâche, libre à vous d'exiger une centaine de jets de dé supplémentaires dans l'espoir de voir se creuser les écarts).

A ce moment du raisonnement, un des joueurs, un Lord londonien du nom de Brett Sinclair[41], a investi dans le goudron et les plumes et a envie de les utiliser. Mais, en parfait gentleman, il trouve trop cavalier d'exiger 100 jets de plus de la part d'un potentiel innocent. Il tient alors le raisonnement suivant : le présent test montre que la distribution n'est pas compatible avec une distribution au hasard. Cependant, elle n'indique pas où se trouve la différence. Sachant que William L. B. ne joue qu'au Craps, il a tout intérêt à truquer ses dés pour que le 3 et le 4 apparaissent plus souvent. Si Brett Sinclair pouvait montrer à coup raisonnablement sûr que ce sont spécifiquement ces deux valeurs qui apparaissent le plus souvent, il aurait alors un argument supplémentaire pour s'amuser avec le goudron.

Plusieurs méthodes permettent de distinguer les écarts spécifiques au sein de chaque cellule. Mais Brett Sinclair a opté pour celle développée par Haberman (1973) dont je tiens l'explication de l'ouvrage de Sheskin (2007) qui le cite. Cette méthode consiste à analyser les résidus pour chaque cellule et diviser ces résidus par son écart-type. Cela revient à standardiser le résidu. En principe, ces résidus suivent une distribution normale (l'erreur est distribuée normalement, c'est une des conditions d'application du Khi-carré dont nous

[41] Dont les exploits sont racontés dans la série des années 70 "*Amicalement Vôtre*".

avons discuté en début de chapitre). Dès lors, tout résidu standardisé ayant une valeur supérieure ou égale à 1,96 sera considérée comme anormale (avec une probabilité de 5% de se tromper, pour rappel 1,96 est la valeur critique que l'on retrouve dans la table 7.7 de la loi normale). Le Tableau 9.3 reprend l'entièreté des données du Tableau 9.1 et les ajustements correspondant à l'erreur standardisée et au calcul du Khi-carré.

Tableau 9. 1. : Jet de dés. On retrouve une anomalie au niveau de la valeur "3" qui sort de manière significativement plus fréquemment que les autres valeurs. La valeur "4" n'est également pas loin d'être anormale, mais le risque de se tromper en le décrétant est supérieur à 5%.

Value	Fréquence observée	Fréquence attendue	Erreur Standardisée	Erreur Standardisée au carré = χ^2
	O	E	$z_{res} = \frac{(O-E)}{\sqrt{E}}$	$z_{res}^2 = \frac{(O-E)^2}{E}$
1	4	10	-1,9	3,6
2	6	10	-1,26	1,6
3	17	10	**2,21**	4,9
4	16	10	1,9	3,6
5	8	10	-0,63	0,4
6	9	10	-0,32	0,1
Σ	60	60	0	14,2

Devant ces résultats, Brett Sinclair considère maintenant que non seulement la distribution est significativement différente de celle correspondant à un dé bien équilibré mais qu'en plus le trois sort significativement plus souvent que les autres. En outre, il conserve de sérieux doutes concernant la valeur quatre qui pourrait très bien être concernée elle aussi. En utilisant la table normale 7.7, vous devriez être en mesure de déterminer que la valeur 1,9 contient 94,26% des scores de la distribution en bilatéral. De même la valeur "*1*" sort particulièrement peu souvent. Brett va même un pas plus loin. Il se dit qu'il n'y avait aucune chance que la valeur "*3*" ou "*4*" sorte moins souvent que les autres (ce qui n'aurait pas permis à William L. B. de tricher), dès lors, il s'autorise à réaliser son test en unilatéral, c'est-à-dire à répartir le risque de 5% uniquement sur un seul côté de la courbe. La valeur correspondant à cette stratégie est de 1,65 (vérifiez sur la table 7.7). Fort de cet argument, Brett considère

donc la situation sous un autre oeil : le dé est mal équilibré, le 3 et le 4 sortent plus souvent qu'ils ne le devraient ce qui correspond point pour point à un dé de tricheur. Indépendamment des vociférations de William L. B., le goudron et les plumes sont appliquées par le Lord anglais heureux de son investissement.

9.4. Exercices de fin de chapitre

T.P. 11 - CHAPITRE 9 : Exercice 1
Test Khi-Carré d'ajustement : Bases théoriques

1. Donnez une définition (la plus complète possible) du test Khi-carré d'ajustement.

2. De quel type doit être la variable dépendante pour l'application du test Khi-carré d'ajustement ? Quelles sont les caractéristiques de ce type de variable ? Donnez un exemple.

3. A partir de quand rejette-t-on l'hypothèse d'égalité des distributions théorique et observée ? « Jusqu'à quand » impute-t-on les écarts entre ces deux distributions à des fluctuations d'échantillonage (= ne rejette-t-on pas l'hypothèse d'égalité)?

4. Quel(s) critère(s) importe(nt) lorsqu'on cherche une valeur dans la table de distribution Khi-carré (Tableau 9.2. du cours théorique) ?

5. Pourquoi est-il impossible d'obtenir une valeur négative dans une distribution chi-carré ?

6. La valeur minimale que l'on peut obtenir, en calculant la variance, est dès lors 0. Dans quel cas cela arrive-t-il ?

7. Quelle est la valeur maximale d'une distribution Khi-carré ?

8. Généralement en psychologie, on tolère un α de .05, mais il est parfois utile de réduire la valeur tolérée de l'alpha ; pouvez-vous dire dans quel cas ?

T.P. 11 : Exercice 2

Test Chi-Carré d'ajustement : Mise en situation

1. Recherchez les valeurs théoriques associées aux études suivantes dans votre table de la distribution Khi-carré :

   a. Etude qui répartit les sujets dans des catégories selon leur couleur de cheveux (= variable dépendante) : bruns, blonds, châtains, noirs, roux. Pour un risque α de 5 % et de 0,5%.

   b. Etude sud américaine qui répartit les sujets selon leur nationalité (= variable dépendante) : Argentine, Bolivie, Brésil, Chili, Colombie, Costa Rica, Cuba, Equateur, Guatemala, Honduras, Mexique, Nicaragua, Panama, Paragaguay, Pérou, République dominicaine, Salvador, Uruguay, Venezuela. Pour un risque de 0,01 et de 0,1.

2. Des psychologues[42] s'intéressant au couple se sont demandés si le fait de mettre par écrit ses sentiments dans un « mini-journal intime » avait un impact sur la durée de la relation. Leur hypothèse était que le simple fait de mettre ses sentiments par écrit ampliefierait ces derniers. Ils ont recruté 50 jeunes couples et ont demandé à un des deux partenaires de passer 20 minutes par jour à écrire ses pensées et sentiments profonds au sujet de la relation amoureuse, et ce, pendant 10 jours. Trois mois plus tard, 39 des 50 partenaires ayant mis leurs sentiments par écrit étaient toujours en couple. Pour tester leur hypothèse, ils vont comparer leurs résultats à ceux rapportés par plusieurs études antérieures s'intéressant au couple de façon « globale » (sans écriture des sentiments) selon lesquelles, après 3 mois, 50% des couples rompraient.

   a. Quelle est la variable indépendante et quelle est la variable dépendante ? Sur quelle échelle se mesure la variable dépendante ? Quelles sont les différentes valeurs possibles de cette variable dépendante ?

[42] D'après Richard B. Slatcher and James W. Pennebaker (2006). How Do I Love Thee? Let Me Count the Words : The Social Effects of Expressive Writing. *Psychological Science, 17 (8)*, 660-663.

b. A quoi devrez-vous faire attention lorsque vous comparerez les fréquences observées aux fréquences attendues de cette étude ?

c. Etablissez le tableau reprenant les valeurs possibles de la variable, les fréquences attendues, les fréquences observées. Ensuite, calculez la différence *O – E* pour chaque catégorie et sommez ces différences.

	Fréquences observées O	Fréquences attendues E	$(O-E)$	$Z^2_{rés} = \frac{(O-E)^2}{E}$	$Z_{rés} = \frac{(O-E)}{\sqrt{E}}$
Maintien du couple					
Rupture du couple					
∑					

d. Que vaut $\sum(O-E)$? Pourquoi ? Qu'est-ce que cela implique pour la comparaison des deux distributions ? A quelle situation pourriez-vous comparer ce phénomène ?

e. Selon vous, est-ce que ces résultats confirment l'hypothèse des auteurs (avec un risque d'erreur ne dépassant pas 5%) ? Réalisez le test adéquat pour répondre à cette question.

f. Quel est le risque d'erreur le plus faible qui conclurait à une différence significative entre les deux distributions ? Commentez.

g. Dans le tableau ci-dessus, à quoi correspond $Z_{rés} = \frac{(O-E)}{\sqrt{E}}$? Pourquoi cet indice se note-t-il à l'aide de la lettre *Z* ? Que permet-il de faire ?

h. L'erreur standardisée associée à la catégorie « maintien du couple » vaut 2,8. Peut-on considérer cette valeur comme étant significativement différente de

la distribution théorique avec un risque d'erreur ne dépassant pas 5% d'une part, et 1% d'autre part ? Si oui, dans quel sens va cette différence ?

i. L'erreur standardisée associée à la catégorie « maintien du couple » vaut 2,8. Quel pourcentage des observations de la distribution contient cette valeur en unilatéral d'une part, et en bilatéral d'autre part ? A votre avis, peut-on réaliser un test unilatéral ?

3. A un âge bien précis, on a pu déterminer que 50% des bébés marchent, 40 % ne marchent pas et 10 % présentent une ébauche de marche. Nous aimerions savoir si des bébés nés prématurément développent la marche de la même manière que les bébés nés à terme. Afin de répondre à cette question, on observe 80 bébés nés prématurément. Au même âge précis, 35 de ces 80 bébés marchent, 40 ne marchent pas et 5 présentent une ébauche de marche. Peut-on affirmer que les enfants nés prématurément développent différemment l'aptitude à la marche ?

	Observed N	Expected N
Marche	35	
Ne marche pas	40	
Ebauche de marche	5	
Total	80	

a. Quel test allons-nous effectuer pour répondre à cette question ? Veuillez expliquer pourquoi.

b. Pour la cellule « marche », décrivez en français la signification des 2 chiffres du tableau.

c. Calculez manuellement la statistique de test χ^2, ainsi que ses degrés de liberté

d. Tirez vos conclusions à l'aide de la table du χ^2.

4. Même énoncé. Nous avons un échantillon de 80 enfants. On sait que la variance de cet échantillon vaut 69.23. Déterminez l'IC de la variance bilatéral et unilatéral à 95%.

5. Une chercheuse s'intéresse à la perception des adolescents belges de la loi concernant l'adoption par les homosexuels et part de l'hypothèse qu'en absence de toute manipulation expérimentale, les participants favorables, défavorables ou mitigés se retrouveront en nombre équivalent. Afin de s'en assurer, elle recrute 90 participants et leur présente à tous la question suivante : « Que pensez-vous de la loi autorisant l'adoption par les homosexuels ? ».

   a. Quel test pourra-t-elle utiliser pour s'en assurer ?

   b. 30 répondent qu'autoriser une telle loi est vraiment mal, 22 affirment trouver cela moyennement mal et enfin, 38 déclarent trouver cela parfaitement juste. Sachant qu'en psychologie, on tolère généralement un risque d'erreur de 5%, ces résultats vont-ils permettre à la chercheuse de confirmer son hypothèse ?

6. On décide de lancer 900 fois un dé. Les fréquences suivantes sont obtenues:

Face	Fréquence
1	190
2	120
3	170
4	180
5	115
6	125

Peut-on considérer que ce dé est bien équilibré ? Dans le cas contraire, veuillez déterminer à quels endroits nous pouvons détecter une anomalie avec respectivement

- 5% de chances de se tromper
- 1% de chances de se tromper

7. Afin de s'assurer qu'un dé (traditionnel à 6 faces) est bien équilibré, on le lance 120 fois en l'air et prend note pour chaque lancer du résultat obtenu. Veuillez répondre aux questions suivantes :

	V	F
a) Nous le considérerons équilibré dans l'unique cas où nous obtenons exactement 20 fois chacune des valeurs du dé ; un léger écart (de maximum ± 3) par rapport à cette valeur attendue sera le signe d'un très léger déséquilibre.		
b) Dans cet énoncé, il y aura 5 degrés de liberté		
c) augmenter le nombre de lancer augmentera le nombre de degrés de liberté		

CHAPITRE 10 : VERIFICATION DES CONDITIONS D'APPLICATION ET ALTERNATIVES - LE TEST DE WILCOXON

10.1. Introduction

Nous avons parcouru l'entièreté de la matière que nous devions voir avant d'entrer dans des considérations plus complexes. Vous avez maintenant les connaissances nécessaires pour développer vos modèles, ce que nous ferons l'année prochaine. Vous connaissez les rudiments des probabilités qui vous permettent de comprendre les notions de densité de probabilité, d'indépendance entre variables, ainsi que les notions plus épistémologiques de pensée inductive et probabiliste. Vous êtes capables de transnumériser une série statistique de manière à la représenter graphiquement et de caractériser la distribution à l'aide de quelques paramètres, essentiellement la moyenne et la variance. Vous êtes également capables, de réaliser les premières étapes, les plus fondamentales, de l'inférence statistique, à savoir décrire une distribution d'échantillonnage, trouver un intervalle de confiance autour d'une moyenne, ou d'une variance. Vous comprenez également la différence et les similitudes entre l'approche par intervalle de confiance et l'approche par comparaison de modèles. Vous connaissez bien un certain nombre de distributions de probabilités : la distribution binomiale, normale, de Student, de Fisher-Snedecor et Khi-carré. Vous comprenez l'utilité de chacune et les conditions dans lesquelles on utilisera l'une plutôt que l'autre. Enfin, lorsque les variables ne sont pas mesurées sur des échelles d'intervalle ou plus, mais bien sur des échelles nominales, vous êtes capables d'adapter votre comportement et de comparer deux distributions à l'aide d'un test Khi-carré d'ajustement. Bref, vous avez bien travaillé (sauf si vous n'en êtes pas là, auquel cas, vous n'avez pas terminé votre étude)!

Il reste néanmoins quelques considérations à envisager. Les questions que nous allons aborder dans ce chapitre sont : comment vérifier si nous sommes dans les conditions pour utiliser la moyenne et la variance comme indicateur? ; si nous ne pouvons pas, quelle est l'alternative pour tester notre modèle?

10.2. Vérifier la normalité de la distribution de l'erreur

Au chapitre 8, nous avons envisagé les différentes conditions nécessaires pour faire de l'inférence sur la moyenne. Une condition fondamentale était que l'erreur soit distribuée

normalement. Creusons à présent ce problème. Selon l'approche par modèles, la réalité est simplifiée par un modèle auquel est associée une erreur. Or dans les exemples que nous avons considérés, le modèle était soit une valeur théorique constante (comme c'était le cas pour le modèle compact utilisé) soit il s'agissait de la moyenne (comme pour le modèle augmenté). Dans les deux cas, le modèle est une constante et seule l'erreur varie. Dès lors, dire que l'erreur suit une distribution normale équivaut à dire que les scores suivent une distribution normale. En effet, dans la mesure où les deux sont liés par l'ajout d'une constante, cela ne change en rien la forme de la distribution. Vous pouvez en faire vous-même l'expérience mathématique : prenez les valeurs du Tableau 6.1, ajoutez-leur la valeur 10 et représentez la distribution correspondant à ces nouvelles valeurs (cela vous fera un bon exercice de récapitulation du chapitre 6). Vous constaterez que la forme de la distribution ne change en rien.

En conséquence, vérifier la normalité de la distribution de l'erreur correspond à comparer la forme de la distribution d'une série de données à une distribution normale. Une telle comparaison pourrait éventuellement se faire à l'aide d'une distribution Khi-carré. Cependant, pour plusieurs raisons, on ne le fait jamais. Tout d'abord, la distribution Khi-carré est conçue pour des variables discrètes, bien que ce soit possible de l'utiliser avec une variable continue. De plus, elle est nettement moins puissante que d'autres tests, donc on pourrait rater la bonne conclusion. En conséquence, on préfèrera d'autres types de tests. Il en existe de nombreux, mais je vais en aborder deux qui sont ceux dont l'information est la plus facile à recueillir avec le logiciel SPSS que vous utiliserez l'année prochaine en travaux pratiques. Le plus fréquemment utilisé est le test de Kolmogorov-Smirnov, mais un autre test, le Shapiro-Wilk, semble encore plus adapté (selon Thode, 2002 cité par Sheskin, 2007). Nous n'allons pas aborder la mathématique de ces tests mais retenez l'information.

Enfin, en première approche, nous pouvons, comme nous l'avons déjà vu, considérer les moments (coefficients d'aplatissement et d'asymétrie) de la distribution. Rappelez-vous que ces coefficients sont adaptés à la forme d'une distribution normale. Il existe des tests permettant de définir la probabilité que ces indicateurs soient significativement différents de zéro. Cependant, nous n'allons pas les utiliser. Je vous conseille plutôt de calculer ces deux indicateurs (ou plutôt de demander aux logiciels de le faire) pour considérer l'information intuitivement, en regardant en même temps la représentation graphique de votre distribution. Puis, si vous devez faire de l'inférence statistique, faites rapidement le test de Shapiro-Wilk (ou de Kolmogorov-Smirnov, mais ils sont généralement effectués

ensemble par SPSS) pour vérifier que vous êtes dans les conditions pour utiliser les tests paramétriques.

Supposons maintenant que les tests réalisés montrent que la distribution ne peut être comparée à une distribution normale. Que faire? Si vous vous rappelez du point 5.5 sur les distributions, vous saurez que pour certaines, la moyenne n'est pas le meilleur indicateur. Par exemple, en cas d'asymétrie forte, la médiane peut remplacer avantageusement la moyenne comme mesure de la tendance centrale. Nous allons voir au point suivant comment résoudre ce problème.

10.3. Test non-paramétrique de Wilcoxon

10.3.1. Principe du Test

Pour utiliser ce test, il est nécessaire que les variables soient au minimum ordinales. L'hypothèse sous-jacente au test est que la médiane (θ) de l'échantillon concerné provienne d'une population dont la médiane ait une valeur théorique spécifique. En cas de test significatif, nous devrons donc en conclure que la médiane de l'échantillon est significativement différente de la médiane théorique spécifiée. C'est donc bien un test équivalent au test-t pour échantillon unique, mais adapté à la médiane.

Le test de Wilcoxon est ce qu'on appelle un test de rang, qui se base sur les rangs de la différence entre les scores de la série et la valeur théorique de la médiane. La méthode de rang est une méthode utilisée dans de nombreux tests dont il est important de comprendre le mécanisme.

10.3.2. La méthode des rangs

Cette méthode consiste à transformer les données en ne tenant pas compte de leur valeur initiale mais bien de la position qu'elles occupent dans une distribution. Par exemple, une série statistique telle que 3, 235, 17 pourrait être ordonnée en deux temps. Le premier consiste à ranger les trois valeurs par ordre croissant : 3, 17 et 235. Le second est d'attribuer un rang à chacune des valeurs : 3 = 1 ; 17 = 2 ; 235 = 3. Nous aurions donc la série 1, 2, 3.

En cas d'égalité, on attribuera le rang moyen à chacune des valeurs égales. Imaginons par exemple la série statistique suivante (elle est déjà triée par ordre croissant) : 13, 32, 32, 123. Il y a quatre valeurs, donc quatre rangs à attribuer. Cependant, deux des valeurs sont égales, la valeur qui occuperait le rang 2 et la valeur qui occuperait le rang 3. Mais il n'y pas de raison d'attribuer un rang inférieur à une des deux valeurs de 32 et un rang supérieur à l'autre. Donc, nous attribuerons à chacune le rang moyen (2+3)/2 = 2,5. La série ordonnée devient : 1 ; 2,5 ; 2,5 ; 4.

10.3.3. Approche du test par l'exemple

Supposons un thérapeute féru de statistique qui voudrait tester son efficacité. Parmi les nombreux indicateurs qu'il évalue, il établi un lien entre l'efficacité de ses thérapies et le nombre de séances auxquelles participent ses patients. Il estime qu'une efficacité optimale est atteinte lorsque les patients sont vus durant environ 10 séances. En-deçà, il estime que la thérapie est trop courte pour porter ses fruits, au-delà il estime avoir été inefficace puisque le patient a toujours besoin de ses services. Pour tester cet indicateur d'efficacité, il prélève aléatoirement 10 de ses patients de ces 5 dernières années et évalue le nombre de séances auxquelles ils ont participé. Il obtient la série suivante : 18, 20, 16, 8, 16, 6, 0, 20, 10, 10. Le zéro correspond à un patient qui a téléphoné plusieurs fois pour prendre rendez-vous mais qui ne s'y est jamais présenté.

L'hypothèse de ce thérapeute est donc que la médiane est égale à 10. Le Tableau 10.1 représente la série statistique ainsi que les transformations nécessaires pour effectuer le test de Wilcoxon. La deuxième colonne reprend la série statistique initiale. La troisième colonne calcule la différence (D) entre les scores initiaux et la médiane. La quatrième colonne attribue un rang à chacune des valeurs de |D|(la différence en valeurs absolues). Attention, on n'attribue aucun rang à la valeur zéro, ce qui correspond à éliminer de l'analyse les sujets qui sont venus exactement 10 fois. La dernière colonne est identique à la quatrième mais à laquelle on attribue le signe correspondant au sens de la différence observée à la troisième colonne. En bas de cette dernière colonne, on rapporte la somme des rangs positifs et la somme des rangs négatifs. Si les écarts par rapport à la médiane étaient identiques en négatif et en positif, nous aurions bien une médiane de 10. En revanche, s'il y a plus d'écarts d'un signe (négatif ou positif) que de l'autre, cela signifie que la médiane théorique ne correspond pas à la médiane observée.

Tableau 10.1. : Informations nécessaires pour réaliser un test de rangs de Wilcoxon.

Sujets	Xi	D = Xi - θ	Rang de |D|	Rang avec signe
1	18	18-10 = 8	8	5
2	20	20-10 = 10	10	7
3	16	16-10 = 6	6	3,5
4	8	8-10 = -2	2	-1
5	16	16-10 = 6	6	3,5
6	6	6-10 = -4	4	-2
7	0	0-10 = -10	10	-7
8	20	20-10 = 10	10	7
				ΣR+ = 26
				ΣR- = -10

La valeur que nous utiliserons pour réaliser le test de Wilcoxon est, par convention (c'est sur cette base que les tables ont été construites), la plus petite des valeurs de la somme des rangs positifs (ΣR+ = 26) et de la somme des rangs négatifs (ΣR- = -10). Ici, nous considérerons donc la valeur -10, dont nous considérerons la valeur absolue : 10. Il existe bien entendu une table (voir Tableau 10.2) qui reprend les valeurs critiques pour le test de Wilocoxon. On y voit que la valeur critique pour une erreur α de 5% et un échantillon de n = 8 (puisqu'on a ôté deux sujets de l'analyse) est de 3. Pour que le test soit significatif, il faudrait que la plus petite somme soit plus petite que 3, or elle est de 10. Nous ne pouvons donc pas conclure que la médiane observée est significativement différente de la valeur théorique 10 (attention ce 10 est la valeur théorique de la médiane, pas la valeur absolue de la ΣR-). Le thérapeute peut-il donc être satisfait de son travail? L'examen de la série pourrait nous inspirer la prudence. Le fait que 4 des 10 patients aient assisté à beaucoup plus de séances que les 10 attendues et la prise en compte du patient qui n'est jamais venu laisse penser que s'il avait pris 20 patients au lieu de 10 il aurait pu atteindre une conclusion différente. Finalement, seuls deux patients ont effectivement assisté à 10 séances.

Remarque sur le rejet de l'hypothèse (bis)

Nous mettons ici, par l'exemple, un problème fréquent d'interprétation en évidence (que nous avons déjà discuté au point 8.5.3, dans l'encadré) : notre thérapeute n'aurait pas pu s'assurer que la médiane était comparable à une médiane de 10. Tout ce qu'il pouvait faire c'est rejeter ou non son hypothèse d'égalité. Mais ne pas la rejeter ne veut pas dire l'accepter. En effet, si le test avait été significatif, il aurait pu être raisonnablement sûr que ses thérapies comportaient trop (ou pas assez) de séances par rapport à ses exigences de 10. En revanche, le fait de ne pas pouvoir rejeter l'hypothèse d'égalité ne veut pas forcément dire que la médiane de la population est égale à 10. Cela veut juste dire que, étant données les informations dont il dispose et la méthodologie qu'il a utilisé, il ne peut pas considérer que cette médiane est différente. Cela peut donc être par manque d'informations (pas assez de sujets considérés) ou parce que la médiane de la population est effectivement égale à 10.

Il pourrait vous sembler anormal, si vous n'avez pas suivi parfaitement bien le raisonnement (en ce qui me concerne, il m'a fallu du temps pour comprendre), que plus la valeur est grande, moins on en conclut que les médianes sont différentes. Cela s'explique parce que nous sommes en train de comparer la somme ΣR- à la valeur qu'elle aurait SI ELLE ETAIT IDENTIQUE A LA ΣR+. Pour un effectif donné, la somme des valeurs absolues de ces deux sommes est donnée par la relation :

$$\Sigma R+ + |\Sigma R-| = n(n+1)/2 \text{ donc, ici, } 26 + 10 = 8*9/2 = 36$$

Si les sommes étaient les mêmes, elles seraient toutes deux égales à 18. Or l'une est égale à 26 et l'autre à 10. Pour être considérées comme anormalement différentes l'une de l'autre, il aurait fallu que la plus petite soit égale à 3 (et donc l'autre à 33). Vous penserez sans doute, qu'il faut vraiment une grande différence entre les valeurs pour qu'elle devienne significative. C'est effectivement le cas, bien que plus le *n* augmente, plus la différence est facilement visible puisque, comme pour les autres tests, on augmente la puissance.

Tableau 10.2. : Table des valeurs critiques pour le test de rangs de Wilcoxon (extrait de la table publiée par Sheskin, 2007, p. 1662)

	Niveau α Unilatéral					Niveau α Unilatéral			
	0,05	0,025	0,01	0,005		0,05	0,025	0,01	0,005
	Niveau α bilatéral					Niveau α bilatéral			
	0,10	0,05	0,02	0,01		0,10	0,05	0,02	0,01
n					*n*				
5	0	-	-	-	14	25	21	15	12
6	2	0	-	-	15	30	25	19	15
7	3	2	0	-	16	35	29	23	19
8	5	3	1	0	17	41	34	27	23
9	8	5	3	1	18	47	40	32	27
10	10	8	5	3	19	53	46	37	32
11	13	10	7	5	20	60	52	43	37
12	17	13	9	7	21	67	58	49	42
13	21	17	12	9	22	75	65	55	48

10.4. Exercices de fin de chapitre

T.P. 11 - CHAPITRE 10 : Exercice 1
Test de Wilcoxon : Bases théoriques

1. Pourquoi n'utilise-t-on pas la distribution khi-carré pour vérifier la normalité de la distribution de l'erreur ? (2 raisons)

2. Dans quel cas utilisera-t-on un test non paramétrique (plutôt qu'un test paramétrique) ?

3. Admettons qu'un chercheur s'intéresse aux sports les plus pratiqués par les adolescents belges et qu'il observe que la distribution de ses données (de son échantillon) ne peut être comparée à une courbe normale. Pourra-t-il réaliser un test de Wilcoxon sur ses données ?

4. Obtenir un résultat non significatif à un test permet-il d'affirmer avec certitude à une absence d'effet (ou dans le cas du test de Wilcoxon, à une absence de différence entre la médiane et la valeur théorique spécifiée ?)

T.P. 11 : Exercice 2
Test de Wilcoxon : Mise en situation

1. Un chercheur s'intéresse à l'obésité. Certaines personnes rapportent ne pas manger plus ni plus gras que la normal et pourtant continuent de grossir. Il décide alors de suivre ces personnes pendant un mois afin de noter la quantité de calories qu'elles mangent chaque jour. A la fin du mois, il calcule pour chaque personne la quantité moyenne de calories qu'elle consomme par jour et obtient les résultats suivants :

i	Nombre de calories par jour
1	2050

2	2150
3	1976
4	2347
5	2203
6	1901
7	2288
8	2100
9	2267
10	2301

Il pense que ces personnes se trompent probablement et sous-estiment le nombre de calories qu'elles consomment. Il désire alors vérifier si le nombre de calories moyen consommé par ces personnes est supérieur ou non à la valeur « normale » de 2100[43] calories. Il s'apprête à faire un test-t pour un échantillon mais constate que la distribution ne suit pas une courbe normale.

a. Quel test peut-il faire ? D'après l'énoncé, s'agit-il d'un test unilatéral ou bilatéral ? Faites-le ensuite.

b. Peut-on conclure que les membres de ce groupe consomment plus de calories qu'une personne « normale » contrairement à ce qu'ils affirment (avec un risque d'erreur de 5%) ?

c. Peut-on conclure que les membres de ce groupe ne consomment pas plus de calories qu'une personne « normale » comme ils l'affirment ?

2. Un psychologue veut tester si le jardin d'enfants influence le comportement social des élèves. En effet, la pédagogie de jeu utilisée dans les jardins d'enfants favorise et encourage l'expression personnelle et le développement de la créativité. Les activités favorisent le développement global de l'enfant tout en respectant son développement personnel. Une précédente étude avait mis en évidence un lien entre le temps passé en garde d'enfants (sans quantifier la qualité) et les comportements négatifs comme l'impulsivité et l'agressivité. Plus les enfants passaient de temps en jardin d'enfants, moins ils étaient susceptibles

[43] La valeur n'est pas tout à fait exacte car dépend de plusieurs autres paramètres individuels mais pour faciliter le calcul, on s'en contentera.

d'avoir des comportements à problèmes comme la désobéissance, l'agressivité, et ce, dès l'école maternelle. Selon un certain psychologue, le comportement social est bénéfiquement influencé à partir du score de 50 (sur une échelle X mesurant un score social : plus les notes sont élevées, plus le comportement social est adéquat). Les notes de score social obtenues par les enfants sont les suivantes :

jardin
82
69
73
43
58
56
76
65

Statistiques

VAR00001

N	Valide	8
	Manquante	0
Asymétrie		-,551
Aplatissement		-,057

a. Que pouvez-vous dire sur G_1 et G_2 au vu de la boîte à moustaches ci-dessus ? Quelles conséquences cela aura-t-il sur l'analyse des données et le test à appliquer ?

b. Pour déterminer si la fréquentation du jardin d'enfant a un effet positif sur le comportement social des enfants, formulez les hypothèses nulles et alternatives et réalisez le test adapté.

3. Voici une série de 15 données :

X_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	53	86	74	77	68	65	62	59	70	70	74	66	69	75	76

a. Pouvez-vous déterminer si la médiane de cet échantillon diffère significativement de 70 ?

CHAPITRE 11 : PRESENTATION DES RESULTATS

Lorsque l'on transmet les résultats statistiques à une audience, il est une série de conventions à respecter pour se faire facilement comprendre. En psychologie, l'entièreté de ces conventions sont reprises dans un manuel établi par l'American Psychological Association (APA). Les standards conseillés par cette association sont actuellement adoptés par la communauté scientifique mondiale en psychologie (à peu de choses près). Régulièrement, des aménagements sont faits et quelques procédures changent. Le manuel est donc réédité. Nous en sommes actuellement à la sixième version. Ces standards ne concernent pas seulement la présentation des résultats, ils énoncent les règles bibliographiques, la manière de présenter les figures, les tables, les titres, le corps du texte, bref, tout ce qui concerne la mise en page d'un article scientifique. Vous pourriez vous dire que ce point n'est pas très important pour la compréhension des statistiques, mais ce n'est pas si évident : les statistiques sont, en soi, à la fois un outil incontournable et à la fois une difficulté. Il est donc important de minimiser les inconvénients associés à cette difficulté. Lorsque vous lirez des articles scientifiques, il vous sera agréable de ne pas devoir, préalablement à la compréhension du contenu, décrypter les habitudes conventionnelles de la revue. Par ailleurs, si un jour vous devez publier par vous-mêmes, votre production ne sera jamais éditée si vous ne vous pliez pas à l'utilisation de ces conventions. Je vous enjoins donc à y être attentifs dès à présent et à vous familiariser avec ce manuel. Vous en trouverez des exemplaires à la bibliothèque. Il est inutile pour ce cours de les consulter, puisque je reprends l'essentiel de ce qui est nécessaire dans les points suivants, mais vous en aurez sûrement besoin lors de votre cursus. Dans la mesure où l'essentiel de la littérature est anglophone (même les revues francophones ont de plus en plus tendance à donner une version française et une version anglaise) je vous donnerai les habitudes anglophones et traduirai lorsque c'est nécessaire.

11.1. Les *p*-valeurs

Les p-valeurs ne sont rien d'autre que la spécification du risque α de commettre l'erreur de première espèce. Plusieurs formes sont acceptées. Par exemple, on peut indiquer si le risque est plus grand ou plus petit que l'habituelle exigence de 5%. On écrira dans ce cas : $p < .05$ ou $p > .05$ (dans ce dernier cas, on pourra également indiquer *NS* pour "*non significant*"). Remarquez que le "*p*" ou le "*NS*" sont en italique, remarquez également qu'on utilise le point et non la virgule comme séparateur entre l'unité et les décimales (à l'inverse de ce que j'ai

fait dans l'entièreté de ce syllabus pour des questions de confort). Il vaut parfois la peine d'indiquer au lecteur que la p-valeur n'est pas suffisante, mais qu'elle n'est pas loin de l'être. Cette indication peut se donner lorsque la p-valeur est plus grande que .05 mais plus petite que .10. On dit alors que le test est marginalement significatif (*marginally significant*) et on écrit $p < .10$. Lorsque le lecteur lit cela, il sait que la probabilité est plus grande que .05, sinon ce serait la première notation qui serait utilisée. Lorsque les résultats sont "particulièrement significatifs", les auteurs l'indiquent souvent : $p < .01$ ou $p < .001$. En ce qui me concerne, je n'aime pas tellement cette pratique. En effet, lorsque vous faites un test statistique, le plus souvent vous décidez PREALABLEMENT votre risque α. Dès lors, même si vous auriez pu rejeter votre hypothèse nulle en étant beaucoup plus exigeant que cela, c'est trop tard (et il ne faut pas le regretter puisqu'être plus exigeant implique de diminuer la puissance). Ceci dit, il est vrai que les logiciels actuels vous donnent la vraie *p*-valeur en un clic. Tout le monde sait donc que l'exigence est de 5% mais que la vraie p-valeur est rapportée.

Cette situation est devenue si évidente qu'il est de plus en plus conseillé (et c'est très certainement ce que je vous conseille personnellement) d'adopter une autre option : **rapporter systématiquement la vraie p-valeur**. Vous arrêterez donc d'indiquer $p < .05$ mais vous opterez pour $p = .038$ (par exemple). Vous avez le choix entre l'utilisation de deux ou de trois chiffres de précision. Personnellement j'estime que deux suffisent amplement.

Dans le cas où la significativité est telle que la *p*-valeur est infinitésimale, par exemple, 0.00000000000000377676, on ne rapporte évidemment pas un tel nombre. On se contentera alors de dire que $p < .001$ qui est la précision minimale qui nous occupe. Attention, n'écrivez jamais $p = .000$, cela suggérerait une probabilité nulle qui n'est pas concevable en inférence statistique, puisqu'une hypothèse doit, rappelez-vous, être falsifiable : on ne fait pas de probabilités sur des événements certains.

11.2. Statistiques descriptives

Les deux principales statistiques descriptives que vous devez transmettre sont évidemment la moyenne et l'écart-type (qu'on préfère, le plus souvent, à la variance). Elles peuvent apparaître dans le texte ou dans un tableau. Nous n'utiliserons jamais de symbole bizarre comme nous pouvons en utiliser dans ce syllabus. La moyenne se rapporte par l'initiale "*M*" et l'écart-type par les initiales "*SD*" en anglais (pour "*Standard Deviation*") ou "*ET*" en français (devinez pour quoi). Le niveau de précision est toujours de deux chiffres. Pour

indiquer cette convention, un nombre entier se notera néanmoins avec deux zéros de décimales.

Exemple 1

La moyenne des cotes du cours sur cent points est de 60.00 (*ET* = 24.70).

Exemple 2

Quatorze participants ont effectué cet examen ($M_{âge}$ = 19.32 ; *ET* = 4,23).

Dans un tableau, le principe est le même. Le tableau 11.1 suivant montre un exemple. Remarquez que d'une part, on écrit table et pas tableau comme je l'ai fait dans ce syllabus. D'autre part, tous les articles sont écrits en police de caractère "Times New Roman" et non "Constancia" comme ici. La taille est de 12 points. Il faut un double interligne (dans ce syllabus il n'y a qu'un interligne et demi).

Table 11.1

Résultats complètement fictifs de deux variables quelconques pour les hommes et pour les femmes

	Variable A		Variable B	
	M	*ET*	*M*	*ET*
Hommes (*n* = 20)	100	10	120	12
Femmes (*n* = 20)	120	12	100	10

11.3. Intervalle de confiance

Les intervalles de confiance, lorsqu'ils sont rapportés, se mettent entre crochets. Il est évidemment nécessaire d'informer votre public de votre niveau d'exigence. Par exemple, vous écrirez : l'intervalle de confiance à 95% est de [2.08 ; 3.78]. Cependant, il vous est toujours loisible de l'indiquer autour d'une moyenne : la moyenne de l'échantillon est de 60.00 ± 3.42 (α = 5%).

11.4. Test-t

La principale donnée dont a besoin le lecteur pour vérifier votre statistique est le nombre de degrés de liberté. Il doit donc être indiqué entre parenthèses. Par ailleurs, vous devrez également rapporter la p-valeur associée. Enfin, nous verrons l'année prochaine, qu'il existe plusieurs types de test-t. Cette année nous n'avons considéré que le test-t pour un échantillon. Le type de test-t est en général tout à fait évident lorsque le lecteur lit votre section résultat, vous ne devrez donc rien spécifier à ce sujet. Encore une fois la lettre "*t*" doit être en italique. Ce sera la même chose pour toutes les lettres se rapportant à une statistique.

Exemple

Les 30 participants à cette étude ont une moyenne d'âge de 32.40 (*ET* = 12.60) et étaient significativement plus âgés que la moyenne d'âge de la population habituelle des étudiants de cette année qui est de 20.4 ans, $t(29) = 5.22$, $p < .001$.

11.5. Statistique F

La statistique *F* est un rapport entre deux variances. Il est donc nécessaire de connaître les degrés de liberté du numérateur et du dénominateur. La p-valeur doit également figurer. Nous avons déjà vu comment rapporter une telle statistique sous forme de tableau (voir Tableau 8.8, par exemple). Dans le texte, la statistique se rapporte de la manière suivante :

Exemple

Les 12 participants dans la condition de dosage élevée ont un temps de réaction moyen de 12.30 secondes (*ET* = 4.10) ; les 9 participants dans la condition de dosage modérée ont un temps de réaction de 7.40 secondes (*ET* = 2,30) ; et les 8 participants dans la condition contrôle ont un temps moyen de 6.60 secondes (*ET* = 3.10). L'effet du dosage est donc significatif, $F(2, 26) = 8.76$, $p = .012$.

Source : http://my.ilstu.edu/~mshesso/apa_stats.htm, retrouvé le 1/11/11.

11.6. Test Khi-Carré

De manière identique aux autres tests, on indiquera la valeur de la statistique, le nombre de degrés de liberté et la p-valeur. Prenons l'exemple du chapitre 9.

Exemple

La distribution des 60 jets de William L. B. n'est pas compatible avec une distribution uniforme que l'on serait en droit d'attendre de la part d'un dé bien équilibré, $\chi^2_{(5)} = 14.2$, $p < .05$ (*ou la vraie p-valeur si vous l'avez, mais dans ce cas, nous n'avons pas utilisé de logiciel, il est donc plus difficile de l'obtenir ce qui explique probablement que cette écriture soit encore autorisée*).

11.7. Test de Wilcoxon

Il n'y a pas, dans le manuel de l'APA, de règle spécifique à ce test (qui est d'ailleurs très rarement utilisé). Je vous suggère de garder la même habitude que pour les autres tests. Le test de Wilcoxon est un test qui dépend essentiellement du degré de liberté, lui aussi, et donc cette information doit être rapportée. En outre, la table est également fonction du degré de risque α exigé, rapporté lui aussi. Vous utiliserez la lette "W" en italique. De sorte que vous obtiendrez par exemple un phrase dans laquelle se situera : $W(20) = 3$, $p < .001$.

11.8. Exercices Récapitulatifs de tous les chapitres

T.P. 12 : EXERCICES DE REVISION

Exercice 1 : Hypothèses et Variables

1. Mettez une croix devant les phrases que vous considérez être des hypothèses. Justifiez votre réponse et précisez, le cas échéant, s'il s'agit d'une hypothèse théorique (T) ou opérationnelle (O). S'il s'agit d'une hypothèse théorique, transformez-la en hypothèse opérationnelle et vice-versa. S'il ne s'agit pas d'une hypothèse, transformez-la si possible en hypothèse. S'il s'agit d'une hypothèse, donnez la variable indépendante et la variable dépendante.

	H_o ?	T ou O
Les enfants qui ont de grands pieds sont meilleurs en calcul mental que ceux qui ont de petits pieds.		
Justification :		
VI et VD :		
Hypothèse transformée :		
Les hommes sont machos.		
Justification :		

VI et VD :		
Transformation :		
Les femmes veulent toujours avoir raison.		
Justification :		
VI et VD :		
Transformation :		
Est-ce que les bébés allaités « font leurs nuits » plus tard que les bébés qui ne le sont pas ?		
Justification :		
Transformation :		
VI et VD :		
Les enfants marchent à quatre pattes avant de marcher et ils marchent avant de courir.		

Justification :
Transformation :
VI et VD :

2. Des chercheurs s'intéressent à l'obésité chez les jeunes. Donnez deux exemples de variables (et leur codage) quantitatives et qualitatives qui pourraient être utilisées dans le cadre de cette étude.

3. Des chercheurs s'intéressent à l'obésité chez les jeunes. Donnez deux exemples de variables indépendantes et dépendantes qui pourraient être utilisées dans le cadre de cette étude.

4. Veuillez déterminer deux variables qui pourraient influencer le choix vestimentaire d'une jeune adolescente (celles qui vous paraissent les plus pertinentes). De quel type de variable s'agit-il ?

5. Construisez deux hypothèses opérationnelles à l'aide de ces variables.

T.P. 12 : Exercice 2

Logique déductive et Syllogisme d'Aristote

Soit, le syllogisme suivant :

Tous les oiseaux ont des ailes.
Or, aucun chien n'est un oiseau.
Donc, aucun chien n'a d'ailes.

S'agit-t-il d'un raisonnement valide ? Veuillez justifier votre réponse

T.P. 12 : Exercice 3

Article

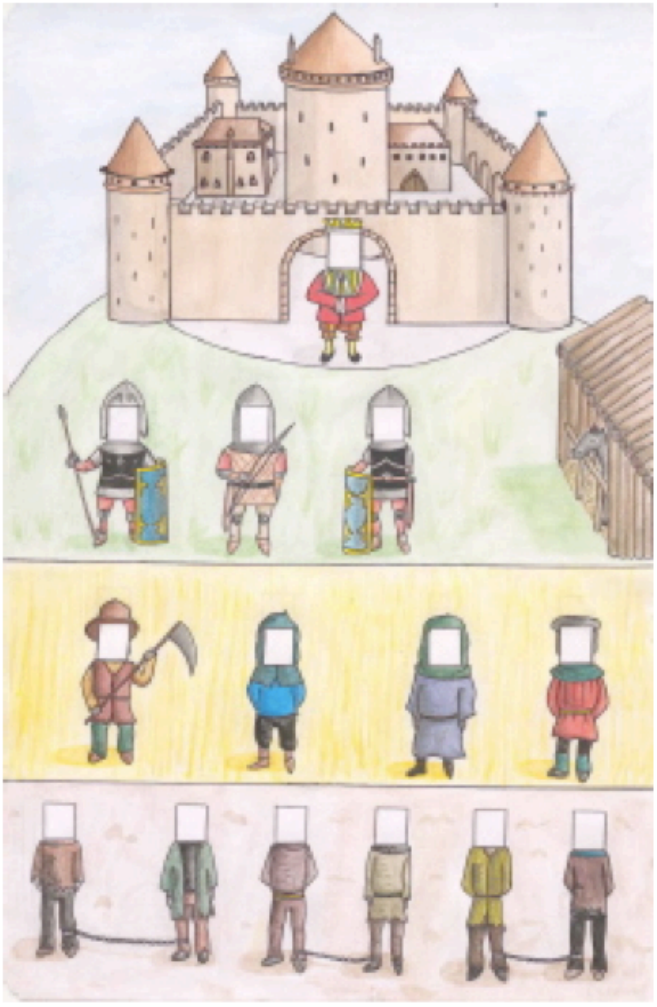
RESUME

Le problème soulevé dans cette étude est de savoir si des enfants âgés de 9 ans ont une connaissance des inégalités sociales entre groupes ethniques et s'ils les perpétuent. Le dessin créé pour tester ce phénomène représente une hiérarchie sur laquelle les enfants avaient pour tâche de placer des visages blancs ou noirs. Conformément à notre hypothèse, les résultats montrent que les enfants ont placé essentiellement des visages blancs en haut de la hiérarchie et des visages noirs en bas. Nous pouvons donc conclure que les enfants valorisent l'ethnie majoritairement mise en avant dans notre société et dévalorisent l'ethnie minoritaire.

INTRODUCTION

En France, dans de nombreux domaines, l'ethnie la plus valorisée est celle englobant les individus de type européen. Ainsi, dans les médias et dans la plupart des institutions, les « blancs » sont souvent représentés au sommet de la hiérarchie. Les enfants quant à eux sont confrontés à ce phénomène de valorisation sociale par l'intermédiaire des contes, de l'école, et toutes sortes de programmes télévisuels qui leurs sont destinés. Cependant, nous pouvons nous demander si les préjugés, les stéréotypes et la discrimination sont transmis à nos enfants par ces moyens de socialisation. Autrement dit, est-ce que la valorisation d'un groupe ethnique dans une société influence la valeur qu'attribue un enfant à son propre groupe ethnique ?
Allport en 1954 (cité par Bourhis, Gagnon, Moïse, 1999, p.162) définit le préjugé comme « une attitude négative ou une prédisposition à adopter un comportement négatif envers un groupe, ou envers les membres de ce groupe, qui repose sur une généralisation erronée et rigide. […] On classe souvent les préjugés selon la catégorie sociale qui est l'objet de la généralisation. ». Dans l'étude de Kaylin, Rayko 1978 (cité par Bourhis, Gagnon, Moïse, 1999) des sujets anglo-canadiens avaient pour tâche de simuler une embauche en tant que directeur du personnel. Ils écoutaient des extraits d'entretiens dans lesquels les postulants avaient un accent anglo-canadien ou un des trois autres accents. L'origine ethnique était la seule information dont disposaient les sujets. Après l'écoute, ils lui attribuaient un des quatre postes au statut hiérarchiquement différent. Les résultats de cette étude montrent que les participants adoptaient un comportement discriminatoire envers les membres de l'exogroupe et favorisaient les membres de l'endogroupe. La discrimination est définie par Dovidio et Gaertner en 1986 (cité par Bourhis, Gagnon, Moïse, 1994, p.164) comme étant «un comportement négatif envers des individus membres d'un exogroupe envers lequel nous entretenons des préjugés ». On sait également par l'intermédiaire de nombreuses expériences que les enfants sont enclins à avoir des préjugés, des stéréotypes et des comportements discriminatoires. Les expériences de Maras (1993) et de Powlishta, Serbin, Doyle et White, citées par Guimond et Dambrun en 2003, montrent que les enfants dès le plus jeune âge sont capables de catégoriser les groupes stigmatisés. Les mêmes auteurs citent l'étude de Yee et Brown de 1992, dans laquelle des enfants participent à un jeu. Ils étaient séparés en deux

groupes : celui des « lents » et des « rapides ». Les résultats montrent que les « enfants sont sensibles aux différences de statut intergroupe ». De la même manière, les enfants montrent des comportements de favoritisme envers le groupe à fort statut.
La littérature, traitant des comportements des enfants quant à leurs comportements intergroupes, ne donne pas d'exemple quant à leur capacité de hiérarchisation. L'expérience qui suit a directement été inspirée de l'étude de Kaylin et Rayko en 1978, précédemment citée, et nous nous sommes intéressées à la possibilité que ces résultats soient observables chez les enfants. Nous en sommes donc venues à supposer qu'en raison d'une valorisation différente des ethnies en France, tous les enfants valoriseront l'origine ethnique majoritairement mise en avant par la société, c'est-à-dire « les blancs », et déprécieront l'origine ethnique minoritaire, « les noirs ». Ainsi, par l'intermédiaire d'un dessin fabriqué pour les besoins de l'expérience, nous avons représenté quatre niveaux hiérarchiques : un roi, trois chevaliers, quatre paysans et six esclaves, auquel nous avons ajouté une série de visages blancs ainsi qu'une série de visages noirs. Nous prédisons que les enfants mettront très fréquemment des visages blancs au niveau Roi, un peu moins fréquemment au niveau Chevaliers, peu fréquemment au niveau Paysans, et très peu fréquemment au niveau Esclaves.



1. Soulignez dans le texte ci-dessus les arguments mis en avant par les auteurs pour justifier leur hypothèse.

2. Formulez l'hypothèse générale et l'hypothèse opérationnelle.

3. Ces hypothèses ont-elles bien les propriétés indispensables à toute hypothèse ? Développez.

Voici la méthode et les résultats de l'étude :

METHODE

Participants :

Il s'agit de 43 enfants : deux classes de CM1, une élève de CE2 et un élève de 6e d'une école dont l'âge moyen était de 9 ans et 4 mois. Le groupe était exclusivement composé d'enfants d'origine européenne (à une exception près : asiatique). Le groupe était également composé de 19 garçons et de 24 filles. Les participants n'étaient pas rémunérés.

RESULTATS

Après avoir quotté par 1 quand l'enfant répondait par un visage blanc et 0 pour noir, nous avons comptabilisé le nombre de réponses « visage blanc » en fonction du nombre total de réponses, les pourcentages obtenus sont conformes à l'hypothèse soulevée précédemment : les enfants ont placé majoritairement des visages blancs en haut de la hiérarchie.

On constate que plus on descend dans la hiérarchie, plus on observe une diminution du nombre de visages blancs : au niveau Roi le score est de 81.395%, au niveau Chevaliers il est de 67.441%, au niveau Paysans il atteint 52.325%, enfin, le niveau Esclaves compte 37.984% de visages blancs (cf. Figure 1).

4. Quelle est la question de recherche ? Quelle est la variable dépendante ? Sur quelle échelle de mesure est-elle mesurée ?

5. Quelle en est la variable indépendante ? Sur quelle échelle de mesure est-elle mesurée ?

Les auteurs dressent le graphique suivant :

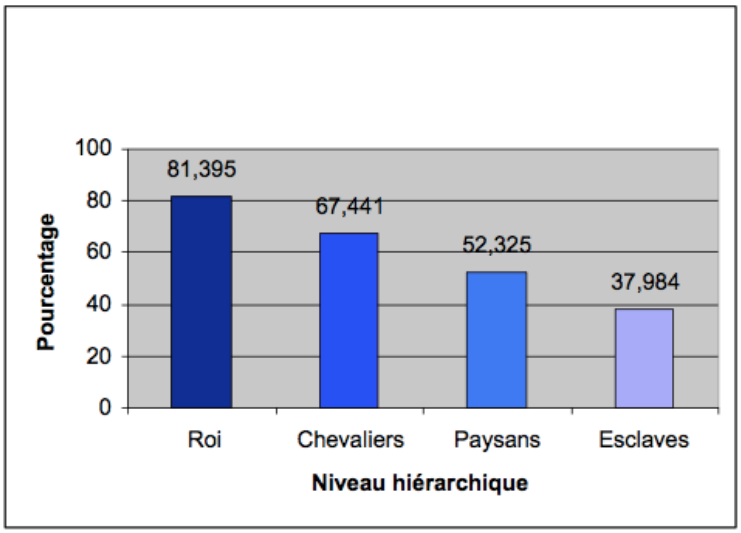


Figure 1: Pourcentage des visages blancs en fonction du niveau hiérarchique

6. De quel genre de graphe s'agit-il ? Pourquoi l'usage de ce graphique est-il requis ? Donnez-en les caractéristiques. Commentez-le.

7. Que pensez-vous de cette étude ? Vous paraît-elle scientifiquement pertinente ?

T.P. 12 : Exercice 4

Séries statistiques et distributions statistiquesReprésentation graphique et quantiles

Afin d'établir un bilan de productivité individuelle des 80 ouvriers d'une entreprise bruxelloise, un relevé du nombre de pièces effectuées en une journée par chacun d'eux est effectué.

70	77	83	91
70	78	83	91
71	78	83	92
71	79	83	92
71	80	84	92
71	80	85	92
71	80	86	92
72	80	87	92
72	81	87	94

72	81	89	94
73	82	89	96
74	82	89	96
74	82	89	96
75	82	89	97
75	82	90	98
76	82	90	98
76	83	90	100
76	83	90	100
77	83	90	100
77	83	91	101

3. Etablissez le tableau de distribution des fréquences correspondant à ces données. Comment appelle-t-on une telle transformation des données ?

4. Représentez visuellement ces données sous forme de graphique en bâton. Sur ce graphique, veuillez estimer la localisation de la médiane.

5. Quelle est la valeur modale ?

6. Combien de pièce produira un individu situé sur le $35^{ème}$ percentile ?

7. Sur quel percentile se situera un individu qui a produit :

   a. 71 pièces ?

   b. 98 pièces ?

8. Lorsque le nombre de données est important, il peut être intéressant de recourir au groupement en classe. Veuillez établir un histogramme contenant 4 classes d'étendues identiques.

T.P. 12 : Exercice 5

Exploration algébrique des données à une dimension

Voici une distribution statistique correspondant à l'âge des étudiants en BA2 :

j	Valeurs de la variable : x_j	Fréquences absolues :	Fréquences relatices : f_j	Fréquences relatives cumulées
1	18	1		
2	19	49		
3	20	46		
4	21	31		
5	22	25		
6	23	11		
7	24	4		
8	25	2		
9	26	2		
10	27	2		
11	29	2		
12	32	2		
13	33	2		
14	40	1		
15	42	1		
		n=181		

1. En quoi une distribution est-elle différente d'une série?

2. Complétez dans le tableau ci-dessus les fréquences relatives et les fréquences relatives cumulées associées à cette distribution.

3. Calculez les différents indicateurs de la tendance centrale de cette distribution (en considérant que la variable est discrète).

4. Calculez les différents indicateurs de la dispersion de cette distribution.

5. Calculez les coefficients de symétrie et d'aplatissement de cette distribution. Qu'en concluez-vous?

6. Vérifiez les informations algébriques par une représentation graphique de vos données.

7. Estimez la moyenne et la variance de la population correspondante.

8. Quelle est cette population? Justifiez votre réponse.

9. Quelles seraient les implications de ne pas tenir compte des valeurs aberrantes?

T.P. 12 : Exercice 6

Article : Discussion

L'impact des prix à « terminaison 9 » sur le comportement d'achat a été peu étudié alors qu'il s'agit d'une pratique largement utilisée. Une seule recherche expérimentale a montré qu'elle ne contribuait pas à augmenter le nombre d'acheteurs mais permettait d'augmenter le panier moyen. Après une présentation synthétique des recherches sur ce thème et de l'état théorique sur ce thème, une nouvelle expérience a été conduite dans laquelle des vendeurs à domicile vendaient des pâtisseries au profit d'une association humanitaire. Les prix, donnés oralement, étaient soit à terminaison « 9 » (1.99 € la demi-douzaine de crêpes) soit à terminaison « pleine » (2.00 € la demi-douzaine de crêpes).

Le tableau chiffré suivant correspond aux résultats de l'étude :

	Prix « 9 » 1.99 €	Prix Plein 2.00 €
Taux d'achat (en %)	47.4 % (117/247)	58.2 % (152/261)
Nombre moyen de paquets achetés		
Moyenne	1.24	1.31
Ecart-type	0.36	0.32
N	117	152

Tableau 1. *Taux d'achat et moyenne du nombre de produits achetés selon les conditions de présentation des prix.*

Commentez ce tableau.

T.P. 12 : Exercice 7

Probabilités

A. Dans un village de vacances, trois stages sont proposés aux adultes et aux enfants. Ils ont lieu dans la même plage horaire ; leurs thèmes sont : la magie, le théâtre et la photo numérique. 150 personnes dont 90 adultes se sont inscrites à l'un de ces stages.

Parmi les 150 personnes inscrites, on relève que :

- la magie a été choisie par la moitié des enfants et 20% des adultes ;
- 27 adultes ont opté pour la photo numérique ainsi que 10% des enfants.

1. Complétez le tableau suivant :

	Magie	Théâtre	Photo numérique	Total
Adultes				
Enfants				

On appelle au hasard une personne qui s'est inscrite à un stage. On pourra utiliser les notations suivantes :

- A l'événement " la personne appelée est un adulte " ;
- M l'événement " la personne appelée a choisi la magie " ;
- T l'événement " la personne appelée a choisi le théâtre " ;
- N l'événement " la personne appelée a choisi la photo numérique ".

2. a) Quelle est la probabilité que la personne appelée soit un enfant ?

   b) Quelle est la probabilité que la personne appelée ait choisi la photo sachant que c'est un adulte?

   c) Quelle est la probabilité que la personne appelée soit un adulte ayant choisi le théâtre ?

3. Montrer que la probabilité que la personne appelée ait choisi la magie est 0,32.

4. Le directeur du village désigne une personne ayant choisi la magie. Il dit qu'il y a deux chances sur trois pour qu'elle soit un enfant. A-t-il raison ? Justifier votre réponse.

B. 1200 élèves se présentent à un examen. Ces étudiants peuvent choisir une, deux ou aucune des options artistiques facultatives : musique et dessin.
On sait que parmi ceux-ci :

- 510 vont passer l'option Musique (certains d'entre eux passeront aussi l'autre option)
- 410 vont passer l'option Dessin (même remarque que ci-dessous)
- 390 ne passeront aucun de ces deux options.

On définit les événements suivants :
M : « l'élève choisi passera l'option Musique »
D : « l'élève choisi passera l'option Dessin »

On choisit un élève au hasard parmi ces 1200 élèves.

1. Calculez P(M), P(D)?

2. Définissez en français les événements $(\sim M \cap \sim D), (M \cup D)$ et $(M \cap D)$. Calculez la probabilité de ces trois événements.

3. Ecrivez en notation ensembliste les événements suivants :
   a. « L'élève choisi passera uniquement la Musique »
   b. « L'élève choisi passera uniquement le dessin »
   c. « L'élève choisi passera exactement une des deux options »

   Calculez les probabilités de ces trois événements.

C. Le tableau suivant donne la répartition de 150 stagiaires en fonction de la langue choisie et de l'activité sportive choisie.

	Tennis	Equitation	Voile
Anglais			
Allemand			

On choisit un élève au hasard.

On définit les événements suivants :`

- An : l'élève choisi étudie l'anglais;
- Al : l'élève choisi étudie l'allemand ;
- T : l'élève choisi pratique le tennis ;
- E : l'élève choisi pratique l'équitation ;
- V : l'élève choisi pratique la voile

a. Les événements « étudier l'allemand » et « pratiquer le tennis » sont-ils indépendants ?

b. Les événements « étudier l'anglais » et « pratiquer la voile » sont-ils indépendants ?

c. Exprimez en français les événements suivants et calculez leur probabilité (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées).

$Al\|T$: $\mathrm{P}(Al\|V)$ =
$V\|An$: $\mathrm{P}(V\|An)$ =

d. Répondez aux questions a) et b), justifiez vos réponses à l'aide d'une formule en utilisant uniquement les probabilités calculées au point c.

Evénements	**Indépendants ou non ?**	**Justification**
Al et T		

V et An		

e. Déterminez, à partir des résultats des points a) et b), les probabilités des événements suivants (exprimez les résultats sous forme de fractions irréductibles ou totalement simplifiées et justifiez en indiquant la formule utilisée).

Probabilité	**Résultat numérique**	**Formule**
$\mathrm{P}(V\|Al)$		
$\mathrm{P}(Al\|V)$		

T.P. 12 : Exercice 8

Standardisation - Score Z

1. La distribution des notes à un examen se distribue normalement autour d'une moyenne de 60 et d'un écart type de 4. Quelles formules utilisez-vous pour répondre aux questions suivantes :
    a. Quel est le score Z d'un élève qui obtient la note de 66 ?
    b. Quel est le score Z d'un élève qui obtient la note de 50 ?
    c. A quelle note correspond un score Z de + 2 ? Indiquez la formule que vous utilisez.

2. Si la moyenne d'une population pour une variable qui se distribue normalement vaut 50 et qu'une note de 43 correspond à un score Z de – 1. Quel est l'écart type de cette population ?

3. Les notes à un examen se distribuent normalement avec $\overline{X}$ = 68 et S_X = 6. Quelle est la probabilité qu'un étudiant ait une note supérieure à 72 ? Donnez les étapes du calcul.

4. Une distribution est normalement distribuée avec $\overline{X}$ = 100 et S_X = 10. Quel est le rang percentile pour X = 114 ? Quel est le rang percentile pour X = 92

5. Les scores d'un test psychologique se distribuent normalement avec $\overline{X}$ = 500 et S_X = 100. Quel est le score minimum nécessaire pour être dans les 15% supérieur de la distribution de ce test ?

6. Pour cette même distribution, quelle est la probabilité qu'un individu obtienne un score situé entre 600 et 650 ? En d'autres termes : $P(600 < X < 650)$.

7. Pour une distribution normale avec une moyenne de 500 et un écart type de 100, calculez :
   a. Quel score sépare les 40% supérieurs des 60% inférieurs de la distribution ?
   b. Quel est le score minimum nécessaire pour être dans le top 5% supérieur de la distribution ?

9. Une population se distribue normalement avec $\overline{X}$ = 60 et S_X = 5. Pour cette population, que vaut le 34ème percentile ?

T.P. 12 : Exercice 9
Approximation normale de la loi binomiale et correction pour la continuité

Soit la distribution binomiale suivante :

N	x	p 0.5
20	0	.0000
	1	.0000
	2	.0002
	3	.0011
	4	.0046
	5	.0148
	6	.0370
	7	.0740
	8	.1201
	9	.1602
	10	.1762
	11	
	12	
	13	
	14	
	15	
	16	
	17	
	18	
	19	
	20	

1. Complétez les cases vides du tableau, sans effectuer de nouveaux calculs. Justifiez la méthode utilisée.

2. Peut-on utiliser l'approximation de la binomiale par la normale ?

3. Calculez de deux manières différentes (en utilisant respectivement la binomiale et son approximation par la normale), la probabilité que le nombre de réussites soit compris 5 et 10 (5 et 10 inclus). Comparez les résultats.

T.P. 12 : Exercice 10

Chi-carré

En 2000, un chercheur s'est intéressé aux enfants qui présentaient des symptômes de phobie scolaire. Il a pris 50 enfants entre 7 et 12 ans et les a réparti en 3 catégories selon qu'ils ne présentaient aucun symptôme, des symptômes légers ou des symptômes importants. L'assignation d'un enfant dans l'une ou l'autre catégorie dépendant du résultat à un test élaboré par le chercheur. En 2011, ce même chercheur se demande si le nombre d'enfants présentant des symptômes graves n'est pas plus élevé qu'il y a 11 ans. Il décide de refaire une étude en prenant à nouveau 50 enfants et en leur faisant passer le même test puis de comparer les 2 résultats.

Il arrive aux résultats suivants :

	Année 2000	Année 2011
Aucun symptôme	41	34
Symptômes légers	7	11
Symptômes importants	2	5
Total	50	50

a. Peut-il conclure avec un risque d'erreur ne dépassant pas 5% que la répartition des enfants dans ces 3 catégories a changé depuis 2000 ? Et avec un risque d'erreur ne dépassant pas 1% ?

b. Peut-il conclure qu'il y a effectivement plus d'enfants qui présentent des symptômes importants qu'en 2000 ?

T.P. 12 : Exercice 11

EXERCICE RECAPITULATIF

A Monaco, une psychologue a récolté, pour 10 de ces patients, le nombre de troubles obsessionnels compulsifs (TOC) sur une semaine, sachant que tous ont suivi un programme d'intervention mis en place par la principauté et visant à réduire leurs TOC de tous les patients présentant des TOC à Monaco. En effet, ce programme suggère au spécialiste, ici

notre psychologue, de recevoir 3 fois par semaine les patients durant un mois, de les suivre une journée au quotidien afin de mieux comprendre les moments d'apparition des TOC et pour finir de leur conseiller une séance de relaxation en groupe.

Voici le nombre de TOC de ces 10 patients durant la semaine qui a suivi la fin du programme d'intervention : 35, 70, 35, 40, 105, 105, 100, 40, 35, 105.

1. Remarque importante : dans un premier temps, la psychologue souhaite utiliser uniquement des tests paramétriques pour faire ses analyses statistiques. Qu'entend-elle par là ?

2. Calculez la moyenne des TOC pour les 10 patients.

3. Que constituent ces 10 mesures ?

4. Comment pouvons-nous savoir si la moyenne de notre échantillon est représentative de notre population, sachant que la variance de la population vaut 90,25 (hé oui Monaco est assez petite pour pouvoir imaginer connaitre la variance) et qu'on souhaite se tromper avec un risque de 10% ? Effectuez le calcul.

5. Dans la question précédente, nous parlons d'un risque de se tromper, donc de probabilité et d'expérience aléatoire ; sur quel type de distribution avez-vous travaillé et que représente-t-elle ?

6. Quelle est la probabilité d'avoir une moyenne inférieure ou égale à 82,68 ?

7. Comment pouvons-nous déterminer si la moyenne de notre échantillon est représentative de notre population, sachant que nous avons perdu le document sur lequel la variance de la population était notée, que nous ne connaissons que la variance de l'échantillon (hélas, la variance de la population n'est que très rarement connue) et que l'on souhaite se tromper avec un risque de 5% ? Sur quel type de distribution allez-vous travailler ? Effectuez le calcul.

Ce programme d'intervention mis en place par la principauté et visant à réduire les TOC de tous les patients présentant des TOC à Monaco existe en fait depuis 3 ans. La première année, le nombre de TOC moyen calculé pour la population était de 85.

D'après les nouveaux résultats (cf. la moyenne de l'échantillon), peut-on dire que le programme est devenu plus efficace et réduit mieux le nombre de TOC qu'à ses débuts ?

1. Quels seraient les modèles compact et augmenté pour répondre à cette question?

2. Calculer la proportion de réduction de l'erreur (PRE) ?

3. D'après vous, ce résultat pour la PRE sera-t-il suffisant ?

4. Calculez la valeur du F (et de toutes les valeurs intermédiaires nécessaire pour y parvenir) afin de vérifier si vous aviez raison pour la question précédente?

5. Peut-on dire que le programme est devenu plus efficace et réduit mieux le nombre de TOC qu'à ses débuts ?

6. Pour en être totalement convaincu et vérifier que vos calculs précédents sont corrects, construisez à l'aide de la distribution F (ou t au choix) un intervalle de confiance autour de la valeur du modèle compact. L'intervalle de confiance recouvre-t-il la moyenne de l'échantillon (le paramètre du modèle augmenté) ?

7. Construisez maintenant un intervalle de confiance autour de la valeur du modèle augmenté. L'intervalle de confiance recouvre-t-il la moyenne de l'échantillon (le paramètre du modèle augmenté) ?

8. Quelle approche est à privilégier selon vous ?

9. L'intervalle de confiance est rarement utilisé dans la littérature scientifique. Souvent, on lui préfère une alternative, quelle est-elle ? Veuillez effectuer les calculs adéquats relatifs à cette méthode, toujours en acceptant un risque de 10% de se tromper.

La collègue de notre première psychologue travaille aussi dans le domaine des TOCs, et souhaite pour sa part, déterminer si ses interventions ont eu un impact sur les troubles obsessionnels compulsifs (TOC) de ses patients. En effet, après avoir expérimenté le programme d'intervention utilisé par sa collègue, elle espère idéalement que le nombre de TOC sera réduit à 70 (ou moins) par semaine (ou encore, en moyenne, 10 par jour).
Elle décide d'interroger une dizaine de patients parmi la Înb ntaine de patients qu'elle a suivis depuis ces 5 dernières années.

Voici le nombre de TOC pour ces 10 patients : 35, 70, 35, 40, 105, 105, 100, 40, 35, 105.

1. Que pouvez-vous dire sur la forme de la distribution de cet échantillon ?

2. Suite à votre réponse au point 1, quel type de test statistique allez-vous utiliser ? Paramétrique ou non paramétrique.

3. Quel test statistique allez-vous réaliser ? Justifiez votre réponse.

4. Allez-vous effectuer un test unilatéral ou bilatéral ?

5. Effectuez ce test de manière rigoureuse et commentez-en les différentes étapes.

La psychologue poursuit ses analyses, et a lu un article sur le taux de guérison que l'on peut espérer pour ces personnes. Dans cette étude, sur 200 patients, 44 sont en voie de guérison, et 156 présentent encore trop de TOCs que pour parler déjà de guérison. Lorsqu'elle s'interroge sur la rémission de ses 100 patients, elle obtient la répartition suivante : 16 peuvent être considérés comme guéris, et 84 doivent encore suivre le programme d'intervention pour obtenir des résultats satisfaisants.

1. Quelle est la variable dépendante? Sur quelle échelle se mesure-t-elle? Quelles sont ses valeurs possibles ? Quel test allez-vous donc utiliser ?

2. A quels résultats la psychologue devrait-elle s'attendre si elle pense que la répartition de ses patients suit celle de l'article ? Dressez le tableau des valeurs observées et attendues.

3. Avec un risque = 5%, testez l'adéquation entre les résultats de la psychologue et ceux de l'étude. Utilisez la table adéquate pour conclure.

REFERENCES

Bruner, J. S. (1960). *The Process of Education*. Cambridge Harvard

Coulter, F. (1979). Homework: A neglected research area. *British Journal of Education, 5*, 21-33.

Fiske, D.W. (1949). Consistency of the factorial structures of personality rating from different sources. *Journal of Abnormal Social Psychology, 44*, 329-344.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical Distributions*. New Jersey: Wiley & Sons.

Freedman, D., Pisani, R., & Purves, R. (1997). *Statistics (3rd. edition).* New York: Norton.

Gadeau, L., & Billon-Galland, I. (2003). La demande d'aide auprès des psychologues scolaires. Une enquête relative à 383 signalements scolaires. *Psychologie et Education, 54*, 13-27.

Gauvrit, N. (2005). *Stats pour Psycho.* Bruxelles : De Boek.

Goldberg, J. L., Landau, M. J., Pyszczynski, T., Cox, C., Greenberg, J., Solomon, S., & Dunnam, H. (2003). Gender-Typical responses to sexual and emotional infidelity as a function of morality salience induced self-esteem striving. *Personality and Social Psychology Bulletin, 12*, 1585-1595.

Hacking, I. (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Hacking, I., & Dufour, M. (2004). *L'Ouverture au Probable. Eléments de Logique Inductive.* Paris: Armand Colin.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics, 29*, 205-220.

Howell, D. C. (1999). *Fundamental Statistics for the Behavioral Sciences*. Belmont, CA: Thomson Wadsworth.

Judd, C. M., McClelland, G. H., Ryan, C. S., Muller, D., & Yzerbyt, V. (2010). *Analyse des Données: Une Approche par Comparaison de Modèles*. Bruxelles: De Boeck.

McCrae, R.R., & Costa, P.T. (1987) Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*, 81-90.

Rozin, P., Kabnick, K., Pete, E., Fischler, C., & Shields, C. (2003). The ecology of eating : Smaller portion sizes in France than in the United States help explain the French paradox. *Psychological Science, 14*, 450-454.

Sanders, D. H., Murph, A. F., & Eng, R. J. (1984). *Les Statistiques une Approche Nouvelle*. Montréal: McGraw-Hill.

Schmid Mast, M., & Hall, J. A. (2003). Anybody can be a boss but only certain poeple make good subordinates : Behavioral impacts of striving for dominance and dominance aversion. *Journal of Personality, 71*, 871-892.

Schultz, D. P. (1969). The human subject in psychological research. *Psychological Bulletin, 72*, 214-228.

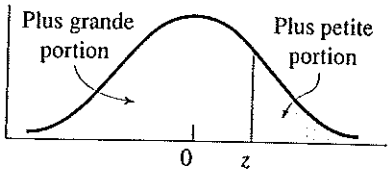
Sheskin, D. J. (2007). *Handbook of Parametric and Nonparametric Statistical Procedures.* Boca Raton: Chapman & Hall/CRC.

Thode, H. C., (2002). *Testing for Normality*. New York: Marcel Dekker.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading MA: Addison-Wiley.

ANNEXES - TABLES

Annexe *z* : la distribution normale (*z*)



z	De la moyenne à *z*	Plus grande portion	Plus petite portion	*y*	*z*	De la moyenne à *z*	Plus grande portion	Plus petite portion	*y*
.00	.0000	.5000	.5000	.3989	.36	.1406	.6406	.3594	.3739
.01	.0040	.5040	.4960	.3989	.37	.1443	.6443	.3557	.3725
.02	.0080	.5080	.4920	.3989	.38	.1480	.6480	.3520	.3712
.03	.0120	.5120	.4880	.3988	.39	.1517	.6517	.3483	.3697
.04	.0160	.5160	.4840	.3986	.40	.1554	.6554	.3446	.3683
.05	.0199	.5199	.4801	.3984	.41	.1591	.6591	.3409	.3668
.06	.0239	.5239	.4761	.3982	.42	.1628	.6628	.3372	.3653
.07	.0279	.5279	.4721	.3980	.43	.1664	.6664	.3336	.3637
.08	.0319	.5319	.4681	.3977	.44	.1700	.6700	.3300	.3621
.09	.0359	.5359	.4641	.3973	.45	.1736	.6736	.3264	.3605
.10	.0398	.5398	.4602	.3970	.46	.1772	.6772	.3228	.3589
.11	.0438	.5438	.4562	.3965	.47	.1808	.6808	.3192	.3572
.12	.0478	.5478	.4522	.3961	.48	.1844	.6844	.3156	.3555
.13	.0517	.5517	.4483	.3956	.49	.1879	.6879	.3121	.3538
.14	.0557	.5557	.4443	.3951	.50	.1915	.6915	.3085	.3521
.15	.0596	.5596	.4404	.3945	.51	.1950	.6950	.3050	.3503
.16	.0636	.5636	.4364	.3939	.52	.1985	.6985	.3015	.3485
.17	.0675	.5675	.4325	.3932	.53	.2019	.7019	.2981	.3467
.18	.0714	.5714	.4286	.3925	.54	.2054	.7054	.2946	.3448
.19	.0753	.5753	.4247	.3918	.55	.2088	.7088	.2912	.3429
.20	.0793	.5793	.4207	.3910	.56	.2123	.7123	.2877	.3410
.21	.0832	.5832	.4168	.3902	.57	.2157	.7157	.2843	.3391
.22	.0871	.5871	.4129	.3894	.58	.2190	.7190	.2810	.3372
.23	.0910	.5910	.4090	.3885	.59	.2224	.7224	.2776	.3352
.24	.0948	.5948	.4052	.3876	.60	.2257	.7257	.2743	.3332
.25	.0987	.5987	.4013	.3867	.61	.2291	.7291	.2709	.3312
.26	.1026	.6026	.3974	.3857	.62	.2324	.7324	.2676	.3292
.27	.1064	.6064	.3936	.3847	.63	.2357	.7357	.2643	.3271
.28	.1103	.6103	.3897	.3836	.64	.2389	.7389	.2611	.3251
.29	.1141	.6141	.3859	.3825	.65	.2422	.7422	.2578	.3230
.30	.1179	.6179	.3821	.3814	.66	.2454	.7454	.2546	.3209
.31	.1217	.6217	.3783	.3802	.67	.2486	.7486	.2514	.3187
.32	.1255	.6255	.3745	.3790	.68	.2517	.7517	.2483	.3166
.33	.1293	.6293	.3707	.3778	.69	.2549	.7549	.2451	.3144
.34	.1331	.6331	.3669	.3765	.70	.2580	.7580	.2420	.3123
.35	.1368	.6368	.3632	.3752	.71	.2611	.7611	.2389	.3101

Source : Howell, 1999

ANNEXE z (Suite)

z	De la moyenne à z	Plus grande portion	Plus petite portion	y	z	De la moyenne à z	Plus grande portion	Plus petite portion	y
.72	.2642	.7642	.2358	.3079	1.17	.3790	.8790	.1210	.2012
.73	.2673	.7673	.2327	.3056	1.18	.3810	.8810	.1190	.1989
.74	.2704	.7704	.2296	.3034	1.19	.3830	.8830	.1170	.1965
.75	.2734	.7734	.2266	.3011	1.20	.3849	.8849	.1151	.1942
.76	.2764	.7764	.2236	.2989	1.21	.3869	.8869	.1131	.1919
.77	.2794	.7794	.2206	.2966	1.22	.3888	.8888	.1112	.1895
.78	.2823	.7823	.2177	.2943	1.23	.3907	.8907	.1093	.1872
.79	.2852	.7852	.2148	.2920	1.24	.3925	.8925	.1075	.1849
.80	.2881	.7881	.2119	.2897	1.25	.3944	.8944	.1056	.1826
.81	.2910	.7910	.2090	.2874	1.26	.3962	.8962	.1038	.1804
.82	.2939	.7939	.2061	.2850	1.27	.3980	.8980	.1020	.1781
.83	.2967	.7967	.2033	.2827	1.28	.3997	.8997	.1003	.1758
.84	.2995	.7995	.2005	.2803	1.29	.4015	.9015	.0985	.1736
.85	.3023	.8023	.1977	.2780	1.30	.4032	.9032	.0968	.1714
.86	.3051	.8051	.1949	.2756	1.31	.4049	.9049	.0951	.1691
.87	.3078	.8078	.1922	.2732	1.32	.4066	.9066	.0934	.1669
.88	.3106	.8106	.1894	.2709	1.33	.4082	.9082	.0918	.1647
.89	.3133	.8133	.1867	.2685	1.34	.4099	.9099	.0901	.1626
.90	.3159	.8159	.1841	.2661	1.35	.4115	.9115	.0885	.1604
.91	.3186	.8186	.1814	.2637	1.36	.4131	.9131	.0869	.1582
.92	.3212	.8212	.1788	.2613	1.37	.4147	.9147	.0853	.1561
.93	.3238	.8238	.1762	.2589	1.38	.4162	.9162	.0838	.1539
.94	.3264	.8264	.1736	.2565	1.39	.4177	.9177	.0823	.1518
.95	.3289	.8289	.1711	.2541	1.40	.4192	.9192	.0808	.1497
.96	.3315	.8315	.1685	.2516	1.41	.4207	.9207	.0793	.1476
.97	.3340	.8340	.1660	.2492	1.42	.4222	.9222	.0778	.1456
.98	.3365	.8365	.1635	.2468	1.43	.4236	.9236	.0764	.1435
.99	.3389	.8389	.1611	.2444	1.44	.4251	.9251	.0749	.1415
1.00	.3413	.8413	.1587	.2420	1.45	.4265	.9265	.0735	.1394
1.01	.3438	.8438	.1562	.2396	1.46	.4279	.9279	.0721	.1374
1.02	.3461	.8461	.1539	.2371	1.47	.4292	.9292	.0708	.1354
1.03	.3485	.8485	.1515	.2347	1.48	.4306	.9306	.0694	.1334
1.04	.3508	.8508	.1492	.2323	1.49	.4319	.9319	.0681	.1315
1.05	.3531	.8531	.1469	.2299	1.50	.4332	.9332	.0668	.1295
1.06	.3554	.8554	.1446	.2275	1.51	.4345	.9345	.0655	.1276
1.07	.3577	.8577	.1423	.2251	1.52	.4357	.9357	.0643	.1257
1.08	.3599	.8599	.1401	.2227	1.53	.4370	.9370	.0630	.1238
1.09	.3621	.8621	.1379	.2203	1.54	.4382	.9382	.0618	.1219
1.10	.3643	.8643	.1357	.2179	1.55	.4394	.9394	.0606	.1200
1.11	.3665	.8665	.1335	.2155	1.56	.4406	.9406	.0594	.1182
1.12	.3686	.8686	.1314	.2131	1.57	.4418	.9418	.0582	.1163
1.13	.3708	.8708	.1292	.2107	1.58	.4429	.9429	.0571	.1145
1.14	.3729	.8729	.1271	.2083	1.59	.4441	.9441	.0559	.1127
1.15	.3749	.8749	.1251	.2059	1.60	.4452	.9452	.0548	.1109
1.16	.3770	.8770	.1230	.2036	1.61	.4463	.9463	.0537	.1092

Source : Howell, 1999

ANNEXE z (Suite)

z	De la moyenne à z	Plus grande portion	Plus petite portion	y	z	De la moyenne à z	Plus grande portion	Plus petite portion	y
1.62	.4474	.9474	.0526	.1074	2.07	.4808	.9808	.0192	.0468
1.63	.4484	.9484	.0516	.1057	2.08	.4812	.9812	.0188	.0459
1.64	.4495	.9495	.0505	.1040	2.09	.4817	.9817	.0183	.0449
1.65	.4505	.9505	.0495	.1023	2.10	.4821	.9821	.0179	.0440
1.66	.4515	.9515	.0485	.1006	2.11	.4826	.9826	.0174	.0431
1.67	.4525	.9525	.0475	.0989	2.12	.4830	.9830	.0170	.0422
1.68	.4535	.9535	.0465	.0973	2.13	.4834	.9834	.0166	.0413
1.69	.4545	.9545	.0455	.0957	2.14	.4838	.9838	.0162	.0404
1.70	.4554	.9554	.0446	.0940	2.15	.4842	.9842	.0158	.0396
1.71	.4564	.9564	.0436	.0925	2.16	.4846	.9846	.0154	.0387
1.72	.4573	.9573	.0427	.0909	2.17	.4850	.9850	.0150	.0379
1.73	.4582	.9582	.0418	.0893	2.18	.4854	.9854	.0146	.0371
1.74	.4591	.9591	.0409	.0878	2.19	.4857	.9857	.0143	.0363
1.75	.4599	.9599	.0401	.0863	2.20	.4861	.9861	.0139	.0355
1.76	.4608	.9608	.0392	.0848	2.21	.4864	.9864	.0136	.0347
1.77	.4616	.9616	.0384	.0833	2.22	.4868	.9868	.0132	.0339
1.78	.4625	.9625	.0375	.0818	2.23	.4871	.9871	.0129	.0332
1.79	.4633	.9633	.0367	.0804	2.24	.4875	.9875	.0125	.0325
1.80	.4641	.9641	.0359	.0790	2.25	.4878	.9878	.0122	.0317
1.81	.4649	.9649	.0351	.0775	2.26	.4881	.9881	.0119	.0310
1.82	.4656	.9656	.0344	.0761	2.27	.4884	.9884	.0116	.0303
1.83	.4664	.9664	.0336	.0748	2.28	.4887	.9887	.0113	.0297
1.84	.4671	.9671	.0329	.0734	2.29	.4890	.9890	.0110	.0290
1.85	.4678	.9678	.0322	.0721	2.30	.4893	.9893	.0107	.0283
1.86	.4686	.9686	.0314	.0707	2.31	.4896	.9896	.0104	.0277
1.87	.4693	.9693	.0307	.0694	2.32	.4898	.9898	.0102	.0270
1.88	.4699	.9699	.0301	.0681	2.33	.4901	.9901	.0099	.0264
1.89	.4706	.9706	.0294	.0669	2.34	.4904	.9904	.0096	.0258
1.90	.4713	.9713	.0287	.0656	2.35	.4906	.9906	.0094	.0252
1.91	.4719	.9719	.0281	.0644	2.36	.4909	.9909	.0091	.0246
1.92	.4726	.9726	.0274	.0632	2.37	.4911	.9911	.0089	.0241
1.93	.4732	.9732	.0268	.0620	2.38	.4913	.9913	.0087	.0235
1.94	.4738	.9738	.0262	.0608	2.39	.4916	.9916	.0084	.0229
1.95	.4744	.9744	.0256	.0596	2.40	.4918	.9918	.0082	.0224
1.96	.4750	.9750	.0250	.0584	2.41	.4920	.9920	.0080	.0219
1.97	.4756	.9756	.0244	.0573	2.42	.4922	.9922	.0078	.0213
1.98	.4761	.9761	.0239	.0562	2.43	.4925	.9925	.0075	.0208
1.99	.4767	.9767	.0233	.0551	2.44	.4927	.9927	.0073	.0203
2.00	.4772	.9772	.0228	.0540	2.45	.4929	.9929	.0071	.0198
2.01	.4778	.9778	.0222	.0529	2.46	.4931	.9931	.0069	.0194
2.02	.4783	.9783	.0217	.0519	2.47	.4932	.9932	.0068	.0189
2.03	.4788	.9788	.0212	.0508	2.48	.4934	.9934	.0066	.0184
2.04	.4793	.9793	.0207	.0498	2.49	.4936	.9936	.0064	.0180
2.05	.4798	.9798	.0202	.0488	2.50	.4938	.9938	.0062	.0175
2.06	.4803	.9803	.0197	.0478	2.51	.4940	.9940	.0060	.0171

Source : Howell, 1999

ANNEXE z (Suite)

z	De la moyenne à z	Plus grande portion	Plus petite portion	y	z	De la moyenne à z	Plus grande portion	Plus petite portion	y
2.52	.4941	.9941	.0059	.0167	2.81	.4975	.9975	.0025	.0077
2.53	.4943	.9943	.0057	.0163	2.82	.4976	.9976	.0024	.0075
2.54	.4945	.9945	.0055	.0158	2.83	.4977	.9977	.0023	.0073
2.55	.4946	.9946	.0054	.0154	2.84	.4977	.9977	.0023	.0071
2.56	.4948	.9948	.0052	.0151	2.85	.4978	.9978	.0022	.0069
2.57	.4949	.9949	.0051	.0147	2.86	.4979	.9979	.0021	.0067
2.58	.4951	.9951	.0049	.0143	2.87	.4979	.9979	.0021	.0065
2.59	.4952	.9952	.0048	.0139	2.88	.4980	.9980	.0020	.0063
2.60	.4953	.9953	.0047	.0136	2.89	.4981	.9981	.0019	.0061
2.61	.4955	.9955	.0045	.0132	2.90	.4981	.9981	.0019	.0060
2.62	.4956	.9956	.0044	.0129	2.91	.4982	.9982	.0018	.0058
2.63	.4957	.9957	.0043	.0126	2.92	.4982	.9982	.0018	.0056
2.64	.4959	.9959	.0041	.0122	2.93	.4983	.9983	.0017	.0055
2.65	.4960	.9960	.0040	.0119	2.94	.4984	.9984	.0016	.0053
2.66	.4961	.9961	.0039	.0116	2.95	.4984	.9984	.0016	.0051
2.67	.4962	.9962	.0038	.0113	2.96	.4985	.9985	.0015	.0050
2.68	.4963	.9963	.0037	.0110	2.97	.4985	.9985	.0015	.0048
2.69	.4964	.9964	.0036	.0107	2.98	.4986	.9986	.0014	.0047
2.70	.4965	.9965	.0035	.0104	2.99	.4986	.9986	.0014	.0046
2.71	.4966	.9966	.0034	.0101	3.00	.4987	.9987	.0013	.0044
2.72	.4967	.9967	.0033	.0099	...	...	...	...	...
2.73	.4968	.9968	.0032	.0096	3.25	.4994	.9994	.0006	.0020
2.74	.4969	.9969	.0031	.0093	...	...	...	...	...
2.75	.4970	.9970	.0030	.0091	3.50	.4998	.9998	.0002	.0009
2.76	.4971	.9971	.0029	.0088	...	...	...	...	...
2.77	.4972	.9972	.0028	.0086	3.75	.4999	.9999	.0001	.0004
2.78	.4973	.9973	.0027	.0084	...	...	...	...	...
2.79	.4974	.9974	.0026	.0081	4.00	.5000	1.0000	.0000	.0001
2.80	.4974	.9974	.0026	.0079					

Source : Howell, 1999

Annexe *F* : valeurs critiques de la distribution *F*

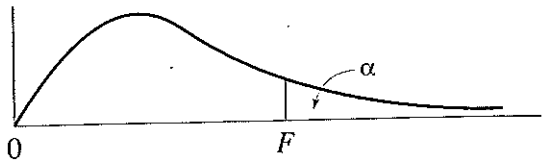
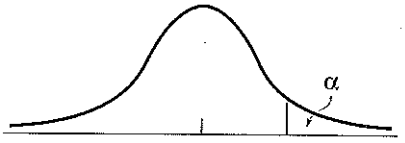


TABLE 1 $\alpha = .05$

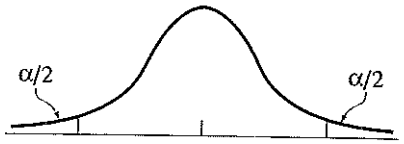
Degrés de liberté pour le dénominateur	Degrés de liberté pour le numérateur															
	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50
1	161.4	199.5	215.8	224.8	230.0	233.8	236.5	238.6	240.1	242.1	245.2	248.4	248.9	250.5	250.8	252.6
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.44	19.46	19.47	19.48	19.48
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.97	1.94	1.89	1.86
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.94	1.90	1.85	1.82
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.79
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.76
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.66
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87	1.78	1.73	1.69	1.63	1.60
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.69	1.65	1.59	1.56
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.75	1.66	1.60	1.55	1.50	1.46
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.72	1.62	1.56	1.52	1.46	1.41
500	3.86	3.01	2.62	2.39	2.23	2.12	2.03	1.96	1.90	1.85	1.69	1.59	1.53	1.48	1.42	1.38
1000	3.85	3.01	2.61	2.38	2.22	2.11	2.02	1.95	1.89	1.84	1.68	1.58	1.52	1.47	1.41	1.36

Source : Howell, 1999

Annexe *t* : points de pourcentage supérieurs de la distribution *t*



Test unilatéral



Test bilatéral

	Seuil de signification pour le test unilatéral								
	.25	.20	.15	.10	.05	.025	.01	.005	.0005
	Seuil de signification pour le test bilatéral								
dl	.50	.40	.30	.20	.10	.05	.02	.01	.001
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657	636.620
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.403	2.678	3.496
100	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.390
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Source : Howell, 1999

Table A5 Table of Critical *T* Values for Wilcoxon's Signed-Ranks and Matched-Pairs Signed-Ranks Test

	One-tailed level of significance					One-tailed level of significance			
	.05	.025	.01	.005		.05	.025	.01	.005
	Two-tailed level of significance					Two-tailed level of significance			
	.10	.05	.02	.01		.10	.05	.02	.01
n					*n*				
5	0	–	–	–	28	130	116	101	91
6	2	0	–	–	29	140	126	110	100
7	3	2	0	–	30	151	137	120	109
8	5	3	1	0	31	163	147	130	118
9	8	5	3	1	32	175	159	140	128
10	10	8	5	3	33	187	170	151	138
11	13	10	7	5	34	200	182	162	148
12	17	13	9	7	35	213	195	173	159
13	21	17	12	9	36	227	208	185	171
14	25	21	15	12	37	241	221	198	182
15	30	25	19	15	38	256	235	211	194
16	35	29	23	19	39	271	249	224	207
17	41	34	27	23	40	286	264	238	220
18	47	40	32	27	41	302	279	252	233
19	53	46	37	32	42	319	294	266	247
20	60	52	43	37	43	336	310	281	261
21	67	58	49	42	44	353	327	296	276
22	75	65	55	48	45	371	343	312	291
23	83	73	62	54	46	389	361	328	307
24	91	81	69	61	47	407	378	345	322
25	100	89	76	68	48	426	396	362	339
26	110	98	84	75	49	446	415	379	355
27	119	107	92	83	50	466	434	397	373

Source : Sheskin, 2007

Table A6 Table of the Binomial Distribution, Individual Probabilities

n	x	π .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
1	0	.9500	.9000	.8500	.8000	.7500	.7000	.6500	.6000	.5500	.5000
	1	.0500	.1000	.1500	.2000	.2500	.3000	.3500	.4000	.4500	.5000
2	0	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0135	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1	.2036	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1562
	2	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0000	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0312
	2	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039

Source : Sheskin, 2007

Table A6 Table of the Binomial Distribution, Individual Probabilities (continued)

n	x	π = .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
9	0	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.0629	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010
11	0	.5688	.3138	.1673	.0859	.0422	.0198	.0088	.0036	.0014	.0004
	1	.3293	.3835	.3248	.2362	.1549	.0932	.0518	.0266	.0125	.0055
	2	.0867	.2131	.2866	.2953	.2581	.1998	.1395	.0887	.0513	.0269
	3	.0137	.0710	.1517	.2215	.2581	.2568	.2254	.1774	.1259	.0806
	4	.0014	.0158	.0536	.1107	.1721	.2201	.2428	.2365	.2060	.1611
	5	.0001	.0025	.0132	.0388	.0803	.1321	.1830	.2207	.2360	.2256
	6	.0000	.0003	.0023	.0097	.0268	.0566	.0985	.1471	.1931	.2256
	7	.0000	.0000	.0003	.0017	.0064	.0173	.0379	.0701	.1128	.1611
	8	.0000	.0000	.0000	.0002	.0011	.0037	.0102	.0234	.0462	.0806
	9	.0000	.0000	.0000	.0000	.0001	.0005	.0018	.0052	.0126	.0269
	10	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0007	.0021	.0054
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0005
12	0	.5404	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3413	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.0988	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0173	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0021	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0002	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0000	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7	.0000	.0000	.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8	.0000	.0000	.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9	.0000	.0000	.0000	.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10	.0000	.0000	.0000	.0000	.0000	.0002	.0008	.0025	.0068	.0161
	11	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0029
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002

Source : Sheskin, 2007

Table A6 Table of the Binomial Distribution, Individual Probabilities (continued)

n	x	π = .05	.10	.15	.20	.25	.30	.35	.40	.45	.50
13	0	.5133	.2542	.1209	.0550	.0238	.0097	.0037	.0013	.0004	.0001
	1	.3512	.3672	.2774	.1787	.1029	.0540	.0259	.0113	.0045	.0016
	2	.1109	.2448	.2937	.2680	.2059	.1388	.0836	.0453	.0220	.0095
	3	.0214	.0997	.1900	.2457	.2517	.2181	.1651	.1107	.0660	.0349
	4	.0028	.0277	.0838	.1535	.2097	.2337	.2222	.1845	.1350	.0873
	5	.0003	.0055	.0266	.0691	.1258	.1803	.2154	.2214	.1989	.1571
	6	.0000	.0008	.0063	.0230	.0559	.1030	.1546	.1968	.2169	.2095
	7	.0000	.0001	.0011	.0058	.0186	.0442	.0833	.1312	.1775	.2095
	8	.0000	.0000	.0001	.0011	.0047	.0142	.0336	.0656	.1089	.1571
	9	.0000	.0000	.0000	.0001	.0009	.0034	.0101	.0243	.0495	.0873
	10	.0000	.0000	.0000	.0000	.0001	.0006	.0022	.0065	.0162	.0349
	11	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0012	.0036	.0095
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
14	0	.4877	.2288	.1028	.0440	.0178	.0068	.0024	.0008	.0002	.0001
	1	.3593	.3559	.2539	.1539	.0832	.0407	.0181	.0073	.0027	.0009
	2	.1229	.2570	.2912	.2501	.1802	.1134	.0634	.0317	.0141	.0056
	3	.0259	.1142	.2056	.2501	.2402	.1943	.1366	.0845	.0462	.0222
	4	.0037	.0349	.0998	.1720	.2202	.2290	.2022	.1549	.1040	.0611
	5	.0004	.0078	.0352	.0860	.1468	.1963	.2178	.2066	.1701	.1222
	6	.0000	.0013	.0093	.0322	.0734	.1262	.1759	.2066	.2088	.1833
	7	.0000	.0002	.0019	.0092	.0280	.0618	.1082	.1574	.1952	.2095
	8	.0000	.0000	.0003	.0020	.0082	.0232	.0510	.0918	.1398	.1833
	9	.0000	.0000	.0000	.0003	.0018	.0066	.0183	.0408	.0762	.1222
	10	.0000	.0000	.0000	.0000	.0003	.0014	.0049	.0136	.0312	.0611
	11	.0000	.0000	.0000	.0000	.0000	.0002	.0010	.0033	.0093	.0222
	12	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0019	.0056
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0009
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
15	0	.4633	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3658	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.1348	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.0307	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0049	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417
	5	.0006	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0000	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0000	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8	.0000	.0000	.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9	.0000	.0000	.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10	.0000	.0000	.0000	.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11	.0000	.0000	.0000	.0000	.0001	.0006	.0024	.0074	.0191	.0417
	12	.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0016	.0052	.0139
	13	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010	.0032
	14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000

Source : Sheskin, 2007