Université PARIS-VI Pierre et Marie Curie
Faculté de Médecine Pitié-Salpêtrière

Statistiques

PCEM1

2001 - 2002

J.F. BOISVIEUX
J.L. GOLMARD
A. MALLET
V. MORICE

Mise à jour : 15 janvier 2002
Relecture : V. Morice et S. Tezenas

Sommaire

Introduction

Les statistiques constituent, en médecine, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence au médecin :

1. Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?
2. Quelle est la fiabilité d'un examen complémentaire ?
3. Quel est le risque de complication d'un état pathologique, et quel est le risque d'un traitement ?
4. Le traitement A est-il plus efficace que le traitement B ?

1 La variabilité et l'incertain

Toutes ces questions, proprement médicales, reflètent une propriété fondamentale des systèmes biologiques qui est leur variabilité. Cette variabilité est la somme d'une variabilité expérimentale (liée au protocole de mesure) et d'une variabilité proprement biologique. On peut ainsi décomposer la variabilité d'une grandeur mesurée en deux grandes composantes :

variabilité totale = variabilité biologique + variabilité métrologique

- La variabilité biologique peut être elle-même décomposée en deux termes : d'une part la variabilité intra-individuelle, qui fait que la même grandeur mesurée chez un sujet donné peut être soumise à des variations aléatoires ; et d'autre part la variabilité inter-individuelle qui fait que cette même grandeur varie d'un individu à l'autre.

  variabilité biologique = variabilité intra-individuelle + variabilité inter-individuelle

  La variabilité intra-individuelle peut être observée lors de la mesure de la performance d'un athlète qui n'est pas capable des mêmes performances à chaque essai, mais qui se différencie des autres athlètes (variabilité inter-individuelle). En général, la variabilité intra est moindre que la variabilité inter.
- La variabilité métrologique peut être elle aussi décomposée en deux termes : d'une part les conditions expérimentales dont les variations entraînent un facteur d'aléas ; et d'autre part les erreurs induites par l'appareil de mesure utilisé.

  variabilité métrologique = variabilité expérimentale + variabilité appareil de mesure

  La mesure de la pression artérielle peut grandement varier sur un individu donné suivant les conditions de cette mesure ; il est ainsi recommandé de la mesurer après un repos d'au moins 15 minutes, allongé, en mettant le patient dans des conditions de calme maximal. Cette recommandation vise à minimiser la variabilité due aux conditions expérimentales. La précision de l'appareil de mesure est une donnée intrinsèque de l'appareil, et est donnée par le constructeur.

2 La décision dans l'incertain

Pour prendre une décision diagnostique ou thérapeutique le médecin doit avoir des éléments lui permettant de prendre en compte cette variabilité naturelle, pour distinguer ce qui est normal de ce qui est pathologique (décision à propos d'un patient) et pour évaluer la qualité d'un nouvel examen, ou d'une nouvelle thérapeutique (décision thérapeutique). La compréhension des méthodes statistiques, de leur puissance et de leurs limites, est essentielle pour un médecin de nos jours. Tout résultat de recherche médicale résulte d'une expérimentation (clinique ou biologique) qui s'appuie sur une méthodologie statistique rigoureuse, et dont les résultats sont analysés en termes statistiques.
De même la démarche statistique permet d'évaluer les risques (ou les bénéfices) d'une prescription, de déterminer dans une situation donnée l'examen qui apportera la meilleure information diagnostique.
Nous voyons donc l'importance de la maîtrise de l'outil et de la démarche statistique :

- Pour permettre les progrès de la connaissance médicale : c'est le domaine de la recherche clinique qui ne peut s'accomplir convenablement (définition de la question, mise en place du protocole expérimental, analyse des résultats) qu'en suivant une méthodologie statistique rigoureuse.
- Pour mieux connaître l'état de santé d'une population, la fréquence et la gravité d'une épidémie (penser au SIDA), etc. Cette connaissance se fera à partir d'échantillons convenablement choisis et de calculs basés sur les outils de la statistique. Il sera alors possible de rechercher les stratégies de prévention les mieux adaptées, d'en évaluer leur impact. Il s'agit là des applications relevant de l'épidémiologie et de la santé publique.
- Pour améliorer la pratique médicale dans ses aspects décisionnels, à savoir choisir le meilleur examen (clinique ou para-clinique) pour aboutir le plus rapidement et le plus sûrement au diagnostic. Pour optimiser la thérapeutique, choisir le traitement le mieux adapté à un patient donné (choix du médicament, posologie, etc).

L'objectif de ce cours est de vous fournir les bases indispensables permettant de comprendre les méthodes utilisées, d'interpréter correctement les résultats de nouvelles recherches, et d'adopter un mode de raisonnement qui soit à même d'optimiser la décision dans l'exercice de la médecine.
Plus précisément nous étudierons successivement :

1. Les bases de calcul de probabilités, qui sont indispensables à la compréhension et à l'utilisation des méthodes statistiques.
2. La statistique descriptive qui permet de représenter et de quantifier la variabilité d'une ou plusieurs grandeurs observées.
3. La statistique inductive qui inclura les tests statistiques permettant de retenir une hypothèse A plutôt qu'une hypothèse B à partir de données expérimentales (comme dans le cas de la comparaison de deux traitements, où l'hypothèse A est que les deux traitements sont équivalents et l'hypothèse B est qu'ils sont différents).
4. Les applications des méthodes statistiques à l'épidémiologie, à l'aide à la décision thérapeutique et diagnostique, et les applications aux essais thérapeutiques.

Chapitre 1

Statistique(s) et Probabilité(s)

Nous commencerons par définir les termes et les concepts importants.

1.1 Statistique

Le terme statistique désigne à la fois un ensemble de données d'observations, et l'activité qui consiste en leur recueil, leur traitement et leur interprétation. Les termes *statistique*, ou *statistique**s*** (au pluriel) englobent ainsi plusieurs notions distinctes :

1. D'une part le recensement de grandeurs d'intérêt comme le nombre d'habitants d'un pays, le revenu moyen par habitant, le nombre de séropositifs dans la population française. Nous voyons que la notion fondamentale qui se dégage de cette énumération est celle de *Population*. Une population est un ensemble d'objets, d'êtres vivants ou d'objets abstraits (ensemble des mains de 5 cartes distribuées au bridge...) de même nature.
2. La statistique en tant que science s'intéresse aux propriétés des populations naturelles. Plus précisément elle traite de nombres obtenus en comptant ou en mesurant les propriétés d'une population. Cette population d'objets doit en outre être soumise à une variabilité, qui est due à de très nombreux facteurs inconnus (pour les populations d'objets biologiques qui nous intéressent ces facteurs sont les facteurs génétiques et les facteurs environnementaux).
3. A ces deux acceptions du terme statistiques (au pluriel) il faut ajouter le terme statistique (au singulier) qui définit toute grandeur calculée à partir d'observations. Ce peut être la plus grande valeur de la série statistique d'intérêt, la différence entre la plus grande et la plus petite, la valeur de la moyenne arithmétique de ces valeurs, etc.

1.2 Population et échantillon

On appelle *population* P un ensemble généralement très grand, voire infini, d'individus ou d'objets de même nature. Tous les médecins de France constituent une population, de même que l'ensemble des résultats possibles du tirage du loto. Une population peut donc être réelle ou fictive.
Il est le plus souvent impossible, ou trop coûteux, d'étudier l'ensemble des individus constituant une population ; on travaille alors sur une partie de la population que l'on appelle *échantillon*. Cet échantillon, s'il est convenablement sélectionné, permettra l'étude de la variabilité des caractéris-

tiques d'intérêt de la population. On dira qu'on a extrait un *échantillon représentatif*. Si par exemple on souhaite déterminer les caractéristiques « moyennes » du poids et de la taille des prématurés masculins on tirera au hasard[1] un certain nombre de sujets parmi les naissances de prématurés de l'année.
Chaque individu, ou unité statistique, appartenant à une population est décrit par un ensemble de caractéristiques appelées *variables* ou *caractères*. Ces variables peuvent être quantitatives (numériques) ou qualitatives (non numériques) :

quantitatives
: pouvant être classées en variables continues (taille, poids) ou discrètes (nombre d'enfants dans une famille)

qualitatives
: pouvant être classées en variables catégorielles (couleurs des yeux) ou ordinales (intensité d'une douleur classée en nulle, faible, moyenne, importante).

1.3 Statistique et probabilité

La théorie (ou le calcul) des probabilités est une branche des mathématiques qui permet de modéliser les phénomènes où le hasard intervient (initialement développée à propos des jeux de hasard, puis progressivement étendue à l'ensemble des sciences expérimentales, dont la physique et la biologie).
Cette théorie permet de construire des modèles de ces phénomènes et permet le calcul : c'est à partir d'un modèle probabiliste d'un jeu de hasard comme le jeu de dés que l'on peut prédire les fréquences d'apparition d'événements comme le nombre de fois que l'on obtient une valeur paire en jetant un dé un grand nombre de fois. Les éléments de calcul des probabilités indispensables à la compréhension des statistiques seront traités dans la première partie du cours.
Sous jacente à la notion de statistiques se trouve la notion de Population dont on souhaite connaître les propriétés (plus précisément les régularités), permettant en particulier de savoir si deux populations sont identiques ou non. Ce cas est celui du cadre des essais thérapeutiques, où l'on considère 2 populations (patients traités avec le médicament A ou avec le médicament B) dont on souhaite savoir si elles diffèrent ou non (c'est le cas le plus simple des essais cliniques). Pour ce faire il est nécessaire de modéliser les populations, en utilisant des modèles probabilistes. Un modèle de ce type est par exemple de considérer que la taille des individus suit une distribution gaussienne. A partir de ce modèle on peut calculer les propriétés d'échantillons ; c'est ce qu'on appelle une déduction qui va du modèle vers l'expérience. A l'inverse considérant un échantillon d'une population on peut essayer de reconstruire le modèle de la population.
Cette démarche est calquée sur la démarche scientifique habituelle. Le scientifique est capable, en utilisant les mathématiques, de prédire le comportement d'un modèle donné (c'est par exemple une « loi » de la physique) : c'est la démarche déductive. A l'inverse, observant des faits expérimen-

1. Nous reviendrons sur cette méthode permettant d'obtenir un échantillon représentatif de la population étudiée. Cela consiste en gros à sélectionner les individus sur la base d'un tirage analogue à celui qui consiste à tirer des noms dans une urne qui contiendrait tous les noms possibles.

taux il va tenter de dégager des propriétés générales du phénomène observé qu'il va en général représenter sous forme d'un modèle (toutes les lois de la physique et de la chimie sont des modèles mathématiques les plus généraux possibles des faits expérimentaux) : c'est la construction inductive de la théorie. Cette démarche générale va plus loin car le modèle permet de prédire des expériences non réalisées. Si les prédictions ainsi réalisées sont contradictoires avec les résultats expérimentaux alors on pourra avec certitude réfuter le modèle (on dit aussi qu'on l'a falsifié) ; dans le cas contraire on garde le modèle mais on n'est pas certain qu'il soit « vrai ». Autrement dit, à l'issue d'un tel test on ne peut avoir de certitude que si on a trouvé des éléments permettant de réfuter le modèle. Nous verrons dans la suite que cette approche se transpose exactement dans la démarche statistique, en particulier dans le domaine des tests.

Chapitre 2

Rappels mathématiques

2.1 Ensembles, Eléments

On appelle ensemble, toute liste ou collection d'objets bien définis, explicitement ou implicitement ; on appelle éléments ou membres de l'ensemble les objets appartenant à l'ensemble et on note :

- $p \in A$ si p est un élément de l'ensemble A
- B est partie de A, ou sous ensemble de A, et l'on note $B \subset A$ ou $A \supset B$, si $x \in B \Rightarrow x \in A$

On définit un ensemble soit en listant ses éléments, soit en donnant la définition de ses éléments :

- $A = \{1, 2, 3\}$
- $X = \{x : x \text{ est un entier positif}\}$

Notations :

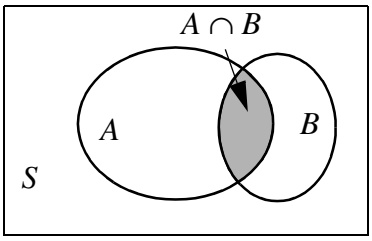
- la négation de $x \in A$ est $x \notin A$
- $\varnothing$ est l'ensemble vide
- S est l'ensemble universel.

2.2 Opérations sur les ensembles, diagrammes de Venn

Soient A et B deux ensembles quelconques.

Intersection

L'intersection de A et B, notée $A \cap B$, est l'ensemble des éléments x tels que $x \in A$ et $x \in B$. Soit :
$A \cap B = \{ x : x \in A \text{ et } x \in B \}$
Le terme « et » est employé au sens $x \in A$ et B si x appartient à la fois à A et à B



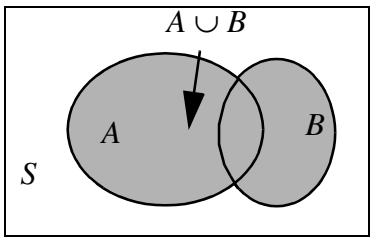
Cas particulier : si $A \cap B = \varnothing$, on dit que A et B sont **disjoints**.

Réunion

La réunion de A et B, notée $A \cup B$, est l'ensemble des éléments x tels que $x \in A$ ou $x \in B$. Soit :

$A \cup B = \{ x : x \in A \text{ ou } x \in B \}$

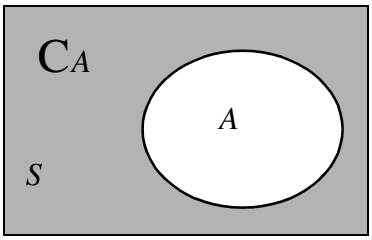
Le terme « ou » est employé au sens $x \in A$ ou B si x appartient à A, ou à B, ou à A et B (car $x \in A$ et B signifie $x \in A$ et $x \in B$).



Complémentaire

Le complémentaire de A est l'ensemble des éléments qui n'appartiennent pas à A.

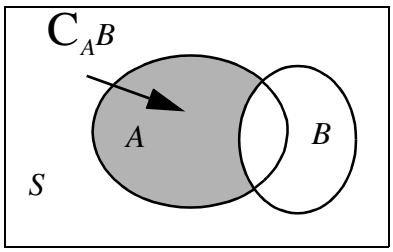
$\complement A = \{x : x \notin A\}$



Différence

La différence entre A et B, ou complémentaire de B relatif à A, est l'ensemble des éléments de A qui n'appartiennent pas à B.

$A - B = \complement_A B = \{x : x \notin B \text{ et } x \in A\}$



Algèbre des ensembles

$A \cup A = A$	$A \cap A = A$
$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
$A \cup B = B \cup A$	$A \cap B = B \cap A$
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
$A \cup \varnothing = A$	$A \cap S = A$
$A \cup S = S$	$A \cap \varnothing = \varnothing$
$A \cup \mathrm{C}A = S$	$A \cap \mathrm{C}A = \varnothing$
$\mathrm{C}\mathrm{C}A = A$	$\mathrm{C}S = \varnothing, \mathrm{C}\varnothing = S$
$\mathrm{C}(A \cup B) = \mathrm{C}A \cap \mathrm{C}B$	$\mathrm{C}(A \cap B) = \mathrm{C}A \cup \mathrm{C}B$

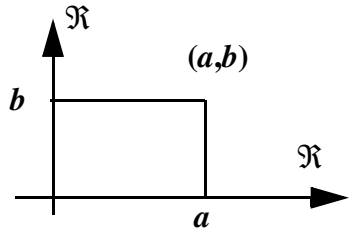
2.3 Ensembles finis, dénombrables, non dénombrables

- Un ensemble est **fini** s'il est vide ($\varnothing$) ou s'il contient un nombre fini d'éléments ; sinon, il est infini :
  $A = \{a_1, a_2, a_3\}$ est fini ;
  $I = \{x \in [0,1]\}$ est infini.
- Un ensemble infini est dit **dénombrable** si on peut faire correspondre de façon unique chaque élément de l'ensemble à un entier naturel et un seul :
  $A = \{n : n$ est un entier pair$\}$ est infini dénombrable.
- Un ensemble infini est **non dénombrable** dans le cas contraire. Dans la pratique, les seuls ensembles infinis non dénombrables que nous rencontrerons seront des intervalles de $\Re$: $\{x \in [a,b]\}$ ou des intervalles de $\Re^2$: $\{(x, y) : x \in [a,b], y \in [c,d]\}$.

2.4 Ensembles produits

Soient A et B deux ensembles ; l'ensemble produit de A et de B, noté $A \times B$, est l'ensemble de tous les couples ordonnés (a, b), avec $a \in A$ et $b \in B$.
Exemples :

- $A = \{a, b, c\}$; $B = \{1, 2\}$
  $A \times B = \{ (a, 1), (a, 2), (b, 1), (b, 2), (c, 1), (c, 2) \}$
- $\Re \times \Re$ est le plan cartésien, chaque élément de $\Re \times \Re$ étant défini par son abscisse et son ordonnée :



2.5 Familles d'ensembles

Les éléments d'un ensemble peuvent eux-mêmes être des ensembles. On dit alors que ces ensembles font partie de la même classe ou de la même famille.

Parties

Soit un ensemble A quelconque. On appelle famille des parties de A l'ensemble des sous-ensembles de A.
Exemple : $A = \{1, 2\}$
$\mathcal{P}(A) = \{\varnothing, \{1\}, \{2\}, \{1,2\}\}$

Partition

Une partition d'un ensemble A est une subdivision de A en sous-ensembles disjoints dont la réunion forme A.

Notation

Soit une famille d'ensembles $\{A_i\} = \{A_1, A_2, ..., A_n, ...\}$ qui peut être finie ou non. On note :

$$\bigcup_i A_i = A_1 \cup A_2 \cup ... \cup A_n \cup ...$$

$$\bigcap_i A_i = A_1 \cap A_2 \cap ... \cap A_n \cap ...$$

2.6 Autres rappels mathématiques

2.6.1 Rappel sur les sommes

Soit $\{a_i\}$ une suite de termes a_i. On note $\sum_{i=1}^{n} a_i = a_1 + a_2 + ... + a_n$.
Propriétés :

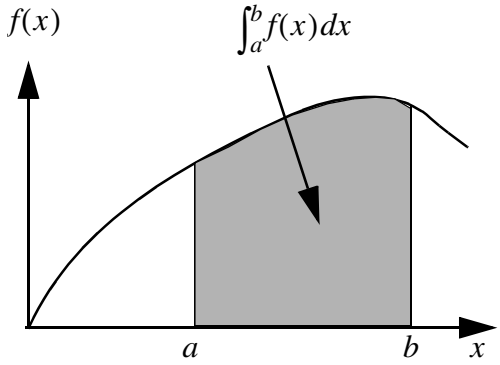
1. $\sum_i (a_i + b_i) = \sum_i a_i + \sum_i b_i$
2. $\sum_i (k a_i) = k \sum_i a_i$

   Si k est une constante (indépendante de i), elle peut être sortie de la somme.

2.6.2 Rappel sur les intégrales

Définition

Soit f une fonction réelle. L'intégrale définie de cette fonction sur l'intervalle $[a,b]$ est l'aire sous la courbe de f sur l'intervalle $[a,b]$.
Elle est notée $\int_a^b f(x)dx$.



Propriétés

1. $\int_a^b (f(x) + g(x))dx = \int_a^b f(x)dx + \int_a^b g(x)dx$
2. $\int_a^b k f(x)dx = k \int_a^b f(x)dx$
3. $\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx$

Fonction primitive

Soit f une fonction réelle. L'aire sous la courbe sur l'intervalle $]-\infty, x]$ varie lorsqu'on fait varier x de $-\infty$ à $+\infty$. Cette aire est une fonction F de x, appelée fonction primitive de f. Elle est définie par :
$F(x) = \int_{-\infty}^{x} f(\tau)d\tau$

Noter l'utilisation de la variable d'intégration τ. On peut utiliser n'importe quel nom de variable (il s'agit d'une variable muette), différent de la borne d'intégration x.

Propriétés

1. Si $F(x) = \int_{-\infty}^{x} f(\tau)d\tau$, alors $f(x) = \frac{dF(x)}{dx}$

   Donc F se déduit de f par intégration, et f se déduit de F par dérivation.
2. $\int_a^b f(x)dx = F(b) - F(a)$

Notation

On écrit souvent $F(x) = \int f(x)dx$ en omettant les bornes d'intégration.

Chapitre 3

Eléments de calcul des Probabilités

3.1 Introduction

Le calcul des probabilités est une théorie mathématique, et donc fondée axiomatiquement, qui permet de modéliser des phénomènes aléatoires, ou non déterministes.
De tels phénomènes sont bien représentés par les jeux de hasard dont l'étude a initié le calcul des probabilités. Considérons le cas du jeu de dés ; lorsqu'on jette un dé on est certain qu'il va tomber sur la table (phénomène déterministe), mais on n'est pas capable de prédire la valeur qui va sortir (phénomène aléatoire).
Un phénomène déterministe est un phénomène dont on peut prévoir le résultat ; les lois de la physique classique sont des modèles permettant de prédire le résultat d'une expérience donnée. La loi d'Ohm permet de prédire la valeur de l'intensité du courant connaissant la résistance et la tension aux bornes. Les lois de la physique mettent en évidence une régularité qui permet de prédire les résultats d'une expérience lorsqu'on contrôle les causes.
Les phénomènes aléatoires exhibent un autre type de régularité. Prenons le cas des lois de Mendel. Mendel était un biologiste qui étudiait les résultats du croisement de deux espèces de plantes ; plus précisément, il étudiait la transmission de caractères comme la couleur, l'aspect, etc. Une observation typique de régularité d'un nouveau type est d'observer que, sur une série suffisamment grande de croisements de deux espèces A et B, on observait par exemple, dans 1/4 des cas, les caractères de A, et dans 3/4 des cas, les caractères de B. Une telle régularité fréquentielle a donné lieu à ce qu'on appelle les lois de Mendel. Cette régularité permet de prédire la fréquence d'apparition d'un phénomène, ce qui est plus « faible » que la prédiction déterministe. L'étude et la modélisation de tels phénomènes (la recherche de lois) est le champ d'application du calcul des probabilités.

3.2 Ensemble fondamental et événements

Ensemble fondamental

Pour une expérience donnée, l'ensemble des résultats possibles est appelé l'ensemble fon-

damental, que nous noterons S dans la suite du cours. Chaque **résultat** d'expérience est un point de S ou un élément de S.

Evénement

Un événement A est un sous ensemble de S, c'est-à-dire un ensemble de résultats.

L'événement $\{a\}$, constitué par un seul point de S, donc par un seul résultat $a \in S$, est appelé **événement élémentaire**.

L'ensemble vide $\varnothing$ ne contient aucun des résultats possibles : il est appelé **événement impossible**.

L'ensemble S contient tous les résultats possibles : c'est l'**événement certain**.

Si S est fini, ou infini dénombrable, tout sous-ensemble de S est un événement ; ce n'est pas vrai si S est non dénombrable (ceci sort du cadre de ce cours).

Exemple

On jette un dé et on observe le résultat obtenu. L'ensemble fondamental est formé par les 6 résultats possibles :

$S = \{1, 2, 3, 4, 5, 6\}$

L'événement correspondant à l'apparition d'un nombre pair est $A = \{2, 4, 6\}$, qui est bien un sous ensemble de S.

L'événement correspondant à l'apparition d'un nombre premier est $B = \{1, 2, 3, 5\}$, et l'événement correspondant à l'apparition d'un 3 est $C = \{3\}$.

3.3 Opérations sur les événements

Les événements peuvent se combiner entre eux pour former de nouveaux événements. Si A et B sont deux événements, les opérations de combinaison sont :

1. $A \cup B$ est l'événement qui se produit si A ou B (ou les deux) est réalisé. Il est parfois noté $A + B$.
2. $A \cap B$ est l'événement qui se produit si A et B sont réalisés tous les deux. Il est parfois noté $A \cdot B$.
3. $\complement A$ est l'événement qui se produit quand A n'est pas réalisé. On l'appelle aussi négation de A. Il est parfois noté « non A », ou $\bar{A}$.

On dit que l'ensemble des événements, muni des opérations précédentes, forme une *algèbre*.

Evénements incompatibles

Quand deux événements A et B sont tels que $A \cap B = \varnothing$, ils ne peuvent être réalisés simultanément. On dit qu'ils s'**excluent mutuellement**, ou qu'ils sont **incompatibles**.

Système complet d'événements

On dit que les événements $A_1, A_2, ..., A_n$ forment une famille complète si les A_i constituent une partition de S, c'est-à-dire si :

1. les événements sont deux à deux disjoints : $\forall (i \neq j), (A_i \cap A_j = \varnothing)$
2. ils couvrent tout l'espace : $\bigcup_i A_i = S$

Exemples

1. Reprenons l'exemple précédent du jeu de dés :
   $S = \{1, 2, 3, 4, 5, 6\}$, $A = \{2, 4, 6\}$, $B = \{1, 2, 3, 5\}$, $C = \{3\}$.
   $A \cup B = \{1, 2, 3, 4, 5, 6\}$ = apparition d'un nombre pair ou premier
   $A \cap B = \{2\}$ = apparition d'un nombre pair et premier
   $\mathrm{C}C = \{1, 2, 4, 5, 6\}$ = apparition d'un nombre autre que 3
   $A \cap C = \varnothing$: A et C s'excluent mutuellement.
2. Dans l'exemple précédent S était fini et donc dénombrable ; S peut être infini dénombrable comme dans le cas suivant. On jette une pièce de monnaie jusqu'à ce qu'on obtienne pile ; l'ensemble fondamental correspondant est la suite des nombres entiers $S = \{1, 2, 3, ..., n, ...\}$ puisqu'on peut avoir un pile au bout d'un jet, de 2 jets, de n jets, n étant aussi grand que l'on veut.
3. On vise avec une fléchette une cible suffisamment grande ; si on admet que la fléchette est très fine, comme le serait un point de la géométrie, l'espace fondamental est la surface de la cible qui est constituée de points et donc infinie et non dénombrable.

3.4 Règles du calcul des probabilités

Soit un ensemble fondamental S. Nous introduisons une fonction P qui, à tout événement A, associe un nombre réel positif ou nul.
P est dite fonction de probabilité, et $P(A)$ est appelée probabilité de l'événement A, si les conditions ou axiomes suivants sont satisfaits :

1. $P(A) \geq 0$ pour tout événement A : une probabilité est positive ou nulle
2. $P(S) = 1$: la probabilité de l'événement certain est 1
3. $(A \cap B = \varnothing) \Rightarrow (P(A \cup B) = P(A) + P(B))$: permet le calcul de la probabilité de la réunion de deux événements **disjoints**
4. Soit un ensemble dénombrable (fini ou non) d'événements A_i deux à deux disjoints $(A_i \cap A_j = \varnothing)$, alors $P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$.

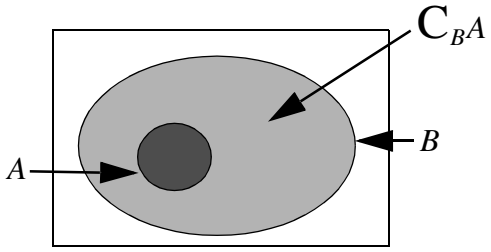
   Cette quatrième condition est proche de la troisième. Elle ne peut cependant pas s'en déduire dans le cas d'un ensemble d'événements infini dénombrable.

Propriétés déduites des quatre conditions précédentes :

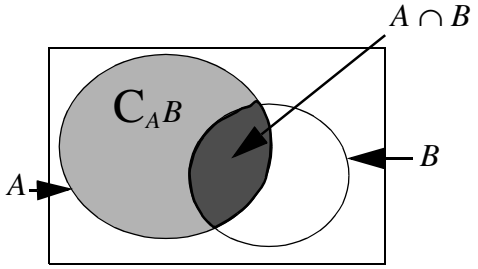
1. $P(\varnothing) = 0$
   Soit A un événement quelconque. A et $\varnothing$ sont évidemment disjoints puisque $A \cap \varnothing = \varnothing$; donc $P(A \cup \varnothing) = P(A) + P(\varnothing)$. Or $A \cup \varnothing = A$; donc $P(A \cup \varnothing) = P(A)$. D'où $P(\varnothing) = 0$.
2. $P(A) \leq 1$
   A et son complémentaire $\mathrm{C}A$ sont disjoints, et leur réunion forme S, de probabilité 1. Donc $P(S) = 1 = P(A \cup \mathrm{C}A) = P(A) + P(\mathrm{C}A)$. Toute probabilité étant positive ou nulle, on

obtient bien $P(A) \leq 1$.

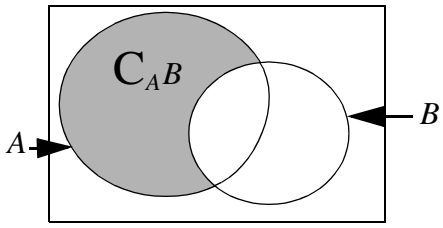
3. $P(\complement A) = 1 - P(A)$
   A démontrer en exercice, en notant que $S = A \cup \complement A$.
4. Si $A \subset B$, alors $P(A) \leq P(B)$.
   A démontrer en exercice, en notant que $B = A \cup \complement_B A$.



5. $\complement_A B = P(A) - P(A \cap B)$
   A démontrer en exercice, en remarquant que $A = \complement_A B \cup (A \cap B)$.



6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
   A démontrer en exercice, en remarquant que $(A \cup B) = \complement_A B \cup B$.



3.5 Remarque

Alors que $P(\varnothing) = 0$, il existe des événements non vides qui peuvent avoir une probabilité nulle. Dans le cas d'un ensemble infini non dénombrable, un tel événement n'est pas nécessairement impossible : il est alors dit « presque impossible ».

Exemple

Considérons l'expérience qui consiste à choisir au hasard un point sur une feuille de papier quadrillé avec une pointe de compas infiniment fine. La probabilité de l'événement *piquer dans un carré donné* a une certaine valeur (par exemple celle du rapport de la surface du

carré avec celle de la feuille de papier) ; en revanche, si on réduit le carré à un point (carré infiniment petit) la probabilité deviendra zéro alors que l'événement (piquer dans ce carré si petit qu'il est devenu un point) n'est pas impossible mais a une probabilité nulle.

De même un événement de probabilité 1 peut ne pas être certain. Il est alors qualifié de « presque certain ».

3.6 Illustration de quelques ensembles probabilisés

3.6.1 Ensemble probabilisé fini

Soit $S = \{a_1, a_2, ..., a_n\}$ un ensemble fondamental fini. On probabilise cet ensemble en attribuant à chaque point a_i un nombre p_i, probabilité de l'événement élémentaire $\{a_i\}$, tel que :

1. $p_i \geq 0$
2. $p_1 + p_2 + ... + p_n = 1$

La probabilité d'un événement quelconque A est la somme des probabilités des a_i qu'il contient : $P(A) = \sum_{a_i \in A} p_i$

Exemple

On jette 3 pièces de monnaie et on compte le nombre de faces obtenues. L'ensemble fondamental correspondant à cette expérience est $S = \{0, 1, 2, 3\}$ puisqu'on peut obtenir comme résultat de l'expérience : 0 face (3 « piles »), 1 face (2 « piles »), 2 faces, ou 3 faces.
On probabilise cet ensemble fini en donnant une valeur p_0, p_1, p_2 et p_3 aux événements $\{0\}$, $\{1\}$, $\{2\}$ et $\{3\}$; comme par exemple $p_0 = 1/8$, $p_1 = 3/8$, $p_2 = 3/8$ et $p_3 = 1/8$.
Considérons l'événement A tel qu'on ait au moins 2 faces, $A = \{a_2, a_3\}$:
$P(A) = p_2 + p_3 = 3/8 + 1/8 = 4/8 = 1/2$

3.6.2 Ensemble fini équiprobable

C'est un ensemble fini probabilisé tel que tous les événements élémentaires ont la même probabilité. On dit aussi qu'il s'agit d'un espace probabilisé uniforme.
$S = \{a_1, a_2, ..., a_n\}$ et $P(\{a_1\}) = p_1, P(\{a_2\}) = p_2, ..., P(\{a_n\}) = p_n$
avec $p_1 = p_2 = ... = p_n = 1/n$
Les jeux de hasard - dés, cartes, loto, etc. - entrent précisément dans cette catégorie :

- jeu de dés : $S = \{1, 2, 3, 4, 5, 6\}$; $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = 1/6$
- jeu de cartes : $S = \{$ensemble des cartes d'un jeu de 52 cartes$\}$; $p_i = 1/52$

Propriété

Dans un ensemble fini équiprobable, la probabilité d'un événement A est égale au rapport du nombre de résultats tel que A est vrai, sur le nombre d'événements de S.

Remarque

Quand on dit qu'on tire « au hasard », on sous-entend que l'ensemble probabilisé considéré est équiprobable.

Exemple

On tire « au hasard » une carte dans un jeu de 52 cartes.
Quelle est la probabilité pour tirer un trèfle ?

$$P(\text{tirer un trèfle}) = \frac{\text{nombre de trèfles}}{\text{nombre de cartes}} = \frac{13}{52} = \frac{1}{4}$$

Quelle est la probabilité de tirer un roi ?

$$P(\text{tirer un roi}) = \frac{\text{nombre de rois}}{\text{nombre de cartes}} = \frac{4}{52} = \frac{1}{13}$$

Quelle est la probabilité de tirer un roi de trèfle ?

$\frac{1}{52}$.

Remarque

Le cas des ensembles finis équiprobables est le plus simple à appréhender. Il faut insister sur le fait que l'équiprobabilité n'est qu'un cas particulier des ensembles probabilisés ; **ce n'est (de loin) pas le plus utile en médecine**.

3.6.3 Ensembles probabilisés infinis

3.6.3.1 Cas dénombrable

On a alors un ensemble fondamental de la forme $S = \{a_1, a_2, ..., a_n, ...\}$ comme dans le cas fini. Cet ensemble fondamental est probabilisé en affectant à chaque élément a_i une valeur réelle p_i telle que :

$$p_i \geq 0 \text{ et } \sum_{i=1}^{\infty} p_i = 1 .$$

La probabilité d'un événement quelconque est alors la somme des p_i correspondant à ses éléments.

Exemple 1

$A = \{a_{25}, a_{31}, a_{43}\}$
$P(A) = p_{25} + p_{31} + p_{43}$

Exemple 2

Si on reprend l'expérience consistant à jeter une pièce et à compter le nombre de jets jusqu'à ce qu'on obtienne un résultat « pile » (c'est un espace infini dénombrable), on peut

construire un espace probabilisé en choisissant :

$$p_1 = \frac{1}{2}, p_2 = \frac{1}{4}, ..., p_n = \frac{1}{2^n}, ..., p_\infty = 0$$

Remarque :
Le choix des p_i est arbitraire ; en réalité, il est justifié soit par des considérations a priori (dans le cas de l'expérience précédente on suppose que chaque jeté constitue une expérience avec $P(\text{pile}) = P(\text{face}) = 1/2$ et que le résultat d'un jet n'influe pas sur le suivant). Il peut être aussi estimé ; c'est le problème des statistiques qui, à partir de nombreuses réalisations de l'expérience, permet d'approcher les valeurs p_i (ce point sera revu dans la suite du cours et constitue l'objet de l'approche statistique).

3.6.3.2 Cas d'un ensemble probabilisé infini non dénombrable

Pour illustrer ce cas, on peut prendre l'exemple de la chute d'un satellite en fin de vie (ce fut le cas, en octobre 1993 pour un gros satellite chinois dont on parla beaucoup dans la presse). Dans l'état actuel des connaissances sur l'orbite de ce satellite, on n'est pas capable de prédire l'endroit de la chute ; l'hypothèse retenue est alors celle d'un espace de probabilité uniforme. Dans ce cas, le satellite a la même chance de tomber dans n'importe quelle parcelle du monde et on peut calculer la probabilité qu'il tombe sur Paris comme le rapport de la surface de Paris sur la surface du globe.
Lorsqu'on se rapprochera de l'échéance, on pourra avoir des hypothèses plus précises, et on pourra prédire par exemple que le point de chute aura un maximum de probabilité dans une région, la probabilité autour de cette région étant d'autant plus petite qu'on s'éloigne de ce maximum.
Il s'agit bien sûr d'un espace infini non dénombrable puisqu'on peut réduire (au moins par l'esprit) la taille de l'élément de la région considérée à celle d'un point. Des probabilités peuvent donc être associées à chaque région de taille non nulle, mais la probabilité d'une chute en un point donné est nulle, puisque sa surface est nulle. Nous verrons dans la suite que les probabilités se calculent généralement à partir d'une densité (de probabilité) associée à chaque point : lorsque les points d'une région ont une densité élevée, la probabilité de chute dans cette région est élevée.

Chapitre 4

Probabilité Conditionnelle ; Indépendance et Théorème de Bayes

4.1 Probabilité conditionnelle

Soient A et B deux événements quelconques d'un ensemble fondamental S muni d'une loi de probabilité P. On s'intéresse à ce que devient la probabilité de A lorsqu'on apprend que B est déjà réalisé, c'est-à-dire lorsqu'on restreint l'ensemble des résultats possibles S à B.
La probabilité conditionnelle de A, sachant que l'événement B est réalisé, est notée $P(A/B)$ et est définie par la relation suivante :

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Equation 1 : probabilité conditionnelle

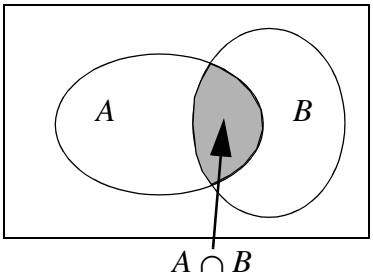


Figure 1 : probabilité conditionnelle

Cette relation générale pour tout espace probabilisé s'interprète facilement dans le cas où S est un espace équiprobable (mais cette relation est vraie pour un espace non-équiprobable !). En notant $|A|$ le nombre d'éléments de A :

$$P(A \cap B) = \frac{|A \cap B|}{|S|}, P(B) = \frac{|B|}{|S|}, P(A/B) = \frac{|A \cap B|}{|B|}$$

$P(A/B)$ traduit le rapport de la surface de $A \cap B$ sur la surface de B dans la figure 1.
Toujours dans le cas où S est équiprobable, on a

$$P(A/B) = \frac{\text{nombre de réalisations possibles de } A \text{ et } B \text{ en même temps}}{\text{nombre de réalisations de } B}$$

Cette interprétation de la probabilité conditionnelle, facile à appréhender dans le cas d'équiprobabilité, est la définition générale de la probabilité conditionnelle qu'on doit utiliser telle quelle, sans chercher une interprétation fréquentiste dans tous les cas.

Exemple

On jette une paire de dés bien équilibrés (espace équiprobable). On observe une réalisation de l'événement {somme des dés = 6}. Quelle est la probabilité pour qu'un des deux dés ait donné le résultat 2 ?
B = {somme des deux dés = 6}
A = {au moins un des deux dés donne 2}
B = {(2, 4), (4, 2), (1, 5), (5, 1), (3, 3)}
Nombre de réalisations de $A \cap B$ = {(2, 4), (4, 2)} = 2

D'où $P(A/B) = \frac{|A \cap B|}{|B|} = \frac{2}{5}$, alors que $P(A) = \frac{11}{36}$ (à vérifier).

4.2 Théorème de la multiplication

Reprenons l'équation 1, définition des probabilités conditionnelles : $P(A/B) = \frac{P(A \cap B)}{P(B)}$

On en tire immédiatement

$$P(A \cap B) = P(A/B)P(B) = P(B/A)P(A)$$
Equation 2 : théorème de la multiplication

L'équation 2 peut se généraliser facilement. Soient A_1, ..., A_n des événements quelconques d'un espace probabilisé ; à partir de l'équation 2, on montre :
$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2/A_1)P(A_3/(A_1 \cap A_2))...P(A_n/(A_1 \cap A_2 ... \cap A_{n-1}))$

Exemple

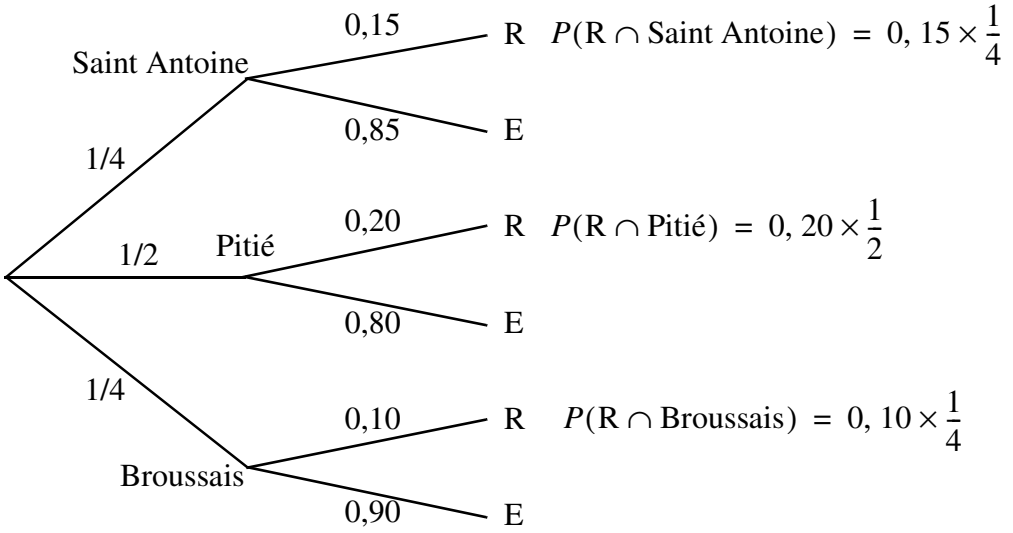
Une boîte contient 10 articles dont 4 sont défectueux. On tire 3 objets de cette boîte. Calculer la probabilité pour que ces 3 objets soient défectueux.
$P(1^{er}$ défectueux) = 4/10
$P(2^{ème}$ défectueux / 1^{er} défectueux) = 3/9
$P(3^{ème}$ défectueux / 1^{er} et $2^{ème}$ défectueux) = 2/8
$P(1^{er}$ et $2^{ème}$ et $3^{ème}$ défectueux) = 4/10×3/9×2/8 = 1/30.

4.3 Diagramme en arbre

On considère une séquence finie d'expériences dont chacune d'entre elles a un nombre fini de résultats possibles. Les probabilités associées aux résultats possibles d'une expérience dépendent du résultat de l'expérience précédente ; il s'agit de probabilités conditionnelles. Pour représenter cette séquence, on utilise une représentation « en arbre », le théorème précédent permettant de calculer la probabilité de chaque feuille de l'arbre.

Exemple

On sait que les taux de réussite au concours dans les trois CHU Pitié, Saint Antoine et Broussais sont respectivement (données arbitraires) de 0,20 ; 0,15 ; et 0,10 (0,20 = P(Réussite/Pitié)) ; on sait que 1/4 des étudiants de Paris VI sont à Saint Antoine, 1/4 à Broussais et 1/2 à la Pitié. Quelle est la probabilité qu'un étudiant de Paris VI soit reçu au concours ?



R signifie réussite et E échec.

$P(\text{R}) = P(\text{R} \cap \text{Saint Antoine}) + P(\text{R} \cap \text{Pitié}) + P(\text{R} \cap \text{Broussais})$

$P(\text{R}) = 0,15 \times 1/4 + 0,20 \times 1/2 + 0,10 \times 1/4$

La probabilité qu'un chemin particulier de l'arbre se réalise est, d'après le théorème de la multiplication, le produit des probabilités de chaque branche du chemin.

Les chemins s'excluant mutuellement, la probabilité d'être reçu est égale à la somme des probabilités d'être reçu pour tout chemin aboutissant à un état R (reçu).

4.4 Théorème de Bayes

En reprenant l'équation 2 page 30 de la section 4.2, on obtient le théorème de Bayes :

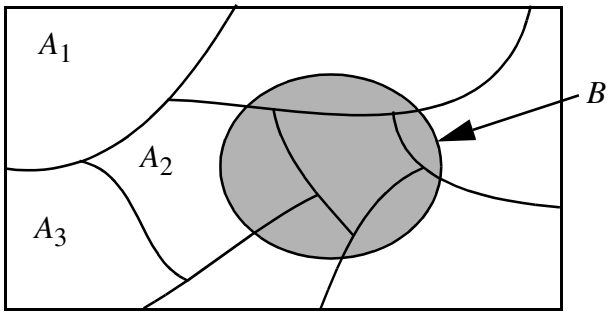
$$P(B/A) = \frac{P(A/B)P(B)}{P(A)}$$

Equation 3 : théorème de Bayes

Ce théorème existe aussi sous une forme développée que nous introduisons maintenant.
Considérons des événements $A_1, ..., A_n$ tels qu'ils forment une **partition** de l'ensemble fondamental S.
Par définition, les Ai s'excluent mutuellement et leur union est S :

$$\forall (i \neq j), (A_i \cap A_j = \varnothing) \; ; \; \bigcup_{i=1}^{n} A_i = S$$

Soit B un événement quelconque



De $S = A_1 \cup A_2 \cup ... \cup A_n$ et de $B \cap S = B$, on tire $B = B \cap (A_1 \cup A_2 \cup ... \cup A_n)$.
Soit, par distributivité, $B = (B \cap A_1) \cup (B \cap A_2) \cup ... \cup (B \cap A_n)$.
En remarquant que les $B \cap A_i$ sont exclusifs, puisque les A_i le sont, et en appliquant le 3ème axiome du calcul des probabilités on obtient la formule dite des « probabilités totales » :

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + ... + P(B \cap A_n)$$

Equation 4 : probabilités totales

En appliquant le théorème de la multiplication :
$P(B) = P(B/A_1)P(A_1) + P(B/A_2)P(A_2) + ... + P(B/A_n)P(A_n)$

Or, par la forme simple du théorème de Bayes, on a $P(A_i/B) = \frac{P(B/A_i)P(A_i)}{P(B)}$
D'où la forme développée du théorème de Bayes :

$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{P(B/A_1)P(A_1) + P(B/A_2)P(A_2) + ... + P(B/A_n)P(A_n)}$$

Equation 5 : formule développée de Bayes

Exemple 1

Reprenons l'exemple des résultats au concours des étudiants de Paris VI.

Comme précédemment, soit R l'événement « un étudiant de Paris VI est reçu ». On a, en notant C_1, C_2, C_3 les 3 CHU Saint Antoine, Pitié et Broussais respectivement :

$P(\mathrm{R}) = P(C_1)P(\mathrm{R}/C_1) + P(C_2)P(\mathrm{R}/C_2) + P(C_3)P(\mathrm{R}/C_3)$

[noter que c'est la même chose que la somme des probabilités des chemins de l'arbre, qui conduisent à un succès]

Le théorème de Bayes permet de répondre à la question duale. Au lieu de chercher la probabilité d'obtenir un étudiant reçu sachant qu'il vient d'un CHU donné, on cherche la probabilité qu'un étudiant soit inscrit à un CHU donné sachant qu'il est reçu (probabilité des causes).

Calculons la probabilité qu'un étudiant reçu soit issu du CHU Pitié-Salpêtrière.

$$P(C_2/\mathrm{R}) = \frac{P(\mathrm{R}/C_2)P(C_2)}{P(\mathrm{R}/C_1)P(C_1) + P(\mathrm{R}/C_2)P(C_2) + P(\mathrm{R}/C_3)P(C_3)}$$

Avec $P(C_1) = 0{,}25$; $P(C_2) = 0{,}50$; $P(C_3) = 0{,}25$;
et $P(\mathrm{R}/C_1) = 0{,}15$; $P(\mathrm{R}/C_2) = 0{,}20$; $P(\mathrm{R}/C_3) = 0{,}10$.

$$\text{D'où } P(C_2/\mathrm{R}) = \frac{0,20 \times 0,50}{0,15 \times 0,25 + 0,20 \times 0,50 + 0,10 \times 0,25} = 0,61$$

Ce qui signifie que, dans ce cas, la probabilité qu'un étudiant appartienne à C_2, s'il est reçu, est plus grande que si l'on ne sait rien (probabilité a priori $P(C_2) = 0{,}50$).

Cette façon de calculer les probabilités des causes connaissant les effets est essentielle en médecine. En effet, le problème du diagnostic peut être posé en ces termes.

Exemple 2

Considérons, pour illustrer notre propos, le problème du diagnostic d'une douleur aiguë de l'abdomen. Il s'agit d'un patient arrivant aux urgences pour un « mal au ventre ».

Si l'on ne sait rien d'autre sur le patient (on n'a pas fait d'examen clinique ou complémentaire), on ne connaît que les probabilités d'avoir tel ou tel diagnostic si on observe une douleur.

Soient D_1, D_2 et D_3 les 3 diagnostics principaux (il y en a en fait au moins une douzaine) et exclusifs ; par exemple D_1 = appendicite, D_2 = perforation d'ulcère, D_3 = autres diagnostics.

Soit un signe s_1 pour lequel on connaît $P(s_1/D_1)$, $P(s_1/D_2)$, et $P(s_1/D_3)$.

Par exemple, s_1 serait « présence d'une fièvre $\geq$ 38,5°C » ; $P(s_1/D_1) = 0{,}90$; $P(s_1/D_2) = 0{,}30$; et $P(s_1/D_3) = 0{,}10$.

Ces probabilités peuvent être estimées sur une population de patients en dénombrant le nombre de sujets ayant le diagnostic D_1 et présentant le signe s_1. De même, on peut connaître $P(D_1)$, $P(D_2)$ et $P(D_3)$.

Le problème diagnostique se pose comme celui de choisir par exemple le diagnostic le plus probable connaissant le signe s_1 ; pour ce faire, on calcule $P(D_1/s_1)$, $P(D_2/s_1)$, $P(D_3/s_1)$ et on retient le diagnostic qui a la plus grande probabilité : c'est l'application de l'approche bayesienne au problème de l'aide au diagnostic.

4.5 Indépendance entre événements

On dit que deux événements A et B sont indépendants si la probabilité pour que A soit réalisé n'est pas modifiée par le fait que B se soit produit. On traduit cela par $P(A / B) = P(A)$.

D'après la définition d'une probabilité conditionnelle, $P(A/B) = \frac{P(A \cap B)}{P(B)}$, on tire la définition :

A et B sont indépendants si et seulement si $P(A \cap B) = P(A)P(B)$.
La symétrie de cette définition implique qu'on a aussi bien $P(A / B) = P(A)$ (A est indépendant de B) que $P(B / A) = P(B)$ (B est indépendant de A) : l'apparition d'un des deux événements n'influe pas sur l'apparition de l'autre.

Note

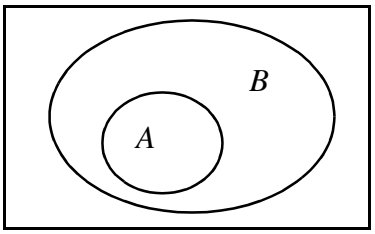
Ce qui est défini précédemment est l'indépendance de deux événements. Si on considère maintenant 3 événements A, B, C, on dira que ces 3 événements sont indépendants :

1. s'ils sont indépendants 2 à 2 : A indépendant de B ; A indépendant de C ; et B indépendant de C
2. et si $P(A \cap B \cap C) = P(A)P(B)P(C)$. Cette condition n'est pas une conséquence des précédentes.

4.6 Indépendance, inclusion et exclusion de deux événements

Considérons deux événements A et B.

1. Si $A \subset B$ (A est inclus dans B) : si A est réalisé, alors B aussi.

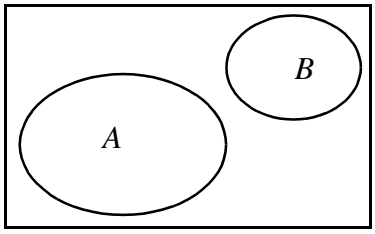


Alors $P(A \cap B) = P(A)$.

D'où $P(B/A) = \frac{P(A \cap B)}{P(A)} = 1$ et $P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)}$.

A et B ne sont **pas indépendants**.

2. Si $A \cap B = \varnothing$ (A et B sont exclusifs) : si A est réalisé, B ne peut pas l'être.



Alors $P(A \cap B) = P(\varnothing) = 0$.

D'où $P(A/B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0}{P(B)} = 0$.

De même A et B ne sont **pas indépendants**.

Chapitre 5

Variables aléatoires

5.1 Définition d'une variable aléatoire

Considérons un ensemble fondamental S correspondant à une certaine expérience. Les éléments de S, résultats possibles de l'expérience, ne sont généralement pas des nombres. Il est cependant utile de faire correspondre un nombre à chaque élément de S, en vue de faire ensuite des calculs. Pour un jet de dé, il semble naturel de faire correspondre à la face obtenue par le jet, le nombre de points qu'elle porte, mais ce n'est pas une obligation. Si on jette 2 dés, on s'intéressera par exemple à la somme des points obtenus. Pour une carte à jouer, il faut convenir d'une valeur pour chaque carte.
Une variable aléatoire X, sur un ensemble fondamental S, est une application de S dans $\Re$: à tout résultat possible de l'expérience (à tout élément de S), la variable aléatoire X fait correspondre un nombre.
Lorsque S est fini ou infini dénombrable, toute application de S dans $\Re$ est une variable aléatoire. Lorsque S est non dénombrable, il existe certaines applications de S dans $\Re$ qui ne sont pas des variables aléatoires. En effet, la définition rigoureuse d'une variable aléatoire X impose que tout intervalle de $\Re$ soit l'image d'un événement de S par l'application X. Cette condition est vérifiée pour toute application X si S est fini ou dénombrable, puisque toute partie de S est un événement. Ce n'est plus vrai si S est non dénombrable. Heureusement, les applications choisies naturellement sont des variables aléatoires.
On parle de variable aléatoire **discrète** lorsque la variable est une application de S dans un sous-ensemble discret de $\Re$, le plus souvent **N** ou une partie de **N**. On parle sinon de variable aléatoire **continue**.

Pour un nombre réel a donné, l'événement constitué de tous les résultats ξ d'expérience tels que $X(\xi) = a$ est noté $[X(\xi) = a]$, ou, en abrégé, $X = a$.
Pour deux nombres réels a et b $(a \leq b)$, l'événement constitué de tous les résultats ξ d'expérience tels que $a \leq X(\xi) \leq b$ est noté $[a \leq X(\xi) \leq b]$ ou, en abrégé, $a \leq X \leq b$.

Si X et Y sont des variables aléatoires définies sur le même ensemble fondamental S, on peut montrer que les fonctions suivantes sont aussi des variables aléatoires :
$(X + Y)(s) = X(s) + Y(s)$ $\quad$ $(X + k)(s) = X(s) + k$
$(kX)(s) = kX(s)$ $\quad$ $(XY)(s) = X(s)\ Y(s)$
pour tout élément s de S.

5.2 Variables aléatoires finies

Considérons maintenant le cas le plus simple d'une variable aléatoire finie, que nous généraliserons dans un second temps à une variable aléatoire infinie dénombrable, puis continue.
Soit X une variable aléatoire sur un ensemble fondamental S à valeurs finies :
$X(S) = \{x_1, x_2, ..., x_n\}$.
$X(S)$ devient un ensemble probabilisé si l'on définit la probabilité $P(X = x_i)$ pour chaque x_i, que l'on note $f(x_i)$. Cette fonction, définie par $f(x_i) = P(X = x_i)$ est appelée distribution de probabilité de X.
Puisque les $f(x_i)$ sont des probabilités sur les événements $\{X=x_1, X=x_2, ..., X=x_n\}$, on a par conséquent :

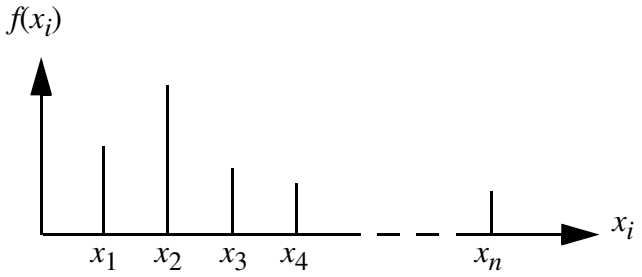
$$(\forall i), f(x_i) \geq 0 \text{ et } \sum_{i=1}^{n} f(x_i) = 1 .$$

5.2.1 Représentation d'une loi de probabilité finie

On peut représenter la loi de probabilité $f(x_i)$ par une table :

x_1	x_2	.....	x_n
$f(x_1)$	$f(x_2)$	.....	$f(x_n)$

Ou par un diagramme en bâtons :



où la hauteur du bâton positionné en x_i a pour valeur $f(x_i)$.

5.2.2 Espérance mathématique d'une loi finie

L'espérance mathématique cherche à traduire la tendance centrale de la loi. Il s'agit d'une moyenne où chacune des valeurs x_i intervient d'autant plus que sa probabilité est importante, c'est-à-dire d'un barycentre ou d'un centre de gravité. On définit alors la **moyenne théorique** (ou **vraie**), ou **espérance mathématique** d'une loi $f(x_i)$ par

$$\mu_X = E(X) = \sum_{i=1}^{n} x_i f(x_i) = x_1 f(x_1) + x_2 f(x_2) + ... + x_n f(x_n) .$$

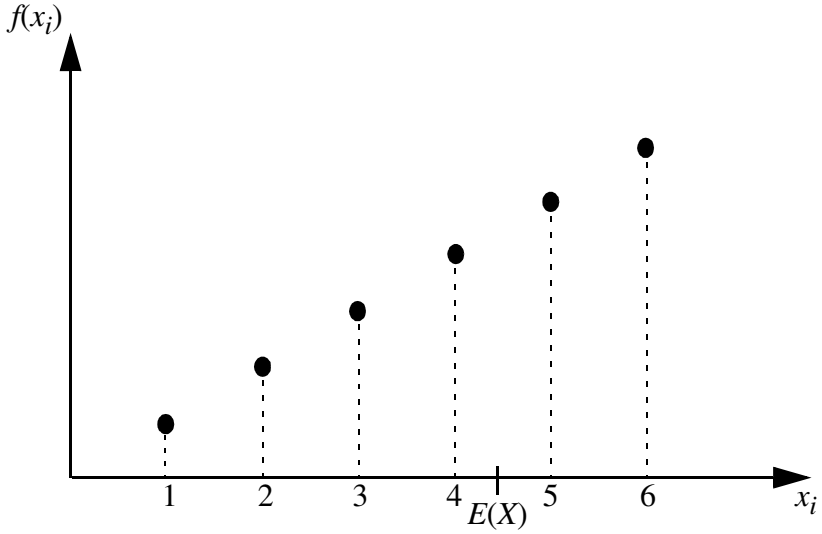
μ_X peut être notée μ s'il n'y a pas de confusion possible.

Exemple

On considère l'expérience qui consiste à jeter deux dés parfaitement équilibrés. L'espace fondamental est constitué par l'ensemble des couples ordonnés
$S = \{(1, 1), (1, 2), (1, 3), ..., (6, 6)\}$
C'est un espace équiprobable (tous les couples résultats élémentaires du tirage sont équiprobables).
Considérons la variable aléatoire définie comme suit : soit $s = (a, b)$ un élément quelconque de S ; on pose $X(s) = X(a, b) = \max(a, b)$
(la valeur de $X(s)$ est égale à a si $a > b$ et à b dans le cas contraire).
X est une variable aléatoire sur S avec $X(S) = \{1, 2, 3, 4, 5, 6\}$,
et la loi de probabilité
$f(1) = P(X = 1) = P(\{(1, 1)\}) = 1/36$;
$f(2) = P(X = 2) = P(\{(1, 2), (2, 1), (2, 2)\}) = 3/36$;
$f(3) = 5/36$; $f(4) = 7/36$; $f(5) = 9/36$; $f(6) = 11/36$.
Soit :

x_i	1	2	3	4	5	6
$f(x_i)$	1/36	3/36	5/36	7/36	9/36	11/36

$E(X) = 1/36 + 6/36 + 15/36 + 28/36 + 45/36 + 66/36 = 161/36 \approx 4{,}47$



Théorèmes

1. Soit X une variable aléatoire et k une constante réelle. On a :
   $E(kX) = kE(X)$
   $E(X + k) = E(X) + k$
2. Soient X et Y deux variables aléatoires définies sur le même espace fondamental S. On a :

$E(X + Y) = E(X) + E(Y)$

On en déduit que pour n variables aléatoires X_i, définies sur le même espace fondamental :

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

(l'espérance de la somme est la somme des espérances).

Exemple

Considérons l'expérience du jeu de dés où $S = \{1, 2, 3, 4, 5, 6\}$ uniforme (équiprobable).
Soit $X(S)$ une première variable aléatoire [noter que l'application définissant X est l'application identité].
On a $X(S) = \{1, 2, 3, 4, 5, 6\}$
et $f(1) = f(2) = f(3) = f(4) = f(5) = f(6) = 1/6$
$E(X) = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 21/6$

Soit $Y(S)$ une seconde variable aléatoire telle que
$Y(S) = 1$ si le chiffre tiré est impair
$Y(S) = 2$ si le chiffre tiré est pair.
Donc $Y(S) = \{1, 2\}$
$f(1) = P(\{1, 3, 5\}) = 1/2$
$f(2) = P(\{2, 4, 6\}) = 1/2$
$E(Y) = 1/2 + 1 = 1,5$

Calculons maintenant la loi de $(X + Y)(S)$
$(X + Y)(s) = X(s) + Y(s)$
Pour $s = 1$, $(X + Y)(1) = X(1) + Y(1) = 1 + 1 = 2$
Pour $s = 2$, $(X + Y)(2) = X(2) + Y(2) = 2 + 2 = 4$
Pour $s = 3$, $(X + Y)(3) = X(3) + Y(3) = 3 + 1 = 4$
Pour $s = 4$, $(X + Y)(4) = X(4) + Y(4) = 4 + 2 = 6$
Pour $s = 5$, $(X + Y)(5) = X(5) + Y(5) = 5 + 1 = 6$
Pour $s = 6$, $(X + Y)(6) = X(6) + Y(6) = 6 + 2 = 8$
On a donc $(X + Y)(S) = \{2, 4, 6, 8\}$ et $f(2) = 1/6$, $f(4) = 2/6$, $f(6) = 2/6$, $f(8) = 1/6$
$E(X + Y) = 2/6 + 8/6 + 12/6 + 8/6 = 30/6$
Or on retrouve bien ce résultat en utilisant $E(X) + E(Y) = 21/6 + 3/2 = 30/6$.

Remarque

Lorsqu'on doit calculer l'espérance d'une fonction $g(X)$, il faut étudier la variable $Y = g(X)$ dont les valeurs sont $y_1 = g(x_1)$, $y_2 = g(x_2)$, ..., $y_n = g(x_n)$. Alors :
$P(Y = y_i) = P[g(X) = g(x_i)] = P(X = x_i) = f(x_i)$
Donc :

$$E(g(X)) = E(Y) = \sum_{i=1}^{n} y_i P(Y = y_i) = \sum_{i=1}^{n} g(x_i) f(x_i)$$

Par exemple, si l'on doit calculer $E(X^2)$, on considère la variable $Y = X^2$ dont les valeurs sont $y_1 = {x_1}^2$, $y_2 = {x_2}^2$, ..., $y_n = {x_n}^2$. Alors :

$$E(X^2) = E(Y) = \sum_{i=1}^{n} y_i P(Y = y_i) = \sum_{i=1}^{n} x_i^2 f(x_i)$$

On constate que pour calculer l'espérance d'un carré, il faut élever les valeurs x_i au carré, mais pas les probabilités $f(x_i)$ associées.

5.2.3 Variance et écart-type

Après avoir traduit la tendance centrale par l'espérance, il est intéressant de traduire la dispersion autour de l'espérance par une valeur (la variance ou l'écart-type).
La variance de X, notée $var(X)$ ou σ_X^2, est définie par :

$$\sigma_X^2 = var(X) = E((X-\mu_X)^2) \text{ où } \mu_X = E(X)$$

L'écart-type de X, noté $\sigma(X)$ ou σ_X, est défini par $\sigma(X) = \sigma_X = \sqrt{var(X)}$.
σ_X peut être notée σ s'il n'y a pas de confusion possible.
Remarques :

1. On démontre facilement que $var(X) = E(X^2) - \mu_X^2$
   En effet :
   $$E((X-\mu_X)^2) = \sum_{i=1}^{n} (x_i-\mu_X)^2 f(x_i) = \sum_{i=1}^{n} (x_i^2 - 2\mu_X x_i + \mu_X^2) f(x_i)$$
   $$E((X-\mu_X)^2) = \sum_{i=1}^{n} x_i^2 f(x_i) - 2\mu_X \sum_{i=1}^{n} x_i f(x_i) + \mu_X^2 \sum_{i=1}^{n} f(x_i)$$
   $$E((X-\mu_X)^2) = \sum_{i=1}^{n} x_i^2 f(x_i) - 2\mu_X^2 + \mu_X^2 = E(X^2) - \mu_X^2$$
2. $\sigma_X^2 \geq 0$, par définition

3. Soit X une variable aléatoire de moyenne μ et de variance σ^2.

   On définit la variable centrée réduite par $Y = \frac{X-\mu}{\sigma}$.

   On peut montrer facilement (faites l'exercice) que $E(Y) = 0$ et $var(Y) = E(Y^2) = 1$.
4. Si a est une constante, on montre que $var(X + a) = var(X)$ et $var(aX) = a^2 var(X)$.

5.2.4 Loi de probabilité produit

Soient X et Y deux variables aléatoires finies sur le même espace fondamental S ayant pour image respective :
$X(S) = \{x_1, x_2, ..., x_n\}$
$Y(S) = \{y_1, y_2, ..., y_m\}$.
Considérons l'ensemble produit
$X(S) \times Y(S) = \{(x_1, y_1), (x_1, y_2), ..., (x_n, y_m)\}$

(ensemble des couples (x_i, y_j) pour $i = 1, ..., n$ et $j = 1, ..., m$)
Cet ensemble produit peut être transformé en ensemble probabilisé si on définit la probabilité du couple ordonné (x_i, y_j) par $P([X = x_i] \cap [Y = y_j])$ que l'on note $p(x_i, y_j)$. Cette loi de probabilité de X, Y est appelée distribution jointe de X et Y.

Y \ X	x_1	x_2	x_3	.....	x_n	$\sum_{i=1,n} x_i$
y_1	$p(x_1, y_1)$	$p(x_2, y_1)$				$g(y_1)$
y_2	$p(x_1, y_2)$					$g(y_2)$
.....						
y_m	$p(x_1, y_m)$					
$\sum_{j=1,m} y_j$	$f(x_1)$	$f(x_2)$				1

Les fonctions $f(x_i) = \sum_{j=1}^{m} p(x_i, y_j)$ et $g(y_j) = \sum_{i=1}^{n} p(x_i, y_j)$
sont souvent appelées lois de probabilité marginales de X et de Y. Il s'agit simplement de leurs distributions.
La loi de probabilité $p(x_i, y_j)$ possède, bien entendu, les propriétés d'une loi :

1. $p(x_i, y_j) \geq 0, \forall i, j$
2. $\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) = 1$

Soient μ_X et μ_Y les espérances de X et de Y, σ_X et σ_Y leurs écart-types. On montre facilement que $var(X + Y) = \sigma_X^2 + \sigma_Y^2 + 2cov(X, Y)$, où $cov(X, Y)$ représente la **covariance de X et Y** et est définie par :

$$cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)] = \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i - \mu_X)(y_j - \mu_Y) p(x_i, y_j)$$

De même que pour la variance (voir section 5.2.3), on a :
$cov(X, Y) = E(X\,Y) - \mu_X \mu_Y$

Une notion dérivée de la covariance est celle de **corrélation** entre X et Y, définie par :

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

On peut vérifier que
$\rho(X, Y) = \rho(Y, X)$
$-1 \leq \rho(X, Y) \leq 1$
$\rho(X, X) = 1$

$\rho(aX + b, cY + d) = \rho(X, Y)$ si a et c non nuls

5.2.5 Variables aléatoires indépendantes

Soient X et Y deux variables aléatoires sur un même espace fondamental S. X et Y sont indépendantes si tous les événements $X = x_i$ et $Y = y_j$ sont indépendants :
$P([X = x_i] \cap [Y = y_j]) = P(X = x_i) \cdot P(Y = y_j)$ pour tous les couples i, j.
Autrement dit, si $f(x_i)$ et $g(y_j)$ sont les distributions respectives de X et Y, les variables sont indépendantes si et seulement si on a
$p(x_i, y_j) = f(x_i)g(y_j)$
(la probabilité conjointe est égale au produit des probabilités marginales).
Il en découle les propriétés importantes suivantes : si X et Y sont indépendantes, on a (attention la réciproque n'est pas toujours vraie)

1. $E(XY) = E(X)E(Y)$
2. $var(X + Y) = var(X) + var(Y)$
3. $cov(X, Y) = 0$ et $\rho(X, Y) = 0$

5.2.6 Fonction de répartition

Si X est une variable aléatoire, on définit sa fonction de répartition $F(x)$ par
$F(x) = P(X \leq x)$ pour tout $x \in \Re$
Si X est une variable aléatoire discrète on a $F(x) = \sum_{x_i \leq x} f(x_i)$
Dans tous les cas, $F(x)$ est une fonction monotone croissante, c'est-à-dire $F(a) \geq F(b)$ si $a \geq b$
De plus
$\lim_{x \to -\infty} F(x) = 0$ et $\lim_{x \to \infty} F(x) = 1$

5.3 Variables infinies dénombrables

Tout ce qui a été vu précédemment dans le cas où S est fini ($S = \{s_1, s_2, ..., s_n\}$) se généralise (nous ne verrons pas les démonstrations) au cas où S est infini dénombrable ; on aura par exemple

$$\mu_X = E(X) = \sum_{i=1}^{\infty} x_i f(x_i)$$

La somme converge à l'infini vers $E(X)$, toutes les autres propriétés sont conservées, les sommes devenant des séries.

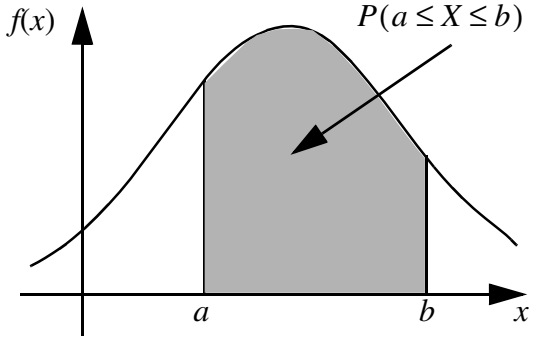
5.4 Variables aléatoires continues

La généralisation au continu est délicate et même difficile si on ne dispose pas d'outils mathématiques hors du champ de ce cours.
Nous nous contenterons de procéder par analogie avec le cas discret.
Une variable aléatoire X dont l'ensemble image $X(S)$ est un intervalle de $\Re$ est une variable aléatoire continue (continue par opposition à discrète, cf supra).
Rappelons que, par définition d'une variable aléatoire, $a \le X \le b$ est un événement de S dont la probabilité est bien définie.
On définit la loi de probabilité de X, ou distribution de X, à l'aide d'une fonction $f(x)$, appelée **densité de probabilité** de X, telle que

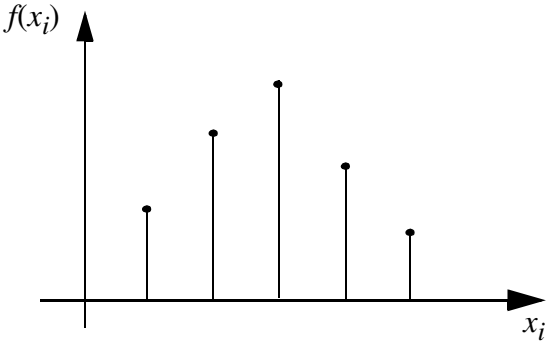
$$\int_a^b f(x)dx \;=\; P(a \le X \le b)$$

Remarques

1. Si f est donnée, la probabilité $P(a \le X \le b)$ est la surface sous la courbe entre a et b



2. Le passage du discret au continu transforme les sommes Σ en intégrales $\int$ et $f(x_i)$ en $f(x)dx$. Ainsi, soit X une variable aléatoire discrète et $f(x_i)$ sa distribution



La formule $P(x_k \le X \le x_n) \;=\; \sum_{i=k}^{n} f(x_i)$ est analogue à $P(a \le X \le b) \;=\; \int_a^b f(x)dx$

En utilisant cette analogie, on admettra les définitions suivantes pour une variable aléatoire X, continue, de distribution $f(x)$:

1. $f(x) \ge 0$ (analogue à $f(x_i) \ge 0$)

2. $\int_{\Re} f(x)dx = 1$ (analogue à $\sum_i f(x_i) = 1$)
3. $\mu_X = E(X) = \int_{\Re} xf(x)dx$ (analogue à $\sum_i x_i f(x_i)$)
4. $\sigma_X^2 = var(X) = \int_{\Re} (x-\mu_X)^2 f(x)dx$ (analogue à $\sum_i (x_i-\mu_X)^2 f(x_i)$)
5. $\sigma_X^2 = var(X) = \int_{\Re} x^2 f(x)dx - \mu_X^2$ (analogue à $\sum_i x_i^2 f(x_i) - \mu_X^2$)
6. $\sigma(X) = \sigma_X = \sqrt{var(X)}$
7. $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(\tau)d\tau$ (analogue à $\sum_{x_i \leq x} f(x_i)$)

   Les propriétés de la fonction de répartition données section 5.2.6 page 43 sont conservées.
8. $P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$

Pour résumer l'analogie entre le cas discret et le cas continu, un point du domaine discret correspond à un intervalle dans le cas continu, la somme discrète correspond à l'intégrale.

Chapitre 6

Exemples de distributions

6.1 Lois discrètes

6.1.1 Loi de Bernoulli

On considère une expérience n'ayant que deux résultats possibles, par exemple succès et échec (ou présence et absence d'une certaine caractéristique). On introduit la variable aléatoire X qui associe la valeur 0 à l'échec (ou à l'absence de la caractéristique) et la valeur 1 au succès (ou à la présence de la caractéristique). Cette variable aléatoire est appelée variable de Bernoulli.

Distribution de *X*

Appelons p la probabilité de l'événement succès :
$P(\{\text{succès}\}) = P(X = 1) = p$
d'où
$P(\{\text{échec}\}) = P(X = 0) = 1 - p$
On note souvent $q = 1 - p$

Espérance de *X*

$$\mu_X = E(X) = \sum x_i P(X = x_i) = 1 \times P(X = 1) + 0 \times P(X = 0) = p$$

Variance de *X*

$$\sigma_X^2 = var(X) = E[(X - \mu_X)^2] = E(X^2) - \mu_X^2$$

$$\sigma_X^2 = [1^2 \times P(X = 1) + 0^2 \times P(X = 0)] - p^2$$

$$\sigma_X^2 = p - p^2 = p(1 - p) = pq$$

6.1.2 Loi binomiale

Définition

Soient les épreuves répétées et indépendantes d'une même expérience de Bernoulli. Chaque expérience n'a que deux résultats possibles : succès ou échec. Comme précédemment,

appelons p la probabilité de l'événement élémentaire succès et $q = 1 - p$ celle de l'événement échec. A cette expérience multiple on associe une variable aléatoire X qui mesure le nombre de succès obtenus.

Distribution de *X*

La probabilité d'avoir k succès lors de n épreuves répétées est

$$P(X = k \text{ pour } n \text{ essais}) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Rappel

$n! = 1 \times 2 \times \ldots \times n$ pour tout n entier positif

$0! = 1$ par définition

Remarques

a. La probabilité de n'avoir aucun succès au cours de n épreuves ($k = 0$) est q^n ; la probabilité d'avoir au moins un succès est donc $1 - q^n$ (un succès ou plus)

b. $\frac{n!}{k!(n-k)!}$ est souvent noté $\binom{n}{k}$ ou C_n^k

Les $\binom{n}{k}$ s'appellent coefficients du binôme.

En effet ils interviennent dans le développement du binôme selon la formule

$$(a+b)^n = \sum_{r=0}^{n} \binom{n}{r} a^{n-r} b^r$$

Exercice :

utiliser cette formule pour vérifier que $(a+b)^4 = a^4 + 4a^3b + 6a^2b^2 + 4ab^3 + b^4$

c. En appliquant la formule du binôme précédente on retrouve que la somme des probabilités pour toutes les valeurs de X est égale à 1 :

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1^n = 1$$

Exemples

1. On jette 6 fois une pièce bien équilibrée ; on suppose que face est un succès. On a donc $p = q = 1/2$ et $n = 6$

   a. Probabilité que l'on ait exactement 2 faces

   $$P(2 \text{ faces parmi } 6 \text{ jets}) = \frac{6!}{2!4!} \cdot \left(\frac{1}{2}\right)^2 \cdot \left(\frac{1}{2}\right)^4 = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6}{1 \times 2 \times 1 \times 2 \times 3 \times 4} \cdot \frac{1}{4} \cdot \frac{1}{16}$$

   $$P(2 \text{ faces parmi } 6 \text{ jets}) = \frac{5 \times 6}{2 \times 4 \times 16} = \frac{15}{4 \times 16} = \frac{15}{64}$$

   b. Probabilité d'avoir 4 faces ou plus (au moins 4 faces)
   C'est aussi la probabilité d'avoir 0, 1 ou 2 piles

$$p_4 = P(4 \text{ faces}) = \frac{6!}{2!4!} \cdot \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^2 = \frac{1\times2\times3\times4\times5\times6}{1\times2\times1\times2\times3\times4} \cdot \frac{1}{16} \cdot \frac{1}{4} = \frac{15}{64}$$

$$p_5 = P(5 \text{ faces}) = \frac{6!}{1!5!} \cdot \left(\frac{1}{2}\right)^5 \cdot \frac{1}{2} = \frac{1\times2\times3\times4\times5\times6}{1\times2\times3\times4\times5} \cdot \frac{1}{32} \cdot \frac{1}{2} = \frac{6}{64}$$

$$p_6 = P(6 \text{ faces}) = \frac{6!}{6!} \cdot \left(\frac{1}{2}\right)^6 \cdot \left(\frac{1}{2}\right)^0 = \frac{1}{64}$$

$$P(\text{au moins 4 faces}) = p_4 + p_5 + p_6 = \frac{15}{64} + \frac{6}{64} + \frac{1}{64} = \frac{11}{32}$$

2. On jette 7 fois un dé équilibré et on considère que tirer 5 ou 6 est un succès. Calculer

   a. la probabilité pour qu'on ait 3 succès exactement

$$P(\text{succès}) = P(\{5,6\}) = \frac{2}{6} = \frac{1}{3}$$

$$P(3 \text{ succès}) = \frac{7!}{3!4!}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^4 = \frac{560}{2187}$$

   b. la probabilité de n'avoir aucun succès

$$P(\text{aucun succès}) = q^7 = \left(\frac{2}{3}\right)^7 = \frac{128}{2187}$$

Propriétés

La fonction de probabilité $P(X=k)$ dépend des 2 paramètres (ou constantes) n et p ; on la note b(k ; n, p). C'est une distribution discrète qui prend les valeurs suivantes :

k	0	1	2	.....	n
$P(X=k)$	q^n	$\binom{n}{1}q^{n-1}p$	$\binom{n}{2}q^{n-2}p^2$		p^n

On dit que X est distribuée selon une loi binomiale B(n, p).
On peut montrer que

Distribution binomiale B($\boldsymbol{n}$, $\boldsymbol{p}$)	
Espérance	$\mu = np$
Variance	$\sigma^2 = npq$
Ecart-type	$\sigma = \sqrt{npq}$

6.2 Lois continues

6.2.1 Loi normale

6.2.1.1 Définition

La distribution normale, ou de Laplace-Gauss, appelée aussi gaussienne, est une distribution continue qui dépend de deux paramètres μ et σ. On la note $N(\mu, \sigma^2)$. Le paramètre μ peut être quelconque mais σ est positif. Cette distribution est définie par :

$$f(x;\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

C'est une des lois les plus importantes, sinon la plus importante comme vous le verrez à l'occasion du théorème central limite.

6.2.1.2 Propriétés

a. La loi normale, notée $N(\mu, \sigma^2)$, est symétrique par rapport à la droite d'abscisse μ.
Exemples :

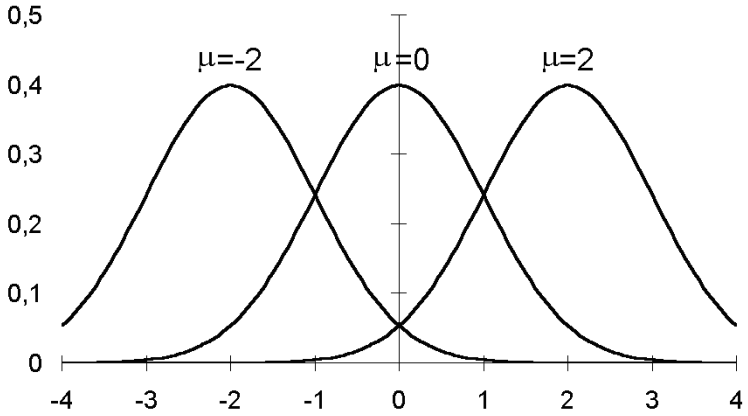


Figure 2 : N(μ, 1) pour les valeurs de μ -2 ; 0 et 2

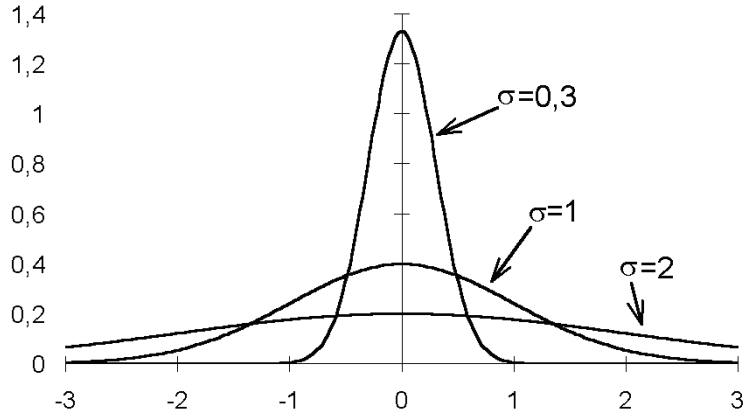


Figure 3 : N(0, σ^2) pour les valeurs de σ 0,3 ; 1 et 2

b. Caractéristiques

Loi normale N(μ, σ^2)	
Espérance	μ
Variance	σ^2
Ecart-type	σ

c. Distribution normale centrée réduite

On dit que la distribution est centrée si son espérance μ est nulle ; elle est dite réduite si sa variance σ^2(et son écart-type σ) est égale à 1. La distribution normale centrée réduite N(0, 1) est donc définie par la formule

$$f(t; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}$$

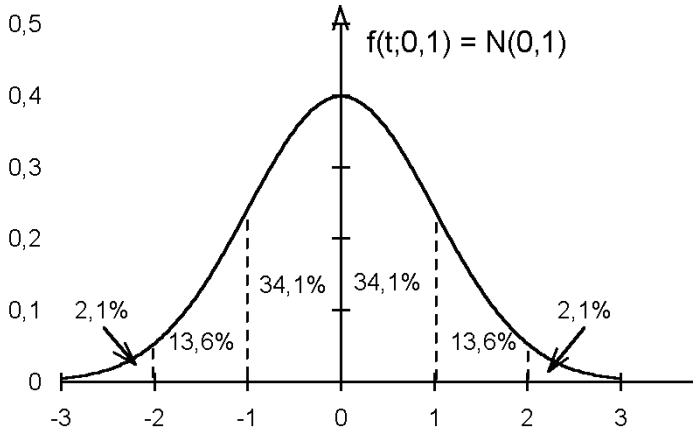


Figure 4 : loi normale centrée réduite N(0, 1)

Les probabilités correspondant aux divers intervalles ont été calculées et regroupées dans une table numérique. Ainsi la table A.1 (en fin de polycopié) permet, à partir d'une probabilité α donnée, de trouver les bornes $-u_\alpha$, $+u_\alpha$ d'un intervalle symétrique autour de 0, tel que $P(t \notin [-u_\alpha ; +u_\alpha]) = \alpha$

ou encore, à partir de u_α, de trouver α.
D'où par exemple :
$P(t \in [-u_\alpha ; +u_\alpha]) = 1 - \alpha$
$P(t > u_\alpha) = P(t < -u_\alpha) = \alpha/2$
On observe ainsi que 68,2 % de la surface est comprise entre (-1 et +1), 95,4 % entre (-2 et +2) et 99,6 % entre (-3 et +3).

d. Transformation en une loi N(0, 1)
Soit une variable X distribuée selon une loi normale d'espérance μ et d'écart-type σ.

Alors la variable $t = \dfrac{X - \mu}{\sigma}$ est distribuée selon une loi normale centrée réduite.

Les probabilités obtenues pour la loi centrée réduite permettent de calculer les probabilités pour une loi normale quelconque, à l'aide de cette transformation :

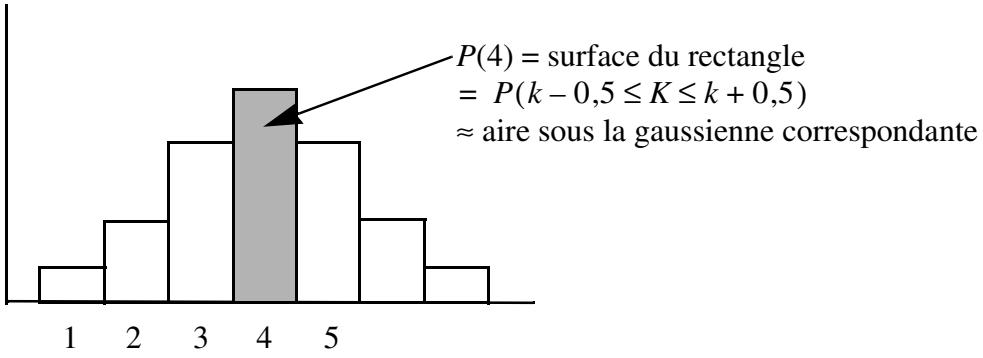
$$t = \frac{X - \mu}{\sigma}.$$

Soit par exemple à calculer $P(a \le X \le b)$. Par la transformation, on a $P(a \le X \le b) = P(c \le t \le d)$ avec

$$c = \frac{a - \mu}{\sigma} \text{ et } d = \frac{b - \mu}{\sigma}.$$

La probabilité cherchée, sur la variable X, revient donc à lire sur la table de la loi centrée réduite (variable t), la probabilité de se trouver entre c et d.
On remarque en particulier que $P(-2 \le t \le 2) = P(\mu - 2\sigma \le X \le \mu + 2\sigma) \approx 0{,}95$

e. Approximation de la distribution binomiale par la loi normale
Lorsque n est grand, et que p et q ne sont pas trop proches de 0 (en pratique si $np \ge 5$ et $nq \ge 5$), alors on constate que la distribution binomiale tend vers la distribution normale de moyenne np et de variance npq ; plus précisément, pour une variable K distribuée selon une loi binomiale $B(n, p)$ et une variable X distribuée selon une loi normale $N(\mu = np, \sigma^2 = npq)$, on a :
$P(K = k) = P(k) = P(k - 0{,}5 \le K \le k + 0{,}5) \approx P(k - 0{,}5 \le X \le k + 0{,}5)$
On choisit l'artifice de représenter graphiquement $P(k)$ par un rectangle dont la base est $[k - 0{,}5, k + 0{,}5]$ et la surface est $P(k)$ pour comparer la loi discrète $P(k)$ et la loi normale continue.



6.2.2 Loi du χ^2 (chi-2)

6.2.2.1 Définition

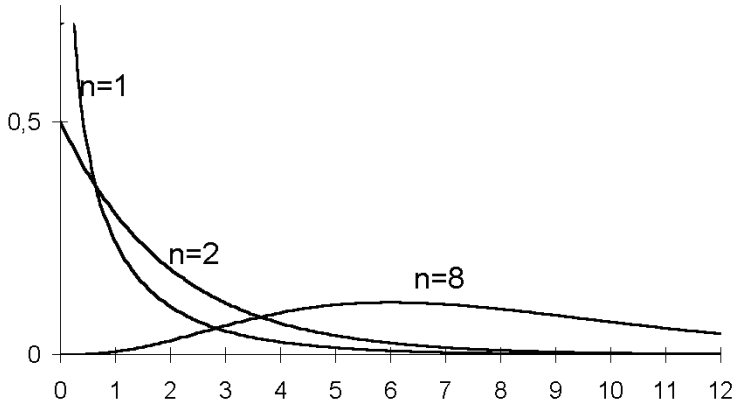
C'est une loi dérivée de la loi normale, très importante pour ses applications en statistiques comme nous le reverrons dans les tests.
Soient $X_1, ..., X_n$ des variables aléatoires indépendantes, chacune étant distribuée selon une loi normale centrée réduite :
$\forall i, X_i \sim \mathrm{N}(0, 1)$

La distribution de $S = X_1^2 + X_2^2 + \ldots + X_n^2$ (somme des carrés des X_i) est appelée loi de χ^2 à n degrés de liberté (en abrégé d. d. l.), que l'on note $\chi^2(n)$ où n est le nombre de d. d. l., seul paramètre de la loi.

Loi du $\chi^2(n)$	
Espérance	n
Variance	$2n$
Ecart-type	$\sqrt{2n}$

6.2.2.2 Propriétés

a. Allure de la distribution de $\chi^2(n)$ pour différentes valeurs de n



Pour $n = 1$, la courbe décroît de $+\infty$ vers zéro de façon monotone ; pour $n = 2$, la courbe décroît de façon monotone de 0,5 à zéro ; pour $n > 2$, la courbe part de 0, a son maximum pour $x = n - 2$, puis redescend vers zéro.

b. Propriété asymptotique
La loi d'une variable X suivant un $\chi^2(n)$ tend vers une loi normale lorsque $n \to +\infty$. On a donc, après avoir centré et réduit cette variable :

$$\frac{X-n}{\sqrt{2n}} \sim \mathrm{N}(0, 1)$$

NB : Dans la pratique, on utilise plutôt la variable $Y = \sqrt{2X} - \sqrt{2n-1}$ dont on montre qu'elle est à peu près distribuée selon une loi normale centrée réduite dès que $n > 30$.

c. Tables
De même que pour la loi normale centrée réduite, une table existe pour la loi du χ^2 (voir en fin de polycopié, table A.3). Cette table indique pour une probabilité α donnée, et un degré de liberté n donné, la valeur χ^2_α telle que $P(X > \chi^2_\alpha) = \alpha$.

6.2.3 Loi de Student

Il s'agit encore d'une loi dérivée de la loi normale, très utilisée dans les tests statistiques. On considère une première variable aléatoire X, distribuée selon une loi normale centrée réduite, puis une seconde variable Y, indépendante de X, distribuée selon un χ^2 à n degrés de liberté.

Alors la variable aléatoire $Z = \sqrt{n}\frac{X}{\sqrt{Y}}$ est distribuée selon une loi de Student à n degrés de liberté, notée t(n).
La courbe correspondante est symétrique autour de 0, et son allure est proche de celle de la loi normale.

Cette loi est centrée, mais non réduite : la variance, $\frac{n}{n-2}$, est supérieure à 1.

Lorsque n croît, en pratique pour $n > 30$, la variance peut être prise égale à 1, et la distribution as-

Loi de Student t(n)	
Espérance	0
Variance	$\frac{n}{n-2}$
Ecart-type	$\sqrt{\frac{n}{n-2}}$

similée à celle d'une loi normale centrée réduite.
La table A.2, en fin de polycopié, indique, pour une probabilité α donnée, et un degré de liberté n donné, la valeur t_α telle que $P(Z \notin [-t_\alpha ; +t_\alpha]) = \alpha$.

6.2.4 Loi exponentielle

Cette loi décrit par exemple le processus de mortalité dans le cas où le « risque instantané » de décès est constant. La loi correspondante est :

$f(x) = \lambda e^{-\lambda x}$ avec $\lambda > 0$ et $x \geq 0$
où x est la durée de vie.

Loi exponentielle	
Espérance	$1/\lambda$
Variance	$1/\lambda^2$
Ecart-type	$1/\lambda$

AVERTISSEMENT
On peut entreprendre dès maintenant la lecture des chapitres 15 et 16 en comprenant le terme « estimateur » dans le sens intuitif de « valeur approchée »

Chapitre 7

Statistiques descriptives

Les statistiques descriptives visent à représenter des données dont on veut connaître les principales caractéristiques quantifiant leur variabilité.

7.1 Rappels et compléments

On suppose que l'on s'intéresse à une caractéristique particulière observable chez des individus issus d'une population ; cette caractéristique sera appelée variable ; si cette caractéristique peut varier entre les individus, on l'appellera **variable aléatoire**. On s'intéresse donc à une variable aléatoire. Cette définition imagée est compatible avec la définition du chapitre 5.

Rappel

Il existe deux grands groupes de variables :

a. Les variables **quantitatives** qui sont des variables ordonnées, productives de nombres. Exemples : nombre d'enfants dans une famille, glycémie, taille d'un individu, nombre de colonies bactériennes dans un milieu de culture.

   Parmi ces variables quantitatives, certaines prennent un continuum de valeurs (entre deux valeurs possibles, il existe toujours une troisième valeur possible) ; ces variables sont dites **continues**. D'autres ne prennent que des valeurs discontinues ; elles sont dites **discrètes**, finies ou non.

b. Les variables **qualitatives** produisant des valeurs non numériques. Exemples : sexe, couleur des cheveux, appartenance au groupe des fumeurs ou des non fumeurs, présence ou absence d'une maladie.

   Les valeurs peuvent être ordonnées ; on parle alors de variable qualitative ordinale. Exemple : intensité d'une douleur (faible, moyenne, forte).

Remarque

L'individu évoqué ci-dessus, sur lequel on observe les caractéristiques d'intérêt, la variable, n'est pas nécessairement un individu physique. C'est l'entité sur laquelle s'opère l'observation de la variable d'intérêt. Exemples : famille, colonies bactériennes. Cette entité s'appelle l'unité statistique.

Définition

L'entité sur laquelle peut s'observer la variable aléatoire s'appelle l'**unité statistique**.
Connaître le phénomène mettant en jeu cette variable, ou connaître cette variable, c'est connaître la probabilité pour qu'un individu tiré au hasard dans la population présente telle va-

leur de la variable. On peut apprécier la probabilité d'un événement aléatoire grâce à l'interprétation suivante de la notion de probabilité. Cette interprétation est cohérente avec les cours précédents.
On **interprétera** la probabilité d'un événement aléatoire comme la valeur limite de la fréquence avec laquelle l'événement se réalise au cours d'un nombre **croissant** de répétitions de l'expérience. Autrement dit comme la valeur limite du rapport du nombre de fois où l'événement s'est réalisé et du nombre de répétitions de l'expérience.

Remarques

- Ce qui précède peut être vu comme une interprétation de la notion probabilité (voire comme une définition).
- En dépit de cette interprétation, la probabilité d'un événement aléatoire reste
  - — une fiction
  - — du domaine théorique.

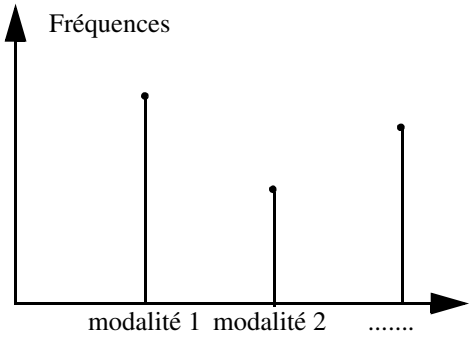
  Mais cette interprétation a deux conséquences :
  - — pour approcher une probabilité on est amené à répéter une expérience,
  - — les fréquences se substituent aux probabilités ; elles seront les contreparties des probabilités.

  On va donc répéter une expérience un nombre fini de fois, noté n ; on aura donc observé une sous-population appelée **échantillon**. Chaque expérience produit un résultat x_i ; on disposera donc de x_1, ..., x_n, ensemble appelé **échantillon** de valeurs de la variable.

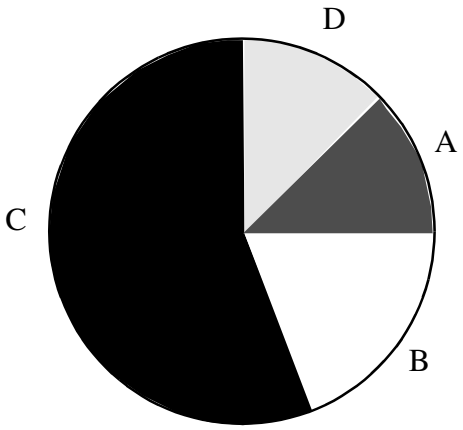
7.2 Représentation complète d'une série d'expériences

7.2.1 Cas d'une variable qualitative

La variable est décrite par la suite des probabilités des différentes modalités. Si l'on connaissait ces probabilités, on produirait le diagramme en bâtons (ou répartition vraie) de cette variable ; on va produire la **répartition expérimentale** par substitution aux probabilités inconnues des fréquences observées. Si la variable est ordinale, on respectera cet ordre dans l'énumération des modalités portées en abscisses.

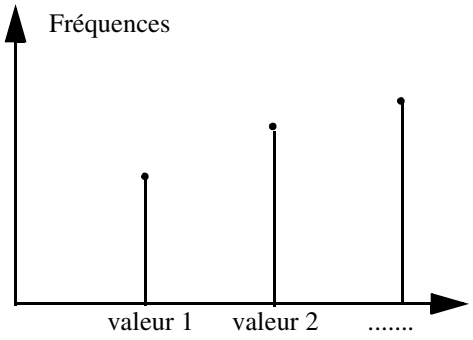


Un autre type de représentation est utilisé : la représentation en camembert où les différentes modalités sont représentées par secteurs angulaires d'angles au centre proportionnels aux fréquences observées.



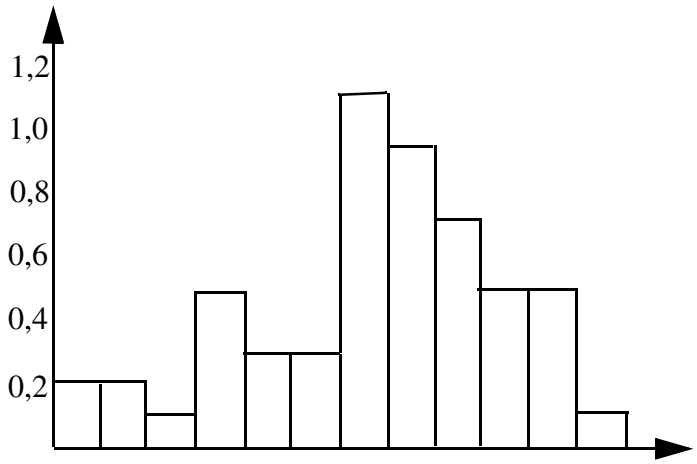
7.2.2 Cas d'une variable quantitative discrète

La situation est similaire si ce n'est qu'il existe un ordre et une échelle naturels en abscisses ; la répartition expérimentale se nomme également histogramme en bâtons.



7.2.3 Cas d'une variable quantitative continue. Notion d'HISTOGRAMME

Dans le cas de variables continues, on va choisir de représenter les données graphiquement d'une façon qui soit proche de la représentation d'une densité de probabilité d'une variable aléatoire continue. Pour cela on découpe l'ensemble du domaine des valeurs possibles de la variable étudiée en un nombre choisi d'intervalles contigus dont on choisit également les bornes. Afin d'obtenir une représentation proche d'une densité de probabilité, on décide de représenter indirectement la fréquence des valeurs observées comprises entre deux bornes consécutives par la surface d'un rectangle dont la base sera précisément cet intervalle. Autrement dit la hauteur de ce rectangle sera le rapport de la fréquence expérimentale de ces valeurs et de la différence entre ces bornes (différence également appelée largeur de la classe).



Les bornes sont choisies arbitrairement ; néanmoins, pour que l'histogramme ait un sens il est nécessaire que la taille de chaque classe constituant un intervalle comprenne un nombre suffisamment grand de valeurs observées, de telle façon que la surface d'un rectangle élémentaire puisse être interprétée comme approchant la probabilité pour que la variable prenne une valeur comprise dans l'intervalle du rectangle. Si la taille de l'échantillon croît, la surface de chaque rectangle tend vers la probabilité que la variable ait une valeur incluse dans l'intervalle correspondant. De plus, si la taille n de l'échantillon est grande, on peut alors sans inconvénient construire un plus grand nombre de classes, c'est-à-dire construire par exemple deux fois plus de rectangles, chacun ayant un support deux fois plus petit. En répétant cette opération, n croissant, on peut comprendre que l'histogramme tend (d'une façon que nous ne préciserons pas ici) vers la densité de probabilité de la loi qui a généré l'échantillon.

7.3 Représentation simplifiée d'une série d'expériences

On a défini certains indicateurs pour représenter, de façon plus résumée que ci-dessus, un échantillon de valeurs issues d'une variable aléatoire.

7.3.1 Indicateurs de localisation des valeurs

Mode

Le mode d'un échantillon est sa valeur la plus fréquente. Si la variable est qualitative, par valeur il faut entendre modalité. Si la variable est continue, par valeur il faut entendre « petit intervalle de valeurs », et la définition du mode perd alors de sa rigueur.

Les autres indicateurs présentés ci-dessous ne concernent que les variables quantitatives.

Médiane

C'est la valeur qui partage l'échantillon en deux groupes de même effectif ; pour la calculer, il faut commencer par ordonner les valeurs (les ranger par ordre croissant par exemple)
Exemple : soit la série 12 3 24 1 5 8 7
on l'ordonne : 1 3 5 **7** 8 12 24
7 est la médiane de la série

Moyenne expérimentale

C'est l'indicateur de localisation le plus fréquemment utilisé. La moyenne expérimentale d'un échantillon de n valeurs x_1, ..., x_n est définie comme la moyenne arithmétique de ces valeurs ; on la note souvent $\bar{x}$:

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

7.3.2 Indicateurs de dispersion des valeurs

Variance expérimentale

La variance expérimentale d'un échantillon $\{x_i\}$ $i = 1, ..., n$ est donnée par

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

Attention : on divise par n -1 et non par n - ceci pour que la variance expérimentale soit un bon estimateur de la variance théorique de la loi (nous reverrons ce point dans la suite).
Une autre expression de s^2, équivalente, est indiquée dans le résumé de ce chapitre.

Ecart-type expérimental

L'écart-type expérimental, noté s, est défini par $s = \sqrt{s^2}$.

7.4 Reformulation de la moyenne et de la variance expérimentales

7.4.1 Reformulation de la moyenne expérimentale

Prenons le cas d'une variable quantitative discrète.
Les données sont notées x_1, ..., x_n.
Les k valeurs possibles de la variable sont notées val_1,val_2, ..., val_k.

Exemple d'un jet de dé : $\text{val}_1 = 1$, ..., $\text{val}_6 = 6$

Chaque donnée x_i coïncide avec une certaine valeur val_j

Par exemple pour le jet de dé, on peut avoir

- jet n°1 ; $x_1 = 1 = \text{val}_1$
- jet n°2 ; $x_2 = 1 = \text{val}_1$
- jet n°3 ; $x_3 = 4 = \text{val}_4$
- jet n°4 ; $x_4 = 3 = \text{val}_3$
- jet n°5 ; $x_5 = 6 = \text{val}_6$
- jet n°6 ; $x_6 = 1 = \text{val}_1$
- jet n°7 ; $x_7 = 2 = \text{val}_2$
- jet n°8 ; $x_8 = 5 = \text{val}_5$
- jet n°9 ; $x_9 = 6 = \text{val}_6$

Alors : $\sum_{i=1}^{n} x_i = \sum_{j=1}^{k} n_j \text{val}_j$
où n_j est le nombre de fois où une observation coïncide avec val_j

Dans notre exemple du jet de dé, on a : $n_1 = 3$, $n_2 = 1$, $n_3 = 1$, $n_4 = 1$, $n_5 = 1$, $n_6 = 2$

Finalement $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \sum_{j=1}^{k} \frac{n_j}{n} \text{val}_j$
Mais $\frac{n_j}{n}$ est une approximation de $P(\text{face marquée} = \text{val}_j)$

Ainsi $\bar{x}$ est une estimation expérimentale - une appréciation - de :

$\sum_j \text{val}_j P(\text{valeur de la variable} = \text{val}_j)$
c'est-à-dire une appréciation de l'espérance mathématique de la variable.

On raccorde ainsi une moyenne expérimentale à une grandeur descriptive du phénomène étudié, à

une grandeur dite « théorique » ou vraie.
On peut dire ceci : la répétition des expériences vise à estimer P(valeur de la variable = certain niveau). La moyenne expérimentale permet d'estimer quelque chose de plus grossier, une combinaison de toutes ces probabilités, précisément l'espérance mathématique
$\mu = \sum_j \text{val}_j P(\text{valeur de la variable} = \text{val}_j)$
C'est la raison pour laquelle dans la suite on utilisera également la terminologie **MOYENNE VRAIE** ou **MOYENNE THEORIQUE** de la variable pour parler de l'espérance mathématique.

Retenons :

ESPERANCE MATHEMATIQUE,
MOYENNE VRAIE,
MOYENNE THEORIQUE
sont SYNONYMES. Ce sont des grandeurs théoriques.

Remarque

La même analyse peut être faite - mais l'expression est un peu plus délicate - dans le cas d'une variable quantitative continue. La moyenne expérimentale approxime là encore l'espérance mathématique.

7.4.2 Reformulation de la variance expérimentale

De la même façon on peut obtenir le résultat suivant : s^2 est une approximation de la grandeur
$\sigma^2 = \sum_j (\text{val}_j - \mu)^2 P(\text{valeur de la variable} = \text{val}_j)$
Cette expression, introduite dans le chapitre 5 sous le nom de variance sera souvent dénommée dans la suite **VARIANCE VRAIE** ou **VARIANCE THEORIQUE** de la variable.

Dans le cas d'une variable continue, la variance expérimentale s^2 approxime :
$\sigma^2 = \int_{\Re} (x-\mu)^2 f(x)\,dx$

CES NOTIONS DE MOYENNE ET VARIANCE VRAIES, DE MOYENNE ET VARIANCE EXPERIMENTALES SONT **ESSENTIELLES** ; NOUS ENGAGEONS LE LECTEUR A BIEN LES COMPRENDRE AVANT DE POURSUIVRE.

7.5 Cas particulier d'une variable à deux modalités - Proportion

On est très souvent amené à considérer des variables à deux modalités, c'est-à-dire des expériences

aléatoires à deux événements.
Exemples :

- maladie : maladie présente - maladie absente
- signe clinique : présent - absent
- traitement : individu traité - individu non traité

Or on peut transformer une telle variable en variable quantitative, sans restriction de généralité, par un artifice de codage :

- une des modalités est codée avec la valeur numérique 0 ;
- l'autre modalité est codée avec la valeur numérique 1.

Une telle variable s'appelle variable de **Bernoulli**.

Notons X cette variable.
Elle est complètement décrite par la donnée de P(valeur de la variable = 1) car
P(valeur de la variable = 1) + P(valeur de la variable = 0) = 1.
On utilise la notation conventionnelle suivante : P(valeur de la variable = 1) SE NOTE p.

7.5.1 Expression de la moyenne vraie de *X*

Utilisant l'expression générale de la moyenne vraie, et remarquant que $\text{val}_1 = 0$, $\text{val}_2 = 1$, on obtient :

$$\mu = \sum_j \text{val}_j P(\text{valeur de la variable} = \text{val}_j) = 0 \times (1-p) + 1 \times p = p$$

Ainsi, $\mu = p = P$(valeur de la variable = 1) = probabilité de la modalité codée 1 = PROPORTION VRAIE des individus présentant la modalité 1.

7.5.2 Expression de la variance vraie de *X*

$$\sigma^2 = \sum_j (\text{val}_j - \mu)^2 P(\text{valeur de la variable} = \text{val}_j) = (0-p)^2(1-p) + (1-p)^2 p = p(1-p)$$

7.5.3 Interprétation de la moyenne expérimentale

$$\bar{x} = \frac{1}{n}\sum_i x_i = \frac{1}{n}[0 + 0 + 1 + 0 + 1 + 1 + \ldots] = \frac{\text{nombre de fois où } X = 1}{n}$$

Ainsi, $\bar{x}$ coïncide avec la fréquence expérimentale de la modalité codée 1. Cette fréquence sera notée p_0 et s'appelle de façon naturelle PROPORTION OBSERVEE d'individus présentant la modalité 1.

Exemple

Dans le cas de l'étude d'un signe clinique, en codant 1 la présence du signe clinique, $\bar{x}$ (donc p_0) sera la fréquence expérimentale de la présence du signe ou encore le pourcentage des individus présentant le signe (à un facteur 100 près).

En résumé

- si X est une variable de Bernoulli,
  - — sa moyenne vraie = p
  - — sa variance vraie = $p(1 - p)$
- UNE PROPORTION OBSERVEE EST UNE MOYENNE EXPERIMENTALE.

7.6 Conclusion : la variable aléatoire moyenne expérimentale

On a jusqu'ici associé une valeur de moyenne expérimentale à une série de n réalisations d'une variable aléatoire quantitative X. Mais chaque expérience consistant à recueillir n réalisations de la variable X permet de calculer une valeur, différente à chaque expérience, de moyenne expérimentale. Autrement dit, la moyenne expérimentale doit être vue comme une nouvelle variable aléatoire ; on la notera $\bar{X}$. Dans certains cas, afin de rappeler que cette variable dépend de n, on notera $\bar{X}_n$ la variable construite à partir de n réalisations de X. On utilisera la terminologie suivante :

on dira que $\bar{X}$ (ou $\bar{X}_n$ si nécessaire) est la VARIABLE ALEATOIRE MOYENNE EXPERIMENTALE DEDUITE DE LA VARIABLE ALEATOIRE X, FONDEE SUR n REPETITIONS

ou, de façon équivalente que $\bar{X}$ (ou $\bar{X}_n$ si nécessaire) est la VARIABLE ALEATOIRE MOYENNE EXPERIMENTALE ASSOCIEE A LA VARIABLE ALEATOIRE X, FONDEE SUR n REPETITIONS

Remarque

Dans le cas où X est une variable de Bernoulli, $\bar{X}_n$ sera noté p_{0n} (et $\bar{X}$ simplement p_0). Il s'agit d'une proportion observée dont on connaît déjà pratiquement la distribution puisque $np_{0n} \sim \mathrm{B}(n, p)$ (voir section 6.1.2 page 47).

Résumé du chapitre

1. **Une variable aléatoire** est une variable observable au cours d'une expérience et dont la valeur peut varier d'une expérience à l'autre de façon non prévisible.
2. **Représentation d'une variable**

	répartition d'un échantillon	**représentation de la population**
variable qualitative	répartition expérimentale	répartition vraie
variable quantitative discrète	histogramme en bâtons	répartition vraie
variable quantitative continue	histogramme	densité de probabilité

3. **Moyennes (variables quantitatives + variables de Bernoulli)**

	moyenne expérimentale	**moyenne vraie**
variable discrète	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\mu = \sum_{j=1}^{k} \text{val}_j P(\text{variable} = \text{val}_j)$
variable continue	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\mu = \int_{\Re} x f(x) dx$
variable de Bernoulli	$\bar{x}$ est notée p_0	$\mu = P(\text{variable} = 1)$ est notée p

4. **Variances (variables quantitatives)**

	variances expérimentales	**variances vraies**
variable discrète	$s^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right]$	$\sigma^2 = \sum_{j=1}^{k} (\text{val}_j - \mu)^2 P(\text{variable} = \text{val}_j)$
variable continue	$s^2 = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right]$	$\sigma^2 = \int_{\Re} (x-\mu)^2 f(x) dx$

5. **Variables centrée et centrée réduite associées à une variable *X***
   Si X est une variable aléatoire de moyenne μ et de variance σ^2,
   - la variable $(X - \mu)$ est dite variable centrée associée à X,
   - la variable $\frac{X-\mu}{\sigma}$ est dite variable centrée réduite associée à X.

Chapitre 8

Fluctuations de la moyenne expérimentale : la variable aléatoire moyenne expérimentale

On conserve le contexte d'étude du chapitre précédent, c'est-à-dire l'examen de la variabilité d'une grandeur (variable aléatoire) dans une population d'individus ou unités statistiques. Mais on s'intéresse ici à la variable aléatoire « moyenne expérimentale ».

8.1 Première propriété de la moyenne expérimentale

8.1.1 Un exemple

Prenons à nouveau le cas d'une variable discrète pouvant prendre les deux valeurs 0 et 1 [c'est-à-dire variable associée à présence-absence ou oui-non]. Supposons que l'on ait des raisons de penser que P(X = 0) = P(X = 1) = 1/2. On a vu qu'une telle variable a pour moyenne vraie 1/2, pour variance vraie 1/4.
On peut, par le calcul, pronostiquer le résultat d'une répétition d'expériences. En particulier, calculer la répartition de la variable « moyenne expérimentale calculée sur un échantillon de deux individus », notée $\overline{X}_2$, ici deux lancers de pièce.
On isole cette variable. Quelles valeurs peut-elle prendre, avec quelles probabilités ?

jet 1 : résultats	Proba	jet 2 : résultats	Proba	Proba jet1, jet2	$\bar{X}_2$
0	1/2	0	1/2	1/4	1/2(0+0) = 0
0	1/2	1	1/2	1/4	1/2(0+1) = 1/2
1	1/2	0	1/2	1/4	1/2(1+0) = 1/2
1	1/2	1	1/2	1/4	1/2(1+1) = 1

Ainsi, $P(\bar{X}_2 = 0) = \frac{1}{4}$, $P\left(\bar{X}_2 = \frac{1}{2}\right) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, $P(\bar{X}_2 = 1) = \frac{1}{4}$
Alors :

- moyenne vraie de $\bar{X}_2 = 0 \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{4} = \frac{1}{2} =$ moyenne vraie de X
- variance vraie de $\bar{X}_2 = \left(0 - \frac{1}{2}\right)^2 \times \frac{1}{4} + \left(\frac{1}{2} - \frac{1}{2}\right)^2 \times \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \times \frac{1}{4} = \frac{1}{8} = \frac{1}{2} \times \frac{1}{4}$

Ainsi la variance vraie de la moyenne expérimentale est plus faible que la variance vraie de la variable d'origine (la moitié ici). La moyenne vraie reste inchangée. Et ainsi vont les choses si la taille des échantillons (ici 2) qui constituent les unités statistiques augmente. La dispersion de $\bar{X}$ diminue au fur et à mesure que $\bar{X}$ se trouve calculée sur la base d'un échantillon de taille croissante. Le « comment » de cette situation peut être résumé ainsi : les valeurs de la moyenne expérimentale deviennent de plus en plus probables dans un voisinage de la moyenne vraie car le nombre de situations pouvant donner une valeur expérimentale proche de la moyenne vraie augmente dans ce voisinage. Cela est dû au fait que la moyenne vraie est « au milieu » des valeurs possibles. On le voit sur l'exemple ci-dessus où la moyenne vraie est obtenue dans les deux cas (0, 1) et (1, 0). C'est encore plus perceptible sur l'exemple d'un dé. Pour que la moyenne expérimentale calculée sur deux jets de dé soit 6, il faut obtenir le résultat (6, 6) ; pour qu'elle soit 3, il faut un total de 6, c'est-à-dire (5, 1), (4, 2), (3, 3), (2, 4), (1, 5), soit un événement 5 fois plus probable.
Il est possible de quantifier tout cela. On peut généraliser ce qui a été obtenu avec deux jets de pièces et on obtient, quelle que soit la distribution de la variable étudiée - qu'elle soit continue ou discrète - les résultats fondamentaux suivants.

8.1.2 Généralisation

i. La moyenne vraie de la moyenne expérimentale calculée sur un échantillon de taille n coïncide avec la moyenne vraie de la variable étudiée, ce que l'on peut résumer par :

$$\mu(\overline{X_n}) = \mu(X)$$

ii. La variance vraie de la moyenne expérimentale calculée sur un échantillon de taille n est égale à la variance vraie de la variable **DIVISEE PAR** n, ce que l'on peut résumer par :

$$\sigma^2(\overline{X_n}) = \frac{1}{n}\sigma^2(X)$$

d'où la relation entre écarts-types :

$$\sigma(\overline{X_n}) = \frac{1}{\sqrt{n}}\sigma(X)$$

8.2 Seconde propriété de la moyenne expérimentale : le théorème central limite

On souhaiterait comparer, par curiosité, les distributions de plusieurs moyennes expérimentales, correspondant à diverses variables aléatoires. Par exemple la taille, la glycémie. Ces distributions sont différentes, ne serait-ce qu'à cause des différences entre moyennes et variances vraies. Pour s'abstraire de ces premières différences, considérons la variable centrée réduite associée, soit pour chaque variable considérée :

$$\frac{\overline{X}_n - \mu(\overline{X}_n)}{\sigma(\overline{X}_n)} \text{ soit } \frac{\overline{X}_n - \mu(X)}{\frac{\sigma(X)}{\sqrt{n}}}$$

Maintenant toutes ces variables ont en commun leur moyenne vraie (0) et leur variance (1). Il se passe quelque chose d'extraordinaire : lorsque n est suffisamment grand, elles finissent par avoir en commun leur distribution, leur densité de probabilité.
Cela signifie que les distributions de toutes ces variables (moyennes expérimentales centrées réduites issues de variables aléatoires différentes) finissent par coïncider, lorsque n est suffisamment grand, avec une distribution particulière unique. Cette distribution s'appelle **LOI NORMALE**, et puisque sa moyenne vraie est nulle et sa variance vraie est 1, on l'appelle LOI NORMALE CENTREE REDUITE ou encore distribution de Gauss ou de Laplace-Gauss (1800). On la notera schématiquement N(0, 1) où 0 rappelle la valeur de la moyenne vraie, 1 la valeur de la variance vraie.
Donc la propriété ci-dessus - connue sous le nom de théorème central limite - s'énonce :

THEOREME CENTRAL LIMITE
Soit X une variable aléatoire quantitative d'espérance mathématique μ, de variance vraie σ^2. Soit $\overline{X}_n$ la variable aléatoire moyenne expérimentale associée à X construite sur n répétitions.

La distribution limite de la variable aléatoire $\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ est la distribution normale centrée réduite notée N(0,1).

Il faut bien mesurer la portée de cette propriété. Quel que soit le phénomène étudié - apprécié par la variable aléatoire que l'on étudie - il suffit de connaître la moyenne et la variance de la variable pour déduire la **distribution** (la densité de probabilité) - c'est-à-dire l'expression la plus achevée des propriétés de variabilité - de la moyenne expérimentale calculée sur un échantillon de taille suffisante. Nous reviendrons plus loin sur cette notion vague « taille suffisante ». Or c'est peu de connaître moyenne, variance (ou écart-type) seulement - ex. : pour le poids à la naissance $\mu = 3$ kg, $\sigma = 1{,}2$ kg.

8.3 Etude de la distribution normale (rappel)

La distribution limite que l'on a mise en évidence dépeint une variable aléatoire de moyenne vraie 0, de variance vraie 1, que l'on a appelée distribution normale centrée réduite ou N(0, 1).

C'est une fonction dont l'équation est $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ et l'allure est représentée sur la figure 5.

Ses principales caractéristiques morphologiques sont les suivantes :

- elle est symétrique,
- elle présente deux points d'inflexion en $x = 1$ et $x = -1$

Par ailleurs, pour faciliter les calculs de probabilité relatifs à cette variable, des tables ont été construites qui donnent le lien entre α et u_α, où ces valeurs ont le sens suivant (voir figure 5) :
$P(X \notin [-u_\alpha ; +u_\alpha]) = \alpha$
En particulier, pour $\alpha = 0{,}05$, la valeur u_α lue dans la table est 1,96, d'où $u_{0,05} = 1{,}96$
On peut voir facilement que toute probabilité $P(X \in [a,b])$ s'obtient à partir d'une telle table, quelles que soient les valeurs de a et b.

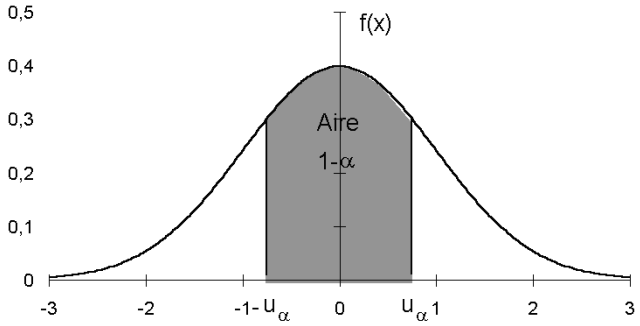


Figure 5 : loi normale centrée réduite

Remarque

Sur la base de cette loi centrée réduite, on définit toute une famille de lois de la façon suivante :

Si X est distribuée selon une loi normale centrée réduite (notation $X \sim N(0, 1)$),
alors la variable $Y = \sigma X + \mu$, dont la moyenne vraie est μ et la variance vraie σ^2, est distribuée selon une loi normale de moyenne μ et de variance σ^2.
On écrit $Y \sim N(\mu, \sigma^2)$

A l'inverse, si on dit que $X \sim N(\mu, \sigma^2)$

cela veut dire que $\frac{X-\mu}{\sigma} \sim N(0, 1)$ (variable centrée réduite associée).

Exemple

La figure 6. présente l'aspect de deux distributions normales l'une N(0, 1), l'autre N(2,9 , 4).

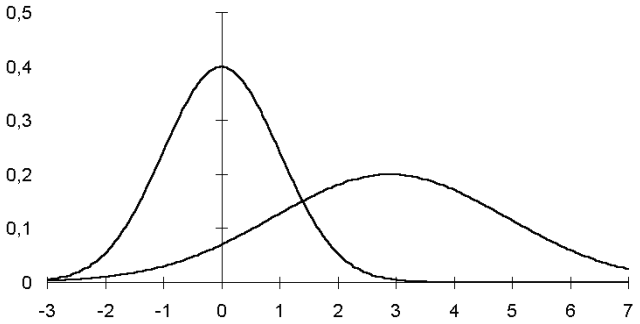


Figure 6 : exemple de lois normales

Résumé et précisions (théorème central limite)

Si n est suffisamment grand, X ayant pour moyenne vraie μ, pour variance vraie σ^2, alors :

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \text{ (à peu près)}$$

ou, de façon équivalente, $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ (à peu près)

où la notation ~ se lit : « est distribué comme » ou « suit une distribution ».

a. Cette propriété est exacte quel que soit n si X elle-même est gaussienne (i.e. est distribuée normalement).

b. si X n'est pas gaussienne :

- si X est continue, la condition de validité usuelle est **$n \geq 30$**
- si X est une variable de Bernoulli (valeurs 0 et 1), la condition usuelle de validité est

$$\begin{cases} np \geq 5 \text{ et} \\ n(1-p) \geq 5 \end{cases}$$

En outre dans ce cas, $\mu = p$, $\sigma^2 = p\ (1 - p)$ si bien que l'on aura :

$$\frac{p_{0n} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1) \text{ (à peu près)}$$

ou, de façon équivalente, $p_{0n} \sim N\left(p, \frac{p(1-p)}{n}\right)$ (à peu près)

8.4 Application du théorème central limite. Intervalle de Pari (I. P.)

8.4.1 Définition de l'intervalle de pari (I. P.) d'une moyenne expérimentale

On considère une variable aléatoire de moyenne vraie μ et de variance vraie σ^2.
On sait que pour n grand ($n \geq 30$, ou np et $n(1 - p) \geq 5$) :

la variable $u = \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ est approximativement distribuée selon N (0, 1).

On se pose le problème suivant. On s'apprête à réaliser une série d'expériences, c'est-à-dire à mesurer la variable X sur un échantillon de n individus. Peut-on construire un intervalle $[a, b]$ tel que la probabilité pour que la moyenne expérimentale que l'on s'apprête à calculer appartienne à cet intervalle, ait un niveau donné ? Il s'agit donc de construire un intervalle qui contienne avec une probabilité fixée la valeur expérimentale que l'on va obtenir.

Il s'agit donc de trouver deux valeurs a et b telles que $P(a \leq \bar{X}_n \leq b) = \text{valeur donnée} = 1 - \alpha$.

Exemple : $P(a \leq \bar{X}_n \leq b) = 0,95$

Un tel intervalle $[a, b]$ s'appelle **INTERVALLE DE PARI (I. P.)** de niveau 1 - α, ou encore intervalle de pari au risque α.

La figure 7 illustre le problème posé.

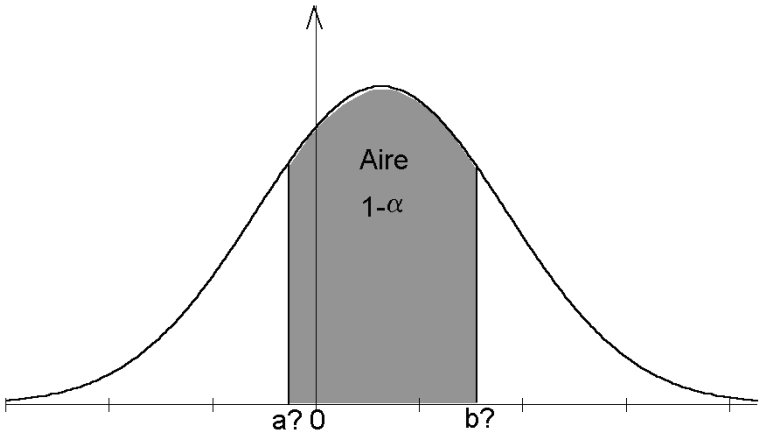


Figure 7 : le problème de l'intervalle de pari

Ce problème admet plusieurs solutions : on choisit généralement un intervalle symétrique autour de μ.
Résolution : $a = \mu - \lambda\frac{\sigma}{\sqrt{n}}$ et $b = \mu + \lambda\frac{\sigma}{\sqrt{n}}$
La valeur λ inconnue doit vérifier :

$$P\left(\mu - \lambda\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + \lambda\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-\lambda\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n - \mu \leq \lambda\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$P\left(-\lambda \leq \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \lambda\right) = 1 - \alpha$$

Si le théorème central limite s'applique, l'expression ci-dessus suit une loi N(0, 1) ; notons-la u. Alors λ doit vérifier $P(-\lambda \leq u \leq \lambda) = 1 - \alpha$. C'est le u_α de la table.
Finalement : $\lambda = u_\alpha$

$$P\left(\mu - u_\alpha\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + u_\alpha\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \text{ et}$$

$$IP_{1-\alpha} = \left[\mu - u_\alpha \frac{\sigma}{\sqrt{n}} ; \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

Intervalle de Pari (I. P.) de la moyenne expérimentale d'une variable de moyenne vraie μ, de variance vraie σ^2 construite sur un échantillon de taille n

Exemple : $\alpha = 0,05$ $u_\alpha = 1,96$ $IP_{0,95} = \left[\mu - 1,96 \frac{\sigma}{\sqrt{n}} ; \mu + 1,96 \frac{\sigma}{\sqrt{n}}\right]$

Les conditions de validité de cette construction sont celles du théorème central limite, c'est-à-dire $n \geq 30$ pour les variables continues non normales et $np, n(1 - p) \geq 5$ pour les variables de Bernoulli.

Cas d'une variable de Bernoulli : μ est notée p, $\sigma^2 = p\ (1 - p)$. Donc

$$IP_{0,95} = \left[p - 1,96 \sqrt{\frac{p(1-p)}{n}} ; p + 1,96 \sqrt{\frac{p(1-p)}{n}}\right]$$

L'interprétation de l'intervalle de pari est fondamentale. Si cet intervalle est bien calculé, on est sûr, avec une probabilité 1 - α, d'obtenir une valeur de la moyenne expérimentale comprise dans cet intervalle. En pariant que la valeur va tomber dans cet intervalle, on se trompera dans 5 pour cent des expériences.

Exemple :
On a des raisons de penser que la fréquence d'une maladie dans la population est $p = 0,2$. L'intervalle de pari de la moyenne expérimentale (proportion observée) calculée sur 64 individus au niveau 0,95 est :

$$IP_{0,95} = \left[0,2 - \frac{1,96\sqrt{0,2(1-0,2)}}{\sqrt{64}} ; 0,2 + \frac{1,96\sqrt{0,2(1-0,2)}}{\sqrt{64}}\right] = [0,10 ; 0,30]$$

Il y a 95 chances sur 100 pour que la proportion observée « tombe » dans cet intervalle.

8.4.2 Les facteurs de dépendance de la longueur de l'intervalle de pari (IP)

La longueur de l'IP est $2u_\alpha \frac{\sigma}{\sqrt{n}}$

- la longueur dépend de α
  La longueur de $IP_{1-\alpha'}$ est supérieure à la longueur de $IP_{1-\alpha}$ si $\alpha' < \alpha$

  Exemple
  $\alpha = 0,05$ $u_{0,05} = 1,96$
  $\alpha = 0,01$ $u_{0,01} = 2,57$

- la longueur dépend de n
  La longueur de $IP_{1-\alpha}$ décroît avec n. C'est le reflet du fait connu selon lequel les fluctuations d'échantillonnage s'estompent avec n

Exemple

Dans le cas ci-dessus, si on remplace $n = 64$ par $n = 6400$, on obtient $IP_{0,95} = [0,19 ; 0,21]$

Remarque

Pour réduire dans un rapport 2 la longueur de l'IP, il faut un échantillon 4 fois plus grand (2^2).

8.4.3 L'intervalle de pari d'une variable aléatoire

Ce que l'on a dit pour une moyenne expérimentale peut s'envisager pour une variable X quelconque dont on connaît la distribution.
L'IP de niveau 1 - α est l'intervalle $[a, b]$ tel que $P(a \leq X \leq b) = 1 - \alpha$.

Exemple :
$X \sim N(0, 1)$
$IP_{1-\alpha} = [-u_\alpha ; u_\alpha]$

Une valeur numérique à retenir :
pour une variable aléatoire normale centrée réduite $IP_{0,95} = [-1,96 ; 1,96]$

Résumé du chapitre

1. Propriétés de la moyenne expérimentale $\bar{X}_n$ d'une variable aléatoire X, moyenne calculée sur n unités statistiques :

$$\text{moyenne vraie de } \bar{X}_n = \text{moyenne vraie de } X$$

$$\text{variance vraie de } \bar{X}_n = \frac{\text{variance vraie de } X}{n}$$

2. **Théorème central limite**
Si X a pour moyenne vraie μ, pour variance vraie σ^2, $\bar{X}_n$ est, lorsque n est suffisamment grand ($n \geq 30$, ou np et $n(1 - p) \geq 5$), à peu près distribuée comme une variable normale de moyenne vraie μ et de variance vraie σ^2/n, ce que l'on écrit :

$$\bar{X}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \text{ ou } \frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0, 1)$$

3. **Intervalle de pari (I. P.)**
Lorsque les conditions ci-dessus sont satisfaites, l'intervalle

$$\mathrm{IP}_{1-\alpha} = \left[\mu - u_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \mu + u_\alpha \frac{\sigma}{\sqrt{n}}\right]$$

a la propriété suivante :

$$P(\bar{X}_n \in \mathrm{IP}_{1-\alpha}) = 1 - \alpha$$

Cet intervalle s'appelle intervalle de pari (I. P.) de niveau 1-α, ou intervalle de pari au risque α.

Chapitre 9

Le premier problème d'induction statistique : les tests d'hypothèses. Principes

Les tests d'hypothèses sont fondés sur les intervalles de pari.
Ce chapitre traite du principe des tests ; des précisions concernant leur usage sont indiquées au chapitre 13.

9.1 Un exemple concret (emprunté à Schwartz)

Une variété de souris présente des cancers spontanés avec un taux (une fréquence ou proportion dans la population) constant bien connu, 20 %. On se demande si un traitement donné modifie ce taux (en plus ou en moins), c'est-à-dire est actif. Pour répondre à cette question on procède à une expérience sur 100 souris ; il s'agira, au vu du pourcentage observé p_0 d'animaux cancéreux, de dire si le traitement est actif. Il n'est pas possible de répondre au sens strict à cette question.

Supposons que le traitement soit sans effet ; alors chaque souris traitée aura toujours 20 chances sur 100 de devenir cancéreuse. Mais le pourcentage de souris cancéreuses, calculé sur un échantillon de 100 souris sera soumis aux fluctuations d'échantillonnage que l'on a étudiées. Le pourcentage expérimental (moyenne expérimentale) pourra prendre a priori, c'est-à-dire avant expérience, plusieurs valeurs, même si les valeurs voisines de 0,2 sont les plus probables. Des valeurs de 0 ou 100 % pourraient même être observées. Ainsi même si le pourcentage observé est très différent de 20 %, il est possible que le traitement soit sans effet.

Supposons maintenant que le traitement soit actif ; la probabilité de cancer pour chaque souris (ou la proportion vraie de souris cancéreuses dans une population fictive de souris traitées) est p', différente de 0,2. Encore à cause des fluctuations d'échantillonnage, on pourra très bien, peut être de façon peu probable, obtenir une fréquence observée égale à 20 %. Ainsi même si le pourcentage

observé est 20 %, il est possible que le traitement soit actif.

On ne peut donc répondre avec certitude à la question posée.

Pourtant ne pas répondre serait renoncer à considérer tous les problèmes liés à la variabilité, c'est-à-dire à « tous » les problèmes biologiques. Alors on répondra, mais en acceptant un risque d'erreur. Répondre correspond à la démarche que chacun adopterait ; par exemple, déclarer le traitement actif si le taux expérimental de cancers après traitement s'écarte « nettement » de 20 %. C'est le sens que l'on peut donner à ce « nettement » qui est le fondement du principe des tests.

Dans le cas étudié, on aurait tendance à s'y prendre de la façon suivante. Deux hypothèses sont en présence :

- le traitement est inactif,
- le traitement est actif.

La première hypothèse est plus « fine » que la seconde car elle porte en elle une interprétation numérique : le pourcentage vrai de souris cancéreuses parmi les souris traitées est 0,2 - l'autre hypothèse indiquant seulement que ce pourcentage est différent de 0,2 ; ce qui est plus vague. Supposons alors vraie l'hypothèse la plus fine. Il devient possible de faire des déductions : sachant ce qui se passe au niveau de la population des souris traitées on peut en déduire ce qui se passera au niveau d'un échantillon. En particulier, on sait construire les intervalles de pari centrés de niveau $1 - \alpha$ pour la fréquence observée.

Par exemple, prenant $\alpha = 0{,}05$, on obtient $IP_{0,95} = [0{,}12 ; 0{,}28]$
Cela signifie, rappelons-le, que si $p = 0{,}2$ (fréquence supposé vraie), 95 % des valeurs des moyennes expérimentales calculées sur 100 individus appartiendront à l'intervalle [0,12 ; 0,28].

On adopte alors la stratégie suivante : si la valeur expérimentale de la fréquence de souris cancéreuses parmi les 100 traitées appartient à cet intervalle, on considère que cette valeur est compatible avec les fluctuations d'échantillonnage et l'activité du traitement n'est pas prouvée. Si la valeur expérimentale n'appartient pas à cet intervalle, le traitement sera considéré comme actif. Dans ce dernier cas le raisonnement est le suivant. Cet événement (la fréquence expérimentale est à l'extérieur de l'intervalle de pari) avait moins de 5 chances sur 100 de se produire et pourtant il s'est produit ; donc je ne crois plus à l'hypothèse qui m'a permis de déduire ces 5 % de chances.

Remarque : reformulation des calculs

Notons p_0 la proportion observée de souris traitées développant un cancer, sur les n souris traitées.

Le résultat du test sera de conclure ou non à l'activité du traitement selon que $p_0 \notin$ ou $\in IP_{1-\alpha}$ c'est-à-dire :

$$p_0 \notin \text{ ou } \in \left[p_h - u_\alpha \sqrt{\frac{p_h(1-p_h)}{n}} \; ; \; p_h + u_\alpha \sqrt{\frac{p_h(1-p_h)}{n}} \right]$$

où p_h est la proportion hypothétique (0,2 dans l'exemple) et u_α la borne de l'intervalle de pari au risque α de p_0.

On suppose ici que les conditions du théorème central limite sont satisfaites. On conclut donc selon que

$$p_0 - p_h \notin \text{ ou } \in \left[-u_\alpha \sqrt{\frac{p_h(1-p_h)}{n}} \; ; \; u_\alpha \sqrt{\frac{p_h(1-p_h)}{n}} \right]$$

ou encore selon que

$$\frac{p_0 - p_h}{\sqrt{\frac{p_h(1-p_h)}{n}}} \notin \text{ ou } \in [-u_\alpha \; ; \; u_\alpha]$$

On reconnaît dans la dernière expression l'intervalle de pari $IP_{1-\alpha}$ d'une variable aléatoire N(0, 1), intervalle indépendant de l'expérience projetée.

C'est comme cela que l'on abordera généralement les tests ; on cherchera à construire une variable aléatoire dont on connaisse, si l'hypothèse fine est vraie, la distribution, pour pouvoir construire un intervalle de pari ; ici il s'agirait de la variable aléatoire z déduite de la variable aléatoire moyenne expérimentale selon :

$$z = \frac{p_0 - p_h}{\sqrt{\frac{p_h(1-p_h)}{n}}}$$

avec $p_h = 0{,}2$ (transcription de l'hypothèse).

Une telle variable aléatoire s'appelle usuellement « paramètre » du test et est notée conventionnellement u ou t ou z. Ici on sait que $z \sim N(0, 1)$ et l'on construit l'intervalle de pari de niveau 1 - α pour z. Par exemple avec $\alpha = 0{,}05$, $IP_{0,95} = [-1{,}96 \; ; \; 1{,}96]$.

Puis on réalise l'expérience ce qui permet d'obtenir p_0 donc une valeur expérimentale de z, notée z_c comme $z_{\text{calculée}}$; on pourrait alors s'exprimer comme ceci (une terminologie plus précise sera indiquée plus loin) :

- si $z_c \in IP_{0,95}$ on ne peut dire que le traitement est actif
- si $z_c \notin IP_{0,95}$ le traitement est actif.

Nous allons, à la lumière de cet exemple, énumérer les étapes de mise en œuvre d'un test et revenir sur différents aspects (sens de α par exemple) avant de donner d'autres exemples de tests usuels

9.2 Principe général des tests d'hypothèses

La mise en œuvre d'un test statistique nécessite plusieurs étapes.

9.2.1 Les étapes de mises en œuvre

Etape 1

Avant le recueil des données.

Définir avec précision les deux hypothèses en présence H_0 et H_1. H_0 et H_1 jouent toujours des rôles dissymétriques.
Une des hypothèses doit être précise, ou fine. Elle engage une égalité généralement ; c'est elle qui sera H_0 et on l'appellera **hypothèse nulle**,

H_0 : hypothèse nulle

Exemple : la fréquence vraie d'apparition du cancer chez les souris traitées est 0,2, ce qui se transcrit par $p = 0{,}2$ (généralement $p = p_h$).

Le principe des tests est d'admettre cette hypothèse H_0 sauf contradiction flagrante entre ses conséquences et les résultats expérimentaux.

L'autre hypothèse est toujours plus vague ; **elle regroupe toutes les hypothèses, hormis H_0**. C'est H_1 et on l'appellera **hypothèse alternative**,

H_1 : hypothèse alternative

Exemple : la fréquence vraie d'apparition du cancer chez les souris traitées est différente de 0,2, qui se transcrit par $p \neq 0{,}2$ (généralement $p \neq p_h$).

Remarque : la formulation de ces hypothèses nécessite généralement une traduction et une simplification du problème médical sous-jacent.

Etape 2

Avant le recueil des données.

On suppose que H_0 est vraie et on cherche à définir une variable aléatoire (ou paramètre) dont on connait alors la distribution. En d'autres termes, on cherche à construire une fonction des données à venir dont on connait la distribution si H_0 est vraie. Soit z cette variable aléatoire.

Exemple : $z = \dfrac{p_0 - p_h}{\sqrt{\dfrac{p_h(1-p_h)}{n}}} \sim N(0, 1)$

Si possible, vérifier les conditions de validité.

Etape 3

Avant le recueil des données.

Choisir un seuil. Typiquement $\alpha = 0{,}05$ (une quasi obligation en pratique)
Construire un intervalle de pari (pour le paramètre z) de niveau 1 - α, noté $IP_{1-\alpha}$. Rappelons qu'il s'agit d'un intervalle tel que si H_0 est vraie, alors
$P(z \in IP_{1-\alpha}) = 1 - \alpha$

Exemple : $IP_{1-\alpha}$ pour z ci-dessus = [-1,96 ; 1,96]

Définition : l'extérieur de l'intervalle de pari $IP_{1-\alpha}$ s'appelle **région critique du test au seuil α**.

Etape 4

Avant le recueil des données.

Définir la règle de décision. Les données vont permettre de calculer une valeur de z, que l'on note z_c.

Exemple : $z_c = \dfrac{p_{\text{réellement observé}} - p_h}{\sqrt{\dfrac{p_h(1-p_h)}{n}}}$

Alors décider que :

- si z_c appartient à la région critique, remettre en cause H_0,
  la **rejeter**, et conclure **H_1 est vraie**, ou dire : « au risque α, H_0 est rejetée ».
- si z_c n'appartient pas à la région critique, mais à l'intervalle de pari $IP_{1-\alpha}$, dire que l'on ne conclut pas, ou dire que l'on ne rejette pas l'hypothèse nulle H_0.

Etape 5

Recueil des données

Réaliser l'expérience. On recueille les données $x_1, ..., x_n$; calculer z_c et conclure.
Si non fait à l'étape 2, vérifier les conditions de validité.

Etape 6

Interprétation des résultats

Cette étape concerne l'interprétation des résultats en des termes compatibles avec le problème médical initialement soulevé, et concerne en particulier le problème de la causalité. Ce point sera détaillé au chapitre 13.

Exemple : dans le cas des souris, et en cas de conclusion au rejet de l'hypothèse nulle, la question serait de savoir si ce rejet exprime véritablement une activité du traitement.

9.2.2 Justification de la règle de décision. Choix de α

9.2.2.1 Interprétation de α

On a déjà vu une interprétation de α avec l'exemple des souris. De façon générale, α est la probabilité pour que la valeur expérimentale - ou calculée - z_c appartienne à la région critique si H_0 est vraie. Si cet événement se réalise, on rejette H_0. Cela ne se justifie que si α est petit car alors on dit : voilà un événement qui avait 100×α % chances de se réaliser (5 % par exemple) - donc peu de chances - et qui pourtant s'est réalisé : les résultats ne sont pas conformes à l'hypothèse ⇒ α doit être petit.
Une autre interprétation de α montre encore mieux que α doit être petit. A nouveau, lorsque H_0 est vraie, la probabilité d'obtenir un résultat z_c dans la région critique est α. Mais alors on dit « H_1 est vraie ». Donc

⇒ α = « probabilité » de conclure H_1 alors que H_0 est vraie

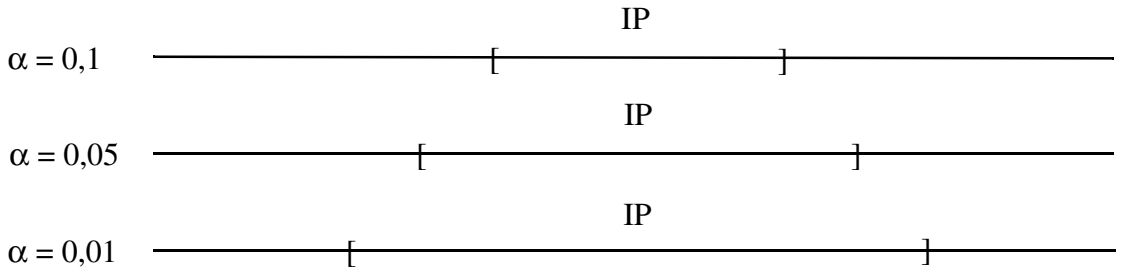
C'est un risque d'erreur qu'il convient de situer dans des valeurs acceptables (petites).

Cette valeur α s'appelle **RISQUE DE PREMIERE ESPECE**.

Cela veut dire que sur un grand nombre d'expériences, en admettant α, on conclura à tort dans 100×α % des cas (5 % des cas par exemple). Pourquoi alors ne pas choisir un α microscopique ?

9.2.2.2 Effet d'un changement de valeur de α

Les intervalles de pari croissent lorsque leur niveau augmente, c'est-à-dire lorsque α diminue.



Donc, toutes choses égales par ailleurs, la région critique diminue lorsque α décroît. Donc on rejette moins fréquemment H_0.
A vouloir commettre moins d'erreurs, on conclut plus rarement.
On s'expose donc à un autre risque : celui de ne pas conclure alors qu'il le faudrait car H_0 est fausse. A la limite, si on se fixe α = 0, on ne conclut jamais, H_0 n'est jamais rejetée.

Prendre une décision, c'est accepter un risque.

Pour finir avec ce problème de α il faut retenir :

- La valeur de α doit être fixée a priori : jamais en fonction des données
- Pire que cela, on choisit la valeur $\alpha = 0{,}05$ qui est un compromis entre le risque de conclure à tort et la faculté de conclure, compromis adopté par l'ensemble de la communauté scientifique.

9.2.3 Justification des conclusions du test. Puissance d'un test

On comprend maintenant la partie de la règle de décision conduisant au rejet de H_0 lorsque la valeur calculée du paramètre n'appartient pas à l'intervalle de pari. On a par ailleurs indiqué (voir l'étape 4 de mise en œuvre des tests) que lorsque la valeur calculée du paramètre appartient à l'intervalle de pari, c'est-à-dire lorsque les résultats expérimentaux ne sont pas contradictoires avec l'hypothèse nulle, on s'exprime avec beaucoup de précautions oratoires puisqu'on demande de dire : « on ne conclut pas » ou « on ne rejette pas l'hypothèse nulle ». Pourquoi ne pas affirmer plus directement « l'hypothèse nulle est vraie » ?

Premier élément

En faisant cela, on adopte une démarche qui s'apparente à la démarche scientifique qui consiste à admettre une théorie jusqu'à la preuve de son échec. Lorsque l'on dit « admettre » on ne signifie pas que la théorie est vraie mais qu'elle rend compte pour l'instant - jusqu'à plus ample informé - des expériences.

Exemples

— la mécanique générale admise jusqu'à la théorie de la relativité
— la mécanique céleste

Second élément

Supposons que l'on mette en parallèle les deux tests suivants :

$H_0 : p = 0{,}2$ $\qquad$ $H_0 : p = 0{,}200001$

$H_1 : p \neq 0{,}2$ $\qquad$ $H_1 : p \neq 0{,}200001$

Les paramètres calculés, soit

$$z_c = \frac{p_{\text{réellement observé}} - p_h}{\sqrt{\frac{p_h(1-p_h)}{n}}}$$

(p_h : 0,2 ou 0,200001)

seront extrêmement voisins, donc les conclusions pratiquement toujours les mêmes.

Considérons alors une expérience au cours de laquelle $z_c \in IP_{0,95}$ pour les deux valeurs calculées. Peut-on conclure à la fois $p = 0{,}2$ et $p = 0{,}200001$? Pourtant on peut remarquer qu'il n'y a pas de vice de fond au niveau de la formulation des hypothèses car il existe bien une valeur vraie, c'est-à-dire qu'il y a vraiment une hypothèse vraie du type p = quelque chose.

On retient : les tests ne sont pas faits pour « démontrer » H_0, mais pour la rejeter.
Cela ne veut pas dire que l'on est toujours content de rejeter H_0.

Exemples

- cas des souris traitées. Là on aimerait probablement rejeter H_0, c'est-à-dire conclure à l'activité du traitement.
- cas d'un test d'homogénéité. On vous livre un nouveau lot de souris ou des souris d'un autre élevage. Vous voulez continuer vos recherches. La première chose à faire est de tester l'hypothèse selon laquelle ces nouvelles souris sont similaires aux précédentes vis-à-vis du taux de cancer, $\Rightarrow H_0 : p = 0{,}2$. Mais là vous espérez bien ne pas rejeter H_0. C'est à cette condition que vous pouvez continuer.

PUISSANCE D'UN TEST

Revenons à la conclusion « l'activité du traitement n'est pas démontrée ». Sous entendu compte tenu de l'expérience effectuée. Cela n'a de sens de s'exprimer comme cela que s'il est pensable qu'une autre expérience, plus complète par exemple, puisse montrer cette efficacité si elle existe.
C'est le cas, en effet. L'aptitude d'un test à rejeter l'hypothèse nulle alors qu'elle est fausse est limitée. Précisément :

On appelle **PUISSANCE D'UN TEST** la probabilité de rejeter l'hypothèse nulle alors qu'elle est fausse.

La valeur complémentaire à 1 de cette puissance, c'est-à-dire la probabilité de ne pas rejeter l'hypothèse nulle alors que l'hypothèse alternative est vraie, s'appelle le **RISQUE DE DEUXIEME ESPECE** et se note conventionnellement β.

Le calcul de la puissance d'un test est une opération complexe. La difficulté tient essentiellement au fait que l'hypothèse alternative est vague. Pour contourner cette difficulté et apprécier plus étroitement cette notion de puissance, considérons le cas d'une hypothèse alternative fine. Par exemple, reprenant l'exemple des souris, supposons que l'hypothèse H_1 soit $p = 0{,}3$, l'hypothèse H_0 restant inchangée, c'est-à-dire $p = 0{,}2$. Dans ces conditions, il est possible de calculer la distribution de la proportion observée, non plus seulement sous l'hypothèse nulle, mais également sous l'hypothèse alternative. On obtient :

- sous l'hypothèse nulle ($p = 0{,}2$) : $p_0 \sim N\left(0{,}2, \frac{0{,}2(1-0{,}2)}{n}\right)$
- sous l'hypothèse alternative ($p = 0{,}3$) : $p_0 \sim N\left(0{,}3, \frac{0{,}3(1-0{,}3)}{n}\right)$

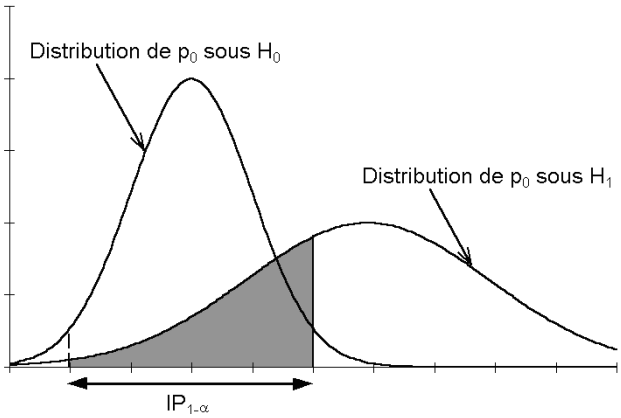


Figure 8 : risque de deuxième espèce d'un test

La figure 8 présente les deux distributions correspondantes, pour une certaine valeur de n. Supposons alors juste l'hypothèse H_1 ; la valeur expérimentale p_0 sera issue de la distribution de droite, et l'on conclura à tort au non rejet de H_0 avec une probabilité égale à l'aire grisée, puisque cette aire est la probabilité pour que la valeur expérimentale appartienne à l'intervalle de pari associé au test, sachant que cette valeur expérimentale est gouvernée par la distribution associée à H_1. Ainsi la valeur de cette aire grisée exprime le risque de deuxième espèce β, son complémentaire à 1 la puissance du test.
Supposons pour fixer les idées que la valeur de cette aire soit 0,4. Cela signifie que si les hypothèses sont $p = 0{,}2$ et $p = 0{,}3$, on aura « 6 chances sur dix » seulement de rejeter l'hypothèse $p = 0{,}2$ lorsque p sera égal à 0,3. Autrement dit, 4 fois sur dix, on sera incapable de détecter que p vaut 0,3 et non 0,2.

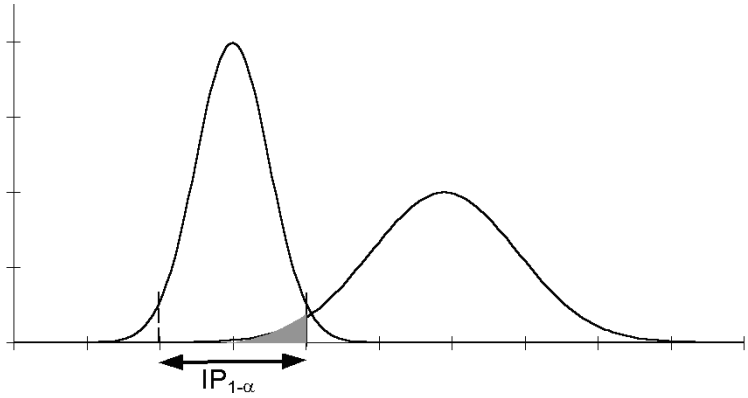


Figure 9 : risque de deuxième espèce d'un test

Par ailleurs, on perçoit que plus les hypothèses H_0 et H_1 sont contrastées (par exemple les hypothèses $p = 0{,}2$, $p = 0{,}4$ sont plus contrastées que les hypothèses $p = 0{,}2$, $p = 0{,}3$), plus les distributions de p_0 sous ces deux hypothèses sont « éloignées », et plus la puissance est grande. C'est la raison pour laquelle on dit souvent que la notion de puissance est proche de la notion de pouvoir discriminant entre hypothèses.

La figure 9 reproduit les conditions de la figure 8, mais avec une valeur de n accrue. Autrement dit le même test est mis en œuvre, mais sur un nombre d'unités statistiques supérieur. On constate sur cette figure que le risque de deuxième espèce est très faible. Ce résultat est général :

TOUTES CHOSES EGALES PAR AILLEURS, LA PUISSANCE D'UN TEST AUGMENTE AVEC LA TAILLE DE L'ECHANTILLON

Remarque
Les calculs de puissance ébauchés ci-dessus, joints au résultat précédent, permettent de répondre à des questions du type :

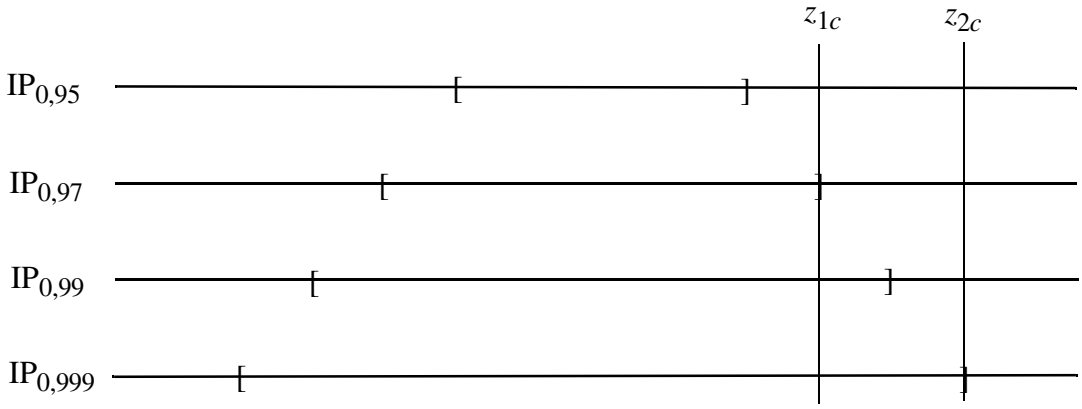
- combien de sujets est-il nécessaire d'inclure dans un essai pour avoir de bonnes chances (9 chances sur dix par exemple) de mettre en évidence une différence entre proportions vraies d'au moins 0,1 ?
- si je dispose de 100 sujets, quelle différence minimum entre proportions vraies suis-je capable de détecter avec une probabilité de 0,9 ?

Les développements ci-dessus montrent que lorsque vous n'avez pas rejeté l'hypothèse nulle, vous pouvez toujours dire que c'est un **manque de puissance du test** puisque H_0 est sans doute fausse (pensons à $p = 0{,}2$ exactement). On peut donc dire qu'avec un plus grand nombre d'individus vous auriez rejeté H_0. Cela justifie l'expression « l'activité du traitement n'est pas démontrée ».
Cependant il faut être réaliste : reprenons l'exemple des souris traitées ou non traitées. Vous avez réalisé votre expérience sur un échantillon de 1000 souris. Résultat du test : non rejet de H_0 c'est-à-dire l'activité n'est toujours pas démontrée. Il n'est pas raisonnable dans ces conditions d'évoquer un manque de puissance du test ; ce résultat suggère plutôt une très faible activité du traitement, si elle existe.

9.2.4 Amélioration de l'interprétation du rejet de H_0

9.2.4.1 Notion de degré de signification

Supposons que l'on réalise un test au risque ou seuil $\alpha = 0{,}05$.
Considérons deux expériences conduisant au rejet de H_0, pour lesquelles on a obtenu des valeurs calculées du paramètre z_{1c} et z_{2c} représentées ci-dessous.
On aurait envie de rejeter plus fortement H_0 dans le second cas que dans le premier. En effet, considérons des intervalles de pari pour z, de niveau croissant à partir de 0,95.



On observe que z_{1c} est à l'extérieur des intervalles de pari jusqu'au niveau 0,97, que z_{2c} est à l'extérieur des intervalles de pari jusqu'au niveau 0,999. Cela signifie que, en ce qui concerne la première expérience, H_0 aurait été rejetée même si on avait limité le risque d'erreur à 1 - 0,97 = 0,03 (soit 3 %), et que, en ce qui concerne la seconde, H_0 aurait été rejetée même si on avait limité le risque d'erreur à 1 - 0,999 = 0,001 (soit 1‰). C'est ce pseudo risque d'erreur que l'on appelle **degré de signification** et qui mesure la force avec laquelle on rejette H_0.
Ce degré de signification est noté p : plus il est petit, plus confortable est le rejet.

Si l'on veut une définition plus précise :
Définition
Lorsque H_0 est rejetée, on appelle degré de signification d'un test le risque associé au plus grand intervalle de pari qui ne contient pas le paramètre calculé z_c.

Calcul pratique du degré de signification
On cherche dans la table la valeur de p pour laquelle $u_p = z_c$, u_p étant du type u_α
Exemple: $z_c = 2{,}43$.
On trouve dans la table $u_{0,02} = 2{,}32$ et $u_{0,01} = 2{,}57$
alors $p \in [0{,}01 ; 0{,}02]$
La valeur exacte ne se trouve pas dans la table : on dira $p < 0{,}02$. Le plus grand intervalle de pari ne contenant pas z_c est de niveau > 0,98, ou au risque < 0,02.

La plupart des résultats de tests s'expriment avec ce degré de signification :

- On réalise le test (avec un risque $\alpha = 0{,}05$)
- Si H_0 est rejetée, on calcule ou on évalue le degré de signification p
- Si H_0 n'est pas rejetée, on ne calcule pas p.

9.2.4.2 Orientation du rejet

Le rejet de H_0 correspond généralement à l'une des deux situations :
rejet car z_c est trop petit (inférieur à la borne inférieure de l'intervalle de pari)
ou car z_c est trop grand (supérieur à la borne supérieure de l'intervalle de pari)

Dans le cadre de l'exemple précédent, chacune de ces situations correspond généralement à des commentaires radicalement différents. Par exemple :
z_c est trop petit $\Leftrightarrow$ le traitement est efficace
z_c trop grand $\Leftrightarrow$ le traitement est nuisible

Résumé du chapitre

A. Etapes de mise en œuvre des tests :

1. Examiner le problème médical, aboutir à une formulation sous forme d'une question simple mettant en jeu deux hypothèses H_0 (précise, dite hypothèse nulle) et H_1 (contraire de H_0, dite hypothèse alternative). Enoncer ces hypothèses.
2. Construire un paramètre dépendant des données à venir dont on connaisse la distribution si H_0 est juste.
3. Choisir le seuil α ; $\alpha = 0{,}05$
4. Mettre en place la règle de décision sur la base d'un intervalle de pari au risque α.
5. Faire l'expérience, les calculs et conclure sur le plan statistique. En particulier indiquer le degré de signification du test en cas de rejet de l'hypothèse nulle.
6. Se livrer à une interprétation médicale des résultats du test (ce point sera revu au chapitre 13).

Vérifier les conditions de validité à l'étape 2 ou l'étape 5.

B. Mettre en œuvre un test c'est accepter deux risques d'erreur :

— le risque de première espèce, α, chiffrant la probabilité de rejeter H_0 alors qu'elle est vraie,

— le risque de deuxième espèce, β, chiffrant la probabilité de ne pas rejeter H_0 alors qu'elle est fausse.

La valeur $1-\beta$ s'appelle la puissance du test et mesure l'aptitude du test à détecter un écart entre la réalité et l'hypothèse nulle. Cette puissance augmente avec la taille des échantillons sur lesquels a été mis en œuvre le test.

Chapitre 10

Quelques tests usuels

10.1 Test d'égalité d'une proportion vraie à une valeur donnée (ou test de comparaison d'une proportion observée à une valeur donnée)

10.1.1 Mise en place du test

Exemple : les souris du chapitre précédent

1. Les hypothèses en présence

   H_0 (hypothèse nulle) : la proportion vraie (dans la population des souris traitées) est égale à p_h (proportion hypothétique ou supposée qu'on se donne pour le test).
   H_1 (hypothèse alternative) : la proportion vraie est différente de p_h.
   Notations :
   $H_0 : p = p_h$
   $H_1 : p \neq p_h$

2. Définition du paramètre

   $$z = \frac{p_0 - p_h}{\sqrt{\frac{p_h(1-p_h)}{n}}}$$

   où p_0 représente la variable aléatoire proportion expérimentale.
   Sous H_0, z est à peu près distribuée selon N(0, 1)

   [conditions de validité : $np_h \geq 5$ et $n(1 - p_h) \geq 5$]

3. Choix d'un seuil de signification α

Construction de l'intervalle de pari de niveau 1 - α : $IP_{1-\alpha}$
Exemple : $\alpha = 0{,}05$ $IP_{0{,}95}$= [-1,96 ; 1,96] (lu dans la table de la distribution normale)

4. Mise en place de la procédure de décision

   Lorsque les données seront disponibles on obtiendra une valeur du paramètre z, soit z_c.

   Si $z_c \notin IP_{1-\alpha}$ on rejette H_0 et on dit : au risque α l'hypothèse d'égalité de la proportion vraie et de la valeur donnée est fausse ; ou, au risque α, la proportion vraie est différente de la valeur donnée.
   Si $z_c \in IP_{1-\alpha}$ on ne rejette pas H_0 ou « on ne conclut pas ».

5. Recueil des données. Conclusion

   Rappelons les conditions de validité: $np_h \geq 5$ et $n(1 - p_h) \geq 5$

10.1.2 Autre interprétation du paramètre z_c

Regardons la forme du paramètre z_c. On conclut (c'est-à-dire on rejette H_0) si $z_c \notin [-u_\alpha ; u_\alpha]$ c'est-à-dire si $|z_c| > u_\alpha$ soit si :

$$|p_0 - p_h| > u_\alpha \sqrt{\frac{p_h(1-p_h)}{n}}$$

c'est-à-dire si la proportion observée p_0 est suffisamment différente de p_h. Voilà pourquoi on dit que l'on compare p_0 et p_h. C'est pourquoi on dit aussi, lorsque H_0 est rejetée :

> La proportion observée est **significativement** différente de la valeur donnée, au risque α (0,05), ou encore : la différence entre p_0 et p_h est significative. Ce qui indique une différence entre la valeur donnée et la proportion vraie p.
> Lorsque H_0 n'est pas rejetée, on dit : la proportion observée n'est pas significativement différente de la valeur donnée.

Très important : une même différence $|p_0 - p_h|$ peut être ou non significative selon la valeur de n. Si l'on vous demande : $p_0 = 0{,}25$ et 0,2, sont-elles significativement différentes, ne répondez-pas ; demandez : quelle est la taille de l'échantillon sur lequel p_0 a été calculé, à quel risque ?

10.2 Test d'égalité d'une moyenne vraie à une valeur donnée (ou test de comparaison d'une moyenne observée à une valeur donnée)

Ce cas concerne les variables quantitatives continues.

Exemple : la moyenne vraie de la taille des individus dans une sous-population coïncide-t-elle avec la moyenne vraie de la taille des individus dans la population générale, cette taille moyenne étant connue par ailleurs.
Il convient là de distinguer deux cas

— le cas des grands échantillons ($n \geq 30$)
— le cas des petits échantillons

10.2.1 Cas des grands échantillons

1. Les hypothèses en présence :

   H_0 : la moyenne vraie coïncide avec la valeur donnée μ_h : $\mu = \mu_h$
   H_1 : $\mu \neq \mu_h$

2. Construction du paramètre

   $$z = \frac{\bar{x} - \mu_h}{\sqrt{\frac{s^2}{n}}}$$

   z est à peu près distribué selon N(0, 1). Cela résulte du théorème central limite, à ceci près que s^2 est utilisé à la place de σ^2. On admettra que z est tout de même distribué selon une distribution normale.

3. Choix du seuil ; $\alpha = 0{,}05$

   Construction de l'intervalle de pari centré $IP_{1-\alpha}$
   $IP_{1-\alpha} = [-u_\alpha ; u_\alpha]$; $u_{0,05} = 1{,}96$

4. Définition de la règle de décision

   La règle de décision est tout à fait similaire au cas des proportions.
   Si $z_c \notin IP_{1-\alpha}$, rejet de H_0. On dit alors : au risque α la moyenne vraie diffère de la valeur donnée ou, pour les mêmes raisons que pour les proportions : la moyenne expérimentale observée est significativement différente, au risque α, de la valeur donnée ; ou encore : la

moyenne observée et la valeur donnée sont significativement différentes, au risque α.
Si $z_c \in IP_{1-\alpha}$, on ne conclut pas. La moyenne observée n'est pas significativement différente de la valeur donnée.

10.2.2 Cas des petits échantillons ($n < 30$)

1. Les hypothèses restent les mêmes $\mu = \mu_h$, et $\mu \neq \mu_h$

2. Construction du paramètre

   Là encore, deux cas se présentent :

   i. la variable que l'on s'apprête à observer, X, a une distribution quelconque : alors on ne sait pas franchir cette étape.
   ii. la variable est normale (gaussienne), alors la densité de probabilité de la variable

   $$t = \frac{\bar{x} - \mu_h}{\sqrt{\frac{s^2}{n}}}$$

   peut se calculer ; cette densité s'appelle loi de Student. En réalité cette densité reste dépendante de la taille de l'échantillon sur lequel on s'apprête à réaliser l'expérience, soit n, si bien que l'on est amené - pour faire face à toutes les situations expérimentales - à considérer une famille de distributions indexées par une valeur entière que l'on appelle le nombre de degrés de liberté de la loi de Student.
   On dit de la variable t ci-dessus qu'elle suit une **LOI DE STUDENT A** (n-1) **DEGRES DE LIBERTE**.

3. Choix de α. Construction de l'intervalle de pari $IP_{1-\alpha}$

   Comme dans le cas de la loi normale, des tables ont été construites qui permettent d'obtenir les intervalles de pari de niveau 1-α. Ces intervalles de pari sont symétriques par rapport à zéro, c'est-à-dire de la forme [$-t_\alpha$; t_α]. Pour se rappeler la dépendance de t_α vis-à-vis du nombre de degrés de liberté, on note la valeur t_α selon $t_\alpha(n\text{-}1)$.
   Exemples :
   $n = 10 \Rightarrow 9$ ddl $\Rightarrow IP_{0,95} = [-2,26 ; 2,26]$
   $n = 15 \Rightarrow 14$ ddl $\Rightarrow IP_{0,95} = [-2,14 ; 2,14]$
   $n = 20 \Rightarrow 19$ ddl $\Rightarrow IP_{0,95} = [-2,09 ; 2,09]$

   Remarque : ddl est une abréviation de « degrés de liberté ».

4. Règle de décision : comme d'habitude selon que t_c appartient ou non à $IP_{1-\alpha}$

5. Recueil des données. Conclusion

Remarque : lorsqu'on réalise ce test on dit que l'on utilise un test de Student. Ce test est utilisable

même pour les grands échantillons mais alors t_α est très peu différent de u_α

10.3 Test d'égalité de deux proportions vraies (ou test de comparaison de deux proportions observées)

Reprenons l'exemple des souris mais en supposant maintenant que l'on ne connaît plus la fréquence vraie de cancer chez les souris non traitées (le 0,2 d'alors). On se pose toujours la même question relative à l'activité du traitement. On est amené à reformuler légèrement le problème et identifier l'absence d'activité du traitement à l'égalité des proportions vraies de souris cancéreuses dans deux populations, l'une traitée l'autre non traitée, et l'activité à une différence entre ces deux pourcentages. On notera A et B les deux populations, p_A et p_B les fréquences vraies de souris cancéreuses dans ces deux populations, n_A et n_B les tailles des échantillons sur lesquels on calculera p_{0A} et p_{0B}, les fréquences expérimentales correspondantes. Mettons en place le test.

1. Les hypothèses en présence

   H_0 hypothèse nulle : les fréquences vraies coïncident $p_A = p_B$
   H_1 hypothèse alternative : les fréquences vraies sont différentes $p_A \neq p_B$

2. Construction d'un paramètre dont on connaisse la loi sous l'hypothèse nulle (i.e. si H_0 est vraie)

   C'est une étape un peu délicate (le lecteur peu curieux peut passer rapidement sur ces développements). Essayons de nous ramener à un cas connu : comparaison d'un pourcentage observé à une valeur donnée, problème associé aux hypothèses suivantes :
   $H_0 : p = p_h$
   $H_1 : p \neq p_h$
   On y parvient en reformulant les hypothèses
   $H_0 : p_A - p_B = 0$
   $H_1 : p_A - p_B \neq 0$
   Il s'agit donc de comparer à 0 la différence $p_A - p_B$.

   Auparavant on formait le paramètre $\dfrac{p_0 - p_h}{\sqrt{\dfrac{p_h(1-p_h)}{n}}}$

   qui peut s'interpréter comme $\dfrac{\%\text{ expérimental} - \text{valeur théorique}}{\text{écart-type du }\%\text{ expérimental}}$

   Alors on va former $\dfrac{\text{différence des }\%\text{ expérimentaux} - \text{valeur théorique}}{\text{écart-type des différences des }\%\text{ expérimentaux}}$

soit $\dfrac{p_{0A} - p_{0B}}{\text{écart-type des différences des \% expérimentaux}}$

La difficulté est de former l'expression de l'écart type des différences des % expérimentaux. Remarquons d'abord que les variables aléatoires p_{0A} et p_{0B} sont indépendantes ; cette indépendance résulte du fait que ce n'est pas parce que l'on a trouvé une souris cancéreuse dans la population des souris traitées que l'on a plus ou moins de chances de trouver une souris cancéreuse ou non dans la population non traitée.
Alors : $var(p_{0A} - p_{0B}) = var(p_{0A}) + var(-p_{0B}) = var(p_{0A}) + var(p_{0B})$ (voir chapitre 5)
Par ailleurs, sous l'hypothèse nulle, les moyennes vraies de p_{0A} et p_{0B} coïncident avec une valeur p - inconnue. D'où :

$$var(p_{0A}) = \frac{p(1-p)}{n_A} \text{ et } var(p_{0B}) = \frac{p(1-p)}{n_B}$$

si n_A et n_B sont les tailles des échantillons sur lesquels p_{0A} et p_{0B} sont calculées.

Donc : $var(p_{0A} - p_{0B}) = \dfrac{p(1-p)}{n_A} + \dfrac{p(1-p)}{n_B}$

Maintenant, p reste inconnu ; il s'agit de la valeur vraie commune des pourcentages. Le mieux pour l'estimer est de mélanger les deux populations - elles contiennent sous H_0 le même pourcentage de souris cancéreuses - et dire :

$$p \text{ proche de } \hat{p} = \frac{\text{nombre de souris cancéreuses dans les deux échantillons}}{\text{nombre total de souris}}$$

soit : $\hat{p} = \dfrac{n_A p_{0A} + n_B p_{0B}}{n_A + n_B}$

Finalement on adopte le paramètre suivant :

$$z = \frac{p_{0A} - p_{0B}}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n_A} + \dfrac{\hat{p}(1-\hat{p})}{n_B}}}$$

avec $\hat{p} = \dfrac{n_A p_{0A} + n_B p_{0B}}{n_A + n_B}$

Sous l'hypothèse nulle z est à peu près distribuée selon N(0, 1).
Conditions de validité :

$$\begin{cases} n_A\hat{p} \geq 5, n_A(1-\hat{p}) \geq 5 \\ n_B\hat{p} \geq 5, n_B(1-\hat{p}) \geq 5 \end{cases}$$

3. Choix d'un seuil de signification α ($\alpha = 0{,}05$).

   Construction de l'intervalle de pari $IP_{1-\alpha}$ lu dans une table.
   ex. : $IP_{0,95} = [-1{,}96 ; 1{,}96]$

4. Mise en place de la procédure de décision

Si z_c, dont on connaîtra la valeur une fois l'expérience réalisée
$\in$ $IP_{0,95}$ on ne conclut pas
$\notin$ $IP_{0,95}$ on rejette H_0 : une proportion est alors plus grande que l'autre.

5. Réalisation de l'expérience, calcul de z_c, conclusion.

10.4 Test d'égalité de deux moyennes vraies (ou test de comparaison de deux moyennes observées)

Exemple : la moyenne vraie de la taille des individus dans une sous-population A coïncide-t-elle avec la moyenne vraie de la taille des individus dans une autre sous-population B, ces moyennes vraies n'étant pas connues. On va réaliser une expérience mettant en jeu deux échantillons issus des deux populations, à l'issue de laquelle on disposera de deux séries de valeurs de taille (les nombres de valeurs observées sont notés respectivement n_A et n_B).

Là encore il convient de distinguer deux cas.

10.4.1 Cas des grands échantillons (n_A et $n_B \geq 30$)

Il s'agit d'un problème très proche du précédent

1. Les hypothèses en présence

   H_0 hypothèse nulle : les moyennes vraies dans les deux populations coïncident $\mu_A = \mu_B$
   H_1 hypothèse alternative : $\mu_A \neq \mu_B$

2. Construction du paramètre : cette construction suit les mêmes lignes que précédemment et on obtient

$$z = \frac{\overline{x_A} - \overline{x_B}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

où

$$\overline{x_A} = \frac{1}{n_A}\sum_{i=1}^{n_A} x_{iA} \text{ et } s_A^2 = \frac{1}{n_A - 1}\sum_{i=1}^{n_A} (x_{iA} - \overline{x_A})^2$$

$$\overline{x_B} = \frac{1}{n_B}\sum_{i=1}^{n_B} x_{iB} \text{ et } s_B^2 = \frac{1}{n_B - 1}\sum_{i=1}^{n_B} (x_{iB} - \overline{x_B})^2$$

les x_{iA} et x_{iB} étant les valeurs de tailles observées dans les échantillons des populations A et B respectivement.
Alors z est à peu près distribuée selon N(0, 1).

3. Choix d'un seuil de signification (0,05)

   Construction de l'intervalle de pari $IP_{1-\alpha}$ ($IP_{0,95}$)

4. Règle de décision

5. Mise en œuvre de l'expérience ; conclusion.

10.4.2 Cas des petits échantillons (n_A ou $n_B < 30$)

1. Les hypothèses en présence restent les mêmes. Seul change le paramètre car le théorème central limite ne s'applique plus. Pour pouvoir continuer à faire quelque chose, il faut supposer :
   - que les caractères étudiés - les variables aléatoires étudiées - sont distribués **normalement** ;
   - que les variances des variables sont égales.
2. On montre qu'alors, sous l'hypothèse nulle, la loi du paramètre suivant est connue :

   $$t = \frac{\overline{x_A} - \overline{x_B}}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}}$$

   où s^2 est une approximation de la variance supposée commune des variables étudiées.
   Précisément, on montre qu'il convient de calculer s^2 selon :

   $$s^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)}$$

   Dans ces conditions, t est à nouveau une variable de Student (voir la section 10.2.2 consacrée à la comparaison d'une moyenne à une valeur théorique), cette fois-ci à ($n_A + n_B$ - 2) degrés de liberté.

3. Choix de α : $\alpha = 0,05$

   Construction de l'intervalle de pari $IP_{0,95}$ lu dans une table
   ex. : $n_A = 8$, $n_B = 13$, $n_A + n_B - 2 = 19 \Rightarrow IP_{0,95} = [-2,09 ; 2,09]$

4. Procédure de décision habituelle.

5. Réalisation de l'expérience. Décision.

10.5 Test de comparaison de deux moyennes. Cas des séries appariées

Jusqu'à présent on a supposé que les tirages (la constitution) des échantillons des populations *A* et *B* étaient indépendants. Il arrive que cette condition ne soit pas vérifiée, que les individus des deux échantillons soient liés. Ceci se produit dans les exemples suivants :

- pour comparer le niveau de sévérité de deux examinateurs, on fait corriger 100 copies par chacun d'eux, c'est-à-dire chacun corrigeant chacune de ces cent copies, et il s'agit de comparer les notes moyennes.
- pour comparer deux méthodes de dosage de la glycémie on dose 100 prélèvements de sang par chacune de ces deux méthodes et l'on souhaite comparer les valeurs moyennes vraies.

La procédure indiquée plus haut ne convient plus. A un moment de la mise en place des tests on avait à calculer la variance de la différence des moyennes expérimentales. On avait dit qu'elle coïncide avec la somme des variances de chacune des moyennes. Ici, c'est faux ; on peut s'en convaincre facilement. Supposez qu'un correcteur accorde systématiquement un point de plus que son collègue à toutes les copies. Alors, quoi qu'il arrive, la différence des moyennes expérimentales sera 1, donc cette différence n'est pas soumise aux fluctuations d'échantillonnage ; sa variance est nulle, donc n'a rien à voir avec les variances de chacune des moyennes qui, elles - ces variances - reflètent les différences de qualité entre les copies.
On montre que le bon abord du problème est de travailler sur les différences obtenues (différence des notes, différence des glycémies par individu) et de mettre à l'épreuve la nullité de la moyenne de ces différences. Finalement c'est plus simple car cela revient au problème de la comparaison d'une moyenne (moyenne des différences) à zéro. Voilà un test que l'on connaît.

On **note *d* la variable aléatoire différence** entre résultats pour un même sujet.
Les étapes de mise en œuvre du test sont les suivantes :

1. H_0 : la moyenne vraie de *d* est nulle, soit $\mu = 0$.
   H_1 : la moyenne vraie de *d* est non nulle, soit $\mu \neq 0$.
2. Construction du paramètre
   i. **Cas des grands échantillons $n \geq 30$**

$$z = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}}$$

où s^2 est la variance expérimentale des différences, soit $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$
n est le nombre de paires

et $\bar{d}$ est la moyenne expérimentale des différences.

On montre que z est à peu près distribuée selon N(0, 1).

ii. **Cas des petits échantillons $n < 30$**

Si les différences sont distribuées normalement, z est une variable de Student à $(n - 1)$ degrés de liberté.

Les étapes se succèdent alors de façon ordinaire.

Remarque

Si les notes attribuées par chacun des correcteurs varient généralement dans le même sens - c'est-à-dire une copie mieux notée qu'une autre par le premier examinateur le sera également par le second - alors la valeur absolue de z calculée sur la base de l'appariement est supérieure à la valeur absolue que l'on aurait obtenue en « oubliant » l'appariement. Ainsi, toutes choses égales par ailleurs, on conclura plus fréquemment au rejet de l'hypothèse nulle : le test ainsi mis en place est plus puissant. On a exploité plus d'information. On a gommé une source de fluctuations, celle liée à la disparité de la qualité des copies. Si cet effet de variation dans le même sens n'est pas réel (ex.: lorsque l'un note la copie x, l'autre la note $20 - x$) le problème dans son ensemble n'a plus beaucoup de sens.

Résumé du chapitre

1. Comparaison d'une proportion observée à une valeur donnée

$$z = \frac{p_0 - p_h}{\sqrt{\frac{p_h(1-p_h)}{n}}} \sim \mathrm{N}(0, 1)$$; validité $np_h \geq 5$ et $n(1 - p_h) \geq 5$

2. Comparaison d'une moyenne observée à une valeur donnée

$$z = \frac{\bar{x} - \mu_h}{\sqrt{\frac{s^2}{n}}} \sim \mathrm{N}(0, 1)$$; validité $n \geq 30$

$$t = \frac{\bar{x} - \mu_h}{\sqrt{\frac{s^2}{n}}} \sim \text{Student } (n-1)$$; validité : normalité

3. Comparaison de deux proportions observées

$$z = \frac{p_{0A} - p_{0B}}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_A} + \frac{\hat{p}(1-\hat{p})}{n_B}}} \sim \mathrm{N}(0, 1) \; ; \; \hat{p} = \frac{n_A p_{0A} + n_B p_{0B}}{n_A + n_B}$$

validité : $n_A\hat{p} \geq 5$, $n_A(1-\hat{p}) \geq 5$, $n_B\hat{p} \geq 5$, $n_B(1-\hat{p}) \geq 5$

4. Comparaison de deux moyennes observées

$$z = \frac{\overline{x_A} - \overline{x_B}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \sim \mathrm{N}(0, 1)$$; validité n_A et $n_B \geq 30$

$$t = \frac{\overline{x_A} - \overline{x_B}}{\sqrt{\frac{s^2}{n_A} + \frac{s^2}{n_B}}} \sim \text{Student } (n_A + n_B - 2) \; ; \; s^2 = \frac{(n_A-1)s_A^2 + (n_B-1)s_B^2}{(n_A-1) + (n_B-1)}$$

validité : normalité, variances égales

5. Comparaison de deux moyennes sur séries appariées
on travaille sur la différence des variables, d

$$z = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}} \sim \mathrm{N}(0, 1)$$; validité n ≥ 30

$$t = \frac{\bar{d}}{\sqrt{\frac{s^2}{n}}} \sim \text{Student } (n-1)$$; validité la variable d est normale

Chapitre 11

Tests concernant des variables qualitatives

Introduction

On a jusqu'à présent complètement négligé les variables qualitatives à plus de deux modalités. On a en effet toujours parlé de **moyenne**, et cette notion n'existe pas pour les variables qualitatives, sauf pour celles à deux modalités grâce à un artifice de codage. Il n'y a pas d'instrument permettant de résumer la distribution d'une variable qualitative ; il faut considérer la distribution dans son ensemble, c'est-à-dire l'ensemble des probabilités pour que telle ou telle modalité se réalise. Pourtant des problèmes de choix d'hypothèses se posent également dans le cas de telles variables ou tels caractères (ex : la répartition [distribution] de la couleur des cheveux est-elle la même chez les habitants de tel département et de tel autre ?). Si la répartition du caractère est connue dans une des deux populations, on aura à comparer une répartition « observée » à une répartition donnée. Si les deux répartitions sont inconnues, on aura à comparer deux répartitions « observées ». Ces problèmes sont respectivement les homologues des tests de comparaison d'une moyenne à une valeur donnée, de comparaison de deux moyennes. Il existe des tests adaptés à chacun de ces cas.

11.1 Comparaison d'une répartition observée à une répartition donnée ou test du χ^2 d'ajustement

Supposons que l'on souhaite savoir si la répartition de la couleur des cheveux dans la population des habitants du département A coïncide avec la répartition de la couleur des cheveux dans la population française, cette dernière répartition étant supposée donnée. Supposons qu'il y ait k couleurs répertoriées. On est alors amené à considérer une variable qualitative à k modalités. Notons p_i la probabilité de survenue de l'événement « la $i^{\text{ème}}$ modalité est observée ».

Exemple :
p_1 = probabilité qu'un individu tiré au hasard dans le département A ait les cheveux blonds
p_2 = probabilité qu'un individu tiré au hasard dans le département A ait les cheveux bruns

etc...
Notons par ailleurs p_{hi} la proportion vraie de la modalité i dans la population française.
On s'apprête à réaliser une expérience sur n individus à l'issue de laquelle on disposera d'un ensemble de O_i (O_i = nombre d'individus présentant la modalité i du caractère étudié, parmi les individus de l'échantillon).

11.1.1 Les étapes de mise en œuvre

1. Les hypothèses en présence

   Deux hypothèses sont en présence :

   i. la répartition vraie de la variable dans la population étudiée coïncide avec la répartition donnée (hypothèse nulle H_0)
   ii. les répartitions diffèrent (hypothèse alternative H_1)

   Avec les notations précédemment introduites, cela s'écrit :

   H_0 : hypothèse nulle : $p_i = p_{hi}$ pour tous les i de 1 à k.
   H_1 : hypothèse alternative : $p_i \neq p_{hi}$ pour au moins une modalité, c'est-à-dire pour au moins un i.

2. Construction du paramètre

   On a déjà mis en place ce test dans le cas d'une variable (0 - 1) c'est-à-dire d'une variable à deux modalités. Dans ce cas, les hypothèses en présence étaient bien du type ci-dessus c'est-à-dire
   $H_0 : p = p_{h1}$ et $1 - p = p_{h2} = 1 - p_{h1}$

   ce qui s'écrit avec nouvelles notations :
   $p_1 = p_{h1}$ et $p_2 = 1 - p_{h1}$

   Mais on n'avait retenu que la condition $p = p_{h1}$ (en fait $p = p_h$) car dans ce cas les deux conditions ci-dessus sont redondantes.
   Le paramètre retenu était :

$$z = \frac{p_0 - p_{h1}}{\sqrt{\frac{p_{h1}(1-p_{h1})}{n}}}$$

   Calculons son carré

$$z^2 = \frac{n(p_0 - p_{h1})^2}{p_{h1}(1-p_{h1})} = \frac{n(p_0 - p_{h1})^2}{p_{h1}} + \frac{n(p_0 - p_{h1})^2}{1-p_{h1}}$$

$$z^2 = \frac{(np_0 - np_{h1})^2}{np_{h1}} + \frac{(n(1-p_0) - n(1-p_{h1}))^2}{n(1-p_{h1})} = \frac{(np_0 - np_{h1})^2}{np_{h1}} + \frac{(n(1-p_0) - np_{h2})^2}{np_{h2}}$$

Or np_0 = nombre d'individus observés présentant la valeur 1 c'est-à-dire la modalité 1 de la variable ; or sous H_0 la probabilité de cette modalité est p_{h1}. On s'attend donc à observer np_{h1} individus présentant cette valeur. Ce nombre d'individus attendu s'appellera effectif calculé de la première modalité et sera noté C_1.
De la même façon, $n(1 - p_0)$ = nombre d'individus observés présentant la valeur 0 c'est-à-dire la modalité 2 de la variable ; or sous H_0 la probabilité de cette modalité est $p_{h2} = 1 - p_{h1}$. On s'attend donc à observer np_{h2} individus présentant cette valeur. Ce nombre d'individus attendu s'appellera effectif calculé de la seconde modalité et sera noté C_2.

D'où $z^2 = \dfrac{(O_1 - C_1)^2}{C_1} + \dfrac{(O_2 - C_2)^2}{C_2}$

où les O_i représentent les effectifs observés dans les différentes modalités, les C_i représentent les effectifs np_{hi} dits ou prévus ou **CALCULES** dans les différentes modalités.

GENERALISATION
Lorsque les variables considérées ont plus de deux modalités, on généralise le calcul ci-dessus et on retient le paramètre suivant :

$$K^2 = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i}$$

où la somme s'étend à toutes les k modalités de la variable.
On rappelle que les O_i sont les effectifs observés, et que les C_i valent np_{hi}.

On remarque que K^2 chiffre l'écart entre ce qui est prévu théoriquement et ce qui est obtenu ; cet écart se fonde naturellement sur les différences O_i - np_{hi} car np_{hi} est le nombre attendu d'individus présentant la modalité i.

Exemple : si p_{hi} = 0,4, sur 100 individus on en attend 40 présentant la modalité i. C'est le nombre que l'on aurait si la distribution d'échantillonnage coïncidait avec la distribution théorique.

Par ailleurs on a pu montrer (résultat dû à Pearson) que sous H_0 (et si tous les $C_i \geq 5$) ce paramètre a une distribution qui ne dépend que du nombre de modalités, k. Cette distribution porte le nom de **DISTRIBUTION DE χ^2**.

Si bien que l'on peut former - grâce encore à une table - un intervalle de pari de niveau donné relatif à cette variable.

RETENONS :
CONDITIONS DE VALIDITE : TOUS LES C_i DOIVENT ETRE AU MOINS EGAUX A 5

3. Intervalle de pari α étant choisi (0,05), construction de l'intervalle de pari $IP_{1-\alpha}$
La variable χ^2 a l'allure présentée figure 10. On remarque qu'il serait stupide de choisir l'intervalle de pari centré dessiné sur cette figure car alors des valeurs numériques voisines de

zéro pour K_c^2 seraient dans la région critique du test ; or des valeurs proches de zéro sont plutôt compatibles avec H_0 d'où le choix suivant (voir figure 11) :

$IP_{1-\alpha} = [0 ; \chi^2_\alpha]$

C'est cette valeur, notée χ^2_α qui est lisible directement dans une table.

Remarque : Notez que l'intervalle n'est pas symétrique.

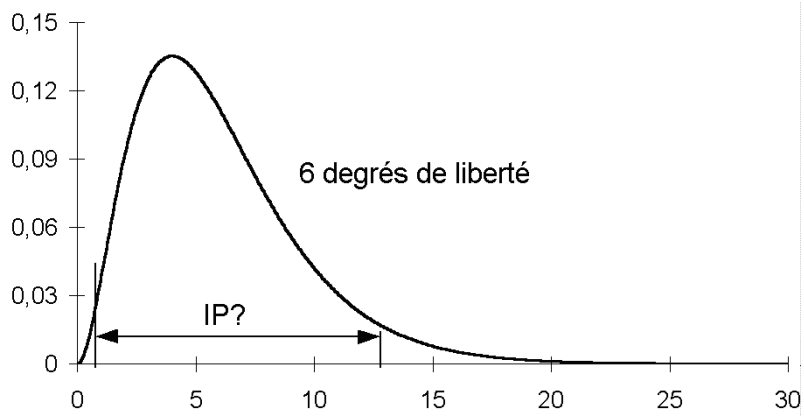


Figure 10 : distribution de χ^2

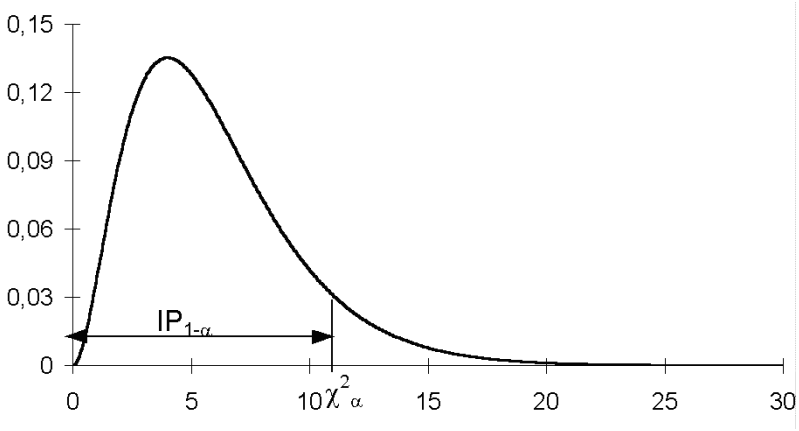


Figure 11 : distribution de χ^2

Usage de la table

Cette table comporte - comme celle du *t* de Student - une entrée entière appelée nombre de degrés de liberté. Pour rappeler cette dépendance, la borne de l'intervalle de pari χ^2_α se note souvent $\chi^2_\alpha(d)$ où *d* est le nombre de degrés de liberté. On montre que pour le test envisagé ici

nombre de degrés de liberté = nombre de modalités - 1

Exemple : $\chi^2_{0,\,05}$ (5 ddl, si 6 modalités) = 11,07

La suite de la mise en place de ce test est usuelle.

4. Règle de décision

   Si $K_c^2 \leq \chi_\alpha^2$ on ne conclut pas

   Si $K_c^2 > \chi_\alpha^2$ H_0 est rejetée. Cela signifie que l'on conclut que la répartition du caractère étudié (par exemple la couleur des cheveux dans le département A) **ne coïncide pas** - ou **ne s'ajuste pas** - avec la répartition donnée (par exemple la répartition de la couleur des cheveux dans la population française). On admet, en formulant cette conclusion, un risque d'erreur égal à α.

5. Recueil des données et conclusion

   Exemple numérique : le tableau ci-dessous présente une application numérique de l'exemple considéré.

	couleur des cheveux			
	blonds	**bruns**	**roux**	**total**
effectifs observés (O_i)	25	9	3	37 (n)
effectifs calculés ($C_i = np_{hi}$)	14,8	11,1	11,1	37
répartition donnée (p_{hi})	0,4	0,3	0,3	1

Les conditions de validité sont vérifiées ($C_i \geq 5$).

On obtient ici :

$$K_c^2 = \frac{(25-14{,}8)^2}{14{,}8} + \frac{(9-11{,}1)^2}{11{,}1} + \frac{(3-11{,}1)^2}{11{,}1} = 13{,}3$$

On sait que K^2 est distribué selon un χ^2 à (3-1) degrés de liberté ; on lit dans la table : $\chi^2_{0,\,05}$ (2 d.d.l.) = 5,99.

Ainsi, la valeur calculée n'appartient pas à l'intervalle de pari : on conclut que la répartition du caractère ne coïncide pas avec la répartition donnée.

11.1.2 Cas particulier : variable à deux modalités

On a vu que le paramètre du test K^2 généralise l'expression du carré du paramètre z utilisé pour

la comparaison d'une proportion observée à une valeur donnée. Dans le cas d'une variable à deux modalités ($k = 2$), ces deux paramètres coïncident : $K^2 = z^2$.

En outre, et sinon il y aurait incohérence, on peut vérifier l'égalité suivante :
$\chi^2_\alpha(1 \text{ ddl}) = u^2_\alpha$

Exemple : pour $\alpha = 0{,}05$ $\chi^2_{0,\,05}(1 \text{ d.d.l.}) = 3{,}84 = (1{,}96)^2$

Ainsi, pour comparer une répartition observée à une répartition donnée, dans le cas d'une variable à deux modalités, on dispose de 2 tests équivalents, l'un fondé sur la distribution normale, l'autre fondé sur la distribution du χ^2 à 1 d.d.l. (qui est en fait la distribution du carré de N(0, 1)).
On peut utiliser l'un ou l'autre de ces tests indifféremment.

Exemple : Reprenons l'exemple du chapitre 9
Une race de souris présente un taux de cancers spontanés de 0,2. Sur 100 souris traitées on observe 34 cancers soit $p_0 = 0{,}34$. La différence est elle significative ?

- test de comparaison :

$$z_c = \frac{0{,}34 - 0{,}2}{\sqrt{\frac{0{,}2 \times 0{,}8}{100}}} = 3{,}5$$

- test du χ^2 :

	cancer	**absence de cancer**	
répartition théorique	0,2	0,8	
effectif calculé	20	80	
effectif observé	34	66	100 (effectif total)

$$K_c^2 = \frac{(34 - 20)^2}{20} + \frac{(66 - 80)^2}{80} = 12{,}25 = (3{,}5)^2$$

Remarque : On parle souvent de ce test sous la terminologie « test du χ^2 d'ajustement » pour exprimer qu'il met à l'épreuve l'ajustement - la compatibilité - entre une répartition expérimentale et une répartition donnée.

11.2 Comparaison de deux répartitions observées ou test du χ^2 d'homogénéité

On reprend l'exemple précédent concernant la répartition de la couleur des cheveux mais sans plus supposer que l'une de ces répartitions est connue ; il s'agit par exemple des répartitions de ce caractère dans deux départements. On souhaite donc comparer deux répartitions observées. Pour cela, on s'apprête à réaliser une expérience mettant en jeu deux échantillons, un échantillon de n_1 individus issu de la population des habitants du département 1, et un échantillon de n_2 individus issu de la population des habitants du département 2. A l'issue de cette expérience on disposera d'un ensemble d'effectifs observés, notés de la façon suivante :

- O_{1i} est le nombre d'individus du premier échantillon présentant la modalité i de la variable.
- O_{2i} est le nombre d'individus du second échantillon présentant la modalité i de la variable.

Le test se met en place de la façon suivante :

1. Les hypothèses en présence

   H_0 : les répartitions vraies de la variable sont identiques dans les deux populations
   H_1 : les répartitions vraies sont différentes

   Ces hypothèses se schématisent par :

   $H_0 : p_{1i} = p_{2i}$ pour toutes les modalités i.
   $H_1 : p_{1i} \neq p_{2i}$ pour au moins une modalité i.

2. Construction du paramètre

   C'est encore ici le point délicat. La solution ressemble dans son approche à celle du problème de la comparaison de deux pourcentages. **Clé du principe** : on mélange les deux populations pour calculer une pseudo-répartition théorique. On se retrouve alors pratiquement dans la situation du paragraphe précédent. Cela se verra mieux sur un exemple. On va faire, pour des raisons de simplicité de calcul, une petite entorse à notre façon de procéder, et directement évaluer le paramètre dont on connaît la loi.

   i. On construit ce que l'on appelle un **tableau de contingence** qui contient les résultats expérimentaux.

      On a procédé à une expérience portant sur 37 individus issus de la population 1 et 40 individus issus de la population 2. Les résultats sont les suivants :

Tableau 1 : effectifs observés (O_{1i} et O_{2i})

	blonds	**bruns**	**roux**	**nombre total**
échantillon 1	25	9	3	$37 = n_1$
échantillon 2	13	17	10	$40 = n_2$

ii. On construit une pseudo-répartition de référence, en mélangeant les résultats expérimentaux, c'est-à-dire en oubliant leur origine (population 1 ou population 2).

On obtient les résultats suivants, en termes d'effectifs (première ligne), puis en termes de fréquences (deuxième ligne).

Tableau 2 : répartition de « référence »

	blonds	**bruns**	**roux**	**nombre total**
mélange	38	26	13	77
fréquences	38/77 = 0,49	26/77 = 0,34	13/77 = 0,17	

Ces trois fréquences, 0,49, 0,34, 0,17, vont jouer maintenant le rôle des probabilités hypothétiques p_{hi} de la section 11.1. Pour la commodité de l'écriture, on les note respectivement p_1, p_2, p_3.

iii. On forme le tableau des effectifs calculés.

Si l'hypothèse nulle est juste, c'est-à-dire si les répartitions de la couleur des cheveux coïncident dans les deux départements, on s'attend à trouver des effectifs calculés comme suit :

effectif attendu pour la modalité i (modalité 1 = blond, modalité 2 = brun, modalité 3 = roux) dans l'échantillon j (j = 1 ou 2) : n_j multiplié par p_i

Par exemple le nombre attendu d'individus bruns dans l'échantillon de la première population est : $37 \times 0{,}34 = 12{,}6$.

En effectuant systématiquement ces calculs, on obtient le tableau des **EFFECTIFS CALCULES**.

Tableau 3 : effectifs calculés (C_{1i} et C_{2i})

	blonds	**bruns**	**roux**
échantillon 1	18,1 (n_1p_1)	12,6 (n_1p_2)	6,3 (n_1p_3)
échantillon 2	19,6 (n_2p_1)	13,6 (n_2p_2)	6,8 (n_2p_3)

iv. On calcule finalement le paramètre du test

On montre que le paramètre adapté à ce test est :

$$K^2 = \sum_{i=1}^{k} \frac{(O_{1i} - C_{1i})^2}{C_{1i}} + \sum_{i=1}^{k} \frac{(O_{2i} - C_{2i})^2}{C_{2i}}$$

où k demeure le nombre de modalités de la variable.

On a souvent recours à une expression plus compacte de l'expression ci-dessus et on écrit :

$$K^2 = \sum_{j=1}^{\text{nombre de cases du tableau}} \frac{(O_j - C_j)^2}{C_j}$$

MAIS ICI LA SOMMATION S'ETEND A **TOUTES LES CASES** DES TABLEAUX, numérotées grâce à l'indice j.

Exemple : dans l'exemple traité il s'agira donc de calculer une somme de 6 termes.
On montre que, si H_0 est vraie, K^2 est distribué comme un χ^2 à $(3 - 1) \times (2 - 1)$ degrés de liberté [3 est le nombre de modalités, et 2 le nombre de répartitions]
La VALIDITE de ce résultat suppose que tous les **effectifs calculés C_j soient au moins égaux à 5**.

GENERALISATION

Les calculs ci-dessus se généralisent à un nombre quelconque de modalités k, à un nombre quelconque de populations m.
Le paramètre K^2 à calculer a alors la forme ci-dessus, où la somme comprend $k \times m$ termes.
La distribution de K^2, sous H_0 est alors un **χ^2 à $(k - 1) \times (m - 1)$ degrés de liberté**.
Les conditions de validité du test sont : $C_j \geq 5$, $1 \leq j \leq km$

3. La suite des étapes de mise en œuvre est classique.

La valeur expérimentale de K^2, soit $K_c{}^2$, sera comparée à la valeur $\chi^2_{0,\,05}$:

- si $K_c^2 \leq \chi^2_{0,05}$ on ne conclut pas. Il n'est pas démontré que les deux répartitions vraies diffèrent.
- si $K_c^2 > \chi^2_{0,05}$ on conclut que les deux répartitions observées diffèrent significativement.

Suite de l'exemple : on obtient :

$$K_c^2 = \frac{(25 - 18{,}1)^2}{18{,}1} + \frac{(9 - 12{,}6)^2}{12{,}6} + \frac{(3 - 6{,}3)^2}{6{,}3} + \frac{(13 - 19{,}6)^2}{19{,}6} + \frac{(17 - 13{,}6)^2}{13{,}6} + \frac{(10 - 6{,}8)^2}{6{,}8}$$

soit : $K_c^2 = 9{,}96$

Or : $\chi^2_{0,\,05}$ (2 d.d.l.) = 5,99 $\Rightarrow$ rejet de H_0. Les répartitions observées de la couleur des cheveux diffèrent significativement dans les deux populations.

Remarque 1 : Ce test s'appelle aussi test du χ^2 **d'homogénéité** de plusieurs répartitions.

Remarque 2 : **Cas particulier de deux variables à deux modalités** : dans le cas où l'on considère deux variables à deux modalités, c'est-à-dire dans le cas où le tableau de contingence est à deux lignes et deux colonnes, on observe que le problème se réduit à un problème de comparaison de deux proportions observées. On montre que, dans ce cas, la valeur de K^2 coïncide avec le carré de la valeur de z, z étant le paramètre formé pour comparer directement ces proportions (voir chapitre 10).

11.3 Test d'indépendance entre deux variables qualitatives

Reprenons l'exemple précédent et supposons que les populations 1 et 2, plutôt que de correspondre à des individus habitant le département 1 et le département 2, soient en fait :

— population 1 : population des individus ayant les yeux bleus
— population 2 : population des individus ayant les yeux verts

La question que l'on aurait résolue dans le paragraphe précédent aurait été :
la répartition de la couleur des cheveux diffère-t-elle dans les populations d'individus aux yeux bleus ou verts. Ou encore, la répartition de la couleur des cheveux diffère-t-elle selon la couleur des yeux ? Autrement dit : la variable couleur des cheveux dépend-elle statistiquement de la variable couleur des yeux ?
Maintenant supposons que l'on veuille répondre à cette question. Plutôt que de prendre un échantillon de la population des individus aux yeux bleus et un autre échantillon issu de la population des individus aux yeux verts, autant prendre un échantillon de la population générale (c'est-à-dire quelle que soit la couleur de ses yeux) et observer **conjointement** la couleur des cheveux et la couleur des yeux. Vues comme cela, les deux variables jouent bien des rôles symétriques et le problème est donc de mettre à l'épreuve leur indépendance.

1. Les hypothèses en présence.

   On formule naturellement deux hypothèses :

   Hypothèse H_0
   les deux variables étudiées (couleur des cheveux, couleur des yeux) sont indépendantes. Sous cette hypothèse, le fait d'avoir observé chez un individu la couleur de ses cheveux (respectivement la couleur de ses yeux) n'apporte aucune information sur la couleur de ses yeux (respectivement la couleur de ses cheveux).
   On pourra se reporter au chapitre 5 dans lequel ont été commentées ces notions d'indépendance.
   On notera que, comme dans tous les cas rencontrés jusqu'ici, cette hypothèse est une hypothèse fine qui engage un ensemble d'égalités.

En effet, on sait que l'indépendance s'exprime par :
P(la modalité de la couleur des cheveux est l et la modalité de la couleur des yeux est c) = P(la modalité de la couleur des cheveux est l) × P(la modalité de la couleur des yeux est c), et ceci pour tous les choix possibles de l et c.

Remarque : on pourra vérifier que parmi les (nombre de modalités de la couleur des cheveux × nombre de modalités de la couleur des yeux) qui en résultent, certaines égalités sont redondantes, et que (nombre de modalités de la couleur des cheveux - 1) × (nombre de modalités de la couleur des yeux - 1) suffisent à exprimer les mêmes conditions.

Hypothèse H_1

les deux variables étudiées ne sont pas indépendantes.
Cette hypothèse exprime le contraire de H_0.

TRES IMPORTANT (des erreurs sont souvent commises)
HYPOTHESE NULLE : LES DEUX VARIABLES SONT INDEPENDANTES
HYPOTHESE ALTERNATIVE : LES DEUX VARIABLES SONT LIEES

2. Le paramètre du test

Le paramètre est encore K^2, et s'exprime exactement comme précédemment, c'est-à-dire :

$$K^2 = \sum_{j=1}^{\text{nombre de cases du tableau}} \frac{(O_j - C_j)^2}{C_j}$$

Ici le nombre de cases du tableau de contingence est égal au produit du nombre de modalités de la première variable et du nombre de modalités de la seconde variable.
Les effectifs calculés s'obtiennent exactement comme dans le cas du paragraphe précédent, ainsi qu'on peut le voir sur l'exemple numérique ci-dessous.

Un exemple numérique

Le tableau ci-dessous montre un exemple de tableau de contingence (Schwartz, 3e édition, p79) ; cet exemple est similaire aux précédents, si ce n'est que l'on a considéré un plus grand nombre de modalités pour la variable couleur des cheveux, et que la nouvelle variable introduite (couleur des yeux) comporte trois modalités. Ces modalités remplacent les échantillons considérés dans la section 11.2 page 109. Ainsi, la modalité « bleu » par exemple peut être lue : « échantillon issu de la population des individus aux yeux bleus ». La taille de cet échantillon n'est cependant plus maîtrisée.

Couleur des yeux	Couleur des cheveux					fréquence
	blonds	**bruns**	**roux**	**noirs**	**total**	
bleus	25	9	7	3	**44**	**44/124**
gris	13	17	7	10	**47**	**47/124**
marrons	7	13	5	8	**33**	**33/124**
total	**45**	**39**	**19**	**21**	**124**	
fréquence	**45/124**	**39/124**	**19/124**	**21/124**	**124/124**	

Les effectifs calculés s'obtiennent comme précédemment. Ainsi, l'effectif calculé relatif au couple « blonds, marrons » sera : 45/124×33/124×124 = 11,9.

REMARQUES

i. Pour alléger les calculs, on peut remarquer que l'effectif calculé relatif à la cellule localisée ligne l, colonne c est égal au rapport
   - du produit du total de la ligne l et du total de la colonne c,
   - et du total général.

ii. La somme des effectifs calculés, soit en ligne, soit en colonne, coïncide avec les mêmes sommes de nature expérimentale. Cette remarque permet une vérification partielle des calculs.

iii. Dans la présentation des calculs, on a procédé au « mélange » des résultats sans plus tenir compte de la couleur des yeux (ce qui conduit à sommer les lignes du tableau). On peut de façon équivalente mélanger les résultats expérimentaux sans plus tenir compte de la couleur des cheveux, ce qui conduira à sommer les colonnes du tableau de contingence pour obtenir la répartition de référence. On pourra vérifier que les résultats du calcul sont strictement les mêmes, ce que l'on attend compte tenu du rôle symétrique joué par les deux variables étudiées.

SOUS L'HYPOTHESE NULLE D'INDEPENDANCE entre les deux variables, K^2
EST DISTRIBUE SELON un χ^2 à :

(nombre de modalités de la première variable - 1) × (nombre de modalités de la seconde variable - 1)

DEGRES DE LIBERTE.

Les CONDITIONS DE VALIDITE sont encore : $C_j \geq 5$.

3. La suite des étapes est habituelle

En particulier, la règle de décision s'établit comme suit :

- si la valeur calculée de K^2, notée ${K_c}^2$, est inférieure à χ^2_α (ddl), on ne rejette pas l'hypothèse d'indépendance des deux variables.
- si la valeur calculée ${K_c}^2$ est supérieure à χ^2_α (ddl), on rejette l'hypothèse d'indépendance des deux variables. On dira alors que les deux variables sont liées, au risque α.

Exemple :
Dans l'exemple ci-dessus, la valeur de ${K_c}^2$, résultant de la sommation de 12 termes, est 15,1. Le nombre de degrés de liberté est : $(4 - 1)\times(3 - 1) = 6$, la valeur de $\chi^2_{0,\,05}$ associée étant 12,6 (lue dans une table). On rejette donc ici l'hypothèse d'indépendance : couleur des cheveux et couleur des yeux sont liées, ou encore sont dépendantes. Voyons une illustration de cette dépendance. Sur la base des données expérimentales on a :
P(yeux bleus) = 44/124 = 0,35
P(yeux bleus / cheveux blonds) = 25/45 = 0,56
La connaissance de la couleur des cheveux (ici la modalité « blond ») modifie la répartition de la couleur des yeux (ici la fréquence de la modalité « bleu » qui évolue de 0,35 à 0,56). Le test indique que cette modification est significative. En réalité la valeur de ${K_c}^2$ ci-dessus chiffre dans leur ensemble les différences entre $P(A\ /\ B)$ et $P(A)$, c'est-à-dire les écarts de $P(A$ et $B)$ par rapport au produit $P(A)P(B)$, où A est un événement relatif à la couleur des yeux et B un événement relatif à la couleur des cheveux.

Résumé du chapitre

Tests du χ^2. Effectifs observés O_j, effectifs calculés C_j.
Conditions de validité générales : $C_j \geq 5$
Paramètre général :

$$K^2 = \sum_{j=1}^{\text{nombre de cases du tableau}} \frac{(O_j - C_j)^2}{C_j}$$

Comparaison d'une répartition observée à une répartition donnée (ajustement)

H_0 : La répartition vraie s'ajuste à la répartition donnée
H_1 : La répartition vraie ne s'ajuste pas à la répartition donnée
Nombre de cases = nombre de modalités
$K^2 \sim \chi^2$(nombre de modalités -1)

Comparaison de plusieurs répartitions observées (homogénéité)

H_0 : Les répartitions coïncident
H_1 : Les répartitions diffèrent
Nombre de cases = nombre de modalités × nombre de répartitions
$K^2 \sim \chi^2$((nombre de modalités -1) × (nombre de répartitions -1))

Test d'indépendance de deux variables qualitatives

H_0 : Les deux variables sont indépendantes
H_1 : Les deux variables sont liées
$K^2 \sim \chi^2$((nb de modalités de 1ère variable - 1) × (nb de modalités de 2ème variable - 1))

Dans les deux derniers cas, si l est le nombre de lignes, c le nombre de colonnes du tableau de contingence, le nombre de degrés de liberté des χ^2 est $(l - 1)(c - 1)$.

Chapitre 12

Liaison entre deux variables continues : notion de corrélation

12.1 Introduction

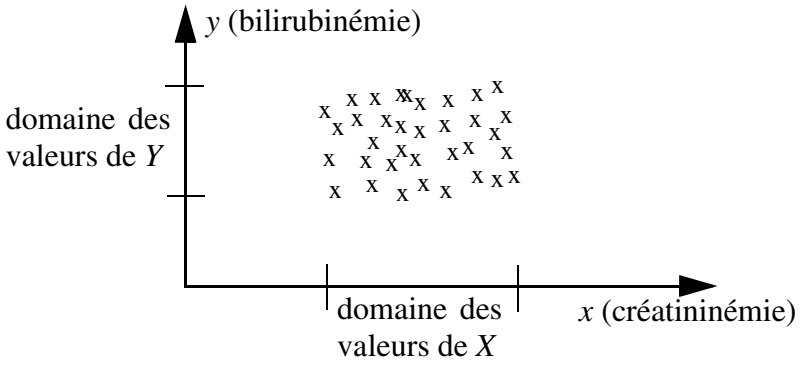
Nous avons rappelé dans le chapitre précédent la notion fondamentale d'indépendance entre deux variables qualitatives et vu la façon dont cette indépendance pouvait être mise à l'épreuve lors d'une expérience. Dans le chapitre 10, les tests mis en œuvre faisaient intervenir une variable quantitative continue et une variable qualitative encore jugées dans leurs interdépendances. Il se trouve qu'il existe une autre classe de problèmes mettant en jeu encore deux variables aléatoires, mais cette fois-ci, deux variables continues. Considérons, par exemple, deux variables aléatoires, l'insuffisance rénale (avec deux valeurs ou modalités présence-absence) et l'insuffisance hépatique (avec les deux mêmes modalités). Supposons que l'on connaisse un indicateur de la fonction rénale (ou de certains de ses aspects), la clairance à la créatinine par exemple et un indicateur de la fonction hépatique (ou de certains de ses aspects) la bilirubinémie et que le diagnostic d'insuffisance rénale soit porté lorsque la clairance est inférieure à un seuil, celui d'insuffisance hépatique lorsque la bilirubinémie est supérieure à un autre seuil. On sait résoudre (voir chapitre 11) la question de savoir si les variables insuffisance rénale et insuffisance hépatique sont indépendantes ou liées. Toutefois, compte tenu des précisions données sur l'origine des diagnostics d'insuffisance rénale et d'insuffisance hépatique, on est tenté de reformuler le problème posé en ces termes : y a-t-il un lien entre les variables aléatoires *clairance à la créatinine* et *bilirubinémie* ? Un niveau élevé de l'une est-il « annonciateur » d'un niveau élevé de l'autre ? Ou encore : la connaissance du niveau de l'une modifie-t-elle l'idée que l'on se fait du niveau de l'autre, non encore observée ? Cette dernière formulation est très proche de la formulation utilisée pour discuter de l'indépendance entre événements : la connaissance du fait qu'un événement s'est réalisé (maintenant un niveau de clairance connu) modifie-t-elle la plausibilité d'un autre événement (maintenant la bilirubinémie) ?

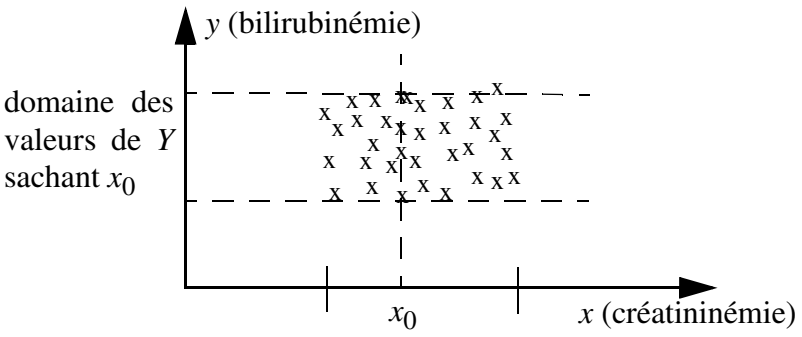
Les situations dans lesquelles on se pose naturellement la question de savoir si deux variables continues sont liées sont extrêmement fréquentes. Voilà quelques exemples :

— la consommation de cigarettes (quotidienne ou cumulée) et la capacité respiratoire sont-elles liées ?
— la gastrinémie et la quantité de cellules ECL sont-elles liées ?
— les valeurs de glycémie obtenues selon deux méthodes de dosage sur les mêmes échantillons sanguins sont-elles liées [ici, il faut l'espérer].

12.2 Abord du problème

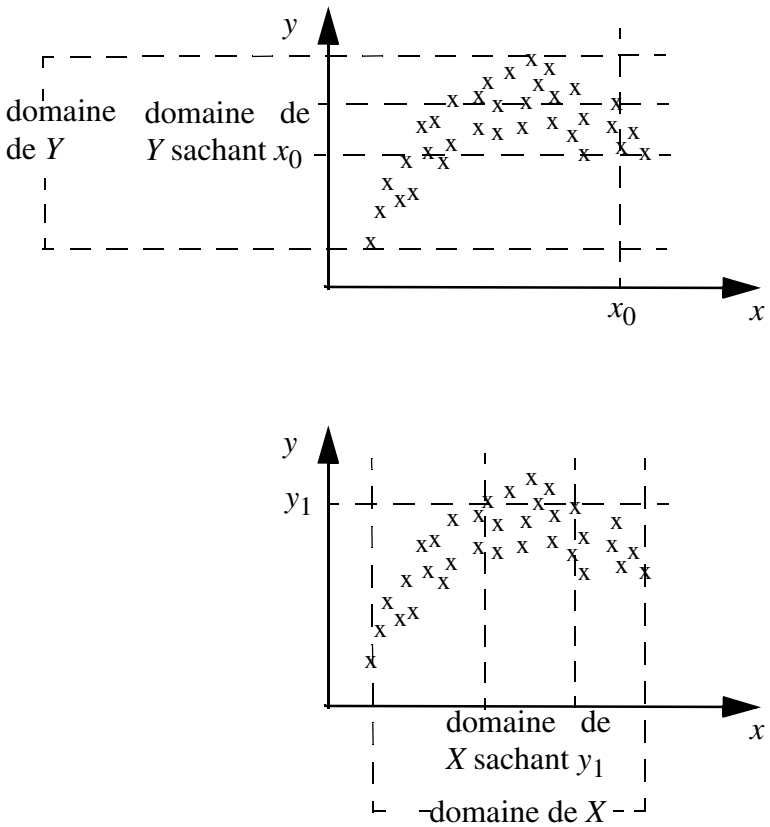
Considérons deux variables aléatoires continues X (créatininémie) et Y (bilirubinémie). Imaginons que nous ayons réalisé une expérience consistant en l'observation conjointe du niveau de ces deux variables sur un ensemble (échantillon) de n sujets. On dispose ainsi d'un ensemble de couples de valeurs x_i, y_i. La représentation naturelle - sinon la meilleure - de ces résultats est donnée dans la figure ci-dessous ; chaque couple de valeurs obtenu chez chaque individu est représenté par un point de coordonnées (créatininémie-bilirubinémie).
On lit sur un tel dessin, au moins grossièrement, le domaine des valeurs possibles de X, le domaine des valeurs possibles de Y.
Intéressons nous à un nouvel individu ; ne mesurons chez lui que la valeur de la créatininémie, x_0. Que peut-on dire alors, sur la base de cette connaissance et sur la base de l'expérience ci-dessus concernant le domaine des valeurs possibles de Y **pour ce même individu** ? On peut proposer la réponse géométrique ou visuelle indiquée sur la figure ci-dessous.





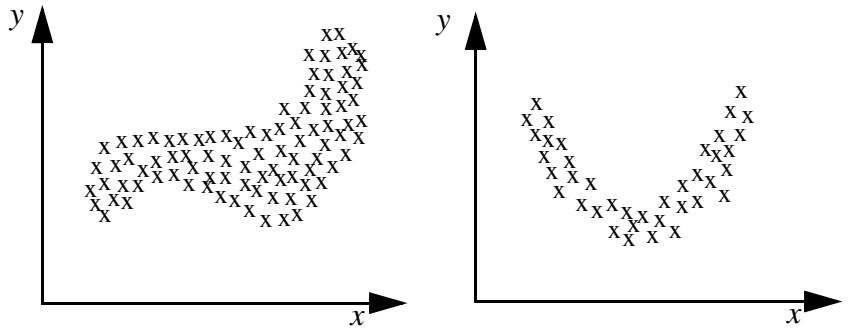
Le nouveau domaine possible - sachant x_0 - est très voisin du domaine initial ; ceci se reproduit pour toute valeur de x_0. Il est alors clair que dans cet exemple, la connaissance de X n'apporte pas d'information sur celle de Y. On a ici une situation visuelle d'un cas où les deux variables X et Y sont indépendantes. On pourrait renverser le rôle de X et Y, la conclusion serait la même.

Considérons maintenant le cas où les résultats expérimentaux produisent la représentation de la figure ci-dessous.
Dans ce cas, au contraire, on voit clairement que la connaissance de x_0 (respectivement y_1) modifie le domaine des valeurs possibles, donc attendues de Y (respectivement X) ; les deux variables X et Y sont liées.

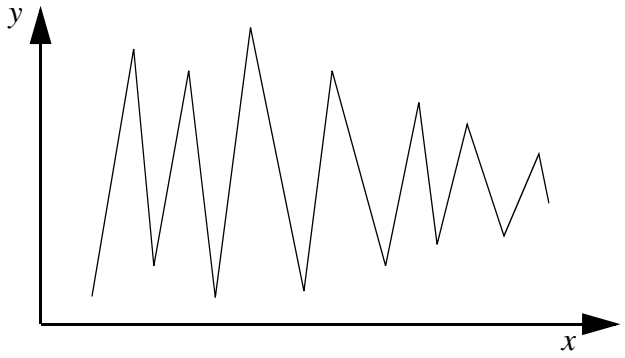


La modification ici concerne aussi bien l'amplitude du domaine que sa localisation en termes de valeurs.
L'appréciation visuelle de la dépendance correspond à l'appréciation de « l'épaisseur » de l'ensemble des points. Plus les points expérimentaux ont tendance à se répartir sur une courbe - non horizontale ni verticale - plutôt qu'à remplir une partie du plan, plus les variables sont liées.
Peut-on trouver un indicateur numérique de la force d'une telle liaison ? Au sens strict, la réponse est non.
Quelques situations de dépendance - c'est-à-dire de liaison - sont représentées sur les figures ci-

dessous.



On ne sait pas, en toute généralité, résumer en un seul nombre exprimant la liaison entre deux variables continues les résultats d'une expérience.
On ne connaît qu'un indicateur général prenant en compte non pas le degré de proximité à une courbe quelconque mais le degré de proximité à une droite : c'est le coefficient de corrélation [linéaire].
Il faut voir cependant que dans la plupart des situations réelles au cours desquelles on s'intéresse à l'examen de la liaison entre deux variables, la possibilité d'interprétation des résultats est largement fonction du caractère monotone, sinon rectiligne, de la dépendance ; que dire en termes d'interprétation d'une dépendance figurée schématiquement sur la figure ci-dessous ?



12.3 Un indicateur de covariation : le coefficient de corrélation

Cherchons alors à quantifier un phénomène de covariation, c'est-à-dire un phénomène de variation couplée entre X et Y.
On impose naturellement à l'indicateur recherché une invariance par translation : les phénomènes productifs de X et Y demeurent fondamentalement inaltérés s'ils produisent $X + a$, $Y + b$. Ainsi l'indicateur se fondera-t-il sur les valeurs $x_i - \bar{x}$ et $y_i - \bar{y}$. Par ailleurs, on souhaite que l'indicateur ne dépende pas des unités exprimant X et Y ; alors on travaillera sur

$$x_{ri} = \frac{x_i - \bar{x}}{s_X} \text{ et } y_{ri} = \frac{y_i - \bar{y}}{s_Y}$$

Maintenant si Y et Y présentent un caractère de covariation, c'est que de façon fréquente, sinon systématique

- soit les variables varient dans le même sens, c'est-à-dire lorsque x_i est grand (i.e. x_{ri} positif par exemple), y_i l'est également (i.e. y_{ri} positif), que lorsque x_i est petit ($x_{ri} < 0$) y_i l'est également ($y_{ri} < 0$) ; dans ce cas, le produit $x_{ri}y_{ri}$ est positif.
- soit les variables varient en sens contraire : lorsque x_i est grand, y_i est petit, lorsque x_i est petit, y_i est grand ; dans ce cas le produit $x_{ri}\ y_{ri}$ est fréquemment négatif.

Compte tenu de l'analyse précédente, on choisit pour indicateur de la covariation ou corrélation le nombre :

$$r = \frac{1}{n-1}\sum_i x_{ri}y_{ri}$$

Ainsi

- si r est grand, c'est le signe d'une covariation dans le même sens de X et Y ;
- si r est petit (c'est-à-dire grand en valeur absolue et négatif), c'est le signe d'une covariation de X et Y en sens contraire ;
- si r est voisin de zéro, c'est le signe d'une absence de covariation.

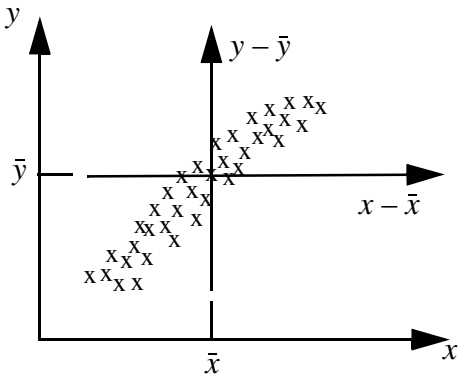
Retenons, exprimé sur la base des observations expérimentales :

$$r = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}$$

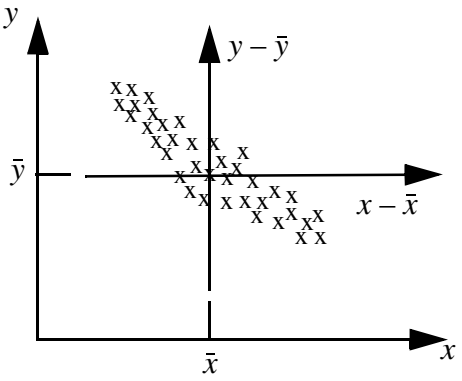
Le numérateur de cette expression est appelé la covariance expérimentale des deux variables X et Y, notée $cov(X, Y)$, dont on montre qu'elle s'exprime aussi sous la forme

$$cov(X, Y) = \frac{n}{n-1}\left(\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}\right)$$

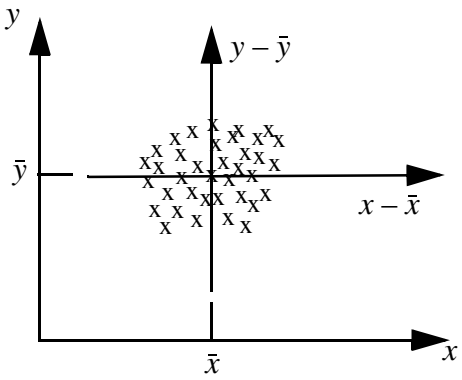
Les figures ci-dessous présentent diverses situations relativement au coefficient de corrélation expérimental.



$r>0$, grand



$r<0$, $|r|$ grand



r voisin de zéro

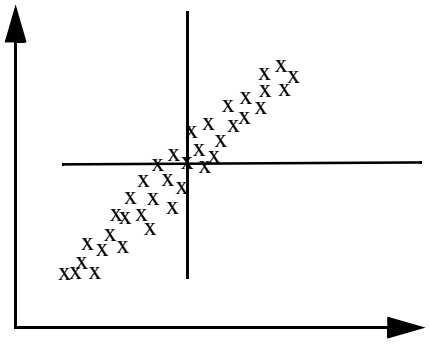
Propriétés numériques fondamentales de r :

- r a toujours une valeur comprise entre -1 et 1 ;
- r prend la valeur -1 (respectivement 1) si et seulement si pour une certaine valeur de a et b on a pour tout i $y_i = ax_i + b$ avec a négatif (respectivement $a > 0$).

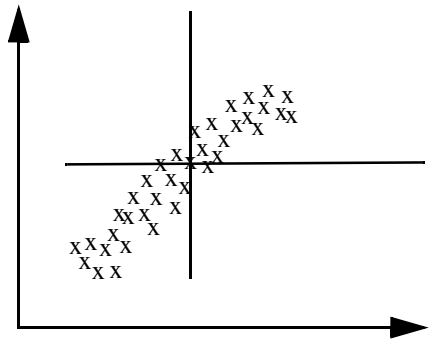
Remarques :

- plus r est grand en valeur absolue, plus les variables sont dites corrélées,
- la valeur absolue de r décroît,
  - — lorsque s'estompe le caractère rectiligne du « nuage » des observations expérimentales,
  - — lorsque s'épaissit ledit nuage,
- une valeur absolue très faible du coefficient de corrélation ne permet pas de conclure à l'indépendance de deux variables. Deux variables indépendantes présenteront en revanche un coefficient de corrélation expérimental très faible en valeur absolue.

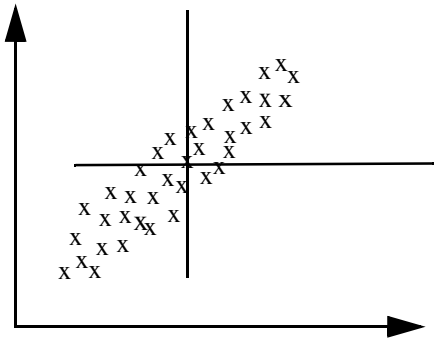
Quelques exemples sont présentés ci-dessous pour fixer les idées.



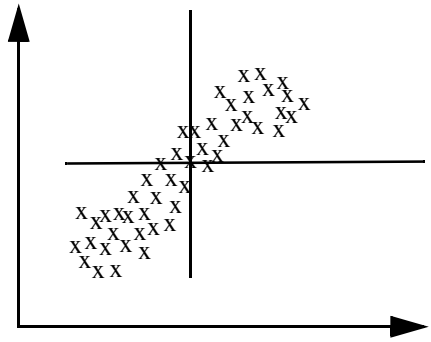
$r \approx 0{,}9$



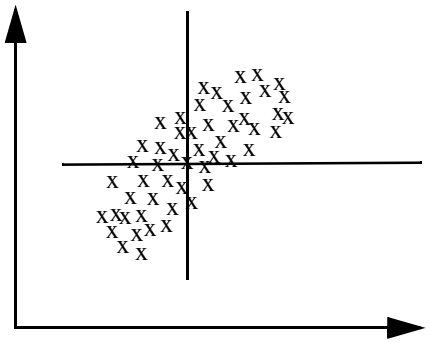
$r \approx 0{,}7$



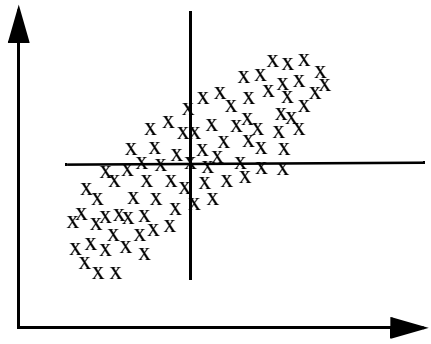
$r \approx 0{,}7$



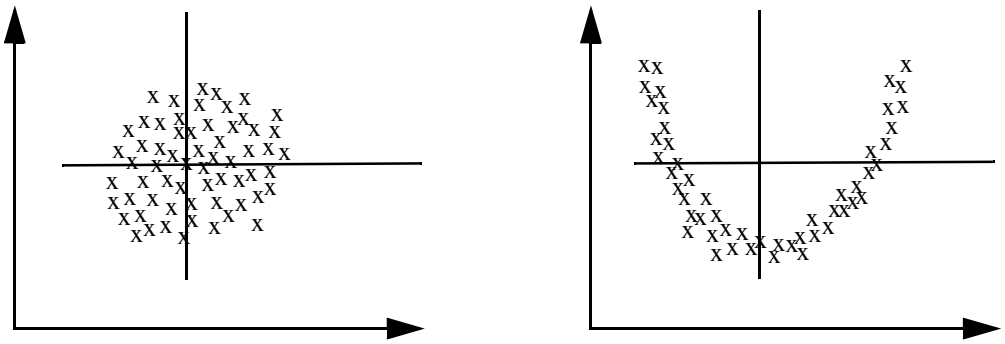
$r \approx 0{,}6$



$r \approx 0{,}5$



$r \approx 0{,}5$



$r \approx 0$ $\qquad\qquad$ $r \approx 0$

Remarque complémentaire

Le coefficient de corrélation linéaire est, au même titre que toute statistique, soumis aux fluctuations d'échantillonnage. La question se pose alors de savoir que faire de cet indicateur en termes d'inférences. Par exemple, avant de conclure que les deux variables sont corrélées, peut-on se garantir du risque de l'observation d'un coefficient de corrélation nul sur une plus grande série d'observations ? On se retrouve dans le contexte des tests d'hypothèses avec ici une difficulté supplémentaire qui tient au fait que l'on n'a pas quitté le niveau expérimental, le niveau intuitif. Il convient de trouver une contrepartie **vraie** à ce coefficient de corrélation expérimental r.

12.4 Le coefficient de corrélation vrai

Cherchons à substituer de la façon la plus naturelle possible des grandeurs vraies aux grandeurs expérimentales constitutives de r. On note l'apparition au dénominateur de s_X et s_Y auxquelles on substitue naturellement σ_X et σ_Y, les écarts types vrais de X et Y. Au numérateur on remarque $\bar{x}$ et $\bar{y}$ auxquels on substitue $E(X)$ et $E(Y)$ les moyennes vraies de X et Y. Reste au numérateur une moyenne expérimentale (lisons n à la place de n-1) ; on lui substitue une moyenne vraie : moyenne vraie du produit $[X - E(X)][Y - E(Y)]$, soit $E\{[X - E(X)][Y - E(Y)]\}$.
Cette moyenne vraie dépendant de X et Y à la fois s'appelle **covariance vraie** de X et Y.
Finalement, on obtient la contrepartie vraie notée ρ :

$$\rho(X, Y) = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sigma_X \sigma_Y}$$

Remarque : à propos des notions d'espérance, de covariance vraie, de coefficient de corrélation vrai, voir le chapitre 5.

12.5 Mise à l'épreuve de la nullité du coefficient de corrélation vrai ρ

Des calculs théoriques complexes, et imposant un certain nombre de restrictions, qui, dépassant le cadre de ce cours ne seront pas mentionnés, permettent de calculer la distribution de r sous l'hypothèse - retenue comme hypothèse nulle - de nullité du coefficient de corrélation vrai ρ. Il s'agit d'une famille de distributions indexées par un entier appelé nombre de degrés de liberté. La mise en œuvre du test est alors conventionnelle :

- $H_0 : \rho = 0$ [les variables ne sont pas corrélées],
  $H_1 : \rho \neq 0$ [les variables sont corrélées]
- Paramètres du test : coefficient de corrélation expérimental

$$r = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y}$$

- sous H_0, r suit une distribution connue, dite du coefficient de corrélation à n-2 degrés de liberté où n est le nombre de couples (xi, yi) expérimentaux. L'intervalle de pari pour r est de la forme
  $IP_{1-\alpha} = [-corr_\alpha(n-2) ; corr_\alpha(n-2)]$, $corr_\alpha(n-2)$ étant lue dans une table.

  Conditions de validité
  Les conditions de validité sont complexes et expriment que toute combinaison linéaire des variables X et Y est distribuée selon une loi normale. Autrement dit, toute variable $aX + bY$ où a et b sont deux nombres quelconques doit être normale.
  Pour la commodité de l'expression, on énoncera les conditions de validité sous le néologisme « distribution de (X, Y) binormale ».
- la suite de la mise en œuvre est standard.

Quelques exemples numériques
Au risque 5 % :
n = 10, $IP_{0,95}$ = [-0,632 ; 0,632], ddl = 8
n = 20, $IP_{0,95}$ = [-0,444 ; 0,444], ddl = 18
n = 50, $IP_{0,95}$ = [-0,280 ; 0,280], ddl = 48

Ainsi, par exemple, pour pouvoir conclure à la corrélation, lorsque l'on dispose de 20 observations (20 couples (x_i, y_i)), le coefficient de corrélation expérimental doit être supérieur à 0,444, ou inférieur à -0,444.

Résumé du chapitre

1. La corrélation entre deux variables aléatoires quantitatives X et Y se mesure à l'aide du coefficient de corrélation vrai :

   $$\rho(X, Y) = \frac{E\{[X - E(X)][Y - E(Y)]\}}{\sigma_X \sigma_Y}$$

   Propriétés :

   - $\rho(X, Y) \in [-1 ; 1]$
   - Si X, Y indépendantes, alors $\rho(X, Y) = 0$

2. Disposant d'un échantillon de n couples (x_i, y_i) on définit le coefficient de corrélation expérimental :

   $$r = \frac{\frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_X s_Y} = \frac{\frac{n}{n-1}\left(\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}\right)}{s_X s_Y}$$

   Propriété : $r \in [-1 ; 1]$
3. Il existe un test de nullité du coefficient de corrélation vrai dont le paramètre est r.
4. Indépendance et corrélation sont des notions différentes ; deux variables dont le coefficient de corrélation vrai est nul peuvent être liées.

Chapitre 13

A propos des tests d'hypothèses

13.1 Rappels et précisions

1. LES TESTS PRENNENT EN COMPTE DES HYPOTHESES SYNTHETIQUES

   On a vu que les tests reposent sur l'énoncé de deux hypothèses exclusives. Il y a parfois beaucoup de chemin à parcourir entre la formulation d'un problème médical et sa formulation en termes statistiques. Reprenons l'exemple des souris du chapitre 9. Le problème fondamental est celui de l'activité du traitement. Cette activité peut avoir bien d'autres manifestations que la modification de la fréquence d'apparition des cancers. On peut penser à un effet portant sur l'âge de survenue de la maladie, portant sur la vitesse de développement des tumeurs etc… On ne peut répondre simultanément à toutes ces questions, par l'intermédiaire d'un test du moins : les tests ne permettent de répondre qu'à des questions simples.

2. ON NE CHOISIT PAS LE SEUIL DE SIGNIFICATION

   Que dirait-on d'un médecin annonçant : j'aime le risque alors j'ai choisi un risque α de 0,4 et le traitement que je propose est efficace (ou actif) à ce risque ?

   $\alpha = 0,05$ est conventionnel

3. ON NE DIT PRATIQUEMENT JAMAIS : L'EXACTITUDE DE L'HYPOTHESE NULLE EST DEMONTREE

4. ON N'ENCHAINE PAS LES TESTS DE FAÇON INCONSIDEREE

   En effet, les risques de conclusion à tort augmentent alors.
   Par exemple, supposons que l'on veuille tester l'égalité à une valeur donnée de deux proportions (ex : succès d'une intervention chirurgicale dans deux services hospitaliers, le pourcentage de succès sur la France étant par ailleurs connu (données de l'année précédente par exemple)). Que se passe-t-il si l'on effectue deux tests successifs dont les hypothèses nulles

sont :
service 1 : $p_1 = p_h$; puis service 2 : $p_2 = p_h$.
Le risque de première espèce global de la procédure exprime la probabilité de dire au moins une fois (soit au cours du premier test soit au cours du second) H_1 alors que H_0 est vraie les deux fois :

P(conclure H_1 au moins une fois si H_0 est vraie) = 1 - P(ne rejeter H_0 aucune des deux fois si H_0 est vraie)

Or P(ne pas rejeter H_0 si H_0 est vraie) = 1 - α

Donc P(ne rejeter H_0 aucune des deux fois si H_0 est vraie) = $(1 - \alpha)^2$

d'où un risque total = $1 - (1 - \alpha)^2$

Exemple
Si $\alpha = 0{,}05$, le risque global est environ 0,10.
Cette situation s'aggrave si le nombre de tests s'accroît. Ainsi, dans le cas de

— 3 services le risque global est 0,14
— 10 services le risque global est 0,40
— 100 services le risque global est 0,994

Cela signifie par exemple que dans le cas où 10 services sont à comparer à une référence il y a 4 chances sur 10 pour qu'au moins une fréquence expérimentale s'écarte de façon significative de la valeur de référence, alors qu'en réalité tous les résultats sont homogènes. Si l'on prend la fréquence expérimentale la plus différente de la valeur de référence, le test permettra de conclure, à tort, avec une probabilité supérieure à 0,4.

En fait, lorsque l'on désire faire des comparaisons multiples, des tests spécifiques doivent être utilisés de façon que les conclusions puissent être tirées avec un risque d'erreur α global de 5 %.

5. IL EST DANGEREUX ET ERRONE DE CHOISIR LES HYPOTHESES AU VU DES DONNEES

Lorsque l'on opère de cette façon, on a en réalité réalisé plus ou moins consciemment un nombre indéterminé de tests que l'on a jugés non concluants.

LA STRATEGIE D'ANALYSE DES DONNEES DOIT ETRE FIXEE CLAIREMENT AVANT LA REALISATION DE L'EXPERIENCE

13.2 Jugement d'interprétation - La causalité

Lorsqu'un test permet de conclure, le premier jugement que l'on tire est un **jugement de signification** (au sens de différences significatives).
Peut-on se livrer à des interprétations plus fines, s'exprimer en termes de causalité ? Il s'agit là du **jugement d'interprétation**. La question est ici de savoir si c'est la présence ou l'absence d'un caractère qui cause - est à l'origine de - ces différences ? C'est un problème de bon sens fondamentalement mais qui suppose également un abord spécifique.

Caractère contrôlé ; caractère aléatoire

On dit d'un **caractère** qu'il est **contrôlé** lorsque sa détermination nous appartient.

Exemple : on s'intéresse à l'effet d'un traitement sur la survenue d'un type de cancer chez des souris. Le caractère absence ou présence du traitement peut être contrôlé.

Dans le cas contraire, on dit que le **caractère** est **aléatoire**.

Exemple : couleur des cheveux, couleur des yeux.

Lorsqu'on envisage un problème de liaison entre deux variables (cela recouvre tous les problèmes que l'on a rencontrés) un au plus des caractères peut être contrôlé.

Démarche expérimentale

Lorsque l'expérience se conduit avec un facteur contrôlé, on dit que l'on suit une **démarche expérimentale**. Dans ce cas, au cours de la constitution de l'échantillon qui permettra de mettre en œuvre les tests, on reste libre d'un caractère (par exemple la $x^{\text{ème}}$ souris sera ou ne sera pas traitée).

Démarche d'observation

Lorsque l'expérience se conduit sur la base de deux facteurs aléatoires, on dit que l'on suit une **démarche d'observation**.

PRINCIPE FONDAMENTAL

La discussion de la causalité ne se conçoit pas sans contrôle d'un des deux caractères étudiés.
Autrement dit, on ne peut affirmer la causalité hors d'une démarche expérimentale.
Seule cette démarche, en effet, permet d'assurer que les individus constituant l'échantillon sont comparables (homogènes) sauf pour ce qui concerne le caractère contrôlé. Encore faut-il assurer cette homogénéité par **tirage au sort**. On parle aussi de **randomisation**.

Quelques exemples.

i. On veut comparer les pourcentages de complications à l'accouchement dans deux maternités, l'une (1) dotée de moyens chirurgicaux les plus modernes, l'autre (2) dotée d'un plateau technique plus modeste. On effectue une étude d'observation au cours de laquelle on obtient des pourcentages expérimentaux de 80 % (100 accouchements) et 30 % (150 accouchements). La différence est significative au risque 5 %. Les fréquences vraies de complications sont différentes au seuil 5 %. C'est incontestable. On ne saurait pourtant en conclure que pour diminuer

les risques de complication il suffit de réduire le plateau technique ! Les recrutements sont très probablement différents dans ces deux maternités, les grossesses à risque se rencontrant plus fréquemment dans la maternité (1). Si l'on veut mettre à l'épreuve cette causalité, il faut adopter une démarche expérimentale randomisée, c'est-à-dire affecter par tirage au sort chaque femme d'un échantillon à l'une ou l'autre maternité et refaire l'analyse.

ii. **Les essais thérapeutiques**

Dans le cas de la comparaison de deux traitements, ou de la mise en évidence de l'effet d'un traitement, c'est-à-dire dans le contexte des essais thérapeutiques, des précautions et une méthodologie particulières doivent être appliquées en ce qui concerne le déroulement de l'expérience. En particulier, il ne faut pas méconnaître l'effet dit effet placebo (« je plairai » en latin) résultant de l'administration d'un traitement inactif (le placebo) à un malade. Cet effet est complexe à analyser mais il faut autant que possible en tenir compte dans l'appréciation de l'effet d'un traitement. C'est la raison pour laquelle en règle générale, pour mettre en évidence l'effet d'un traitement, on constituera deux groupes de patients, l'un recevant le traitement, l'autre un placebo administré dans les mêmes conditions.
Le groupe recevant le placebo se nomme groupe **témoin**.
En outre, le malade ne devra pas savoir s'il reçoit le traitement ou le placebo : on parle de procédure d'**insu** ou « d'**aveugle** ». L'attribution du traitement ou du placebo pourra être effectuée également à l'insu du médecin ; on parlera alors d'essai en **double insu** ou « **double aveugle** ».
Les essais thérapeutiques comparatifs ayant pour objet la comparaison de deux traitements relativement voisins seront réalisés dans les mêmes conditions. Dans de tels essais, l'un des traitements est le meilleur traitement connu au début de l'essai (traitement de référence), l'autre le traitement nouveau, expérimental. On appellera encore groupe témoin l'ensemble des patients recevant le traitement de référence.

Exemple : comparaison d'un traitement anticoagulant et d'un traitement anticoagulant + antiagrégant plaquettaire chez les malades porteurs d'une fibrillation auriculaire.

Les types d'essais évoqués ci-dessus sont dits essais thérapeutiques à visée **explicative**.
Il existe par ailleurs des essais dits **pragmatiques** dont l'objectif est de comparer des traitements éventuellement très différents ; dans ces essais la procédure d'aveugle n'a généralement plus de sens, mais le caractère de répartition au hasard des patients dans les deux groupes de traitement doit être maintenu.

Exemple : comparaison d'un traitement chirurgical et d'un traitement médical dans une certaine maladie.

Pour en savoir plus, voir, dans la bibliographie donnée en fin de polycopié, l'ouvrage « L'essai thérapeutique chez l'homme ».

Chapitre 14

Le second problème d'induction statistique : l'estimation - Intervalle de confiance

14.1 Introduction

Le problème de l'estimation statistique est le suivant : on cherche à connaître les valeurs de certaines grandeurs grâce à des observations réalisées sur un échantillon. Très souvent, ces grandeurs sont des moyennes. On a vu que la moyenne joue un rôle fondamental - comme résumé de la variabilité - dans l'étude des variables quantitatives. Egalement un grand nombre de problèmes statistiques consistent en la détermination de la moyenne vraie, sur la base d'observations réalisées sur un échantillon. Cependant, on peut aussi chercher à connaître d'autres valeurs, comme par exemple les variances (exemple c. ci-dessous).

Exemples :

a. quelle est la fréquence de survenue de tel type de cancer chez les souris ?
b. quelle est la glycémie de ce patient ? dans ce cas on identifie (c'est un modèle, pas la réalité inattaquable) la moyenne vraie des dosages à la vraie valeur de la glycémie.
c. quelle est la variance de la glycémie mesurée chez ce patient ?

Il est bien sûr impossible de répondre à ces questions au sens strict. De la même façon qu'il était impossible de trancher avec certitude entre deux hypothèses.
On apporte généralement deux types de réponses à ces questions :

1. On produit une valeur qui nous semble être la meilleure possible : on parle alors d'**estimation ponctuelle**.
2. On produit un intervalle de valeurs possibles, compatibles avec les observations. C'est la no-

tion d'**intervalle de confiance**.

14.2 Estimation ponctuelle

14.2.1 Définition

A partir des données expérimentales, on construit une nouvelle variable dont la valeur « se rapproche » de celle de la grandeur qu'on cherche à connaître. Cette nouvelle variable est l'**estimateur** de la grandeur. On notera θ la grandeur à estimer et T ou $T(\theta)$ son estimateur.

14.2.2 Propriétés

Les estimateurs sont des fonctions des échantillons : ce sont donc des variables aléatoires qui possèdent une densité de probabilité, et le plus souvent, une moyenne (espérance mathématique) et une variance. Ces deux grandeurs permettent de comparer, dans une certaine mesure, les estimateurs entre eux.

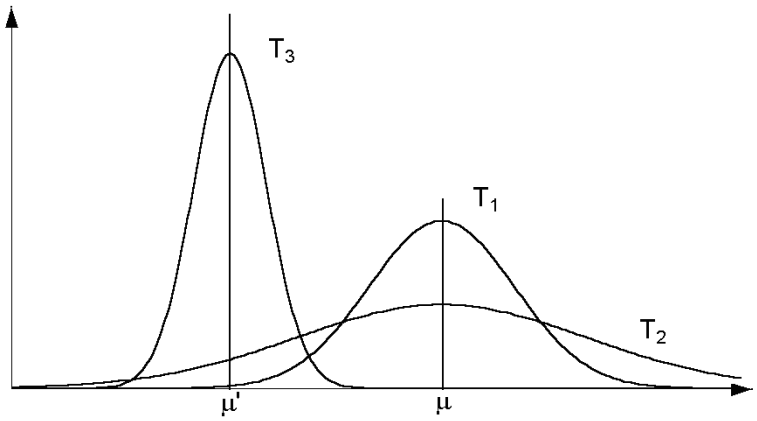


Figure 12 : densité de probabilité de 3 estimateurs T_1, T_2 et T_3

La figure 12 représente les densités de probabilité de 3 estimateurs T_1, T_2 et T_3 d'une moyenne μ.

14.2.2.1 Biais

On voit sur la figure 12 que T_1 et T_2 sont centrés autour de μ, tandis que T_3 a pour moyenne μ' inférieure à μ. Cette notion est définie plus précisément de la manière suivante :
Le **biais** d'un estimateur, noté $B(T)$, est la différence moyenne entre sa valeur et celle de la quantité qu'il estime. On a :

$B(T) = E(T - \theta) = E(T) - \theta$

Ici, on a : $B(T_1) = E(T_1 - \mu) = E(T_1) - \mu = 0$
de même : $B(T_2) = 0$
mais : $B(T_3) = E(T_3 - \mu) = E(T_3) - \mu = \mu' - \mu < 0$

On dit que T_1 et T_2 sont des estimateurs sans biais de μ, et que T_3 est un estimateur biaisé de μ.

14.2.2.2 Variance

La variance d'un estimateur est définie de la manière usuelle :
$var(T) = E[T - E(T)]^2$
Si deux estimateurs sont sans biais, le meilleur est celui qui a la variance la plus petite : en effet, ses valeurs sont « en moyenne » plus proches de la quantité estimée.
Par exemple, sur la figure ci-dessus, on voit que $var(T_1) < var(T_2)$. On peut donc conclure que T_1 est un meilleur estimateur de μ que T_2.
Quand des estimateurs sont biaisés, en revanche, leur comparaison n'est pas aussi simple : un estimateur peu biaisé, mais de variance très faible, pourrait même, en pratique, être préféré à un estimateur sans biais, mais de variance grande.

14.2.2.3 Erreur quadratique moyenne

L'erreur quadratique moyenne est une grandeur permettant de comparer des estimateurs entre eux, qu'ils soient biaisés ou sans biais. Elle est définie de la manière suivante :
$EQM(T) = E[(T - \theta)^2]$

On démontre facilement qu'on peut relier l'erreur quadratique moyenne, l'espérance et la variance d'un estimateur par l'expression suivante :
$EQM(T) = var(T) + [E(T) - \theta]^2 = var(T) + B(T)^2$

En particulier, l'erreur quadratique moyenne des estimateurs sans biais est égale à leur variance.

14.2.3 Exemple

On a souvent utilisé, dans ce cours, les quantités $\bar{x}$, moyenne expérimentale, et s^2, variance expérimentale. La variable aléatoire moyenne expérimentale, notée $\overline{X_n}$, a été étudiée au chapitre 7. De la même manière, on peut considérer la variable aléatoire variance expérimentale S_n^2, définie par :

$$S_n^2 = \frac{n}{n-1}[\overline{(X^2)_n} - \overline{X_n}^2]$$

où $\overline{(X^2)_n}$ est la variable aléatoire « moyenne expérimentale de X^2 ».

On va calculer $E(S_n^2)$. On rappelle que si U est une variable aléatoire, sa moyenne expérimentale

a les propriétés suivantes :

$E(\overline{U_n}) = E(U)$ (1) et $var(\overline{U_n}) = \frac{1}{n} var(U)$ (2)

On a par ailleurs :
$var(U) = E(U^2) - [E(U)]^2$ et donc $E(U^2) = var(U) + [E(U)]^2$ (3).

On peut maintenant calculer $E(S_n{}^2)$. Soit X une variable aléatoire d'espérance $E(X) = \mu$ et de variance $var(X) = \sigma^2$. On a :

$$E(S_n^2) = \frac{n}{n-1}[E(\overline{(X^2)_n}) - E(\overline{X}_n{}^2)]$$

Mais $E(\overline{(X^2)_n}) = E(X^2) = \sigma^2 + \mu^2$ d'après (1) et (3),

et $E(\overline{X}_n{}^2) = var(\overline{X_n}) + [E(\overline{X_n})]^2 = \frac{\sigma^2}{n} + \mu^2$ d'après (3), (2) et (1),

et finalement : $E(S_n^2) = \frac{n}{n-1}\left[\sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2\right] = \sigma^2$.

$S_n{}^2$ est donc un estimateur sans biais de σ^2.

14.3 Intervalle de confiance

Bien que des intervalles de confiance soient définissables pour toute quantité estimée, leur détermination est le plus souvent difficile. Nous nous limiterons donc dans ce cours à la définition des intervalles de confiance des moyennes vraies.

14.3.1 Exemple d'une proportion

L'idée directrice est la suivante : on souhaite associer à une valeur expérimentale p_0 un intervalle appelé INTERVALLE DE CONFIANCE qui ait « de bonnes chances » de contenir la valeur vraie de la proportion. Que signifie de « bonnes chances » ? Si l'on effectue un grand nombre de fois l'expérience - chaque expérience produisant un pourcentage observé p_0 - on construit autant d'intervalles de confiance. On voudrait qu'un grand nombre de ces intervalles contienne la valeur vraie p. Par exemple que 95 % des intervalles en gros contiennent p. On parlera alors d'intervalle de confiance DE NIVEAU 0,95 ou d'intervalle de confiance AU RISQUE 0,05. On considérera généralement des intervalles de confiance de niveau 1-α. La valeur α sera alors le risque - ou la probabilité - pour qu'un intervalle de confiance ne contienne pas la proportion vraie p.

DE FACON GENERALE, L'INTERVALLE DE CONFIANCE AU RISQUE α D'UNE VALEUR QUE L'ON CHERCHE A ESTIMER EST UN INTERVALLE QUI CONTIENT AVEC UNE PROBABILITE 1 - α LA VALEUR CHERCHEE ; IL S'AGIT D'UN INTERVALLE QUE

L'ON DEVRA ETRE EN MESURE DE CONSTRUIRE A L'ISSUE D'UNE EXPERIENCE PORTANT SUR UN ECHANTILLON.

Comment construire de tels intervalles ? C'est facile graphiquement.

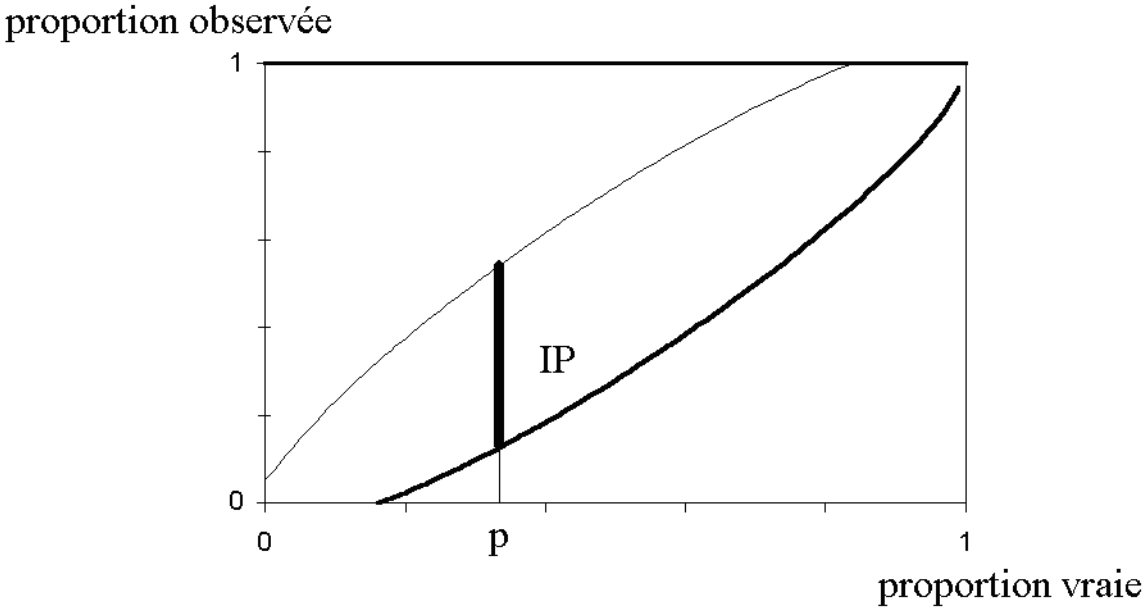


Figure 13

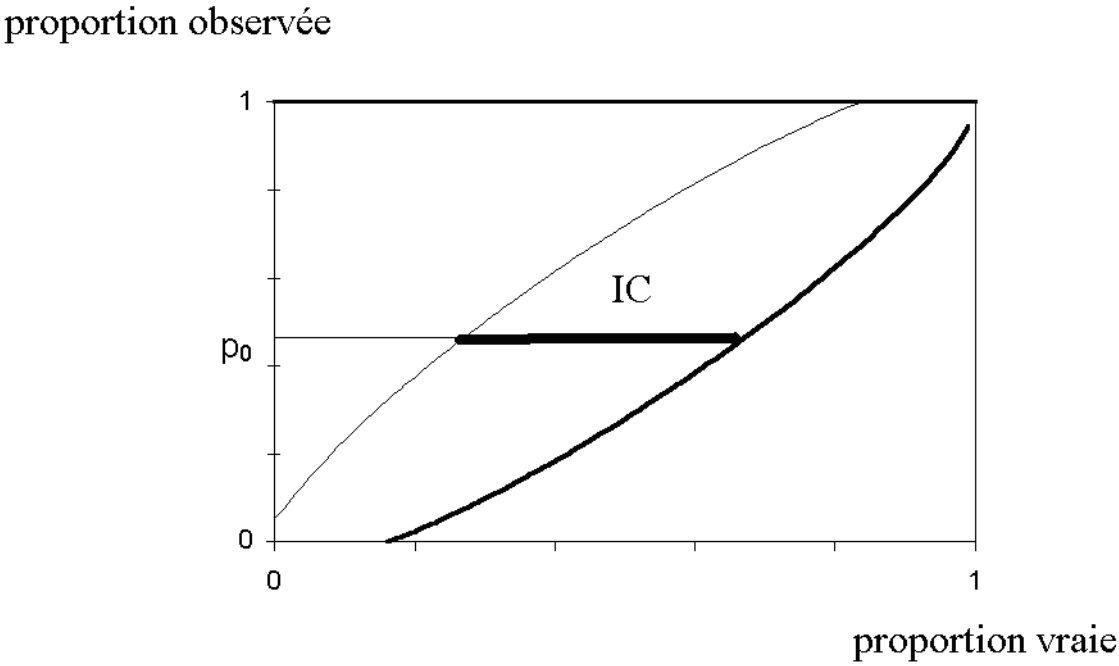


Figure 14

Considérons la figure 13. On a porté en abscisses une échelle 0-1 de mesure de proportions vraies, en ordonnées une échelle de mesure de proportions observées. Donnons nous une valeur de proportion vraie ; on sait associer à cette valeur un intervalle de pari de niveau 0,95 de la proportion expérimentale que l'on est susceptible d'obtenir au cours d'une expérimentation conduite sur n individus. Cet intervalle de pari peut être représenté sur l'échelle verticale. Si l'on opère cette représentation pour toutes les valeurs possibles d'une proportion vraie, on obtient un domaine limité par les deux courbes représentées sur la figure.
Considérons alors un problème mettant en jeu une proportion vraie, p. Supposons que nous fassions un ensemble d'expériences, chaque expérience portant sur n individus étant productive d'une valeur de proportion expérimentale p_0. On peut associer à chacune de ces expériences un point de coordonnées (p, p_0) sur la figure 13. Compte tenu de la construction précédente, on peut affirmer que ces points appartiendront 95 fois sur cent (c'est-à-dire dans 95 % des expériences) au domaine limité par les deux courbes, et ceci quelle que soit la valeur de p.
Maintenant supposons qu'une expérience unique ait été réalisée, produisant une valeur de proportion, p_0. Le problème est, sur la base de cette valeur, de définir un intervalle ayant de bonnes chances de contenir la valeur inconnue de la proportion vraie. La solution, immédiate, est fournie par la figure 14. Il suffit de trancher le domaine limité par les deux courbes DANS L'AUTRE SENS. Cet intervalle contiendra 95 fois sur cent la véritable valeur de la proportion.
Ainsi, si on adopte cette stratégie de construction, on aura pour chaque valeur observée p_0 un intervalle qui contiendra p avec la probabilité 0,95.
Le problème est résolu. Maintenant, ce qui est simple sur un dessin est compliqué en termes de calcul et il existe des tables d'intervalles de confiance et des formules toutes faites permettant de former des intervalles de confiance approchés.

14.3.2 Intervalle de confiance approché d'une proportion vraie

On montre qu'une bonne approximation de l'intervalle de confiance de niveau 1 - α de p, fondé sur la valeur expérimentale p_0, p_0 étant calculée sur n individus, est donnée par l'intervalle ci-dessous :

$$IC_{1-\alpha} = \left[p_0 - u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \; ; \; p_0 + u_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \right]$$

Notons $p_{\min}$ et $p_{\max}$ les bornes de cet intervalle.
Cette approximation n'est jugée satisfaisante que sous les CONDITIONS DE VALIDITE suivantes : $np_{\min} \geq 5$, $n(1\text{-}p_{\max}) \geq 5$

LORSQUE LES CONDITIONS DE VALIDITE NE SONT PAS REMPLIES, IL FAUT AVOIR RECOURS AUX TABLES.

Exemple : $n = 100$, $\alpha = 0{,}05$, $p_0 = 0{,}12$

$$IC_{0,95} = \left[0{,}12 - 1{,}96 \sqrt{\frac{0{,}12 \times 0{,}88}{100}} \; ; \; 0{,}12 + 1{,}96 \sqrt{\frac{0{,}12 \times 0{,}88}{100}} \right] = [0{,}06 \; ; \; 0{,}18]$$

conditions de validité
$100 \times 0{,}06 = 6 \geq 5$.
$100 \times (1 - 0{,}18) = 82 \geq 5$.

14.3.3 Intervalle de confiance approché d'une moyenne vraie (variable continue)

De même, il existe une expression approchée pour l'intervalle de confiance de niveau $1 - \alpha$ d'une moyenne vraie μ, intervalle fondé sur la valeur expérimentale $\bar{x}$ obtenue après une expérience portant sur n individus. Le calcul de cet intervalle suppose en outre le calcul de la variance expérimentale s^2. L'expression est la suivante :

$$IC_{1-\alpha} = \left[\bar{x} - u_\alpha \frac{s}{\sqrt{n}} \; ; \; \bar{x} + u_\alpha \frac{s}{\sqrt{n}} \right]$$

L'approximation ci-dessus n'est jugée satisfaisante que sous la
CONDITION DE VALIDITE : $n \geq 30$.

Lorsque cette condition n'est pas remplie, on ne sait plus former d'intervalle de confiance sauf si l'on peut supposer que la variable primitive X d'intérêt est normale.
Si la variable étudiée est NORMALE, alors, et sans autre condition de validité, un intervalle de confiance de niveau $1 - \alpha$ a pour expression :

$$IC_{1-\alpha} = \left[\bar{x} - t_\alpha \frac{s}{\sqrt{n}} \; ; \; \bar{x} + t_\alpha \frac{s}{\sqrt{n}} \right]$$

où t_α est associé à la distribution de Student à $(n-1)$ degrés de liberté.

Remarque (pour une variable normale encore)
Si la variance vraie de la variable étudiée, σ^2, est connue, l'intervalle de confiance a la forme suivante :

$$IC_{1-\alpha} = \left[\bar{x} - u_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; \bar{x} + u_\alpha \frac{\sigma}{\sqrt{n}} \right]$$

14.3.4 Applications

L'intervalle de confiance exprime fondamentalement, comme son nom l'indique, la confiance que l'on peut attribuer à un résultat expérimental.
IDEALEMENT TOUT PROBLEME D'ESTIMATION DEVRAIT ETRE PRODUCTIF D'UN INTERVALLE DE CONFIANCE. Ne donner qu'une estimation ponctuelle masque l'incertitude qui accompagne tout résultat.

Exemple : supposons qu'étudiant la fréquence d'un événement, on ait obtenu une fréquence observée p_0 égale à 0,12.

Supposons que cette valeur ait été obtenue sur la base de 8 individus (l'événement étudié s'est donc réalisé une fois). On peut lire dans une table spécialisée que l'intervalle de confiance de la fréquence vraie est, au risque 0,05 [0,003 ; 0,527]. Cela signifie que cette valeur observée de 12 % sur si peu d'individus ne fait qu'indiquer ceci : la fréquence vraie se situe dans le domaine 3 ‰, 52,7 %.
Supposons que cette même valeur 12 % ait été obtenue sur la base de 100 individus (l'événement étudié s'est réalisé 12 fois au cours des 100 essais). L'intervalle de confiance associé est alors proche de [0,06 ; 0,18]. Sur la base de cette valeur 12 %, on est maintenant en mesure d'affirmer, acceptant toujours un risque d'erreur de 5 pour cent, que la fréquence vraie se situe dans le domaine 6 %, 18 %, domaine beaucoup plus étroit que le précédent.

De façon générale, la longueur de l'intervalle de confiance indique la précision obtenue. Les deux exemples qui suivent montrent l'usage que l'on peut en faire.

14.3.4.1 Précision d'un sondage

Supposons que l'on s'apprête à réaliser un sondage pour estimer la prévalence d'une maladie, c'est-à-dire la proportion de la population atteinte par cette maladie à la date du sondage. On souhaite un résultat précis, c'est-à-dire que l'on souhaite par exemple que l'intervalle de confiance résultant ait une longueur au plus égale à 0,04.
On remarque que la longueur de l'intervalle de confiance ne dépend que d'une seule grandeur contrôlable, le nombre d'individus. La question est donc : combien d'individus faut-il inclure dans le sondage ?

Ce problème est simple, puisque la longueur de l'intervalle de confiance s'établit à :

$2 \times 1{,}96\sqrt{\frac{p_0(1-p_0)}{n}}$ qu'on arrondit ici à $4\sqrt{\frac{p_0(1-p_0)}{n}}$

L'effectif de l'échantillon devra donc être au moins $10000\, p_0(1 - p_0)$.
Toutefois, cet effectif dépend de p_0, inconnu avant l'expérience. L'usage de ces calculs supposera donc que l'on ait une idée du résultat attendu, grâce à un sondage exploratoire par exemple ou grâce à une connaissance préalable du phénomène étudié.

De façon générale, si l'on souhaite obtenir un intervalle de confiance d'une proportion de longueur $2i$, il est nécessaire d'inclure un nombre d'individus au moins égal à :

$4\frac{p_0(1-p_0)}{i^2}$ au risque 0,05 (ou $u_\alpha^2\frac{p_0(1-p_0)}{i^2}$ au risque α)

REMARQUE
Lorsque le sondage est réalisé, un intervalle de confiance lui est associé. Dans le langage courant, les instituts de sondage nomment ces intervalles de confiance des FOURCHETTES.

14.3.4.2 Précision d'une moyenne

Dans le cas où l'on s'intéresse à la moyenne vraie d'une variable quantitative, on peut effectuer le même type de calcul. Pour obtenir un intervalle de confiance de longueur $2i$, il faut inclure un nom-

bre d'individus au moins égal à :

$$n = u_{\alpha}^{2} \frac{s^2}{i^2}$$

L'exploitation de ce calcul nécessite ici une connaissance, même approximative, de la variance de la variable étudiée pour se donner a priori s^2- ou mieux σ^2.

Exemple très important : les problèmes de dosage.

Soit à doser la glycémie ; on a devant soi un échantillon de sang. Quelle est la concentration en glucose ? Si on fait plusieurs dosages, on va obtenir plusieurs résultats. Cela est dû, non à la variabilité de la glycémie, mais aux erreurs analytiques. On assimile la glycémie vraie à la moyenne vraie de la variable aléatoire « résultat du dosage ». Supposons que l'on connaisse la variance des résultats, car on connaît bien la technique analytique. Par exemple, $\sigma = 10$ mg.l^{-1}. Supposons en outre que les résultats expérimentaux soient distribués normalement.
Si on effectue un dosage donnant 90 mg.l^{-1}, on a pour intervalle de confiance approché (σ étant connu) :
$IC_{0,95} = [90 - 2\sigma ; 90 + 2\sigma] = [70 ; 110]$ soit un intervalle de longueur 40.

Si on effectue deux dosages donnant 90 et 96 mg.l^{-1}, on a

$$IC_{0,95} = \left[93 - 2\frac{\sigma}{\sqrt{2}} ; 93 + 2\frac{\sigma}{\sqrt{2}}\right] = [78,9 ; 107,1]$$

soit un intervalle d'amplitude 28,2.

Si l'on effectue trois dosages donnant 90, 96 et 93 mg.l^{-1} on a

$$IC_{0,95} = \left[93 - 2\frac{\sigma}{\sqrt{3}} ; 93 + 2\frac{\sigma}{\sqrt{3}}\right] = [81,5 ; 104,5]$$

soit un intervalle d'amplitude 23,0.

Ces calculs objectivent le fait bien connu selon lequel la répétition des dosages permet d'atténuer les conséquences des erreurs expérimentales. Certains dosages - certaines mesures (tension artérielle) - sont répétés avant qu'une valeur soit indiquée.