1 Quelques noms de la statistique

Pearson, Karl (1857-1936), est un mathématicien et philosophe britannique. Il a mis au point les principales techniques statistiques modernes et les a surtout appliquées aux questions de l'hérédité.

Il est né à Londres et est diplômé de l'université de Cambridge en 1879. il effectue ensuite de brèves études de droit, mais, en réalité, consacrera par la suite toute sa carrière à l'enseignement des mathématiques, des mathématiques appliquées et de la science à l'University College de Londres.

Au début du XX^{e} siècle, Pearson s'intéresse aux travaux du Britannique sir Francis Galton sur la transmission des caractéristiques biologiques entre générations. Par ses recherches, Pearson pose les bases de la théorie statistique moderne, définissant les notions de corrélation, d'analyse de régression et d'écart-type. Promoteur du darwinisme social, il vise par ses recherches à l'amélioration de l'espèce humaine. Auteur prolifique, il a écrit notamment « the Grammar of Science » en 1892, livre sur les méthodes scientifiques.

Gosset William Sealy (1876-1937) est connu sous le pseudonyme « student » : avec Pearson ils ont défini l'écart-type, paramètre essentiel en théorie de la mesure ; il est utilisé comme référence standard et échelle de mesure dans les tests d'hypothèse et dans la définition des intervalles de confiance. Il a aussi développé des méthodes d'analyse statistique adaptées aux petits échantillons et en particulier des tests basés sur la distribution t. Il distinguait l'écart-type, noté S, évalué sur un échantillon prélevé dans la population considérée, et l'écart-type σ associé à la population entière.

Spearman, Charles (1863-1945), est un psychologue anglais, connu pour ses travaux sur l'intelligence. Dès 1907, date de ses débuts dans la carrière universitaire, il s'intéresse vivement aux méthodes de mesure de l'intelligence humaine au moyen de tests non verbaux. Utilisant expérimentalement la méthode des corrélations mise au point par ses prédécesseurs dans la psychologie quantitative comme Thomson, Burt, Galton et Pearson, il constate l'existence de corrélations positives entre plusieurs variables et exploite la signification de ces corrélations. Par la suite, Spearman affinera sa méthode d'analyse factorielle en y

ajoutant les facteurs de groupe. Ses travaux ont contribué à asseoir la validité des tests psychologiques sur les modèles théoriques qu'il a mis au point.

Laplace, Pierre Simon, marquis de (1749-1827), astronome, mathématicien et physicien français qui émit l'hypothèse de la « nébuleuse primitive » sur l'origine du Système solaire. Pierre Simon, marquis de Laplace est né en Normandie, où il fit ses études. En 1767, il devint professeur de mathématiques à l'École royale militaire et, en 1783, il fut élu membre de l'Académie des sciences. Les travaux scientifiques majeures de Laplace concernent la mécanique céleste et le calcul des probabilités. Dans sa Théorie analytique des probabilités (1812), qui contient des calculs très élaborés d'approximation de grands nombres, Laplace indiqua les principes et les applications de la géométrie du hasard. Il fut à l'origine de la loi de Laplace-Gauss, ou loi normale, très utilisée en probabilités. Il donna la fameuse équation de Laplace, équation différentielle de forme $\Delta f = 0$, où f est une fonction deux fois dérivable et Δ l'opérateur appelé laplacien. Il fit de nombreux travaux sur la chaleur spécifique, la capillarité et l'électromagnétisme (loi de Laplace).

Gauss Carl Friedrich est un mathématicien, astronome et physicien allemand né à Brunswick (1777-1855). Il sera surnommé le "Prince des mathématiciens". Il a 19 ans lorsqu'il découvre la construction à la règle et au compas d'un polygone à 17 côtés. C'est le seul pas important accompli dans la théorie euclidienne des polygones réguliers depuis deux mille ans. Ce théorème ne sera communiqué au public qu'en 1801 dans son traité d'arithmétique *Disquisitones arithmeticae* dont il a achevé la rédaction en 1798 alors qu'il est toujours étudiant à l'université de Göttingen. En 1799, il soutient une thèse devant l'université d'Helmstedt sur le théorème fondamental de l'algèbre : *Tout polynôme entier sur le corps des nombres complexes a au moins un zéro*. Son nom reste essentiellement attaché au domaine des probabilités avec la loi de Laplace-Gauss et sa courbe de répartition en cloche dite « loi normale ».

Gauss s'est aussi intéressé au magnétisme et il a imaginé le magnétomètre, dont il formule la théorie mathématique dans son ouvrage *Théorie générale du magnétisme terrestre* (1839). C'est pourquoi l'unité d'induction magnétique porte aujourd'hui son nom.

Kolmogorov Andreï (Moscou 1903-1987) ; fils d'un agronome statisticien, Andreï Kolmogorov est certainement un des mathématiciens les plus importants du

XX^{e} siècle. Son nom est associé à la Théorie des Probabilités, à la Théorie des Systèmes dynamiques, à la Théorie de l'Information et à la Topologie. Il a formalisé la théorie des *probabilités* dans l'article « *Grundbegriffe der Warscheinligkeitsrechnung » (*Les fondements du calcul des probabilités) en 1933. Il développe notablement cette théorie à propos des processus de **Markov** (nommés d'après Andreï Markov, 1856-1922).

Il travaille dans de nombreux domaines : la *topologie*, ou il développe la théorie de l'homologie et de la cohomologie, la *théorie des systèmes dynamiques* où il propose avec Arnol et Moser la théorie KAM, la théorie de *l'Information* où il développe le concept de "suite n'obéissant pas à une loi", en utilisant le concept d'algorithme, ou plutôt d'absence d'algorithme. Il fait aussi intervenir le concept d'*entropie* dans ce domaine de recherche.

Il a résolu en partie le Sixième problème de Hilbert (l'axiomatisation des Probabilités) mais aussi totalement le treizième problème.

2 Glossaire

Les termes contenus dans ce glossaire sont ceux nécessaires à la compréhension du cours qui va suivre

Ajustement *:* au moment de l'analyse d'un essai, subdivision de la population selon des critères pronostiques pour augmenter la puissance du test. Il permet d'étudier la liaison existant entre deux variables en tenant compte d'une troisième.

Aléatoire *:* caractère ou paramètre dont la détermination n'appartient pas à l'expérimentateur (voir contrôlé).

Appariement **:** Technique permettant de rendre comparables deux ou plusieurs groupes, en particulier par rapport à certains facteurs de confusion déjà connus dont on veut neutraliser les effets.

Biais *:* erreur systématique. Plus communément, on dit qu'une étude est "biaisée" lorsque les groupes ne sont pas comparables au moment de l'analyse (à cause d'un nombre trop important de perdus de vue, par exemple).

Biais de classement **:** Biais dans la mesure du facteur de risque ou dans la certitude de la maladie. Cette erreur est quasi inévitable puisqu'aucun outil de mesure (interrogatoire, examen, test) n'est parfait. *Exemple : un comportement à risque minimisé par le malade, ou simplement non recherché dans le questionnaire.*

Biais de confusion **:** Biais provoqué par un facteur de confusion interagissant avec le facteur de risque étudié dans l'étude du lien entre ce facteur et la maladie.

Biais de sélection : Biais dans la constitution de l'échantillon, qui va se retrouver non représentatif de la population générale pour des facteurs liés au problème étudié (d'où le biais).

Biais de mémorisation : Type de biais de classement lorsque l'information sur l'exposition a été obtenue a posteriori après que le diagnostic des cas a été établi (cas-témoin).

Bilatéral : (test) test dont la zone de rejet de l'hypothèse nulle est située à l'extérieur et de part et autre d'un intervalle. On cherche à répondre à la question A est-il différent de B.

Contrôlé *:* caractère ou paramètre dont la détermination appartient à l'expérimentateur (voir aléatoire).

Corrélation *:* liaison entre deux variables quantitatives lorsqu'elles sont toutes deux aléatoires (voir régression).

Degré **de signification** : c'est la probabilité que la différence observée entre deux groupes de sujets soit due aux seules fluctuations d'échantillonnage, c'est-à-dire au hasard (voir hasard et seuil de signification).

DDL : abréviation de nombre de degrés de liberté ce nombre est associé à la significativité d'un test.

Ecart type *:* Pour une variable qualitative s'exprimant par un pourcentage p, sur n sujets, c'est : $\sqrt{\frac{p \times q}{n}}$ (où q = 1 - p) et pour une variable quantitative s'exprimant par une moyenne m, sur n sujets, c'est : $\sqrt{\sum (x-m)^2/n-1}$ (où x est la valeur de chaque observation). Voir déviation standard et intervalle de confiance.

Echantillon *:* c'est la fraction d'une population; si elle est déterminée par tirage au sort, c'est un sondage. Son étude permet d'en déduire des renseignements sur la population (voir estimation et sondage).

Erreur **(de 1ère espèce ou risque α)** : lors d'un test, conclure à tort à une différence entre deux populations ;

Erreur (de 2ème espèce ou risque β) : lors d'un test conclure à tort à l'égalité de deux populations (voir hypothèse et puissance d'un test).

Estimation *:* à partir d'un échantillon, permet d'émettre un jugement sur la valeur d'une donnée de la population d'origine en y associant un intervalle de confiance dans lequel on a une certaine garantie de trouver cette valeur.

Etendue : pour une série de données c'est l'écart qui sépare la plus petite de la plus grande valeur.

Facteur de confusion : on dit qu'il y a confusion lorsque l'on teste une liaison entre 2 variables et que l'on suspecte que la nature de cette liaison puisse être différente à l'intérieur des sous groupes d'une des 2 variables, les sous groupes étant définis par un 3e facteur. Ce 3e facteur est appelé facteur d'ajustement ou *facteur de confusion*.

Fréquence *:* (d'un événement) : pourcentage de fois où un événement est survenu. Par extension, c'est la probabilité de survenue d'un événement (voir incidence et prévalence).

Hasard *:* survenue aléatoire d'un événement, mais le hasard peut être contrôlé par tirage au sort (voir degré de signification).

Hypothèse (nulle) : accepter l'égalité de deux populations (c'est l'hypothèse à tester) ; (alternative) : rejeter l'égalité de deux populations et accepter une différence. Voir erreur.

Incidence *:* fréquence ou probabilité de survenue d'un événement dans une population par unité de temps, un an par exemple (voir fréquence et prévalence).

Intervalle **de confiance** : intervalle dans lequel on peut situer avec une certaine garantie un pourcentage ou une moyenne inconnue avec un risque d'erreur connu. L'intervalle de confiance à 95 % est : ± 1,96 × écart-type (voir écart-type et estimation).

Interaction : On dit qu'il y a interaction si le facteur étudié a des effets significativement différents sur le critère de jugement en fonction des sous groupes définis selon le facteur d'ajustement.

Multivariée : méthode utilisée pour étudier simultanément les liaisons existant entre plus de 2 variables, il faut recourir aux études multidimensionnelles.

***Normale* (distribution)** : Dispersion particulière des valeurs d'une variable autour de la moyenne suivant une loi dite "normale" ou loi de Laplace Gauss. Ici, normal n'est pas le contraire d'anormal.

Odds ratio ou rapport des cotes : la cote est, dans un groupe ou échantillon, le rapport du nombre de patients ayant présenté un événement sur le nombre de patients n'ayant pas présenté ce même événement.

Population : ensemble des individus porteurs du caractère étudié.

Prévalence *:* fréquence ou probabilité de survenue d'un événement dans une population à une date donnée (voir fréquence et incidence).

***Prospective* (étude)** : les attendus de l'étude sont définis avant la collection des données.

***Puissance* (d'un test)** : test d'autant plus puissant que le risque de conclure à tort à l'absence de différence entre deux populations est faible : puissance = 1 - β.

***Qualitative* (variable)** : dont les valeurs sont discrètes, peuvent se ranger en classes. La plus simple est la variable dichotomique (présence ou absence d'un caractère).

***Quantitative* (variable)** : dont les valeurs sont continues et mesurables.

Randomiser : tirer au sort pour permettre une répartition au hasard, aléatoire, des sujets dans deux ou plusieurs groupes

Rapport de vraisemblance : Le rapport de vraisemblance d'un résultat X d'un test à la recherche d'une maladie M est défini par le rapport de la probabilité d'un résultat X parmi les malades (M) par la probabilité du même résultat X parmi les patients non malades (m).

Régression *:* liaison entre deux variables quantitatives lorsque l'une d'elles est contrôlée (voir corrélation).

***Rétrospective* (étude)** : les attendus de l'étude sont définis après la collection des données.

Risque relatif : c'est le rapport de la probabilité de maladie chez les sujets exposés au facteur étudié (groupe test+) à la probabilité de maladie chez les sujets non exposés (groupe test-). Il mesure une association entre ce facteur et la maladie que cette association soit de nature causale ou non.

Sensibilité *:* rapport du nombre de malades qui présentent un signe à l'ensemble des malades. C'est la fréquence du signe dans la maladie.

***Seuil* (de signification)** : limite supérieure du degré de signification pour admettre une différence significative. Habituellement : 0,05 (voir degré de signification et significative).

***Significative* (différence statistiquement)** : la probabilité que la différence observée soit due aux simples fluctuations d'échantillonnage (degré de signification) est inférieure au seuil de signification (voir degré de signification).

Sondage *:* tirage au sort d'un échantillon d'une population pour en induire des renseignements sur la population (voir échantillon et estimation).

Spécificité *:* rapport du nombre de sujets qui ne présentent pas le signe et indemnes de la maladie à l'ensemble des sujets indemnes. C'est la rareté du signe chez les non malades.

Stratification *:* au moment du tirage au sort dans un essai thérapeutique, subdivision de la population selon des critères pronostiques pour augmenter la puissance du test statistique (voir ajustement).

Unilatéral *: (test)* test dont la zone de rejet de l'hypothèse nulle est située d'un seul coté d'un intervalle. On cherche à répondre à la question A est > à B ou A est < B.

Univariée : méthode utilisée pour analyser la liaison entre 2 variables.

Valeur prédictive **négative** : rapport du nombre de sujets indemnes de la maladie et qui ne présentent pas le signe à l'ensemble des sujets qui ne présentent pas le signe. C'est l'estimation de la probabilité de ne pas avoir la maladie si on n'a pas le signe.

Valeur prédictive **positive** : rapport du nombre de malades qui présentent le signe à l'ensemble des sujets qui présentent le signe. C'est l'estimation de la probabilité d'avoir la maladie si on a le signe.

Variance *:* c'est le carré de l'écart-type.

3 Exercice

3.1 Facteurs de risque et pathologie vasculaire

Certains paramètres biologiques comme l'augmentation du cholestérol ou des triglycérides, comme la C. R. P. (protéine C réactive), certaines situations cliniques comme la surcharge pondérale et certaines habitudes comme la consommation de tabac sont considérés comme des facteurs de risque de pathologie vasculaire artérielle, notamment coronarienne. L'OMS admet que les concentrations de cholestérol ou des triglycérides sont satisfaisantes si elles sont inférieures à 2 g/l et que l'apoprotéine B (l'apoprotéine B est une protéine qui participe au transport du cholestérol) est normale si elle est inférieure à 1,30 g/l.

Dans le cadre d'un projet de recherche, des médecins souhaitent redéfinir la place de certains des paramètres du bilan lipidique (cholestérolémie, triglycéridémie, Apoprotéine B) dans différents groupes de population. La deuxième partie de leur travail consiste à évaluer les effets de plusieurs stratégies de traitements, dont le régime alimentaire et un traitement médicamenteux ou leur association sur ces paramètres. Ces stratégies sont considérées comme efficaces si l'on observe une diminution des concentrations des différents paramètres lipidiques et du poids. Dans le suivi d'efficacité, il est aussi pris en compte l'apparition de complications. Par ailleurs, à la fin de la séquence de traitement, le cholestérol a été dosé par 2 méthodes de dosage appelées méthodes I et II.

Un groupe de 66 volontaires (32 femmes et 34 hommes), âgés de 18 à 89 ans est réuni par tirage au sort. Dans ce groupe 4 classes d'âge ont été définies : moins de trente ans, de trente à 49 ans, de 50 à 65 ans et soixante-cinq ans ou plus. Pour chaque participant il est aussi noté s'il y a ou non consommation de tabac. Chacun va recevoir le traitement associé ou non au régime.

Avant le début de l'essai une évaluation clinique a permis de classer ces sujets en 2 groupes : à risque de complications en particulier cardiaque **R+** (n = 30) et pas de risque reconnu **R-** (n = 36).

Avant de répondre aux questions qui seront posées, il peut être intéressant de réaliser la description des variables qualitatives et quantitatives (catégorisation des groupes, nombre d'individus par groupe [à vous de définir les groupes], moyenne et écart-type pour les variables quantitatives).