Université de Liège

Faculté de Médecine

BIOSTATISTIQUE

Adelin Albert
Professeur ordinaire

Edition 2005

Préface

La statistique joue un rôle central dans la recherche et les activités des sciences de la Santé : médecine, dentisterie, pharmacie, sciences biomédicales, kinésithérapie, sciences de la motricité et Santé publique mais aussi dans d'autres domaines des sciences (biologie, biochimie, géographie, agronomie, psychologie, sociologie, économie). Qu'il s'agisse d'un plan d'expérience, d'une enquête de santé, d'un essai clinique randomisé, ou d'une étude observationnelle, on a recours aux méthodes de la biostatistique moderne pour analyser et interpréter les résultats obtenus. Il est d'ailleurs conseillé de consulter un statisticien avant d'entreprendre un travail et non pas d'attendre que l'on ait récolté l'ensemble des données. En effet, un plan d'expérience bien conçu, une enquête de santé correctement planifiée, un essai clinique clairement élaboré, ou un étude observationnelle aux objectifs précis permet d'économiser des ressources, d'améliorer l'efficacité statistique, et d'éviter des biais éventuels dans les conclusions. Par ailleurs, depuis quelques décennies, les grandes revues scientifiques internationales se sont adjoint les services de statisticiens qui veillent à l'utilisation correcte des tests statistiques par les auteurs des articles. Enfin, les universités, en particulier les facultés de médecine, ont créé des centres ou départements de biostatistique destinés à former les étudiants, encadrer les chercheurs et assurer la consultation statistique en général. La biostatistique trouve donc naturellement sa place dans le curriculum des étudiants de médecine, dentisterie, pharmacie, sciences biomédicales, kinésithérapie, sciences de la motricité et Santé publique, mais aussi dans la formation complémentaire des chercheurs de ces mêmes disciplines.

Cet ouvrage peut être considéré comme une introduction avancée à la biostatistique. Il se décompose en 18 chapitres que l'on peut regrouper en trois parties : statistique descriptive (Chapitres 1 à 7), échantillonnage et probabilité (Chapitres 8 à 13), et statistique inférentielle (Chapitres 14 à 18). Le Chapitre 1 introduit les notions de base, en particulier celles de population, échantillon, variables et données. Les méthodes de présentation graphique des données sont présentées au Chapitre 2. Les deux concepts fondamentaux de moyenne et écart-type sont introduits au Chapitre 3 ainsi que quelques-unes de leurs applications médicales. D'autres paramètres de position et de dispersion, en particulier les percentiles et l'écart interquartiles, sont introduits au Chapitre 4. Le Chapitre 5 est consacré à un domaine de la plus haute actualité, celui des courbes de survie de Kaplan-Meier dont la littérature scientifique médicale abonde. Les Chapitres 6 et 7 envisagent l'association entre deux variables, selon qu'elles sont observées simultanément (concept de corrélation) ou qu'une des deux variables est contrôlée par l'expérimentateur (problème de régression).

Les Chapitres 8 à 13 constituent la transition entre la statistique descriptive et la statistique inférentielle. Ainsi, au Chapitre 8, on introduit le concept d'erreur type d'une estimation statistique, qui joue un rôle fondamental dans l'interprétation des résultats, mais dont la compréhension n'est pas aisée. Les intervalles de confiance, fréquemment sollicités dans les publications scientifiques, sont définis au Chapitre 9. Les Chapitres 10 et 11 abordent de façon simple les concepts de probabilité et de probabilité conditionnelle. On évoque au Chapitre 11 le célèbre théorème de Bayes, dont l'importance dans l'aide à la décision médicale n'est plus à démontrer. Enfin, les deux chapitres suivants sont consacrés aux variables aléatoires discrètes (Chapitre 12) et aux variables aléatoires continues (Chapitre 13). En particulier, on y présente succinctement quelques lois de probabilité remarquables, à savoir la loi Binomiale et la loi de Poisson, les lois Normale, Chi-carré, t de Student et F de Snedecor, grâce auxquelles on peut résoudre la grande majorité des problèmes statistiques.

Les Chapitres 14 à 18 couvrent le domaine de l'inférence statistique, c'est-à-dire des tests d'hypothèses. Au Chapitre 14, on présente la structure générale des tests d'hypothèses qui sera utilisée dans les chapitres suivants. Les tests sur les corrélations sont décrits au Chapitre 15. Le Chapitre 16 aborde l'important problème de l'analyse des tables de comptage, appelées aussi tables de contingence. Enfin, les tests de comparaison de moyennes sont décrits en détail au Chapitre 17 dans le cas d'échantillons indépendants et au Chapitre 18 dans le cas d'échantillons appariés.

Les annexes reprennent quelques-uns des fichiers de données qui ont permis d'illustrer les différents chapitres. Enfin, l'ouvrage se termine par la présentation de sept tables statistiques (Tables A à G), indispensables pour interpréter les tests statistiques. Il s'agit des tables de la loi Normale Z, Chi-carré, t de Student, F de Snedecor, test de corrélation, test de Mann-Whitney, et test des rangs signés de Wilcoxon.

La majorité des exemples qui ont servi à illustrer les méthodes présentées dans cet ouvrage sont des applications réelles prélevées dans les nombreux projets conduits avec les cliniciens et chercheurs de la Faculté de Médecine. J'ai pris la liberté de remplacer la virgule décimale par le point décimal afin d'améliorer la clarté de présentation des résultats, notamment dans les tables ou dans l'expression des quantiles. Enfin, il m'a paru utile de donner la traduction anglaise des principaux termes utilisés en statistique, la littérature scientifique étant essentiellement anglo-saxonne.

Qu'il me soit permis de remercier outre les nombreux Collègues, cliniciens, chercheurs et étudiants avec lesquels j'ai eu la chance de travailler, mes collaborateurs directs Paul Gérard, Laurence Seidel, Walthère Dewé, Corinne Jamoul, Bernard Vrijens, Frank Jeusette, Daniel Gillain, Liying Zhang, Gisèle Mersch †, Laetitia Comté, David Magis, Anne-Françoise Donneau et Sophie Vanbelle, tous statisticiens, Anna Marchetta, ma secrétaire, ainsi que Nicole Dumont et Danielle Bartholoméus qui se sont chargées de la dactylographie finale du travail. Enfin, une reconnaissance particulière s'adresse à Attilio Ceccato, docteur en sciences pharmaceutiques, dont les notes claires prises à mon cours ont servi durant plusieurs années de support aux étudiants.

Février 2000, revu en juillet 2005

Adelin Albert

Chapitre 1

Notions de base

1.1 Définition de la statistique

Selon Ronald Aylmer Fisher (1890-1962), considéré comme le plus célèbre statisticien du 20e siècle, la statistique est la discipline qui étudie

- les méthodes de réduction de données
- la variabilité
- les populations.

La biostatistique est l'application de la statistique aux sciences de la vie.

1.1.1 Les méthodes de réduction de données

Dans la pratique quotidienne, la statistique est avant tout utilisée pour résumer sous forme claire et synthétique un ensemble de données. C'est le domaine couvert par la *statistique descriptive* ou exploratoire (descriptive or exploratory statistics). On parle aussi d'*analyse de données* (data analysis). Qu'il s'agisse d'une enquête de santé, d'un sondage d'opinion, d'une étude d'observation, d'une étude expérimentale, ou d'un essai clinique, la première mission du statisticien est de présenter les résultats de façon compréhensible pour le lecteur, soit à l'aide de graphiques, soit au moyen de paramètres ou d'indicateurs simples.

Le mathématicien et physicien belge Adolphe Quetelet (1796-1874) fut l'un des premiers à reconnaître la nécessité et l'importance des méthodes de statistique descriptive dans les problèmes de recensement de populations et les données d'Etat.

1.1.2 La variabilité

Si la variabilité n'existait pas, il n'y aurait pas de statistique. Dans la vie, les choses sont éminemment variables. Les étudiants d'une même classe varient en sexe, taille, poids, groupe sanguin, pression artérielle, cholestérol sanguin et selon une multitude d'autres caractéristiques. Mieux, au cours d'une même journée, la pression artérielle d'un individu peut varier de façon considérable. Il en est de même du taux de cholestérol au

cours du temps. Le nombre d'enfants varie d'un ménage à l'autre. Le taux de pollution urbain diffère d'un quartier à l'autre, à l'intérieur d'une même ville, d'une ville à l'autre, d'un pays à l'autre. Le taux de réussite en première candidature varie d'une année académique à l'autre, mais aussi selon les facultés ou les universités. En répétant plusieurs fois le dosage du glucose sur un même tube de sang, on ne trouve pas toujours la même concentration. On est confronté à la variabilité analytique. La durée d'hospitalisation n'est pas la même pour chaque patient, ni la gravité de sa maladie, ni sa durée de vie. La variabilité est donc inévitable et la statistique a pour objectif d'expliquer en partie cette variabilité, de la caractériser par des constantes. Dans la suite de ce cours, nous distinguerons la variabilité d'échantillon (variabilité analytique, expérimentale, biologique) de la variabilité due à l'échantillonnage, c'est-à-dire au mécanisme aléatoire de prélèvement des échantillons d'une population.

1.1.3 L'inférence statistique

La statistique ne se contente pas de résumer des informations ou des données recueillies au cours d'une étude, d'une enquête ou d'une expérience. Elle permet en outre de tirer des conclusions sur l'ensemble de la population d'où les données sont extraites. En ce sens, la statistique est un puissant outil de recherche, moteur de l'inférence scientifique. C'est le domaine de la *statistique inférentielle* (inferential statistics) comportant l'ensemble des tests d'hypothèses. Le sondage d'opinion constitue un bel exemple de la statistique inférentielle, où, à partir des réponses de quelques centaines ou d'un millier de personnes, on tire des conclusions précises sur l'ensemble d'une population qui en comporte plusieurs millions. En fait, la statistique "peut faire beaucoup à partir de peu" ! Si la statistique existe, c'est aussi parce qu'en général les populations sont tellement grandes qu'il est impossible d'analyser chaque individu en particulier et qu'il faut en conséquence recourir à un échantillon de taille plus petite.

1.2 Population et échantillon

Au préalable de toute étude statistique, il convient de définir clairement la population à laquelle on s'intéresse et de laquelle on tire un échantillon.

1.2.1 Population

Une population est une collection ou ensemble d'individus (ou d'objets) ayant au moins une caractéristique en commun. On désigne par N l'effectif (size) de la population. En général, N est tellement grand qu'on l'assimile à l'infini ($N \simeq \infty$).

Citons quelques exemples de populations.

- Les étudiants de l'Université de Liège
- Les médecins généralistes wallons
- Les pharmaciens de la ville de Liège
- Les patients du Centre Hospitalier Universitaire (CHU) de Liège

– Les malades atteints d'un cancer du poumon
– Les valves cardiaques
– Les hôpitaux belges
– Les êtres humains
– Les médicaments prescrits aux personnes âgées

Tant que la population à laquelle on s'intéresse n'a pas été clairement définie, il subsistera un problème pour l'analyse statistique. Notons enfin que l'*unité statistique* (statistical unit) est l'élément (sujet ou objet) de la population que l'on observe ou sur lequel on effectue les mesures.

1.2.2 Echantillon

Un échantillon (sample) est un sous-ensemble (partie) de la population étudiée. On désigne par n l'effectif (sample size) de l'échantillon. Par définition, n est toujours fini sinon on ne pourrait analyser l'échantillon dans son ensemble. La façon de tirer un échantillon d'une population est un exercice difficile et délicat (voir théorie de l'échantillonnage). L'objectif poursuivi est que l'échantillon soit "représentatif" de la population et que les conclusions tirées sur la population à partir de l'échantillon soient non biaisées (unbiased).

A titre d'exemples, on peut étudier un échantillon de 200 étudiants de l'Université de Liège (au lieu des 14.000 existants), 355 médecins généralistes wallons, 20 pharmaciens de Liège, 1000 patients du CHU, 73 patients atteints d'un cancer du poumon, 60 valves cardiaques, 150 hôpitaux belges, 2000 êtres humains, 30 étudiants de 1ère candidature en médecine, 10.000 prescriptions de médicaments aux personnes âgées.

1.3 Variables

1.3.1 Définition

En statistique, les variables sont les caractéristiques (critères ou facteurs) des éléments de la population auxquelles on s'intéresse. Comme pour la population, il est essentiel de définir clairement les variables du problème statistique.

Si l'on ne s'intéresse qu'à une seule variable, que l'on note X, on parle de *statistique univariée* (univariate statistics) ; par contre, s'il y en a plusieurs, on dit que l'on a affaire à un problème de *statistique multivariée* (multivariate statistics). On note alors $X_1, \ldots, X_p$ les p variables.

Citons à titre d'illustration pour les populations ci-dessus (voir 1.2.1) :
– La réussite ou non de l'étudiant en fin d'année
– La patientèle (nombre de patients) du médecin généraliste
– L'âge du pharmacien ou de la pharmacienne
– Le stade du cancer du poumon au moment du diagnostic
– La longévité de la valve cardiaque
– Le nombre d'admissions par an de l'hôpital

– La durée de vie de l'être humain
– Le nombre de médicaments prescrits à la personne âgée

On distingue différents *types de variables* en statistique. Il est essentiel de préciser ce type pour chaque variable étudiée car les méthodes statistiques utilisées en dépendent.

1.3.2 Variables qualitatives

Une variable est *qualitative* (qualitative variable) lorsqu'elle exprime une qualité. Les valeurs prises par une variable qualitative sont appelées des "modalités" qui portent des noms. C'est la raison pour laquelle, on parle aussi de variable "nominale". Une variable qualitative s'observe, on ne la mesure pas.

Citons à titre d'exemples :
– l'état-civil (célibataire, marié, veuf, divorcé, séparé) qui comporte 5 modalités
– le groupe sanguin (A, B, AB, O) : 4 modalités
– la localisation d'un infarctus (antérieur, postérieur, inférieur)
– l'université d'origine du pharmacien (ULg, UCL, ULB)
– la race d'un être humain (blanc, noir, jaune, rouge)
– le sexe (homme, femme)

Si on désigne par q le nombre de modalités d'une variable qualitative et par $\{m_1, \ldots, m_q\}$ les q modalités, on peut numéroter celles-ci de 1 à q mais ces nombres n'ont aucune valeur numérique. En d'autres termes, il est interdit d'y faire des calculs ! Cette numérotation est purement arbitraire et peut varier d'un expérimentateur à l'autre.

Variables ordinales

Il arrive qu'il existe une relation d'ordre sur les modalités de la variable qualitative, soit $m_1 < m_2 < \cdots < m_q$. Dans ce cas, on parle de variable *ordinale* (ordinal variable). Citons à titre d'exemples :
– le grade aux examens (Aj, S, D, GD, PGD)
– le stade d'un cancer (I, II, III, IV)
– la pratique d'un sport (jamais, rarement, occasionnel, fréquent, quotidien)
– l'état d'un patient (détérioration, condition stable, amélioration)
– l'issue d'un patient traumatisé à 6 mois selon la "Glasgow Outcome Scale (GOS)" (1 = bonne récupération, 2 = incapacité légère, 3 = incapacité sévère, 4 = état végétatif persistant, 5 = décès).

Dans le cas d'une variable ordinale, il n'est pas totalement erroné de numéroter les modalités et d'y effectuer des calculs comme sur des variables discrètes.

Variables catégorisées

Dans les questionnaires, il arrive fréquemment que des variables quantitatives soient catégorisées pour des raisons de facilité de réponse. Par exemple, la patientèle d'un médecin est classée en quatre groupes : 0-30, 31-60, 61-90, plus de 90 patients/semaine,

et il suffit de choisir la catégorie qui convient. On parle alors de variable *catégorisée* (categorical variable), dont le traitement statistique n'est pas toujours aisé.

1.3.3 Variables quantitatives

Par opposition aux variables qualitatives, une variable *quantitative* (quantitative variable) exprime une quantité. Les valeurs sont généralement numériques, résultat d'une mesure. On parle dès lors de variables "numériques" ou "mesurables". Ce sont les variables les plus fréquentes dans les sciences de la Santé et leur traitement statistique est plus aisé que celui des variables qualitatives.

Il convient à ce stade de distinguer les variables discrètes des variables continues.

Variables discrètes

Une variable quantitative est *discrète* (discrete variable) lorsqu'elle ne prend qu'un nombre fini (k) de valeurs. On parle aussi de variable de comptage (count variable). Citons à titre d'exemples :

- le nombre d'enfants dans un ménage (0, 1, 2, ... k)
- le nombre de garçons dans une famille de quatre enfants (0, 1, 2, 3, 4)
- le nombre d'hospitalisations d'un patient (0, 1, 2,...)
- le nombre de médicaments différents prescrits à un malade
- le nombre d'implants dentaires
- le nombre de cas de tuberculose par province

Comme il s'agit de nombres, on peut y faire tous les calculs souhaités.

Variables continues

Une variable quantitative *continue* (continuous variable) prend toutes les valeurs possibles dans un intervalle (ou continuum) donné. Le nombre de valeurs possibles est donc infini (voire infiniment non dénombrable). Les exemples sont nombreux car il s'agit de la majorité des variables étudiées :

- le poids d'un sujet
- le taux de cholestérol sanguin
- la durée de vie d'une valve cardiaque
- le taux de pollution urbain
- la pression artérielle systolique.

En pratique, une variable continue peut apparaître comme discrète car on est limité par la précision de l'appareil de mesure. Néanmoins, la variable doit être considérée comme continue et il convient d'en préciser les unités ! Ainsi, si une balance mesure au kilo près, un individu qui pèse 85 kg présente en réalité un poids situé entre 84,5 kg et 85,5 kg. La pression artérielle systolique d'un patient valant 113 mmHg se situe en réalité entre 112,5 et 113,5 mmHg. Un individu âgé de 70 ans a un âge (inconnu) supérieur à 70 ans mais inférieur à 71 ans. En conclusion, tout ce qui se mesure est considéré comme une variable continue dont les valeurs ne sont connues qu'avec une

précision limitée due à l'appareillage de mesure. Noton enfin, que certaines variables discrètes sont parfois considérées et traitées comme des variables continues parce que le nombre de valeurs distinctes qu'elles prennent est très élevé. Citons à titre d'exemple, le nombre de globules rouges ou de globules blancs par mm^3.

1.3.4 Variables binaires

Entre les variables qualitatives et quantitatives, il convient d'évoquer les variables binaires car elles sont fréquemment utilisées en médecine.

D'une part, une variable *binaire* (binary variable) est une variable qualitative à $q = 2$ modalités, que l'on peut numéroter 1 et 2 ou 0 et 1. Par ailleurs, une variable binaire peut être vue comme une variable discrète à $k = 2$ valeurs 0 et 1, sur lesquelles on peut faire des calculs. Dès lors, les variables binaires sont à l'intersection des variables qualitatives et quantitatives. Citons à tire d'exemples :

- sexe (0 = homme, 1 = femme)
- tabac (0 = non fumeur, 1 = fumeur)
- issue (0 = vie, 1 = décès)
- récurrence d'un cancer (0 = non, 1 = oui)
- remboursement INAMI d'un médicament (0 = non, 1 = oui)
- résultat d'un étudiant (0 = réussite, 1 = échec)
- cholestérol (0 = normal, 1 = anormal)
- symptôme (0 = absent, 1 = présent)

1.4 Données

Les variables observées ou mesurées sur les individus (ou objets) d'un échantillon conduisent à des *données* (data). Ces données nominales et numériques feront l'objet d'une analyse statistique (statistical data analysis).

Les données posent parfois problème.

1.4.1 Données manquantes

Dans de nombreuses études, malgré les précautions prises, certaines données sont manquantes (missing data). En clair, elles n'ont été ni observées, ni mesurées. Les raisons sont multiples : le sujet ne s'est pas présenté à la prise de sang, le patient est décédé, l'appareil de mesure n'était pas disponible, le sujet n'a pas daigné répondre à la question, etc. La présence d'un grand nombre de données manquantes dans un échantillon de données pose problème et affecte la qualité globale des résultats et des conclusions (perte de puissance statistique). En général, dans les logiciels statistiques, on laisse la case "vide" (EXCEL) ou on met un code (par exemple le point "." en SAS, "NA" en S-PLUS), de manière à signaler au logiciel que la donnée est manquante.

1.4.2 Données aberrantes

Il arrive qu'un échantillon (ou fichier) de données contiennent des valeurs totalement inhabituelles, extrêmes ou impossibles. On parle alors de valeurs aberrantes (outliers, freak values) qu'il convient de détecter et d'examiner avec le plus grand soin. Ainsi, en encodant les données dans un ordinateur, on a introduit une taille de 718 cm au lieu de 178 cm. Il arrive qu'un technicien de laboratoire effectue un dosage et oublie d'appliquer un facteur de correction (158 g/L de cholestérol au lieu de 1.58 g/L). Attention, une valeur extrême n'est pas nécessairement aberrante ! Ainsi, s'il est connu que la durée de séjour en milieu hospitalier est 8-10 jours, un polytraumatisé de la route peut séjourner 420 jours dans un hôpital. De même, la facture des produits pharmaceutiques d'un patient hospitalisé peut atteindre des montants considérables (plus d'un million de francs en hématologie clinique) sans qu'il ne s'agisse d'une erreur d'écriture de la part du pharmacien hospitalier.

Les données aberrantes peuvent fausser gravement les résultats et conclusions d'une analyse statistique.

1.4.3 Données censurées

Un problème dont les statisticiens ont pris conscience plus récemment est celui des données censurées (censored data). Une donnée est dite censurée (à gauche et/ou à droite) lorsque la valeur vraie n'est pas connue mais qu'une limite inférieure et/ou supérieure est fournie en lieu et place. Donnons deux exemples classiques.

Considérons un appareil de dosage de laboratoire incapable de détecter une concentration dans le sang inférieure à 5 μg/L par exemple. Dans ce cas, si un individu a une concentration sérique de 3 μg/L, elle ne pourra être détectée et on écrira comme résultat < 5 μg/L.

Un autre exemple, fréquent dans le domaine des essais cliniques en cancérologie, concerne la survie d'un patient depuis le diagnostic de son cancer. Si le patient est tué dans un accident de voiture 58 mois après le diagnostic du cancer ou qu'à la même date il ne se présente plus chez le médecin qui le traite pour toute autre raison, on dit qu'il est "perdu de vue" (lost-to-follow, withdrawn) ; on ne connaîtra jamais sa durée de vie exacte et on écrira comme résultat > 58 mois.

1.5 L'enquête du CUMG

Les données à l'Annexe I reprennent les résultats d'une enquête menée par le Centre Universitaire de Médecine Générale (CUMG) de Liège auprès d'un échantillon de 355 médecins généralistes de la Région wallonne en 1993 (Taziaux et al., 1996). Cette enquête visait à étudier la prescription médicamenteuse des médecins généralistes (MG) wallons aux personnes âgées.

Pour chaque médecin, on dispose des onze variables suivantes, en plus du numéro du médecin dans l'étude :

1. Sexe (1 = homme, 2 = femme)
2. Age (années)
3. Expérience (ancienneté) professionnelle (années)
4. Université d'origine (1 = ULg, 2 = UCL, 3 = ULB, 4 = autre)
5. Médecin agréé (0 = non, 1 = oui)
6. Province où le médecin travaille (1 = Brabant wallon, 2 = Hainaut, 3 = Liège, 4 = Luxembourg, 5 = Namur)
7. Patientèle (nombre de patients/semaine) : 0-30, 31-60, 61-90, > 90
8. Nombre moyen de médicaments prescrits par patient âgé (Nméd)
9. Nombre moyen de problèmes présentés par patient âgé (Nprob)
10. Variance du nombre de médicaments prescrits/patient (Vméd)
11. Variance du nombre de problèmes/patient (Vprob).

Ce fichier qui contient au total $11 \times 355 = 3905$ nombres constitue une base de données considérable qu'on ne peut interpréter sans avoir recours à une analyse statistique des données. On constate qu'il existe quelques données manquantes dans le fichier.

En réalité, chaque médecin devait fournir des données pour au maximum 10 patients choisis au hasard dans sa clientèle âgée de plus de 75 ans. En particulier, au moment de l'enquête, pour chaque patient, il devait fournir le nombre de problèmes que présentait le patient, le nombre de médicaments prescrits au patient et le nom de ces médicaments. Ces données ont permis de construire les variables 8 à 11. Cette enquête s'est soldée par des informations pour

- 355 médecins wallons
- 3384 patients âgés de plus de 75 ans (9,5 patients/médecin)
- 16117 prescriptions de médicaments (4,8 médicaments/patient)

Cet exemple montre la variété des variables étudiées (qualitative, continue, binaire, discrète, catégorisée, etc.).

Chapitre 2

Statistique descriptive graphique

2.1 Introduction

Désignons par X la variable à laquelle on s'intéresse et par $\{x_1, \ldots, x_n\}$ les observations ou mesures de la variable X réalisées sur un échantillon de n sujets (ou objets) extraits de la population à laquelle on s'intéresse. Ainsi donc x_1 est l'observation de la variable X chez le sujet 1, x_2 est l'observation de la variable X chez le sujet 2, ainsi de suite, et x_n l'observation de la variable X chez le sujet n.

Comme on l'a vu, la statistique est la discipline qui étudie les méthodes de réduction de données. Une manière classique de présenter les données en statistique est sous forme graphique. On dit souvent "qu'un petit dessin vaut mieux qu'un long discours". C'est en général sous forme graphique que les statistiques sont présentées dans la presse et les médias.

Procédons par ordre selon le type de variable envisagée.

2.2 Variable qualitative

2.2.1 Tableau recensé

Lorsque X est une variable qualitative (ordinale ou non) à q modalités $m_1, m_2, \ldots, m_q$, on recense pour chaque modalité m_i le nombre r_i de sujets qui ont cette modalité. On calcule ensuite la fréquence (relative) $f_i = r_i/n$ de chaque modalité ($i = 1, \ldots, q$). Il est souhaitable de multiplier f_i par 100 afin d'exprimer les résultats en pourcent.

A titre d'illustration (voir Annexe I), le tableau recensé de la variable "Université d'origine du médecin" (4 modalités) est repris ci-dessous.

Tableau 2.1 Enquête CUMG. Distribution statistique de l'université d'origine du médecin

Université d'origine	Répétition	Fréquence (%)
ULg	139	39.2
UCL	165	46.5
ULB	46	13.0
Autre	5	1.4
Total	355	100 %

2.2.2 Représentation graphique

Le tableau recensé peut se présenter graphiquement sous diverses formes. La plus courante est le diagramme de fréquence (bar diagram) où l'on met en abscisse les modalités et en ordonnée les fréquences (voir Figure 2.1). Une autre approche consiste à faire une présentation en forme de "camembert" (pie chart) (voir Figure 2.2). On parle chaque fois de la distribution statistique de la variable. Notons que l'ordre des modalités n'a aucune importance, sauf s'il s'agit d'une variable ordinale ou catégorisée (par exemple, la patientèle du médecin).

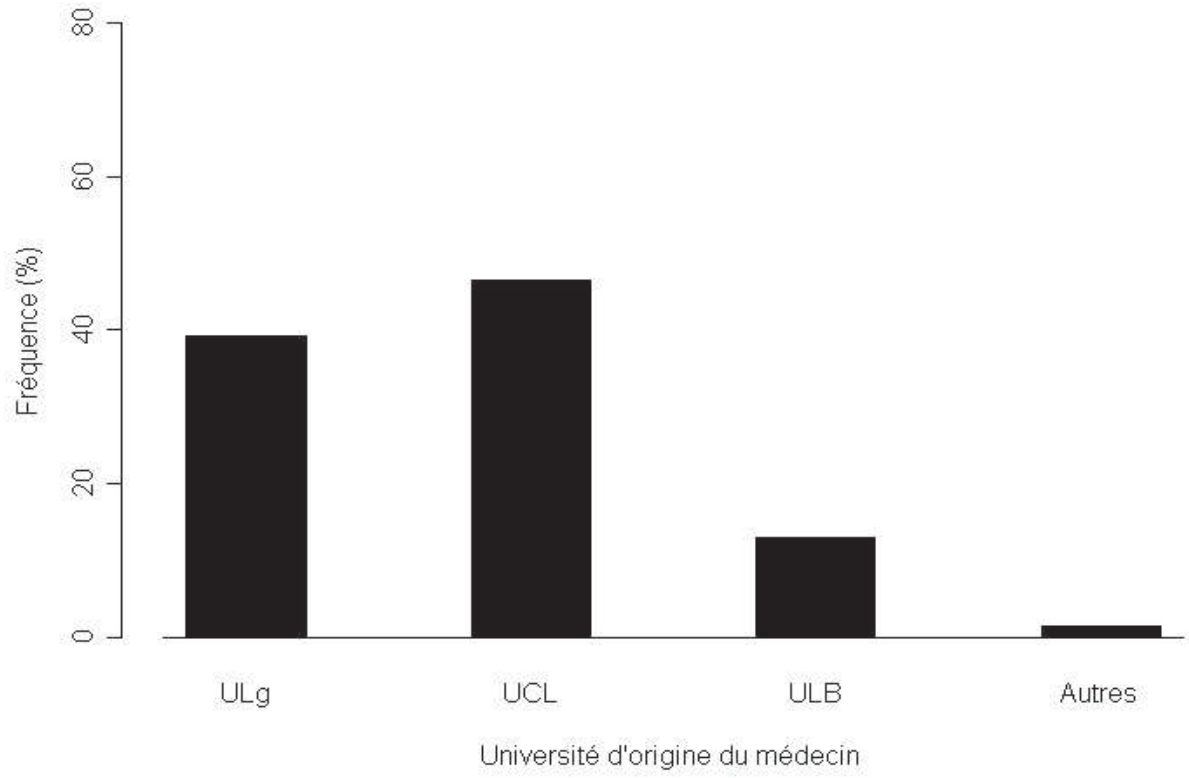

Figure 2.1 Répartition des médecins en fonction de leur université d'origine (Enquête du CUMG, $n = 355$ médecins)

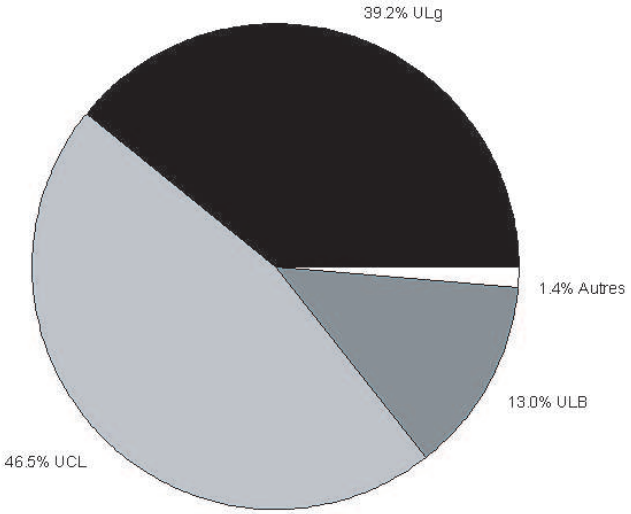

Figure 2.2 Répartition des médecins en fonction de leur université d'origine (Enquête du CUMG, $n = 355$ médecins). Présentation en forme de camembert.

Pour une variable binaire, il n'y a que deux catégories et le graphique se simplifie. A titre d'exemple, la Figure 2.3 représente la distribution statistique du sexe du médecin. On observe une majorité d'hommes dans la profession (83.1% d'hommes contre 16.9% de femmes).

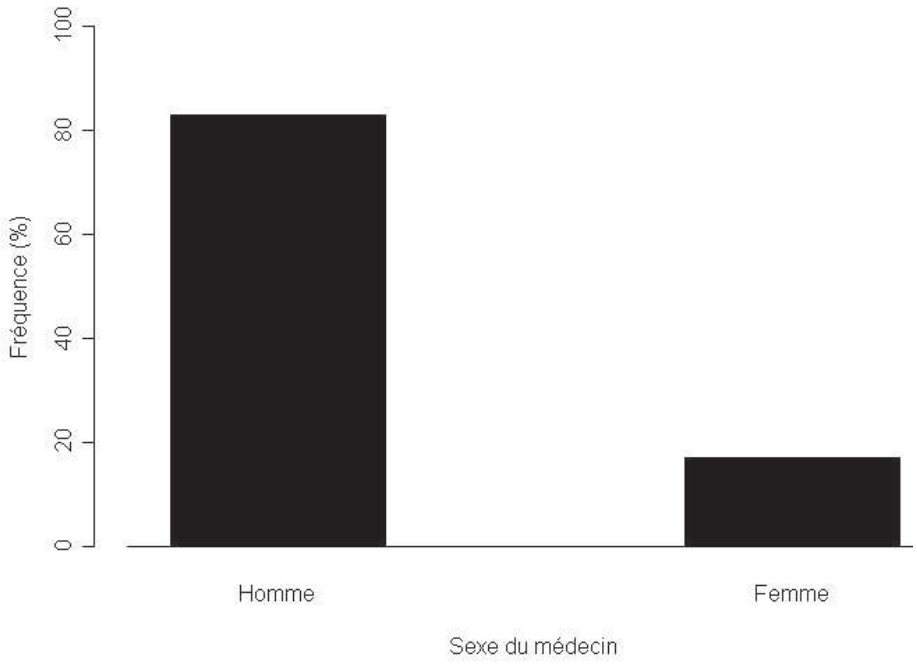

Figure 2.3 Répartition des médecins en fonction de leur sexe (enquête du CUMG, $n = 355$ médecins).

2.3 Variable discrète

2.3.1 Tableau recensé

Lorsque X est une variable quantitative discrète, la démarche à suivre est fort semblable à celle d'une variable qualitative. Notons k le nombre de valeurs distinctes prises par la variable et $a_1, \ldots, a_k$ ces valeurs. Pour chaque valeur a_i de la variable, on recense le nombre r_i de fois qu'elle est présente dans l'échantillon et on calcule la fréquence (relative) correspondante $f_i = r_i/n$, souvent exprimée en pourcent. On établit alors un tableau recensé comme précédemment. Lorsque l'échantillon de données $\{x_1, \ldots, x_n\}$ est grand (n élevé), il convient de le trier par ordre croissant (on obtient alors un *tableau ordonné*), sans quoi le recensement devient fastidieux voire impossible si l'on ne dispose pas d'un ordinateur.

A titre d'exemple, on a étudié le nombre d'enfants par ménage (variable X) dans un échantillon de $n = 133$ ménages. Les observations sont triées par ordre croissant comme l'indique le tableau ordonné ci-dessous (Tableau 2.2).

Tableau 2.2 Etude statistique du nombre d'enfants par ménage (variable X) portant sur un échantillon de $n = 133$ ménages (tableau ordonné)

0	0	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
3	3	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	5	5	5
5	5	5	5	5	5	5	5	5	5
5	6	6	6	6	6	6	6	6	6
6	6	6	6	6	6	6	6	7	7
9	9	10							

Le Tableau 2.3 donne la distribution statistique du nombre d'enfants par ménage après recensement des observations.

Tableau 2.3 Recensement du nombre d'enfants par ménage dans un échantillon de 133 ménages (tableau recensé)

Enfants/ménage a_i	Répétition r_i	Fréquence (%) f_i	Fréquence cumulée (%) c_i
0	2	1.5	1.5
1	8	6.0	7.5
2	10	7.5	15.0
3	52	39.1	54.1
4	25	18.8	72.9
5	14	10.5	83.4
6	17	12.8	96.2
7	2	1.5	97.7
8	0	0	97.7
9	2	1.5	99.2
10	1	0.8	100.0
Total	$n = 133$	100 %	

Comme on peut le constater, la valeur $X = 8$ n'a pas été observée dans l'échantillon (il n'y avait pas de ménages avec 8 enfants). Ceci ne signifie pas qu'il n'y a pas de ménages avec 8 enfants dans la population !

La dernière colonne du Tableau 2.3 donne les "fréquences cumulées". Celles-ci s'obtiennent en cumulant pour chaque valeur distincte $X = a_i$ les fréquences relatives f_i jusqu'à cette valeur a_i, c'est-à-dire $c_i = f_1 + f_2 + \cdots + f_i$ $(i = 1, \ldots, k)$. On voit que $c_i = c_{i-1} + f_i$. En d'autres termes, $c_1 = f_1$, $c_2 = f_1 + f_2 = c_1 + f_2$, $c_3 = c_2 + f_3$, etc. Bien évidemment, $c_k = 100\%$.

2.3.2 Représentation graphique

Le diagramme en bâtons (bar diagram) est la représentation graphique classique des variables discrètes. On y reporte les fréquences (relatives) f_i en fonction des valeurs de la variable X. Attention, contrairement aux variables qualitatives, l'ordre des valeurs doit être respecté, puisqu'il s'agit d'une variable numérique. La Figure 2.4 donne la distribution observée du nombre d'enfants par ménage. L'autre graphique (Figure 2.5) donne le diagramme cumulatif où l'on reporte les fréquences cumulées c_i en fonction des valeurs de la variable.

Le diagramme cumulatif est moins aisé à comprendre mais il permet d'estimer ce que l'on appelle les "queues" de la distribution. Par exemple, on peut répondre aux questions suivantes : Quelle est la proportion de ménages qui ont plus de 5 enfants ? (Réponse : 16.6%) ou quelle est la probabilité de trouver un ménage avec moins de trois enfants ? (Réponse : 15%).

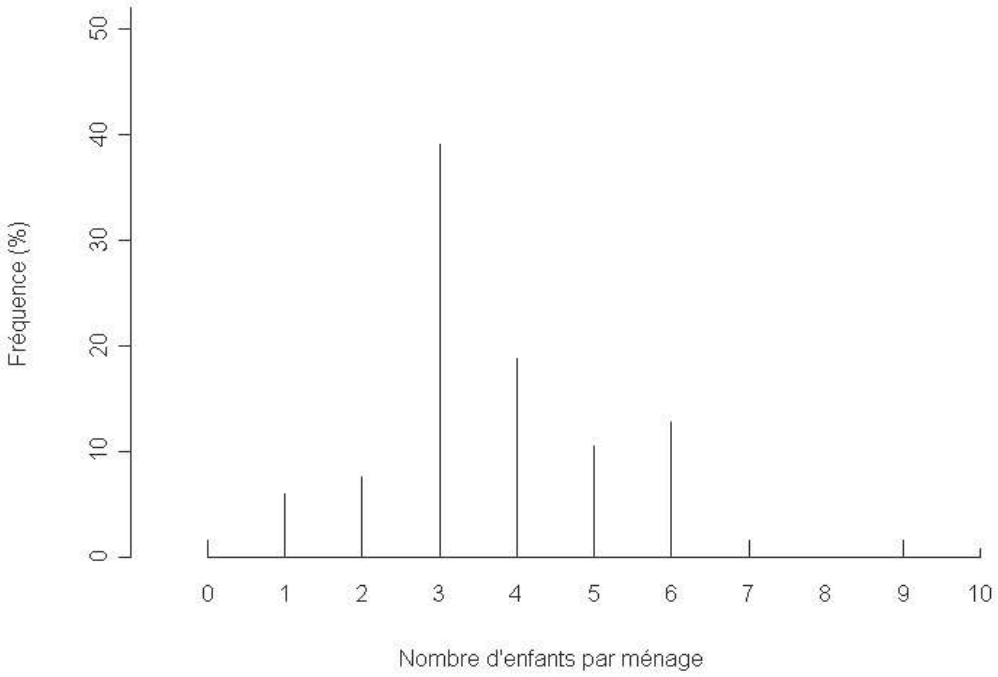

Figure 2.4 Distribution du nombre d'enfants par ménage dans un échantillon de 133 ménages

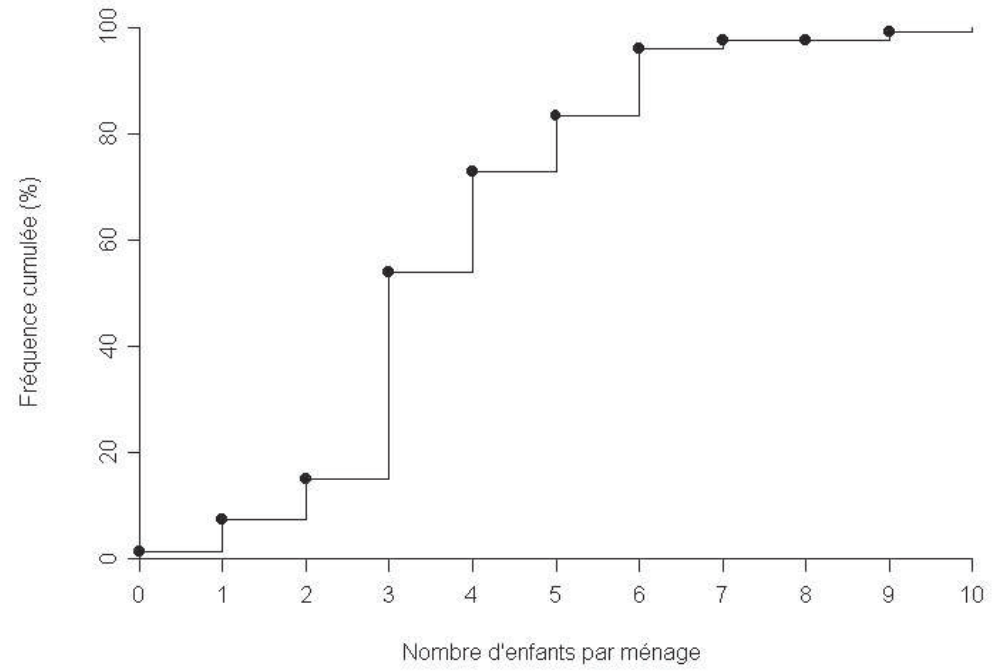

Figure 2.5 Diagramme cumulatif du nombre d'enfants par ménage dans un échantillon de 133 ménages.

La Figure 2.6 donne la distribution du nombre de médicaments prescrits par patient âgé de plus de 75 ans par les médecins généralistes wallons ($n = 3384$ patients).

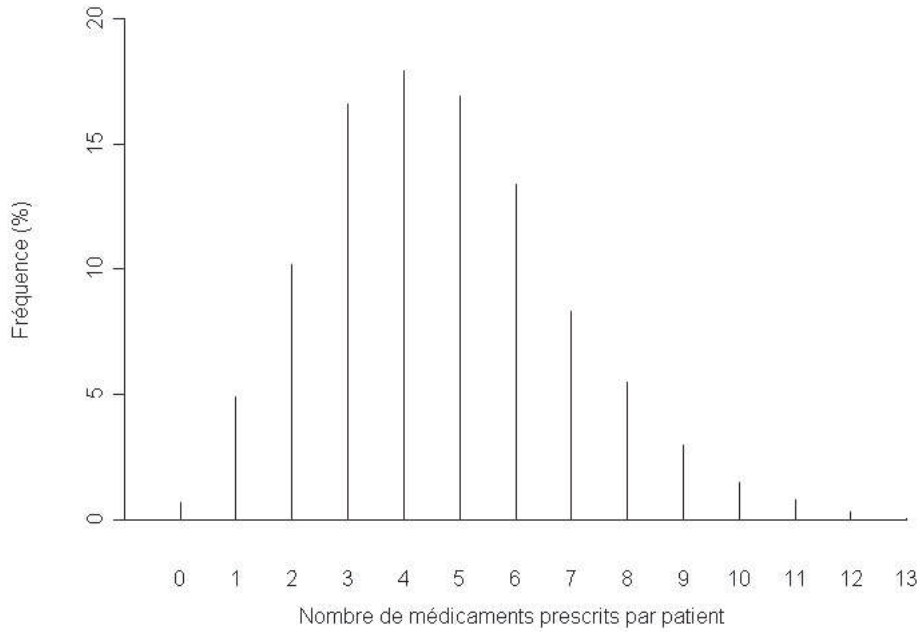

Figure 2.6 Distribution du nombre de médicaments prescrits par les médecins généralistes wallons aux personnes âgées. (Enquête du CUMG, $n = 3384$ patients).

2.4 Variables continues

2.4.1 Tableau de classes

Lorsque la variable X est continue, on pourrait construire un tableau recensé comme dans le cas des variables discrètes. Cette approche n'a guère de sens car les valeurs observées d'une variable continue peuvent être toutes distinctes et dès lors le diagramme en bâtons n'apporte aucune information, même si le diagramme cumulatif présente un intérêt. On procède de préférence par l'élaboration d'un "tableau de classes".

On définit un certain nombre (q) de classes définies par des limites inférieures et supérieures, de manière à couvrir l'ensemble du domaine des valeurs de la variable continue X. Pour chaque classe, on définit le centre de classe (C_i), on recense "le nombre" r_i d'observations de l'échantillon tombant dans la classe, on calcule la fréquence (relative) $f_i = r_i/n$ et la fréquence cumulée correspondante $c_i = c_{i-1} + f_i$ $(i = 1, \ldots, q)$ avec $c_0 = 0$.

Le Tableau 2.4 donne l'âge (années) de 100 patients admis dans un hôpital spécialisé pour les maladies chroniques durant une période d'un mois.

Tableau 2.4 Age (années) de 100 patients admis dans un hôpital spécialisé pour les maladies chroniques durant une période d'un mois

10	22	24	42	37	77	89	85	28	63
9	10	7	51	2	1	52	7	48	54
32	29	2	15	46	48	39	6	72	14
36	69	40	61	12	21	54	53	58	32
27	33	1	25	22	6	81	11	56	5
63	53	88	48	52	87	71	51	52	33
46	33	85	22	5	87	28	2	85	61
16	42	69	7	10	53	33	3	85	8
51	60	58	9	14	74	24	87	7	81
30	76	7	6	27	18	17	53	70	49

On construit ensuite le tableau de classes (voir Tableau 2.5).

Tableau 2.5 Distribution de l'âge (années) dans un échantillon de 100 patients admis dans un hôpital spécialisé pour maladies chroniques sur une période d'un mois (tableau de classes)

Classe d'âge [1]	Centre C_i	Répétition r_i	Fréquence (%) f_i	Fréquence cumulée (%) c_i
0-10	5	22	22	22
10-20	15	8	8	30
20-30	25	13	13	43
30-40	35	10	10	53
40-50	45	8	8	61
50-60	55	16	16	77
60-70	65	7	7	84
70-80	75	5	5	89
80-90	85	11	11	100
Total		$n = 100$	100%	

[1] 0-10 signifie de 0 exclu à 10 inclus, 10-20 de 10 exclu à 20 inclus, etc.

En général, les classes sont de même largeur (ici de 10 en 10 ans) mais ce n'est pas une obligation. En cas de classes de largeur inégale, il faut être attentif dans la construction de l'histogramme. On calcule à cet effet la hauteur de classe $h_i = f_i/(k_i \times u)$, où $u = 1$ est l'unité de largeur de classe choisie et k_i est le nombre d'unités de largeur de la classe i. Par exemple, si u représente 10 ans et que l'on regroupe les deux dernières classes, on constate que $h_i = 8\%$ puisque f_i vaut $5+11 = 16\%$ et $k_i = 2$ car la largeur de la nouvelle classe vaut 20 ans $= 2 \times 10$ ans ! Le nombre de classes est laissé à l'appréciation de celui qui fait l'analyse statistique. Trop de classes risque de conduire à des classes vides, trop peu de classes risque de masquer des aspects importants de la distribution. Une règle

heuristique consiste à prendre $q \simeq \sqrt{n}$. D'autres auteurs proposent $q \simeq 1 + 1.44\,\ell n(n)$. Par exemple, pour $n = 100$, dans le premier cas on a $q = 10$ et dans l'autre $q = 8$.

2.4.2 Histogramme et diagramme cumulatif

L'histogramme (des fréquences) consiste à reporter pour chaque classe un rectangle dont la base est la largeur de classe et l'aire, la fréquence (relative) correspondante. L'histogramme est donc un diagramme d'aire. Les hauteurs des rectangles ne sont proportionnelles aux fréquences f_i que si les classes sont de même largeur, sinon il faut calculer les hauteurs h_i.

Le diagramme cumulatif consiste à reporter les fréquences cumulées en fonction des classes. En général, on part de 0 à la limite inférieure de la première classe et on termine à 1 (ou 100 %) à la limite supérieure de la dernière classe. On rejoint les points par des segments de droite.

L'histogramme et le diagramme cumulatif relatifs aux données reprises au Tableau 2.5 (âge des patients) sont repris aux Figures 2.7 et 2.8. Comme on peut le voir, la distribution de l'âge présente plusieurs pics (0-10 ans, 20-30 ans, 50-60 ans, 80-90 ans).

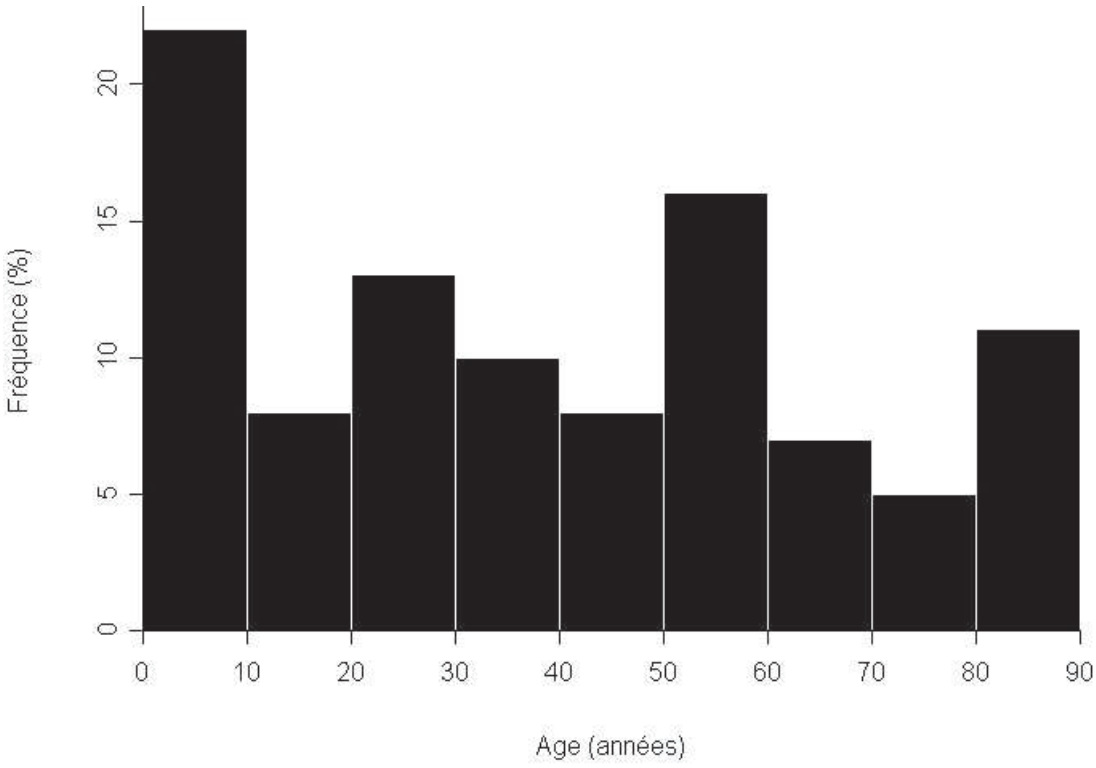

Figure 2.7 Histogramme de l'âge des patients admis dans un hôpital spécialisé pour les maladies chroniques sur une période d'un mois ($n = 100$ patients)

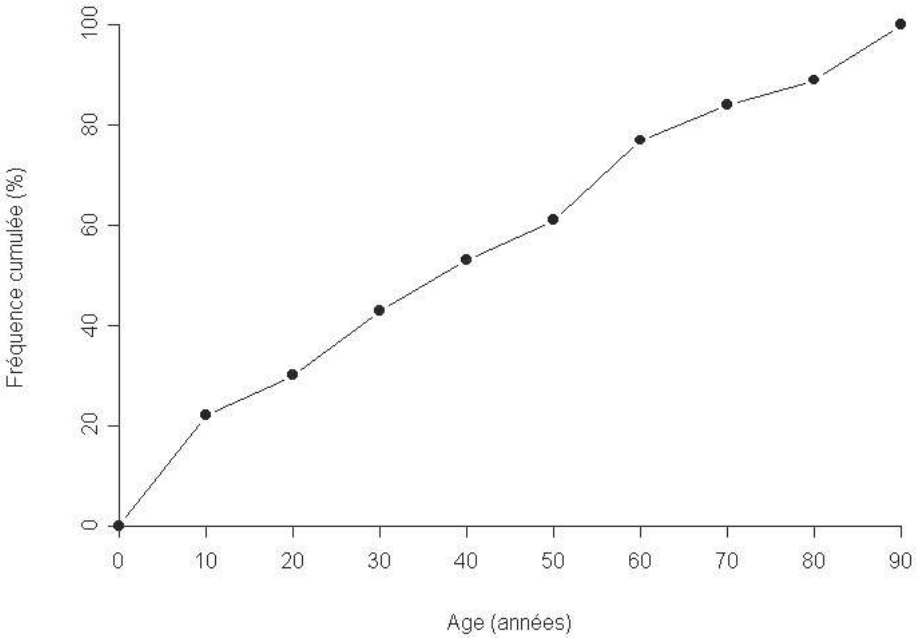

Figure 2.8 Diagramme cumulatif de l'âge des patients admis dans un hôpital spécialisé pour les maladies chroniques sur une période d'un mois ($n = 100$ patients)

Les Figures 2.9 et 2.10 donnent les histogrammes de l'âge (années) et de l'expérience professionnelle (années) du médecin, dans l'enquête du CUMG sur la prescription de médicaments aux personnes âgées en Wallonie (voir Annexe I).

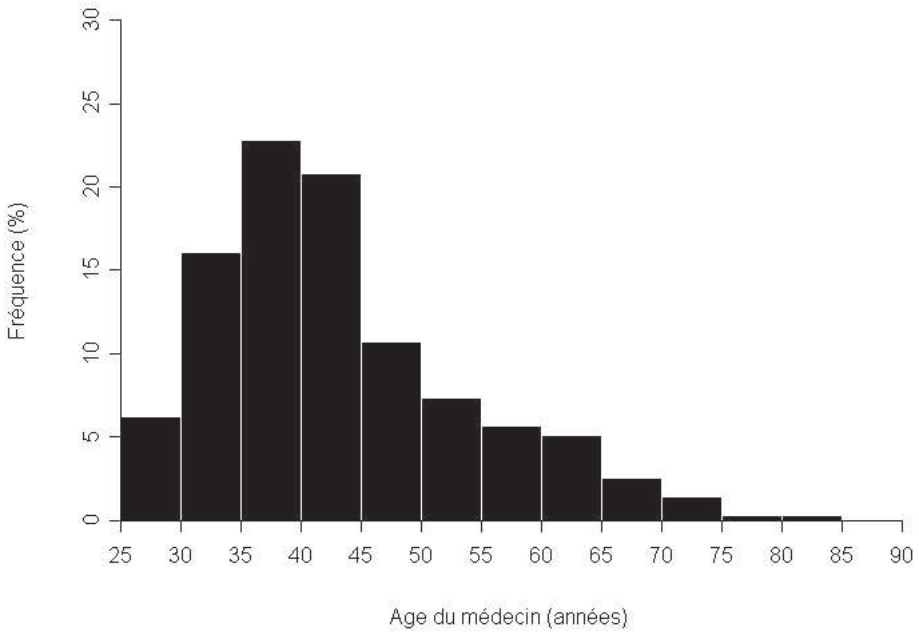

Figure 2.9 Histogramme de l'âge des médecins (enquête du CUMG, $n = 352$ médecins)

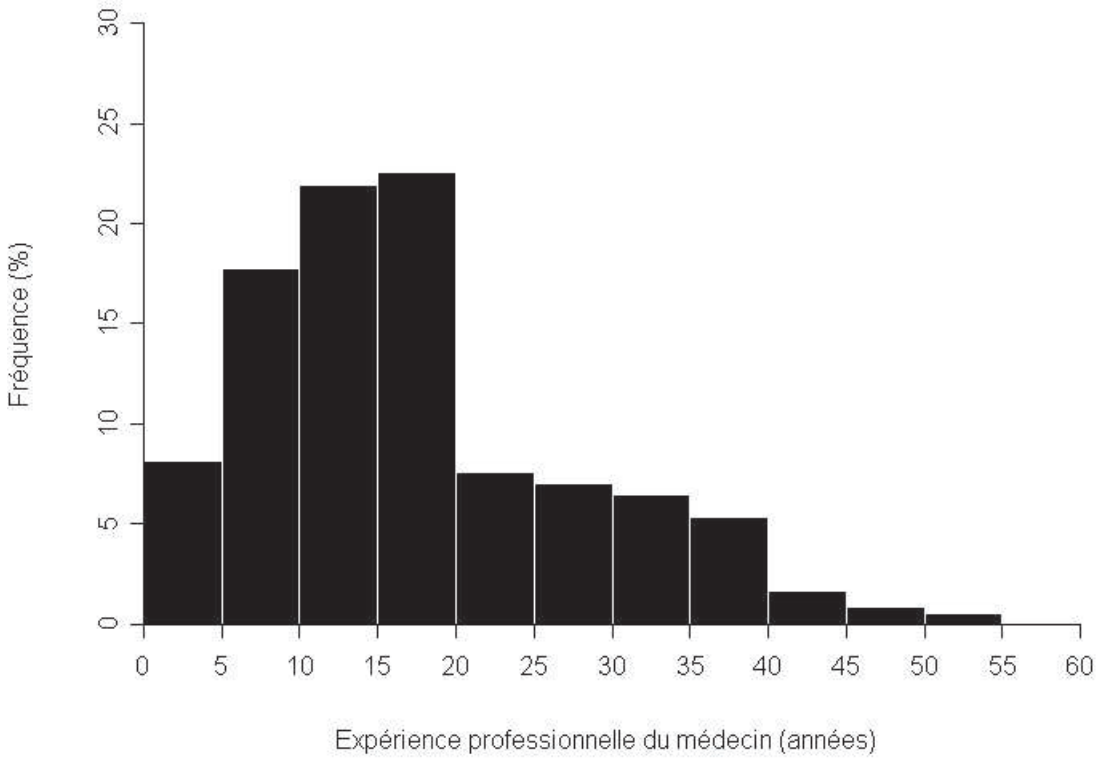

Figure 2.10 Histogramme de l'expérience professionnelle des médecins (enquête du CUMG, $n = 352$ médecins)

2.5 Représentation bivariée et multivariée

Lorsqu'on dispose des observations de deux variables quantitatives (X et Y) dans un échantillon de n sujets, on peut en obtenir une représentation graphique en reportant une variable par rapport à l'autre. Ainsi, en abscisse on met la variable X et en ordonnée la variable Y. Ce type de graphique bidimensionnel est utile et couramment utilisé.

A titre d'exemple, la Figure 2.11 reporte l'expérience professionnelle du médecin en fonction de son âge dans l'enquête du CUMG. On observe une relation évidente entre les deux variables. De même, la Figure 2.12 donne pour chaque médecin le nombre moyen de médicaments prescrits par patient en fonction du nombre moyen de problèmes (pathologies) présentés par patient. On observe également une relation croissante entre les deux variables.

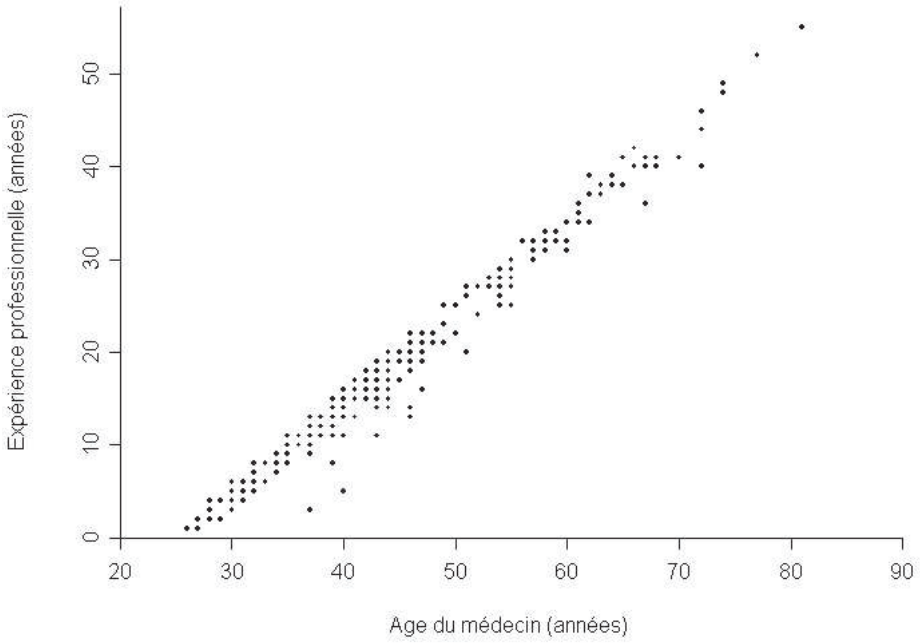

Figure 2.11 Relation entre l'expérience professionnelle et l'âge du médecin (enquête du CUMG, $n = 352$ médecins)

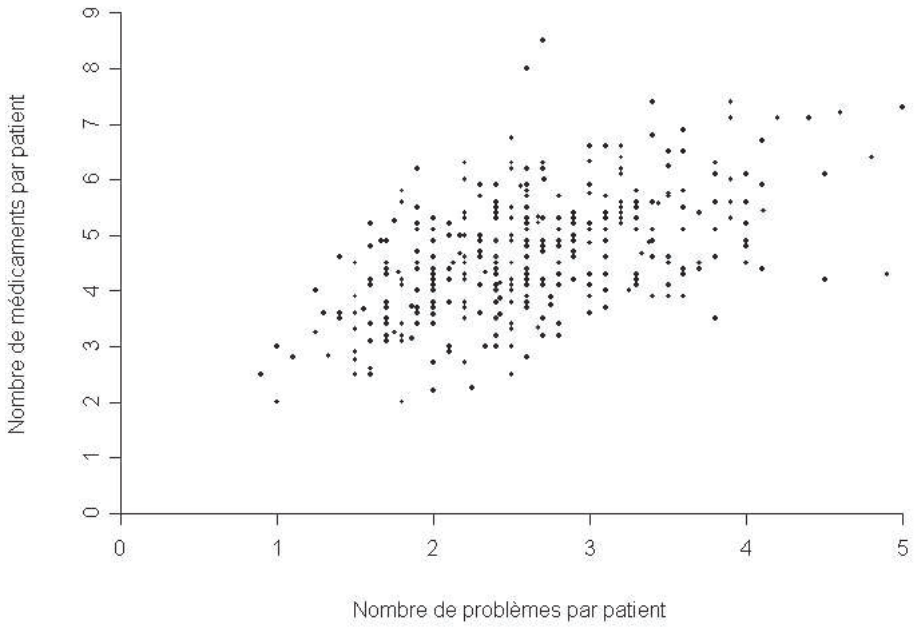

Figure 2.12 Relation entre le nombre moyen de médicaments prescrits par patient et le nombre moyen de problèmes présentés par patient chez 355 médecins ayant participé à l'enquête du CUMG

Il est possible aujourd'hui d'obtenir des représentations à deux dimensions d'observations dans un espace à p-dimensions. On reporte sur ce type de graphique, communément appelé "biplot", non seulement la position de chaque observation mais également celle des variables $X_1, \ldots, X_p$ du problème. La Figure 2.13 donne le biplot des 355 médecins généralistes wallons, où l'on a tenu compte des variables sexe (X_1), âge (X_2), expérience professionnelle (X_3), agrément (X_4), nombre de médicaments prescrits/patient (X_5), nombre de problèmes/patient (X_6), variance du nombre de médicaments prescrits/patient (X_7) et variance du nombre de problèmes/patient (X_8).

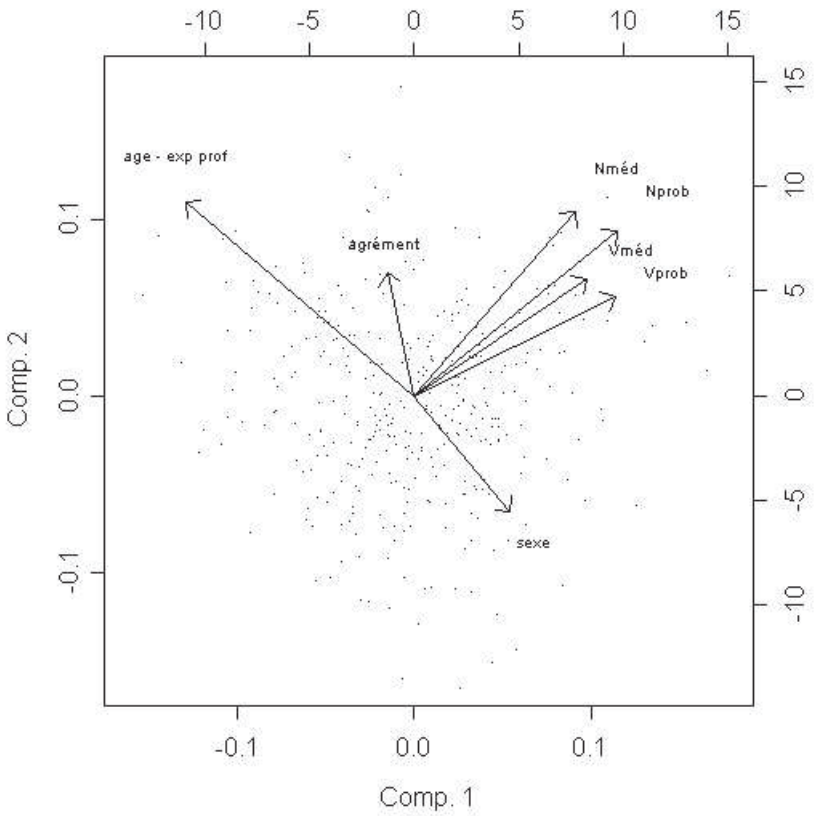

Figure 2.13 Enquête du CUMG. Exemple de biplot avec $n = 355$ médecins et $p = 8$ variables

2.6 Conclusion

La représentation graphique d'un échantillon de données apporte des informations essentielles sur la distribution statistique de la variable dans la population étudiée, plus précisément sur

- la présence ou non d'une valeur centrale (pic de fréquences)
- la dispersion des données

– la forme de la distribution (unimodale, multimodale, symétrique, dissymétrique à gauche ou à droite).

Par ailleurs, la comparaison de la distribution d'une même variable dans deux ou plusieurs populations peut révéler des différences importantes entre ces populations.

Enfin, les représentations bivariées et multivariées permettent de mettre en évidence des associations entre variables et la présence éventuelle de données aberrantes.

Chapitre 3

Moyenne et écart-type

3.1 Introduction

Il n'est pas toujours possible ni opportun de donner une représentation graphique des données (histogramme, diagramme de fréquences, diagramme cumulatif, biplot). C'est vrai notamment lorsque le nombre de variables est élevé comme c'est le cas dans de nombreuses études, ou pour des raisons de place, par exemple dans un mémoire de fin d'études ou une publication scientifique. Il faut alors trouver d'autres moyens pour synthétiser un échantillon de données. Plutôt qu'une approche graphique, on peut avoir recours à une approche "numérique" par laquelle on résume un échantillon de données par d'autres nombres, appelés caractéristiques, paramètres ou indicateurs d'échantillon.

Lorsqu'on a affaire à une variable qualitative, il suffit, comme on l'a vu, de calculer les fréquences de chacune des modalités (voir Tableau 2.1). En effet, on ne peut effectuer d'autres calculs puisque les modalités sont nominales.

Pour les variables quantitatives, il est habituel de résumer la distribution statistique des données à l'aide de deux paramètres, la moyenne et l'écart-type. Il arrive parfois que ces paramètres ne soient pas les plus appropriés (voir Chapitre 4). Comme précédemment, désignons par $\{x_1, \ldots, x_n\}$ un échantillon de n données numériques.

3.2 Moyenne

La moyenne (mean ou average) est le premier grand concept fondamental en statistique. Elle permet de localiser l'échantillon sur l'échelle des valeurs de la variable.

3.2.1 Définition

Par définition, la moyenne arithmétique d'un échantillon d'effectif n est notée $\bar{x}$ (ou m) et est définie par l'expression

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.1}$$

où, par convention, $\sum_{i=1}^{n} x_i = x_1 + \cdots + x_n$, la somme des n observations. Lorsque le contexte le permet, on écrit plus simplement $\sum x$.

Dans le cas d'une variable discrète à k valeurs, la formule (3.1) peut aussi s'écrire

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} a_i r_i \tag{3.2}$$

où a_i et r_i désignent respectivement les valeurs distinctes de la variable et les répétitions correspondantes (voir Tableau recensé).

Dans le cas d'un tableau de classes, on peut calculer une moyenne arithmétique "approchée" à l'aide de la formule

$$\bar{x}_{app} = \frac{1}{n} \sum_{i=1}^{q} C_i \times r_i \tag{3.3}$$

où q est le nombre de classes, C_i les centres de classes et r_i les répétitions correspondantes.

3.2.2 Propriétés

- La moyenne arithmétique est un concept simple et intuitif. Elle est facile et rapide à calculer.
- La moyenne est le centre de gravité de la distribution des observations $\{x_1, \ldots, x_n\}$ au sens physique du terme, c'est-à-dire $\sum(x - \bar{x}) = 0$. On montre d'ailleurs qu'elle divise le diagramme cumulatif (exact) en deux régions telles que l'aire à gauche de $\bar{x}$ et sous le diagramme cumulatif est égale à l'aire à droite de $\bar{x}$ et au-dessus du diagramme cumulatif.
- Si $y_i = a + bx_i$ $(i = 1, \ldots, n)$, alors $\bar{y} = a + b\bar{x}$.
- La moyenne arithmétique est particulièrement sensible aux valeurs aberrantes car elle fait intervenir toutes les valeurs de l'échantillon. On dit qu'elle est peu robuste aux valeurs extrêmes. Il convient donc de détecter ces valeurs au risque de fausser la moyenne.
- La moyenne arithmétique est un paramètre dit de *position* dans la mesure où il situe le centre de la distribution statistique des données.

3.2.3 Proportion

On montre facilement qu'une proportion est une moyenne arithmétique. En effet, soit X une variable binaire prenant les valeurs de 0 et 1. Dans ce cas, l'échantillon d'effectif n est composé de 0 et de 1. Si on désigne par r le nombre de valeurs égales à 1 et donc par $(n - r)$ le nombre de valeurs nulles, l'équation (3.2) s'écrit successivement

$$\begin{aligned} \bar{x} &= \frac{1}{n} \left[0 \times (n - r) + 1 \times r\right] \\ &= \frac{r}{n} \\ &= p \end{aligned}$$

où p $(0 \leq p \leq 1)$ représente la proportion de sujets ayant la valeur $X = 1$ dans l'échantillon. En conséquence, toute proportion est une moyenne et obéit aux mêmes propriétés.

3.2.4 Exemples

L'application de la formule (3.2) à l'échantillon de 133 ménages montre que le nombre moyen d'enfants par ménage vaut $\bar{x} = 3.74$.

L'application de la formule (3.1) à l'échantillon des 100 patients admis en hospitalisation conduit à un âge moyen de $\bar{x} = 39.2$ ans. Par contre, en utilisant la formule (3.3) et le tableau de classes (voir Tableau 2.5), on trouve une moyenne d'âge approchée égale à $\bar{x}_{app} = 39.1$ ans, proche de la moyenne arithmétique exacte.

Enfin, dans l'étude sur la prescription des médicaments aux personnes âgées en Wallonie (Annexe I), on montre que la proportion (moyenne) de médecins de sexe masculin est égale à $p = 295/355$, soit 83.1%. La proportion de femmes vaut donc 16.9%.

3.3 Ecart-type

L'écart-type (standard deviation SD) est le deuxième plus important concept de base en statistique. Il fournit des informations sur la variabilité des données de l'échantillon autour de la moyenne. C'est donc un indicateur de dispersion. Or, nous l'avons vu, la variabilité est un élément fondamental dans les sciences de la vie.

3.3.1 Définition

Par définition, l'écart-type d'un échantillon de données d'effectif n est noté s (ou SD) et est défini par l'expression

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \tag{3.4}$$

où le numérateur $\sum (x - \bar{x})^2 = (x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2$ est appelé "somme des carrés" et le dénominateur $(n - 1)$ "degrés de liberté".

Pour des raisons de précision de calculs, il est préférable de développer le numérateur sous la racine carré comme suit : $\sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2/n$. La formule 3.4 devient dès lors

$$s = \sqrt{\frac{\sum x^2 - (\sum x)^2/n}{n - 1}}. \tag{3.5}$$

On voit immédiatement que pour calculer l'écart-type, on a besoin de la somme des x (nécessaire pour la moyenne) mais aussi de la somme des x^2 (termes du second degré).

3.3.2 Propriétés

- L'écart-type est un indicateur de *dispersion* non négatif ($s \geq 0$). En réalité, en l'absence de variabilité, $s = 0 \Leftrightarrow x_i = \bar{x}$ $(i = 1, \ldots, n)$.
- Pour deux échantillons ayant la même moyenne ($\bar{x}_1 = \bar{x}_2$), si $s_1 > s_2$ le premier échantillon est plus dispersé que le second, et inversement si $s_1 < s_2$.
- Le carré de l'écart-type, s^2, est appelé la "variance" de l'échantillon.
- Si $y_i = a + bx_i$ $(i = 1, \ldots, n)$, alors $s_y = bs_x$. Il en résulte que l'écart-type est invariant par rapport aux translations mais non aux facteurs d'échelle.
- L'écart-type s possède les mêmes unités que la variable X.
- Comme la moyenne, l'écart-type est sensible aux observations aberrantes qui peuvent dès lors en fausser complètement la valeur.
- Notons qu'à partir de l'expression (3.5), on peut recalculer $\sum x^2$ connaissant s et $\bar{x}$ puisque $\sum x^2 = (n-1)s^2 + n\bar{x}^2$.
- Dans le cas d'une variable binaire ($X = 0$ ou 1), on montre que

$$s = \sqrt{p(1-p)}. \tag{3.6}$$

  et qu'en conséquence, connaissant la moyenne p on connaît l'écart-type. De plus, on voit que $s = 0$ lorsque $p = 0$ ou $p = 1$. En effet, si dans une classe de 100 étudiants, tout le monde fume ($p = 1$), il n'y a pas de variabilité. Il en serait de même si personne dans la classe ne fumait ($p = 0$). La formule (3.6) est fréquemment utilisée en épidémiologie.
- Comme pour la moyenne, dans le cas d'une variable discrète, on peut se servir du tableau recensé et écrire

$$s = \sqrt{\frac{\sum_{i=1}^{k} a_i^2 r_i - (\sum_{i=1}^{k} a_i r_i)^2/n}{n-1}} \tag{3.7}$$

  et ainsi se faciliter la tâche. De même, on peut définir un écart-type approché à partir d'un tableau de classes et on a

$$s_{app} = \sqrt{\frac{\sum_i C_i^2 r_i - (\sum C_i r_i)^2/n}{n-1}} \tag{3.8}$$

  où la somme s'effectue sur les q classes du tableau.

3.3.3 Exemples

La valeur de l'écart-type du nombre d'enfants par ménage pour l'échantillon des 133 ménages (Tableau 2.3) vaut $s = 1.68$, puisque $\sum x = \sum a_i r_i = 498$ et $\sum x^2 = \sum a_i^2 r_i = 2238$.

En ce qui concerne l'échantillon des 100 patients admis en hospitalisation (Tableau 2.4), on a $\sum x = 3920$ et $\sum x^2 = 224452$. L'écart-type de l'âge vaut donc $s = 26.74$ ans.

Notons enfin qu'en se servant des données du Tableau 2.5, l'écart-type approché vaut $s_{app} = 26.89$ (proche de la valeur exacte) avec $\sum C_i r_i = 3910$ et $\sum C_i^2 r_i = 224500$.

3.4 Présentation des résultats

Classiquement, pour résumer un échantillon de données dans un travail ou une publication scientifique, on a recours aux concepts de moyenne (ou de proportion) et d'écart-type. A titre d'illustration, l'analyse statistique des données recueillies sur les médecins généralistes ayant participé à l'enquête du CUMG sur la prescription de médicaments aux personnes âgées (Annexe I) est présentée au Tableau 3.1.

Tableau 3.1 Caractéristiques des 355 médecins généralistes ayant participé à l'enquête du CUMG sur la prescription médicamenteuse à la personne âgée en Wallonie

Variable		Moyenne ± écart-type	Nombre	(%)
Sexe	Hommes		295	(83)
	Femmes		60	(17)
Age (années)		43.8 ± 10.7		
Expérience professionnelle (années)		18.0 ± 10.5		
Université d'origine	ULg		138	(39.3)
	UCL		162	(46.2)
	ULB		46	(13.1)
	Autre		5	(1.4)
Médecin agréé	oui		326	(93)
	non		26	(7)
Province	Brabant wallon		32	(9.0)
	Hainaut		123	(34.6)
	Liège		131	(36.9)
	Luxembourg		21	(5.9)
	Namur		48	(13.5)
Patientèle (patients/semaine)				
< 30			27	(8)
31-60			70	(20)
61-90			84	(24)
> 90			171	(48)
Nombre moyen de médicaments prescrits		4.66 ± 1.09		
Nombre moyen de problèmes présentés		2.60 ± 0.74		
Variance du nombre de médicaments prescrits		4.14 ± 2.66		
Variance du nombre de problèmes présentés		1.28 ± 0.95		

3.5 Intervalle de référence

Une des plus belles applications de la moyenne et de l'écart-type en médecine, et plus particulièrement en biologie clinique, est la construction d'intervalles de référence (reference intervals). Par le passé, on disait aussi "valeurs normales". Toutefois, cette appellation est aujourd'hui abandonnée en raison de la confusion entre le caractère "Normal" (gaussien) d'une distribution statistique et l'état "normal" d'un individu au sens médical du terme.

En effet, on peut montrer que si la distribution de la variable X est symétrique et approximativement Normale (gaussienne), comme c'est le cas pour la distribution des erreurs de mesure ou celle de nombreux paramètres biologiques, alors 95% des observations (et donc des sujets) sont compris dans l'intervalle défini par les limites $\bar{x} \pm 1.96s$, c'est-à-dire

$$\bar{x} \pm 2s. \tag{3.9}$$

Cet intervalle est appelé intervalle de référence (ou intervalle de tolérance) car seulement 5% des sujets tombent en dehors des limites (voir figure 3.1). En statistique, une proportion de 5% est considérée comme très faible. Le résultat précédent est remarquable dans la mesure où la seule connaissance des deux caractéristiques de l'échantillon, $\bar{x}$ et s, permet de construire l'intervalle. Ces intervalles de référence sont utilisés surtout en biométrie, biologie clinique et médecine.

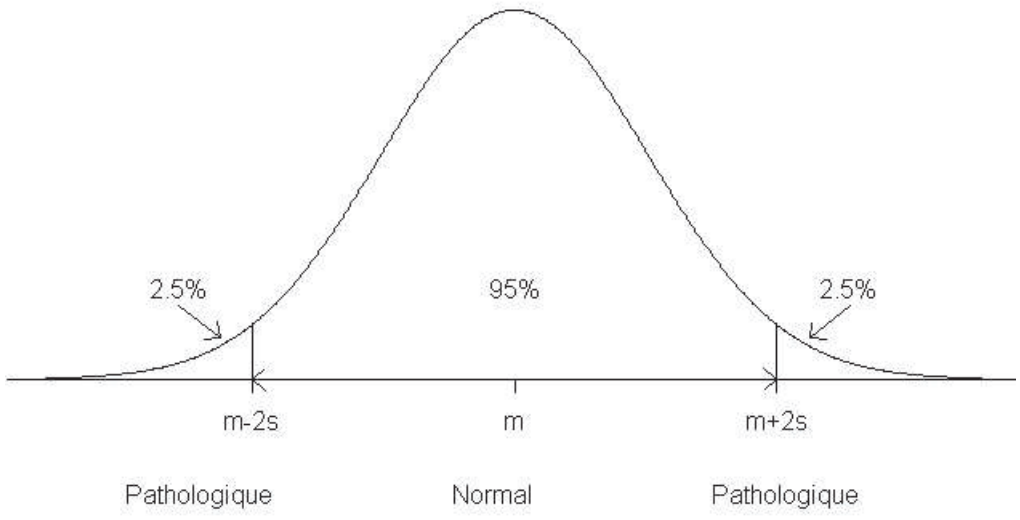

Figure 3.1 Détermination d'un intervalle de de référence contenant 95% des observations pour une distribution Normale

A titre d'exemple, les valeurs de référence de l'urée (mmol/L) ont été déterminées comme suit. Un échantillon de 284 sujets présumés en bonne santé ont subi une prise de sang afin de doser la concentration d'urée. La moyenne et l'écart-type obtenus valent respectivement $\bar{x} = 5.1$ mmol/L et $s = 1.1$ mmol/L. En utilisant la formule (3.9), on

détermine les limites de l'intervalle de référence, soient $5.1 \pm 2 \times 1.1 = 5.1 \pm 2.2$. En conclusion, l'intervalle de référence pour l'urée vaut $[2.9 - 7.3]$ mmol/L. Ce sont ces fourchettes de référence que l'on trouve sur les protocoles de laboratoire.

Lorsqu'un individu se présente au laboratoire pour un dosage d'urée, le résultat peut être considéré comme "normal" s'il tombe dans l'intervalle de référence, sinon il est "pathologique". En effet, il est très rare que l'on tombe en dehors de l'intervalle (seulement 2.5% des cas en-dessous et 2.5% des cas au-dessus) lorsqu'on est en bonne santé. Il est donc probable que l'individu présente une perturbation au niveau physiologique.

Notons enfin que le recours à la formule (3.9) peut s'avérer catastrophique si la distribution des données s'écarte nettement d'une loi Normale.

3.6 Standardisation des données

Soit un échantillon d'effectif n $\{x_1, \ldots, x_n\}$ de moyenne $\bar{x}$ et d'écart-type s.
L'observation transformée

$$z_i = \frac{x_i - \bar{x}}{s} \qquad (i = 1, \ldots, n) \tag{3.10}$$

est appelée "valeur centrée réduite". Elle mesure la distance de l'observation x_i à la moyenne $\bar{x}$ en termes d'écart-type s. Ainsi, si $z_i = \pm 2$, l'observation x_i est à ± 2 écarts-types de la valeur moyenne. Elle se situe donc sur la limite inférieure ou supérieure de l'intervalle de référence.

Il est facile de montrer que $\bar{z} = 0$ et $s_z = 1$.

La valeur z_i peut donc être considérée comme une distance standardisée à la moyenne. En calculant z_i, on voit immédiatement si l'observation se situe à l'intérieur ou à l'extérieur de l'intervalle de référence, et de quel côté de la moyenne.

En termes d'interprétation biologique, on pourrait dire qu'un sujet dont la valeur $|z_i|$ est élevée est "atypique" de la population. Ainsi, seulement 5% des individus de référence ont une valeur supérieure à $|z_i| = 2$ et seulement 3 sujets sur mille ont une valeur supérieure à $|z_i| = 3$.

Pour reprendre l'exemple de l'urée, un sujet présentant une valeur d'urée de 8.2 mmol/L a une valeur z égale à $(8.2 - 5.1)/1.1 = 2.82$, valeur qui n'est dépassée que par 2 sujets sains sur 1000 (utilisation de la table de la loi Normale). On peut donc considérer cette valeur comme très pathologique.

3.7 Coefficient de variation d'une technique

Le calcul de la moyenne et de l'écart-type se révèle aussi utile pour déterminer la précision d'une technique ou d'un appareil de mesure (ou de dosage).

Par définition, le coefficient de variation (exprimé en pourcent) est donné par la formule (on suppose $\bar{x} > 0$)

$$CV = \frac{s}{\bar{x}} \times 100\%. \tag{3.11}$$

Le coefficient de variation (coefficient of variation) exprime donc la variabilité d'une série d'observations $\{x_1, \ldots, x_n\}$ en fonction de leur moyenne $\bar{x}$. Plus il est faible, plus la technique de mesure est reproductible, c'est-à-dire précise. Par contre, un coefficient de variation élevé correspond à une technique de dosage peu précise.

On définit la précision en termes de répétabilité (précision à court-terme) et de reproductibilité (précision à long-terme).

- La *répétabilité* (repeatability) d'une technique de mesure est la variabilité observée dans des dosages répétés d'un même spécimen sur une courte période sans recalibration de l'appareil.
- La *reproductibilité* (reproducibility) d'une technique de mesure est la variabilité observée dans des dosages répétés d'un même spécimen sur une longue période impliquant plusieurs recalibrations de l'appareil.

Illustrons ces deux concepts par un exemple. On a dosé à 10 reprises et consécutivement le glucose (mmol/L) dans un même échantillon de sang à l'aide d'un appareil de laboratoire. Les résultats obtenus sont 3.21, 3.22, 3.20, 3.25, 3.22, 3.23, 3.24, 3.21, 3.22, 3.21 mmol/L. La moyenne vaut 3.22 mmol/L et l'écart-type 0.015 mmol/L. En conséquence, le coefficient de variation (répétabilité) de l'appareil vaut 0.47%, valeur inférieure à 1%. Par contre, en dosant le glucose dans le même échantillon de sang durant 10 jours consécutifs, on a obtenu les résultats suivants : 3.15, 3.22, 3.32, 3.27, 3.28, 3.12, 3.19, 3.17, 3.36, 3.14. La moyenne et l'écart-type valent respectivement $\bar{x} = 3.22$ mmol/L et $s = 0.082$ mmol/L. Donc le coefficient de variation (reproductibilité) s'élève à 2.5%.

3.8 Contrôle de qualité

Une autre application de la moyenne et de l'écart-type concerne le domaine du contrôle de qualité (quality control) dans les laboratoires de biologie clinique, par exemple.

On sait, en effet, que si un processus analytique (de mesure ou de dosage) est "sous contrôle", c'est-à-dire qu'il fluctue autour d'une moyenne $\bar{x}$ avec une dispersion égale à s, 95% des valeurs observées tombent dans les limites $\bar{x} \pm 2s$ (appelées *limites d'avertissement*) et 99.7% des valeurs tombent dans les limites $\bar{x} \pm 3s$ (appelées *limites d'action*).

En contrôle de qualité, trois cas de figure peuvent se présenter selon la position de la valeur fournie par le processus à un moment donné, soit x_t.

1. x_t tombe dans l'intervalle $\bar{x} \pm 2s$. On dit que le processus est sous contrôle.
2. x_t tombe en dehors des limites $\bar{x} \pm 2s$ mais endéans des limites $\bar{x} \pm 3s$. On dit qu'il y a avertissement et on suspecte un problème dans le processus.
3. x_t tombe en dehors des limites $\bar{x} \pm 3s$. On dit que le processus est hors-contrôle et il faut agir pour rectifier la situation.

En général, les valeurs $\bar{x}$ et s sont calculées à partir d'un échantillon de n valeurs observées au cours du temps. On établit alors une "carte de contrôle" (quality control

chart) avec des limites à $\bar{x} \pm 2s$ et à $\bar{x} \pm 3s$, sur laquelle on reporte les observations journalières x_t (voir Figure 3.2). Cette carte de contrôle est appelée "Carte de contrôle de Levey-Jennings".

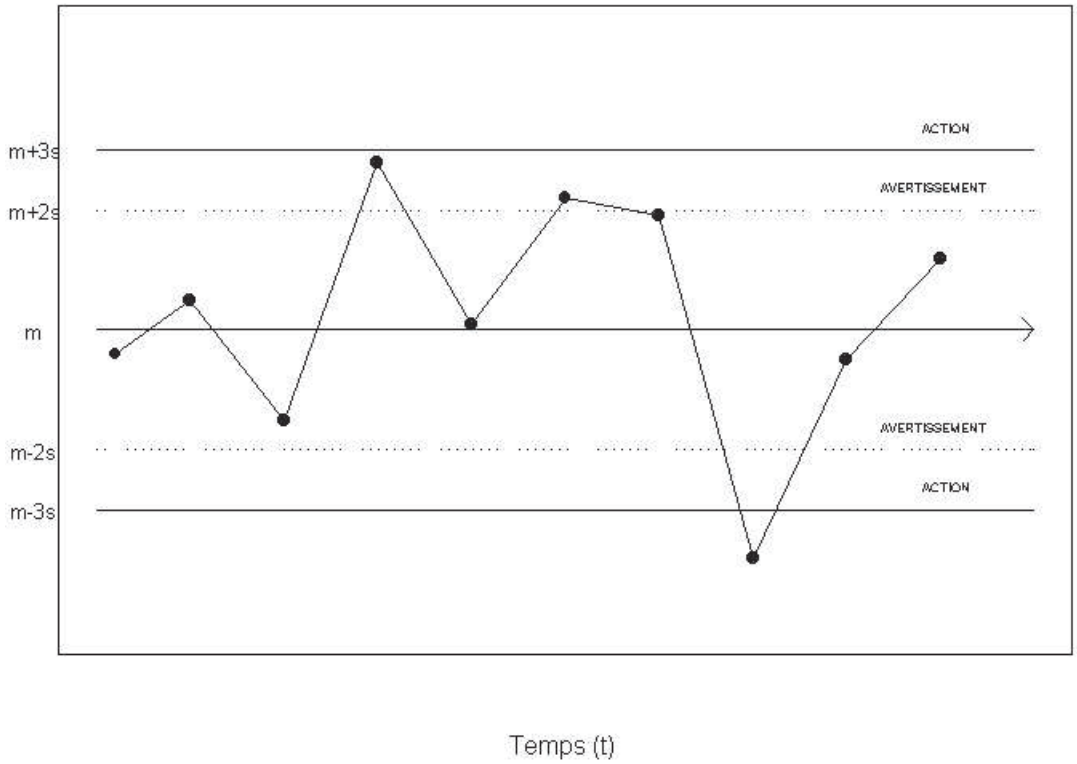

Figure 3.2 Carte de Contrôle de Levey-Jennings avec ses limites d'avertissement et d'action. Les points reportés sur la carte sont par exemple les observations quotidiennes

Chapitre 4

Percentiles

4.1 Introduction

La moyenne et l'écart-type sont les deux principaux paramètres de position et de dispersion utilisés en pratique quotidienne et dans la recherche scientifique. Toutefois, comme signalé au Chapitre 3, ils sont sensibles aux valeurs aberrantes. Par ailleurs, pour certains types de distribution statistique, ils ne sont pas nécessairement les paramètres les plus appropriés pour caractériser la valeur centrale et la variabilité des observations. Dans ce chapitre, nous décrivons d'autres paramètres de position et de dispersion.

4.2 Paramètres de position

Les paramètres de position (location parameters) ont pour but de caractériser la "valeur centrale" de la distribution des valeurs. Outre la moyenne arithmétique définie au Chapitre 3, on distingue le mode, la médiane et le percentile.

4.2.1 Mode

Le mode (mode en anglais) d'un échantillon de données est la valeur la plus fréquente dans l'échantillon. Par exemple, pour une variable discrète, le mode correspond à la valeur a_i pour laquelle f_i est maximum.

Dans le cas d'une variable qualitative, le mode correspond à la modalité m_i la plus fréquente. Pour une variable continue, le mode correspond à la classe la plus fréquente. On parle alors de "classe modale".

Le mode peut ne pas exister, par exemple lorsque toutes les valeurs sont différentes. Au contraire, il peut y avoir un seul mode (distribution unimodale) ou plusieurs modes (distribution multimodale).

A titre d'exemple, pour l'échantillon des 133 ménages, le mode est égal à 3 enfants par ménage car c'est la valeur la plus fréquente. Dans le cas des patients admis en hospitalisation, la classe modale correspond à la classe 0-10 ans. On distingue cependant

dans l'histogramme (voir Figure 2.7) d'autres classes modales, par exemple 20-30 ans, 50-60 ans et 80-90 ans. En ce qui concerne l'enquête sur les médecins généralistes (Annexe I), l'université d'origine la plus fréquente est l'UCL (48%).

4.2.2 Médiane

La médiane (median) est la valeur centrale par excellence car elle divise la distribution en deux parties égales. On la note M. Par définition, la médiane d'un échantillon est la valeur qui laisse 50% des observations en-dessous et 50% des observations au-dessus. On l'appelle aussi "Percentile 50" noté P50 (voir 4.2.3).

On peut déterminer la médiane graphiquement à partir du diagramme cumulatif (variables discrètes et continues). Il suffit de tracer une horizontale à hauteur de 50% de l'ordonnée et de recherche l'abscisse du point d'intersection entre cette l'horizontale et le diagramme cumulatif. A titre d'exemple, pour l'âge des patients hospitalisés, on trouve approximativement $M = 36$ ans (Figure 4.1).

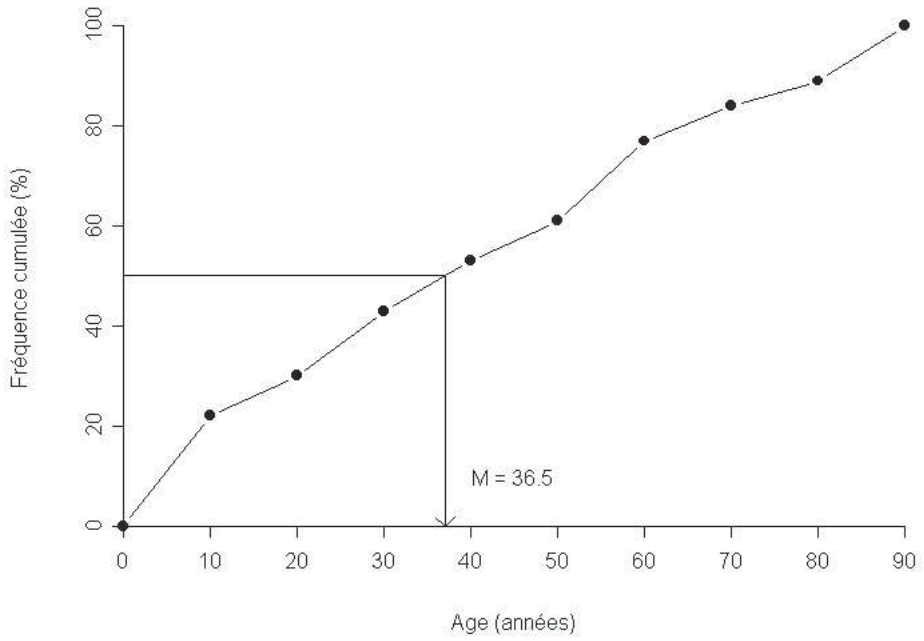

Figure 4.1 Détermination de la médiane à partir du diagramme cumulatif.

La détermination de la médiane à partir de l'histogramme est beaucoup plus difficile, puisqu'il faut trouver le point qui divise l'histogramme en deux parties d'aires égales.

La médiane peut aussi se calculer numériquement à partir des données de l'échantillon $\{x_1, \ldots, x_n\}$. On procède comme suit :

– Trier l'échantillon par ordre croissant, ce que l'on note

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}.$$

Le nombre entre parenthèses en indice est appelé le "rang" (rank) de l'observation dans la série triée par ordre croissant, c'est-à-dire sa position dans la série. Ainsi, $x_{(1)}$ est l'observation de rang 1 (la plus petite), $x_{(2)}$ l'observation de rang 2 (la 2e plus petite), $\ldots, x_{(n)}$ l'observation de rang n (la plus grande).

– Si n est impair, alors la médiane est l'observation de rang $(n+1)/2$

$$M = x_{(\frac{n+1}{2})} \tag{4.1}$$

– Si n est pair, alors la médiane est la moyenne arithmétique des observations de rang $n/2$ et $(n/2)+1$. En clair,

$$M = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}. \tag{4.2}$$

La médiane est une caractéristique particulièrement robuste puisqu'elle ne fait pas intervenir les valeurs extrêmes de la distribution. En effet, on ne se sert que de l'observation du milieu de la distribution au travers des rangs. Quelle que soit la valeur la plus élevée ou la moins élevée de la série, la médiane reste inchangée.

Dans le cas d'une variable discrète, vu le caractère discret des valeurs, la médiane n'a pas toujours de sens. Dans le cas du nombre d'enfants par ménage, puisqu'il y a 133 observations (nombre impair), la médiane est l'observation de rang $134/2 = 67$, soit $M = 3$, confirmant le caractère central de cette valeur.

Dans le cas de l'âge des patients admis en hospitalisation, $n = 100$ est pair et il convient d'employer la formule (4.2). En triant les observations par ordre croissant, on constate que l'observation de rang 50 est 36 et celle de rang 51 vaut 37. Par conséquent, $M = (36+37)/2 = 36.5$ ans. Pour rappel, la moyenne valait 39.1 ans.

Remarque : La comparaison de la moyenne arithmétique et de la médiane est intéressante à plusieurs égards. Trois cas de figure se présentent :

– la moyenne est voisine de la médiane ($\bar{x} \simeq M$) : ceci indique une certaine symétrie de la distribution ;
– la moyenne est sensiblement supérieure à la médiane ($\bar{x} >> M$) : ceci peut indiquer la présence de valeurs aberrantes élevées dans l'échantillon ou au contraire signifie que la distribution est dissymétrique à droite (c'est le cas des durées de vie, par exemple) ;
– la moyenne est sensiblement inférieure à la médiane ($\bar{x} << M$) : ceci peut suggérer la présence de valeurs aberrantes basses dans l'échantillon ou au contraire signifie que la distribution est dissymétrique à gauche (situation plus rare en pratique).

4.2.3 Percentile ou Quantile

Par définition, le percentile $P\alpha$ ou "quantile d'ordre α" est la valeur de l'échantillon qui laisse α% des observations en-dessous et $(100-\alpha)$% au-dessus. On voit immédiatement que la médiane est le percentile $P50$. En théorie, $0 \leq \alpha \leq 1$ mais on l'exprime généralement en pourcent.

A titre d'exemple, le percentile $P25$ divise l'histogramme en deux parties dont les aires valent respectivement 25% et 75%. De même, pour le percentile $P75$, les aires valent 75% et 25%, respectivement. Les percentiles les plus utilisés sont les "quartiles" $P25$, $P50$ (médiane) et $P75$ car ils divisent la distribution en 4 parties d'aire égale à 25% chacune. Les "déciles" $P10$, $P20$, ..., $P80$, $P90$ sont aussi caractéristiques car ils divisent l'histogramme en 10 parties d'aire égale à 10% chacune.

Notons que

$$\begin{array}{rcl} P0 & = & x_{(1)} = \min\{x_1, \ldots, x_n\} \\ P100 & = & x_{(n)} = \max\{x_1, \ldots, x_n\}. \end{array} \tag{4.3}$$

Les percentiles ont l'avantage d'être robustes, car peu sensibles aux valeurs extrêmes.

On peut déterminer les percentiles à partir du diagramme cumulatif comme on l'a vu pour la médiane. Toutefois, en pratique, on utilise la formule suivante ($0 \leq \alpha \leq 1$)

$$P\alpha = x_{(r_1)} + [x_{(r_1+1)} - x_{(r_1)}] \times (r - r_1) \tag{4.4}$$

où $r = (n+1) \times \alpha$ et r_1 est l'entier directement inférieur ou égal à r, soit $r_1 \leq r$.

Pour les quartiles $P25$ et $P75$, il est préférable d'utiliser la procédure suivante. Soient r_1 et r_2 les deux plus grands entiers tels que $r_1 \leq (n+1)/2$ et $r_2 \leq r_1/2$. On calcule $r_3 = r_2 + 1$ et $r_4 = r_1 - r_2$, $r_5 = n + 1 - r_1 + r_2$ et $r_6 = n - r_2$. Dans ces conditions,

$$\begin{array}{rcl} P25 & = & (x_{(r_3)} + x_{(r_4)})/2 \\ P75 & = & (x_{(r_5)} + x_{(r_6)})/2 \end{array} \tag{4.5}$$

Remarque. Dans de nombreux logiciels statistiques, on résume un échantillon de données $\{x_1, \ldots, x_n\}$ par cinq quantiles, $P0$, $P25$, $P50$, $P75$ et $P100$, c'est-à-dire les valeurs extrêmes et les trois quartiles. Ces cinq quantités permettent de construire un graphique, appellé "boxplot".

4.3 Paramètres de dispersion

Les paramètres de dispersion ont pour but de mesurer la variabilité des observations d'un échantillon. Un indicateur de dispersion est un nombre non négatif (≥ 0) d'autant plus élevé que la variabilité dans un échantillon est grande. Lorsque l'indicateur de dispersion est nul, toutes les valeurs de l'échantillon sont égales et réciproquement.

En plus de l'écart-type introduit au Chapitre 3, on distingue l'étendue, la variance, et l'écart interquartiles.

4.3.1 Etendue

L'étendue ou amplitude (range) est la différence entre la plus grande et la plus petite observation de l'échantillon.

$$E = x_{(n)} - x_{(1)}. \tag{4.6}$$

Elle est peu utilisée vu son caractère particulièrement sensible aux valeurs extrêmes. Dans un rapport statistique, cependant, il est classique de rapporter les valeurs extrêmes de l'échantillon.

4.3.2 Variance

Comme on l'a vu au Chapitre 3, la variance est le carré de l'écart-type. Inversement l'écart-type est la racine carrée positive de la variance. La variance joue un rôle fondamental en théorie mais est peu utilisée en pratique dans la présentation des résultats (sauf dans les tables d'analyse de la variance).

Par définition,

$$s^2 = \frac{\sum(x - \bar{x})^2}{(n-1)} \tag{4.7}$$

Une variance est toujours une "somme de carrés" divisée par des "degrés de liberté". Les unités de la variance sont les unités de la variable au carré, ce qui rend son interprétation difficile. Pour rappel, dans le cas d'une variable binaire, $s^2 = p(1-p)$. La variance est particulièrement sensible aux valeurs extrêmes vu la mise au carré des écarts.

4.3.3 Ecart interquartile

Par définition, l'écart interquartile (interquartile range ou H-spread) est donné par l'équation

$$H = P75 - P25. \tag{4.8}$$

Utilisant les 1er et 3ème quartiles, H est un paramètre de dispersion peu sensible aux valeurs aberrantes et donc robuste. Il est communément utilisé dans plusieurs applications de contrôle de qualité.

Par définition, $H \geq 0$ et, plus il est élevé, plus la distribution des données est dispersée.

Dans le cas d'une distribution symétrique approximativement Normale (gaussienne), on peut calculer l'écart-type à partir de H en utilisant la formule

$$s = 0.74H \tag{4.9}$$

A titre d'exemple, l'Annexe II reproduit les résultats du dosage du glucose (mmol/L) sur un même échantillon contrôle envoyé par le Ministère de la Santé publique à tous les laboratoires de Biologie clinique de Belgique ($n = 545$). Les résultats sont triés par ordre croissant. On constate la présence de plusieurs valeurs aberrantes. La moyenne (Eq. 3.1) et l'écart-type (Eq. 3.5) valent respectivement $\bar{x} = 3.97$ mmol/L et $s = 14.624$ mmol/L, conduisant à un coefficient de variation (Eq. 3.11) de 368% !

Le recours aux formules précédentes par contre conduit aux résultats suivants :

$$\begin{aligned} P25 &= 3.08 \text{ mmol/L} \\ P50 &= 3.22 \text{ mmol/L} \\ P75 &= 3.44 \text{ mmol/L} \end{aligned}$$

d'où $H = 3.44 - 3.08 = 0.36$ et $s = 0.74 \times 0.36 = 0.27$ mmol/L. On peut donc conclure que la concentration réelle de glucose dans l'échantillon vaut 3.22 mmol/L et que la variabilité entre les laboratoires exprimée en termes de coefficient de variation vaut

$$CV = \frac{0.27}{3.22} \times 100\% = 8.3\%.$$

On se rend ainsi compte de l'effet désastreux des valeurs aberrantes sur le calcul de la valeur centrale et de la variabilité des résultats.

4.4 Paramètres de forme

A titre d'information, il convient de compléter l'étude de la distribution statistique d'une variable par les paramètres de forme pour lesquels des puissances d'ordre 3 et 4 deviennent nécessaires. Ces paramètres entraînent des calculs longs et fastidieux, une sinécure pour les ordinateurs actuels.

4.4.1 Coefficient d'asymétrie

Le coefficient d'asymétrie (skewness), appelé aussi premier coefficient de Fisher, est défini par l'expression

$$g_1 = \frac{\sum z^3}{n} \tag{4.10}$$

où $z_i = (x_i - \bar{x})/s$, les valeurs centrées réduites définies au Chapitre 3. Trois cas de figures se présentent :

- $g_1 \simeq 0$: distribution symétrique ($\bar{x} \simeq M$)
- $g_1 >> 0$: distribution dissymétrique à droite ($\bar{x} >> M$)
- $g_1 << 0$: distribution dissymétrique à gauche ($\bar{x} << M$).

4.4.2 Coefficient d'aplatissement

Le coefficient d'aplatissement (kurtosis), appelé aussi deuxième coefficient de Fisher, est défini par l'expression

$$g_2 = \frac{\sum z^4}{n} - 3. \tag{4.11}$$

On distingue trois cas de figures :

- $g_2 \simeq 0$: distribution standard gaussienne (mésocurtique)
- $g_2 >> 0$: distribution trop pointue par rapport à la gaussienne (leptocurtique)
- $g_2 << 0$: distribution trop plate par rapport à la gaussienne (platycurtique)

4.5 Intervalle de référence non paramétrique

Nous avons vu au paragraphe 3.5 comment construire un intervalle de référence (ou de tolérance) à partir de la moyenne et de l'écart-type en calculant simplement $\bar{x} \pm 2s$. Toutefois, cette méthode suppose que la distribution est approximativement Normale (gaussienne). Lorsque ce n'est pas le cas, on peut utiliser une approche non paramétrique en se basant sur les percentiles.

En effet, dans le cas d'une distribution Normale, on sait qu'il y a 2.5% des observations en-dessous de la limite inférieure $\bar{x} - 2s$ et 2.5% des observations au-dessus de la limite supérieure $\bar{x} + 2s$. Ceci fait immédiatement penser aux percentiles $P2.5$ et $P97.5$. En effet, par définition, le percentile $P2.5$ laisse 2.5% des observations en-dessous tandis que le percentile $P97.5$ laisse 2.5% des observations au-dessus. En conséquence, pour une loi Normale $P2.5 \simeq \bar{x} - 2s$ et $P97.5 \simeq \bar{x} + 2s$. En réalité, en théorie, le coefficient 2 doit être remplacé par 1.96 mais cela modifie peu les valeurs finales.

Si la distribution n'est pas Normale, on peut continuer à utiliser les percentiles. Par conséquent, l'intervalle de référence à 95% peut être défini par les percentiles $P2.5$ et $P97.5$ puisque 95% des observations tombent entre ces deux limites.

$$\text{Intervalle de référence à } 95\% = [P2.5 - P97.5]. \tag{4.12}$$

En biologie clinique, il arrive que les valeurs basses de certains constituants biochimiques soient "normales" et que le caractère "anormal" ne se situe que vers les valeurs élevées. C'est le cas pour les enzymes et les hormones. Dans ce contexte, on définit un intervalle de référence allant de 0 à $P95$, laissant 5% des valeurs au-dessus et 0% en-dessous. Il suffit donc de calculer le percentile $P95$.

Tableau 4.1 Valeurs d'alanine amino-transférase (ALT) en UI/L chez 240 étudiants en médecine (University of Virginia, Charlottesville, USA). Les valeurs sont triées par ordre croissant.

5	6	6	6	7	8	8	8	8	8	9	9	9	10	10	10	10	11	11	11
11	11	11	11	11	11	11	11	12	12	12	12	12	12	12	12	12	12	12	12
12	13	13	13	13	13	13	13	13	13	13	13	13	13	14	14	14	14	14	14
14	14	14	14	14	14	14	15	15	15	15	15	15	15	15	15	15	16	16	16
16	16	16	16	16	16	16	16	17	17	17	17	17	17	17	17	17	18	18	18
18	18	18	18	18	18	18	19	19	19	19	19	19	19	19	19	19	19	19	19
20	20	20	20	20	20	20	20	20	20	20	20	20	20	20	21	21	21	21	21
21	21	21	21	21	21	22	22	22	22	22	22	22	22	23	23	23	23	23	24
24	24	25	25	25	25	25	25	25	25	25	25	25	26	26	26	26	26	27	28
28	28	28	28	28	29	29	30	30	30	30	30	31	31	31	31	31	32	33	34
34	35	35	36	36	36	36	36	36	37	37	37	38	38	39	39	39	40	40	40
41	42	45	45	46	47	47	48	48	49	51	51	51	53	54	55	55	62	65	69

Le Tableau 4.1 fournit les valeurs d'alanine amino-transférase (ALT ou TGP) en unités internationales par litre (UI/L) mesurées chez 240 étudiants en médecine (120

hommes et 120 femmes) de la University of Virginia (USA) en 1987-1988. En utilisant la formule (4.4) pour $P95$, on a $r = 241 \times 0.95 = 228.95$, de sorte que $r_1 = 228$ et $P95 = 48 + (48 - 48)(228.95 - 228) = 48$ UI/L. On peut donc conclure que l'intervalle de référence pour l'ALT s'écrit : $[0 - 48]$ UI/L. Toute valeur supérieure à 48 UI/L sera considérée comme "pathologique".

4.6 Courbes de percentiles

Les courbes de percentiles sont couramment utilisées en biométrie et en médecine, notamment pour les courbes de croissance. Ces courbes de croissance sont obtenues en calculant pour chaque âge des percentiles caractéristiques : $P10$, $P25$, $P50$, $P75$, $P90$ mais aussi parfois $P3$ et $P97$.

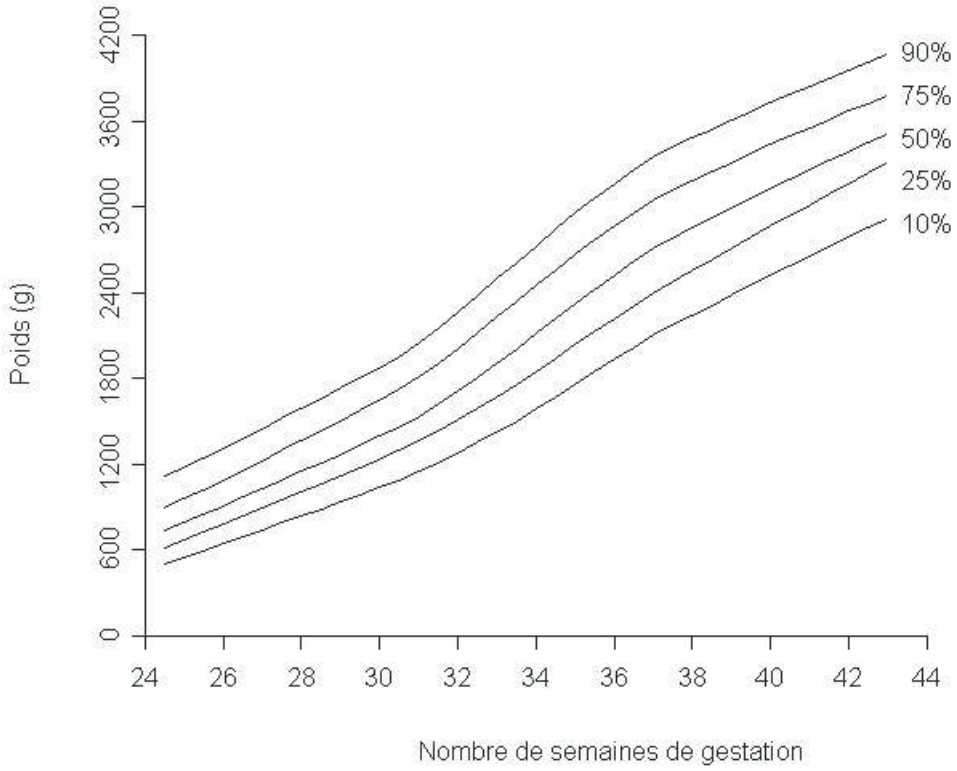

Figure 4.2 Courbes de Lubchenco pour le poids (g) en fonction du nombre de semaines de gestation.

A ce titre, les plus célèbres sont les courbes de Lubchenco (voir Figure 4.2) utilisées en pédiatrie pour caractériser différents paramètres du nouveau-né en fonction du nombre de semaines de gestation (par exemple, la taille, le poids, le périmètre crânien ou l'index pondéral).

Si, pour un âge gestationnel donné, on se situe à l'intérieur des courbes extrêmes, la croissance est normale. Tout dépassement vers le haut ou vers le bas est signe d'un développement atypique le plus souvent anormal (excès ou retard de croissance).

4.7 Normalisation d'une distribution

En statistique, une distribution Normale (gaussienne) présente de nombreux avantages par rapport aux autres distributions. En effet, elle est caractérisée par seulement deux paramètres, la moyenne $\bar{x}$ et l'écart-type s. Elle a une allure symétrique (en forme de cloche) et donc unimodale avec une diminution rapide des fréquences de part et d'autre de la moyenne. En outre, le caractère Normal d'une distribution est souvent une condition indispensable pour poursuivre l'analyse statistique inférentielle (tests statistiques). On peut dès lors comprendre les efforts consentis par les statisticiens pour "Normaliser" une distribution qui n'est pas Normale.

La transformation couramment utilisée pour normaliser une distribution dissymétriques à droite (durées de vie, par exemple) est la fonction logarithmique (Népérien). En clair, si on calcule $y_i = \ell n\, x_i$ $(i = 1, \ldots, n)$, il arrive qu'on puisse rendre symétrique cette distribution et la rapprocher d'une distribution Normale. On définit ainsi l'échantillon transformé $\{y_1, \ldots, y_n\}$ et tous les calculs s'effectuent sur ces valeurs transformées. Dès lors, on peut calculer la moyenne et l'écart-type des observations y_i, soit

$$\begin{aligned} \bar{y} &= \sum y/n \\ s_y &= \sqrt{\frac{\sum y^2 - (\sum y)^2/n}{n-1}} \end{aligned}$$

Puisque la distribution des y_i est symétrique, la médiane M_y est proche de $\bar{y}$ et les coefficients de Fisher g_1 et g_2 sont tous deux voisins de 0.

Notons que si on prend l'exponentielle (transformation inverse) de $\bar{y}$, on définit ce que l'on appelle la "moyenne géométrique" notée m_g et on a

$$m_g = e^{\bar{y}} = e^{\sum \ell n\, x/n} \tag{4.13}$$

qui est proche de la médiane M_x des observations x_i $(i = 1, \ldots, n)$.

De même, on peut déterminer un intervalle de référence à 95% pour la variable X en calculant les limites inférieure $(\bar{y} - 2s_y)$ et supérieure $(\bar{y} + 2s_y)$ et en effectuant la transformation inverse

$$L_{\text{inf}} = e^{\bar{y}-2s_y} \quad \text{et} \quad L_{\text{sup}} = e^{\bar{y}+2s_y}. \tag{4.14}$$

Il n'est pas toujours possible de "Normaliser" une distribution statistique observée. D'autres transformations existent comme la transformation $y = \sqrt{x}$ ou $y = \ell n(x + c)$ et $y = \sqrt{x + c}$ lorsque x peut être nul.

La transformation de normalisation de Box et Cox est souvent utilisée par les spécialistes mais dépasse le cadre d'un cours d'introduction. Elle est définie par l'équation

$$\begin{aligned} y(\lambda) &= \frac{x^\lambda - 1}{\lambda} \quad \text{pour } \lambda \neq 0 \\ &= \ell n\, x \quad \text{pour } \lambda = 0 \end{aligned} \tag{4.15}$$

où λ est un paramètre qu'il faut estimer à partir des données par la méthode dite du "maximum de vraisemblance". Appliquée aux observations, cette transformation permet de réduire toute dissymétrie, qu'elle soit positive $(\lambda \geq 0)$ ou négative $(\lambda < 0)$.

A titre d'exemple, l'utilisation de la méthode de Box-Cox aux données de ALT (voir Tableau 4.1) conduit à une valeur $\lambda = -0.052$, ce qui signe une transformation logarithmique $y_i = \ell n(x_i + c)$ avec $c = -1.6$. La moyenne $\bar{y}$ vaut 2.89 et $s_y = 0.545$. Puisque pour une loi Normale $P95 = \bar{y} + 1.645 s_y$, on obtient $P95 = 2.89 + 1.645 \times 0.545 = 3.787$. En repassant sur l'échelle des valeurs originales par la transformation inverse, on a $P95 = e^{3.787} + 1.6 = 46$ UI/L, valeur quelque peu inférieure au seuil de 48 UI/L trouvé précédemment (voir 4.4.3).

Chapitre 5

Courbes de survie

5.1 Introduction

Les variables "durée de vie" (failure, lifetime variable) sont fréquentes en recherche biomédicale. On parle en statistique de l'analyse des "données de survie" (survival data analysis). L'exemple le plus connu est la survie des patients atteints d'un cancer ou de toute autre maladie grave (par exemple, le SIDA). Il y a cependant bien d'autres situations où de telles données sont récoltées. Par exemple, les économistes de la Santé sont intéressés par la durée d'hospitalisation des malades, en particulier en matière de financement des soins de Santé. Les compagnies d'assurance sont préoccupées par la durée d'invalidité physique de leurs affiliés suite à un accident ou une maladie. La chirurgie cardiaque s'intéresse à la durée de vie d'une valve cardiaque artificielle, l'ophtalmologue à la durée de vie d'un implant intra-oculaire à la suite d'une intervention de la cataracte, le transplanteur au temps de rejet d'un organe greffé. Mais le domaine couvre aussi par exemple la durée d'une intervention chirurgicale, la durée de séjour en soins intensifs, la récurrence d'une affection, le temps de rechute d'un cancer ou l'intervalle entre deux poussées d'une maladie chronique (par exemple, le lupus érythémateux).

D'un point de vue statistique, l'analyse des durées de vie occupe actuellement une position dominante dans la littérature, notamment par le biais des essais cliniques (clinical trials) et des courbes de survie de Kaplan-Meier. Il est dit que l'article de Kaplan-Meier (1958) est un des plus cités de la littérature scientifique en général.

5.2 Durée de vie

Une variable de durée de vie, notée T pour la circonstance, se différencie des autres variables par au moins trois caractéristiques.

5.2.1 Positivité

$T \geq 0$ est une variable non-négative.

5.2.2 Dissymétrie à droite

En général, la distribution statistique de T présente une nette dissymétrie à droite (positive skewness). La moyenne est supérieure à la médiane et le coefficient d'asymétrie g_1 est positif. On ne peut donc la traiter comme une variable symétrique Normale mais il arrive qu'une transformation de type logarithmique $\ell n\, T$ permette de la normaliser. Le caractère dissymétrique provient du fait que certains sujets (ou objets) de la population ont une durée de vie anormalement longue. On a déjà évoqué l'hospitalisation de longue durée d'un traumatisé de la route alors que la majorité des patients séjournent 8-10 jours à l'hôpital. Certains patients atteints d'un cancer grave peuvent continuer à vivre longtemps en dépit de leur cancer.

L'histogramme d'une variable de survie montre donc typiquement un allongement vers la droite.

5.2.3 Censure

La caractéristique la plus importante cependant des données de durée de vie est la censure (à droite en général). Ce point a été évoqué au Chapitre 1. Pour rappel, une censure (censoring) survient lorsque la valeur $T = t$ d'un individu ne peut être observée. On peut cependant affirmer que $T > t^*$, où t^* est le dernier temps où l'individu a été vu "en vie". On dit que t^* est une donnée censurée. La durée de vie réelle de ce sujet est au moins supérieure à t^*.

La censure d'une observation est indiquée de différente manière. Dans ce chapitre, nous utilisons l'astérisque pour dire que la donnée est censurée. En clair, une donnée non censurée n'aura pas d'astérisque.

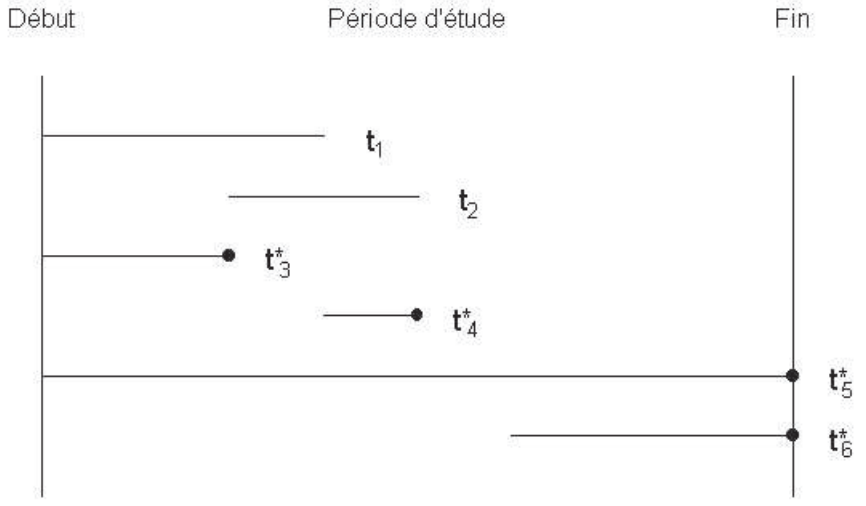

Figure 5.1 Durées de vie observées et censurées durant une étude clinique. Les observations censurées sont marquées d'une astérisque (t représente la durée de vie depuis l'entrée du patient dans l'étude).

La censure peut se produire de plusieurs façons. Par exemple, dans une étude sur le cancer, le patient peut être "perdu de vue" ou décédé d'une autre cause non liée à son cancer. Dès lors, la seule information disponible sur T est la dernière fois qu'il a été vu en vie. La censure peut être provoquée. Par exemple, une étude prospective peut être prématurément clôturée avant que tous les sujets (ou objets) ne soient décédés. Ceux qui sont en vie à ce moment ont des durées de vie censurées.

La Figure 5.1 illustre les différentes possibilités de censures dans un essai clinique limité dans le temps et où les patients entrent de façon aléatoire et sont suivis jusqu'au décès, à moins qu'une censure ne survienne. Les patients 1 et 2 meurent au cours de l'étude bien qu'entrés à des moments différents ; dès lors $T = t_1$ et $T = t_2$ sont des durées de vie observées. Les patients 3 et 4 ont été "perdus de vue" (drop-out) durant l'étude bien qu'entrés à des moments différents ; dès lors, t_3^* et t_4^* sont des données de vie censurées (non observées). Leur durée de vie réelle vaut $T > t_3^*$ et $T > t_4^*$, respectivement. Les patients 5 et 6 ont été enrôlés dans l'étude mais on n'a pu observer leur décès parce que l'étude a été clôturée auparavant. Ils étaient "en vie" au moment de la fin de l'étude. Dès lors, $T > t_5^*$ et $T > t_6^*$.

5.2.4 Données de survie

Un échantillon d'effectif n de données de survie inclut en général des données censurées et on peut le noter de deux manières : $\{t_1, t_2^*, \ldots, t_i, \ldots, t_j^*, \ldots, t_n^*\}$ ou $\{(t_i, \delta_i), i = 1, \ldots, n\}$ où δ_i est un "indicateur de censure" égal à 0 si t_i est une valeur réellement observée et 1 si t_i est une valeur censurée.

Le calcul de la moyenne et de l'écart-type d'un échantillon contenant des données censurées est complexe et ne peut se faire à la main. Il faut avoir recours à un programme informatique. Il n'est pas recommandable d'éliminer les valeurs censurées, au risque d'obtenir des résultats erronés. Pour caractériser graphiquement la distribution d'une variable de survie, on utilise la méthode de Kaplan-Meier.

5.3 Courbe de survie de Kaplan-Meier

5.3.1 Courbe de survie théorique

Dans l'analyse statistique d'une variable de durée de vie T, ce qui importe c'est l'estimation de la courbe de survie, notée $S(t)$.

Par définition, la survie $S(t)$ à l'instant t est la probabilité (c'est-à-dire la chance) de vivre au moins jusqu'au temps t, ce que l'on note $P(T > t)$. En d'autres termes, c'est la probabilité de "décéder" après le temps $T = t$. On comprend aisément que (voir Figure 5.2).

- $S(0) = 1$
- $S(t)$ est décroissant, soit $S(t) \leq S(t')$ si $t > t'$
- $\lim_{t\to\infty} S(t) = 0$.

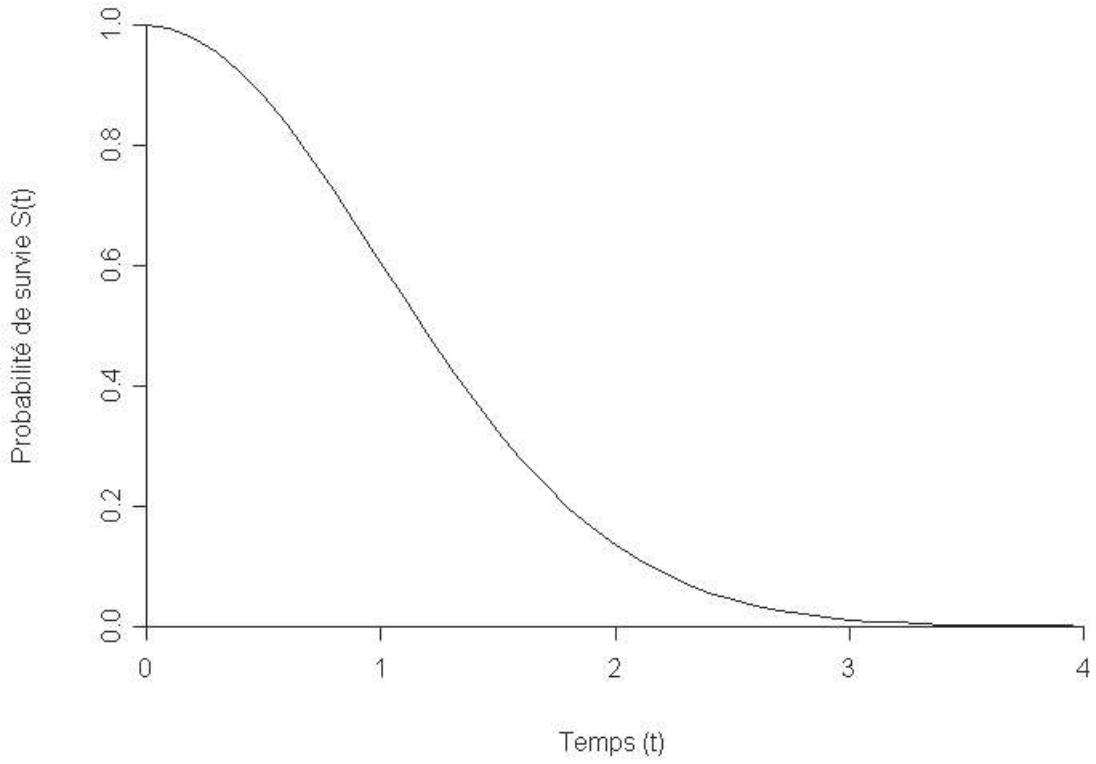

Figure 5.2 Courbe de survie théorique $S(t)$

5.3.2 Courbe de survie estimée

Pour estimer la courbe de survie $S(t)$ à partir d'un échantillon d'effectif n, on a recours à la méthode de Kaplan-Meier. Cette méthode procède comme suit :

- On trie les n données de survie par ordre croissant. Dans ce classement, une durée censurée *suit* toujours une donnée non censurée de même valeur.
- On ne retient que les k valeurs non censurées *distinctes*, notées $t_1 < t_2 < \cdots < t_{k-1} < t_k$.
- Pour chaque valeur t_i, on note ℓ_i le nombre de sujets toujours en vie juste avant ce moment et d_i le nombre de sujets qui décèdent juste en t_i. Attention, une durée censurée avant t_i n'est plus comptabilisée puisqu'on ignore ce qu'elle est devenue.
- Alors, la valeur de la courbe de survie en $T = t_i$, vaut

$$\hat{S}(t_i) = \prod_{j=1}^{i} \frac{\ell_j - d_j}{\ell_j} \quad (i = 1, \ldots, k) \tag{5.1}$$

ou encore

$$\hat{S}(t_i) = \hat{S}(t_{i-1}) \times \frac{\ell_i - d_i}{\ell_i} \tag{5.2}$$

avec $\hat{S}(t_0) = \hat{S}(0) = 1$ par convention.

– On répertorie les opérations précédentes dans un tableau par souci de clarté et de rigueur.

En résumé, on travaille par une formule de récurrence

$$\begin{aligned} \hat{S}(0) &= 1 \\ \hat{S}(t_1) &= 1 \times \frac{\ell_1 - d_1}{\ell_1} \\ \hat{S}(t_2) &= \hat{S}(t_1) \times \frac{\ell_2 - d_2}{\ell_2} \\ &\dots \end{aligned}$$

On constate que la courbe de survie n'est définie qu'aux points $t_1, \dots, t_k$. On reporte les valeurs $\hat{S}(t_i)$ pour chaque t_i et on trace des horizontales jusqu'au point suivant. On obtient donc une fonction en escaliers ! C'est la courbe de survie de Kaplan-Meier.

5.3.3 Exemple

Les données reprises à l'Annexe III décrivent les durées de vie (en mois), l'âge et le sexe, de patients atteints d'un adénocarcinome du rectum et répartis en deux groupes selon l'intensité du traitement radiothérapeutique pré-opératoire reçu, soit < 5000 rads (groupe 1), soit ≥ 5000 rads (groupe 2). Le premier groupe comporte 21 patients, le second 35 patients.

Déterminons la courbe de Kaplan-Meier du premier groupe de patients en procédant comme indiqué au paragraphe précédent. On constate qu'il y a 21 données dont 14 sont des durées de vie réelles (non censurées) et 7 sont des données censurées.

– Les données sont déjà triées par ordre croissant. Aucune donnée censurée n'est égale à une donnée non censurée.
– Parmi les 14 données non censurées, $k = 10$ sont distinctes, à savoir, 7, 9, 12, 19, 23, 24, 34, 41, 54, et 78.
– Juste avant la première durée de vie ($t_1 = 7$), il y a $\ell_1 = 21$ patients en vie ; en outre, un décès se produit à 7 mois, donc $d_1 = 1$. Dès lors en appliquant la formule (5.2), on obtient, sachant que $\hat{S}(0) = 1$,

$$\hat{S}(t_1) = 1 \times \frac{21 - 1}{21} = \frac{20}{21} = 0.9524.$$

Pour $t_2 = 9$, $\ell_2 = 20$ puisque un patient est décédé à 7 mois, et $d_2 = 1$ puisque un patient décède à 9 mois ; dès lors

$$\begin{aligned} \hat{S}(t_2) &= \hat{S}(t_1) \times \frac{20 - 1}{20} \\ &= 0.9524 \times \frac{19}{20} = 0.9048. \end{aligned}$$

Pour $t_3 = 12$, $\ell_3 = 19$ et $d_3 = 2$, d'où

$$\begin{aligned} \hat{S}(t_3) &= \hat{S}(t_2) \times \frac{19-2}{19} \\ &= 0.9048 \times \frac{17}{19} = 0.8096. \end{aligned}$$

On continue ainsi de suite jusqu'au terme des 10 durées de vie distinctes. Ces calculs sont résumés dans le Tableau 5.1. Notons qu'au temps $t_7 = 34$, le nombre de patients en vie juste avant ce temps est égal à $\ell_7 = 10$ car au temps précédent ($t_6 = 24$), il y avait eu quatre décès ($d_6 = 4$) et qu'entretemps, un patient a été perdu de vue ($t = 29^*$). Le même raisonnement s'applique au dernier temps $t_{10} = 78$.

Tableau 5.1 Courbe de survie de Kaplan-Meier pour le groupe de patients avec cancer du rectum et ayant reçu une radiothérapie pré-opératoire < 5000 rad.

Temps réel t_i	En vie ℓ_i	Décès d_i	Survie $\hat{S}(t_i)$
0	–	–	1
7	21	1	$1 \times 20/21 = 0.9524$
9	20	1	$0.9524 \times 19/20 = 0.9048$
12	19	2	$0.9048 \times 17/19 = 0.8096$
19	17	1	$0.8096 \times 16/17 = 0.7620$
23	16	1	$0.7620 \times 15/16 = 0.7144$
24	15	4	$0.7144 \times 11/15 = 0.5239$
34	10	1	$0.5239 \times 9/10 = 0.4715$
41	9	1	$0.4715 \times 8/9 = 0.4191$
54	8	1	$0.4191 \times 7/8 = 0.3667$
78	6	1	$0.3667 \times 5/6 = 0.3056$

La courbe de Kaplan-Meier correspondant au Tableau 5.1 est reprise à la Figure 5.3.

Les programmes informatiques actuels permettent de calculer une région de confiance pour la courbe de survie théorique $S(t)$. Sur la Figure 5.3, cette région est définie par les deux courbes en pointillé.

Notons enfin que les courbes de survie de Kaplan-Meier relatives aux patients traités par radiothérapie pré-opératoire < 5000 rad (groupe 1) et > 5000 rad (groupe 2) sont illustrées à la Figure 5.4.

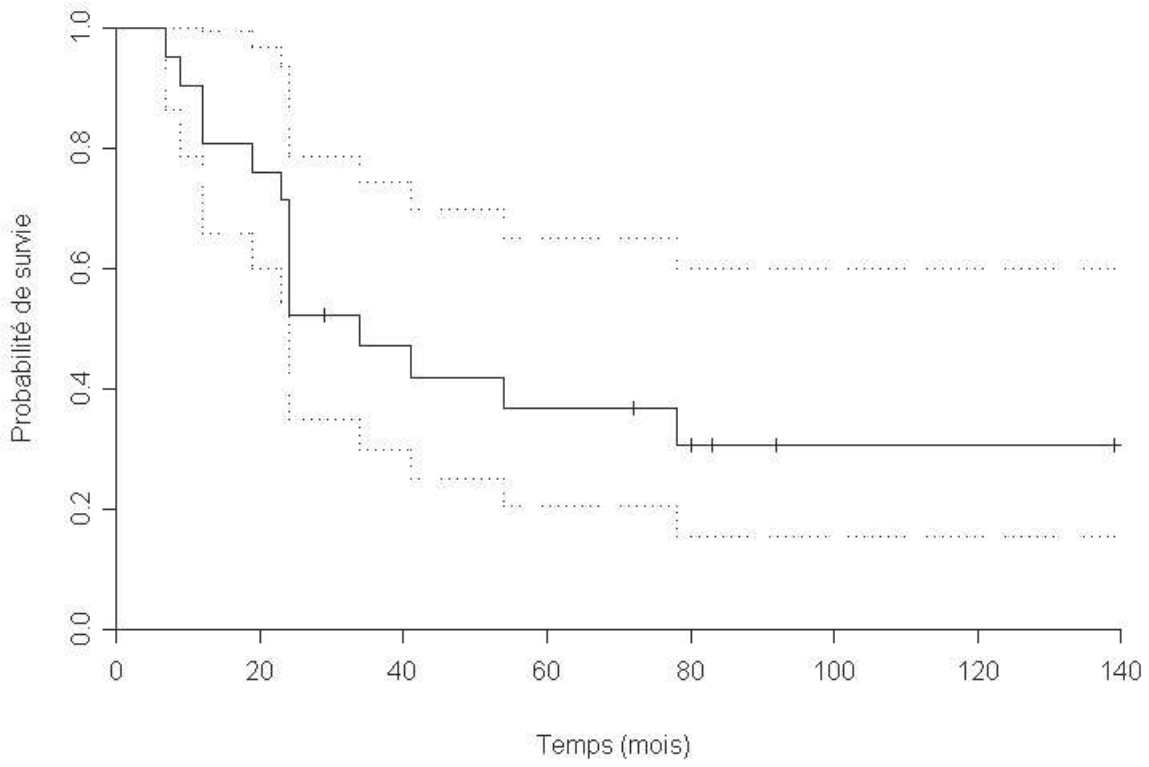

Figure 5.3 Courbe de Kaplan-Meier (avec région de confiance) pour le groupe de patients avec cancer du rectum et ayant reçu une radiothérapie pré-opératoire < 5000 rad.

5.3.4 Propriétés

- Il est de coutume d'indiquer sur la courbe de survie les données censurées par un trait vertical, de manière à voir où elles se sont produites.
- Si la durée de vie la plus élevée de l'échantillon initial est censurée, la courbe de Kaplan-Meier ne tombe pas à zéro et on prolonge la courbe par une ligne horizontale en pointillé. Si au contraire, la durée de vie la plus élevée est un décès réel, alors la courbe de Kaplan-Meier se termine sur l'abscisse.
- Toute valeur censurée précédant la première observation non censurée n'est pas prise en compte. Il faut donc l'éliminer et diminuer l'effectif n en conséquence.
- Si deux courbes de Kaplan-Meier sont reportées sur un même graphique et que l'une est toujours supérieure à l'autre, $\hat{S}_2(t) > \hat{S}_1(t)$, cela signifie que les sujets du groupe 2 ont une durée de vie plus longue que ceux du groupe 1. (voir Figure 5.4).

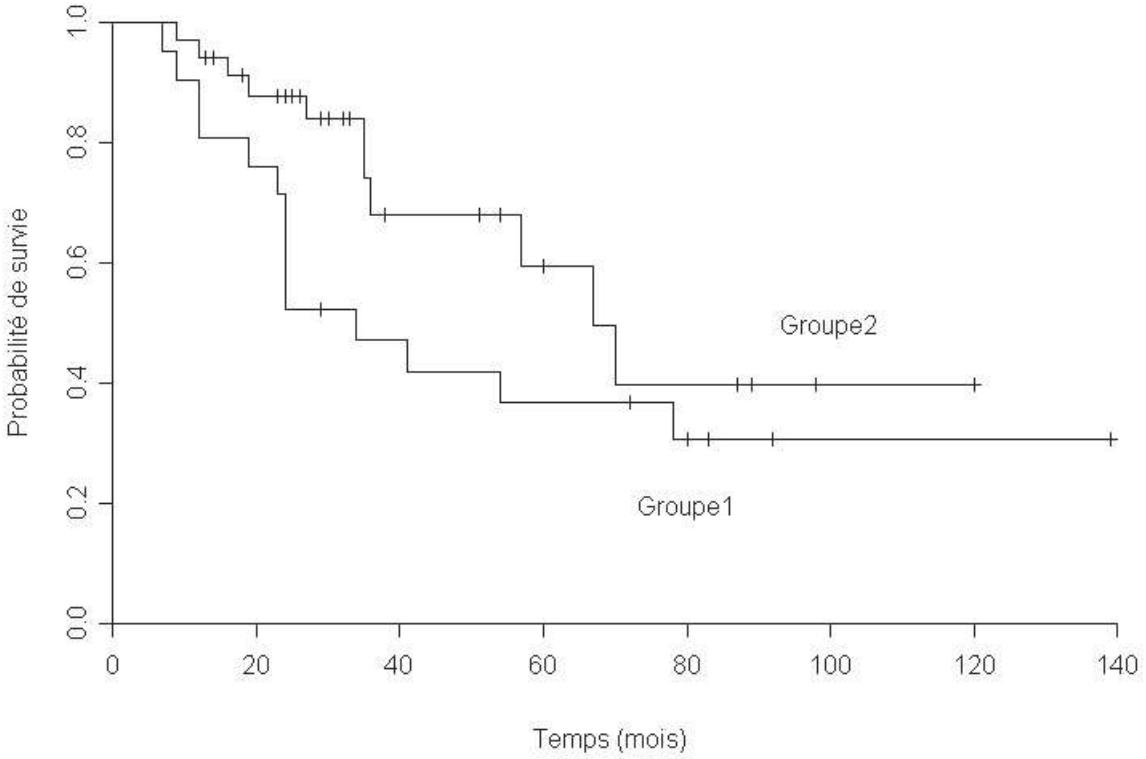

Figure 5.4 Courbe de Kaplan-Meier pour deux groupes de patients avec cancer du rectum et ayant reçu une radiothérapie pré-opératoire < 5000 rad (groupe 1) ou > 5000 rad (groupe 2).

– A partir de la courbe de survie, on peut estimer la durée médiane de survie ($P50$), en recherchant le temps t où $\hat{S}(t) = 50\%$. Il peut arriver que cette médiane n'existe pas. On peut procéder de la même manière avec tout autre percentile. Inversement, on peut rechercher pour quelle durée de vie une proportion définie de patients sont décédés (par exemple, quelle est la durée de vie à 5 ans ou à 10 ans ?).

Chapitre 6

Corrélation

6.1 Introduction

La mesure de l'association (ou de la relation) entre les observations simultanées de deux variables faites chez n sujets (ou objets) est une préoccupation fréquente en pratique. Par exemple, y a-t-il une relation entre le poids et la taille chez l'Homme ? La pression artérielle varie-t-elle avec l'âge des individus ? Peut-on dire qu'il y a une association entre le taux de cholestérol et le poids ? Le cancer du poumon est-il lié (ou dû) au tabagisme ? La consommation de médicaments est-elle corrélée avec la profession exercée ? La durée de vie de patients cancéreux augmente-t-elle en fonction de la dose de chimiothérapie reçue ? Ces questions illustrent les innombrables situations où une relation ou association entre deux variables est recherchée.

On désigne par X et Y les deux variables considérées. D'emblée il convient de distinguer deux cas de figures.

- Lorsque les variables X et Y sont observées simultanément pour chaque élément de l'échantillon, on a affaire à un problème de *corrélation* (étude observationnelle).
- Lorsque la variable X est fixée par l'expérimentateur et que l'on observe Y pour chaque valeur de X fixée, on a affaire à un problème de *régression* (étude expérimentale – plan d'expérience).

Nous supposons que les variables X et Y sont quantitatives, le plus souvent continues. Au terme du chapitre, nous dirons quelques mots sur les autres types de variables.

6.2 Coefficient de corrélation

6.2.1 Echantillon bivarié

Supposons donc que deux variables X et Y soient observées simultanément chez les sujets (ou objets) d'un échantillon d'effectif n extrait d'une population donnée. Celui-ci s'écrit comme suit :

Tableau 6.1 Echantillon bivarié

N° obs	X	Y
1	x_1	y_1
2	x_2	y_2
$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n

On dispose ainsi d'un "échantillon bivarié" d'effectif n que l'on peut aussi écrire $\{(x_i, y_i),\, i = 1, \ldots, n\}$.

Nous avons vu au Chapitre 2 comment obtenir une représentation graphique de ces données en reportant la variable Y en fonction de la variable X.

Comme décrit au Chapitre 3, on peut calculer pour chaque variable la moyenne et l'écart-type des n observations. On note $\bar{x}$ et $\bar{y}$ les moyennes, s_x et s_y, les écarts-types correspondants. Pour ce faire, on a besoin de $\sum x$, $\sum y$, $\sum x^2$ et $\sum y^2$.

6.2.2 Calcul de la corrélation

Pour mesurer l'association entre les deux variables X et Y, on calcule le coefficient de corrélation (correlation coefficient), dit de Bravais-Pearson. C'est le troisième concept le plus important en statistique.

Définition : Le coefficient de corrélation, notée $corr(X, Y)$ ou de préférence r, est donné par la formule

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{[\sum(x - \bar{x})^2][\sum(y - \bar{y})^2]}}. \tag{6.1}$$

En général, on préfère utiliser une autre formule en raison des erreurs d'arrondis qu'implique le calcul des sommes dans la formule (6.1). En développant les numérateurs et les dénominateurs comme au Chapitre 3 pour l'écart-type, on obtient :

$$r = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}}. \tag{6.2}$$

Au dénominateur, on retrouve le produit des variances de X et de Y (au facteur $(n-1)$ près). Le numérateur par contre multiplie les observations x_i et y_i. En divisant ce numérateur par $(n-1)$, on obtient la covariance de X et Y, notée $cov(X, Y)$ ou s_{xy}.

Propriétés : Le coefficient de corrélation r possède des propriétés remarquables.
- r est une mesure de la relation linéaire entre X et Y et non pas de la relation curviligne entre les deux variables
- r est un nombre pur (dépourvu d'unités)
- On a toujours $-1 \leq r \leq +1$

- Lorsque r est positif (négatif), on dit qu'il y a une corrélation positive (négative) entre X et$Y$
- $r > 0$ indique une relation croissante entre X et$Y$
  $r < 0$ indique une relation décroissante entre X et Y
  $r \simeq 0$ indique l'absence de relation croissante ou décroissante entre X et Y
- Pour calculer r, on a besoin de $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ et $\sum xy$, soit cinq quantités. Attention, toutes les sommes se calculent sur l'ensemble des paires complètes de données ! Il faut donc éliminer les paires incomplètes.
- Notons que la corrélation d'une variable avec elle-même, $corr(X, X)$, vaut 1 (corrélation parfaite).

Exemple

A titre d'exemple, considérons l'âge (années) et l'expérience professionnelle (années) des 15 premiers médecins du fichier à l'Annexe I (voir Tableau 6.2 et Figure 6.1). Calculons le coefficient de corrélation r.

Tableau 6.2 Age (années) et expérience professionnelle (années) chez les 15 premiers médecins généralistes du fichier à l'Annexe I

N° Médecin	Age (années)	Expérience professionnelle (années)
1	40	5
2	46	20
3	38	13
4	34	9
5	33	8
6	47	22
7	44	20
8	55	27
9	41	16
10	31	6
11	45	20
12	42	17
13	60	34
14	42	15
15	32	8

On trouve aisément que

$$\begin{array}{ll} \sum x = 630 & \sum y = 240 \\ \sum x^2 = 27394 & \sum y^2 = 4778 \\ \sum xy = 10965 & \end{array}$$

En utilisant la formule (6.2), on obtient

$$r = 0.9455$$

On constate qu'il existe une corrélation positive entre les deux variables comme on pouvait s'y attendre.

A titre indicatif, la corrélation entre l'âge et l'expérience professionnelle sur l'ensemble des $n = 352$ médecins (3 valeurs manquantes) vaut $r = 0.9889$ (voir Figure 6.2).

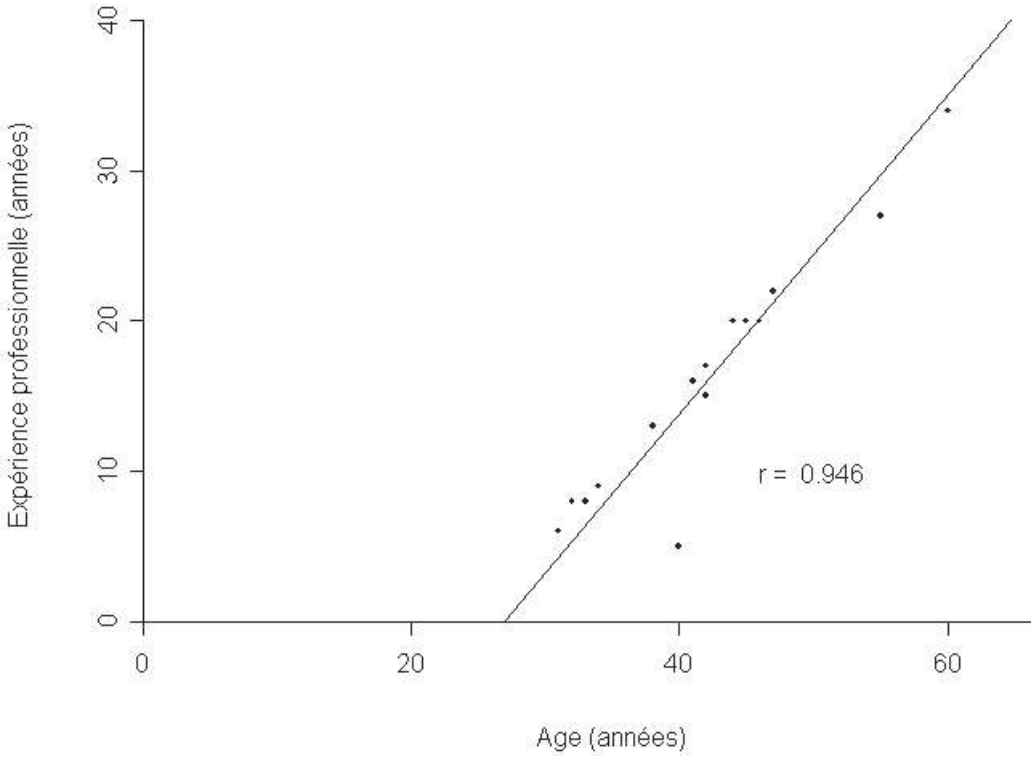

Figure 6.1 Relation entre l'âge et l'expérience professionnelle chez les 15 premiers médecins de l'enquête du CUMG (Annexe I)

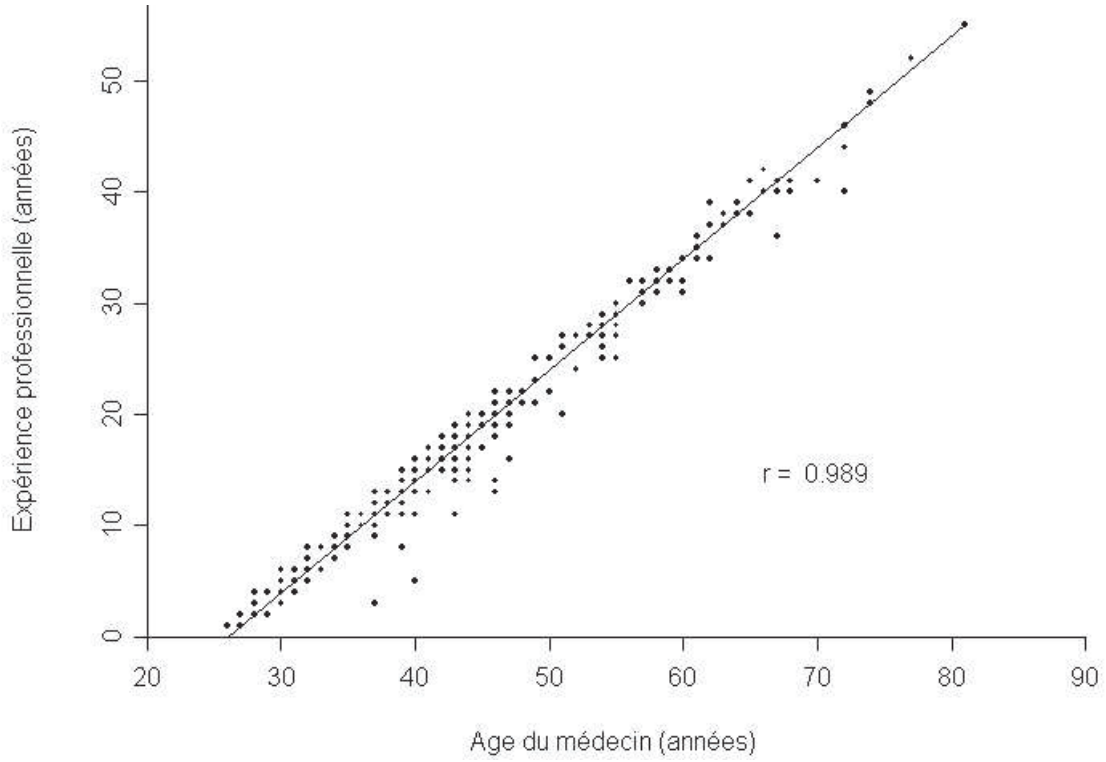

Figure 6.2 Relation entre l'expérience professionnelle et l'âge chez l'ensemble des médecins ayant participé à l'enquête du CUMG

Précautions

– En général, il est recommandé de ne pas calculer un coefficient de corrélation sans avoir examiné le graphique bivarié Y vs X. En effet, le coefficient de corrélation peut être faussé par la présence de valeurs aberrantes (non nécessairement extrêmes). A titre d'exemple, si l'on permute l'âge et l'expression professionnelle pour le médecin N°15 dans l'exemple précédent, la corrélation se détériore et on obtient $r = 0.2051$. Le point aberrant est ainsi visible sur la Figure 6.3.
En conséquence, une corrélation élevée aberrante peut parfois masquer l'absence de corrélation entre deux variables et une corrélation faible peut cacher une forte association entre les variables.

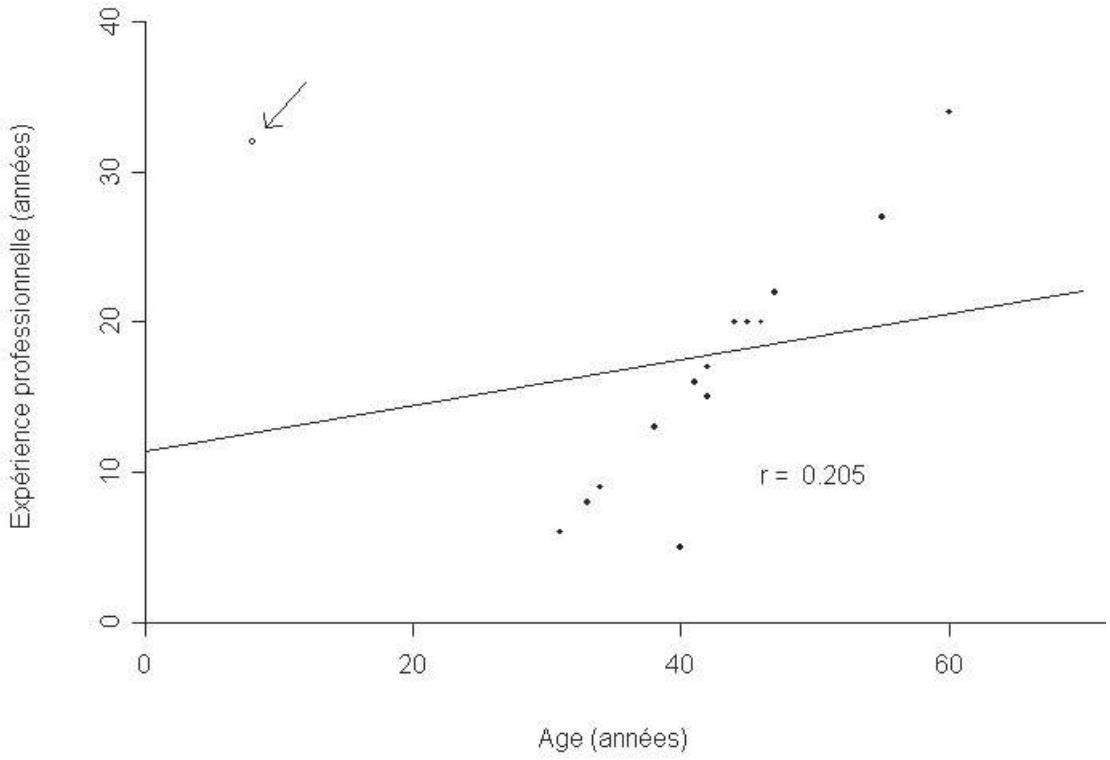

Figure 6.3 Relation entre l'expérience professionnelle et l'âge chez les 15 premiers médecins de l'enquête du CUMG (Annexe I) avec présence d'une valeur aberrante.

– Un autre aspect qui mérite attention est le caractère plus ou moins Normal (gaussien) des distributions des variables X et Y. Par exemple, si la variable X est dissymétrique à droite, il vaut mieux en prendre le logarithme ($\ell n\, X$) et calculer la corrélation entre $\ell n\, X$ et Y. On fait de même pour la variable Y si sa distribution est fortement dissymétrique. La transformation logarithmique n'est qu'un artifice mathématique pour normaliser la distribution et n'a pas d'implication biologique ou autre.

6.3 Matrice des corrélations

Lorsque plusieurs variables, notées $X_1, \ldots, X_p$, sont mesurées simultanément chez n sujets, on peut calculer les corrélations entre toutes les paires de variables. On note $r_{ij} = corr(X_i, X_j)$.

Au total, puisque $r_{ii} = 1$ et $r_{ij} = r_{ji}$, $\forall i \neq j$, il y a $p(p-1)/2$ corrélations différentes à calculer. On construit ainsi une matrice de corrélations, c'est-à-dire un tableau à p lignes et p colonnes. La matrice ci-dessous donne les corrélations entre les $p = 4$ variables, âge (X_1), expérience professionnelle (X_2), nombre moyen de médicaments prescrits par patient (X_3) et nombre moyen de problèmes par patient (X_4) pour l'échantillon des médecins généralistes (voir Tableau 6.3). On note 3 données manquantes, de sorte que $n = 352$.

Tableau 6.3 Matrice des corrélations entre l'âge (X_1), l'expérience professionnelle (X_2), le nombre moyen de médicaments prescrits par patient (X_3) et le nombre moyen de problèmes par patient (X_4) observés chez les médecins généralistes wallons dans l'enquête du CUMG ($n = 352$)

	X_1	X_2	X_3	X_4
X_1	1.0			
X_2	0.9889	1.0		
X_3	−0.0391	−0.0370	1.0	
X_4	−0.1333	−0.1369	0.5421	1.0

6.4 Coefficient de corrélation de Spearman

6.4.1 Définition et calcul

Lorsque les variables X et Y ne suivent pas une distribution Normale (gaussienne) ou que les variables X et Y sont ordinales plutôt que continues, il y a lieu de calculer le coefficient de corrélation de Spearman, noté r_S.

Celui-ci consiste simplement à calculer le coefficient de corrélation de Bravais-Pearson (6.1) sur les rangs des observations de X et de Y et non sur les valeurs observées.

On procède comme suit :

- On remplace chaque valeur x_i par son rang noté $rang\,(x_i)$ et on remplace chaque valeur y_i par son rang noté $rang\,(y_i)$.
  Pour rappel, le rang est la position de l'observation dans la suite des observations triées par ordre croissant. En cas d'ex-aequo, on remplace les rangs correspondants par leur moyenne.
- On calcule $d_i = rang\,(x_i) - rang\,(y_i)$ pour tout $i = 1, \ldots, n$.
- On utilise ensuite la formule

$$r_S = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}. \tag{6.3}$$

Le coefficient de corrélation de Spearman s'interprète comme le coefficient de corrélation classique. Il est cependant moins sensible aux valeurs aberrantes, du fait que l'on travaille sur les rangs des observations, et donc plus robuste.

6.4.2 Exemple

A titre d'illustration, le Tableau 6.4 reproduit les données du Tableau 6.1 pour le calcul du coefficient de corrélation de Spearman.

Tableau 6.4 Calcul du coefficient de corrélation de Spearman entre l'âge et l'expérience professionnelle chez les 15 premiers médecins généralistes du fichier à l'Annexe I

N° Médecin	Age	(rang)	Exp. prof.	(rang)	d	d^2
1	40	(6)	5	(1)	5	25
2	46	(12)	20	(11)	1	1
3	38	(5)	13	(6)	−1	1
4	34	(4)	9	(5)	−1	1
5	33	(3)	8	(3.5)	−0.5	0.25
6	47	(13)	22	(13)	0	0
7	44	(10)	20	(11)	−1	1
8	55	(14)	27	(14)	0	0
9	41	(7)	16	(8)	−1	1
10	31	(1)	6	(2)	−1	1
11	45	(11)	20	(11)	0	0
12	42	(8.5)	17	(9)	−0.5	0.25
13	60	(15)	34	(15)	0	0
14	42	(8.5)	15	(7)	1.5	2.25
15	32	(2)	8	(3.5)	−1.5	2.25
				Total	$\sum d^2 = 36.0$	

D'où on obtient

$$r_S = 1 - \frac{6 \times 36.0}{15(15^2 - 1)} = 0.9357$$

confirmant la forte corrélation positive entre les deux variables. Si on permute les valeurs observées pour le médecin N°15 comme on l'a fait précédemment, le coefficient de corrélation de Spearman devient

$$r_s = 1 - \frac{6 \times 207.5}{15(15^2 - 1)} = 0.6295$$

et on constate qu'il est moins affecté que précédemment !

6.5 Cas particuliers

Nous avons vu que le coefficient de corrélation était calculé pour mesurer l'association entre deux variables quantitatives X et Y, voire continues. Qu'en est-il dans les autres cas ?

6.5.1 Coefficient de corrélation bisérial de point

Lorsqu'une des deux variables, X par exemple, est binaire et l'autre quantitative, l'équation (6.1) se simplifie et on obtient le coefficient de corrélation bisérial de point :

$$r = \frac{\sqrt{n_0 n_1}\,(\bar{y}_1 - \bar{y}_0)}{n s_y} \tag{6.4}$$

où n_0 et n_1 correspondent aux nombres d'observations pour lesquelles X vaut 0 et 1, respectivement, $\bar{y}_0$ et $\bar{y}_1$, la moyenne des observations y_i pour $X = 0$ et $X = 1$ respectivement et s_y l'écart-type des observations y_i $(i = 1, \ldots, n)$. On constate immédiatement que si $\bar{y}_1 > \bar{y}_0$, $r > 0$ et la droite qui passe dans le plan par les points moyens est croissante. Par contre, si $\bar{y}_1 < \bar{y}_0$, la droite est décroissante. A titre d'exemple, pour le fichier repris à l'Annexe I, le coefficient de corrélation bisérial de point entre le sexe et l'âge des médecins vaut $r = -0.2912$, puisque $n_0 = 292$, $n_1 = 60$, $\bar{y}_0 = 45.2$, $\bar{y}_1 = 37.0$ et $s_y = 10.662$. On constate que la corrélation est négative, indiquant par là que les femmes sont plus jeunes que les hommes.

6.5.2 Coefficient de corrélation de point

Lorsque les deux variables X et Y sont binaires, on obtient le coefficient de corrélation de point et la formule (6.2) devient

$$r = \frac{n_{00}n_{11} - n_{01}n_{10}}{\sqrt{(n_{00} + n_{01})(n_{10} + n_{11})(n_{00} + n_{10})(n_{01} + n_{11})}} \tag{6.5}$$

où n_{ij} est le nombre d'observations où $X = i$ et $Y = j$ avec $i, j = 0$ ou 1. A titre d'exemple, pour le fichier des médecins généralistes, on a corrélé le sexe $(0 = H,\ 1 = F)$ avec l'agrément (0 = non, 1 = oui). On obtient le Tableau 6.5.

Tableau 6.5 Association entre le sexe du médecin et son agrément INAMI

	Sexe		
Agrément INAMI	Homme	Femme	Total
Non	19	7	26
Oui	273	53	326
Total	292	60	352

Dès lors,

$$r = \frac{19 \times 53 - 273 \times 7}{\sqrt{26 \times 326 \times 292 \times 60}} = -0.0742.$$

On peut donc conclure qu'il n'y a pas d'association entre ces deux variables.

6.5.3 Coefficient d'association

Lorsqu'on veut voir l'association entre deux variables qualitatives, respectivement X à ℓ modalités et Y à c modalités, on détermine une table $\ell \times c$ qui croise les deux variables.

On effectue alors un test d'indépendance comme au Chapitre 10. On vérifie s'il y a indépendance ou dépendance entre les deux variables. Par exemple, pour l'échantillon des médecins, on a croisé l'université d'origine et la Province où le médecin travaille (Tableau 6.6).

Tableau 6.6 Croisement entre l'université d'origine du médecin et la province où il travaille (Enquête CUMG)

	Université				
Province	ULg	UCL	ULB	Autres	Total
Brabant wallon	2	19	11	0	32
Hainaut	13	82	26	2	123
Liège	108	19	2	2	131
Luxembourg	6	15	0	0	21
Namur	10	30	7	1	48
Total	131	165	46	5	355

On montre qu'il existe une association hautement significative entre ces deux facteurs qualitatifs ($\chi^2 = 182$ à 12 degrés de liberté).

6.6 Coefficient Kappa de Cohen

Un problème qui est proche de la corrélation est celui du degré de concordance (ou d'accord) entre deux observateurs. Par exemple, deux radiologues examinent chacun n radiographies de patients et doivent conclure s'il y a ou non présence d'un nodule (absence = 0, présence = 1). On peut en outre leur demander s'il s'agit d'un nodule bénin ou malin (avis selon 3 catégories). Deux professeurs peuvent juger les capacités de n étudiants en 4 catégories (insuffisant, moyen, bon, excellent). On obtient chaque fois deux séries de données "appariées" (paired data) car les appréciations sont faites sur les mêmes sujets. On montre dans ce cas que le coefficient de corrélation n'est pas l'indicateur idéal et qu'il est préférable de calculer le coefficient Kappa de Cohen qui mesure l'accord entre les deux observateurs. On parle d'ailleurs d'accord inter-observateurs ("inter-observer agreement").

Supposons que deux observateurs A et B évaluent n sujets (ou objets) selon une échelle à k critères. On obtient ainsi une série de couples de valeurs $\{(a_i, b_i), i = 1, \ldots, n\}$, où a_i est le critère donné par l'observateur A au sujet i et b_i le critère donné par l'observateur B au sujet i. Dans ces conditions, on croise les évaluations des observateurs A et B et on établit un tableau $k \times k$ (Tableau 6.7) reprenant les fréquences d'accord ou de

désaccord dans chaque cellule. La fréquence p_{ij} représente la proportion de sujets (sur le total de n sujets) qui ont été classés dans le critère a_i par l'observateur A et dans le critère b_j par l'observateur B. Bien sûr, $\sum_{ij} p_{ij} = 1$.

Tableau 6.7 Croisement des évaluations à l'aide d'une échelle à k critères réalisées par deux observateurs A et B sur n sujets

	Observateur B				
Observateur A	1	2	...	k	Total
1	p_{11}	p_{12}	...	p_{1k}	$p_{1.}$
2	p_{21}	p_{22}	...	p_{2k}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
k	p_{k1}	p_{k2}	...	p_{kk}	$p_{k.}$
Total	$p_{.1}$	$p_{.2}$	...	$p_{.k}$	1

Le coefficient Kappa de Cohen compare la proportion d'accords observés p_0 avec la proportion d'accords attendus p_e si les deux observateurs évaluaient au hasard. On a donc la formule

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{6.6}$$

où $p_0 = \sum_{i=1}^{k} p_{ii}$ et $p_e = \sum_{i=1}^{k} p_{i.}p_{.i}$.

La formule (6.6) est évidemment plus simple dans le cas $k = 2$ (évaluations à 2 critères).

A titre d'exemple, supposons que 2 psychiatres aient diagnostiqué 100 sujets en trois catégories (psychotique, neurotique, organique) et qu'on ait obtenu le Tableau 6.8

Tableau 6.8 Classification de 100 sujets par deux psychiatres en 3 catégories (psychotique, neurotique, organique)

	Psychiatre B			
Psychiatre A	Psychotique	Neurotique	Organique	Total
Psychotique	0.75	0.01	0.04	0.80
Neurotique	0.05	0.04	0.01	0.10
Organique	0	0	0.10	0.10
Total	0.80	0.05	0.15	1.00

En utilisant la formule (6.6), on obtient successivement

$$p_0 = 0.75 + 0.04 + 0.10 = 0.89$$

et

$$p_e = 0.80 \times 0.80 + 0.10 \times 0.05 + 0.10 \times 0.15 = 0.66$$

d'où

$$\kappa = \frac{0.89 - 0.66}{1 - 0.66} = 0.68.$$

Remarque : Plus le coefficient κ est voisin de 1, meilleur est l'accord entre les deux observateurs. Une valeur de κ voisine de 0 correspond à une absence de concordance entre les deux évaluateurs. Il peut arriver que κ soit négatif (désaccord total).

6.7 Coefficient de détermination

Le coefficient de détermination (determination coefficient) n'est autre que le carré du coefficient de corrélation. On le note donc r^2. Par définition $0 \leq r^2 \leq 1$. Il s'interprète de la manière suivante : le coefficient de détermination mesure le pourcentage de la variabilité d'une variable expliquée par l'autre variable. A titre d'exemple, si la corrélation entre le poids et la taille vaut $r = 0.80$, le coefficient de détermination vaut $r^2 = 0.64$. En d'autres termes, 64% de la variabilité du poids peut être expliquée par la taille. Si on connaît la taille de quelqu'un, en moyenne, on a déjà une excellente idée de son poids. Cet exemple montre qu'une corrélation n'a de réelle signification que si elle est au moins supérieure ou égale à 0.7 puisque $r^2 = 0.7 \times 0.7 = 0.49$ (50% de la variabilité d'une variable est expliquée par l'autre variable).

Chapitre 7

Régression linéaire

7.1 Introduction

Comme on l'a vu au Chapitre 6, lorsqu'on étudie la relation entre deux variables X et Y dont l'une est fixée (on dit aussi "contrôlée") par l'expérimentateur, on n'a plus affaire à un problème de corrélation. On parle alors de problème de régression linéaire ou de peréquation rectiligne (regression analysis).

Supposons donc que l'on fixe n valeurs de la variable X, soient $x_1, \ldots, x_n$, et que pour $X = x_i$ on observe la valeur $Y = y_i$, $(i = 1, \ldots, n)$. On obtient ainsi un échantillon bivarié $\{(x_i, y_i),\ i = 1, \ldots, n\}$ comme celui repris au Tableau 6.1. Si l'apparence est la même, la façon dont les données ont été récoltées est totalement différente et il faut en tenir compte.

Notons que certaines valeurs x_i peuvent être les mêmes, ce qui revient à dire que l'on fait plusieurs observations de la variable Y en $X = x_i$. On parle alors d'un modèle à mesures répétées.

Le fait de contrôler la variable X présente certains avantages :

- écarter un domaine des valeurs de la variable
- balayer l'ensemble des valeurs de la variable
- faire davantage de mesures en certains points.

La méthode de régression est utilisée quotidiennement dans les laboratoires (courbe d'étalonnage ou de calibration) mais aussi dans les plans d'expérience en recherche (courbe dose-réponse), en biométrie (courbe de croissance), ou en Santé publique (courbe de morbidité en fonction de l'exposition à des facteurs pathogènes).

Historiquement, le terme "régression" est dû à Sir Francis Galton (1822-1911), médecin et statisticien anglais. Intéressé par les problèmes d'hérédité, il étudia la taille des fils (progéniture) en fonction de la taille des pères (parents) et observa une "régression" des tailles vers la valeur centrale de la distribution (qu'il appela lui-même "mediocrity"). En effet, la taille des fils dont les pères étaient grands avait tendance à se réduire, alors que les pères de petite taille avaient une propension à avoir des fils plus grands.

7.2 Modèle linéaire

En régression linéaire, on suppose que les valeurs moyennes de la variable "observée" Y en fonction de la variable "contrôlée" X se situent approximativement sur une droite. Il s'agit d'une condition qui n'est pas toujours réalisée (régression non linéaire), mais une transformation de la variable Y (et/ou de X) permet parfois de se ramener à un modèle linéaire.

Le modèle linéaire s'écrit donc

$$y_i = a + b.x_i + \varepsilon_i \qquad (i = 1, \ldots, n) \tag{7.1}$$

où x_i est la valeur fixée par l'expérimentateur, y_i la valeur observée et ε_i l'écart entre la valeur observée y_i et la valeur $\hat{y}_i = a + bx_i$ située sur la droite en $X = x_i$. En effet, on ne peut espérer faire passer une droite par l'ensemble des points observés !

Le modèle s'écrit aussi sous la forme de l'équation d'une droite, appelée "droite de régression de Y sur X",

$$\hat{Y} = a + b.x \tag{7.2}$$

où $\hat{Y}$ est la moyenne de la variable Y en $X = x$, b est la pente (slope) et a l'ordonnée à l'origine (intercept) de la droite de régression. $\hat{Y}$ est aussi appelée la "prédiction" de Y pour $X = x$.

L'ordonnée à l'origine (a) et la pente (b) sont appelés les paramètres du modèle de régression (7.2). Par ailleurs, les valeurs observées de Y fluctuent autour de la droite de régression avec une variabilité que l'on suppose *constante* quelle que soit la valeur choisie $X = x_i$. Or la variabilité se mesure en termes d'écart-type (ou de variance). Dans ce chapitre, on note $s_{y|x}$, l'écart-type des valeurs de Y pour $X = x$. Il mesure la variabilité des observations autour de la droite (7.2) et est appelé "écart-type résiduel" (residual standard deviation). Il convient de ne pas confondre cette valeur avec l'écart-type de Y, noté s_y, obtenu sans tenir compte de X. On est confronté ici à un problème bivarié et non univarié.

7.3 Calcul de la droite de régression

Il existe une infinité de droites d'équation (7.2) passant au travers du nuage de points $\{(x_i, y_i),\ i = 1, \ldots, n\}$. On recherche la droite qui s'ajuste "le mieux" aux observations, en utilisant le principe "des moindres carrés" (least-squares method). Ce principe consiste à minimiser la somme des carrés des écarts ou résidus ε_i entre les valeurs observées y_i et prédites $\hat{y}_i = a + bx_i$.

En d'autres termes, il convient de rechercher le minimum de la fonction

$$\begin{aligned} L(a, b) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \end{aligned} \tag{7.3}$$

On montre aisément que le minimum est atteint pour les valeurs suivantes :

$$b = \frac{\sum xy - (\sum x)(\sum y)/n}{\sum x^2 - (\sum x)^2/n} \tag{7.4}$$

et

$$a = \bar{y} - b.\bar{x} \tag{7.5}$$

où $\bar{x}$ et $\bar{y}$ représentent respectivement les moyennes des observations x_i et y_i, $(i = 1, \ldots, n)$.

Comme pour la corrélation, on constate que pour déterminer la droite de régression, il faut calculer $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ et $\sum xy$.

Par ailleurs, l'écart-type résiduel vaut

$$s_{y|x} = \sqrt{\frac{[\sum y^2 - (\sum y)^2/n] - b^2[\sum x^2 - (\sum x)^2/n]}{n-2}}. \tag{7.6}$$

qui mesure la variabilité des observations autour de la droite de régression.

Connaissant a et b, on peut calculer les résidus (residuals)

$$\begin{aligned} \hat{\varepsilon}_i &= y_i - \hat{y}_i \\ &= y_i - (a + bx_i) \end{aligned} \tag{7.7}$$

que l'on peut reporter sur un graphique en fonction des valeurs x_i. On obtient ainsi un graphique des résidus (residual plot), fort utile pour juger de l'adéquation du modèle ou pour voir si certaines observations ne sont pas aberrantes.

7.4 Application

Illustrons la méthodologie de la droite de régression à l'aide de données récoltées dans un plan d'expérience, où on a étudié le temps d'une réaction chimique complète (minutes) en fonction de la température (°C). Ces données sont reprises au Tableau 7.1. La variable X fixée par l'expérimentateur est la "température" et la variable Y observée est le "temps de réaction chimique complète".

Tableau 7.1 Temps de réaction chimique complète (min)
en fonction de la température (° Celsius). Plan d'expérience

Numéro de l'expérience	Température (° C)	Temps de réaction complète (min)
1	25	0.64
2	45	1.27
3	55	0.95
4	85	1.85
5	115	2.81
6	125	2.80
7	150	3.42
8	165	4.30
9	175	4.54
10	200	4.70

Déterminons d'abord les sommes nécessaires au calcul de la pente et de l'ordonnée à l'origine de la droite de régression. On a successivement :

$$\begin{aligned} \sum x &= 1140 \\ \sum y &= 27.28 \\ \sum x^2 &= 162100 \\ \sum y^2 &= 94.972 \\ \sum xy &= 3912.80 \end{aligned}$$

Comme $n = 10$, on a $\bar{x} = 114°\text{C}$ et $\bar{y} = 2.73$ min.

En utilisant les formules (7.4) et (7.5), la pente vaut $b = 0.02498$ et l'ordonnée à l'origine $a = 2.73 - 0.02498 \times 114 = -0.1198$.

La droite de régression a pour équation :

$$\hat{Y} = -0.1198 + 0.02498x \tag{7.8}$$

et est représentée à la Figure 7.1 avec les données observées.

Enfin, l'écart-type résiduel, obtenu à partir de la formule (7.6), vaut

$$\begin{aligned} s_{y|x} &= \sqrt{\frac{[94.972 - (27.28)^2/10] - (0.02498)^2[162100 - (1140)^2/10]}{8}} \\ &= 0.2488. \end{aligned}$$

L'équation (7.8) peut aussi s'écrire

$$\hat{Y} = -0.1198 + 0.02498x \ \ (\pm 0.249) \tag{7.9}$$

de manière à préciser la variabilité des valeurs de Y autour de la droite de régression. A titre d'exemple, que vaut le temps de réaction chimique complète pour une température de 70°C ? En utilisant l'équation (7.9), on a

$$\begin{aligned} \hat{Y} &= -0.1198 + 0.02498 \times 70 \ \ (\pm 0.249) \\ &= 1.63 \ (\pm 0.249) \text{ minutes.} \end{aligned}$$

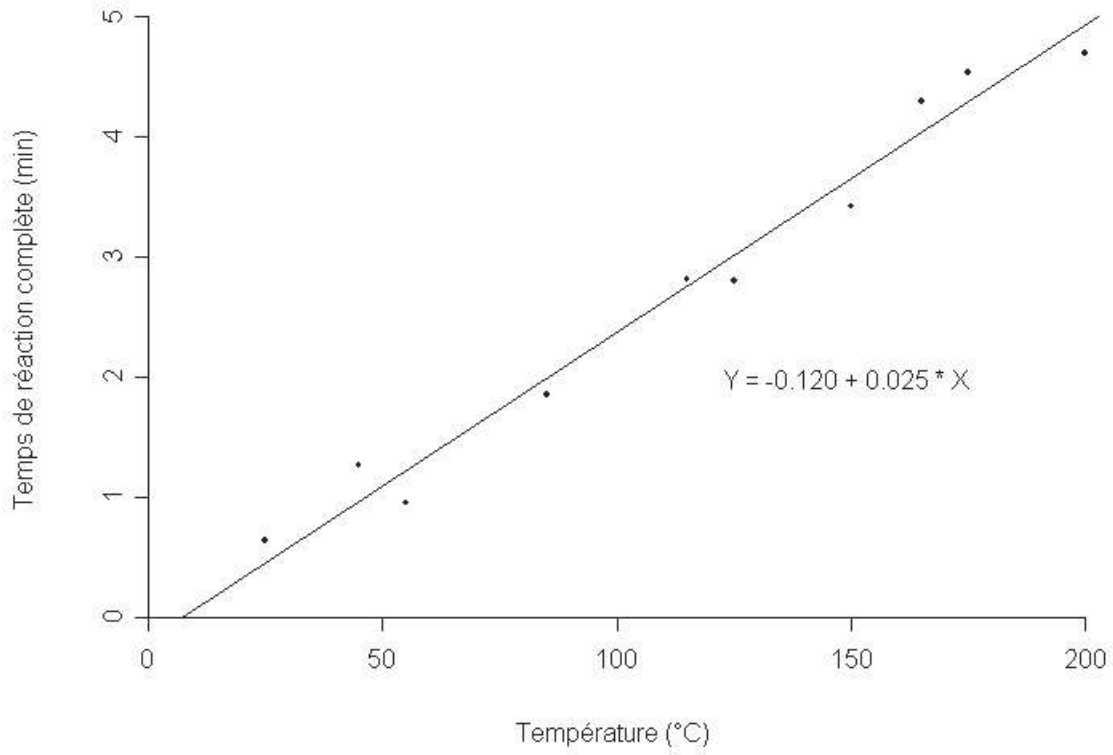

Figure 7.1 Droite de régression du temps de réaction chimique complète (min) en fonction de la température (°C)

En utilisant un argument similaire à celui utilisé pour les intervalles de référence, on peut dire qu'en répétant un grand nombre de fois l'expérience à 70°C, 95% des temps de réaction chimique observés se situent dans l'intervalle $1.63 \pm 2 \times 0.249$, soit de 1.13 à 2.13 min ! On peut donc aller jusqu'à une minute de différence pour une même température.

La Figure 7.2 reporte les résidus ε_i en fonction des valeurs fixées x_i $(i = 1, \ldots, n)$. On voit qu'ils se répartissent au hasard autour de l'axe des abscisses, indiquant par là un bon ajustement linéaire. D'ailleurs, $\sum_{i=1}^{n} \varepsilon_i = 0$.

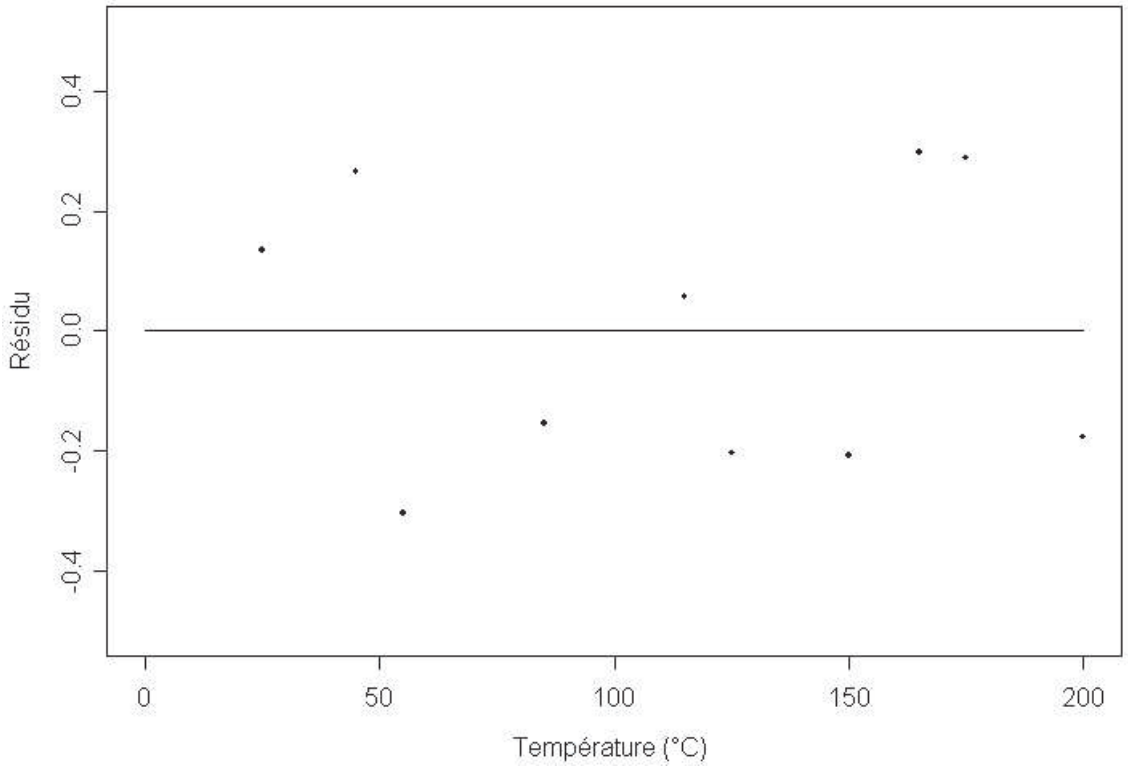

Figure 7.2 Relation entre le temps de réaction chimique complète (min) en fonction de la température (°C). Grahique des résidus.

7.5 Coefficient de détermination

Le coefficient de détermination (determination coefficient), noté r^2, est un indicateur de la qualité de l'ajustement de la droite de régression sur le nuage de points. Il mesure le rapport entre la variabilité de Y attribuée à la régression sur la variabilité totale de Y.

Il est donné par la formule

$$r^2 = \frac{[\sum xy - (\sum x)(\sum y)/n]^2}{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]} \tag{7.10}$$

et on constate qu'il est en tous points identique au carré du coefficient de corrélation défini par l'équation (6.1). Toutefois, en régression, r^2 ne doit pas être considéré comme le carré de r, terme banni, puisque une des deux variables est fixée par l'utilisateur.

Comme nous l'avons vu, $0 \leq r^2 \leq 1$, la valeur $r^2 = 1$ correspond à un ajustement parfait (tous les points observés sont situés sur la droite) et la valeur $r^2 = 0$ à une droite de régression horizontale, quelle que soit la variabilité des points autour de cette horizontale ; dans ce dernier cas, Y ne varie pas en fonction de X.

Le coefficient de détermination donne le pourcentage de variabilité de Y expliqué par la droite de régression. On montre d'ailleurs qu'on a approximativement

$$s^2_{y|x} = (1 - r^2)s^2_y. \tag{7.11}$$

Donc si $r^2 = 1$ (régression parfaite), $s^2_{y|x} = 0$ et il n'y a pas de variabilité résiduelle autour de la droite. Tout est expliqué par la régression. A l'inverse, si $r^2 = 0$, la variabilité résiduelle contient toute la variabilité de Y et rien n'est expliqué par la régression, donc par X.

Dans l'exemple du paragraphe précédent,

$$\begin{aligned} r^2 &= \frac{[3912.8 - (1140 \times 27.28)/10]^2}{[162100 - (1140)^2/10][94.972 - (27.28)^2/10]} \\ &= \frac{644616.3}{32140 \times 20.552} \\ &= 0.9759 \end{aligned}$$

On peut donc dire qu'approximativement 98% de la variabilité du temps de réaction chimique complète peut être attribuée à la température.

Notons enfin qu'en utilisant la formule (7.11), la valeur de $r^2 = 1 - s^2_{y|x}/s^2_y$ est légèrement différente. En effet, on a $r^2 = 1-(0.0619/2.2836) = 0.9727$. Cette estimation est en général meilleure que celle obtenue avec l'équation (7.10).

7.6 Régression et corrélation

Si on ne peut parler de corrélation dans un problème de régression, l'inverse n'est pas vrai. En effet, dans un problème de corrélation, après avoir calculé la corrélation entre deux variables, on peut calculer la droite de régression de Y sur X en utilisant la méthode décrite dans ce chapitre. Il est facile de montrer que la pente et l'ordonnée à l'origine de la droite de régression $\hat{Y} = a + bx$ sont données par les relations :

$$\begin{aligned} b &= r\,\frac{s_y}{s_x} \\ a &= \bar{y} - b\bar{x} \end{aligned} \tag{7.12}$$

On peut même calculer la droite de régression de X sur Y, soit $\hat{X} = a' + b'y$, qui est différente de celle de Y sur X.

On a

$$\begin{aligned} b' &= r\,\frac{s_x}{s_y} \\ a' &= \bar{x} - b'\bar{y} \end{aligned} \tag{7.13}$$

et on constate que $b.b' = r^2$, le coefficient de détermination. Observons donc que b' n'est pas l'inverse de b mais que les deux pentes ont le même signe ! A titre d'exemple, on a

calculé les droites de régression de l'expérience professionnelle Y sur l'âge X et celle de l'âge sur l'expérience professionnelle. Ces deux droites de régression ont pour équation :

$$\begin{aligned} \hat{Y} &= -25.01 + 0.978x \\ \hat{X} &= 25.97 + 0.9996y \end{aligned}$$

Remarque : La droite d'équation $\hat{Y} = a'' + b''.x$, où $b'' = \pm s_y/s_x$ (on choisit le signe de b et b') et $a'' = \bar{y} - b''.\bar{x}$ est la droite d'allométrie ou des "moindre rectangles". Elle est intermédiaire entre les deux droites de régression et est parfois utilisée dans les laboratoires pour comparer deux techniques de mesure.

7.7 Régression non linéaire

Lorsque la régression de Y sur X est curviligne, on dit qu'on a affaire à un problème de régression non linéaire. Un exemple simple est l'ajustement d'un polynôme du second degré. On a le modèle non linéaire

$$Y = a + b.x + c.x^2 \tag{7.14}$$

On recherche les paramètres a, b et c qui minimisent l'expression

$$\begin{aligned} L(a,b,c) &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum [y_i - (a + bx_i + cx_i^2)]^2 \end{aligned}$$

en résolvant un système de 3 équations à 3 inconnues. Ces équations sont

$$\left\{ \begin{array}{l} an + b\sum x + c\sum x^2 = \sum y \\ a\sum x + b\sum x^2 + c\sum x^3 = \sum xy \\ a\sum x^2 + b\sum x^3 + c\sum x^4 = \sum x^2 y \end{array} \right. \tag{7.15}$$

Un autre cas classique est la fonction exponentielle $Y = ae^{bx}$ que l'on peut linéariser en posant $Y^* = \ell n\, Y$. En effet, $Y^* = \ell n\, a + bx$. De même, pour la fonction puissance $Y = ax^b$, la fonction logarithmique conduit au modèle $Y^* = a + b\, \ell n\, x$.

7.8 Remarque finale

La notion de régression n'a pas vraiment de sens si la variable X est binaire (0 ou 1). Toutefois, rien n'empêche d'appliquer la méthode et de déterminer la régression de Y sur X, à savoir $\hat{Y} = a + bx$. On voit immédiatement que pour $X = 0$, $\hat{Y} = a$ et pour $X = 1$, $\hat{Y} = a + b$. La pente vaut $b = \bar{y}_1 - \bar{y}_0$. En calculant la droite de régression de l'âge sur le sexe (Annexe I), on obtient l'équation

$$\text{Age } = 53.5 - 8.25 \times \text{ sexe},$$

qui donne la moyenne de l'âge chez les médecins de sexe masculin (sexe = 0) et féminin (sexe =1). Dans le cas où X est une variable qualitative à k modalités, on a affaire à un problème d'analyse de la variance à 1 critère (voir Chapitre 17). Toutefois, la théorie moderne des modèles linéaires généraux (GLM) permet d'accommoder les différents types de variable X, pour autant que la variable Y soit continue.

Chapitre 8

Erreur type

8.1 Introduction

Le concept d'erreur type (standard error, SE) est la quatrième notion la plus importante en statistique après celles de moyenne, d'écart-type et de corrélation. Elle fait la transition (difficile) entre la statistique descriptive et la statistique inférentielle. Sa compréhension n'est pas aisée. En effet, l'erreur type est une mesure de la variabilité non pas des observations dans un échantillon de données mais d'une caractéristique d'échantillon, par exemple la moyenne arithmétique, sur l'ensemble des échantillons possibles que l'on aurait pu tirer d'une population. Elle mesure donc ce que l'on appelle la variabilité d'échantillonnage (sampling variability). La difficulté de compréhension provient du fait qu'on ne peut aisément concevoir la variabilité d'une caractéristique d'échantillon, par exemple la moyenne, lorsqu'on n'a qu'un *seul* échantillon. En effet, nous l'avons vu aux Chapitres 3 et 4, pour calculer l'écart-type ou tout autre paramètre de dispersion, il faut au moins deux observations. Dans le cas de l'erreur type, il faudrait donc au moins deux échantillons ! Certains résultats mathématiques et de nouvelles approches apparues récemment dans la littérature permettent cependant de contourner cette difficulté.

La notion d'erreur type est intimement liée à celle d'échantillonnage d'une population et de paramètres de population. L'erreur type mesure aussi la "précision statistique" de l'estimation d'un paramètre de population. Supposons que dans un sondage d'opinion portant sur 1000 individus lors d'une élection présidentielle où deux candidats s'opposent, on obtienne 52% pour le premier candidat et 48% pour le second ; quelle est la fiabilité statistique de ces chiffres ? Que deviendraient ces chiffres, si on recommençait le sondage sur 1000 autres sujets ? Trouverait-on des valeurs voisines de celles obtenues ou au contraire des scores fort différents ? Sur base des résultats obtenus, 52% et 48%, avec quel degré de certitude peut-on affirmer que le premier candidat sera le futur président ? Telles sont les questions fondamentales qu'il faut se poser dans tout problème de statistique inférentielle. On s'interroge ainsi sur la signification qu'il faut attribuer "aux statistiques". Ne dit-on pas que "on peut faire dire aux statistiques tout ce que l'on veut !".

8.2 Paramètres de population

Au Chapitre 1, nous avons défini la population comme une collection de sujets (ou d'objets) ayant au moins une propriété en commun. Par ailleurs, l'effectif (souvent infini) d'une population est noté N.

Désignons par X la variable (quantitative) à laquelle on s'intéresse dans la population, par exemple la pression artérielle systolique, le poids, la taille, l'âge ou le nombre de médicaments pris par jour.

8.2.1 Moyenne

Chaque individu de la population étant porteur d'une valeur x de la variable X, on peut calculer la moyenne (mean) de X pour l'ensemble des sujets de la population, soit

$$\mu = \frac{\sum x}{N} \tag{8.1}$$

Lorsque N est infini, cette quantité est remplacée par une intégrale. On dit que μ est un "paramètre de population". Par convention, les paramètres de population sont souvent désignés par des lettres grecques. En général, μ est un paramètre *inconnu* sinon il n'y aurait pas de problème statistique ni la nécessité de tirer un échantillon. C'est parce que μ est inconnu que la statistique inférentielle a sa raison d'être.

8.2.2 Proportion

Une proportion est la moyenne d'une variable binaire. Donc si X est une variable binaire (0 ou 1), la moyenne de X dans la population est la proportion

$$\pi = \frac{\sum x}{N}. \tag{8.2}$$

8.2.3 Ecart-type

Dans la population, tous les individus n'ont pas la même valeur de la variable X. La variabilité dans la population peut être mesurée par l'écart-type (standard deviation) de population défini comme suit :

$$\sigma = \sqrt{\frac{\sum (x-\mu)^2}{N}}. \tag{8.3}$$

Il s'agit aussi d'un paramètre de population inconnu. Notons au dénominateur la division par N et non pas par $N-1$. Enfin, la variance de population n'est autre que le carré de l'écart-type, σ^2. Notons enfin, que dans le cas d'une variable binaire, l'écart-type s'écrit

$$\sigma = \sqrt{\pi(1-\pi)}. \tag{8.4}$$

8.2.4 Autres paramètres de population

Il y a bien sûr une multitude d'autres paramètres de population que la moyenne (ou proportion) et l'écart-type. Tous les paramètres (ou caractéristiques) d'échantillon que nous avons définis aux Chapitres 4 à 7 ont leur équivalent "théorique", c'est-à-dire de population.

Ainsi on peut définir les "percentiles" $P\alpha$ ($0 \leq \alpha \leq 100\%$) de la population, en particulier la médiane $M = P50$.

Dans le cas de deux variables X et Y, on peut définir le coefficient de corrélation théorique

$$\rho = \frac{\sum xy - (\sum x)(\sum y)/N}{\sqrt{[\sum x^2 - (\sum x)^2/N][\sum y^2 - (\sum y)^2/N]}} \tag{8.5}$$

et montrer que $-1 \leq \rho \leq 1$.

De même, si on souhaite mesurer l'accord entre deux observateurs A et B sur l'ensemble des sujets d'une population, on note κ le coefficient Kappa de Cohen théorique et on a

$$\kappa = \frac{\pi_0 - \pi_1}{1 - \pi_1} \tag{8.6}$$

avec $\pi_0 = \sum_i \pi_{ii}$ et $\pi_1 = \sum_i \pi_{i.} \times \pi_{.i}$, où π_{ij} est la proportion théorique d'avis où l'observateur A a attribué le critère i et l'observateur B le critère j ($i, j = 1, \ldots, k$).

Par ailleurs, dans un problème de régression, la droite de régression s'écrit

$$\mu_y(x) = \alpha + \beta.x \tag{8.7}$$

où $\mu_y(x)$ est la moyenne de la variable Y dans la population de sujets ayant $X = x$, α et β sont respectivement l'ordonnée à l'origine et la pente (inconnues) de la droite de régression.

Lorsque X est une variable qualitative à k modalités, on désigne par $\pi_1, \ldots, \pi_k$, les proportions de chacune des modalités dans la population, avec $\pi_1 + \cdots + \pi_k = 1$.

8.3 Distributions théoriques

Lorsqu'on parle de population, l'histogramme d'une variable continue X s'apparente davantage à une courbe, appelée "fonction de densité" (density function) en raison du caractère infini du nombre de valeurs prises par la variable et du nombre de sujets dans la population (voir section 13.2). L'aire sous la courbe de densité vaut 1 (ou 100%). Pour une variable discrète, le diagramme en bâtons reste d'application mais les proportions sont à présent théoriques ; on parle alors de la distribution théorique (distribution function) de X (voir section 12.3).

8.4 Echantillon et Estimation

Les paramètres de population, de même que les distributions théoriques, étant inconnus, il convient de les estimer à partir d'un échantillon extrait de la population. En

effet, il est impensable (voire impossible) en général d'étudier tous les individus de la population.

En conséquence, un paramètre de population n'est jamais connu avec certitude, mais on dispose d'une estimation connue (estimation ou estimate) plus ou moins fiable de celui-ci. Ainsi, à partir d'un échantillon d'effectif n extrait de la population (voir Chapitres 2 à 7), on peut dire que

- $\bar{x}$ estime μ
- p estime π
- s estime σ
- r estime ρ
- a estime α
- b estime β
- $\hat{P}_\alpha$ estime P_α, etc.

On constate qu'on utilise les lettres latines pour désigner les estimations. Faute de lettres grecques pour les paramètres de population, on utilise un accent circonflexe "ˆ" sur le paramètre théorique pour désigner l'estimation, du moins si le contexte l'impose ! Ainsi, si θ est un paramètre de population, $\hat{\theta}$ désigne une estimation de θ.

De même, le diagramme en bâtons "estime" la distribution des fréquences théoriques, l'histogramme la fonction de densité, et le diagramme cumulatif de l'échantillon celui de la population, appelé "fonction de répartition".

8.5 Echantillonnage

L'extraction d'un échantillon (sample) d'une population est une opération difficile, périlleuse mais capitale en statistique. On souhaite que l'échantillon soit "représentatif" de la population afin d'éviter tout biais (bias) ! L'Homme étant un "piètre échantillonneur", il faut faire appel à des mécanismes particuliers d'échantillonnage afin de limiter au maximum l'intervention humaine.

8.5.1 Définition

L'échantillonnage (sampling) est un mécanisme (on dit aussi "procédé") qui permet de tirer des échantillons d'une population. Il y a différents types d'échantillonnage mais le plus classique est appelé l'échantillonnage simplement fortuit.

L'échantillonnage simplement fortuit (simple random sampling) répond à trois critères :

C1. L'effectif n de l'échantillon est *fixé* à l'avance.

C2. Les tirages successifs des individus (ou objets) de la population se font *au hasard* et de façon *indépendante.*

C3. Les tirages successifs se font d'une population *invariante.*

Notons que les termes "au hasard", "fortuit" et "aléatoire" signifient la même chose. Le caractère aléatoire d'un échantillonnage est capital, laissant au seul hasard (chance) de tirer les individus de la population. Il n'y a donc pas de choix délibéré !

Le critère C3 est toujours satisfait lorsque la population est infinie ($N = \infty$). Lorsque la population est finie ($N < \infty$), pour satisfaire à la condition C3, il faut replacer le sujet dans la population après l'avoir extrait afin de ne pas modifier l'effectif de la population. Ainsi donc, un même sujet peut être retiré deux ou plusieurs fois. On parle alors d'échantillonnage *avec remplacement*; dans le cas contraire, l'échantillonnage se fait *sans remplacement*. A titre d'exemple, le Lotto n'est pas un échantillonnage simplement fortuit car il ne satisfait pas au critère C3.

Il existe d'autres méthodes d'échantillonnage. Citons à titre d'exemple, l'échantillonnage stratifié (stratified sampling), l'échantillon systématique (systematic sampling) ou l'échantillonnage en grappes (cluster sampling). Ce dernier est fréquemment utilisé dans les enquêtes de Santé, par exemple dans le "Health Interview Survey (HIS)" réalisé en Belgique en 1997.

8.5.2 Méthode pratique

En pratique, supposons que les individus de la population soient numérotés de 1 à N. Pour tirer un échantillon simplement fortuit, on peut avoir recours à des "tables de nombres aléatoires" (random numbers). Un exemplaire d'une page d'une telle table se trouve à l'Annexe IV. Ces nombres ont été générés au hasard par la RAND Corporation et publiés dans une série de volumes. Il suffit de choisir au hasard une page d'un volume et de pointer au hasard un endroit sur la page qui constituera le point de départ de la sélection des n nombres de 1 à N.

De nos jours, on a davantage recours à l'ordinateur pour générer des nombres au hasard. Ces nombres sont dits "pseudo-aléatoires" car l'ordinateur utilise un algorithme pour les générer et nécessite au préalable de la part de l'utilisateur l'introduction d'un premier nombre, appelé "germe (seed)" de la série. En utilisant le même germe, on génère chaque fois la même série, ce qui peut présenter des avantages notamment dans les méthodes de simulation. Bien sûr, si le germe est connecté à l'horloge interne de l'ordinateur, les séries générées sont toutes différentes.

8.6 Nombres d'échantillons possibles

Tout étudiant ou chercheur doit être conscient du nombre extraordinairement élevé d'échantillons d'effectif n que l'on peut tirer d'une population d'effectif N. Ce nombre, que nous noterons $N(n)$, dépasse souvent l'imagination et il est utile de s'en souvenir.

On distingue quatre cas de figures selon que l'échantillonnage est fait sans ou avec remplacement et que l'on tient compte ou non de l'ordre de tirage des individus (voir Tableau 8.1). Notons que le cas le plus fréquent est l'échantillonnage sans remplacement et sans tenir compte de l'ordre.

Tableau 8.1 Nombre d'échantillons possibles d'effectif n extraits d'une population d'effectif N en fonction du type d'échantillonnage

	Remplacement	
	Sans	Avec
Ordre du tirage		
Non	C_N^n	C_{N+n-1}^n
Oui	A_N^n	N^n

Pour rappel,

$$C_N^n = \frac{N!}{n!(N-n)!} \tag{8.8}$$

$$A_N^n = \frac{N!}{(N-n)!} \tag{8.9}$$

avec $N!$ (factoriel de N) $= 1 \times 2 \times \cdots \times N$ et par convention $0! = 1$.

A titre d'exemples, supposons que l'on dispose d'une population d'effectif $N = 40$ et que l'on souhaite y tirer un échantillon d'effectif $n = 10$.

- Pour un échantillonnage sans remplacement, si l'on ne tient pas compte de l'ordre, on a $C_{40}^{10} = 847.660.528$ possibilités différentes. En tenant compte de l'ordre des tirages, ce nombre passe à $A_{40}^{10} = 3,076 \times 10^{15}$ possibilités !
- Pour un échantillonnage avec remplacement, si on ne tient pas compte de l'ordre, on a $C_{49}^{10} = 8.217.822.536$ possibilités différentes ! En tenant compte de l'ordre des tirages, ce nombre passe à $40^{10} = 1,049 \times 10^{16}$ possibilités !

Compte tenu de la petite taille de la population, la magnitude de ces nombres laisse rêveur. Est-il possible d'imaginer le nombre d'échantillons d'effectif $n = 1000$ que l'on peut tirer d'une population de plusieurs millions d'habitants (comme dans un sondage d'opinion, par exemple) ?

8.7 Erreur type d'une moyenne arithmétique

Introduisons le concept d'erreur type pour la moyenne d'échantillon $\bar{x}$. La méthodologie est semblable pour d'autres caractéristiques d'échantillons (proportion, écart-type, corrélation, pente de régression, etc.).

Soit X la variable étudiée, μ et σ sa moyenne et son écart-type dans la population d'effectif N. Pour rappel, μ et σ sont inconnus.

Soit $\bar{x}$ la moyenne d'*un* échantillon simplement fortuit d'effectif n extrait de la population et s l'écart-type de l'échantillon. Il s'agit de l'échantillon que l'on a obtenu en pratique.

8.7.1 Définition

Considérons l'ensemble des $N(n)$ échantillons possibles que l'on peut extraire de la population. Nous l'avons vu, ce nombre est considérable et nous ne disposons en pratique que d'un seul échantillon !

Notons $\bar{x}_i$ la moyenne du i-ème échantillon ($i = 1, \ldots, N(n)$). On peut montrer (Théorème de la théorie de l'échantillonnage) que la moyenne des $\bar{x}_i$ vaut μ, la moyenne inconnue de la population. Donc, en utilisant la formule (8.1),

$$\mu(\bar{x}) = \frac{\sum \bar{x}}{N(n)} = \mu. \tag{8.10}$$

Ce résultat est remarquable car il montre que la moyenne d'échantillon est un estimateur fidèle (unbiased estimator) de la moyenne de la population.

Il est évident que les moyennes d'échantillon $\bar{x}_i$ ($i = 1, \ldots, N(n)$) varient entre elles, car les échantillons varient de l'un à l'autre. On peut mesurer cette variabilité d'échantillonnage (sampling variation) des moyennes en calculant l'écart-type par la formule (8.3) appliquée aux moyennes $\bar{x}_i$. On a

$$\sigma(\bar{x}) = \sqrt{\frac{\sum(\bar{x} - \mu)^2}{N(n)}}. \tag{8.11}$$

Toutefois, le théorème de la théorie de l'échantillonnage montre que

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}. \tag{8.12}$$

La quantité $\sigma(\bar{x})$ est appelée l'erreur type (standard error ou SE) de $\bar{x}$. On dit aussi "précision statistique". L'équation (8.12) est remarquable dans la mesure où elle montre que la variabilité entre les moyennes des échantillons est directement proportionnelle à la variabilité des sujets dans la population et inversement proportionnelle à la racine carrée de l'effectif de l'échantillon. Donc pour diminuer $\sigma(\bar{x})$ de moitié, il faut quadrupler l'effectif n.

8.7.2 Estimation de l'erreur type

L'erreur type de la moyenne arithmétique $\bar{x}$ donnée par l'équation (8.12) n'est pas connue puisque σ est inconnu. En pratique, nous avons vu que l'écart-type de l'échantillon s est une estimation de σ. On peut dès lors remplacer σ et par s et obtenir une estimation de $\sigma(\bar{x})$, notée $s(\bar{x})$ ou $SE(\bar{x})$, par la relation

$$s(\bar{x}) = \frac{s}{\sqrt{n}}. \tag{8.13}$$

On dit dès lors que l'erreur type de la moyenne arithmétique $\bar{x}$ d'un échantillon d'effectif n est égale à $s/\sqrt{n}$. On voit qu'elle est égale à l'écart-type de l'échantillon divisé par $\sqrt{n}$.

Si l'erreur type est grande, cela signifie que les moyennes d'échantillon sont très variables les unes des autres. La moyenne $\bar{x}$ en tant qu'estimation de μ est peu précise. Par contre, si l'erreur type est faible, il y a peu de variabilité d'un échantillon à l'autre. La moyenne $\bar{x}$ en tant qu'estimation de μ est donc très précise. Pour augmenter la précision de l'estimation, c'est-à-dire diminuer l'erreur type, il suffit d'augmenter n. Malheureusement, ceci n'est pas toujours possible en raison du manque de ressources humaines, matérielles ou financières.

Ce qu'il y a de remarquable dans l'équation (8.13), c'est qu'il est possible d'estimer la variabilité des $\bar{x}_i$ $(i = 1, \ldots, N(n))$, alors qu'on ne dispose en pratique que d'un seul échantillon.

8.8 Erreur type d'autres estimateurs

La notion d'erreur type est non seulement associée à la moyenne arithmétique mais à toute autre caractéristique d'échantillon. Malheureusement, il n'existe pas toujours une formule simple qui donne l'erreur type, comme c'est le cas pour la moyenne. Il faut alors recourir à d'autres approches.

8.8.1 Formules connues

En plus de la moyenne, l'erreur type (SE) est connue pour d'autres caractéristiques d'échantillon, parfois sous certaines conditions supplémentaires. Outre l'expression (8.13), on dispose aussi des formules suivantes :

– proportion (p) :

$$SE(p) = \sqrt{\frac{p(1-p)}{n}} \tag{8.14}$$

– médiane (M) d'une loi Normale (avec $\pi = 3.1416\ldots$) :

$$SE(M) = s\sqrt{\frac{\pi}{2n}} \tag{8.15}$$

– variance (s^2) d'une loi Normale :

$$SE(s^2) = s^2\sqrt{\frac{2}{n-1}} \tag{8.16}$$

– pente de la droite de régression (b) :

$$SE(b) = \frac{s_{y|x}}{\sqrt{\sum(x-\bar{x})^2}} \tag{8.17}$$

– ordonnée à l'origine de la droite de régression (a) :

$$SE(a) = s_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}} \tag{8.18}$$

– Kappa de Cohen ($\hat{\kappa}$) :

$$SE(\hat{\kappa}) = \frac{1}{(1-p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum_{i=1}^{k} p_{i.}p_{.i}(p_{i.} + p_{.i})} \tag{8.19}$$

En comparant les formules (8.13) et (8.15), on constate que, dans le cas d'une loi Normale, la médiane M est un estimateur moins précis de la moyenne μ que la moyenne arithmétique $\bar{x}$, puisque $SE(M) > SE(\bar{x})$.

8.8.2 Méthode du Bootstrap

Lorsqu'il n'existe pas de formule analytique pour calculer l'erreur type d'une estimation, c'est-à-dire d'une caractéristique d'échantillon, on peut utiliser la méthode du "Bootstrap" développée par Bradley Efron (1970).

Soit un échantillon simplement fortuit d'effectif n que nous notons $\omega = \{x_1, \ldots, x_n\}$ et $\hat{\theta}$ une caractéristique d'échantillon (par exemple, $\hat{\theta}$ est le percentile $P25$, la corrélation r entre deux variables, l'écart interquartile H, voire l'étendue E). Comment calculer $SE(\hat{\theta})$?

La méthode du bootstrap procède comme suit :

B1. On rééchantillonne aléatoirement et avec remplacement l'échantillon ω jusqu'à obtenir un nouvel échantillon de même effectif n. Notons ω_j l'échantillon ainsi obtenu. Calculons la caractéristique d'échantillon $\hat{\theta}_j$ pour cet échantillon.
Notons que le nombre possible d'échantillons "bootstrapés" obtenus de cette manière à partir de l'échantillon initial ω est égal à (voir Tableau 8.1).

$$N(n) = n(n) = C_{N+n-1}^{n} = C_{2n-1}^{n} \tag{8.20}$$

Ce nombre devient vite formidable. Par exemple, pour un échantillon initial d'effectif $n = 15$, il y a $C_{29}^{15} = 77.558.760$, soit plus de 77 millions de possibilités !!

B2. On répète l'opération B1 un grand nombre de fois. Notons B ce nombre, que l'on choisit en général entre 200 et 300. A chaque échantillonnage, on calcule la caractéristique d'échantillon $\hat{\theta}_j$. On obtient ainsi un échantillon $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$.

B3. L'erreur type de $\hat{\theta}$, notée $SE(\hat{\theta})$, n'est autre que l'écart-type de l'échantillon $\{\hat{\theta}_1, \ldots, \hat{\theta}_B\}$ et en utilisant la formule (3.5), on a

$$SE(\hat{\theta}) = \sqrt{\frac{\sum_j \hat{\theta}_j^2 - (\sum_j \hat{\theta}_j)^2/B}{B-1}}. \tag{8.21}$$

8.9 Exemples

Calculons l'erreur type pour quelques-uns des exemples vus précédemment.

L'erreur type de l'âge moyen des médecins généralistes est égale à (voir Tableau 3.1)

$$SE(\bar{x}) = \frac{s}{\sqrt{n}} = \frac{10.7}{\sqrt{355}} = 0.568.$$

Par contre, l'erreur type de la proportion de médecins agréés est égale à

$$SE(p) = \sqrt{p(1-p)/n} = \sqrt{0.93 \times 0.07/352} = 0.0136.$$

Dans le problème de régression linéaire (voir Section 7.4), on voit aisément en appliquant les formules (8.17) et (8.18) que les erreurs types de la pente et de l'ordonnée à l'origine valent respectivement

$$\begin{aligned} SE(b) &= \frac{s_{y|x}}{\sqrt{\sum(x-\bar{x})^2}} = 0.00139 \\ SE(a) &= s_{y|x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x-\bar{x})^2}} = 0.177. \end{aligned}$$

A titre d'illustration de la méthode du Bootstrap, nous avons calculé l'erreur type du coefficient de corrélation r entre l'âge du médecin et son expérience professionnelle obtenu à partir des observations du Tableau 6.2.

En fixant $B = 250$, l'écart-type des 250 coefficients de corrélation obtenus par bootstrap de l'échantillon initial vaut 0.0631, de sorte que $SE(r) = 0.0631$. En fixant $B = 1000$, on a obtenu $SE(r) = 0.0597$.

Chapitre 9

Intervalle de confiance

9.1 Introduction

La notion d'intervalle de confiance (confidence interval) est intimement liée à celle d'erreur type. Elle permet d'apprécier avec un niveau de confiance donné (en général 95%) la proximité entre un paramètre de population (θ) et son estimation ($\hat{\theta}$) obtenue à partir d'un échantillon d'effectif n. Plus l'intervalle de confiance est étroit, plus on est proche de la valeur "vraie" de la population. Un intervalle de confiance est souvent appelé dans le jargon une "fourchette", dans la mesure où il fournit deux limites devant en principe contenir le paramètre de population θ. Les limites de l'intervalle de confiance sont donc calculées à partir de l'échantillon.

L'intervalle de confiance fournit davantage d'informations que l'estimation seule car il informe le lecteur sur les marges de variabilité des résultats de l'étude, sur la précision avec laquelle le paramètre θ est estimé. Il permet aussi dans une certaine mesure de comparer des populations ou de voir si le paramètre de population θ est égale à une valeur donnée θ_0.

9.2 Théorie de l'échantillonnage

Afin de simplifier la situation, reprenons l'exemple de la moyenne d'une population (plus précisément la moyenne de la variable X dans la population). Donc $\theta = \mu$. Comme précédemment, notons σ l'écart-type de X dans la population.

Soit un échantillon simplement fortuit d'effectif n extrait de la population que l'on note $\omega = \{x_1, \ldots, x_n\}$.

Comme on l'a fait précédemment (section 8.7), considérons l'ensemble des $N(n)$ échantillons possibles d'effectif n que l'on peut extraire de la population et calculons chaque fois la moyenne $\bar{x}_i$ $(i = 1, \ldots, N(n))$.

La théorie de l'échantillonnage nous apprend que (voir Equations 8.10 et 8.11) :

$$\begin{aligned} R1 &: \mu(\bar{x}) = \mu \\ R2 &: \sigma(\bar{x}) = \sigma/\sqrt{n}. \end{aligned}$$

Par ailleurs, on démontre que lorsque n tend vers l'infini ($n \to \infty$) la distribution statistique (histogramme) des $\bar{x}$ tend vers une loi Normale (gaussienne) dont la moyenne et l'écart-type sont donnés par les expressions ci-dessus, ce que l'on écrit

$$R3 : \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right). \tag{9.1}$$

Ces trois résultats de la théorie de l'échantillonnage, dont le troisième ($R3$) est asymptotique (loi des grands nombres), ont une implication pratique considérable.

Comme nous le verrons ultérieurement, la propriété $R3$ implique que 95 fois sur 100 (donc pour 95% des échantillons extraits de la population) la moyenne $\bar{x}$ tombe dans l'intervalle $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$, ce que l'on écrit

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1.96\frac{\sigma}{\sqrt{n}}. \tag{9.2}$$

Cette double inégalité peut aussi s'écrire

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \tag{9.3}$$

permettant ainsi d'isoler le paramètre de population dans l'intervalle $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$.

L'intervalle (9.3) est appelé "*intervalle de confiance à 95%*" pour la moyenne μ de la population. Il s'interprète comme suit : "dans 95% des cas (c'est-à-dire pour 95% des échantillons extraits de la population), la fourchette $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ recouvre μ". En conséquence, dans 5% des cas, la fourchette ne contient pas μ.

En clair, on est "presque sûr" que l'intervalle de confiance (9.3) contient μ !

9.3 Intervalle de confiance à 95%

En pratique, dans l'intervalle de confiance (9.3), le paramètre σ n'est pas connu. Toutefois, comme on l'a fait pour l'erreur type, on dispose d'une estimation s de σ. On peut donc remplacer dans (9.3) σ par s et ainsi obtenir l'intervalle de confiance à 95% estimé, soit

$$IC\ 95\% : \bar{x} - 1.96\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{s}{\sqrt{n}}. \tag{9.4}$$

Le lecteur attentif constatera immédiatement que l'intervalle de confiance est obtenu en retirant et en ajoutant à $\bar{x}$ l'erreur type de $\bar{x}$, soit $SE(\bar{x})$, multiplié par le facteur 1.96, soit

$$\bar{x} - 1.96\ SE(\bar{x}) \leq \mu \leq \bar{x} + 1.96\ SE(\bar{x}). \tag{9.5}$$

Donc si l'on connaît l'erreur type de $\bar{x}$, on peut immédiatement déterminer l'intervalle de confiance à 95%. Sous certaines conditions, cette proposition est applicable en toutes généralités et on peut écrire

$$IC\ 95\% : \hat{\theta} - 1.96\ SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + 1.96\ SE(\hat{\theta}). \tag{9.6}$$

Remarque : En pratique, pour des raisons de facilité, on remplace souvent le facteur 1.96 par 2 ! En réalité, 1.96 est le percentile à 97.5% de la distribution Normale de moyenne 0 et d'écart-type 1.

9.4 Intervalle de confiance pour une proportion

Une des plus importantes applications de la théorie des intervalles de confiance est celle de l'estimation d'une proportion π. Pour rappel, π est la moyenne d'une variable binaire ($X = 0$ ou 1).

Pour une proportion estimée p, nous avons vu que l'erreur type vaut (Equation 8.14)

$$SE(p) = \sqrt{p(1-p)/n}.$$

Dès lors, en appliquant (9.6), où $\theta = \pi$ et $\hat{\theta} = p$, on obtient un intervalle de confiance à 95% pour π.

$$IC\ 95\% : p - 1.96\sqrt{\frac{p(1-p)}{n}} \leq \pi \leq p + 1.96\sqrt{\frac{p(1-p)}{n}}. \tag{9.7}$$

A titre d'exemple, dans une étude sur le tabagisme à l'université, on s'est proposé d'estimer la proportion π d'étudiants fumeurs. La population estudiantine était trop importante ($N > 15000$), on a interrogé un échantillon simplement fortuit de $n = 200$ étudiants sur leurs habitudes tabagiques. Cinquante-huit ont répondu qu'ils fumaient, soit une proportion estimée $p = 58/200 = 0.29$. L'erreur type de cette estimation vaut

$$SE(p) = \sqrt{p(1-p)/n} = \sqrt{0.29 \times 0.71/200} = 0.0319.$$

On peut dès lors affirmer avec une confiance de 95% (ou avec une incertitude de 5%) que l'intervalle

$$\begin{aligned} 0.29 - 1.96 \times 0.0319 &\leq \pi \leq 0.29 + 1.96 \times 0.0319 \\ 0.29 - 0.0624 &\leq \pi \leq 0.29 + 0.0624 \\ 0.228 &\leq \pi \leq 0.352 \end{aligned}$$

contient la vraie valeur de la proportion de fumeurs à l'université. Celle-ci se situe donc plus que vraisemblablement entre 22.8% et 35.2%.

Si l'on se souvient qu'une étude faite il y a 20 ans sur la quasi totalité des étudiants avait donné une proportion de 41% de fumeurs, on peut conclure qu'aujourd'hui il y a "significativement" moins de fumeurs, puisque la valeur 41% ne se situe pas dans la fourchette $[22.8 - 35.2\%]$.

Revenons au problème de l'élection présidentielle, où le candidat A avait reçu 52% des suffrages et le candidat B 48% sur base d'un sondage auprès de 1000 personnes : peut-on affirmer que le candidat A est vainqueur ? Puisque $p = 0.52$, $SE(p) = \sqrt{0.52 \times 0.48/1000} =$

0.0158 et $1.96 \times SE(p) = 1.96 \times 0.0158 = 0.031$. Dès lors, l'intervalle de confiance à 95% a pour limites 0.52 ± 0.031, soient 0.489 et 0.551, respectivement. On peut donc écrire $48.9\% \leq \pi_A \leq 55.1\%$ et le candidat peut encore être battu puisque la valeur $\pi_A = 50\%$ est dans la fourchette. Avec les mêmes proportions observées, 52% et 48%, il eut fallu disposer d'un effectif de $n = 2397$ sujets pour affirmer avec une confiance de 95% que le candidat A l'emportât. Pour cela, il suffit de résoudre l'équation $1.96\sqrt{0.52 \times 0.48/n} = 0.02$.

9.5 Généralisation

Dans les formules précédentes, on a défini un intervalle de confiance à 95% (ou au niveau d'incertitude α de 5%), conduisant au facteur de multiplication $Q_Z(0.975) = 1.96$.

Si l'on veut augmenter le niveau de confiance à 99% (niveau d'incertitude de 1%), par exemple, il faut changer le facteur de multiplication et utiliser $Q_Z(0.995) = 2.58$. On a alors

$$IC\ 99\% : \hat{\theta} - 2.58\ SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + 2.58\ SE(\hat{\theta}). \tag{9.8}$$

En toutes généralités, l'intervalle de confiance bilatéral au niveau de confiance $(1-\alpha)$ pour θ s'écrit

$$IC\ (1-\alpha)\% : \hat{\theta} - Q_Z(1-\alpha/2)\ SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + Q_Z(1-\alpha/2)\ SE(\hat{\theta}) \tag{9.9}$$

Dans cette expression, α est le niveau d'incertitude et $Q_Z(1-\alpha/2)$ le quantile $(1-\alpha/2)$ de la loi Normale (voir Tableau 13.1 et Table A en annexe).

Il y a deux manières de réduire un intervalle de confiance :

- en augmentant n, l'effectif de l'échantillon
- en augmentant α le niveau d'incertitude, c'est-à-dire en diminuant $1-\alpha$ le niveau de confiance.

Remarque : Il ne faut pas confondre les notions d'intervalle de référence (ou de tolérance) comme on l'a vu au Chapitre 3 et d'intervalle de confiance telle que définie dans ce chapitre. L'intervalle de référence définit un espace de variabilité des valeurs individuelles d'un échantillon tandis qu'un intervalle de confiance définit un espace de variabilité d'une caractéristique d'échantillon sur l'ensemble des échantillons possibles.

Chapitre 10

Probabilité

10.1 Introduction

La théorie des probabilités (probability theory) permet le passage de la statistique descriptive à la statistique inférentielle. Elle constitue en réalité la base de l'inférence statistique. Pour le mathématicien, la probabilité est une branche mathématique comme l'algèbre ou l'analyse fonctionnelle. A fortiori, la statistique n'est qu'une application de la théorie des probabilités qui en comporte d'autres comme les processus stochastiques (aléatoires), les séries temporelles, la théorie de la décision ou celle des files d'attente. L'histoire des probabilités est ancienne et on rapporte que le premier mathématicien à calculer une probabilité est Girolano Cordano, un italien qui vécut de 1501 à 1576.

Le terme "probabilité" fait partie de notre langage quotidien, au même titre que ses synonymes de "chance" ou de "risque". Il est intimement lié aux concepts d'incertitude et de hasard. On dira d'un accidenté grave qu'il a de fortes chances de s'en tirer, d'une complication qu'elle est peu probable, d'une intervention chirurgicale qu'elle est à haut risque (autrement dit que ses chances de succès sont faibles). En réalité, sans le savoir, nous calculons continuellement des probabilités dans notre vie quotidienne, chaque fois que nous sommes confrontés à des situations incertaines. Nous savons qu'un événement de probabilité faible a peu de chances de se produire et qu'un événement de probabilité élevée se produira presque certainement.

Dès lors, un chapitre consacré à la théorie des probabilités apparaît indispensable dans un cours de statistique pour se familiariser avec ce concept important et ainsi mieux comprendre la puissance de l'outil statistique dans la démarche scientifique.

Sans entrer dans des considérations mathématiques trop évoluées, nous suivons le cheminement proposé par H. Breny (1922-1991) pour définir la probabilité.

10.2 Phénomène fortuit

L'expérience commune montre qu'il convient d'admettre qu'il existe des phénomènes fortuits (on dit aussi "processus aléatoires"), c'est-à-dire des phénomènes qui ont plusieurs résultats possibles, dont un seul se produit à chaque réalisation du phénomène,

sans qu'aucune cause ne rende raison de l'occurrence de tel résultat plutôt que de tel autre. On dit que le résultat du phénomène "est dû au hasard". On désigne par $\mathcal{F}$ le phénomène fortuit (random process) auquel on s'intéresse.

On appelle "Théorie des probabilités", la théorie mathématique des phénomènes fortuits.

Les phénomènes fortuits les plus connus sont les "jeux de hasard" (pile/face, dés, cartes, roulette, loterie, autres jeux de casino). En Biologie, citons la transmission des caractères héréditaires (loi de Mendel 1865 et théorie des chromosomes 1900). La physique des micro-particules, en particulier les émissions radioactives, sont des phénomènes aléatoires. Les variations associées à la production en série relèvent aussi de ce domaine. En statistique, le phénomène fortuit central est l'échantillonnage (sampling process) qui comme on l'a vu au Chapitre 8 est un mécanisme aléatoire qui permet de tirer des échantillons d'une population.

10.3 Catégorie d'épreuve

La catégorie d'épreuve (world) est l'ensemble des résultats possibles d'un phénomène fortuit $\mathcal{F}$. On la note Ω ou $\Omega(\mathcal{F})$. Un élément de la catégorie d'épreuve est appelé singleton (ou événement élémentaire) et noté ω.

A titre d'exemples, citons

- jet d'une pièce de monnaie : $\Omega = \{P, F\}$
- jet d'une dé : $\Omega = \{1, 2, 3, 4, 5, 6\}$
- jet de deux dés discernables : $\Omega = \{(i, j);\ i, j = 1, \ldots, 6\}$
- jet de deux dés non discernables : $\Omega = \{(i, j) : i \leq j,\ i, j = 1, \ldots, 6\}$
- lotto : $\Omega = \{(i_1 < i_2 < i_3 < i_4 < i_5 < i_6),\ i_j \in \{1, \ldots, 42\},\ j = 1, \ldots, 6\}$
- échantillonnage simplement fortuit d'effectif n d'une population d'effectif N : $\Omega = \{\omega_i,\ i = 1, \ldots, N(n)\}$.

On constate que même pour des phénomènes fortuits d'apparence simple, la catégorie d'épreuve peut être complexe. On note $N(\Omega)$ le cardinal (l'effectif) de la catégorie d'épreuve Ω. Dans la suite de ce chapitre, nous prendrons comme exemple le jet de deux dés discernables (par exemple, un dé Rouge et un dé Bleu). La catégorie d'épreuve comporte $N(\Omega) = 36$ résultats possibles. Au lotto, $N(\Omega) = 5.245.786$.

10.4 Evénement

Un événement (event), noté E, est un sous-ensemble de la catégorie d'épreuve Ω. Rappelons que le singleton ω est un événement élémentaire. En général, le cardinal d'un événement $N(E) > 1$. Comme dans la théorie des ensemble, $\emptyset$ est l'ensemble vide et $N(\emptyset) = 0$.

Si, lors de la réalisation du phénomène fortuit $\mathcal{F}$, $\omega \in E$, on dit que l'événement E *s'est produit.*

A titre d'exemple, dans le jet de deux dés discernables, on peut considérer l'événement $E = \{(i,j) \in \Omega : i = j\}$, "les deux dés donnent le même résultat". On constate immédiatement que $N(E) = 6$. Si, en jetant les deux dés, on obtient (3,3) l'événement E s'est produit. Si au contraire, on obtient (6,2), il ne s'est pas produit !

10.5 Partition

On dit que les événements $E_1, E_2, \ldots, E_k$ forment une "partition" de Ω si
C1 : les événements sont disjoints deux à deux : $E_i \cap E_j = \emptyset$, $\forall i \neq j$
C2 : l'union des événements est égale à la catégorie d'épreuve

$$\bigcup_{i=1}^{k} E_i = \Omega.$$

La partition est une "vue macroscopique" de la catégorie d'épreuve. Il existe de multiples façons de partitionner une catégorie d'épreuve. Par exemple, les événements $E_1 = \{(i,j) \in \Omega : i = j\}$, $E_2 = \{(i,j) \in \Omega : i > j\}$ et $E_3 = \{(i,j) \in \Omega : i < j\}$, constituent une partition de Ω dans le jet de deux dés discernables. Mais il en existe d'autres.

10.6 Probabilité

10.6.1 Définition

La probabilité (probability) est une notion associée aux événements d'une catégorie d'épreuve $\Omega(\mathcal{F})$.

Par définition, la probabilité d'un événement E, notée $P(E)$, est une mesure de sa propension à l'occurrence. En d'autres termes, c'est une mesure de la chance qu'a l'événement E de se produire.

10.6.2 Propriétés

– Tout événement a une probabilité non négative de se produire

$$\forall E \subset \Omega, \quad P(E) \geq 0. \tag{10.1}$$

– Si deux événements sont disjoints, la probabilité que l'un ou l'autre se produise est la somme des probabilités (axiome d'additivité des probabilités)

$$\begin{array}{c} \forall E_i,\ E_j \subset \Omega\ ,\ E_i \cap E_j = \emptyset \\ P(E_i \cup E_j) = P(E_i) + P(E_j) \end{array} \tag{10.2}$$

– La probabilité de la catégorie d'épreuve Ω (événement certain) est égale à 1,

$$P(\Omega) = 1. \tag{10.3}$$

Il en résulte que

$$\forall E \subset \Omega, \quad 0 \leq P(E) \leq 1. \tag{10.4}$$

La probabilité d'un événement est donc toujours comprise entre 0 et 1 (0 et 100% si on exprime la probabilité en pourcent).

Lorsqu'un événement E a une probabilité nulle, $P(E) = 0$, on dit que l'événement E est "impossible". Au contraire, si $P(E) = 1$, on dit que l'événement E est "certain".

10.7 Calcul de probabilité

10.7.1 Approche mathématique

Pour calculer la probabilité $P(E)$ d'un événement $E \subset \Omega$, on peut avoir recours à une approche mathématique (théorie des ensembles). Si on suppose que tous les singletons $\omega \in \Omega$ sont équiprobables (principe qui résulte de la définition même d'un phénomène fortuit $\mathcal{F}$), alors vu les propriétés (10.2) et (10.3),

$$P(E) = \frac{N(E)}{N(\Omega)}. \tag{10.5}$$

La probabilité est donc le rapport entre le nombre de résultats "favorables" à l'événement E et le nombre total de résultats possibles du phénomène fortuit.

A titre d'illustration dans le jet de deux dés discernables, si $E = \{(i, j) : i = j\}$, alors $P(E) = 6/36 = 1/6$. De même, si $E = \{(i, j) : i > j\}$, $P(E) = 15/36 = 5/12$.

10.7.2 Approche empirique

Lorsqu'il n'y a pas de modèle mathématique du processus aléatoire ou que les singletons ne sont pas tous équiprobables (par exemple, une pièce de monnaie non balancée, un dé non homogène, une différence entre le poids des boules au lotto, etc.), on peut "calculer" la probabilité d'un événement E de façon empirique, en procédant comme suit :

On réalise n fois le phénomène fortuit et on comptabilise le nombre $n(E)$ de réalisations "favorables" à l'événement E. Dès lors

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}. \tag{10.6}$$

Pour autant que n soit élevé, $P(E) \simeq n(E)/n$.

Quelle est la probabilité de décéder d'un infarctus du myocarde (IM) dans les 10 premiers jours d'hospitalisation ? Sur $n = 300$ cas consécutifs d'infarctus admis au CHU, on a constaté $n(E) = 24$ décès endéans dix jours. Dès lors, en utilisant l'équation (10.6), $P(E) \simeq 24/300 = 0.08$, soit 8%. On peut donc dire qu'un patient IM a 8% de chance de décéder endéans les 10 jours.

Chapitre 11

Théorème de Bayes

11.1 Introduction

La probabilité d'un événement peut changer en fonction de la connaissance que nous avons d'autres informations en rapport avec l'événement. Ainsi, l'hémophilie étant une maladie rare, la probabilité d'en être atteint pour un individu donné est faible. Toutefois, si la mère du sujet est porteuse de l'anomalie chromosomique, la probabilité pour le descendant de sexe masculin d'être atteint s'en trouve considérablement modifiée.

La séropositivité conserve une prévalence faible dans les pays occidentaux. Néanmoins, elle est beaucoup plus élevée dans certains pays africains.

Ce chapitre introduit la notion de probabilité conditionnelle (conditional probability). Le théorème de Bayes en est une belle illustration et présente de nombreuses applications en médecine.

11.2 Probabilité conditionnelle

Tout phénomène fortuit $\mathcal{F}$ contient généralement des phénomènes fortuits subordonnés $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_q\}$ qui peuvent ou non s'influencer mutuellement.

Si les phénomènes fortuits $\mathcal{F}_i$ $(i = 1, \ldots, q)$ ne s'influencent pas mutuellement, on dit qu'ils sont "indépendants". Par exemple, dans le jet simultané de deux dés homogènes discernables $(\mathcal{F})$, il n'y a aucune raison que le jet du premier dé $(\mathcal{F}_1)$ influence le jet du second dé $(\mathcal{F}_2)$. Il y a indépendance.

Si au contraire les phénomènes fortuits $\mathcal{F}_i$ $(i = 1, \ldots, q)$ interagissent les uns sur les autres, on dit qu'ils sont "dépendants". Par exemple, dans deux tirages successifs sans remise d'une boule dans une urne qui contient 3 boules rouges et une boule blanche, le deuxième tirage $(\mathcal{F}_2)$ dépend fortement du résultat du premier tirage $(\mathcal{F}_1)$ car celui-ci modifie la composition de l'urne.

11.2.1 Définition

Considérons un phénomène fortuit $\mathcal{F}$ de catégorie d'épreuve Ω et composé de deux phénomènes fortuits subordonnés $\mathcal{F}_1$ et $\mathcal{F}_2$ de catégories d'épreuve Ω_1 et Ω_2, respectivement.

Soient E_1 un événement du phénomène fortuit $\mathcal{F}_1$ et E_2 un événement du phénomène fortuit $\mathcal{F}_2$.

Par définition, la probabilité de l'événement E_2 *conditionnelle à l'événement* E_1 est la probabilité de l'événement E_2 sachant que l'événement E_1 s'est produit. Elle est donnée par l'expression

$$P(E_2 \mid E_1) = \frac{P(E_1 \cap E_2)}{P(E_1)} \tag{11.1}$$

où $P(E_1) \neq 0$ est la probabilité de l'événement E_1 et $P(E_1 \cap E_2)$ est la probabilité que les événements E_1 *et* E_2 se produisent simultanément. L'événement $E_1 \cap E_2$ est évidemment un événement du phénomène fortuit $\mathcal{F}$.

L'expression (11.1) peut aussi s'écrire

$$P(E_1 \cap E_2) = P(E_2 \mid E_1).P(E_1) \tag{11.2}$$

qui permet de calculer la probabilité que deux événements relatifs à des phénomènes fortuits différents se produisent ensemble.

On définit de la même manière la probabilité de E_1 conditionnelle à E_2 et on a

$$P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} \tag{11.3}$$

d'où

$$P(E_1 \cap E_2) = P(E_1 \mid E_2).P(E_2). \tag{11.4}$$

En égalant les expressions (11.2) et (11.4), on obtient aussi

$$P(E_2 \mid E_1) = \frac{P(E_1 \mid E_2).P(E_2)}{P(E_1)} \tag{11.5}$$

expression dans laquelle n'interviennent plus que des probabilités conditionnelles et des probabilités d'événements simples.

Remarque : Que l'on ait affaire à un phénomène fortuit $\mathcal{F}$ simple ou complexe, si deux événements E_i et E_j ne sont pas disjoints, $(E_i \cap E_j) \neq \emptyset$, la probabilité que l'un ou l'autre événement se produise s'écrit

$$P(E_i \cup E_j) = P(E_i) + P(E_j) - P(E_i \cap E_j) \tag{11.6}$$

Lorsque les événements sont disjoints, on retrouve l'axiome (10.2).

11.2.2 Exemple

Considérons une unité de soins où sont hospitalisés 24 malades (18 hommes et 6 femmes), de laquelle on sélectionne au hasard et successivement deux personnes. Quelle est la probabilité d'avoir un homme (H) et une femme (F) dans l'échantillon obtenu ? Notons que $P(H \cap F) = P(H_1 \cap F_2) + P(F_1 \cap H_2)$. En utilisant la formule (11.2), on a

$$\begin{aligned} P(H_1 \cap F_2) &= P(F_2 \mid H_1).P(H_1) \\ &= \frac{6}{23} \times \frac{18}{24} = 0.1957. \end{aligned}$$

De même,

$$P(F_1 \cap H_2) = P(H_2 \mid F_1).P(F_1) = \frac{18}{23} \times \frac{6}{24} = 0.1957.$$

En conséquence, $P(H \cap F) = 2 \times 0.1957 = 0.3914$.

Cette probabilité aurait aussi pu être obtenue à partir du phénomène fortuit $\mathcal{F}$. En effet, $P(H \cap F) = 108/276 = 0.3914$ mais l'approche précédente est en général plus simple et plus rapide.

11.3 Axiome de multiplication des probabilités

Si deux phénomènes fortuits $\mathcal{F}_1$ et $\mathcal{F}_2$ sont indépendants, c'est-à-dire s'ils ne s'influencent pas mutuellement, les événements correspondants sont aussi indépendants. On a donc $P(E_2 \mid E_1) = P(E_2)$ puisque la connaissance de E_1 ne modifie pas la probabilité de E_2. L'expression (11.2) peut alors s'écrire $P(E_1 \cap E_2) = P(E_2).P(E_1)$.

Plus généralement,

$$\forall E_i \subset \Omega_1 \text{ et } E_j \subset \Omega_2, \quad P(E_i \cap E_j) = P(E_i).P(E_j). \tag{11.7}$$

C'est l'*axiome de multiplication* des probabilités.

Par exemple, dans le jet de deux dés homogènes discernables, si E_1 désigne l'événement "Premier dé > 4" et E_2 l'événement "Deuxième dé ≤ 5", on peut calculer la probabilité que ces deux événements se produisent simultanément :

$$\begin{aligned} P(D_1 > 4 \cap D_2 \leq 5) &= P(D_1 > 4).P(D_2 \leq 5) \\ &= \frac{2}{6} \times \frac{5}{6} \\ &= \frac{10}{36} = 0.2778. \end{aligned}$$

11.4 Théorème de Bayes

Le Révérend Thomas Bayes (1702-1761) était ministre du culte presbytérien dans la petite ville de Tunbridge Wells, Kent (Angleterre). Il a été rendu célèbre par un article

publié à titre posthume en 1764 par un de ses amis et intitulé "*An essay towards solving a problem in the doctrine of chances*". Il y introduit notamment la notion de "probabilité a priori", ou probabilité subjective, qui ne repose pas nécessairement sur l'existence d'un phénomène fortuit sous-jacent. Cette vision de la probabilité n'est pas partagée par tout le monde et elle a créé une scission au sein de la communauté statistique. Aujourd'hui encore, il y a les statisticiens classiques et les statisticiens bayesiens !

Dans le domaine médical, le théorème de Bayes offre de nombreuses applications en qualité d'outil diagnostic ou pronostic. Il permet par exemple de calculer la probabilité qu'un sujet soit atteint d'une maladie donnée, sachant que le test effectué sur lui s'est avéré "positif".

11.4.1 Définition

Considérons une maladie D et notons $\bar{D}$ l'ensemble des sujets non atteints de cette maladie. Soit un test diagnostic de la maladie dont le résultat est "positif" (T) ou "négatif" ($\bar{T}$).

La probabilité qui intéresse le médecin au premier chef est celle d'être en présence de la maladie chez un sujet "positif", soit $P(D \mid T)$! Il s'agit d'une probabilité conditionnelle.

En utilisant la formule (11.5), on peut écrire

$$P(D \mid T) = \frac{P(T \mid D).P(D)}{P(T)}. \tag{11.8}$$

Or vu l'axiome d'additivité (10.2) et en appliquant deux fois la formule (11.2), on a successivement

$$\begin{aligned} P(T) &= P(T \cap D) + P(T \cap \bar{D}) \\ &= P(T \mid D).P(D) + P(T \mid \bar{D}).P(\bar{D}). \end{aligned}$$

En remplaçant dans (11.8) et en notant que $P(\bar{D}) = 1 - P(D)$, on obtient

$$P(D \mid T) = \frac{P(T \mid D).P(D)}{P(T \mid D).P(D) + P(T \mid \bar{D})[1 - P(D)]}. \tag{11.9}$$

Il s'agit du théorème de Bayes.

11.4.2 Valeur prédictive positive

On peut donner une représentation plus moderne et plus familière de ce théorème. A cet effet, notons que

- $P(T \mid D)$ est la probabilité (ou proportion) de trouver un test "positif" chez les sujets atteints de la maladie D. C'est ce qu'on appelle "la sensibilité" (sensitivity) du test T. On la note Se.

- $P(\bar{T} \mid \bar{D})$ est la probabilité (ou proportion) de trouver un test "négatif" chez les sujets non atteints de la maladie D. C'est ce qu'on appelle "la spécificité" (specificity) du test T. On la note Sp. Dès lors, $P(T \mid \bar{D}) = 1 - P(\bar{T} \mid \bar{D}) = 1 - Sp$ est la "non-spécificité" du test T.
- $P(D \mid T)$ est appelée la "valeur prédictive positive VPP" (positive predictive value PPV) du test T pour la maladie D.
- $P(D)$ est la probabilité a priori ou *prévalence* de la maladie D.

En introduisant ces définitions dans l'expression (11.9), on obtient la "valeur prédictive positive VPP" du test qui s'écrit

$$VPP = \frac{Se.P(D)}{Se.P(D) + (1 - Sp).[1 - P(D)]} \tag{11.10}$$

On définit de même la "valeur prédictive négative VPN" (negative predictive value NPV) qui est égale à la probabilité qu'un patient avec un test "négatif" soit exempt de la maladie, soit $P(\bar{D} \mid \bar{T})$. On montre que

$$VPN = \frac{Sp.[1 - P(D)]}{Sp.[1 - P(D)] + (1 - Se).P(D)} \tag{11.11}$$

11.4.3 Exemple

Le test de Folin-Wu (T) est utilisé comme test de dépistage du diabète (maladie D). Si la glycémie observée 1h après le repas est < 150 mg/ml, le test est "négatif" ($\bar{T}$) ; par contre, si elle est ≥ 150 mg/ml, le test est "positif" (T).

On a appliqué ce test chez 510 sujets non diabétiques et 70 patients diabétiques. Les résultats sont repris au Tableau 11.1

Tableau 11.1 Test de Folin-Wu dans le dépistage du diabète (Renson et Wilkinson, 1961)

Test Folin-Wu	Non diabétiques $\bar{D}$	Diabétiques D
Négatif ($\bar{T}$)	461	14
Positif (T)	49	56
Total	510	70

A partir de cette table, on calcule la spécificité et la sensibilité du test de Folin-Wu.

$$\begin{array}{lcl} \text{Spécificité} & = & Sp = 461/510 = 0.904 \\ \text{Sensibilité} & = & Se = 56/70 = 0.800 \end{array}$$

Supposons que la prévalence du diabète soit de 6%, $P(D) = 0.06$.

En utilisant la formule (11.10), on peut calculer la valeur prédictive positive du test, c'est-à-dire la probabilité d'être diabétique en présence d'un test de Folin-Wu positif. On a

$$\begin{aligned} VPP = P(D \mid T) &= \frac{0.8 \times 0.06}{0.8 \times 0.06 + (1 - 0.904)(1 - 0.06)} \\ &= 0.347. \end{aligned}$$

On conclut que cette probabilité est de 34.7%.

Si la probabilité a priori du diabète est de 20%, $P(D) = 0.20$ (consultation de médecine interne), on obtient

$$\begin{aligned} VPP &= \frac{0.8 \times 0.20}{0.8 \times 0.20 + (1 - 0.904)(1 - 0.20)} \\ &= 0.676. \end{aligned}$$

Ces deux exemples montrent le rôle important de la prévalence de la maladie dans l'application du théorème de Bayes. A la limite, si la prévalence d'une maladie est extrêmement faible, le test diagnostic peut être inutile en dépit d'une bonne sensibilité et d'une bonne spécificité.

L'application de la formule (11.11) au cas précédent conduit au résultat suivant :

$$\begin{aligned} VPN &= \frac{0.904 \times 0.80}{0.904 \times 0.80 + (1 - 0.8)0.20} \\ &= 0.948 \end{aligned}$$

Donc, si le test de Folin-Wu est négatif, on est presque sûr que le sujet n'a pas le diabète.

Chapitre 12

Lois Binomiale et de Poisson

12.1 Introduction

Pour faire de l'inférence statistique, c'est-à-dire passer de la pratique à la théorie, du cas particulier au cas général, il faut modéliser la réalité observée. Est-il possible de trouver une loi mathématique qui permette de prédire le nombre de garçons dans une famille de cinq enfants, le nombre quotidien d'admissions aux urgences d'un hôpital, la durée de survie d'un patient cancéreux, ou la proportion de sujets ayant un taux de cholestérol anormalement élevé ? Comme en physique, on peut décrire les phénomènes fortuits à l'aide de modèles mathématiques, dont seule la confrontation avec la réalité permet de vérifier la validité.

Dans ce chapitre, nous introduisons la notion de "variables aléatoires" associée à celle de phénomène fortuit et montrons qu'il existe quelques distributions (ou lois) statistiques théoriques qui permettent d'expliquer quantités de situations aléatoires rencontrées en pratique. En particulier, nous définissons la loi Binomiale et la loi de Poisson, comme étant les plus caractéristiques des variables discrètes.

12.2 Variable aléatoire

12.2.1 Définition

Une variable aléatoire (random variable) est une fonction à valeurs numériques, définie sur une catégorie d'épreuve Ω. Soit $\mathcal{F}$ un phénomène fortuit et Ω sa catégorie d'épreuve (ensemble des résultats possibles de $\mathcal{F}$), alors la variable aléatoire (V.A.) X est définie comme la fonction

$$\forall \omega \in \Omega : \omega \rightarrow X(\omega) = x \in \mathbb{R}$$

où $\mathbb{R}$ est l'ensemble des réels.

Remarque : Pour les variables qualitatives, les valeurs sont nominales et $X(\omega) = x$ appartient à un ensemble $\mathcal{M}$ de modalités.

Une variable aléatoire est donc une variable dont les valeurs dépendent du hasard. Le poids n'est pas une variable aléatoire mais s'il est mesuré sur un sujet tiré au hasard d'une population, il le devient aussitôt. En effet, on ne peut connaître le poids qu'à partir du moment où on connaît le sujet extrait aléatoirement de la population.

De même, la moyenne arithmétique $\bar{x}$ ou l'écart-type s (ou tout autre caractéristique) d'un échantillon simplement fortuit d'effectif n sont des variables aléatoires car elles dépendent du phénomène fortuit $\mathcal{F}$ = échantillonnage.

12.2.2 Exemple

Dans le jet de deux dés homogènes discernables ($\mathcal{F}$), la catégorie d'épreuve s'écrit $\Omega = \{\omega = (i, j),\ i, j = 1, \ldots, 6\}$. On peut définir sur celle-ci de nombreuses variables aléatoires. Par exemple, $X = i + j$ la somme des points des deux dés. Les valeurs prises par X sont les entiers de 2 à 12. L'ensemble $E = \{\omega : X(\omega) = i + j = 4\}$ est un événement. En effet, c'est l'ensemble des résultats de $\mathcal{F}$ qui conduisent à une somme des deux dés égale à 4. Les éléments de cet événement sont donc (1,2), (2,2) et (3,1). En conséquence, $N(E) = 3$ et comme $N(\Omega) = 36$, la probabilité de E vaut (Eq. 10.5), $P(E) = 3/36 = 1/12$.

12.3 Variable aléatoire discrète

Comme on l'a montré au Chapitre 1, de nombreuses variables étudiées en statistique sont discrètes, c'est-à-dire qu'elles ne prennent qu'un nombre fini (ou infiniment dénombrable) de valeurs $\{a_1, a_2, \ldots, a_k\}$. Si une telle variable est subordonnée à un phénomène fortuit $\mathcal{F}$, elle devient aléatoire. L'exemple de la somme des deux dés en constitue un bel exemple. Les V.A. discrètes sont aussi appelées variables de "comptage" (counting variables).

12.3.1 Loi de probabilité

Une V.A. discrète X est complètement déterminée par sa "loi de probabilité" (probability distribution), c'est-à-dire la probabilité avec laquelle elle prend chacune de ses valeurs. Par définition,

$$P_i = P[X = a_i] \qquad (i = 1, \ldots, k) \tag{12.1}$$

avec $\sum P_i = 1$.

Notons que $P[X = a_i] = P\{\omega : X(\omega) = a_i\}$ et $\sum P_i = 1$ résulte de l'axiome d'additivité des probabilités.

Comme pour les valeurs d'un échantillon ou d'une population, on peut définir la moyenne et l'écart-type (ou la variance) d'une variable aléatoire X. Pour une variable

discrète

$$\mu = \sum_{i=1}^{k} a_i P_i \tag{12.2}$$

$$\sigma = \sqrt{\sum_{i=1}^{k} (a_i - \mu)^2 P_i}. \tag{12.3}$$

La loi de probabilité (théorique) d'une V.A. X discrète peut se représenter graphiquement en reportant P_i en fonction de a_i $(i = 1, \ldots, k)$.

12.3.2 Exemple

Reprenons le cas du jet de deux dés homogènes discernables $(\mathcal{F})$ et considérons la variable X = somme des deux dés. Celle-ci prend $k = 11$ valeurs possibles, les entiers de 2 à 12 inclus. Les probabilités associées à chacune de ces valeurs sont reprises au Tableau 12.1.

Tableau 12.1 Loi de probabilité de la somme de deux dés discernables

Valeur	Probabilité	Probabilité cumulée
2	1/36	1/36
3	2/36	3/36
4	3/36	6/36
5	4/36	10/36
6	5/36	15/36
7	6/36	21/36
8	5/36	26/36
9	4/36	30/36
10	3/36	33/36
11	2/36	35/36
12	1/36	36/36=1

On remarque que la loi de probabilité de X est parfaitement symétrique (voir figure 12.1) autour de la valeur $X = 7$.

La moyenne et l'écart-type de X valent respectivement $\mu = 7$ et $\sigma = 2.41$. En se servant de la loi Normale, on peut dire qu'approximativement 95% des valeurs de X se situent sur l'intervalle $7 \pm 1.96 \times 2.41$, soit dans l'intervalle 2.3 et 11.7.

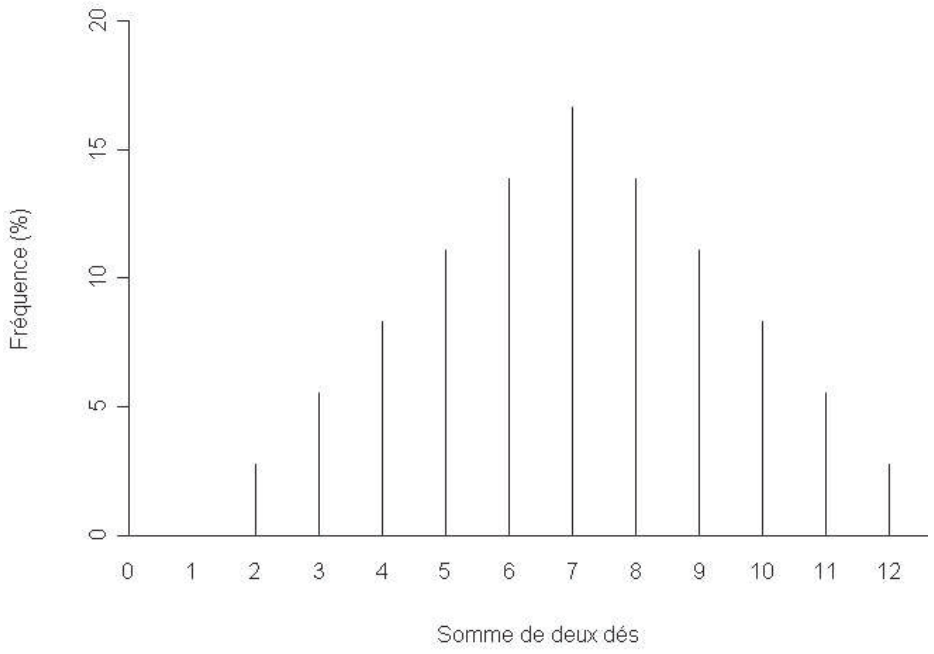

Figure 12.1 Loi de probabilité de deux dés homogènes discernables.

12.4 Loi Binomiale

La loi (distribution ou V.A.) Binomiale est la plus caractéristique des V.A. discrètes.

12.4.1 Définition

Soient n expériences aléatoires indépendantes se soldant chacune par un succès avec probabilité π ou un échec avec probabilité $(1-\pi)$. Le nombre X de succès définit une variable aléatoire Binomiale de signature (n, π), que l'on note $X \sim B(n, \pi)$.

Les valeurs prises par la V.A. Binomiale X sont les entiers de 0 à n. Donc $k = n+1$ et $a_i = i-1$ $(i = 1, \ldots, k)$. La loi de probabilité est donnée par l'équation

$$P_i = P[X = i] = C_n^i \pi^i (1-\pi)^{n-i} \quad (i = 0, \ldots, n) \tag{12.4}$$

où $C_n^i = n!/[i!(n-i)!]$ et $0! = 1$.

On montre facilement que $\mu = n\pi$ et $\sigma = \sqrt{n\pi(1-\pi)}$.

12.4.2 Exemple

Dans une famille de $n = 4$ enfants, quelle est la probabilité d'avoir 4 garçons sachant que la probabilité d'avoir un garçon $\pi = 0.51$ et celle d'avoir une fille $(1-\pi) = 0.49$?

En utilisant la formule (12.4), on a ($i = 4$)

$$\begin{aligned} P[X = 4] &= C_4^4\, 0.51^4\, 0.49^0 \\ &= 0.51^4 = 0.0677. \end{aligned}$$

Il y a donc 6,8% de chances que l'événement se produise ! Donc, en pratique sur 1000 familles de 4 enfants, on devrait trouver à peu près $1000 \times 0.0677 \simeq 68$ familles de 4 garçons ! Si cette valeur théorique diffère fortement de la valeur observée, on pourrait mettre en doute les hypothèses du modèle Binomial pour les naissances successives.

Notons que $\mu = 4 \times 0.51 = 2.04$ et $\sigma = \sqrt{4 \times 0.51 \times 0.49} = 0.9998$.

Attention : En pratique, il ne faut pas confondre n, l'effectif d'un échantillon, et n le nombre qui intervient dans la loi Binomiale. Dans l'exemple qui précède, l'effectif de l'échantillon est de $n = 1000$ familles et $n = 4$ est le nombre de naissances successives (dans des familles de 4 enfants). Il y a aussi N l'effectif de la population des familles de 4 enfants et $i =$ le nombre de garçons dans la famille de 4 enfants. Voici quatre nombres qu'il ne faut pas confondre et qui nécessiteraient des notations différentes. Nous souhaitons toutefois rester conforme à la littérature.

12.5 Loi de Poisson

La loi de Poisson, due au célèbre mathématicien français Simeon Denis Poisson (1781-1840) en 1837, est une autre variable aléatoire discrète fort utilisée en pratique, notamment pour caractériser les événements rares.

12.5.1 Définition

Supposons qu'un événement se produise de façon aléatoire et indépendante dans le temps (ou dans l'espace). Considérons un intervalle de temps donné (ou un volume d'espace donné) pris comme référence unitaire et comptons le nombre X d'événements qui s'y produisent. La variable aléatoire X prend toute les valeurs entières de 0 à l'infini (puisqu'il s'agit d'un comptage) et est appelée V.A. de Poisson de moyenne μ. Le paramètre μ est la moyenne du nombre d'occurrences de l'événement aléatoire dans l'intervalle de temps (ou le volume d'espace). On écrit $X \sim Po(\mu)$.

La loi de probabilité d'une V.A. de Poisson de moyenne μ s'écrit :

$$P_k = P(X = k) = \frac{\mu^k . e^{-\mu}}{k!} \qquad (k = 0, 1, 2, \ldots) \tag{12.5}$$

$$\sum_{k=0}^{\infty} P_k = 1.$$

On montre aisément que μ est la moyenne de la variable X et que son écart-type vaut $\sigma = \sqrt{\mu}$ (donc $\sigma^2 = \mu$). La V.A. de Poisson se caractérise par une variance égale à la moyenne.

L'inverse de la moyenne $\lambda = \mu^{-1}$ est appelé le "taux" d'occurrence de l'événement aléatoire dans l'intervalle de temps.

Il est clair que si l'intervalle de temps unitaire est multiplié par un facteur L, le nombre moyen d'occurrence l'est aussi et la formule (12.5) s'en trouve affectée :

$$P_k(L) = P_k L^k . e^{-(L-1)\mu}. \tag{12.6}$$

De même pour un facteur de volume V : $P_k(V) = P_k . V^k_. e^{-(V-1)\mu}$.

12.5.2 Exemples

Il existe de nombreux exemples de la loi de Poisson. Le plus célèbre est le nombre annuel de cavaliers tués par un coup de sabot de cheval dans l'armée prussienne. Le nombre d'appels téléphoniques arrivant à un central téléphonique par unité de temps suit une loi de Poisson. La distribution du nombre de bactéries par unité de volume de sang obéit généralement à une loi de Poisson (lecture au microscope de lames quadrillées).

Supposons que le nombre d'arrivées aux urgences d'un hôpital soit en moyenne de 3.4 par heure ($\mu = 3.4$) et suive une loi de Poisson. Quelle est la probabilité d'avoir exactement 5 admissions aux urgences durant une heure ? En utilisant la formule (12.5), on a

$$\begin{aligned} P_5 = P[X = 5] &= \frac{(3.4)^5 e^{-3.4}}{5!} \\ &= \frac{454.35 \times 0.0334}{120} = 0.126. \end{aligned}$$

Quelle est la chance de n'avoir aucune urgence dans l'heure qui vient ?

$$\begin{aligned} P_0 = P[X = 0] &= \frac{(3.4)^0 . e^{-3.4}}{0!} \\ &= 0.0334 \end{aligned}$$

soit environ 3.3%.

Par contre, si l'on souhaite connaître la probabilité qu'il n'y ait aucune admission aux urgences durant 4 heures, on obtient, en utilisant l'équation (12.6),

$$\begin{aligned} P_0(4h) &= P_0 \times 4^0 \times e^{-3\times 3.4} \\ &= 0.0334 \times 1 \times 0.0000372 \\ &= 1.2415 \times 10^{-6} \end{aligned}$$

soit une chance sur un million !

Sur 1000 heures d'observations aux urgences, le nombre attendu d'heures où il n'y a aucune admission aux urgences devrait être théoriquement de $1000 \times 0.0334 = 33.4$. En comparant cette valeur à celle obtenue en pratique, on peut confronter le modèle de Poisson à la réalité.

Chapitre 13

Loi Normale et ses dérivées

13.1 Introduction

La loi Normale, dite aussi *loi Gaussienne*, du nom du célèbre mathématicien et physicien Carl Friedrich Gauss (1777-1855), fut d'abord publiée par Abraham De Moivre (1667-1754) le 12 novembre 1733. Elle est considérée comme la loi la plus importante en statistique pour diverses raisons.

(1) Dans la nature, diverses variables se distribuent selon une loi Normale ; par exemple, la taille et les autres mesures biométriques, des constituants sanguins comme les protéines totales, l'urée, l'acide urique, le cholestérol ou les acides gras. En pédagogie, il est connu que les résultats d'un test se distribuent selon la loi de Gauss. En physique, les erreurs de mesure suivent une distribution Normale.

(2) La loi des "grands nombres" montre que la loi Normale est la limite de lois non gaussiennes. Ainsi, si on calcule la moyenne d'échantillons simplement fortuits d'effectif n, nous avons vu que lorsque n tend vers l'infini, la moyenne tend vers une loi Normale.

(3) Enfin, la loi Normale se caractérise seulement par deux paramètres, la moyenne et l'écart-type (ou la variance).

A partir de la loi Normale, on peut en dériver d'autres comme les lois chi-carré, t de Student et F de Snedecor.

13.2 Variable aléatoire continue

Une variable aléatoire continue est une variable aléatoire qui prend toutes les valeurs possibles dans un intervalle (ou continuum) donné. Il n'est plus possible de la définir par la loi de probabilité (12.1) puisque il y a un nombre non dénombrable de valeurs.

13.2.1 Fonction de répartition

Une variable aléatoire continue X est complètement déterminée par sa fonction de répartition (cumulative distribution function), $F(x)$.

Par définition,

$$F(x) = P[X \leq x] \qquad x \in \mathbb{R} \tag{13.1}$$

Il s'agit d'une fonction monotone non décroissante entre 0 et 1, telle que $F(-\infty) = 0$ et $F(+\infty) = 1$.

La fonction de répartition est l'équivalent théorique du diagramme cumulatif d'un échantillon (voir Chapitre 2). Il s'agit d'une courbe continue et non plus d'un diagramme en escaliers (voir Figure 13.1).

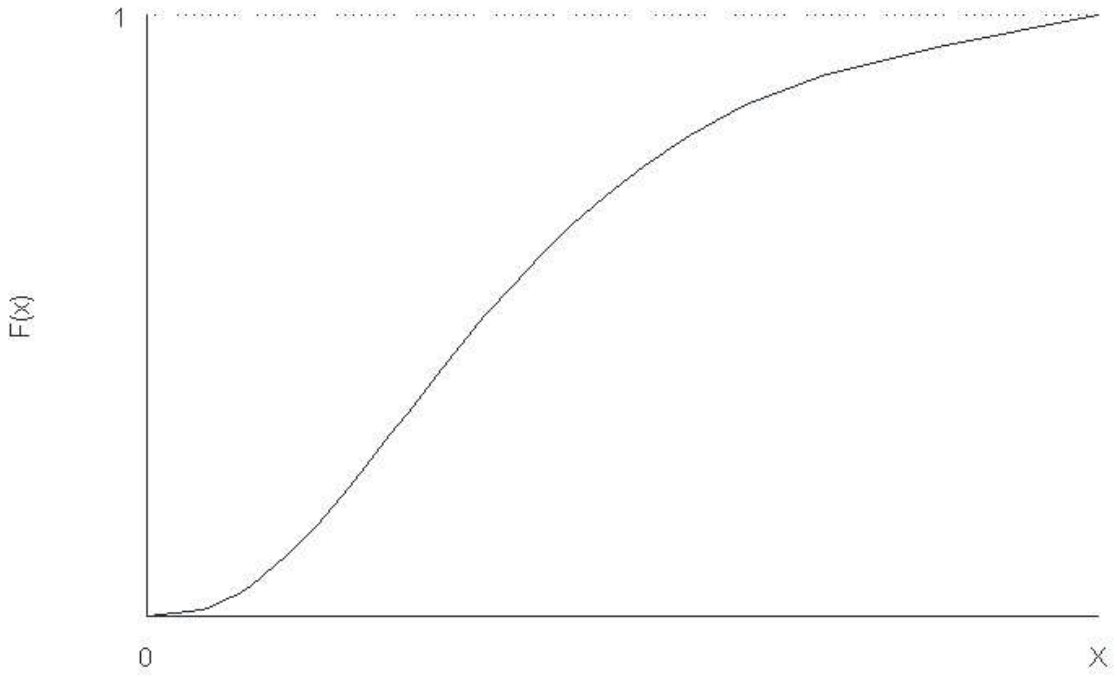

Figure 13.1 Fonction de répartition d'une variable aléatoire continue X

13.2.2 Densité de probabilité

Pour des raisons mathématiques, il est préférable de travailler avec la dérivée de la fonction de répartition $F(x)$, notée $f(x)$. On dit que $f(x)$ est une *densité de probabilité* (probability density) et on a

$$f(x) = \frac{dF(x)}{dx} \qquad x \in \mathbb{R} \tag{13.2}$$

avec

$$f(x) \geq 0$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

La densité de probabilité est l'équivalent théorique de l'histogramme d'un échantillon (voir Chapitre 2). Il s'agit d'une courbe continue et non plus d'un ensemble de rectangles juxtaposés les uns à côté des autres (voir Figure 13.2).

Connaissant la densité de probabilité (13.2), on peut recalculer la fonction de répartition et on a

$$F(x) = \int_{-\infty}^{x} f(t)dt \qquad x \in \mathbb{R} \tag{13.3}$$

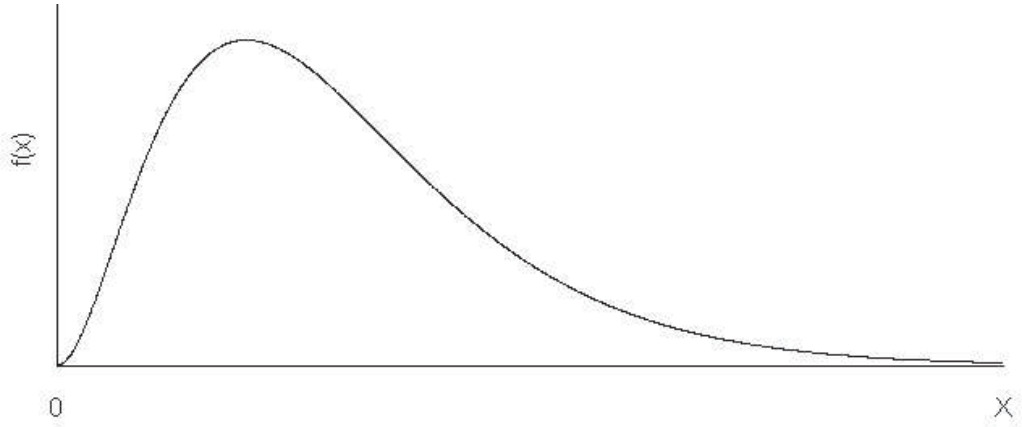

Figure 13.2 Densité de probabilité
d'une variable aléatoire continue X

Donc l'intégration d'une densité de probabilité sur un intervalle de $\mathbb{R}$ donne une probabilité. En d'autres termes, une probabilité correspond à une aire sous la densité de probabilité et à une ordonnée sur la fonction de répartition.

Notons enfin que la moyenne et la variance d'une variable aléatoire continue sont données par les expressions :

$$\mu = \int_{-\infty}^{+\infty} xf(x)dx \tag{13.4}$$

$$\sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx. \tag{13.5}$$

13.2.3 Percentiles

Connaissant la densité de probabilité $f(x)$ d'une variable aléatoire X, deux opérations sont possibles :

1. Etant donné une valeur x, calculer la probabilité de trouver une valeur inférieure ou supérieure à x. On a immédiatement par intégration

$$\begin{array}{rcl} P[X \leq x] & = & F(x) \\ P[X > x] & = & 1 - F(x) \end{array} \tag{13.6}$$

2. Etant donnée une probabilité α ($0 \leq \alpha \leq 1$), calculer le percentile ou quantile P_α correspondant, c'est-à-dire trouver P_α tel que

$$\alpha = P[X \leq P_\alpha].$$

Si α désigne l'aire supérieure, l'aire inférieure vaut $1 - \alpha$ et le quantile $P_{1-\alpha}$ est solution de

$$1 - \alpha = P[X \leq P_{1-\alpha}]. \tag{13.7}$$

ou de

$$\alpha = P[X > P_{1-\alpha}].$$

3. Dans le cas d'une variable aléatoire *symétrique*, on répartit α de moitié des deux côtés de la distribution. Dans ce cas, $P_{\alpha/2}$ est solution de

$$\frac{\alpha}{2} = P[X \leq P_{\alpha/2}] \quad \text{et} \quad P_{1-\alpha/2} = -P_{\alpha/2}. \tag{13.8}$$

13.3 Loi Normale

13.3.1 Définition

Par définition, une variable aléatoire continue X suit la loi Normale (gaussienne) de moyenne μ et d'état-type σ ($\sigma > 0$), et on écrit $X \sim N(\mu, \sigma)$, si sa densité de probabilité a pour équation

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \qquad x \in \mathbb{R}. \tag{13.9}$$

La moyenne et l'écart-type de X valent respectivement μ et σ.

13.3.2 Propriétés

La loi Normale $N(\mu, \sigma)$ possède des propriétés remarquables.
- Elle est complètement déterminée par deux paramètres, μ et σ.
- Mode = médiane = moyenne.
- Elle est symétrique et présente une forme en cloche (bell-shaped distribution).
- Elle présente un point d'inflexion en $X = \mu - \sigma$ et en $X = \mu + \sigma$.
- L'aire comprise entre $\mu \pm \sigma$ est égale à 0.6827, soit environ 2/3 des valeurs.
- L'aire comprise entre $\mu \pm 1.96\sigma$ est égale à 0.95 (95%).
- L'aire comprise entre $\mu \pm 2.58\sigma$ est égale à 0.99 (99%).

13.3.3 Calcul des aires

En effectuant la transformation "centrée réduite"

$$Z = \frac{X - \mu}{\sigma} \tag{13.10}$$

on obtient la loi Normale de moyenne 0 et d'écart-type 1, notée $Z \sim N(0, 1)$. C'est la distribution Normale standardisée, à partir de laquelle on peut obtenir toutes les autres lois Normales. Il suffit de faire la transformation inverse $X = \mu + \sigma Z$. La densité de probabilité de Z s'écrit

$$f(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2} \qquad z \in \mathbb{R} \tag{13.11}$$

et est représentée à la Figure 13.3.

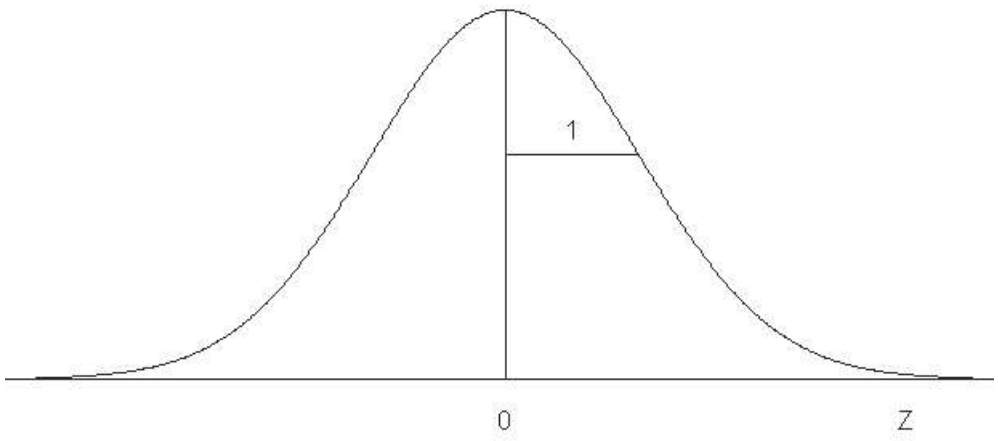

Figure 13.3 Densité de probabilité de la loi Normale $Z \sim N(0, 1)$.

Il existe une table de statistique qui donne la valeur de la densité gaussienne (13.11) pour chaque z mais elle n'est pas fort utile. Par contre, la table des aires sous la courbe gaussienne $N(0, 1)$ est importante car elle donne pour $\forall z \geq 0$, $F(z)$ ou $1 - F(z)$ selon les cas. On profite du caractère symétrique de la distribution. Ces valeurs ont été calculées par ordinateur car il s'agit d'intégrales. Elles sont reprises dans la Table de la loi Normale (voir Table A).

A titre d'exemple, supposons que la distribution de la taille (X) dans une population soit Normale de moyenne $\mu = 165$ cm et d'écart-type $\sigma = 15$ cm. Quelle est la probabilité d'avoir une taille supérieure à 180 cm ? On a successivement en utilisant (13.10)

$$\begin{aligned} P[X > 180] &= P[X - \mu > 180 - 165] \\ &= P\left[\frac{X - \mu}{\sigma} > \frac{180 - 165}{15}\right] \\ &= P[Z > 1] \\ &= 0.159 \ \text{(soit 15.9\%)}. \end{aligned}$$

Quelle est la probabilité de trouver un sujet avec une taille inférieure ou égale à 160 cm ? On a successivement,

$$\begin{aligned} P[X \leq 160] &= P\left[\frac{X-\mu}{\sigma} \leq \frac{160-165}{15}\right] \\ &= P[Z \leq -0.33] \\ &= P[Z > 0.33] \\ &= 0.371 \text{ (soit 37.1\%)}. \end{aligned}$$

13.3.4 Quantiles de la Loi Normale

Le quantile (ou percentile) α ($0 < \alpha < 1$) de la loi Normale est noté $Q_Z(\alpha)$. En vertu du caractère symétrique de la loi Normale, $Q_Z(\alpha) \leq 0$ lorsque $\alpha \leq \frac{1}{2}$ et $Q_Z(\alpha) > 0$ lorsque $\alpha > \frac{1}{2}$. De plus,

$$Q_Z(1-\alpha) = -Q_Z(\alpha). \tag{13.12}$$

Les quantiles les plus importants sont donnés dans le Tableau 13.1. Il s'agit des quantiles supérieurs de la loi Normale.

Tableau 13.1 Quantiles supérieurs de la loi Normale $N(0,1)$

α	$1-\alpha$	$Q_Z(1-\alpha)$
0.05	0.95	1.65
0.025	0.975	1.96
0.01	0.99	2.33
0.005	0.995	2.58
0.001	0.999	3.09
0.0005	0.9995	3.29

13.4 Loi du Chi-carré

La loi du chi-carré (chi-squared distribution) fait partie d'une famille de lois indexées par un paramètre entier ν appelé "degrés de liberté" (dl) (degrees of freedom, df). On la note

$$\chi^2_\nu \qquad \nu = 1, 2, 3 \ldots \tag{13.13}$$

Une variable aléatoire chi-carré n'est jamais négative ($\chi^2_\nu \geq 0$) et on montre que $\chi^2_\nu = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$, la somme de ν V.A. gaussiennes $N(0,1)$ au carré. En particulier, si $\nu = 1$, $\chi^2_1 = Z^2$, une gaussienne au carré.

Les densités de probabilité des V.A. χ^2 présentent généralement une dissymétrie à droite. On montre que la moyenne et l'écart-type d'une V.A. χ^2 à ν degrés de liberté valent $\mu = \nu$ et $\sigma = \sqrt{2\nu}$, respectivement. Il convient de signaler qu'en statistique, les lois Normale et chi-carré sont utilisées dans le cas de grands échantillons (n élevé) !

On n'a pas tabulé les probabilités associées aux V.A. chi-carrés comme on l'a fait pour la loi Normale. En effet, il faudrait une table pour chaque degré de liberté ! Par contre, en pratique, on utilise surtout les quantiles de la loi chi-carré. En particulier, les quantiles supérieurs, notés $Q_{\chi^2}(1-\alpha;\nu)$, où $\alpha = 0.05$ et 0.01 sont les seuils les plus fréquemment utilisés (voir Table B). On s'intéresse peu aux percentiles inférieurs !

A titre d'exemple, $Q_{\chi^2}(0.95;1) = 3.84 = (1.96)^2$, $Q_{\chi^2}(0.99;17) = 33.4$, $Q_{\chi^2}(0.95;8) = 15.5$.

Il est donc essentiel pour un étudiant ou un chercheur de savoir lire correctement les quantiles d'une loi χ^2 dans la Table B.

Remarque : Pour calculer la probabilité $P[\chi^2_\nu \leq a]$ ou $P[\chi^2_\nu > b]$, il faut avoir recours à une calculatrice ou à un ordinateur.

13.5 Loi t de Student

Comme la loi chi-carré, la loi t de Student fait partie d'une famille de lois indexées par un paramètre entier ν appelé "degrés de liberté". On la note

$$t_\nu \qquad \nu = 1, 2, \ldots \tag{13.14}$$

La loi t de Student fut proposée par William Sealy Gosset (1876-1937) en 1908. Comme la variable aléatoire gaussienne $Z \sim N(0,1)$, les V.A. t de Student sont symétriques autour de l'origine 0 et sont définies comme le rapport entre une gaussienne et la racine carrée d'un chi-carré divisé par ses degrés de liberté. En clair,

$$t_\nu = \frac{Z}{\sqrt{\chi^2_\nu/\nu}}. \tag{13.15}$$

On montre aussi que $\lim_{\nu\to\infty} t_\nu = Z$. Donc, pour des degrés de liberté élevés, la distribution du t de Student est proche de celle de la loi Normale $N(0,1)$.

Comme pour la V.A. Normale $N(0,1)$, les quantiles supérieurs les plus importants de la loi t de Student correspondent aux valeurs de $\alpha = 0.05$, 0.025, 0.01, 0.005, 0.001 et 0.0005 (voir Table C). On les note $Q_t(1-\alpha;\nu)$.

A titre d'exemples : $Q_t(0.975;6) = 2.45$, $Q_t(0.99;16) = 2.58$, $Q_t(0.95;68) = 1.67$, $Q_t(0.975;14) = 2.14$. Notons que la dernière ligne de la Table C ($\nu = \infty$) correspond aux quantiles de la loi Normale (voir aussi Tableau 13.1).

En statistique, les lois t de Student sont utilisées dans le cas de "petits échantillons" (n petit) !

13.6 Loi F de Snedecor

La loi F de Snedecor fait partie d'une famille de lois indexées par deux paramètres entiers ν_1 et ν_2, appelés "degrés de liberté". On la note

$$F_{\nu_1,\nu_2} \qquad \nu_1, \nu_2 = 1, 2, \ldots \tag{13.16}$$

ν_1 est appelé le "premier degré de liberté" et ν_2 "le second degré de liberté". Il convient de ne pas les intervertir, car

$$F_{\nu_1,\nu_2} \neq F_{\nu_2,\nu_1}.$$

Comme son nom l'indique, la loi F est due au statisticien américain William George Snedecor mais fut en réalité déjà obtenue par R.A. Fisher. Par définition, la V.A. F de Snedecor est définie comme le rapport de deux variables chi-carré divisées par leurs degrés de liberté, soit

$$F_{\nu_1,\nu_2} = \frac{(\chi^2_{\nu_1}/\nu_1)}{(\chi^2_{\nu}/\nu_2)}. \tag{13.17}$$

Il s'agit d'une V.A. non négative ($F_{\nu_1,\nu_2} \geq 0$) et dont la distribution est dissymétrique à droite. Seuls les quantiles supérieurs $Q_F(1-\alpha;\nu_1,\nu_2)$ sont utilisés avec $\alpha = 0.05$ ou 0.01 (voir Table D).

En statistique, la loi F de Snedecor est utilisée pour les petits échantillons (n petits) comme la loi t de Student. D'ailleurs, on constate que si $\nu_1 = 1$ et $\nu_2 = \nu$,

$$F_{1,\nu} = t^2_{\nu} \tag{13.18}$$

A titre d'exemple, $Q_F(0.95;5,2) = 19.3$ alors que $Q_F(0.95;2,5) = 5.79$, $Q_F(0.99;10,5) \simeq 10$ et $Q_F(0.99;1,36) \approx 7.40$.

13.7 Remarque finale

Les lois Normale (Z), t de Student (t_ν), Chi-carré (χ^2_ν) et F de Snedecor (F_{ν_1,ν_2}) sont les *quatre lois fondamentales* en statistique et permettent de répondre à la quasi-totalité des tests statistiques utilisés en pratique.

Chapitre 14

Tests d'hypothèses

14.1 Introduction

La statistique n'est pas simplement un outil pour résumer des données mais c'est aussi une méthodologie pour estimer des paramètres de population et tester des hypothèses (statistique inférentielle). Le but des tests d'hypothèses (hypothesis testing) est d'aider le clinicien, le chercheur ou le gestionnaire à prendre une décision au sujet d'une population à partir d'un échantillon extrait de cette population.

Considérons le développement d'un nouveau médicament contre l'hypertension. Ce médicament est-il efficace ? Est-il meilleur que d'autres médicaments sur le marché ? Après une campagne de sensibilisation anti-tabac, peut-on affirmer que la consommation de cigarettes a réellement diminué ? La survie de patients cancéreux est-elle différente selon que le patient est traité par chimiothérapie ou par radiothérapie ? Les femmes fument-elles aujourd'hui plus que les hommes ? Y a-t-il une association entre le cancer du poumon et l'exposition à des substances toxiques ? Le taux de réussite dans les différentes facultés de l'université est-il homogène ? Dans la comparaison de la reproductibilité de deux appareils de laboratoire, peut-on affirmer que l'un est meilleur (plus précis) que l'autre ?

Les exemples sont aussi nombreux que les situations réelles auxquelles on est confronté. La statistique est donc un outil indispensable dans la pratique quotidienne, la recherche, et les prises de décision.

La statistique inférentielle permet de conclure ou de décider en situation d'incertitude. Contrairement au raisonnement mathématique (vrai/faux), le raisonnement statistique doit tenir compte d'un continuum de possibilités et toute conclusion se prend avec un risque non nul de se tromper.

La structure générale de résolution des tests (on dit aussi "épreuves") d'hypothèses peut être présentée selon une procédure en six étapes :

1. Définition des hypothèses
2. Collecte des données
3. Fixation du niveau d'incertitude

4. Calcul du test statistique
5. Détermination du seuil de décision (p-value)
6. Conclusion

Chaque étape est décrite en détail dans les sections suivantes.

14.2 Les hypothèses

En statistique inférentielle, une hypothèse peut être considérée comme "une proposition au sujet d'une ou plusieurs populations". En général, cependant, les hypothèses portent sur les paramètres θ des populations concernées. Les paramètres de population sont en général la moyenne μ ou la proportion π, l'écart-type σ ou la variance σ^2, le coefficient de corrélation ρ, la pente et l'ordonnée d'une droite de régression.

On distingue deux types d'hypothèses, l'hypothèse nulle et l'hypothèse alternative.

14.2.1 Hypothèse nulle

L'hypothèse que l'on teste est appelée "hypothèse nulle" (null hypothesis) et est notée H_0. On s'arrange en général pour faire en sorte qu'on rejette l'hypothèse nulle. Ainsi, si on suspecte que la proportion de fumeurs chez les hommes est différente de celle chez les femmes, on postule qu'elles sont les mêmes ! Si on veut démontrer qu'un médicament est efficace (par exemple, qu'il diminue la pression artérielle systolique), on postule qu'il est inefficace (pas d'effet sur la pression artérielle). Si on souhaite mettre en évidence la supériorité du traitement A sur le traitement B dans le cancer de la prostate, par exemple, on suppose que les deux traitements sont équivalents. En conséquence, l'inverse de ce que le chercheur ou l'expérimentateur veut montrer devient l'hypothèse nulle. Dans les épreuves d'hypothèses, l'hypothèse nulle H_0 est rejetée ou n'est pas rejetée. Il est abusif de dire qu'elle est "acceptée" même si on utilise ce terme pour une meilleure compréhension. Lorsqu'une hypothèse nulle n'est pas rejetée, on dit que les données n'ont pas permis d'infirmer H_0 ou ne supportent pas H_0.

A titre d'exemples, voici quelques hypothèses nulles H_0 :

- $\rho = 0$ (corrélation nulle entre deux variables)
- $\pi_1 = \pi_2$ (comparaison de deux proportions)
- $\mu_1 = \mu_2 = \cdots = \mu_k$ (comparaison de k moyennes)
- $\kappa = 0$ (absence de concordance entre deux évaluateurs)
- $\beta = 1$ (pente d'une droite de régression égale à 1)
- $\alpha = 0$ (droite de régression passant par l'origine)
- $P(X = x,\ Y = y) = P(X = x).P(Y = y)\ \forall x, y,\ x \in \mathcal{M}_X$ et $y \in \mathcal{M}_Y$ (indépendance entre deux variables qualitatives).

14.2.2 Hypothèse alternative

Par opposition à l'hypothèse nulle, il y a l'hypothèse alternative, celle qui est vraie lorsque H_0 est fausse, celle que l'on accepte lorsque H_0 est rejetée. On la note H_a ou H_1. L'hypothèse alternative est plus générale que l'hypothèse nulle. Elle est aussi moins précise.

Pour chacune des hypothèses nulles de la section 14.2.1, voici les hypothèses alternatives correspondantes H_1 :

- $\rho \neq 0$
- $\pi_1 \neq \pi_2$
- $\exists i, j \in \{1, \ldots, k\}$, $i \neq j : \mu_i \neq \mu_j$
- $\kappa \neq 0$
- $\beta \neq 1$
- $\alpha \neq 0$
- $\exists x, y : P(X = x,\ Y = y) \neq P(X = x).P(Y = y)$.

Dans chaque situation, on a posé une hypothèse alternative "*bilatérale*" (two-sided hypothesis) dans la mesure où, si H_0 est fausse, le sens de l'hypothèse alternative n'est pas précisé (plus grand ou plus petit, négatif ou positif).

Lorsqu'on indique le sens de l'hypothèse alternative H_1, on dit qu'on a affaire à une hypothèse "*unilatérale*" (one-sided hypothesis). Celle-ci doit être définie avant la collecte des données et non pas après. Le fait d'indiquer "dans la procédure" le sens de H_1 constitue une information supplémentaire dont on pourra tirer profit dans la suite. Voici quelques hypothèses alternatives unilatérales : $\rho > 0$, $\pi_1 < \pi_2$, $\kappa > 0$, $\beta < 1$.

En conclusion, si θ désigne un paramètre de population et que l'on souhaite comparer θ dans deux populations, les hypothèses nulles et alternatives s'écrivent :

$$\begin{array}{lll} H_0 : \theta_1 = \theta_2 & \text{versus} & H_1 : \theta_1 \neq \theta_2 \quad \text{(test bilatéral)} \\ H_0 : \theta_1 = \theta_2 & \text{versus} & H_1 : \theta_1 > \theta_2 \quad \text{(test unilatéral)} \\ H_0 : \theta_1 = \theta_2 & \text{versus} & H_1 : \theta_1 < \theta_2 \quad \text{(test unilatéral)} \end{array}$$

14.3 Les données

Pour tester une hypothèse, il faut des données. Ces données peuvent être obtenues à l'issue d'un sondage, d'une enquête, d'une étude clinique, d'un essai médicamenteux, d'une expérience de laboratoire. Elles proviennent de la mesure ou de l'observation d'une variable X (ou de deux variables X et Y) sur les éléments d'un échantillon simplement fortuit d'effectif n (ou de plusieurs échantillons simplement fortuits) extrait de la population.

Dans le cas simple, les données sont notées $\mathcal{D} = \{x_1, \ldots, x_n\}$.

Les données sont les *seuls* éléments d'information dont on dispose pour décider entre les hypothèses H_0 et H_1

14.4 Le niveau d'incertitude

Comme on ne dispose que des informations obtenues à partir d'un échantillon extrait de la population et non pas de l'ensemble des éléments de la population, toute décision statistique est affectée d'un risque d'erreur. En fait, il y a deux risques d'erreur selon que l'hypothèse nulle est vraie ou fausse et selon que les données conduisent à rejeter ou à ne pas rejeter H_0 (voir Table 14.1).

Table 14.1 Risques d'erreur dans les épreuves d'hypothèses

	Hypothèse nulle H_0	
Données	Vraie	Fausse
"Accepter" H_0	$1-\alpha$	β
"Rejeter" H_0	α	$1-\beta$

14.4.1 Risque de 1ère espèce

Rejeter H_0 alors que H_0 est vraie est une erreur. La probabilité de cette erreur est appelée "niveau d'incertitude" (significance level) ou "risque de première espèce" (type I error). On le note α (ne pas confondre avec l'ordonnée à l'origine d'une droite de régression). Donc

$$\alpha = P(\text{Rejeter } H_0 \mid H_0 \text{ vraie}). \tag{14.1}$$

On souhaite que α soit aussi petit que possible. Par convention, il est admis de prendre $\alpha = 0.05$ (5%). Toute autre valeur inférieure $\alpha = 0.01$ ou 0.001 est acceptable mais parfois trop sévère. Une valeur $\alpha = 0.10$ est par contre considérée comme trop élevée !

14.4.2 Risque de 2e espèce

Ne pas rejeter H_0 alors que H_0 est fausse constitue une autre erreur de décision. La probabilité de cette erreur est appelée "risque de second espèce" (type II error) et notée β (ne pas confondre avec la pente d'une droite de régression). Donc

$$\begin{aligned} \beta &= P(\text{Accepter } H_0 \mid H_0 \text{ fausse}) \\ &= P(\text{Accepter } H_0 \mid H_1 \text{ vraie}) \end{aligned} \tag{14.2}$$

On voit que β dépend de l'hypothèse alternative H_1, ce que l'on note $\beta(H_1)$. S'il n'y a qu'une seule hypothèse nulle, il y a par contre une infinité d'hypothèses alternatives.

A nouveau, il est souhaitable que β soit aussi faible que possible. Malheureusement, pour n fixé, lorsqu'on diminue α, on augmente β et vice-versa. La seule façon de diminuer simultanément α et β est d'augmenter l'effectif n de l'échantillon (ce qui n'est pas toujours possible).

La quantité $1 - \beta$ est appelée "la puissance" (power) d'un test statistique.

14.4.3 Calcul de puissance

Aujourd'hui, dans les essais cliniques (clinical trial) de grande ampleur, on effectue ce que l'on appelle un "calcul de puissance" (power computation) dans la mesure où pour une hypothèse alternative H_1 donnée, on recherche l'effectif n de l'échantillon qu'il faut extraire de la population pour que les risques de 1e et 2e espèces atteignent des valeurs fixées par l'utilisateur $\alpha = \alpha_0$ et $\beta = \beta_0$. Dès lors

$$n = n(H_1, \alpha_0, \beta_0). \tag{14.3}$$

Ce calcul peut être simple ou au contraire extrêmement complexe.

14.4.4 Analogie avec la clinique

En médecine, lorsqu'on se sert d'un test biologique ou clinique pour diagnostiquer une maladie (voir Chapitre 9), deux types d'erreurs sont aussi possibles :

- Le test est "positif", alors que le sujet n'est pas malade (H_0). Ce sont les "faux positifs FP" (analogie avec α)
- Le test est "négatif", alors que le sujet est malade (H_1). Ce sont les "faux négatifs FN" (analogie avec β).

Pour diminuer simultanément les FP et FN, il faut disposer d'un test plus performant ou faire d'autres examens cliniques (c'est-à-dire augmenter l'information disponible).

14.5 Le test statistique

Ayant défini les hypothèses H_0 et H_1, collecté les données et fixé le niveau d'incertitude α, on construit à partir des données de l'échantillon un critère, appelé "statistique" ou "test statistique", qui est une grandeur univariée, notée $T = T(x_1, \ldots, x_n)$. Donc, un test statistique est une caractéristique d'échantillon. On peut le considérer comme une variable aléatoire, puisqu'on dispose d'un processus d'échantillonnage sur la population.

Pour l'échantillon obtenu, on note $T = T_0$, la valeur observée du test statistique.

On suppose ensuite que H_0 est vraie et on étudie la distribution de la V.A. T sous H_0. On montre qu'en général on peut se ramener à l'une des quatre distributions classiques (voir Chapitre 13) : Normale (Z), t de Student (t_ν), Chi-carré (χ^2_ν) ou F de Snedecor (F_{ν_1,ν_2}). Ces quatre lois ont été largement tabulées (voir Tables A–D) et sont donc parfaitement connues. Ceci signifie que si H_0 est vraie, la distribution de T est parfaitement connue et il est possible d'y localiser la valeur observée T_0. Si T_0 se situe dans le milieu de

la distribution, on dit que T_0 est "favorable à H_0". Par contre, si T_0 se situe aux extrêmes de la distribution, on dit que T_0 est "défavorable à H_0" (analogie avec le résultat d'un examen biologique par rapport à l'intervalle de référence).

14.6 Seuil de décision

14.6.1 Seuil critique au niveau α

Puisqu'on s'est fixé un niveau d'incertitude α (risque maximum de rejeter H_0 si H_0 est vraie), on peut définir sur l'échelle des valeurs de T sous H_0 les percentiles "bilatéraux" $T_{\alpha/2}$ et $T_{1-\alpha/2}$ dans le cas des distributions gaussienne et t Student (ou le percentile unilatéral $T_{1-\alpha}$ dans le cas des distributions χ^2 et F de Snedecor) et définir ainsi une région de "non rejet de H_0" entre $T_{\alpha/2}$ et $T_{1-\alpha/2}$ et une région de "rejet de H_0" en dehors de cet intervalle, soit $T < T_{\alpha/2}$ ou $T > T_{1-\alpha/2}$. Dans le cas d'un test "unilatéral", on place la zone de rejet d'un seul côté, par exemple $T > T_{1-\alpha}$ ou encore $T < T_{\alpha}$. Ces seuils de décision sont lus dans les tables statistiques.

L'approche que nous venons d'adopter est tout à fait identique à celle décrite pour les intervalles de référence (voir Chapitre 3). Quand on considère les individus en bonne santé (H_0 vraie), on peut définir deux seuils de décision $P2.5$ et $P97.5$ qui déterminent une zone de référence où tombent 95% des sujets sains. En-deçà de $P2.5$ et au-delà de $P97.5$, on rejette le caractère normal du résultat et donc du sujet. Le risque d'erreur de première espèce n'excède pas 5%.

14.6.2 "p-value"

Plutôt que de définir des seuils de décision $T_{\alpha/2}$ et $T_{1-\alpha/2}$ (ou $T_{1-\alpha}$) comme en 14.6.1, on procède aujourd'hui de manière différente en calculant la "p-value" ou "probabilité de dépassement". A cet effet, il faut disposer d'une calculatrice scientifique ou d'un ordinateur car on ne peut calculer la "p-value" à la main.

Par définition, la "p-value", notée "p", est la probabilité de trouver un résultat T plus défavorable à l'hypothèse nulle H_0 que le résultat obtenu T_0 (en supposant que H_0 est vraie). Par définition, lorsque T suit la loi Chi-carré ou F de Snedecor,

$$p = P[T > T_0 \mid H_0] \tag{14.4}$$

Par contre, lorsque T suit la loi gaussienne Z ou t de Student,

$$\begin{aligned} p &= P[T > |T_0|] + P[T < -|T_0|] \\ &= 2P[T > |T_0|]. \end{aligned} \tag{14.5}$$

Dans les publications scientifiques, il est courant de donner la valeur exacte de p (par exemple, $p = 0.18$ ou $p = 0.0031$). On ne descend cependant pas en-dessous d'un dix millième ($p < 0.0001$).

14.7 Conclusion

Le test statistique bilatéral se conclut comme suit :

"On rejette H_0 si $|T_0| \geq T_{1-\alpha/2}$, sinon on ne rejette pas H_0". (14.6)

Sur base de la p-value, la décision est plus simple encore :

"On rejette H_0 si $p \leq \alpha$, sinon on ne rejette pas H_0". (14.7)

On constate donc que la p-value évite le recours aux tables statistiques pour la détermination du seuil critique de décision et permet si nécessaire de baisser le niveau d'incertitude.

Lorsque $p > \alpha$, on dit que le test statistique est *non significatif* (on ne rejette pas H_0), ce que l'on note $N.S.$

Remarque : Dans le cas d'un test "unilatéral", on décide comme suit :

"On rejette H_0 si $T_0 > T_{1-\alpha}$, sinon on ne rejette pas H_0" (*unilatéral à droite*)

"On rejette H_0 si $T_0 < T_\alpha$, sinon on ne rejette pas H_0" (*unilatéral à gauche*).

14.8 Remarque finale

Quelle que soit l'hypothèse nulle que l'on teste, le schéma en six points décrit ci-dessus est toujours d'application et il convient de le respecter scrupuleusement (Tableau 14.2).

Le test statistique T utilisé dépend à chaque fois du problème auquel on est confronté. Comme déjà évoqué, il s'agit d'un test gaussien Z, t de Student, χ^2 ou F de Snedecor. Parfois, il faut avoir recours à d'autres tables, mais c'est plus rare (par exemple, pour les tests non paramétriques).

Tableau 14.2 Test d'hypothèses en six étapes

Hypothèses H_0 vs H_1
↓
Données
↓
Niveau d'incertitude α
↓
Test statistique T (Z, t, χ^2 ou F)
↓
Seuil de décision p-value
↓
Conclusion (Rejet de H_0 si $p \leq \alpha$)

Chapitre 15

Tests sur les corrélations

15.1 Coefficient de corrélation

Le coefficient de corrélation r est une des notions les plus importantes en statistique car elle mesure l'association entre deux variables quantitatives X et Y. Pour rappel, à partir d'un échantillon bivarié d'effectif n, soit $\{(x_i, y_i),\ i = 1, \ldots, n\}$, on calcule les quantités $\sum x$, $\sum y$, $\sum x^2$, $\sum y^2$ et $\sum xy$. Le coefficient de corrélation r s'écrit

$$r = \frac{\sum xy - (\sum x)(\sum y)/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}}. \tag{15.1}$$

Il s'agit d'un nombre pur toujours compris entre -1 et 1. On note ρ le coefficient de corrélation théorique (dans la population).

Lorsqu'on souhaite montrer qu'il existe une corrélation ρ (positive ou négative) entre deux variables, on pose l'hypothèse nulle qu'il n'y a pas de corrélation entre ces variables ($\rho = 0$). Le rejet de cette hypothèse nulle conduira au résultat attendu. Dans le cas contraire, on ne pourra pas affirmer que les deux variables sont corrélées.

Le test d'hypothèse sur une corrélation nulle est l'un des plus utilisés en pratique. Il est donc naturel de commencer les tests d'hypothèses par celui-ci.

15.2 Test pour une corrélation nulle

On procède en six points comme décrit au Chapitre 14.

15.2.1 Hypothèses

Les hypothèses nulle et alternative (bilatérale) s'écrivent

$$H_0 : \rho = 0 \quad \text{vs} \quad H_1 : \rho \neq 0. \tag{15.2}$$

Dans le cas d'un test unilatéral, H_1 s'écrirait $\rho > 0$ ou $\rho < 0$ (dans ce cas, on connaît a priori le sens de la corrélation si elle n'est pas nulle).

15.2.2 Données

Les données se présentent sous la forme d'un échantillon simplement fortuit d'effectif n $\{(x_i, y_i),\ i = 1, \ldots, n\}$ ou d'emblée sous la forme du coefficient de corrélation obtenu r (voir Eq. 15.1).

15.2.3 Niveau d'incertitude

On fixe le niveau d'incertitude $\alpha = P[\text{rejeter } H_0 \mid H_0 \text{ vraie}]$ au seuil conventionnel de 5%, par exemple ($\alpha = 0.05$).

15.2.4 Test statistique

On calcule le test statistique

$$t_{(n-2)} = r\sqrt{\frac{n-2}{1-r^2}} \tag{15.3}$$

car on démontre que si H_0 est vraie ($\rho = 0$), il est distribué selon une loi t de Student à $n-2$ degrés de liberté. Donc, sous H_0, $T \sim t_\nu$ avec $\nu = n-2$. C'est pourquoi dans l'équation 15.3, on écrit d'emblée $t_{(n-2)}$ plutôt que T.

Notons t_0 la valeur de $t_{(n-2)}$ pour l'échantillon obtenu.

15.2.5 Seuil de décision

Les valeurs de $t_{(n-2)}$ défavorables à H_0 sont les valeurs qui s'écartent fortement de l'origine dans un sens au dans l'autre (puisque le t de Student est une distribution symétrique autour de 0).

On définit donc deux seuils au-delà desquels on rejette H_0 et endéans desquels les valeurs sont favorables à H_0. Ces seuils sont les quantiles (ou percentiles) $\alpha/2$ et $1-\alpha/2$ du t de Student à $(n-2)$ degrés de liberté et que l'on note $Q_t(\alpha/2; n-2)$ et $Q_t(1-\alpha/2; n-2)$. Notons que $Q_t(1-\alpha/2; n-2) = -Q_t(\alpha/2; n-2)$.

Par ailleurs, la p-value associée à t_0 s'écrit (voir Eq. 14.5)

$$p = 2P[t_{(n-2)} \geq |t_0|]. \tag{15.4}$$

15.2.6 Décision

On rejette H_0 ($\rho = 0$) si

$$|t_0| \geq Q_t(1-\alpha/2; n-2) \tag{15.5}$$

c'est-à-dire si $t_0 > Q_t(1-\alpha/2; n-2)$ ou $t_0 < Q_t(\alpha/2; n-2)$, sinon on ne rejette pas H_0.

De manière équivalente, on rejette H_0 si $p \leq \alpha$, sinon on ne rejette pas H_0.

15.3 Exemples

– Chez 20 patients atteints d'hépatite chronique, on a calculé le coefficient de corrélation entre deux enzymes du sang, la glutamate déshydrogénase (GlDH) et l'ornithine carbonyl transférase (OCT). Cette corrélation vaut $r = 0.73$. Au niveau d'incertitude $\alpha = 0.05$, peut-on considérer que les deux variables sont significativement corrélées ?
Il suffit donc de calculer l'expression (15.3) avec $\nu = n - 2 = 18$ degrés de liberté et on a

$$\begin{aligned} t_{(18)} &= 0.73\sqrt{\frac{20-2}{1-0.73^2}} \\ &= 4.532. \end{aligned}$$

En consultant la table du t de Student (Table C), on constate que $Q_t(0.975; 18) = 2.10$. Dès lors, en utilisant la règle de décision (15.5), puisque

$$|t_0| = 4.532 > 2.10$$

on rejette l'hypothèse nulle H_0. On peut donc conclure qu'il existe une corrélation positive significative entre les enzymes GlDH et OCT chez les sujets atteints d'hépatite chronique.

Remarque On a effectué un test bilatéral. Si au départ, on savait que la corrélation ne pouvait être que positive si elle n'était pas nulle, on aurait pu utiliser un test unilatéral. Dès lors, on aurait eu recours au seuil critique $Q_t(1 - \alpha; n - 2) = Q_t(0.95; 18) = 1.73$. A posteriori, l'hypothèse nulle eut aussi été rejetée puisque $t_0 = 4.53$ est plus grand que 1.73.

Notons que la p-value associée à la valeur observée du t de Student à 18 degrés de liberté ($t_0 = 4.532$) vaut

$$\begin{aligned} p &= 2 \times P[t_{(18)} \geq |t_0|] \\ &= 2 \times P[t_{(18)} \geq 4.532] \\ &= 2 \times 0.000129 \\ &= 0.0003. \end{aligned}$$

La probabilité de trouver un résultat plus défavorable à H_0 que celui obtenu est de l'ordre de 3 chances sur 10000, ce qui est très petit. Ceci signifie que le résultat obtenu est vraiment très défavorable à H_0 ($p << \alpha$) et qu'il conduit au rejet de cette hypothèse.

– Par ailleurs, chez 100 sujets en bonne santé, la corrélation entre les enzymes GlDH et OCT est égale à $r = 0.083$. Cette corrélation est-elle statistiquement différente de 0 ?

Le calcul du t de Student à $\nu = (n-2) = 98$ degrés de liberté conduit au résultat suivant

$$\begin{aligned} t_{(98)} &= 0.083\sqrt{\frac{100-2}{1-0.083^2}} \\ &= 0.825. \end{aligned}$$

La probabilité de dépassement associée à cette valeur s'écrit

$$\begin{aligned} p &= 2 \times P[t_{(98)} \geq 0.825] \\ &= 2 \times 0.206 = 0.412. \end{aligned}$$

On peut donc conclure que le résultat obtenu $t_0 = 0.825$ n'est pas défavorable à H_0 puisqu'il est possible d'obtenir dans 41 cas sur 100 un résultat plus défavorable. Comme $p = 0.412 > 0.05$, on ne rejette pas H_0.
En utilisant le seuil critique bilatéral à 5%, on obtient la même conclusion. En effet, $Q_t(0.975; 98) = 1.98$. Donc, puisque $|t_0| = 0.825 < 1.98$, on ne rejette pas H_0. En conclusion, il n'y a pas d'association entre GlDH et OCT chez le sujet en bonne santé.

15.4 Coefficient de corrélation de Spearman

Pour tester l'hypothèse d'une corrélation nulle sur base du coefficient de corrélation de Spearman r_S, on peut utiliser la même formule (15.3) que pour le coefficient de corrélation de corrélation classique r. On a

$$t_{(n-2)} = r_S\sqrt{\frac{n-2}{1-r_S^2}}. \tag{15.6}$$

A titre d'exemple, le coefficient de corrélation de Spearman obtenu entre les deux scores psychologiques du Spielberger State-Trait Anxiety Inventory (STAI), à savoir STAI-state et STAI-trait, chez $n = 29$ sujets présumés en bonne santé est égal à $r_S = 0.71$. L'application de la formule (15.6) donne

$$\begin{aligned} t_{(27)} &= 0.71\sqrt{\frac{29-2}{1-0.71^2}} \\ &= 5.239 \end{aligned}$$

et on constate que ce résultat est hautement significatif ($p < 0.0001$), indiquant par là qu'il existe une corrélation importante entre ces deux scores.

Chez 41 transplantés cardiaques, on a obtenu une corrélation de Spearman à peu près identique, $r_S = 0.73$ ($p < 0.0001$).

15.5 Tables de valeurs critiques

Au lieu de faire le test t de Student sur r, on peut consulter une table (voir Table E) qui donne directement les valeurs critiques bilatérales pour r en fonction de α et de n, soit $r^*(n, 1-\alpha/2)$.

Dès lors, la décision s'énonce comme suit :

"On rejette H_0 ($\rho = 0$) si $|r| \geq r^*(n, 1-\alpha/2)$, sinon on ne rejette pas H_0." (15.7)

A titre d'exemple, pour $n = 20$ et $\alpha = 0.05$, on trouve $r^*(20, 0.975) = 0.444$. Dès lors, puisque le coefficient de corrélation observé entre GlDH et OCT vaut $r = 0.73$, on rejette H_0 puisque $|r| > 0.444$.

Pour trouver $r^*(n, 1-\alpha/2)$, il suffit de résoudre l'équation (15.3) en fonction de r. On a

$$r = \frac{t_{(n-2)}}{\sqrt{(n-2) + t^2_{(n-2)}}} \tag{15.8}$$

et donc

$$r^*(n, 1-\alpha/2) = \frac{Q_t(1-\alpha/2; n-2)}{\sqrt{(n-2) + Q_t^2(1-\alpha/2; n-2)}}. \tag{15.9}$$

Par exemple, pour $n = 20$ et $\alpha = 0.05$, on obtient

$$r^*(n, 1-\alpha/2) = \frac{2.1009}{\sqrt{18 + 2.1009^2}} = 0.444.$$

15.6 Test pour une corrélation non nulle

15.6.1 Principe

A titre d'information, il arrive que l'on souhaite tester l'hypothèse que la corrélation théorique ρ est égale à une valeur donnée ρ_0. Dans ce cas, on a les hypothèses

$$H_0 : \rho = \rho_0 \quad \text{vs} \quad H_1 : \rho \neq \rho_0$$

pour un test bilatéral (ou $H_1 : \rho > \rho_0$ ou $\rho < \rho_0$ pour un test unilatéral).

Le test utilisé est dû à R.A. Fisher et n'est réellement applicable que dans le cas de grands échantillons (n élevé). A cet effet, on calcule respectivement

$$z_0 = \frac{1}{2}\, \ell n\, \frac{1+\rho_0}{1-\rho_0} \tag{15.10}$$

$$\hat{z} = \frac{1}{2}\, \ell n\, \frac{1+r}{1-r}. \tag{15.11}$$

Fisher a démontré que si H_0 est vraie ($\rho = \rho_0$), alors le critère

$$Z = (\hat{z} - z_0)\sqrt{n-3} \tag{15.12}$$

suit une loi Normale $N(0, 1)$.

En conséquence,

"On rejette H_0 si $|Z| \geq Q_Z(1-\alpha/2)$, sinon on ne rejette pas H_0."

15.6.2 Exemple

A titre d'exemple, reprenons la corrélation $r = 0.73$ entre la GlDH et l'OCT chez 20 sujets atteints d'hépatite chronique et posons les hypothèses

$$H_0 : \rho = 0.80 \quad \text{vs} \quad H_1 : \rho \neq 0.80.$$

En procédant comme ci-dessus, on a respectivement

$$\begin{aligned} z_0 &= \frac{1}{2}\ \ell n\, \frac{1+0.80}{1-0.80} = 1.099 \\ \hat{z} &= \frac{1}{2}\ \ell n\, \frac{1+0.73}{1-0.73} = 0.929 \\ Z_0 &= (0.929 - 1.099)\sqrt{17} = -0.70. \end{aligned}$$

Puisque $Q_Z(0.975) = 1.96$ et que $|Z_0| = 0.70 \leq 1.96$, on ne rejette pas H_0. On peut donc admettre que la corrélation entre GlDH et OCT vaut au moins 0.80 ($p = 0.48$).

15.7 Tests statistiques en régression

Il ne faut pas confondre corrélation et régression comme on l'a vu aux Chapitres 6 et 7. Toutefois, une analogie formelle de traitement statistique existe entre les deux problèmes.

En régression, les hypothèses portent essentiellement sur la pente (β) et l'ordonnée à l'origine (α) de la droite de régression. Il faut éviter la confusion entre ces paramètres et les risques de 1e et 2e espèces, α et β, respectivement. On désigne par a et b les valeurs estimées de α et β à partir d'un échantillon bivarié d'effectif n. On a donc le modèle

$$\mu_y(x) = \alpha + \beta x \tag{15.13}$$

15.7.1 Test sur la pente

On formule généralement l'hypothèse nulle que la pente a une valeur donnée β_0, par exemple $\beta = 1$ (droite à 45°) ou $\beta = 0$ (pente nulle). Les hypothèses s'établissent comme suit :

$$H_0 : \beta = \beta_0 \quad \text{vs} \quad H_1 : \beta \neq \beta_0.$$

On démontre que, si H_0 est vraie, le critère

$$t_{(n-2)} = \frac{b - \beta_0}{s(b)} \tag{15.14}$$

où (voir Eq. 7.6 et 8.17)

$$s(b) = \sqrt{\frac{[\sum y^2 - (\sum y)^2/n] - b^2[\sum x^2) - (\sum x)^2/n]}{(n-2)[\sum x^2 - (\sum x)^2/n]}} \tag{15.15}$$

est distribué comme un t de Student à $\nu = n - 2$ degrés de liberté. On montre en effet que le test (15.14) est équivalent au test (15.3) pour une corrélation.

Dès lors, on rejette H_0 si la valeur de t observée (soit t_0)

$$|t_0| \geq Q_t(1 - \alpha/2; n - 2)$$

sinon on ne rejette pas H_0.

Notons que si n est très grand, on peut utiliser le seuil $Q_Z(1 - \alpha/2)$ au lieu du seuil du t de Student.

A titre d'exemple, reprenons la relation entre la vitesse de réaction chimique complète et la température (voir Chapitres 7 et 8). Montrons qu'il existe une relation significative entre ces deux variables. Pour rappel, la pente estimée vaut $b = 0.02498$ et $s(b) = 0.00139$.

$$H_0 : \beta = 0 \quad \text{vs} \quad H_1 : \beta \neq 0.$$

On calcule le critère (15.14) et on a

$$t_{(8)} = \frac{0.02498 - 0}{0.00139} = 18.0$$

Comme $Q_t(0.975; 8) = 2.31$, on rejette H_0 puisque $|t_0| = 18.0 > 2.31$. Notons que la p-value vaut $p < 0.0001$.

15.7.2 Test sur l'ordonnée à l'origine

Les hypothèses s'établissent comme suit :

$$H_0 : \alpha = \alpha_0 \quad \text{vs} \quad H_1 : \alpha \neq \alpha_0$$

On démontre que le critère

$$t_{(n-2)} = \frac{a - \alpha_0}{s(a)} \tag{15.16}$$

où

$$s(a) = \sqrt{\frac{[\sum y^2 - (\sum y)^2/n] - b^2[\sum x^2 - (\sum x)^2/n]}{n - 2}} \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{[\sum x^2 - (\sum x)^2/n]}} \tag{15.17}$$

est distribuée comme un t de Student à $\nu = n - 2$ degrés de liberté.

Dès lors, on rejette H_0 si la valeur observée de (15.16), soit t_0,

$$|t_0| \geq Q_t(1 - \alpha/2; n - 2)$$

sinon on ne rejette pas H_0.

Reprenons l'exemple du Chapitre 7 et vérifions si la droite de régression passe par l'origine.

$$H_0 : \alpha = 0 \quad \text{vs} \quad H_1 : \alpha \neq 0.$$

Puisque la valeur observée de l'ordonnée à l'origine et son erreur type valent respectivement $a = -0.1198$ et $s(a) = 0.177$, le calcul du t de Student (15.16) conduit au résultat suivant :

$$t_{(8)} = \frac{-0.1198 - 0}{0.177} = -0.677$$

Comme $Q_t(0.975; 8) = 2.31$, on ne rejette pas H_0 puisque $|t_0| = 0.677 << 2.31$. On peut donc "admettre" que la droite de régression passe par l'origine (ce qui est somme toute logique). Notons enfin que $p = 2P[t_{(8)} > 0.677] = 0.52$.

15.7.3 Analyse de la variance

De nombreux logiciels actuels solutionnent le problème de régression linéaire par analyse de la variance (analysis of variance), souvent abrégée ANOVA. Dans cette analyse de variance, le critère obtenu est distribué comme un F de Snedecor à 1 et $n-2$ degrés de liberté. Il sert à tester l'hypothèse nulle : $H_0 : \beta = 0$ versus l'hypothèse alternative $H_1 : \beta \neq 0$ comme au paragraphe 15.7.1. On constatera facilement que

$$F_{(1,n-2)} = t^2_{(n-2)}. \tag{15.18}$$

Il suffit donc de prendre la racine carrée du F de Snedecor et de lui donner le signe de la pente observée b (négatif ou positif) et de procéder comme au paragraphe 15.7.1. On peut aussi calculer la p-value pour le F observé $= F_0$ et on a

$$\begin{aligned} p &= P[F_{(1,n-2)} \geq F_0] \\ &= 2P[t_{(n-2)} \geq |t_0|] \end{aligned}$$

15.8 Test pour un Kappa de Cohen

Pour le coefficient d'accord (Kappa de Cohen) entre deux évaluateurs , on teste en général l'hypothèse

$$H_0 : \kappa = \kappa_0 \quad \text{vs} \quad H_1 : \kappa \neq \kappa_0$$

où κ_0 désigne une valeur de Kappa préfixée, en général $\kappa_0 = 0$.

A cet effet, on montre que le critère

$$Z = \frac{\hat{\kappa} - \kappa_0}{s(\hat{\kappa})} \tag{15.19}$$

est asymptotiquement (n élevé) distribuée comme une loi Normale $Z \sim N(0, 1)$. Dès lors, on rejette H_0 si le Z observé, soit Z_0, est tel que

$$|Z_0| \geq Q_Z(1 - \alpha/2)$$

sinon on ne rejette pas H_0.

Rappelons que $Q_Z(1 - \alpha/2)$ est le quantile gaussien $1 - \alpha/2$ (par exemple, pour $\alpha = 0.05$, $Q_Z(0.975) = 1.96$).

A titre d'exemple, l'accord observé entre deux psychiatres dans le diagnostic de $n = 100$ patients était de $\hat{\kappa} = 0.68$ (voir Chapitre 6) et l'erreur type $s(\hat{\kappa}) = 0.087$ (Eq. 8.18). Dès lors, en posant les hypothèses

$$H_0 : \kappa = 0 \quad \text{vs} \quad H_1 : \kappa \neq 0$$

et en calculant le critère (15.19), on obtient

$$Z_0 = \frac{0.68 - 0}{0.087} = 7.82$$

résultat "hautement significatif" car largement supérieur au seuil critique 1.96. La p-value $2P[Z > 7.82]$ est de l'ordre de 5.10^{-15}, donc $(p < 0.0001)$.

Chapitre 16

Tables de contingence $r \times c$

16.1 Introduction

L'analyse statistique de données qualitatives n'est pas aisée, en raison du nombre parfois élevé de modalités. Ce chapitre est consacré à un problème fréquemment rencontré en pratique, celui de l'étude simultanée de deux variables qualitatives.

Lorsqu'on croise deux variables qualitatives, respectivement X à r modalités et Y à c modalités, on définit un tableau à r lignes et c colonnes contenant $r \times c$ cellules. Un tel tableau est appelé "table de contingence $r \times c$" (contingency table). Typiquement, les données à l'intérieur de chaque cellule sont des "comptages" (counts) et non des mesures. Ce point est essentiel.

Une table de contingence peut être établie de deux manières selon que les deux variables qualitatives X et Y sont observées simultanément (test d'indépendance) ou que l'expérimentateur contrôle (fixe) une des deux variables (test d'homogénéité). La distinction entre ces deux situations est essentielle même si la méthode de calcul est identique.

Ceci rappelle la distinction entre corrélation (Chapitre 6) et régression (Chapitre 7) opérée dans le cas de variables quantitatives.

Dans un test d'indépendance (independence test), on observe simultanément les variables qualitatives X et Y dans un échantillon de n sujets et on comptabilise le nombre de sujets tombant dans chaque cellule. La question statistique qui se pose est celle de l'association (dépendance) entre les deux variables. Cette situation ressemble à un problème de corrélation. Un exemple classique est l'association entre la couleur des cheveux et la couleur des yeux, ou encore entre le groupe sanguin et la race.

Dans un test d'homogénéité (homogeneity test), on observe la variable qualitative X à r modalités dans c populations distinctes (considérées comme une variable qualitative Y à c modalités) sur base d'un échantillon extrait séparément de chaque population. Cette situation ressemble davantage à un problème de régression.

La question statistique qu'on se pose est de savoir si la distribution de X est la même (est homogène) dans chaque population. Par exemple, la distribution des grades aux examens est-elle la même dans chaque faculté ? L'état de santé des citoyens (classé

selon l'échelle : mauvais, moyen, bon) est-il identique dans chaque province du pays ?

Il n'est pas toujours aisé de distinguer entre les deux problèmes, indépendance et homogénéité. Une manière simple est de vérifier si les deux variables ont été observées simultanément (échantillonnage d'une seule population) ou non (échantillonnage séparé de chaque population).

Notations. Dans ce qui suit, nous noterons O_{ij} le nombre de sujets (objets) tombant dans la cellule (i, j). Le premier indice est celui de la ligne et le second indice celui de la colonne. Dans certains ouvrages statistiques, on utilise la notation n_{ij}. Dans ce chapitre, la lettre O signifie "Observé". O_{ij} désigne un nombre entier, un comptage ! !

Si on calcule la somme des nombres d'une ligne $R_i = \sum_{j=1}^{c} O_{ij}$ $(i = 1, \ldots, r)$ ou d'une colonne $C_j = \sum_{i=1}^{r} O_{ij}$ $(j = 1, \ldots, c)$, on définit ce que l'on appelle les "totaux marginaux" (margins). On note $n = \sum_{ij} O_{ij}$.

16.2 Test d'indépendance

Soit une population de sujets (ou d'objets) pour laquelle on observe simultanément deux variables qualitatives X et Y, à r et c modalités, respectivement. Nous proposons de numéroter les modalités de chaque variable de sorte que $X = 1, \ldots, r$ et $Y = 1, \ldots, c$.

16.2.1 Hypothèses

Le problème étant de voir s'il existe une association entre les deux variables, on va postuler l'hypothèse qu'il y a "indépendance" (c'est-à-dire absence d'association) entre X et Y. Vu la théorie des probabilités conditionnelles et l'axiome de multiplication des probabilités (Chapitre 11), les hypothèses s'écrivent

H_0 : Indépendance entre X et Y

$$\forall i, j : P(X = i \cap Y = j) = P(X = i).P(Y = j). \tag{16.1}$$

H_1 : Dépendance entre X et Y

$$\exists i, j : P(X = i \cap Y = j) \neq P(X = i).P(Y = j). \tag{16.2}$$

16.2.2 Données

Considérons un échantillon simplement fortuit d'effectif n extrait de la population et pour lequel on a observé X et Y. Celui-ci peut être présenté sous la forme d'une table $r \times c$ (Tableau 16.1). On calcule ainsi les totaux marginaux R_i et C_j.

Tableau 16.1 Table de contingence (test d'indépendance)

	Variable Y				
Variable X	1	2	...	c	Total
1	O_{11}	O_{12}	...	O_{1c}	R_1
2	O_{21}	O_{22}	...	O_{2c}	R_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Total	C_1	C_2	...	C_c	n

Notons que dans le tableau 16.1, le total général n est fixé mais les totaux marginaux sont observés (donc ils sont aléatoires, ils dépendent de l'échantillon obtenu).

Exemple : On a observé chez $n = 1500$ sujets d'une population l'absence ou la présence d'une anomalie génétique (variable X à 2 modalités) et le groupe sanguin (variable Y à 4 modalités) ; on obtient la table de contingence 2×4 suivante (Tableau 16.2)

Tableau 16.2 Relation entre une anomalie génétique et le groupe sanguin chez 1500 sujets

	Groupe sanguin				
Anomalie génétique	A	B	AB	O	Total
Absente	543	211	90	476	1320
Présente	72	31	15	62	180
Total	615	242	105	538	1500

On se pose la question d'une association éventuelle entre l'anomalie génétique et le groupe sanguin !

16.2.3 Niveau d'incertitude

On fixe le niveau d'incertitude du test d'hypothèse, par exemple $\alpha = 0.05$. Si on rejette H_0 alors que H_0 est vraie, ceci n'arrivera que dans 5% des cas, donc rarement. On peut donc avoir confiance (95%) dans la conclusion obtenue.

16.2.4 Test statistique

Si on suppose que H_0 est vraie, on peut utiliser l'équation (16.1) pour chaque cellule du tableau 16.1. En particulier, la probabilité de tomber dans la cellule (i, j), estimée à partir des observations peut s'écrire

$$P[X = i, Y = j] = \frac{R_i}{n} \times \frac{C_j}{n}.$$

Donc le nombre attendu (expected) de sujets tombant dans la cellule (i, j) sur n individus est égal à

$$\begin{aligned} E_{ij} &= n \times P[X = i, Y = j] \\ &= n \times \frac{R_i}{n} \times \frac{C_j}{n} \\ &= \frac{R_i \times C_j}{n}. \end{aligned} \tag{16.3}$$

On peut donc calculer pour chaque cellule (i, j) le nombre de sujets attendus dans cette cellule si l'hypothèse d'indépendance est satisfaite. On constate que le calcul des E_{ij} est aisé puisqu'il s'agit de multiplier les totaux marginaux de la ligne et de la colonne passant par la cellule (i, j) et de diviser par n.

On établit ainsi un nouveau tableau des nombres (ou effectifs) attendus, on dit aussi "théoriques" (voir Tableau 16.3). On peut aussi reproduire les valeurs O_{ij} et E_{ij} dans un même tableau. La lettre E signifie "Expected".

Tableau 16.3 Nombres attendus dans un test d'indépendance

	Variable Y				
Variable X	1	2	...	c	Total
1	E_{11}	E_{12}	...	E_{1c}	R_1
2	E_{21}	E_{22}	...	E_{2c}	R_2
⋮	⋮	⋮		⋮	⋮
r	E_{r1}	E_{r2}	...	E_{rc}	R_c
Total	C_1	C_2	...	C_c	n

Les nombres attendus ne sont pas nécessairement des entiers ; néanmoins, aux erreurs d'arrondis près, les totaux marginaux restent inchangés.

Pour l'exemple, on vérifiera facilement que la table 2×4 des nombres attendus est celle donnée dans le Tableau 16.4.

Tableau 16.4 Relation entre anomalie génétique et groupe sanguin. Valeurs attendues ($n = 1500$ sujets)

	Groupe sanguin				
Anomalie génétique	A	B	AB	O	Total
Absente	541.2	212.96	92.4	473.44	1320
Présente	73.8	29.04	12.6	64.56	180
Total	615	242	105	538	1500

Principe

Il paraît naturel de dire que si les nombres observés sont proches des nombres attendus, on se trouve dans une situation *favorable* à H_0 (ce qui est le cas dans l'exemple). Par contre, si les nombres observés et théoriques diffèrent de manière substantielle, ceci ne plaide pas en faveur de H_0 (*défavorable* à H_0) et l'hypothèse d'indépendance doit probablement être rejetée.

Il faut donc calculer une "distance" entre les O_{ij} et les E_{ij} sous H_0. A cet effet, on calcule le critère

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}. \tag{16.4}$$

On démontre que si H_0 est vraie, le critère est distribué comme une loi chi-carré à $\nu = (r-1)(c-1)$ degrés de liberté. Donc, si le chi-carré obtenu est petit ($O_{ij} \simeq E_{ij}$), c'est favorable à H_0. Par contre, si la valeur chi-carré obtenue est grande ($O_{ij} \neq E_{ij}$), ceci plaide en faveur du rejet de H_0

Les quantités

$$\varepsilon_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} \qquad \forall ij \tag{16.5}$$

sont appelées les "résidus" (residuals) des cellules. Le résidu mesure l'écart standardisé entre le nombre observé de sujets et le nombre attendu de sujets dans la cellule (i, j). En comparant (16.4) et (16.5), on peut ainsi écrire

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \varepsilon^2_{ij}. \tag{16.6}$$

Comme pour les nombres observés O_{ij} et les nombres attendus E_{ij}, on peut construire la table des ε_{ij} et voir ainsi dans quelles cellules les observations s'écartent des prédictions (voir Tableau 16.5).

Table 16.5 Relation entre anomalie génétique et groupe sanguin. Résidus ($n = 1500$ sujets)

	Groupe sanguin			
Anomalie génétique	A	B	AB	O
Absente	0.077	-0.134	-0.250	0.118
Présente	-0.210	0.364	0.676	-0.319

Pour l'exemple envisagé, si l'hypothèse nulle est vraie (absence de relation entre l'anomalie génétique et le groupe sanguin), on a un chi-carré à $\nu = (r-1)(c-1) = 1 \times 3 = 3$ degrés de liberté et le calcul de (16.4) donne

$$\chi^2_{(3)} = 0.835.$$

Cette valeur que l'on peut aussi obtenir à partir de la formule (16.6) et du Tableau 16.5 semble "favorable à H_0".

16.2.5 Seuil de décision (p-value)

Comme on a choisi un niveau d'incertitude égal à α et que les valeurs de χ^2 défavorables à H_0 sont celles qui sont très élevées, on décide de prendre comme seuil critique le percentile (ou quantile) de la distribution du chi-carré qui n'est dépassé que dans $\alpha\%$ des cas, c'est-à-dire le quantile $Q_{\chi^2}(1-\alpha; (r-1)(c-1))$.

Dans l'exemple qui nous concerne, puisque $\nu = (r-1)(c-1) = 3$ et $\alpha = 0.05$, on a

$$Q_{\chi^2}(0.95; 3) = 7.82.$$

Donc les valeurs défavorables à H_0 sont celles supérieures à ce seuil critique.

La p-value associée au chi-carré observée s'écrit

$$p = P[\chi^2_{(r-1)(c-1)} \geq \chi^2_{obs}]. \tag{16.7}$$

Pour l'exemple envisagé, on a

$$\begin{aligned} p &= P[\chi^2_{(3)} \geq 0.835] \\ &= 0.841. \end{aligned}$$

16.2.6 Conclusion

Vu ce qui précède,

> "On rejette H_0 (indépendance) si $\chi^2_{obs} \geq Q_{\chi^2}(1-\alpha; (r-1)(c-1))$, sinon on ne rejette pas H_0." (16.8)

De manière équivalente,

"On rejette H_0 si $p \leq \alpha$, sinon on ne rejette pas H_0."

Dans l'exemple, puisque $p = 0.841 > 0.05$, on ne rejette pas l'hypothèse d'indépendance entre l'anomalie génétique et le groupe sanguin. Il n'y a donc pas d'association entre les deux critères qualitatifs. Ceci est confirmé par (16.8) puisque $\chi^2_{obs} = 0.835 << 7.82$, le seuil critique à 5%.

Attention : Lorsque l'hypothèse d'indépendance est rejetée, cela signifie qu'il y a une "association significative" entre X et Y. Par contre, si H_0 n'est pas rejetée, on n'a pas d'association significative entre les deux variables qualitatives.

16.3 Test d'indépendance 2×2

16.3.1 Formule du chi-carré

La table de contingence la plus simple est la table 2×2, obtenue par le croisement de deux variables binaires X et Y. Dans ce cas, la table des observations (Tableau 16.1) se simplifie et on préfère la présenter de la manière suivante (Tableau 16.6)

Tableau 16.6 Table de contingence 2×2

	Variable Y		
Variable X	0	1	Total
0	a	b	$a+b$
1	c	d	$c+d$
Total	$a+c$	$b+d$	n

On montre facilement que le critère (16.4) devient, puisque $\nu = (r-1)\,(c-1) = 1$,

$$\chi^2_{(1)} = \frac{(ad-bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}. \tag{16.9}$$

Il s'agit de la célèbre formule dite du "chi-carré", l'une des plus utilisée en statistique.

Selon nos notations, on écrirait

$$\chi^2_{(1)} = \frac{(O_{11}O_{22} - O_{12}O_{21})^2 n}{R_1 R_2 C_1 C_2}. \tag{16.10}$$

16.3.2 Exemple

On a étudié la relation entre le statut alimentaire (mauvais, bon) et les résultats universitaires (échec, réussite) chez $n = 500$ étudiants. Les résultats sont repris dans le Tableau 16.7.

Tableau 16.7 Relation entre les conditions d'alimentation et les résultats universitaires chez 500 étudiants

	Alimentation		
Résultat	Mauvaise	Bonne	Total
Echec	105	15	120
Réussite	80	300	380
Total	185	315	500

Y a-t-il une relation entre le statut alimentaire et les résultats universitaires chez les étudiants ?

On suppose qu'il y a indépendance entre les deux variables (H_0), l'hypothèse alternative stipule qu'il n'y a pas indépendance et donc qu'il y a une relation entre les deux critères (H_1).

En utilisant la formule (16.9), on obtient

$$\begin{aligned} \chi^2_{(1)} &= \frac{(105 \times 300 - 15 \times 80)^2 \times 500}{120 \times 380 \times 185 \times 315} \\ &= 172.8. \end{aligned}$$

Comme le seuil critique à 5% vaut $Q_{\chi^2}(0.95; 1) = 3.84$, on ne peut conclure qu'au rejet de H_0 puisque $\chi^2_{obs} = 172.8 >> 3.84$. On note que $p < 0.0001$. L'association entre les deux facteurs est donc hautement significative.

16.4 Test d'homogénéité

Soit c (≥ 2) populations de sujets (ou d'objets) chez lesquels on observe une variable qualitative X à r modalités (numérotée de 1 à r).

16.4.1 Hypothèses

La question que l'on se pose est de savoir si la distribution de la variable X est la même dans chaque population. On parle d'homogénéité des distributions. Dans le cas contraire, la distribution de X varie selon les populations ; on dit qu'il y a hétérogénéité. On désigne par π_{ij} la proportion de sujets ayant la modalité i dans la population j.

Puisque X est une variable nominale, pour chaque population, la somme des proportions π_{ij} vaut 1, ce que l'on écrit

$$\sum_{i=1}^{r} \pi_{ij} = 1 \qquad (j = 1, \ldots, c). \tag{16.11}$$

Les hypothèses s'écrivent

H_0 : Homogénéité des populations

$$\pi_{i1} = \pi_{i2} = \cdots = \pi_{ic} \qquad (i = 1, \ldots, r) \tag{16.12}$$

H_1 : Hétérogénéité des populations

$$\exists i \in (1, \ldots, r) \text{ et } j \neq j' \in \{1, \ldots, c\} : \pi_{ij} \neq \pi_{ij'}. \tag{16.13}$$

16.4.2 Données

Considérons un échantillon simplement fortuit d'effectif n_j extrait de chaque population ($j = 1, \ldots, c$). Au total, on a donc $n_1 + n_2 + \cdots + n_c = n$ sujets. Pour chaque sujet, on a observé la variable X. On peut donc représenter les données sous la forme d'une table de contingence $r \times c$ (voir Tableau 16.1). On note O_{ij} le nombre de sujets de l'échantillon j qui ont la modalité i de X ($i = 1, \ldots, r; j = 1, \ldots, c$).

Notons que les totaux marginaux des colonnes sont fixés, puisque en réalité $C_j = n_j$ ($j = 1, \ldots, c$). Par contre, les totaux marginaux des lignes R_i ($i = 1, \ldots, r$) sont observés ! La situation d'échantillonnage est donc fort différente de celle du problème d'indépendance.

Exemple : Dans quatre études cliniques focalisées sur la relation entre le tabagisme et le cancer du poumon, on a compté le nombre de fumeurs chez les patients cancéreux. Peut-on conclure que la proportion de patients fumeurs est la même dans chaque étude clinique ? Les données sont décrites dans le Tableau 16.8.

Tableau 16.8 Nombre de fumeurs et de non fumeurs dans quatre études cliniques impliquant des patients atteints d'un cancer du poumon

	Etude				
Tabagisme	1	2	3	4	Total
Non fumeur	3	3	7	12	25
Fumeur	83	90	129	70	372
Total	86	93	136	82	397

Les populations sont ici les $c = 4$ études scientifiques. La variable qualitative X est le tabagisme à $r = 2$ modalités (non fumeur, fumeur). On a donc une table de contingence 2×4. Les totaux des colonnes sont fixés et correspondent au nombre de patients dans chaque étude. Le nombre de fumeurs est observé dans chaque étude, de sorte que les totaux des lignes sont observés.

Notons que la proportion de fumeurs dans chaque étude vaut 96.5, 96.8, 94.9 et 85.4%, respectivement. Puisqu'il n'y a que deux lignes dans le table de contingence, on revient à comparer ces proportions entre elles !

16.4.3 Niveau d'incertitude

On fixe le niveau d'incertitude α (en général à 5%).

16.4.4 Test statistique

Si on suppose que H_0 est vraie, c'est-à-dire que la distribution de la variable X est la même dans toutes les populations, on peut regrouper les échantillons ensemble. La fréquence de chaque modalité de la variable X peut donc être estimée à partir de la relation

$$\hat{\pi}_i = \hat{P}[X = i] = \frac{R_i}{n} \qquad (i = 1, \ldots, r). \tag{16.14}$$

Dès lors si H_0 est vraie, le nombre attendu de sujets tombant dans la cellule i (c'est-à-dire la modalité i de la variable X) sur les n_j individus de l'échantillon j vaut

$$\begin{aligned} E_{ij} &= \frac{R_i}{n} \times n_j \\ &= \frac{R_i \times C_j}{n}. \end{aligned} \tag{16.15}$$

puisque $C_j = n_j$. On retrouve donc exactement la même formule que dans le test d'indépendance (voir 16.3), même si elle a été obtenue d'une manière totalement différente puisque le problème est différent.

Appliqué à l'exemple sur le tabagisme, le calcul des nombres attendus (théoriques) donne les résultats repris au Tableau 16.9.

Tableau 16.9 Tabac et cancer du poumon dans quatre études. Nombres observés et théoriques (entre parenthèses)

	Etude				
Tabagisme	1	2	3	4	Total
Non fumeur	3	3	7	12	25
	(5.42)	(5.86)	(8.56)	(5.16)	
Fumeur	83	90	129	70	372
	(80.58)	(87.14)	(127.44)	(76.84)	
Total	86	93	136	82	397

On procède ensuite comme pour le test d'indépendance en calculant la distance entre les nombres observés (O_{ij}) et les nombres théoriques (E_{ij}) et on montre que cette distance est distribuée comme un chi-carré à $\nu = (r-1)(c-1)$ degrés de liberté (Eq. 16.4)

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}.$$

Pour l'exemple, on a $\nu = (r-1)(c-1) = 3$ et en utilisant l'équation précédente

$$\chi^2_{(3)} = 12.62.$$

Cette valeur apparaît relativement élevée mais seule sa comparaison au seuil de décision critique permettra de dire si elle conduit au rejet de H_0.

16.4.5 Seuil de décision (p-value)

Le seuil de décision choisi est le quantile $1-\alpha$ de la distribution du chi-carré à $\nu = (r-1)(c-1)$ degrés de liberté, soit $Q_{\chi^2}(1-\alpha;(r-1)(c-1))$. Pour l'exemple, on a comme précédemment $Q_{\chi^2}(0.95;3) = 7.82$. De même, la probabilité de dépassement associée à la valeur observée $\chi^2_{obs} = 12.62$ vaut

$$\begin{aligned} p &= P[\chi^2_{(3)} \geq 12.62] \\ &= 0.0055. \end{aligned}$$

16.4.6 Conclusion

Vu ce qui précède,

> "On rejette H_0 (homogénéité) si $\chi^2_{obs} \geq Q_{\chi^2}(1-\alpha;(r-1)(c-1))$, sinon on ne rejette pas H_0." (16.16)

De manière équivalente,

> "On rejette H_0 si $p \leq \alpha$, sinon on ne rejette pas H_0."

Dans l'exemple, puisque $p = 0.0055 < 0.05$ on rejette H_0 ($\chi^2 = 12.62 >> 7.82$). On est donc forcé de conclure que la proportion de fumeurs dans les 4 études cliniques n'est pas la même, on dit qu'il y a hétérogénéité. Un examen attentif montre que c'est surtout la 4e étude qui contient une proportion de fumeurs moindre (85% contre 96% pour les autres études).

En refaisant le test d'homogénéité sur les 3 premières études (c'est-à-dire en excluant la 4e étude), on voit que l'hypothèse d'homogénéité n'est pas rejetée ($\chi^2_{(2)} = 0.637$; $p = 0.73, NS$).

16.5 Test d'homogénéité 2×2

Comme pour le test d'indépendance, lorsque la variable X est binaire et qu'on n'a que $c = 2$ populations, on obtient une table de contingence 2×2. On utilise les mêmes formules que celle de la section 16.3.

A titre d'exemple, une étude canadienne a étudié la survie à 5 ans chez 402 vétérans de la guerre 1940-1945 fumeurs de pipe et chez 1067 vétérans non fumeurs. Le taux de mortalité est-il le même dans les deux groupes ? Les données sont reprises au tableau 16.10.

Tableau 16.10 Comparaison de la survie à 5 ans des deux groupes de vétérans en fonction du tabagisme

	Vétérans		
Issue à 5 ans	Fumeur de pipe	Non fumeur	Total
En vie	348	950	1298
Décès	54	117	171
Total	402	1067	1469

On note que la proportion de décès chez les fumeurs est de 54/402 (13.43%) et de 117/1067 (10.97%) chez les non fumeurs.

On calcule le test du chi-carré (Eq. 16.9) et on obtient

$$\chi^2_{(1)} = \frac{(348 \times 117 - 950 \times 54)^2 \times 1469}{1298 \times 171 \times 402 \times 1067} = 1.73.$$

Cette valeur est inférieure au seuil critique du chi-carré à 1 degré de liberté, soit $Q_{\chi^2}(0.95; 1) = 3.84$.

Dès lors, on ne rejette pas l'hypothèse nulle d'homogénéité. Les données ne permettent pas de conclure à une différence significative du taux de mortalité entre les deux groupes de vétérans. Ceci est confirmé par la p-value qui vaut $p = P(\chi^2_{(1)} \geq 1.73) = 0.188$, qui est supérieure au seuil $\alpha = 0.05$.

16.6 Odds Ratio

En épidémiologie, on se pose souvent la question de savoir s'il existe une association entre un facteur de risque et une maladie. Par exemple, les travailleurs exposés à des substances toxiques sont-ils plus enclins à développer un cancer pulmonaire que ceux qui ne sont pas exposés à ces substances ? Les patients diabétiques ont-ils davantage tendance à développer une cataracte ? Les sujets dont le taux de cholestérol est élevé ont-ils un risque plus élevé de faire un accident cardio-vasculaire ? Le terme "facteur de risque" (on dit aussi "facteur d'exposition") doit donc être pris au sens large.

Ce type de problème conduit à l'élaboration d'une table 2×2 comme pour les tests d'indépendance et d'homogénéité. Cette table se présente de la manière suivante (Tableau 16.11).

Tableau 16.11 Classement d'un échantillon d'individus en fonction de la présence ou de l'absence d'une maladie et de l'exposition ou non à un facteur de risque

Facteur de risque	Maladie		
	Présente	Absente	Total
Exposé	a	b	$a+b$
Non exposé	c	d	$c+d$
Total	$a+c$	$b+d$	n

16.6.1 Etudes épidémiologiques

On distingue trois types d'études en épidémiologie :

- *Etude prospective* (prospective study). On considère une cohorte de sujets exposés et une cohorte de sujets non exposés au facteur de risque que l'on suit au cours du temps. On comptabilise pour chaque cohorte le nombre de sujets qui développent la maladie (cas d'incidence). Donc dans une étude prospective $(a+b)$ et $(c+d)$ sont fixés et on observe $(a+c)$ et $(b+d)$.
- *Etude transversale* (cross-sectional study). On considère un échantillon de n sujets que l'on classe en fonction de la présence ou de l'absence de la maladie (cas de prévalence) et selon qu'ils sont ou non exposés au facteur de risque. Donc dans une étude transversale, n est fixé mais les totaux marginaux $(a+b)$, $(c+d)$, $(a+c)$ et $(b+d)$ sont observés.
- *Etude rétrospective* (retrospective ou case-control study). On considère un groupe de sujets atteints de la maladie et un groupe de sujets non atteints de la maladie. Pour chaque groupe de sujets, on regarde s'ils ont été ou non exposés au facteur de risque. Donc dans une étude rétrospective, $(a+c)$ et $(b+d)$ sont fixés, tandis que $(a+b)$ et $(c+d)$ sont observés.

Il n'y a pas d'autres situations possibles mais pour chaque type d'étude les données se présentent sous la forme du Tableau 16.11.

16.6.2 Définition de l'odds ratio

Pour mesurer l'association entre le facteur de risque et la maladie, on calcule une quantité appelée "odds ratio (OR)" qui n'a pas de traduction française. Parfois, on parle de "rapport croisé (définition de l'OMS)" ou, pour les études prospectives uniquement, de "risque relatif (RR)".

Par définition,

$$\widehat{OR} = \frac{ad}{bc} \tag{16.17}$$

On constate aisément que $\widehat{OR} \geq 0$. Par ailleurs, l'odds ratio s'interprète de la manière suivante :

- si $\widehat{OR} >> 1$, on dit qu'il y a association positive entre le facteur de risque et la maladie (augmentation du risque de développer la maladie dans le groupe exposé)
- si $\widehat{OR} << 1$, on dit qu'il y a association négative entre le facteur de risque et la maladie (diminution du risque dans le groupe exposé – facteur de protection)
- si $\widehat{OR} \approx 1$, on dit qu'il n'y a pas d'association entre le facteur de risque et la maladie.

16.6.3 Intervalle de confiance

L'odds ratio $\widehat{OR}$ défini par l'équation 16.17 est une estimation de l'odds ratio théorique OR obtenu à partir de l'ensemble de la population. En général, OR est inconnu.

On peut calculer un intervalle de confiance à 95% pour OR en procédant comme suit (méthode de Woolf) :

- On calcule $\widehat{OR}$ et ensuite $\ell n\, \widehat{OR}$
- On estime l'erreur type de $\ell n\, \widehat{OR}$ à partir de l'expression

$$SE(\ell n\, \widehat{OR}) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{16.18}$$

- On détermine un intervalle de confiance à 95% pour $\ell n\, OR$ en utilisant l'erreur type, soit $\ell n\, \widehat{OR} \pm 1.96\ SE(\ell n\, \widehat{OR})$, et on note

$$\begin{array}{rcl} \ell n\, \widehat{OR}_1 & = & \ell n\, \widehat{OR} - 1.96\ SE(\ell n\, \widehat{OR}) \\ \ell n\, \widehat{OR}_2 & = & \ell n\, \widehat{OR} + 1.96\ SE(\ell n\, \widehat{OR}) \end{array} \tag{16.19}$$

- En prenant l'exponentielle des expressions (16.19), on obtient l'intervalle de confiance à 95% pour OR, soit

$$IC\ 95\% : \widehat{OR}_1 \leq OR \leq \widehat{OR}_2 \tag{16.20}$$

16.6.4 Exemple

Une étude rétrospective de Hiller et Kahn (1976) a porté sur la relation entre le diabète et la cataracte. Les données sont reprises au Tableau 16.12. Dans cet exemple, le diabète est le facteur de risque et la cataracte la maladie.

Tableau 16.12 Relation entre le diabète et la cataracte dans une étude rétrospective portant sur des sujets âgés de 50 à 69 ans

	Patients	
Diabète	Avec cataracte	Sans cataracte
+ (Présent)	55	84
− (Absent)	552	1927
Total	607	2011

En utilisant la formule (16.17), l'odds ratio vaut

$$\widehat{OR} = \frac{55 \times 1927}{84 \times 552} = 2.29$$

Il apparaît que le risque de développer une cataracte est 2.3 fois plus élevé chez les patients diabétiques. Toutefois, on peut se poser la question de savoir si cette association est significative, c'est-à-dire si le vrai OR est différent de 1. On teste donc l'hypothèse nulle H_0 ($OR = 1$) versus l'hypothèse alternative H_1 ($OR \neq 1$), ou, ce qui revient au même, on calcule l'intervalle de confiance à 95% pour OR.

En procédant comme au point 16.6.3, on a successivement :

- $\widehat{OR} = 2.29$ et $\ell n\, \widehat{OR} = 0.8286$
- $SE\ (\ell n\, \widehat{OR}) = \sqrt{\dfrac{1}{55} + \dfrac{1}{84} + \dfrac{1}{552} + \dfrac{1}{1927}} = 0.1800$
- $\ell n\, \widehat{OR}_1 = 0.8286 - 1.96 \times 0.1800 = 0.4758$
  $\ell n\, \widehat{OR}_2 = 0.8286 + 1.96 \times 0.1800 = 1.1814$
- Dès lors, en prenant l'exponentielle,

$$e^{0.4758} \leq OR \leq e^{1.1814}$$

ou

$$IC\ 95\% : 1.6 \leq OR \leq 3.3$$

On constate que cet intervalle de confiance ne contient pas la valeur $OR = 1$. On peut donc affirmer que le diabète est un facteur de risque significatif pour le développement de la cataracte.

16.7 Remarques finales

1. Il est bon de rappeler que les nombres qui se trouvent dans une table de contingence sont des "comptages" et non des mesures ou des proportions ! Au cas où la table de contingence contiendrait des pourcentages (%), il conviendrait de la reconstruire afin d'obtenir les comptages originaux.
2. La section 16.4 montre clairement que si on veut comparer c proportions $p_1, \ldots, p_c$, on peut utiliser le test d'homogénéité. Il faut toutefois reconstruire la table de contingence.

   Ainsi, dans la comparaison des études sur le tabac dans le cancer du poumon, on aurait pu se contenter de donner le nombre de sujets de chaque étude et la proportion de fumeurs dans chaque étude, c'est-à-dire

Etude	Nombre de patients	Proportions de fumeurs
1	86	96.5%
2	93	96.8%
3	136	94.9%
4	82	85.4%

Il aurait alors fallu reconstruire le Tableau 16.8 et arrondir à l'unité pour avoir des entiers. Ainsi, le nombre de fumeurs dans l'étude 1 vaut $86 \times 0.965 = 82.99 \simeq 83$ et donc le nombre de non fumeurs $86 - 83 = 3$. On fait de même pour chaque étude.

Chapitre 17

Comparaison d'échantillons indépendants

17.1 Introduction

Dans ce chapitre, nous envisageons la comparaison de la distribution d'une variable quantitative X dans deux ou plusieurs populations distinctes. En statistique, lorsqu'on compare des populations, cela signifie en général qu'on compare leurs moyennes.

La comparaison de deux moyennes non appariées, c'est-à-dire obtenues à partir de deux *échantillons indépendants*, fait partie, avec le test sur une corrélation et les tables de contingence 2×2, des méthodes statistiques les plus utilisées en pratique. Par exemple, on compare le taux de cholestérol chez le sujet sain et le sujet obèse, la durée d'hospitalisation des patients dans deux hôpitaux, la consommation d'alcool en milieu rural et en milieu urbain, le taux de globules rouges chez l'homme et chez la femme, la consommation de médicaments dans deux unités de soins, ou la durée de vie de patients cancéreux avec ou sans métastases hépatiques. Les exemples ne manquent pas et on pourrait en allonger la liste. Chaque fois, on compare deux populations ou deux groupes distincts.

Lorsqu'on envisage le problème plus général de comparer plusieurs ($k \geq 2$) moyennes, on inclut le cas élémentaire de deux groupes ($k = 2$) mais on se trouve en même temps confronté à de nouvelles difficultés. En effet, si les moyennes ne sont pas les mêmes, elles peuvent être toutes différentes ou certaines sont égales et d'autres différentes, ou encore une moyenne diffère de toutes les autres. Il faut donc effectuer des *comparaisons multiples* pour démasquer les groupes qui diffèrent entre eux.

Ce Chapitre se propose de comparer plusieurs échantillons indépendants (ou non appariés), c'est-à-dire de comparer leurs moyennes.

La méthode s'intitule "Analyse de la variance à un critère de classification" (one-way analysis of variance), appelée aussi ANOVA-1. En effet, le nombre k de populations (donc d'échantillons) est fixé et constitue une donnée du problème. On note ici une analogie avec le test d'homogénéité dans les tables de contingence $r \times c$ ou avec le problème de régression.

17.2 Analyse de la variance à un critère

Considérons k (≥ 2) populations (ou groupes) distinctes numérotées de 1 à k. Soit X une variable *quantitative* que nous supposons de distribution *Normale* dans chaque population. Il s'agit d'une condition contraignante sur laquelle on reviendra ultérieurement.

Désignons par μ_i et σ_i^2 la moyenne et la variance de la variable X dans la i-ème population ($i = 1, \ldots, k$). Ces paramètres de population sont en général inconnus (voir Chapitre 8).

Formulons une condition supplémentaire sur les variances (ou écarts-types) de population :

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \tag{17.1}$$

appelée "hypothèse d'*homoscédasticité*" ou d'homogénéité (égalité) des variances. Il s'agit d'une autre condition assez forte sur laquelle on reviendra à la section 17.5.

17.2.1 Hypothèses

On se propose de tester l'hypothèse d'égalité (ou d'homogénéité) des moyennes des k populations. Ceci revient à dire que la distribution de la variable X est la même dans toutes les populations puisque, pour une loi Normale, la distribution est complètement caractérisée par la moyenne et l'écart-type.

Les hypothèses s'écrivent :

H_0 : Homogénéité des moyennes

$$\mu_1 = \mu_2 = \cdots = \mu_k \tag{17.2}$$

H_1 : Hétérogénéité des moyennes

$$\exists i, j \in \{1, \ldots, k\}, \;\; i \neq j : \mu_i \neq \mu_j \tag{17.3}$$

17.2.2 Données

Considérons un échantillon simplement fortuit d'effectif n_i extrait de chacune des k populations ($i = 1, \ldots, k$). Au total, on dispose donc de $n_1 + n_2 + \cdots + n_k = n$ individus. Pour chaque sujet de chaque échantillon, on mesure la variable X. On note x_{ij} l'observation de la variable X chez le sujet j de la population i ($i = 1, \ldots, k$; $j = 1, \ldots, n_i$). Il convient d'insister sur le fait que contrairement aux tables de contingence, les observations x_{ij} sont des mesures et non des comptages ! On présente les données obtenues selon le Tableau 17.1. Les quantités $\bar{x}_i$ et s_i ($i = 1, \ldots, k$) sont les moyennes et écarts-types des différents échantillons. Il faut insister sur le fait que les colonnes du Tableau 17.1 n'ont pas nécessairement toutes la même longueur puisqu'en général les effectifs n_i ($i = 1, \ldots, k$) sont différents.

Tableau 17.1 Observations d'une variable quantitative X dans k échantillons simplement fortuits extraits de populations distinctes

Echantillon (Population)			
1	2	...	k
x_{11}	x_{21}	...	x_{k1}
x_{12}	x_{22}	...	x_{k2}
$\vdots$	$\vdots$		$\vdots$
x_{1,n_1}	x_{2,n_2}	...	x_{k,n_k}
$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_k$
s_1	s_2	...	s_k

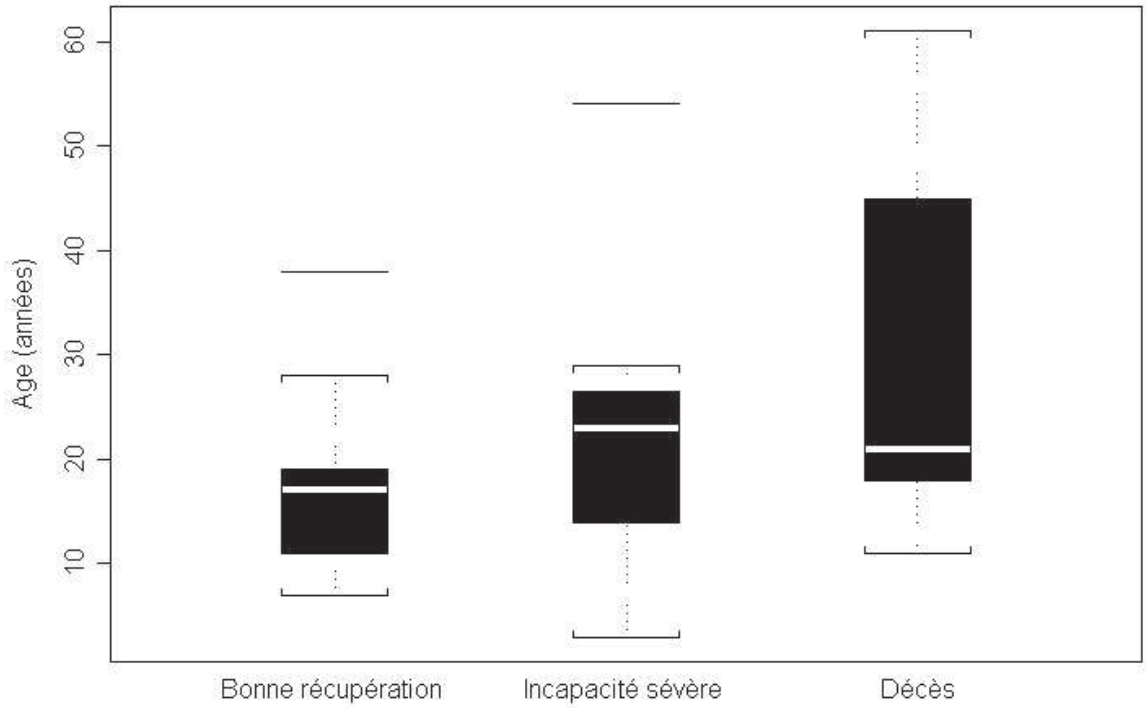

Figure 17.1 Boxplot de l'âge dans trois groupes de patients traumatisés crâniens : bonne récupération ($n = 23$, groupe 1), incapacité sévère ($n = 8$, groupe 2) et décès ($n = 22$, groupe 3)

Exemple : En neurochirurgie, il est habituel de classer les traumatisés crâniens en fonction de leur issue à 6 mois selon la "Glasgow Outcome Scale (GOS)" : bonne récupération (BR), incapacité sévère (IS), état végétatif persistant (EVP) et décès

(D). Les deux premières catégories et les deux dernières sont parfois regroupées. On a répertorié l'âge (années) chez $n_1 = 23$ patients BR, $n_2 = 8$ patients IS et 22 patients D. On se pose la question de savoir si l'âge des patients diffère entre ces trois groupes. L'âge étant une variable quantitative distribuée approximativement selon une loi Normale, on est confronté à un problème d'analyse de la variance à un critère. On pose dès lors l'hypothèse nulle que la moyenne de l'âge est la même dans chaque groupe de patients. Les données sont reprises au Tableau 17.2 et illustrées sous forme de "boxplot" à la Figure 17.1

Tableau 17.2 Observation de l'âge (années) dans trois groupes de patients traumatisés crâniens en fonction de leur issue à 6 mois (échelle GOS)

Groupe de patients		
Bonne récupération	Incapacité sévère	Décès
$n_1 = 13$	$n_2 = 8$	$n_3 = 22$
38	29	16
19	9	22
17	54	59
16	3	19
28	24	45
12	19	51
11	23	11
19	23	50
18		19
17		61
8		61
11		30
7		29
		20
		24
		16
		18
		15
		15
		20
		45
		18
$\bar{x}_1 = 17.0$	$\bar{x}_2 = 23.0$	$\bar{x}_3 = 30.2$
$s_1 = 8.42$	$s_2 = 15.2$	$s_3 = 17.0$

Même si les nombres figurant dans le Tableau 17.2 sont des entiers, il s'agit en réalité des mesures de l'âge, considéré comme une variable continue, arrondies à l'année près (voir Chapitre 1).

Les moyennes d'âge (± écart-type) dans les 3 groupes de patients valent respectivement 17 ± 8.42 ans (BR), 23 ± 15.2 ans (IS) et 30.2 ± 17.0 ans (D). On constate que les écarts-types sont relativement différents mettant peut-être en cause la condition d'homoscédasticité (17.1).

17.2.3 Niveau d'incertitude

On fixe α le niveau d'incertitude global de l'épreuve d'hypothèses, niveau maximum pour le rejet de l'hypothèse d'égalité de toutes les moyennes lorsque cette hypothèse est vraie.

17.2.4 Test statistique

A partir du tableau 17.1, on calcule pour chaque colonne (population), deux quantités : la somme des observations (T_i) et la somme des carrés des observations (S_i), soient ($i = 1, \ldots, k$)

$$T_i = \sum_{j=1}^{n_i} x_{ij} \tag{17.4}$$

$$S_i = \sum_{j=1}^{n_i} x_{ij}^2 \tag{17.5}$$

On calcule ensuite

$$T = \sum_{i=1}^{k} T_i \tag{17.6}$$

$$S = \sum_{i=1}^{k} S_i \tag{17.7}$$

Notons que $\bar{x}_i = T_i/n_i$ et $s_i = \sqrt{(S_i - T_i^2/n_i)/(n_i - 1)}$, $(i = 1, \ldots, k)$ et $\bar{x} = T/n$, la moyenne générale des n observations du tableau.

Pour les patients traumatisés, on a les résultats suivants (voir Tableau 17.3). Notons que $\bar{x} = 1068/43 = 24.9$.

Tableau 17.3 Données des patients traumatisés crâniens : effectifs (n_i), somme des obervations (T_i) et somme des carrés des observations (S_i)

Groupe de patients			
1	2	3	Total
$n_1 = 13$	$n_2 = 8$	$n_3 = 22$	$n = 43$
$T_1 = 221$	$T_2 = 184$	$T_3 = 664$	$T = 1069$
$S_1 = 4607$	$S_2 = 5842$	$S_3 = 26128$	$S = 36577$

Ces quantités sont fondamentales et vont servir pour la suite des calculs. Il est donc important de ne pas se tromper à ce stade.

Principe

Il paraît étonnant de parler d'analyse de la variance alors que l'on compare des moyennes. Toutefois, si l'on examine l'ensemble des données d'âge du tableau 17.2, on constate que ces âges sont variables (rappelons ce que disait Fisher : "*La statistique est la discipline qui étudie la variabilité*"). La variabilité entre les âges peut se décomposer en deux parties : la variabilité à l'intérieur de chaque échantillon (ou groupe), dont on a supposé en théorie qu'elle était la même, soit σ^2, et la variabilité entre les groupes, c'est-à-dire explicable par l'issue (les patients décédés apparaissent plus âgés que les patients ayant bien récupéré). Cette dernière variabilité s'exprime surtout entre les moyennes !

Si la variabilité entre les moyennes est faible par rapport à la variabilité à l'intérieur des groupes, cela signifie que les moyennes sont proches les unes des autres. Ceci plaide en faveur de l'hypothèse nulle.

Par contre, si la variabilité entre les moyennes est beaucoup plus grande que celle à l'intérieur des groupes, cela signifie que les moyennes sont fort distantes les unes des autres. Ceci plaide en défaveur de H_0 et donc en faveur de H_1.

Tel est le principe de l'analyse de la variance. Il faut à présent traduire ces propositions sous forme mathématique, c'est-à-dire exprimer la variabilité totale en variabilité intra-groupes et variabilité entre-groupes.

Sommes de carrés

L'écart de toute observation x_{ij} du tableau 17.1 à la moyenne générale $\bar{x}$ peut s'écrire

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x}). \tag{17.8}$$

Si l'on élève les deux membres de cette égalité au carré et que l'on somme sur i et sur j, on obtient

$$\sum_{i,j}(x_{ij} - \bar{x})^2 = \sum_{i,j}(x_{ij} - \bar{x}_i)^2 + \sum_{i,j}(\bar{x}_i - \bar{x})^2 \tag{17.9}$$

car le terme en double produit disparaît par addition.

On définit ainsi trois sommes de carrés, chacune associée à des "degrés de liberté".

Somme des carrés totale (total sum of squares)

$$SCT = \sum_{i,j}(x_{ij} - \bar{x})^2 = S - \frac{T^2}{n} \tag{17.10}$$

à $n - 1$ degrés de liberté.

Sommes des carrés intra-groupes (within groups sum of squares)

$$SCW = \sum_{i,j}(x_{ij} - \bar{x}_i)^2 = S - \left(\frac{T_1^2}{n_1} + \cdots + \frac{T_k^2}{n_k}\right) \tag{17.11}$$

à $n-k$ degrés de liberté.

Sommes des carrés entre-groupes (between groups sum of squares)

$$SCB = \sum_{i,j}(\bar{x}_i - \bar{x})^2 = SCT - SCW \tag{17.12}$$

à $k-1$ degrés de liberté

Carrés moyens

En divisant chaque somme de carrés par ses degrés de liberté, on définit des variances, appelées aussi "carrés moyens" (mean squares).

Le carré moyen intra-groupes est une estimation pondérée de la variabilité à l'intérieur de chaque groupe (dont on sait qu'elle est la même dans chaque groupe, par hypothèse), soit CMW ou

$$s_p^2 = \frac{SCW}{n-k}. \tag{17.13}$$

Le carré moyen inter-groupes constitue aussi une estimation de la variabilité de la variable X pour autant que H_0 soit vraie, soit

$$CMB = \frac{SCB}{k-1}. \tag{17.14}$$

Rapport des carrés moyens

Si H_0 est vraie, c'est-à-dire si les k populations sont les mêmes et donc si les échantillons obtenus constituent k échantillons indépendants d'une même population, le rapport

$$F = \frac{CMB}{s_p^2} \tag{17.15}$$

doit être voisin de 1. On démontre que le critère (17.15) sous H_0 est distribué comme un F de Snedecor à $\nu_1 = k-1$ et $\nu_2 = n-k$ degrés de liberté.

Ainsi donc, si F est très élevé, les données plaident en défaveur de H_0. Par contre, si F est faible ou voisin de 1, on se trouve dans une situation favorable à H_0.

Table d'analyse de la variance

On résume l'ensemble des calculs précédents dans une table d'analyse de la variance comme suit (voir Tableau 17.4).

Table 17.4 Table d'analyse de la variance

Source de variabilité	Somme de carrés	Degrés de liberté	Carré moyen	F-test
Entre-populations	SCB	$k-1$	CMB	$F = \frac{CMB}{s_p^2}$
Intra-populations	SCW	$n-k$	s_p^2	
Total	SCT	$n-1$		

Exemple (suite)

Pour l'exemple des trois groupes de traumatisés crâniens, on a respectivement

$$\begin{aligned} SCT &= 36577 - (1069)^2/43 = 10001.16 \\ SCW &= 36577 - \left(\frac{221^2}{13} + \frac{184^2}{8} + \frac{664^2}{22}\right) \\ &= 8547.27 \\ SCB &= SCT - SCW = 10001.16 - 8547.27 \\ &= 1453.89 \end{aligned}$$

La table d'analyse de la variance s'établit comme suit (Tableau 17.5).

Tableau 17.5 Table d'analyse de la variance pour les traumatisés crâniens

Variabilité	SC	$d\ell$	CM	F
Entre-groupe	1453.89	2	726.95	3.40
Intra-groupe	8547.27	40	213.68	
Total	10001.16	42		

17.2.5 Seuil de décision (p-value)

Sous l'hypothèse H_0, le critère F étant distribué comme un F de Snedecor à $k-1$ et $n-k$ degrés de liberté, on décide de considérer comme trop élevée toute valeur de F qui excéderait le percentile ou quantile $1-\alpha$, soit $Q_F(1-\alpha; k-1, n-k)$.

Pour l'exemple des traumatisés crâniens, on a $Q_F(0.95; 2, 40) = 3.232$ (voir Table D).

La p-value associée au F observé s'écrit

$$p = P[F_{k-1,n-k} \geq F_{obs}]. \tag{17.16}$$

Pour l'exemple, on a

$$\begin{aligned} p &= P[F_{2,40} \geq 3.40] \\ &= 0.0433 \end{aligned}$$

17.2.6 Conclusion

Vu ce qui précède,

$$\begin{array}{c}\text{“On rejette } H_0 \text{ (égalité des moyennes)} \\ \text{si } F_{obs} \geq Q_F(1-\alpha; k-1, n-k), \\ \text{sinon on ne rejette pas } H_0\text{”}\end{array} \tag{17.17}$$

De manière équivalente,

$$\begin{array}{c}\text{“On rejette } H_0 \text{ si } p \leq \alpha, \\ \text{sinon on ne rejette pas } H_0.\text{”}\end{array}$$

Dans notre exemple, puisque $p = 0.0433 < 0.05$ (ou puisque $F_{obs} = 3.40 > Q_F(0.95; 2, 40) = 3.23$), on rejette H_0. On conclut donc que l'âge moyen des patients diffère de manière significative ($p = 0.0432$) entre les trois groupes de patients traumatisés. On se situe néanmoins juste à la limite du seuil de signification $\alpha = 0.05$.

17.3 Comparaisons multiples

Lorsque le test F de Snedecor est non significatif, on ne rejette pas H_0 et le problème statistique est terminé.

Par contre, si le test F s'avère “significatif”, c'est-à-dire que l'on est amené à rejeter H_0 (comme c'est le cas dans l'exemple), le problème n'est pas terminé. En effet, on s'interroge alors sur le fait de savoir quels groupes diffèrent entre eux. On est donc amené à comparer les groupes 2 à 2, ce que l'on appelle les “comparaisons multiples” (multiple comparisons).

La méthode utilisée est celle dite des “intervalles de confiance simultanés de Scheffé” qui maintient le niveau d'incertitude global des comparaisons multiples à la valeur α fixée.

17.3.1 Différence critique

Pour chaque couple de populations, on teste l'hypothèse nulle $H_0 : \mu_i = \mu_j$ versus $H_1 : \mu_i \neq \mu_j$ ($i \neq j;\ i, j \in \{1, \ldots, k\}$).

A cet effet, on calcule la différence entre les moyennes des deux groupes et la différence critique, soient

$$d_{ij} = \bar{x}_i - \bar{x}_j \tag{17.18}$$

$$d^*_{ij}(\alpha) = \sqrt{(k-1)Q_F(1-\alpha; k-1, n-k)s_p^2\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \tag{17.19}$$

Dès lors, la décision s'énonce comme suit :

$$\begin{array}{c}\text{“On rejette } H_0 : \mu_i = \mu_j \text{ si } |d_{ij}| \geq d^*_{ij}(\alpha), \\ \text{sinon on ne rejette pas } H_0\text{”}\end{array} \tag{17.20}$$

Notons que si les effectifs des échantillons sont égaux ($n_1 = n_2 = \cdots = n_k$), le seuil critique (17.19) ne se calcule qu'une seule fois.

17.3.2 Exemple (suite)

Calculons les quantités (17.18) et (17.19) pour chaque paire de groupes de patients.

BR versus IS

$$\begin{aligned} d_{12} &= -6.0 \\ d^*_{12}(\alpha) &= \sqrt{2 \times 3.232 \times 213.68 \times \left(\frac{1}{13} + \frac{1}{8}\right)} = 16.7 \end{aligned}$$

Donc puisque $|d_{12}| = 6 < 16.7$, on ne rejette pas l'hypothèse d'égalité de l'âge entre les deux groupes ($BR = IS$).

BR versus D

$$\begin{aligned} d_{13} &= -13.2 \\ d^*_{13}(\alpha) &= \sqrt{2 \times 3.232 \times 213.68 \times \left(\frac{1}{13} + \frac{1}{22}\right)} = 13.0 \end{aligned}$$

Puisque $|d_{13}| = 13.2 > 13.0$, on rejette H_0 et donc les groupes BR et D diffèrent significativement en âge ($BR \neq D$).

IS versus D

$$\begin{aligned} d_{23} &= -7.2 \\ d^*_{23}(\alpha) &= \sqrt{2 \times 3.232 \times 213.68 \times \left(\frac{1}{8} + \frac{1}{22}\right)} = 15.3 \end{aligned}$$

Puisque $|d_{23}| = 7.2 < 15.3$, on ne rejette pas H_0 et les deux groupes ne diffèrent pas en âge ($IS = D$).

L'exemple précédent montre que la logique statistique n'est pas la logique mathématique au niveau de la transitivité.

$$BR = IS, \quad IS = D, \quad BR \neq D.$$

Par ailleurs, il peut arriver que le F soit significatif et que les groupes ne diffèrent pas entre eux (il s'agit d'un paradoxe).

On peut donc conclure des données de l'exemple que les patients décédés à 6 mois sont significativement plus âgés que les patients ayant bien récupéré. On ne peut se prononcer sur l'âge des patients avec incapacité sévère qui occupent une position intermédiaire entre les deux groupes extrêmes.

17.4 Comparaison de deux moyennes

17.4.1 Test t de Student

Lorsqu'on ne compare que deux moyennes non appariées, $\bar{x}_1$ et $\bar{x}_2$, obtenues à partir de n_1 et n_2 observations respectivement, on peut faire une analyse de la variance à un critère comme décrit à la section 17.3. Toutefois, dans ce cas ($k = 2$), les calculs se simplifient considérablement et on préfère utiliser un test t de Student. On note s_1 et s_2 les écarts-types des deux échantillons.

On teste l'hypothèse $H_0 : \mu_1 = \mu_2$ vs $H_1 : \mu_1 \neq \mu_2$. Si H_0 est vraie, on montre que le critère

$$t_{n_1+n_2-2} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{17.21}$$

où $s_p = \sqrt{[(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1+n_2-2)}$, l'écart-type pondéré des deux échantillons, est distribué comme un t de Student à $n_1 + n_2 - 2$ degrés de liberté.

Dès lors, on rejette H_0 si

$$|t_{obs}| \geq Q_t(1-\alpha/2; n_1+n_2-2) \tag{17.22}$$

sinon on ne rejette pas H_0.

De même, on peut calculer la probabilité de dépassement en calculant

$$p = 2P[t_{n_1+n_2-2} \geq |t_{obs}|] \tag{17.23}$$

et rejeter H_0 si $p \leq \alpha$.

Remarque : On montre aisément que le F de Snedecor de l'ANOVA a pour valeur le t de Student au carré.

17.4.2 Exemple

A titre d'exemple, supposons qu'on ne dispose que de l'âge dans les groupes 1 (BR) et 2 (IS) de patients traumatisés.

On note que

$$\begin{array}{rrcrr} n_1 = 13 & \bar{x}_1 = 17 & \text{et} & s_1 = 8.42 \\ n_2 = 8 & \bar{x}_2 = 23 & \text{et} & s_2 = 15.2 \end{array}$$

En conséquence, puisque

$$\begin{aligned} s_p &= \sqrt{(12 \times 8.42^2 + 7 \times 15.2^2)/19} \\ &= 11.40 \end{aligned}$$

le test t de Student vaut (Eq. 17.21)

$$\begin{aligned} t_{(19)} &= \frac{17-23}{11.40\sqrt{\frac{1}{13}+\frac{1}{8}}} \\ &= -1.17. \end{aligned}$$

Comme $Q_t(0.975; 19) = 2.09$ et que $|t_{obs}| = 1.17 < 2.09$, on ne rejette pas H_0. Il n'y a donc pas de différence significative de l'âge entre les groupes BR et IS. D'ailleurs, $p = 2P[t_{(19)} > 1.17] = 0.26(NS)$.

Si l'on effectue une ANOVA à 1 critère sur les deux groupes (voir Tableau 17.6), on trouve $F = t^2 = 1.37$ à 1 et 19 degrés de liberté et le quantile $Q_F(0.95; 1, 19) = 2.09^2 = 4.38$.

Tableau 17.6 Table d'analyse de la variance pour la comparaison des deux groupes de patients traumatisés crâniens BR et IS

Variabilité	SC	$d\ell$	CM	F
Entre-groupes	178.29	1	178.29	1.38 ($p = 0.26$)
Intra-groupes	2460.00	19	129.47	
Total	2638.29	20		

17.5 Test d'homogénéité des variances

L'analyse de la variance à un critère repose sur deux conditions : (1) que la distribution de X soit Normale dans chaque population, et (2) que les variances soient homogènes (conditions 16.1). Cette dernière hypothèse peut être éprouvée à l'aide du test de Bartlett dans le cas de $k > 2$ populations et par le test de Fisher dans le cas de deux populations.

17.5.1 Test de Bartlett

Soient les hypothèses

H_0 : Homogénéité des variances

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

H_1 : Hétérogénéité des variances

$$\exists i, j \in \{1, \ldots, k\},\ i \neq j : \sigma_i^2 \neq \sigma_j^2$$

Les données disponibles à cet effet sont les variances (écarts-types) observées dans les différents échantillons, soient $s_1^2, \ldots, s_k^2$, et les effectifs correspondants $n_1, n_2, \ldots, n_k$.

Principe

Le test de Bartlett consiste à calculer la variance pondérée s_p^2 (voir Table d'analyse de la variance 17.4) donnée par la formule (17.13), que l'on peut aussi écrire

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n - k}. \tag{17.24}$$

On calcule ensuite les expression M et C :

$$M = (n-k)\,\ell n\, s_p^2 - \sum_{i=1}^{k}(n_i - 1)\,\ell n\, s_i^2 \tag{17.25}$$

$$C = 1 + \frac{1}{3(k-1)}\left\{\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right\} \tag{17.26}$$

et le critère

$$\chi^2_{(k-1)} = \frac{M}{C} \tag{17.27}$$

distribué comme un chi-carré à $\nu = k - 1$ degrés de liberté si l'hypothèse nulle H_0 est vraie.

Dès lors, on rejette H_0 si $\chi^2_{obs} \geq Q_{\chi^2}(1 - \alpha; k - 1)$ et on ne rejette pas H_0 sinon.

Exemple

Comparons les variances des données recueillies sur les groupes de patients traumatisés (voir Tableau 17.2). On a :

$$\begin{array}{cc} n_1 = 13 & s_1^2 = 70.90 \\ n_2 = 8 & s_2^2 = 231.04 \\ n_3 = 22 & s_3^2 = 289.0 \end{array}$$

La table d'analyse de la variance (Tableau 17.5) montre que $s_p^2 = 213.68$ pour un total de $n = 43$ observations.

Dès lors, en utilisant les équations (17.25) et (17.26),

$$\begin{aligned} M &= 40 \times \ell n\, 213.68 - (12 \times \ell n\, 70.9 + 7 \times \ell n\, 231.04 + 21 \times \ell n\, 289) \\ &= 6.35 \end{aligned}$$

$$\begin{aligned} C &= 1 + \frac{1}{6}\left\{\frac{1}{12} + \frac{1}{7} + \frac{1}{21} - \frac{1}{40}\right\} \\ &= 1.042 \end{aligned}$$

Le test statistique s'écrit donc (puisque $k - 1 = 2$)

$$\chi^2_{(2)} = \frac{M}{C} = 6.09$$

conduisant à une probabilité de dépassement de $p = P[\chi^2_{(2)} \geq 6.09] = 0.048$, juste en-dessous du niveau de signification $\alpha = 0.05$. On est donc obligé de rejeter H_0 et en conséquence, l'homogénéité des variances. Une des conditions d'applicabilité de l'analyse de variance n'est donc pas réalisée. Notons que $Q_{\chi^2}(0.95; 2) = 5.99$.

17.5.2 Test de Fisher

Lorsqu'on n'a que deux variances ($k = 2$) à comparer, on peut bien sûr utiliser le test de Bartlett à $\nu = 1$ degré de liberté. Il est préférable cependant d'utiliser le test de Fisher :

$$F = \frac{s_1^2}{s_2^2} \tag{17.28}$$

où $s_1^2 \geq s_2^2$ (dans le cas contraire, on inverse), distribué sous H_0 comme un F de Snedecor à $\nu_1 = n_1 - 1$ et $\nu_2 = n_2 - 1$ degrés de liberté. On rejette donc l'hypothèse $H_0 : \sigma_1^2 = \sigma_2^2$ si $F \geq Q_p(1 - \alpha : n_1 - 1, n_2 - 1)$.

A titre d'exemple, comparons les variances des groupes BR et IS. Notons que $s_1^2 = 70.90$ et $s_2^2 = 231.04$. Comme $s_2^2 > s_1^2$, le test de Fisher s'écrit

$$F = \frac{s_2^2}{s_1^2} = \frac{231.04}{70.90} = 3.26$$

à $n_2 - 1 = 7$ et $n_1 - 1 = 12$ degrés de liberté. La p-value associée à ce résultat vaut $p = P[F_{7,12} \geq 3.26] = 0.035$. On conclut donc que les deux variances sont significativement différentes au niveau d'incertitude de 5%. Notons que $Q_F(0.95; 7, 12) = 2.91$.

17.6 Test de Kruskal-Wallis

L'applicabilité de l'analyse de la variance implique la distribution Normale de X dans chaque groupe et l'égalité des variances. Si la distribution de X n'est pas Normale, on peut s'efforcer de la normaliser en utilisant une transformation comme celles décrites à la section 4.7. En général, la normalisation de la distribution de X stabilise aussi les variances, c'est-à-dire les rend plus homogènes.

S'il n'est pas possible de répondre aux conditions de l'ANOVA, on peut avoir recours au test non-paramétrique de Kruskal-Wallis. A cet effet, on travaille sur les rangs des observations.

Principes

Le test de Kruskal-Wallis procède en plusieurs étapes :

1. On regroupe (virtuellement) l'ensemble des $n = n_1 + n_2 + \cdots + n_k$ données et on les trie par ordre croissant. On attribue ainsi un rang (rank) à chaque observation. En clair, à chaque x_{ij} on associe son rang $rang\,(x_{ij})$. En cas d'ex-aequo, on attribue la moyenne des rangs correspondants à chaque ex-aequo.

2. On calcule la somme des rangs dans chaque échantillon (groupe), soit

$$R_i = \sum_{j=1}^{n_i} rang\,(x_{ij}) \qquad (i = 1, \ldots, k) \tag{17.29}$$

3. On calcule le critère

$$\chi^2_{(k-1)} = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n+1) \tag{17.30}$$

Kruskal et Wallis ont démontré que ce critère est distribué comme un chi-carré à $\nu = k - 1$ degrés de liberté si H_0 est vraie.

4. Dès lors, on rejette H_0 si

$$\chi^2_{obs} \geq Q_{\chi^2}(1 - \alpha; k - 1) \quad \text{ou si} \quad p \leq \alpha$$

sinon on ne rejette pas H_0.

Exemple

Nous avons appliqué le test de Kruskal-Wallis aux données des patients traumatisés (Tableau 17.2). Comme il y a $n = 43$ observations au total, les rangs s'étalent de 1 à 43. Le Tableau 17.6 reproduit les observations avec leurs rangs dans chacun des 3 groupes.

La somme des rangs dans chaque échantillon vaut

$$\begin{aligned} R_1 &= 35 + 21 + \cdots + 6 + 2 = 191 \\ R_2 &= 32.5 + 4 + \cdots + 27.5 + 27.5 = 181 \\ R_3 &= 12 + 26 + \cdots + 36.5 + 17 = 574 \end{aligned}$$

On vérifiera que la somme $R = R_1 + R_2 + R_3 = 946$ est bien égale à $n(n+1)/2$ avec $n = 43$.

Le critère de Kruskal-Wallis vaut donc (puisque $k - 1 = 2$)

$$\begin{aligned} \chi^2_{(2)} &= \frac{12}{43 \times 44}\left(\frac{191^2}{13} + \frac{181^2}{8} + \frac{574^2}{22}\right) - 3 \times 44 \\ &= 138.8 - 132 = 6.8 \end{aligned}$$

Comme le seuil critique à 5% du chi-carré à 2 degrés de liberté vaut $Q_{\chi^2}(0.95; 2) = 5.99$, l'hypothèse nulle H_0 est rejetée. D'ailleurs, $p = P[\chi^2_{(2)} \geq 6.8] = 0.0334$. On conclut donc que l'âge est significativement différent dans les 3 groupes de patients.

Afin de voir quels groupes diffèrent, il faudrait comparer les groupes 2 à 2 à l'aide du même test de Kruskal-Wallis mais choisir un niveau d'incertitude plus faible pour chaque test, par exemple $\alpha' = \alpha/3$ (correction de Bonferroni). En fait, on divise α par le nombre de tests que l'on effectue ! Dès lors, pour rejeter chaque hypothèse H_0, il faut que p soit $< \alpha'$.

Tableau 17.6 Comparaison de trois groupes de patients traumatisés crâniens. Observations et leurs rangs (entre parenthèses)

Bonne récupération $n_1 = 13$	Incapacité sévère $n_2 = 8$	Décès $n_3 = 22$
38 (35)	29 (32.5)	16 (12)
19 (21)	9 (4)	22 (26)
17 (14.5)	54 (38)	59 (41)
16 (12)	3 (1)	19 (21)
28 (31)	24 (29.5)	45 (36.5)
12 (8)	19 (21)	51 (40)
11 (6)	23 (27.5)	11 (6)
19 (21)	23 (27.5)	50 (39)
18 (17)		19 (21)
17 (14.5)		61 (42.5)
8 (3)		61 (42.5)
11 (6)		30 (34)
7 (2)		29 (32.5)
		20 (24.5)
		24 (29.5)
		16 (12)
		18 (17)
		15 (9.5)
		15 (9.5)
		20 (24.5)
		45 (36.5)
		18 (17)

Attention : Quand on compare deux groupes d'effectif n_1 et n_2, il faut recalculer les rangs sur un total de $n = n_1 + n_2$ observations. On ne peut donc pas se servir des rangs de l'ensemble des groupes.

17.7 Test de Mann-Whitney

Lorsqu'on ne compare que deux groupes, on peut aussi avoir recours au test non-paramétrique U de Mann-Whitney. Ce test est célèbre et fréquemment utilisé en pratique. Il postule cependant les conditions suivantes :

1. Les deux échantillons d'effectif n_1 et n_2 sont indépendants.
2. La variable X est au moins ordinale.
3. Si les populations diffèrent, elles diffèrent quant à leurs médianes.

Principe

On procède comme suit (en supposant que $n_1 \leq n_2$) :

1. On regroupe les observations des deux échantillons et on les trie par ordre croissant.
2. On souligne les observations du groupe le plus petit (n_1).
3. On comptabilise pour chaque observation du groupe le plus grand (n_2), le nombre de fois qu'elle précède une observation du groupe le plus petit. On obtient ainsi un nombre noté U, appelé U de Mann-Whitney.
4. On rejette H_0 au niveau d'incertitude α si $U \leq U_{\alpha/2}$ ou si $U > U_{1-\alpha/2}$, les seuils critiques $U_{\alpha/2}$ et $U_{1-\alpha/2}$ étant repris dans la Table F. Sinon, on ne rejette pas H_0. En réalité, la Table F donne pour n_1 et n_2 le seuil critique supérieur $U_{1-\alpha/2}$. Pour obtenir le seuil critique inférieur, il suffit de calculer $U_{\alpha/2} = n_1 n_2 - U_{1-\alpha/2}$.

Exemple

A titre d'exemple, comparons les deux premiers échantillons de patients traumatisés et permutons-les de manière à avoir $n_1 < n_2$: $n_1 = 8$ et $n_2 = 13$ (voir Tableau 17.2). Trions les observations par ordre croissant en soulignant celles du groupe le plus petit.

$$\underline{3}, 7, 8, \underline{9}, 11, 11, 12, 16, 17, 17, 18, \underline{19}, 19, 19, \underline{23}, \underline{23}, \underline{24}, 28, \underline{29}, 38, \underline{54}$$

La première observation du groupe le plus grand est 7, celle-ci précède 7 observations du groupe le plus petit. La seconde observation du groupe le plus grand est 8, celle-ci précède aussi 7 observations de l'autre groupe. On continue ainsi de suite et on a

$$\begin{aligned} U &= 7 + 7 + 6 + 6 + 6 + 6 + 6 + 6 + 6 + 5 + 5 + 2 + 1 \\ &= 69 \end{aligned}$$

Pour $n_1 = 8$ et $n_2 = 13$ (donc $n = n_1 + n_2 = 21$), le seuil critique supérieur bilatéral au niveau d'incertitude $\alpha = 0.05$ vaut $U_{1-\alpha/2} = U_{0.975} = 80$. Dès lors, le seul critique inférieur vaut $U_{\alpha/2} = n_1 n_2 - U_{1-\alpha/2} = 104 - 80 = 24$. On conclut donc comme pour le t de Student que les deux groupes ne diffèrent pas par rapport à l'âge. En effet, la valeur observée $U = 69$ est comprise dans l'intervalle $24 - 80$.

Chapitre 18

Comparaison d'échantillons appariés

18.1 Introduction

Il est fréquent qu'une même variable quantitative soit mesurée à plusieurs reprises sur les individus d'une même population. On parle alors de mesures "répétées" (repeated measurements). On dit aussi que la variable est mesurée dans différentes "conditions expérimentales" chez les mêmes individus.

Ainsi, on peut mesurer la pression artérielle systolique d'un sujet en position assise, couchée ou debout (trois "conditions expérimentales") et se demander si on observe des différences entre les trois positions. On peut soumettre un groupe de patients obèses à quatre types de régime diététique chacun (régimes suffisamment espacés dans le temps pour qu'il n'y ait pas d'effet de "carry-over") et comparer les régimes entre eux. Le recensement du nombre de patients hospitalisés durant les quatre saisons de l'année (printemps, été, automne, hiver) dans plusieurs hôpitaux permet de voir s'il y a des différences de fréquentation en fonction des saisons.

Dans chacun des exemples cités, la même "unité statistique" (sujet, patient, hôpital) est son propre contrôle, dans la mesure où elle est soumise à chaque condition expérimentale. Cette façon de procéder permet de contrôler le facteur de variabilité entre les unités statistiques.

La mesure répétée d'une même variable X chez n sujets dans k conditions expérimentales donne lieu à un tableau à n lignes et k colonnes. Les lignes sont appelées "les blocs" (blocks) et les colonnes "les traitements" (treatments), même s'il ne s'agit pas toujours d'un traitement au sens premier du terme. En conséquence, chaque bloc (ou élément d'un bloc) est soumis aléatoirement à k traitements. La situation est analogue à celle d'une table de contingence $r \times c$ mais il est fondamental de constater que les cellules ne contiennent pas des comptages relatifs au croisement des modalités de deux variables qualitatives mais des mesures d'une variable quantitative X (qui peut être discrète ou continue).

Le cas particulier où il n'y a que $k = 2$ traitements est particulièrement fréquent en pratique, et on y reviendra par la suite. Ainsi, on peut mesurer le taux de cholestérol avant et après un repas chez des sujets obèses, la pression artérielle d'un patient au

début et à la fin d'une consultation médicale, le poids d'un patient à l'admission et à la sortie de l'hôpital, le score psychologique d'un transplanté cardiaque avant et après la transplantation. On dit qu'on a affaire à des données "appariées" (paired data) car elles ont été obtenues sur les mêmes individus. Il y a donc une corrélation entre elles.

Le schéma expérimental que nous venons de décrire est appelé "le schéma de blocs complets randomisés" (randomized complete block design) qui fut développé par R.A. Fisher en 1925. Il s'agit probablement du plan expérimental le plus puissant et le plus utilisé. Il est donc fortement conseillé en recherche.

Un bloc n'est pas nécessairement un individu ou un objet mais pourrait être constitué de sujets (ou objets) homogènes, c'est-à-dire semblables. Dès lors, si un traitement "détruit" l'unité expérimentale, elle ne peut plus être réutilisée pour un autre traitement. On prend alors une autre unité expérimentale du bloc. Par exemple, dans des expériences de laboratoire où on doit sacrifier l'animal (rat, souris), et que la réponse au traitement peut varier en fonction de la portée, chaque portée sera considérée comme un bloc et les traitements seront appliqués à des éléments différents d'une même portée.

Le traitement statistique des données du plan expérimental de Fisher porte le nom d'analyse de la variance à deux critères (two-way analysis of variance), les deux critères étant "bloc" et "traitement". On appelle aussi cette méthode ANOVA-2.

Remarque : Lorsqu'on mesure un même sujet au cours du temps, on obtient un schéma fort semblable à celui décrit, mais nous pensons qu'il s'agit davantage d'un problème de régression ! En effet, les "traitements" n'ont pas nécessairement un ordre chronologique pré-fixé.

18.2 Analyse de la variance à 2 critères

Considérons donc n blocs de sujets (en général le bloc = le sujet) distincts soumis chacun aléatoirement à k traitements, numérotés de 1 à k. Soit X une variable *quantitative* de distribution Normale et de même variance σ^2 dans chaque cellule "bloc/traitement".

Désignons par μ_{ij} la moyenne de la variable X dans le bloc i soumis au traitement j ($i = 1, \ldots, n;\ j = 1, \ldots, k$). Notons également $\mu_{.j}$ la moyenne de X dans le traitement j et $\mu_{i.}$ la moyenne de X dans le bloc i.

18.2.1 Hypothèses

Dans l'analyse de la variance à deux critères, on n'est pas intéressé par la comparaison des blocs car on sait qu'ils sont en général différents. Par contre, on aimerait savoir s'il y a des différences entre les colonnes, c'est-à-dire entre les traitements.

Les hypothèses s'écrivent donc :

H_0 : Egalité (homogénéité) des traitements

$$\mu_{.1} = \mu_{.2} = \cdots = \mu_{.k} \tag{18.1}$$

H_1 : Différence (hétérogénéité) des traitements

$$\exists j, j' \in \{1, \ldots, k\}, \quad j \neq j' : \mu_{.j} \neq \mu_{.j'} \tag{18.2}$$

18.2.2 Données

Supposons donc que n sujets (ou des sujets différents d'un même bloc) soient soumis à k traitements. Dans chaque cas, on mesure une variable quantitative X. On note x_{ij} la mesure de la variable X chez le sujet i soumis au traitement j ($i = 1, \ldots, n; j = 1, \ldots, k$).

On définit ainsi un tableau à n lignes et k colonnes dont les données sont des mesures et non le nombre de sujets au croisement de la ligne i et de la colonne j (voir Tableau 18.1). Notons que le nombre total d'observations dans le tableau vaut $n \times k$ et non $n!$ (données répétées). Il est utile de compléter ce tableau par les sommes des lignes R_i ($i = 1, \ldots, n$) et les sommes des colonnes C_j ($j = 1, \ldots, k$) ainsi que par les moyennes des lignes $\bar{x}_{i.}$ ($i = 1, \ldots, n$) et les moyennes des colonnes $\bar{x}_{.j}$ ($j = 1, \ldots, k$). Enfin $T = \sum_i R_i$ et $\bar{x} = T/nk$.

Tableau 18.1 Plan expérimental où n sujets sont soumis à k traitements

	Traitement					
Sujet	1	2	...	k	Total	Moyenne
1	x_{11}	x_{12}	...	x_{1k}	R_1	$\bar{x}_{1.}$
2	x_{21}	x_{22}	...	x_{2k}	R_2	$\bar{x}_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$	
n	x_{n1}	x_{n2}	...	x_{nk}	R_n	$\bar{x}_{n.}$
Total	C_1	C_2	...	C_k	T	
Moyenne	$\bar{x}_{.1}$	$\bar{x}_{.2}$	...	$\bar{x}_{.k}$		$\bar{x}$

Exemple : Trois systèmes de préparation des repas (A, B et C) ont été comparés chacun dans 5 hôpitaux. La variable étudiée est le temps moyen de préparation du repas (en minutes). On a pris le "repas de midi". Si on tient compte des différences entre les 5 hôpitaux, quel est le système de préparation de repas qui est le plus rapide ? Les données sont reproduites au Tableau 18.2. Dans ce problème, $n = 5$ et $k = 3$.

On a affaire à une analyse de la variance à 2 critères puisque chaque hôpital sert de référence pour chaque système de préparation de repas.

Chaque temps reproduit au Tableau 18.2 étant la moyenne sur plusieurs centaines de repas, on peut supposer que la distribution de la variable X est Normale et que la variance est homogène dans chaque cellule. Notons enfin que le bloc est l'hôpital mais que l'unité statistique sur laquelle on effectue la mesure dans l'hôpital est le "repas".

Tableau 18.2 Temps moyen (min.) de préparation des repas dans cinq hôpitaux et avec trois systèmes différents

	Système de préparation des repas				
Hôpital	A	B	C	Total	Moyenne
1	7.56	9.68	11.65	28.89	9.63
2	9.98	9.69	10.69	30.36	10.12
3	7.23	10.49	11.77	29.49	9.83
4	8.22	8.55	10.72	27.49	9.16
5	7.59	8.30	12.36	28.25	9.42
Total	40.58	46.71	57.19	144.48	
Moyenne	8.12	9.34	11.44		9.63

$S = 1430.05$

18.2.3 Niveau d'incertitude

On fixe α le niveau d'incertitude global de l'épreuve d'hypothèses, définissant ainsi le seuil maximum de rejet de H_0 si celle-ci est vraie.

18.2.4 Test statistique

Comme nous l'avons déjà évoqué, à partir du tableau des données (Tableau 18.1), on calcule les sommes (totaux) marginales

$$R_i = \sum_{j=1}^{k} x_{ij} \tag{18.3}$$

$$C_j = \sum_{i=1}^{n} x_{ij} \tag{18.4}$$

mais aussi

$$T = \sum_{i=1}^{n} R_i = \sum_{j=1}^{k} C_j \tag{18.5}$$

et

$$S = \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}^2. \tag{18.6}$$

Notons $\bar{x}_{i.} = R_i/k$ $(i = 1, \ldots, n)$, $\bar{x}_{.j} = C_j/n$ $(j = 1, \ldots, k)$ et $\bar{x} = T/nk$, les moyennes des lignes, des colonnes et la moyenne générale, respectivement.

Pour l'exemple, ces données sont reprises au Tableau 18.2. Les quantités (18.3) à (18.6) sont fondamentales car elles vont servir dans les calculs ultérieurs ! Remarquons une nouvelle fois que ces totaux ne sont pas des sommes de comptages mais de mesures : leurs unités sont donc les mêmes que celles de la variable X.

Principe

Comme pour l'analyse de la variance à un critère, on décompose la variabilité totale des $n \times k$ observations x_{ij} en variabilité due aux blocs, en variabilité due aux traitements et en variabilité résiduelle. En effet, la variabilité des temps de préparation des repas peut être expliquée par le système de préparation utilisé, mais aussi par l'hôpital où l'expérience est conduite. On ne peut cependant pas espérer que seuls le système de préparation de repas et l'hôpital expliquent toute la variabilité des données. Il reste une partie de la variabilité inexpliquée, c'est ce qu'on appelle la variabilité "résiduelle".

Si la variabilité entre les traitements est faible par rapport à la variabilité résiduelle, cela signifie qu'en moyenne les traitements sont proches les uns des autres. Ceci plaide en faveur de H_0.

Si par contre, la variabilité entre les traitements est beaucoup plus importante que la variabilité résiduelle, c'est que les traitements sont fort différents et qu'en conséquence on est dans une situation défavorable à H_0.

Il paraît donc logique de comparer la variabilité entre traitements à la variabilité résiduelle. De même, la comparaison de la variabilité entre blocs comparée à la variabilité résiduelle permet de comparer les blocs.

Sommes des carrés

L'écart de toute observation x_{ij} du tableau à la moyenne générale $\bar{x}$ peut s'écrire

$$x_{ij} - \bar{x} = (\bar{x}_{.j} - \bar{x}) + (\bar{x}_{i.} - \bar{x}) + (x_{ij} - \bar{x}_{.j} - \bar{x}_{i.} + \bar{x}). \tag{18.7}$$

Si on élève les deux membres de cette égalité au carré et que l'on somme sur i et sur j, on obtient

$$\sum_{i,j}(x_{ij} - \bar{x})^2 = \sum_{i,j}(\bar{x}_{.j} - \bar{x})^2 + \sum_{i,j}(\bar{x}_{i.} - \bar{x})^2 + \sum_{i,j}(x_{ij} - \bar{x}_{.j} - \bar{x}_{i.} + \bar{x})^2. \tag{18.8}$$

car tous les doubles produits disparaissent.

On définit ainsi trois "sommes de carrés" (sums of squares) auxquelles sont associées chaque fois des degrés de liberté.

Somme des carrés totale (Total sum of squares)

$$SCT = \sum_{i,j}(x_{ij} - \bar{x})^2 = S - \frac{T^2}{nk} \tag{18.9}$$

à $nk - 1$ degrés de liberté (attention, non pas $n - 1$ mais $nk - 1$).

Somme des carrés due aux traitements (Treatment sum of squares)

$$SCTr = \sum_{i,j}(\bar{x}_{.j} - \bar{x})^2 = \sum_{j=1}^{k} \frac{C_j^2}{n} - \frac{T^2}{nk} \tag{18.10}$$

à $k-1$ degrés de liberté.

Somme des carrés due aux blocs (Block sum of squares)

$$SCB\ell = \sum_{i,j}(\bar{x}_{i.} - \bar{x})^2 = \sum_{i=1}^{n} \frac{R_i^2}{k} - \frac{T^2}{nk} \tag{18.11}$$

à $n-1$ degrés de liberté.

Somme des carrés résiduelle, due à l'erreur (Residual sum of squares)

$$\begin{aligned} SCE &= \sum_{i,j}(x_{ij} - \bar{x}_{.j} - \bar{x}_{i.} + \bar{x})^2 \\ &= SCT - SCTr - SCB\ell \end{aligned} \tag{18.12}$$

à $(n-1)(k-1)$ degrés de liberté.

Notons que les degrés de liberté s'additionnent comme les sommes de carrés. Par analogie à (18.8), on a

$$nk - 1 = (k-1) + (n-1) + (n-1)(k-1) \tag{18.13}$$

comme on peut le vérifier aisément.

Carrés moyens

En divisant chaque somme de carrés par les degrés de liberté correspondants, on définit les variances ou "carrés moyens" (mean square). Notons que les variances ne s'additionnent pas, au contraire des sommes de carrés (18.8) et des degrés de liberté (18.13).

Ainsi donc le carré moyen résiduel est une estimation de la variance de la variable X dans chaque cellule, notée σ^2. On écrit

$$s_p^2 = \frac{SCE}{(n-1)(k-1)}. \tag{18.14}$$

Le carré moyen relatif aux traitements (colonnes) donne aussi une estimation de la variabilité de X dans chaque cellule pour autant que H_0 soit vraie, soit

$$CMTr = \frac{SCTr}{k-1}. \tag{18.15}$$

Enfin, on a aussi pour les blocs

$$CMB\ell = \frac{SCB\ell}{n-1}. \tag{18.16}$$

Rapport des carrés moyens

Si H_0 est vraie, c'est-à-dire si les traitements sont équivalents, le rapport

$$F = \frac{CMTr}{s_p^2} \tag{18.17}$$

doit être voisin de 1. On démontre que si H_0 est vraie, le critère (18.17) est distribué comme un F de Snedecor à $\nu_1 = k-1$ et $\nu_2 = (n-1)(k-1)$ degrés de liberté.

Ainsi donc, si F est élevé, ceci plaide en défaveur de H_0. Par contre, si F est voisin de 1, les données sont favorables à H_0.

De même, pour tester l'hypothèse d'égalité entre les blocs, il suffit de calculer le critère F'

$$F' = \frac{CMB\ell}{s_p^2} \tag{18.18}$$

distribué (sous H_0) comme un F de Snedecor à $n-1$ et $(n-1)(k-1)$ degrés de liberté.

Table d'analyse de la variance

On résume l'ensemble des calculs précédents dans une table d'analyse de la variance (voir Tableau 18.3).

Tableau 18.3 Table d'analyse de la variance à 2 critères

Source de variabilité	Somme de carrés	Degrés de liberté	Carré moyen	F-test
Traitement	$SCTr$	$k-1$	$CMTr$	$F = CMTr/s_p^2$
Bloc	$SCB\ell$	$n-1$	$CMB\ell$	$F' = CMB\ell/s_p^2$
Résiduel	SCE	$(n-1)(k-1)$	s_p^2	
Total	SCT	$nk-1$		

Exemple (suite)

Reprenons l'exemple du Tableau 18.2. On a successivement.

$$\begin{aligned} SCT &= 1430.05 - \frac{144.48^2}{15} \\ &= 1430.05 - 1391.63 = 38.420 \end{aligned}$$

$$\begin{aligned} SCTr &= \frac{1}{5}\left(40.58^2 + 46.71^2 + 57.19^2\right) - 1391.63 \\ &= 1419.85 - 1391.63 = 28.22 \end{aligned}$$

$$\begin{aligned} SCB\ell &= \frac{1}{3}\left(28.89^2 + 30.36^2 + 29.49^2 + 27.49^2 + 28.25^2\right) - 1391.63 \\ &= 1393.26 - 1391.63 = 1.63 \end{aligned}$$

$$\begin{aligned} SCE &= SCT - SCTr - SCB\ell \\ &= 38.42 - 28.22 - 1.63 = 8.57 \end{aligned}$$

La table d'analyse de la variance s'établit comme suit (Tableau 18.4).

Tableau 18.4 Table d'analyse de la variance à 2 critères pour la comparaison de 3 systèmes de préparation de repas dans 5 hôpitaux

Variabilité	SC	$d\ell$	CM	F
Systèmes repas	28.22	2	14.11	$F = 13.2$
Hôpitaux	1.63	4	0.408	$F' = 0.38$
Résiduel	8.57	8	1.071	
Total	38.42	14		

18.3 Seuil de décision (p-value)

Si H_0 est vraie, le critère F étant distribué comme un F de Snedecor à $k-1$ et $(n-1)(k-1)$ degrés de liberté, on décide de considérer comme trop élevée toute valeur de F supérieure au quantile $1-\alpha$, soit $Q_F(1-\alpha; k-1, (n-1)(k-1))$.

Pour l'exemple ci-dessus, on a $Q_F(0.95; 2, 8) = 4.46$ (voir Table D).

La p-value associée au F observé est donnée par

$$p = P[F_{k-1,(n-1)(k-1)} \geq F_{obs}]. \tag{18.19}$$

Pour l'exemple, l'utilisation de cette formule conduit à

$$\begin{aligned} p &= P[F_{2,8} \geq 13.2] \\ &= 0.0029. \end{aligned}$$

18.3.1 Conclusion

$$\begin{array}{l}\text{“On rejette } H_0 \text{ (égalité des traitements)}\\ \text{si } F_{obs} \geq Q_F(1-\alpha; k-1, (n-1)(k-1)),\\ \text{sinon on ne rejette pas } H_0\text{”}\end{array} \tag{18.20}$$

De manière équivalente,

“on rejette H_0 si $p(F) \leq \alpha$, sinon on ne rejette pas H_0”.

En ce qui concerne les systèmes de préparation de repas, puisque $p = 0.0029 << 0.05$ (ou puisque $F_{obs} = 13.2 >> Q_p(0.95; 2, 8) = 4.46$), on rejette H_0. On peut donc conclure que le temps moyen de préparation des repas n'est pas le même selon le système de préparation utilisé.

Remarque : On constate au vu du Tableau 18.4 qu'il n'y a pas de différence entre les hôpitaux ($F' = 0.38$ à 4 et 8 degrés de liberté, $p = 0.82$).

18.4 Comparaisons multiples

Lorsque le test F de Snedecor est non significatif, on ne rejette pas H_0 et le problème statistique est terminé.

Par contre, si le test F s'avère “significatif” et que l'on rejette H_0 (comme c'est le cas dans l'exemple), on doit se poser la question de savoir quels traitements diffèrent entre eux. On effectue à cet effet des comparaisons multiples (multiple comparisons) en comparant les groupes 2 à 2. La méthode utilisée est celle des intervalles de confiance simultanés de Scheffé, qui maintient le niveau d'incertitude global des comparaisons multiples à la valeur α fixée.

18.4.1 Différence critique

Pour chaque paire de traitements, on teste l'hypothèse nulle $H_0 : \mu_{.j} = \mu_{.j'}$ versus $H_1 : \mu_{.j} \neq \mu_{.j'}$ $(j \neq j';\ j, j' \in \{1, \ldots, k\})$.

A cet effet, on calcule la différence entre les moyennes des deux traitements et la différence critique, soient

$$d_{jj'} = \bar{x}_{.j} - \bar{x}_{.j'} \tag{18.21}$$

$$d^*(\alpha) = \sqrt{(k-1)Q_F(1-\alpha; k-1, (n-1)(k-1))s_p^2 \, \frac{2}{n}} \tag{18.22}$$

On rejette $H_0 : \mu_{.j} = \mu_{.j'}$ si

$$|d_{jj'}| \geq d^*(\alpha) \tag{18.23}$$

sinon on ne rejette pas H_0.

On procède de la même manière pour chaque comparaison. Observons que la différence critique (18.22) est constante et ne dépend pas de la paire de traitements (j, j') envisagée.

18.4.2 Exemple

Nous avons vu à la section 18.2 que les systèmes de préparation de repas diffèrent quant au temps moyen de préparation. Effectuons dès lors les comparaisons multiples et calculons la différence critique (18.22) :

$$\begin{aligned} d^*(\alpha) &= \sqrt{2 \times 4.46 \times 1.071 \times \frac{2}{5}} \\ &= 1.955 \end{aligned}$$

A versus B

En examinant le Tableau 18.2, on voit que $d_{12} = \bar{x}_{.1} - \bar{x}_{.2} = 8.12 - 9.34 = -1.22$. Puisque $|d_{12}| = 1.22 < 1.955$, on ne rejette pas H_0. Les systèmes de préparation A et B ne diffèrent pas significativement. On a donc $A = B$.

A versus C

On trouve $d_{13} = 8.12 - 11.44 = -3.32$ et puisque $|d_{13}| = 3.32 >> 1.955$, on rejette l'hypothèse d'égalité des temps moyens de préparation des systèmes A et C. On a donc $A \neq C$.

B versus C

En comparant les systèmes B et C, on trouve $d_{23} = 9.34 - 11.44 = -2.10$. Cette valeur est également significative puisque $|d_{23}| = 2.10 > 1.955$. On a donc $B \neq C$.

En conclusion, le système de préparation C diffère significativement des deux autres, mais on ne peut toutefois mettre en évidence une différence significative entre les systèmes A et B.

18.4.3 Remarque

Si l'on souhaite comparer les blocs deux à deux, on compare la différence des moyennes observées entre ces deux blocs ($\bar{x}_{i.}$ et $\bar{x}_{i'.}$) avec la différence critique

$$d^{*'}(\alpha) = \sqrt{(n-1)Q_F(1-\alpha; n-1; (n-1)(k-1))s_p^2 \, \frac{2}{k}} \tag{18.24}$$

qui ne dépend pas non plus des blocs choisis.

18.5 Comparaison de deux moyennes appariées

18.5.1 Test t-Student pour échantillons appariés

Lorsque les sujets d'une même population sont mesurés seulement dans deux conditions expérimentales ($k = 2$ traitements), on dit que les échantillons obtenus sont appariés et on peut comparer les moyennes des deux traitements μ_1 et μ_2 par une analyse de la variance à 2 critères. Toutefois, dans le cas $k = 2$, les calculs se simplifient

considérablement et on effectue en général un test "*t de Student pour échantillons appariés*" (paired t-test).

On teste donc l'hypothèse nulle $H_0 : \mu_1 = \mu_2$ vs l'hypothèse alternative $H_1 : \mu_1 \neq \mu_2$ ou de manière équivalente

$$H_0 : \Delta = \mu_1 - \mu_2 = 0 \quad \text{vs} \quad H_1 : \Delta \neq 0.$$

Les données se présentent sous la forme d'un tableau à deux colonnes qui ressemble fort à celui d'un échantillon bivarié (voir section 6.2.1). Toutefois, la situation est différente puisqu'il s'agit de la même variable X mesurée dans deux conditions expérimentales différentes plutôt que deux variables différentes X et Y mesurées chez les mêmes sujets ! (voir Tableau 18.5).

Tableau 18.5 Cas de deux échantillons appariés résultant de l'observation d'une variable X chez n sujets dans deux conditions expérimentales

	Condition		
Sujet	1	2	Différence
1	x_{11}	x_{12}	$d_1 = x_{11} - x_{12}$
2	x_{21}	x_{22}	$d_2 = x_{21} - x_{22}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{n2}	$d_n = x_{n1} - x_{n2}$

On calcule les différences $d_i = x_{i1} - x_{i2}$ $(i = 1, \ldots, n)$ des observations appariées, obtenant de ce fait une série univariée $\{d_1, \ldots, d_n\}$. On calcule la moyenne et l'écart-type des d_i, soit

$$\bar{d} = \sum d/n \tag{18.25}$$

$$s_d = \sqrt{\frac{\sum d^2 - (\sum d)^2/n}{n-1}} \tag{18.26}$$

On montre que si H_0 est vraie ($\Delta = \mu_1 - \mu_2 = 0$), le critère

$$t_{(n-1)} = \frac{\bar{d}\sqrt{n}}{s_d} \tag{18.27}$$

est distribué comme un t de Student à $n-1$ degrés de liberté.

Dès lors, on rejette H_0 si

$$|t_{obs}| \geq Q_t(1 - \alpha/2; n-1) \tag{18.28}$$

sinon on ne rejette pas H_0.

De même, on peut calculer la probabilité de dépassement associée à la valeur de t observée (voir Eq. 18.27) par la formule

$$p = 2P[t_{(n-1)} \geq |t_{obs}|] \tag{18.29}$$

et rejeter H_0 si $p \leq \alpha$.

Remarque : On montre aisément que le test F obtenu par l'analyse de la variance à 2 critères (voir Eq. 18.17) a pour valeur le t de Student (voir Eq. 18.27) au carré.

18.5.2 Exemple

A titre d'exemple, on a dosé le taux de cholestérol (g/l) chez 12 sujets avant et après un régime diététique combiné à un exercice physique. Peut-on conclure à une modification du cholestérol suite au régime ? Les données sont reprises au Tableau 18.6. On est en présence d'échantillons appariés ($n = 12$ blocs et $k = 2$ traitements).

Tableau 18.6 Taux de cholestérol (g/l) avant et après un régime diététique chez 12 sujets

Sujet	Avant régime	Après régime	Différence
1	2.01	2.00	-0.01
2	2.31	2.36	+0.05
3	2.21	2.16	-0.05
4	2.60	2.33	-0.27
5	2.28	2.24	-0.04
6	2.37	2.16	-0.21
7	3.26	2.96	-0.30
8	2.35	1.95	-0.40
9	2.40	2.07	-0.33
10	2.67	2.47	-0.20
11	2.84	2.10	-0.74
12	2.01	2.09	+0.08

Le calcul de la moyenne et de l'écart-type donne les résultats suivants, puisque $\sum d = -2.42$ et $\sum d^2 = 1.0765$,

$$\begin{aligned} \bar{d} &= \frac{-2.42}{12} = -0.202 \\ s_d &= \sqrt{\frac{1.0765 - (-2.42)^2/12}{11}} = 0.2313. \end{aligned}$$

Le test t de Student (Eq. 18.27) à $\nu = n - 1 = 11$ degrés de liberté vaut

$$\begin{aligned} t_{(11)} &= \frac{-0.202\sqrt{12}}{0.2313} \\ &= -3.02. \end{aligned}$$

Comme $Q_t(0.975; 11) = 2.20$ et que $|t_{obs}| = 3.02 > 2.20$, on rejette H_0. On peut donc affirmer qu'il y a un effet significatif du régime diététique sur le taux de cholestérol des individus. Notons que la p-value vaut $p = 2P[t_{11} \geq 3.02] = 0.0117$.

Si l'on effectue une analyse de la variance à 2 critères, on obtient la table d'analyse de la variance suivante (Tableau 18.7).

Tableau 18.7 Table d'analyse de la variance à 2 critères pour le taux de cholestérol avant et après un régime chez 12 sujets

Variabilité	SC	$d\ell$	CM	F
Régime	0.2440	1	0.244	9.12
Sujets	1.9156	11	0.174	6.51
Résiduel	0.2943	11	0.0268	
Total	2.4539	23		

On constate que $F = 9.12 = t^2 = (-3.02)^2$ et $p = P[F_{1,11} \geq 9.12] = 0.0117$. On voit aussi qu'il existe une différence significative entre les sujets ayant participé à l'expérience, puisque $p = P[F'_{11,11} \geq 6.51] = 0.0022$.

18.6 Test de Friedman

L'analyse de la variance à deux critères suppose que la distribution de la variable X soit Normale dans chaque cellule (intersection des blocs et des traitements) et que la variance y soit homogène (σ^2 constant).

S'il n'est pas possible de normaliser la distribution, on peut faire appel au test non-paramétrique de Friedman. Il faut que la variable X soit au moins ordinale.

18.6.1 Principe

Le test de Friedman procède en plusieurs étapes :

1. On remplace les observations par leur rang dans chaque bloc (ligne). Puisqu'il y a k traitements, la somme des rangs doit chaque fois valoir $k\ (k+1)/2$. En cas d'ex-aequo, on attribue la moyenne des rangs correspondants.

   En clair, pour chaque observation x_{ij} $(j = 1, \ldots, k)$ du bloc i, on associe le rang $rang_i(x_{ij})$, de telle sorte que $\sum_j rang_i(x_{ij}) = k(k+1)/2$.

2. On calcule la somme des rangs dans chaque traitement (colonnes), soit

$$R_j = \sum_{i=1}^{n} rang_i(x_{ij}) \qquad (j = 1, \ldots, k). \tag{18.30}$$

3. On calcule le critère

$$\chi^2_{(k-1)} = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \tag{18.31}$$

appelé test de Friedman et qui est distribué comme un chi-carré à $k-1$ degrés de liberté si H_0 est vraie (égalité des traitements).

4. On rejette H_0 si

$$\chi^2_{obs} \geq Q_{\chi^2}(1-\alpha; k-1) \quad \text{ou si} \quad p \leq \alpha,$$

sinon on ne rejette pas H_0.

18.6.2 Exemple

Reprenons les données du Tableau 18.2 et montrons que le test de Friedman confirme les résultats de l'ANOVA-2. Le Tableau 18.8 donne les rangs des observations pour chaque hôpital (bloc).

Table 18.8 Rangs des observations dans chaque bloc pour l'exemple de la comparaison de trois systèmes de préparation de repas dans cinq hôpitaux

	Système de préparation des repas			
Hôpital	*A*	*B*	*C*	Total
1	1	2	3	6
2	2	1	3	6
3	1	2	3	6
4	1	2	3	6
5	1	2	3	6
Total	6	9	15	30

La somme des rangs dans chaque colonne vaut

$$\begin{aligned} R_1 &= 1+2+1+1+1 = 6 \\ R_2 &= 2+1+2+2+2 = 9 \\ R_3 &= 3+3+3+3+3 = 15. \end{aligned}$$

Le critère de Friedman vaut donc puisque $n = 5$ et $k = 3$ (voir Eq. 18.31) et que le nombre de degrés de liberté $\nu = k - 1 = 2$,

$$\begin{aligned} \chi^2_{(2)} &= \frac{12}{5 \times 3 \times 4} (6^2 + 9^2 + 15^2) - 3 \times 5 \times 4 \\ &= 68.4 - 60 = 8.40. \end{aligned}$$

Comme le seuil critique à 5% du chi-carré à 2 degrés de liberté vaut $Q_{\chi^2}(0.95; 2) = 5.99$, l'hypothèse nulle est rejetée. D'ailleurs, $p = P[\chi^2_{(2)} \geq 8.4] = 0.015$. La différence des temps de préparation entre les 3 systèmes est donc significative.

On peut vérifier quels systèmes diffèrent en calculant le test de Friedman pour chaque paire de systèmes. Attention, il faut redéfinir à chaque fois les rangs !

A titre d'exemple, pour la comparaison A vs B ($n = 5, k = 2$), on a $\nu = k - 1 = 1$, $R_1 = 6$ et $R_2 = 9$, de sorte que le critère (18.31) s'écrit

$$\begin{aligned} \chi^2_{(1)} &= \frac{12}{5 \times 2 \times 3} (6^2 + 9^2) - 3 \times 5 \times 3 \\ &= 46.8 - 45 = 1.80 \end{aligned}$$

avec $p = P[\chi^2_{(1)} \geq 1.80] = 0.18$. On constate comme précédemment que les systèmes de préparation des repas A et B ne diffèrent pas significativement.

18.7 Test des rangs signés de Wilcoxon

Lorsque la variable X n'est mesurée que dans deux ($k = 2$) conditions expérimentales (ou traitements), on peut recourir aussi au test non paramétrique, dit de la somme "des rangs signés" (signed rank test) de Wilcoxon.

Principe

On procède comme suit :

1. On calcule les différences $d_i = x_{i1} - x_{i2}$ ($i = 1, \ldots, n$) comme pour le test t de Student.
2. On élimine les différences nulles éventuelles et on corrige l'effectif n en conséquence.
3. On trie les valeurs absolues $|d_i|$ non nulles ($i = 1, \ldots, n$) par ordre croissant (donc, on ne tient pas compte du signe !).
4. On associe un rang à chaque $|d_i|$, soit $rang\,(|d_i|)$, comme dans toute méthode non paramétrique.
5. On calcule la somme des rangs des observations d_i positives. On fait de même pour la somme des rangs des observations d_i négatives. La plus petite de ces deux sommes est le test de la somme des rangs signés de Wilcoxon. Soit V cette somme.

6. On rejette H_0 au niveau d'incertitude α si $V \leq V_{\alpha/2}$ (test bilatéral) ou si $V \leq V_\alpha$ (test unilatéral), seuil critique que l'on trouve dans la Table G en annexe. Sinon, on ne rejette pas H_0.

Exemple

A titre d'exemple, appliquons ce test aux données du Tableau 18.6. Trions les différences par ordre croissant sans tenir compte du signe (notons que $n = 12$ car il n'y a pas de différences nulles). On a

0.01	0.04	0.05	0.05	0.08	0.20	0.21	0.27	0.30	0.33	0.40	0.74
(1)	(2)	(3.5)	(3.5)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)

Seules deux différences sont positives (rangs soulignés). La somme des rangs d'observations négatives vaut

$$\begin{aligned} V_- &= 1 + 2 + 3.5 + 6 + 7 + 8 + 9 + 10 + 11 + 12 \\ &= 69.5 \end{aligned}$$

Celle des rangs des observations positives vaut

$$V_+ = 3.5 + 5 = 8.5$$

Donc $V = V_+ = 8.5$. Or, pour $n = 12$, le seuil critique bilatéral à $\alpha = 0.05$ vaut $V_{0.025} = 14$ (voir Table G). En conséquence, puisque $V_{obs} = 8.5 < 14$, on rejette H_0 confirmant ainsi le résultat du test t de Student.

Annexe I

Enquête du Centre Universitaire de Médecine Générale (CUMG) de Liège sur la prescription médicamenteuse par les médecins généralistes aux personnes âgées de plus de 75 ans en Région wallonne.

L'enquête a porté sur un échantillon de 355 médecins généralistes wallons. Cette annexe reprend pour chaque médecin, outre le numéro du médecin, onze variables comme suit :

X_1 = le sexe (1 = homme, 2 = femme)
X_2 = l'âge (années)
X_3 = l'expérience professionnelle (années de pratique)
X_4 = l'université d'origine (1 = ULg, 2 = UCL, 3 = ULB, 4 = autres)
X_5 = l'agrément (0 = non agréé, 1 = médecin agréé)
X_6 = la province où le médecin professe (1 = Brabant wallon, 2 = Hainaut, 3 = Liège, 4 = Luxembourg, 5 = Namur)
X_7 = la patientèle (nombre de patients par semaine) en 4 catégories : 1 = $<$ 30 patients, 2 = 31 à 60 patients, 3 = 61 à 90 patients, et 4 = $>$ 90 patients/semaine
X_8 = la moyenne du nombre de médicaments prescrits par le médecin à ses patients âgés (Nméd)
X_9 = la moyenne du nombre de problèmes présentés par les patients du médecin (Nprob)
X_{10} = la variance du nombre de médicaments prescrits par le médecin à ses patients âgés (Vméd)
X_{11} = la variance du nombre de problèmes présentés par les patients du médecin (Vprob)

Les données manquantes sont indiquées par des points.

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	1	40	5	3	1	3	4	4.86	3	3.48	2.33
2	1	46	20	1	1	3	4	3.7	1.7	0.46	0.46
3	2	38	13	2	1	1	4	3.7	2.2	1.57	0.4
4	1	34	9	1	1	3	4	3	2.1	1.56	0.99
5	1	33	8	3	1	2	4	5.5	3.3	3.39	2.23
6	1	47	22	1	1	3	4	5.2	3.2	1.96	1.51
7	1	44	20	2	1	3	1	3.7	1.7	4.68	0.9
8	1	55	27	1	1	2	1	4.2	2.7	6.62	0.46
9	1	41	16	2	1	1	3	4.1	2.3	1.66	1.79
10	2	31	6	2	1	2	3	6.7	4.1	10.6	3.43
11	1	45	20	1	1	3	4	4.3	2.6	2.23	0.93
12	1	42	17	1	1	3	4	6.2	1.9	0.84	0.54
13	1	60	34	1	1	3	2	5.1	2.8	7.43	1.73
14	1	42	15	1	1	3	3	4.6	2.2	9.16	1.73
15	2	32	8	2	1	2	3	5.75	3	5.36	0.86
16	1	41	16	1	1	3	4	5.6	3.4	8.04	3.6
17	1	45	17	2	1	2	4	6.6	3	7.6	0.44
18	1	47	21	2	1	2	4	4.1	2.8	2.1	1.29
19	1	51	27	2	1	2	4	3.2	2.8	3.73	1.51
20	1	72	44	1	1	3	1	3.33	2.67	3	0.75
21	2	44	18	1	1	3	2	4.4	2	6.04	0.89
22	1	38	12	2	1	2	4	7.1	4.4	3.21	2.27
23	1	40	14	2	1	5	4	4.5	2.2	2.06	1.29
24	1	54	26	1	1	5	3	5.1	2.6	3.43	0.93
25	1	39	12	1	1	3	4	4.1	3	1.88	2
26	1	41	17	1	1	3	3	5.4	3.1	11.3	2.77
27	2	36	11	1	1	3	2	4.5	2.8	2.06	0.62
28	1	26	1	2	0	2	3	6.5	3.5	6.5	0.5
29	1	44	20	1	1	3	4	4.89	1.67	1.11	0.5
30	2	36	10	1	1	3	2	4.4	2.6	4.93	0.71
31	1	35	10	1	1	3	3	5.3	2.7	6.01	0.23
32	1	54	29	2	1	2	3	3.2	1.8	1.29	0.62
33	1	46	13	3	1	5	2	3.6	2.3	6.27	2.23
34	2	31	5	1	0	3	1	3.75	2.75	2.25	1.58
35	1	55	30	3	1	1	4	3.8	2.6	1.29	0.49
36	1	42	16	2	1	3	4	4.1	1.6	3.66	0.93
37	1	45	20	2	1	2	4	3.8	1.7	0.4	0.68
38	1	42	17	3	1	2	4	5	2.2	2.89	1.51
39	1	68	41	2	1	1	3	4.6	2.4	3.16	0.49

Données de l'enquête CUMG

N° MG	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
40	1	30	3	1	0	3	2	3.5	2.2	1.61	0.4
41	1	37	12	2	1	1	4	4.8	2.6	6.4	0.71
42	1	55	28	2	1	3	4	4.8	4	4.62	2
43	1	54	28	1	0	3	4	2.9	1.5	0.54	0.28
44	1	72	46	1	1	4	2	6.1	3.8	10.1	2.4
45	1	59	32	2	1	2	2	5.2	1.6	7.96	0.49
46	1	46	21	2	1	2	4	3	2.5	1.56	0.28
47	1	38	11	1	1	3	3	5.1	3	4.99	2.67
48	1	42	17	2	1	2	4	5.1	3.6	4.32	1.38
49	1	67	41	2	0	4	1	2	1.8	2	0.7
50	2	37	13	2	1	2	4	4.5	1.9	1.61	0.77
51	1	66	40	1	1	3	4	5.1	2	2.1	0.89
52	1	36	10	2	1	2	4	3.1	1.8	1.21	0.62
53	1	39	14	2	1	4	3	5.7	2.8	7.57	1.29
54	1	42	18	1	1	3	3	5.4	3.7	2.49	2.46
55	1	63	38	1	1	3	2	4.5	3.7	5.61	0.9
56	1	47	16	1	1	3	2	4.9	2.3	3.43	2.46
57	1	49	23	2	1	1	4	3.57	2	0.95	1.33
58	1	34	8	2	1	3	4	4.6	3.1	3.38	1.66
59	2	43	19	1	1	3	4	5.2	1.9	3.73	0.77
60	1	35	9	3	1	2	4	5.1	3.4	3.66	1.38
61	1	50	25	1	1	3	4	4.4	2	2.04	0.67
62	1	38	13	3	1	2	4	5.5	2.6	4.06	1.6
63	1	57	30	3	1	1	4	4.2	3.3	1.29	0.46
64	1	53	27	1	1	3	4	6.3	2.7	4.9	0.68
65	1	39	13	2	1	2	2	2.7	2.2	2.68	0.4
66	2	40	15	1	1	2	4	4.4	2.1	6.04	0.77
67	1	29	2	1	0	3	3	5.2	1.9	7.73	0.54
68	2	40	15	1	1	3	2	5.7	3.1	2.9	0.77
69	1	43	18	1	1	3	4	5.6	3.2	2.93	1.07
70	1	52	27	1	1	3	4	5.2	4	3.29	4
71	1	44	18	1	1	3	3	7.4	3.9	0.71	0.77
72	1	35	9	2	1	5	4	3.8	2.5	1.07	1.17
73	1	42	17	3	1	2	4	3.5	1.4	2.5	0.27
74	1	47	19	1	1	2	4	4.3	2.6	4.01	0.93
75	1	43	14	2	1	1	4	2.5	2.5	0.72	2.06
76	1	33	6	2	1	4	2	5.22	2.67	7.44	1
77	1	63	38	2	1	5	3	6.3	3.8	5.57	0.84
78	1	38	11	3	1	1	3	5	2.3	8.22	1.12
79	1	43	17	1	1	3	4	3.8	2.6	2.62	1.38

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
80	1	60	34	1	1	3	3	5.3	2.2	2.9	1.51
81	2	40	13	3	1	1	2	4.1	2.3	1.66	1.57
82	1	36	10	2	1	2	3	5.1	3.3	1.21	2.01
83	1	46	21	1	1	2	3	4.8	2.8	5.51	1.96
84	1	43	18	1	1	3	3	4.2	2	2.4	1.56
85	1	51	26	2	1	2	4	5.4	2.9	2.93	0.77
86	1	34	7	1	1	3	4	5.4	2.2	1.16	0.4
87	2	35	9	2	1	5	1	2.25	2.25	2.25	3.58
88	2	61	35	1	1	2	3	5.5	2.5	1.61	0.5
89	1	54	28	1	1	3	4	5.1	2.8	1.88	0.62
90	1	53	28	2	1	2	4	4.5	1.5	1.39	0.28
91	1	47	20	1	1	5	4	6.2	2.6	2.18	1.82
92	1	31	6	2	1	2	4	4.6	2.9	3.38	2.32
93	1	44	17	2	1	4	4	4.6	3.5	4.93	0.28
94	1	47	19	4	1	3	4	3.5	2.7	2.5	0.9
95	1	44	19	2	1	2	4	6.9	3.6	4.1	2.49
96	1	38	12	3	1	5	2	3	2.33	1	0.33
97	1	47	20	1	1	3	1	5.1	2	1.21	0.67
98	1	39	15	3	1	2	4	5.1	2.8	4.1	1.73
99	1	60	32	1	1	3	4	5.7	2.6	1.79	1.6
100	1	38	13	2	1	1	2	6.2	2.5	3.96	1.61
101	1	43	18	2	1	5	3	6.75	2.5	20.5	0.57
102	1	27	2	2	0	5	1	2	1	8	0
103	2	32	6	2	1	2	1	3.14	1.86	1.48	1.48
104	1	43	18	2	1	2	4	3.7	3.1	2.68	0.54
105	1	35	9	2	1	2	4	5.3	2.9	2.68	1.21
106	1	36	10	1	1	3	4	5.1	1.9	5.88	0.32
107	1	55	28	2	1	2	4	3.4	1.9	2.27	0.54
108	1	81	55	2	1	3	1	3.8	2.1	2.84	0.54
109	1	30	5	2	1	3	3	3.4	2.8	2.93	2.62
110	1	54	29	2	1	5	2	6.2	3.2	4.18	1.51
111	1	28	2	2	0	5	3	5.7	2.3	3.12	1.57
112	1	57	31	2	1	2	4	3.2	2.7	3.07	2.23
113	1	35	10	2	1	2	4	6.3	2.2	10	1.51
114	1	38	13	3	1	1	4	5.9	4.1	2.1	0.77
115	1	62	37	1	1	3	4	6.1	3.2	6.77	1.73
116	1	38	13	2	1	3	4	4.9	2.6	8.32	2.04
117	1	37	10	3	1	2	3	5.3	3.9	3.12	1.88
118	2	26	1	2	0	4	1	3	1	.	.
119	2	34	9	3	1	2	4	4.1	2.6	2.54	1.6

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
120	1	42	16	4	1	2	4	4.7	2.3	2.23	0.46
121	1	31	4	3	1	2	3	5.6	4	2.71	2.22
122	1	45	20	2	1	1	2	3.3	2.5	0.68	0.5
123	2	34	9	2	1	2	3	5.4	2.4	4.04	0.71
124	2	44	18	1	1	3	2	3.86	2.43	3.81	0.29
125	1	44	19	3	1	2	4	5.8	3.6	4.4	0.27
126	1	57	32	2	1	2	3	5.6	1.8	6.49	0.4
127	1	55	30	2	1	2	4	4.4	1.7	4.27	0.46
128	1	57	32	1	1	3	4	4.6	2.6	1.16	0.93
129	1	37	11	1	1	3	4	4.6	2.9	3.6	0.99
130	1	50	25	2	1	2	4	5.3	2.4	4.68	1.16
131	1	44	18	2	1	3	4	4.6	2.2	4.49	0.84
132	1	37	9	1	1	3	3	5.5	3.2	4.06	0.4
133	1	61	34	2	1	2	3	3.6	1.5	1.16	0.72
134	1	35	10	2	1	2	4	2.8	1.1	2.18	0.77
135	1	63	37	1	1	3	3	6.5	3.6	5.83	1.82
136	1	.	21	2	.	2	.	4.4	2.6	3.6	0.49
137	2	37	13	2	1	4	1	3.25	1.75	3.93	0.21
138	1	44	19	1	1	3	4	6.1	4.5	7.66	4.94
139	1	30	4	2	1	2	2	5.4	2.2	9.38	0.84
140	1	39	11	1	1	3	2	5	2.1	2.67	0.99
141	2	60	31	2	1	4	1	5.7	2.8	6.23	1.29
142	1	58	31	2	1	3	4	5.2	2.6	6.84	1.16
143	1	39	14	2	1	2	4	5.1	2.8	4.54	1.73
144	1	35	8	2	1	3	4	5.6	3.2	2.27	1.07
145	1	44	19	1	1	3	3	2.8	2.6	2.62	1.6
146	1	42	16	3	1	2	4	4.6	3.8	6.04	3.29
147	1	34	7	2	1	2	4	6.4	4.8	5.6	2.62
148	1	34	9	1	0	3	2	4.67	3.33	1.33	0.33
149	1	55	25	1	1	5	3	4.1	2	6.77	1.11
150	1	35	9	2	1	5	4	7.4	3.4	6.71	1.82
151	1	74	49	1	1	2	4	4.7	1.9	4.23	0.99
152	1	45	19	2	1	4	2	4.5	1.7	5.17	0.46
153	1	43	18	2	1	2	4	6.8	3.4	6.62	1.38
154	1	31	6	3	1	5	3	5.1	3.1	3.43	0.99
155	1	36	10	3	1	5	3	5.89	2.56	9.61	2.03
156	2	32	6	1	1	3	4	3.2	1.7	3.73	0.68
157	1	34	8	2	1	1	2	3.88	2.75	4.13	1.07
158	1	41	16	1	1	2	3	4.3	3.3	9.12	1.12
159	1	45	19	2	1	5	3	6.4	3.2	4.27	1.29

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
160	1	33	6	2	1	2	2	6.25	3.5	4.92	1.67
161	1	42	17	3	1	2	3	2.7	2	1.79	0.89
162	1	48	22	2	1	2	4	3.4	1.8	1.38	0.4
163	1	36	10	2	1	2	4	2.5	0.9	2.06	0.32
164	2	36	10	2	1	2	1	4.9	2.4	2.99	1.16
165	1	68	40	2	0	2	3	3.67	1.56	1.25	0.28
166	1	41	16	2	1	2	3	3.8	2.2	3.73	1.96
167	2	30	5	1	1	3	2	5.25	1.75	2.25	0.92
168	1	65	38	1	0	3	3	3.6	1.4	1.16	0.71
169	1	43	18	1	1	4	4	5.7	2.3	6.01	0.9
170	1	36	11	1	1	3	3	4.6	2.3	5.6	1.34
171	1	44	14	3	1	2	4	4.2	2	3.73	1.33
172	1	55	30	2	1	2	4	5.6	3.3	5.82	0.46
173	1	40	16	2	1	2	4	4.3	4.9	1.57	1.66
174	2	35	8	2	1	4	2	4	1.25	2	0.92
175	1	43	18	3	1	2	4	4.3	2.2	3.12	1.07
176	1	.	26	2	.	1	.	4	3.25	0.67	1.58
177	1	46	20	1	0	3	4	5	2.9	5.33	0.77
178	1	58	32	1	1	3	4	4.8	1.6	4.84	0.49
179	1	46	22	2	1	2	4	4.3	3.1	3.79	1.66
180	1	56	32	2	1	2	4	4.6	2.9	1.6	1.21
181	1	37	11	1	1	2	3	6	3.9	8.22	2.77
182	1	40	15	4	1	3	3	3.9	3	3.21	2.22
183	2	33	8	1	1	3	2	4.8	2.7	4.4	2.23
184	1	62	34	1	1	5	2	4.9	2	4.77	0.67
185	1	39	14	2	1	2	4	5.6	3.2	8.04	1.29
186	2	32	5	1	1	3	3	7.3	5	4.46	2
187	2	38	11	3	1	2	2	5.33	2.67	7.07	1.87
188	1	41	16	2	1	5	2	5.8	2.6	5.51	1.38
189	1	38	11	2	1	3	2	5	2.5	8	0.94
190	1	57	32	1	1	3	3	4.9	1.7	2.99	0.46
191	1	40	15	2	1	4	2	4.9	2.9	5.43	1.66
192	2	37	11	1	1	3	4	7.1	3.9	14.1	3.21
193	1	41	13	2	1	2	4	5.3	2	5.57	1.11
194	1	41	16	1	1	3	4	4.2	1.6	1.07	0.71
195	2	40	15	3	1	1	2	5.56	3.44	9.78	2.03
196	2	29	4	1	0	3	4	4.5	3.5	7.39	3.83
197	1	46	19	1	1	3	4	6.1	3.8	4.54	1.07
198	1	46	21	3	1	2	2	4.4	3.7	5.38	2.01
199	1	49	23	2	1	2	4	4.4	2.4	2.49	1.6

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
200	1	31	5	2	1	2	4	5.3	3.1	6.23	0.54
201	1	50	25	1	1	2	4	3.5	2.4	2.06	1.38
202	1	40	14	2	1	2	4	4.2	2.1	5.96	0.99
203	1	30	3	2	0	2	3	4.5	2.5	6.72	2.06
204	1	57	31	3	1	2	4	3.6	1.9	0.71	1.43
205	1	50	22	1	1	1	4	3.7	1.9	2.23	0.54
206	2	34	8	1	1	3	3	2.6	1.6	2.04	0.71
207	1	48	21	2	1	1	3	4.9	4	6.99	1.56
208	1	42	15	1	1	4	2	3.6	1.3	1.82	0.68
209	1	28	4	2	0	2	3	5.44	4.11	8.03	6.11
210	1	59	33	2	1	5	4	4.1	3.3	2.99	1.57
211	1	44	19	2	1	5	3	8	2.6	3.56	0.71
212	1	39	12	1	1	3	4	3.5	3.8	3.39	1.51
213	1	66	40	2	1	4	2	4	2	3.78	0.89
214	1	52	27	2	1	2	4	2.9	2.1	1.43	1.21
215	1	40	15	2	1	1	4	3.4	2.5	2.93	0.72
216	2	44	18	4	1	5	4	3.7	2.6	4.9	2.04
217	2	32	8	2	1	4	2	4.6	2	4.71	1.33
218	1	67	40	2	1	2	4	3.6	3	3.82	2
219	1	32	7	2	1	5	4	4.9	3.1	6.1	1.21
220	2	27	2	2	0	3	1	3.4	1.6	4.8	0.3
221	1	38	13	2	1	5	4	4	3.1	1.33	0.77
222	1	62	37	2	1	2	3	3.1	1.7	2.54	0.9
223	1	28	3	2	1	3	3	5.6	3.9	9.82	3.88
224	1	46	20	1	1	3	4	4.9	2.8	4.1	1.29
225	1	64	38	1	1	3	4	5.5	2.4	10.2	1.6
226	1	55	29	1	1	3	3	4.1	2.5	3.66	0.72
227	1	46	14	1	1	3	4	7.2	4.6	3.29	3.16
228	1	41	13	1	1	3	4	5.5	1.9	2.72	0.54
229	1	33	6	2	1	2	1	4.6	3.4	4.04	1.38
230	2	26	1	1	0	3	1	3.71	1.86	5.9	1.48
231	1	43	15	1	1	3	2	4.1	1.8	2.99	0.4
232	1	38	13	1	1	2	4	2.7	2	1.57	0.44
233	1	40	14	2	1	5	4	4.1	2.4	3.43	0.49
234	1	39	8	2	1	2	4	3.2	2.4	1.29	1.82
235	1	46	20	2	1	3	1	6.2	2.6	0.7	0.3
236	1	46	19	1	1	3	3	5.3	3.3	9.34	1.12
237	1	43	16	1	1	4	4	4.2	2.6	4.18	0.71
238	1	30	3	3	1	5	2	4.14	2.43	4.14	1.62
239	1	43	18	1	1	3	4	4.9	2.5	2.54	0.94

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
240	1	64	39	1	1	3	2	3.8	2.4	3.96	1.16
241	1	27	1	2	0	5	2	4.4	3.6	3.38	3.6
242	1	35	10	1	1	3	4	5.2	2.9	5.07	1.21
243	1	40	16	1	1	4	4	4	1.9	3.56	0.32
244	1	40	15	1	1	2	3	5.8	1.8	7.73	0.62
245	1	36	11	3	1	2	3	4.8	2.6	2.84	0.27
246	1	44	15	2	1	1	4	2.2	2	0.84	0.44
247	2	40	11	2	1	5	3	3.8	2.2	1.96	0.62
248	1	53	28	2	1	2	4	4.1	2.4	3.66	0.93
249	1	40	15	1	1	5	3	4.3	2.4	3.57	2.27
250	2	42	17	1	1	3	2	3.5	2.4	0.94	0.71
251	2	31	5	1	1	3	2	5.2	3	8.84	2
252	1	43	17	1	1	3	4	3.57	2.43	1.29	1.29
253	1	77	52	3	1	5	2	3.3	1.5	2.23	0.5
254	2	40	15	2	1	2	2	3.5	1.7	0.72	0.46
255	1	40	13	2	1	5	4	6.6	3.2	6.93	3.07
256	1	47	21	2	1	3	3	4.9	3.4	2.32	1.38
257	1	62	39	2	1	2	4	5.5	3.3	4.28	2.23
258	2	58	33	1	1	3	2	8.5	2.7	7.61	1.12
259	1	40	15	2	1	2	4	5.7	3.5	4.46	2.5
260	1	58	33	2	1	2	4	3.9	2.6	3.66	1.6
261	1	36	11	1	0	3	2	3.4	1.7	1.16	0.68
262	1	35	9	1	1	3	4	4	2.4	3.56	1.38
263	1	35	10	1	1	3	2	5.9	2.3	9.43	2.23
264	1	54	29	1	1	5	2	5.9	2.4	2.1	0.71
265	1	58	33	3	1	1	4	3.2	1.7	3.96	1.12
266	2	38	13	2	1	5	1	6	2.71	7.67	1.57
267	1	37	3	1	1	4	2	3	2.4	4	1.38
268	1	37	10	2	1	2	4	4.2	2	3.29	0.67
269	1	37	11	3	1	2	4	7.1	4.2	6.54	2.18
270	1	37	11	1	1	3	3	3.9	3.4	1.66	0.71
271	1	48	21	4	1	2	4	2.75	1.5	0.5	0.57
272	2	49	25	1	1	2	1	3.25	1.25	3.58	0.25
273	1	45	19	2	1	2	4	6.2	2.7	1.51	0.9
274	1	52	24	1	0	3	1	5.9	2.6	4.1	1.82
275	1	35	10	1	1	3	4	3.9	3.5	2.32	2.5
276	1	37	12	2	1	1	2	4.7	2.9	6.01	2.1
277	1	51	20	1	1	3	3	4.3	3	3.57	1.11
278	1	61	36	1	1	3	3	4.9	2.6	7.43	2.04
279	2	36	10	1	1	1	2	6.33	3	2.33	4

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
280	1	63	38	2	0	3	3	4.6	1.4	2.49	0.27
281	2	41	16	2	1	2	2	5.2	2.6	5.29	0.93
282	2	27	2	2	0	1	1	3.4	2	4.04	1.33
283	2	43	11	1	1	3	2	3.9	3.6	3.66	1.82
284	2	27	1	2	0	1	1	4.5	2.13	12.8	0.41
285	1	35	9	2	1	2	4	5.2	2.1	3.96	0.99
286	1	44	17	1	1	5	4	5.3	2.6	6.23	1.16
287	2	40	14	1	1	3	2	5.3	2.8	7.12	1.07
288	1	59	32	2	1	2	3	3.1	1.6	0.77	0.49
289	1	50	25	2	1	3	2	5.3	2.2	3.57	0.4
290	1	39	13	2	1	4	3	3.8	2	1.73	0.44
291	2	37	12	3	1	2	2	4.2	2.7	5.96	0.68
292	1	43	18	2	1	2	4	4.9	2.7	6.32	1.79
293	2	31	6	2	1	2	4	4.2	4.5	5.29	2.94
294	1	72	40	1	1	3	1	6.6	3.1	6.93	1.21
295	2	37	12	3	1	1	4	2.5	1.5	1.17	0.72
296	1	65	38	3	1	1	3	4.8	2.6	1.96	0.49
297	1	35	10	2	1	2	4	4.5	2.4	6.72	1.16
298	1	37	11	1	1	3	4	3.9	1.5	0.99	0.5
299	1	42	16	1	1	3	3	4.4	2.6	3.16	0.71
300	1	45	20	1	1	5	3	5.1	3.8	9.43	3.96
301	1	67	36	2	1	5	2	4.6	2.4	3.6	1.82
302	1	70	41	2	1	2	1	2.6	1.6	1.6	0.71
303	1	41	15	2	1	4	3	5.8	3.3	7.07	1.12
304	1	65	41	1	1	3	4	5.4	2.4	4.04	0.27
305	1	46	21	2	1	5	4	6.3	2.5	7.57	0.94
306	1	34	9	1	1	5	3	2.83	1.33	0.57	0.27
307	1	74	48	2	1	2	2	4.3	2.8	3.79	0.62
308	1	46	19	2	1	2	4	4.1	2.7	7.43	1.79
309	1	44	15	1	1	3	4	6	2.2	6.44	0.62
310	1	29	2	1	0	3	2	4.67	2.17	2.67	0.57
311	1	43	16	1	1	3	2	4	2.5	3.33	0.33
312	1	44	19	2	1	5	3	4.3	3.6	2.01	1.16
313	2	39	14	2	1	3	2	4.3	1.7	3.57	0.23
314	1	30	6	1	1	3	4	4.9	2.4	7.88	0.93
315	1	43	18	2	1	2	4	4.5	2.5	4.06	1.17
316	1	63	38	1	1	3	3	4.2	2.1	2.18	0.54
317	2	39	15	3	1	2	4	4.3	2	2.46	0.67
318	2	38	11	1	1	3	2	4.5	4	10	8.22
319	1	59	32	2	1	2	4	4.8	2.6	1.51	1.38

Données de l'enquête CUMG

$N°\ MG$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
320	2	31	5	2	1	5	1	5.75	3.5	2.25	1.67
321	2	33	8	2	1	5	2	4.4	4.1	3.16	4.32
322	1	49	23	3	1	2	4	4.4	1.9	4.04	0.99
323	1	35	11	3	1	2	3	4.1	3.5	5.66	4.72
324	1	54	27	1	1	5	4	5.7	3.1	6.46	1.43
325	1	46	22	2	1	5	4	3.7	2	0.46	1.11
326	1	32	7	3	1	3	4	5.6	3.8	8.93	3.29
327	1	54	25	2	1	5	4	5.1	2.6	2.99	0.49
328	1	33	8	2	1	5	3	6.1	4	8.1	2.22
329	2	38	12	2	1	1	1	4.88	3.38	2.13	1.98
330	1	40	15	1	1	3	4	6.2	2.7	5.73	1.12
331	1	37	12	1	1	3	4	4.2	2.2	3.07	0.4
332	1	61	35	1	1	3	4	5.9	3	5.88	0.67
333	1	34	8	1	1	3	3	5.6	2.4	10	0.93
334	2	34	9	2	1	1	2	5	2.17	7.2	0.57
335	1	40	15	2	1	2	3	5.4	3.2	3.38	2.18
336	1	45	19	2	1	2	4	5.3	3.2	4.23	1.07
337	1	42	17	2	1	5	4	4.9	2.7	1.43	0.68
338	2	34	9	3	1	1	2	4.33	2.33	6.33	1.33
339	1	.	26	2	.	2	.	3	2.1	1.33	1.88
340	1	49	21	3	1	1	3	4	2	1.56	0.67
341	1	34	9	1	1	2	4	4.6	2.2	5.82	1.96
342	1	42	17	2	1	3	3	4	1.9	2.22	0.54
343	1	43	18	1	1	3	4	5.3	2.2	5.79	0.62
344	1	46	18	1	1	3	3	5.3	2.5	1.34	0.94
345	1	35	8	3	1	2	4	4.9	2.5	3.43	0.5
346	1	40	16	3	1	5	4	2.5	1.6	2.28	0.27
347	1	44	16	1	1	3	3	4.33	1.78	4.5	1.19
348	1	55	30	1	1	3	3	5.1	3.1	3.21	0.77
349	1	27	2	2	0	4	2	4.2	2.5	1.73	1.17
350	1	66	42	2	1	1	3	3.1	1.6	5.66	0.71
351	1	33	8	2	1	2	4	4	2.4	1.33	0.71
352	1	35	8	2	1	2	2	5.5	3.6	9.61	1.16
353	1	40	13	3	1	2	4	4.2	1.8	1.07	0.62
354	1	49	25	3	1	2	2	4.2	2.9	3.29	1.21
355	1	45	20	2	1	5	4	4.7	2.7	7.12	2.68

Annexe II

Dosage du glucose (mmol/l) dans un même échantillon contrôle envoyé par le Ministère de la Santé Publique à 545 laboratoires de biologie clinique en Belgique dans le cadre de l'Evaluation Externe de la Qualité*

0.30	2.14	2.23	2.42	2.42	2.44	2.46	2.47	2.50	2.53	2.55	2.58	2.68	2.69	2.69
2.70	2.75	2.75	2.77	2.79	2.80	2.80	2.81	2.81	2.83	2.83	2.83	2.86	2.86	2.86
2.86	2.86	2.88	2.88	2.89	2.89	2.89	2.90	2.90	2.90	2.90	2.90	2.91	2.91	2.91
2.91	2.91	2.92	2.92	2.92	2.92	2.92	2.92	2.93	2.94	2.94	2.94	2.97	2.97	2.97
2.97	2.97	2.97	2.97	2.97	2.97	2.97	2.97	2.97	2.97	2.97	2.97	2.97	2.99	2.99
3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.01	3.01	3.01	3.02	3.02	3.02	3.02	3.02
3.02	3.02	3.02	3.02	3.02	3.02	3.03	3.03	3.03	3.03	3.03	3.03	3.03	3.03	3.03
3.03	3.03	3.03	3.03	3.03	3.03	3.03	3.03	3.03	3.05	3.05	3.05	3.05	3.05	3.05
3.05	3.05	3.05	3.05	3.07	3.07	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08
3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08	3.08
3.08	3.08	3.08	3.08	3.08	3.09	3.10	3.10	3.10	3.10	3.10	3.10	3.11	3.11	3.11
3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11	3.11
3.11	3.11	3.11	3.11	3.12	3.12	3.12	3.13	3.13	3.13	3.13	3.13	3.13	3.13	3.13
3.13	3.13	3.13	3.13	3.13	3.13	3.13	3.13	3.13	3.13	3.13	3.14	3.14	3.14	3.14
3.14	3.14	3.14	3.14	3.14	3.14	3.14	3.14	3.14	3.14	3.14	3.14	3.14	3.15	3.15
3.15	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16	3.16
3.16	3.16	3.18	3.18	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19	3.19
3.19	3.19	3.19	3.19	3.19	3.19	3.20	3.20	3.20	3.20	3.20	3.22	3.22	3.22	3.22
3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.22	3.23	3.23	3.24	3.24	3.24	3.24
3.24	3.24	3.24	3.24	3.24	3.25	3.25	3.25	3.25	3.25	3.25	3.25	3.25	3.27	3.27
3.27	3.27	3.27	3.27	3.27	3.27	3.27	3.28	3.28	3.30	3.30	3.30	3.30	3.30	3.30
3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30	3.30
3.30	3.30	3.30	3.30	3.30	3.31	3.33	3.33	3.33	3.33	3.33	3.33	3.33	3.33	3.33
3.33	3.33	3.33	3.33	3.33	3.33	3.33	3.34	3.35	3.35	3.35	3.35	3.35	3.35	3.35
3.35	3.35	3.36	3.36	3.36	3.36	3.36	3.38	3.38	3.38	3.39	3.39	3.39	3.39	3.39
3.39	3.39	3.39	3.40	3.40	3.40	3.40	3.41	3.41	3.41	3.41	3.41	3.41	3.41	3.41
3.41	3.41	3.41	3.41	3.41	3.41	3.41	3.41	3.42	3.42	3.43	3.44	3.44	3.44	3.44
3.44	3.44	3.44	3.44	3.44	3.44	3.44	3.44	3.45	3.45	3.46	3.46	3.46	3.46	3.46
3.46	3.46	3.46	3.47	3.47	3.47	3.47	3.47	3.47	3.47	3.48	3.48	3.49	3.49	3.49
3.49	3.49	3.50	3.50	3.50	3.50	3.50	3.50	3.50	3.50	3.51	3.51	3.52	3.52	3.52
3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.52	3.54	3.54	3.57	3.57	3.57
3.57	3.58	3.58	3.58	3.58	3.60	3.60	3.60	3.61	3.61	3.61	3.61	3.63	3.63	3.63
3.63	3.63	3.63	3.63	3.63	3.66	3.66	3.66	3.66	3.68	3.68	3.68	3.68	3.68	3.68
3.68	3.69	3.70	3.72	3.74	3.74	3.74	3.74	3.75	3.77	3.77	3.79	3.79	3.80	3.80
3.81	3.83	3.83	3.85	3.87	3.88	3.89	3.89	3.94	3.96	3.96	3.96	4.00	4.01	4.02
4.03	4.12	4.12	4.16	4.18	4.29	4.29	4.29	4.45	4.56	4.70	4.84	4.94	5.50	8.66
11.4	13.8	13.9	14.0	344										

* Les résultats sont triés par ordre croissant

Annexe III

Durée de vie (mois), sexe ($0 = H$, $1 = F$) et âge (années) de patients atteints d'un adénocarcinome du rectum et traités par une radiothérapie pré-opératoire < 5000 rad ($n_1 = 21$, groupe 1) ou ≥ 5000 rad ($n_2 = 35$, groupe 2). Source : Harris et Albert, 1991

Groupe 1 (< 5000 rad)			Groupe 2 (≥ 5000 rad)		
Survie	Sexe	Age	Survie	Sexe	Age
7	0	68	9	1	77
9	0	69	12	1	55
12	0	68	12*	1	78
12	0	71	13*	1	47
19	1	77	14*	0	69
23	0	70	16	1	68
24	0	67	18*	0	62
24	1	68	19	1	60
24	1	88	23*	1	54
24	1	89	24*	0	62
29*	1	28	25*	1	55
34	1	73	26*	0	39
41	0	60	27	1	50
54	0	60	29*	0	58
72*	1	44	30*	0	75
78	0	82	32*	0	61
80*	0	62	33*	1	53
83*	0	53	33*	0	57
92*	0	66	35	0	50
139*	0	63	35	0	78
139*	1	68	35*	0	55
			35*	0	65
			35*	0	73
			36	1	53
			38*	0	47
			51*	0	60
			54*	0	54
			57	0	66
			60*	1	64
			67	0	60
			70	0	41
			87*	0	58
			89*	0	45
			98*	0	73
			120*	0	63

* Durée de vie censurée

Annexe IV

Nombres aléatoires

10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
63573	32135	05325	47048	90553	57548	28168	28709	83491	25624
73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
11805	05431	33808	27732	50725	68248	29405	24201	52775	67851
83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
15474	45266	95270	79953	59367	83848	82396	10118	33211	59466
94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
42481	16213	97344	08721	16868	48767	03071	12059	25701	46670
23523	78317	73208	89837	68935	91416	26252	29663	05522	82562
04493	52494	75246	33824	45862	51025	61962	79335	65337	12472
00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
35963	15307	26898	09354	33351	35462	77974	50024	90103	39333
59808	08391	45427	26842	83609	49700	13021	24892	78565	20106
46058	85236	01390	92286	77281	44077	93910	83647	70617	42941

Annexe V

Tables statistiques

Table A Loi Normale $Z \sim N(0,1)$.
Probabilités supérieures $\alpha = P[Z > z]$, $z \geq 0$

Table B Loi Chi-carré à ν degrés de liberté χ^2_ν
Quantiles supérieurs $Q_{\chi^2}(1-\alpha;\nu)$

Table C Loi t de Student à ν degrés de liberté t_ν
Quantiles supérieurs $Q_t(1-\alpha;\nu)$
Note : la ligne $\nu = \infty$ correspond aux quantiles supérieurs gaussiens $Q_Z(1-\alpha)$

Table D Loi F de Snedecor à ν_1 et ν_2 degrés de liberté F_{ν_1,ν_2}
Quantiles supérieurs $Q_F(1-\alpha;\nu_1,\nu_2)$ pour $\alpha = 0.05$ et $\alpha = 0.01$

Table E Coefficient de corrélation r
Quantiles supérieurs $r^*(n;1-\alpha)$

Table F Test U de Mann-Whitney
Quantiles supérieurs $U_{1-\alpha} = U(n_1,n_2;1-\alpha)$
Note : $U_\alpha = n_1 n_2 - U_{1-\alpha}$

Table G Test V des rangs signés de Wilcoxon.
Quantiles inférieurs $V_\alpha = V(n;\alpha)$

Table A Loi Normale $Z \sim N(0,1)$
Probabilités supérieures $\alpha = P[Z > z]$, $z \geq 0$

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	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002

Table B Loi Chi-carré à ν degrés de liberté χ^2_ν
Quantiles supérieurs $Q_{\chi^2}(1-\alpha;\nu)$

	$1-\alpha$					
ν	0.90	0.95	0.975	0.99	0.995	0.999
1	2.706	3.841	5.024	6.635	7.879	10.83
2	4.605	5.991	7.378	9.210	10.60	13.82
3	6.251	7.815	9.348	11.34	12.84	16.27
4	7.779	9.488	11.14	13.28	14.86	18.47
5	9.236	11.07	12.83	15.09	16.75	20.52
6	10.64	12.59	14.45	16.81	18.55	22.46
7	12.02	14.07	16.01	18.48	20.28	24.32
8	13.36	15.51	17.53	20.09	21.95	26.12
9	14.68	16.92	19.02	21.67	23.59	27.88
10	15.99	18.31	20.48	23.21	25.19	29.59
11	17.28	19.68	21.92	24.72	26.76	31.26
12	18.55	21.03	23.34	26.22	28.30	32.91
13	19.81	22.36	24.74	27.69	29.82	34.53
14	21.06	23.68	26.12	29.14	31.32	36.12
15	22.31	25.00	27.49	30.58	32.80	37.70
16	23.54	26.30	28.85	32.00	34.27	39.25
17	24.77	27.59	30.19	33.41	35.72	40.79
18	25.99	28.87	31.53	34.81	37.16	42.31
19	27.20	30.14	32.85	36.19	38.58	43.82
20	28.41	31.41	34.17	37.57	40.00	45.31
21	29.62	32.67	35.48	38.93	41.40	46.80
22	30.81	33.92	36.78	40.29	42.80	48.27
23	32.01	35.17	38.08	41.64	44.18	49.73
24	33.20	36.42	39.36	42.98	45.56	51.18
25	34.38	37.65	40.65	44.31	46.93	52.62
26	35.56	38.89	41.92	45.64	48.29	54.05
27	36.74	40.11	43.19	46.96	49.64	55.48
28	37.92	41.34	44.46	48.28	50.99	56.89
29	39.09	42.56	45.72	49.59	52.34	58.30
30	40.26	43.77	46.98	50.89	53.67	59.70
35	46.06	49.80	53.20	57.34	60.27	66.62
40	51.81	55.76	59.34	63.69	66.77	73.40
45	57.51	61.66	65.41	69.96	73.17	80.08
50	63.17	67.50	71.42	76.15	79.49	86.66
60	74.40	79.08	83.30	88.38	91.95	99.61
70	85.53	90.53	95.02	100.4	104.2	112.3
80	96.58	101.9	106.6	112.3	116.3	124.8
90	107.6	113.1	118.1	124.1	128.3	137.2
100	118.5	124.3	129.6	135.8	140.2	149.4

Table C Loi t de Student à ν degrés de liberté t_ν
Quantiles supérieurs $Q_t(1-\alpha;\nu)$

	$1-\alpha$					
ν	0.90	0.95	0.975	0.99	0.995	0.999
1	3.0777	6.3138	12.706	31.821	63.657	318.31
2	1.8856	2.9200	4.3027	6.9646	9.9248	22.327
3	1.6377	2.3534	3.1824	4.5407	5.8409	10.215
4	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732
5	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076
7	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853
8	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247
12	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296
13	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520
14	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874
15	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328
16	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852
35	1.3062	1.6896	2.0301	2.4377	2.7238	3.3400
40	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069
45	1.3006	1.6794	2.0141	2.4121	2.6896	3.2815
50	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614
60	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317
70	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108
80	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953
90	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833
100	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737
120	1.2886	1.6577	1.9799	2.3578	2.6174	3.1595
200	1.2858	1.6525	1.9719	2.3451	2.6006	3.1315
∞	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902

Table D Loi de Snedecor à ν_1 et ν_2 degrés de liberté
Quantiles supérieurs $Q_F(1-\alpha;\nu_1,\nu_2)$ pour $1-\alpha = 0.95$

					ν_1					
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*12*	*24*	∞
1	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91	249.05	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.85	8.74	8.64	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28	3.12	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.85	2.69	2.51	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.77	2.60	2.42	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.37	2.20	2.01	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.31	2.13	1.93	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.29	2.12	1.91	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.28	2.10	1.90	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.10	1.92	1.70	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.02	1.83	1.61	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	1.94	1.75	1.52	1.00

Table D Loi de Snedecor à ν_1 et ν_2 degrés de liberté
Quantiles supérieurs $Q_F(1-\alpha;\nu_1,\nu_2)$ pour $1-\alpha = 0.99$

	ν_1									
ν_2	*1*	*2*	*3*	*4*	*5*	*6*	*8*	*12*	*24*	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.42	99.46	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.49	27.05	26.60	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37	13.93	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89	9.47	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.10	7.72	7.31	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47	6.07	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67	5.28	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11	4.73	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71	4.33	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.74	4.40	4.02	3.61
12	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16	3.78	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.30	3.96	3.59	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.14	3.80	3.43	3.01
15	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67	3.29	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55	3.18	2.76
17	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.46	3.08	2.66
18	8.29	6.01	5.09	4.58	4.25	4.01	3.71	3.37	3.00	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30	2.92	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23	2.86	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17	2.80	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.45	3.12	2.75	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07	2.70	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03	2.66	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.32	2.99	2.62	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96	2.58	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93	2.55	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90	2.52	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87	2.49	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84	2.47	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66	2.29	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50	2.12	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34	1.95	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.51	2.18	1.79	1.00

Table E Coefficient de corrélation r (Test $H_0 : \rho = 0$)
Quantiles supérieurs $r^*(n; 1-\alpha)$

n	0.90	0.95	0.975	0.99	0.995	0.999
3	0.9511	0.9877	0.9969	0.9995	0.9999	1.0000
4	0.8000	0.9000	0.9500	0.9800	0.9900	0.9980
5	0.6870	0.8054	0.8783	0.9343	0.9587	0.9859
6	0.6084	0.7293	0.8114	0.8822	0.9172	0.9633
7	0.5509	0.6694	0.7545	0.8329	0.8745	0.9350
8	0.5067	0.6215	0.7067	0.7887	0.8343	0.9049
9	0.4716	0.5822	0.6664	0.7498	0.7977	0.8751
10	0.4428	0.5494	0.6319	0.7155	0.7646	0.8467
11	0.4187	0.5214	0.6021	0.6851	0.7348	0.8199
12	0.3981	0.4973	0.5760	0.6581	0.7079	0.7950
13	0.3802	0.4762	0.5529	0.6339	0.6835	0.7717
14	0.3646	0.4575	0.5324	0.6120	0.6614	0.7501
15	0.3507	0.4409	0.5140	0.5923	0.6411	0.7301
16	0.3383	0.4259	0.4973	0.5742	0.6226	0.7114
17	0.3271	0.4124	0.4821	0.5577	0.6055	0.6940
18	0.3170	0.4000	0.4683	0.5425	0.5897	0.6777
19	0.3077	0.3887	0.4555	0.5285	0.5751	0.6624
20	0.2992	0.3783	0.4438	0.5155	0.5614	0.6481
21	0.2914	0.3687	0.4329	0.5034	0.5487	0.6346
22	0.2841	0.3598	0.4227	0.4921	0.5368	0.6219
23	0.2774	0.3515	0.4132	0.4815	0.5256	0.6099
24	0.2711	0.3438	0.4044	0.4716	0.5151	0.5986
25	0.2653	0.3365	0.3961	0.4622	0.5052	0.5879
26	0.2598	0.3297	0.3882	0.4534	0.4958	0.5776
27	0.2546	0.3233	0.3809	0.4451	0.4869	0.5679
28	0.2497	0.3172	0.3739	0.4372	0.4785	0.5587
29	0.2451	0.3115	0.3673	0.4297	0.4705	0.5499
30	0.2407	0.3061	0.3610	0.4226	0.4629	0.5415
31	0.2366	0.3009	0.3550	0.4158	0.4556	0.5334
32	0.2327	0.2960	0.3494	0.4093	0.4487	0.5257
33	0.2289	0.2913	0.3440	0.4032	0.4421	0.5184
34	0.2254	0.2869	0.3388	0.3972	0.4357	0.5113
35	0.2220	0.2826	0.3338	0.3916	0.4296	0.5045
36	0.2187	0.2785	0.3291	0.3862	0.4238	0.4979
37	0.2156	0.2746	0.3246	0.3810	0.4182	0.4916
38	0.2126	0.2709	0.3202	0.3760	0.4128	0.4856
39	0.2097	0.2673	0.3160	0.3712	0.4076	0.4797
40	0.2070	0.2638	0.3120	0.3665	0.4026	0.4741
50	0.1843	0.2353	0.2787	0.3281	0.3610	0.4267
60	0.1678	0.2144	0.2542	0.2997	0.3301	0.3912
70	0.1550	0.1982	0.2352	0.2776	0.3060	0.3632
80	0.1448	0.1852	0.2199	0.2597	0.2864	0.3405
90	0.1364	0.1745	0.2072	0.2449	0.2702	0.3215

Table F Test U de Mann-Whitney
Quantiles supérieurs $U_{1-\alpha} = U(n_1, n_2, 1-\alpha)$
Note : $U_\alpha = n_1 n_2 - U_{1-\alpha}$

n_1	n_2	n	0.95	0.975	0.99	0.995
2	*3*	*5*	-	-	-	-
3	*3*	*6*	9	-	-	-
2	*4*	*6*	-	-	-	-
3	*4*	*7*	12	-	-	-
4	*4*	*8*	15	16	-	-
2	*5*	*7*	10	-	-	-
3	*5*	*8*	14	15	-	-
4	*5*	*9*	18	19	20	-
5	*5*	*10*	21	23	24	25
2	*6*	*8*	12	-	-	-
3	*6*	*9*	16	17	-	-
4	*6*	*10*	21	22	23	24
5	*6*	*11*	25	27	28	29
6	*6*	*12*	29	31	33	34
2	*7*	*9*	14	-	-	-
3	*7*	*10*	19	20	21	-
4	*7*	*11*	24	25	27	28
5	*7*	*12*	29	30	32	34
6	*7*	*13*	34	36	38	39
7	*7*	*14*	38	41	43	45
2	*8*	*10*	15	16	-	-
3	*8*	*11*	21	22	24	-
4	*8*	*12*	27	28	30	31
5	*8*	*13*	32	34	36	38
6	*8*	*14*	38	40	42	44
7	*8*	*15*	43	46	49	50
8	*8*	*16*	49	51	55	57

Table F Test U de Mann-Whitney (suite)

n_1	n_2	n	0.95	0.975	0.99	0.995
1	*9*	*10*	-	-	-	-
2	*9*	*11*	17	18	-	-
3	*9*	*12*	23	25	26	27
4	*9*	*13*	30	32	33	35
5	*9*	*14*	36	38	40	42
6	*9*	*15*	42	44	47	49
7	*9*	*16*	48	51	54	56
8	*9*	*17*	54	57	61	63
9	*9*	*18*	60	64	67	70
1	*10*	*11*	-	-	-	-
2	*10*	*12*	19	20	-	-
3	*10*	*13*	26	27	29	30
4	*10*	*14*	33	35	37	38
5	*10*	*15*	39	42	44	46
6	*10*	*16*	46	49	52	54
7	*10*	*17*	53	56	59	61
8	*10*	*18*	60	63	67	69
9	*10*	*19*	66	70	74	77
10	*10*	*20*	73	77	81	84
1	*11*	*12*	-	-	-	-
2	*11*	*13*	21	22	-	-
3	*11*	*14*	28	30	32	33
4	*11*	*15*	36	38	40	42
5	*11*	*16*	43	46	48	50
6	*11*	*17*	50	53	57	59
7	*11*	*18*	58	61	65	67
8	*11*	*19*	65	69	73	75
9	*11*	*20*	72	76	81	83
10	*11*	*21*	79	84	88	92
11	*11*	*22*	87	91	96	100

Table F Test U de Mann-Whitney (suite)

n_1	n_2	n	0.95	0.975	0.99	0.995
1	*12*	*13*	-	-	-	-
2	*12*	*14*	22	23	-	-
3	*12*	*15*	31	32	34	35
4	*12*	*16*	39	41	43	45
5	*12*	*17*	47	49	52	54
6	*12*	*18*	55	58	61	63
7	*12*	*19*	63	66	70	72
8	*12*	*20*	70	74	79	81
9	*12*	*21*	78	82	87	90
10	*12*	*22*	86	91	96	99
11	*12*	*23*	94	99	104	108
12	*12*	*24*	102	107	113	117
1	*13*	*14*	-	-	-	-
2	*13*	*15*	24	25	26	-
3	*13*	*16*	33	35	37	38
4	*13*	*17*	42	44	47	49
5	*13*	*18*	50	53	56	58
6	*13*	*19*	59	62	66	68
7	*13*	*20*	67	71	75	78
8	*13*	*21*	76	80	84	87
9	*13*	*22*	84	89	94	97
10	*13*	*23*	93	97	103	106
11	*13*	*24*	101	106	112	116
12	*13*	*25*	109	115	121	125
13	*13*	*26*	118	124	130	135
1	*14*	*15*	-	-	-	-
2	*14*	*16*	25	27	28	-
3	*14*	*17*	35	37	40	41
4	*14*	*18*	45	47	50	52
5	*14*	*19*	54	57	60	63
6	*14*	*20*	63	67	71	73
7	*14*	*21*	72	76	81	83
8	*14*	*22*	81	86	90	94
9	*14*	*23*	90	95	100	104
10	*14*	*24*	99	104	110	114
11	*14*	*25*	108	114	120	124
12	*14*	*26*	117	123	130	134
13	*14*	*27*	126	132	139	144
14	*14*	*28*	135	141	149	154

Table F Test U de Mann-Whitney (suite)

n_1	n_2	n	0.95	0.975	0.99	0.995
1	*15*	*16*	-	-	-	-
2	*15*	*17*	27	29	30	-
3	*15*	*18*	38	40	42	43
4	*15*	*19*	48	50	53	55
5	*15*	*20*	57	61	64	67
6	*15*	*21*	67	71	75	78
7	*15*	*22*	77	81	86	89
8	*15*	*23*	87	91	96	100
9	*15*	*24*	96	101	107	111
10	*15*	*25*	106	111	117	121
11	*15*	*26*	115	121	128	132
12	*15*	*27*	125	131	138	143
13	*15*	*28*	134	141	148	153
14	*15*	*29*	144	151	159	164
15	*15*	*30*	153	161	169	17
1	*16*	*17*	-	-	-	-
2	*16*	*18*	29	31	32	-
3	*16*	*19*	40	42	45	46
4	*16*	*20*	50	53	57	59
5	*16*	*21*	61	65	68	71
6	*16*	*22*	71	75	80	83
7	*16*	*23*	82	86	91	94
8	*16*	*24*	92	97	102	106
9	*16*	*25*	102	107	113	117
10	*16*	*26*	112	118	124	129
11	*16*	*27*	122	129	135	140
12	*16*	*28*	132	139	146	151
13	*16*	*29*	143	149	157	163
14	*16*	*30*	153	160	168	174
15	*16*	*31*	163	170	179	185
16	*16*	*32*	173	181	190	196

Table F Test U de Mann-Whitney (suite)

n_1	n_2	n	0.95	0.975	0.99	0.995
1	*17*	*18*	-	-	-	-
2	*17*	*19*	31	32	34	-
3	*17*	*20*	42	45	47	49
4	*17*	*21*	53	57	60	62
5	*17*	*22*	65	68	72	75
6	*17*	*23*	76	80	84	87
7	*17*	*24*	86	91	96	100
8	*17*	*25*	97	102	108	112
9	*17*	*26*	108	114	120	124
10	*17*	*27*	119	125	132	136
11	*17*	*28*	130	136	143	148
12	*17*	*29*	140	147	155	160
13	*17*	*30*	151	158	166	172
14	*17*	*31*	161	169	178	184
15	*17*	*32*	172	180	189	195
16	*17*	*33*	183	191	201	207
17	*17*	*34*	193	202	212	219
1	*18*	*19*	-	-	-	-
2	*18*	*20*	32	34	36	-
3	*18*	*21*	45	47	50	52
4	*18*	*22*	56	60	63	66
5	*18*	*23*	68	72	76	79
6	*18*	*24*	80	84	89	92
7	*18*	*25*	91	96	102	105
8	*18*	*26*	103	108	114	118
9	*18*	*27*	114	120	126	131
10	*18*	*28*	125	132	139	143
11	*18*	*29*	137	143	151	156
12	*18*	*30*	148	155	163	169
13	*18*	*31*	159	167	175	181
14	*18*	*32*	170	178	187	194
15	*18*	*33*	182	190	200	206
16	*18*	*34*	193	202	212	218
17	*18*	*35*	204	213	224	231
18	*18*	*36*	215	225	236	243

Table F Test U de Mann-Whitney (suite)

n_1	n_2	n	0.95	0.975	0.99	0.995
1	*19*	*20*	19	-	-	-
2	*19*	*21*	34	36	37	38
3	*19*	*22*	47	50	53	54
4	*19*	*23*	59	63	67	69
5	*19*	*24*	72	76	80	83
6	*19*	*25*	84	89	94	97
7	*19*	*26*	96	101	107	111
8	*19*	*27*	108	114	120	124
9	*19*	*28*	120	126	133	138
10	*19*	*29*	132	138	146	151
11	*19*	*30*	144	151	159	164
12	*19*	*31*	156	163	172	177
13	*19*	*32*	167	175	184	190
14	*19*	*33*	179	188	197	203
15	*19*	*34*	191	200	210	216
16	*19*	*35*	203	212	222	230
17	*19*	*36*	214	224	235	242
18	*19*	*37*	226	236	248	255
19	*19*	*38*	238	248	260	268
1	*20*	*21*	20	-	-	-
2	*20*	*22*	36	38	39	40
3	*20*	*23*	49	52	55	57
4	*20*	*24*	62	66	70	72
5	*20*	*25*	75	80	84	87
6	*20*	*26*	88	93	98	102
7	*20*	*27*	101	106	112	116
8	*20*	*28*	113	119	126	130
9	*20*	*29*	126	132	140	144
10	*20*	*30*	138	145	153	158
11	*20*	*31*	151	158	167	172
12	*20*	*32*	163	171	180	186
13	*20*	*33*	176	184	193	200
14	*20*	*34*	188	197	207	213
15	*20*	*35*	200	210	220	227
16	*20*	*36*	213	222	233	241
17	*20*	*37*	225	235	247	254
18	*20*	*38*	237	248	260	268
19	*20*	*39*	250	261	273	281
20	*20*	*40*	262	273	286	295

Table G Test V des rangs signés de Wilcoxon.
Quantiles inférieurs $V_\alpha = V(n;\alpha)$

n	0.05	0.025	0.01	0.005
5	1	–	–	–
6	2	1	–	–
7	4	2	0	–
8	6	4	2	0
9	8	6	3	2
10	11	8	5	3
11	14	11	7	5
12	17	14	10	7
13	21	17	13	10
14	26	21	16	13
15	30	25	20	16
16	36	30	24	19
17	41	35	28	23
18	47	40	33	28
19	54	46	38	32
20	60	52	43	37
21	68	59	49	43
22	75	66	56	49
23	83	73	62	55
24	92	81	69	61
25	101	90	77	68
26	110	98	85	76
27	120	107	93	84
28	130	117	102	92
29	141	127	111	100
30	152	137	120	109
31	163	148	130	118
32	175	159	141	128
33	188	171	151	138
34	201	183	162	149
35	214	195	174	160
36	228	208	186	171
37	242	222	198	183
38	256	238	211	195
39	271	250	224	208
40	287	264	238	221
45	371	344	313	292
50	466	434	398	373

Table des matières