



**Université PARIS-VI Pierre et Marie Curie  
Facultés de Médecine Pitié-Salpêtrière et Saint Antoine**

# **Biostatistique**

**PCEM1**

**2004 - 2005**

**Responsables : A. Mallet et A.J. Valleron  
Auteurs : J.L. Golmard, A. Mallet, V. Morice**

**Mise à jour : 18 janvier 2005  
Relecture : V. Morice et S. Tezenas**



# Sommaire

3		<b>Sommaire</b>
9		<b>Avant-propos</b>
11		<b>Introduction</b>
11	1	La variabilité et l'incertain
12	2	La décision dans l'incertain
13		<b>Chapitre 1 : Statistique(s) et Probabilité(s)</b>
13	1.1	Statistique
13	1.2	Population et échantillon
14	1.3	Statistique et probabilité
17		<b>Chapitre 2 : Rappels mathématiques</b>
17	2.1	Ensembles, éléments
17	2.2	Opérations sur les ensembles
19	2.3	Ensembles finis, dénombrables, non dénombrables
19	2.4	Ensembles produits
20	2.5	Familles d'ensembles
20	2.6	Autres rappels mathématiques
20	2.6.1	Rappel sur les sommes
21	2.6.2	Rappel sur les intégrales
23		<b>Chapitre 3 : Eléments de calcul des Probabilités</b>
23	3.1	Introduction
24	3.2	Expérience aléatoire, ensemble fondamental et événements
25	3.3	Opérations sur les événements
25	3.4	Règles du calcul des probabilités
27	3.5	Remarque
27	3.6	Illustration de quelques ensembles probabilisés
27	3.6.1	Ensemble probabilisé fini
28	3.6.2	Ensemble fini équiprobable
28	3.6.3	Ensembles probabilisés infinis
28	3.6.3.1	Cas dénombrable
29	3.6.3.2	Cas d'un ensemble probabilisé infini non dénombrable

31      **Chapitre 4 :    Probabilité Conditionnelle ; Indépendance et Théorème de Bayes**

- 31      4.1      Probabilité conditionnelle
- 32      4.2      Théorème de la multiplication
- 33      4.3      Diagramme en arbre
- 34      4.4      Théorème de Bayes
- 36      4.5      Indépendance entre événements
- 36      4.6      Indépendance, inclusion et exclusion de deux événements

39      **Chapitre 5 :    Evaluation de l'intérêt diagnostique des informations médicales**

- 39      5.1      Introduction
- 39      5.1.1      Le diagnostic
- 40      5.1.2      Les informations médicales
- 40      5.1.3      Situation expérimentale et estimation
- 41      5.2      Les paramètres de l'évaluation
- 41      5.2.1      Sensibilité et spécificité
- 42      5.2.2      Valeurs prédictives
- 42      5.2.3      Comparaison des deux couples de paramètres
- 43      5.2.4      Choix d'un seuil : courbes ROC
- 45      5.3      Estimation des paramètres de l'évaluation
- 45      5.3.1      Un échantillon représentatif
- 45      5.3.1.1      Les données
- 45      5.3.1.2      Estimation de la sensibilité et de la spécificité
- 46      5.3.1.3      Estimation des valeurs prédictives
- 47      5.3.2      Deux échantillons représentatifs

49      **Chapitre 6 :    Variables aléatoires**

- 49      6.1      Définition d'une variable aléatoire
- 50      6.2      Variables aléatoires finies
- 50      6.2.1      Représentation d'une loi de probabilité finie
- 50      6.2.2      Espérance mathématique d'une loi finie
- 53      6.2.3      Variance et écart-type
- 53      6.2.4      Loi de probabilité produit
- 55      6.2.5      Variables aléatoires indépendantes
- 55      6.2.6      Fonction de répartition
- 55      6.3      Variables infinies dénombrables
- 56      6.4      Variables aléatoires continues

## 59      **Chapitre 7 : Exemples de distributions**

59	7.1	Lois discrètes
59	7.1.1	Loi de Bernoulli
59	7.1.2	Loi binomiale
61	7.1.3	Loi de Poisson
63	7.2	Lois continues
63	7.2.1	Loi normale
63	7.2.1.1	Définition
63	7.2.1.2	Propriétés
66	7.2.2	Loi du $\chi^2$ (chi-2)
66	7.2.2.1	Définition
67	7.2.2.2	Propriétés
67	7.2.3	Loi de Student (hors programme)
68	7.2.4	Loi exponentielle (hors programme)

## 69      **Chapitre 8 : Statistiques descriptives**

69	8.1	Rappels et compléments
70	8.2	Représentation complète d'une série d'expériences
70	8.2.1	Cas d'une variable qualitative
71	8.2.2	Cas d'une variable quantitative discrète
72	8.2.3	Cas d'une variable quantitative continue. Notion d'HISTOGRAMME
73	8.3	Représentation simplifiée d'une série d'expériences
73	8.3.1	Indicateurs de localisation des valeurs
73	8.3.2	Indicateurs de dispersion des valeurs
74	8.4	Reformulation de la moyenne et de la variance observées
74	8.4.1	Reformulation de la moyenne observée
75	8.4.2	Reformulation de la variance observée
76	8.5	Cas particulier d'une variable à deux modalités - Proportion
76	8.5.1	Expression de l'espérance mathématique de X
76	8.5.2	Expression de la variance de X
77	8.5.3	Interprétation de la moyenne observée
77	8.6	Conclusion : la variable aléatoire moyenne arithmétique
78		Résumé du chapitre

## 81      **Chapitre 9 : Fluctuations de la moyenne observée : la variable aléatoire moyenne arithmétique**

81	9.1	Première propriété de la variable aléatoire moyenne arithmétique
81	9.1.1	Un exemple
82	9.1.2	Généralisation
83	9.2	Seconde propriété de la variable aléatoire moyenne arithmétique : le théorème central limite
84	9.3	Etude de la distribution normale (rappel)

86	9.4	Application du théorème central limite. Intervalle de Pari (I. P.)
86	9.4.1	Définition de l'intervalle de pari (I. P.) d'une moyenne observée
88	9.4.2	Les facteurs de dépendance de la longueur de l'intervalle de pari (IP)
89	9.4.3	L'intervalle de pari d'une variable aléatoire
90		Résumé du chapitre

## 91 **Chapitre 10 : Estimation - Intervalle de confiance**

91	10.1	Introduction
92	10.2	Estimation ponctuelle
92	10.2.1	Définition
92	10.2.2	Propriétés
92	10.2.2.1	Biais
93	10.2.2.2	Variance
93	10.2.2.3	Erreur quadratique moyenne
93	10.2.3	Exemple
94	10.3	Estimation par intervalle - Intervalle de confiance
94	10.3.1	Exemple d'une proportion
96	10.3.2	Intervalle de confiance approché d'une proportion « vraie »
97	10.3.3	Intervalle de confiance approché d'une moyenne « vraie » (variable continue)
97	10.3.4	Applications
98	10.3.4.1	Précision d'un sondage
98	10.3.4.2	Précision d'une moyenne

## 101 **Chapitre 11 : Les tests d'hypothèses. Principes**

101	11.1	Un exemple concret (emprunté à Schwartz)
104	11.2	Principe général des tests d'hypothèses
104	11.2.1	Les étapes de mises en œuvre
106	11.2.2	Justification de la règle de décision. Choix de $\alpha$
106	11.2.2.1	Interprétation de $\alpha$
106	11.2.2.2	Effet d'un changement de valeur de $\alpha$
107	11.2.3	Justification des conclusions du test. Puissance d'un test
110	11.2.4	Amélioration de l'interprétation du rejet de $H_0$
110	11.2.4.1	Notion de degré de signification
111	11.2.4.2	Orientation du rejet
113		Résumé du chapitre

## 115 **Chapitre 12 : Quelques tests usuels**

115	12.1	Tests concernant des variables de Bernoulli
115	12.1.1	Test d'égalité d'une proportion « vraie » à une valeur donnée (ou test de comparaison d'une proportion observée à une valeur donnée)
115	12.1.1.1	Mise en place du test

116	12.1.1.2	Autre interprétation du paramètre $z$
117	12.1.2	Test d'égalité de deux proportions « vraies » (ou test de comparaison de deux proportions observées)
119	12.2	Tests concernant des variables quantitatives
119	12.2.1	Tests de comparaison à une valeur donnée
119	12.2.1.1	Test d'égalité d'une moyenne « vraie » à une valeur donnée (ou test de comparaison d'une moyenne observée à une valeur donnée)
120	12.2.1.2	Test de symétrie d'une variable ( $X$ ) par rapport à une valeur donnée ( $\mu_0$ ) : test de Wilcoxon
121	12.2.2	Tests de comparaison de variables quantitatives
121	12.2.2.1	Test d'égalité de deux moyennes « vraies » (ou test de comparaison de deux moyennes observées)
122	12.2.2.2	Test d'égalité de deux distributions (ou test de comparaison de deux distributions) : test de Mann-Whitney-Wilcoxon
124	12.2.3	Cas des séries appariées
124	12.2.3.1	Test de comparaison de deux moyennes sur séries appariées
125	12.2.3.2	Test de symétrie de la distribution des différences
126		Résumé du chapitre

## 127    **Chapitre 13 : Tests concernant des variables qualitatives**

127	13.1	Comparaison d'une répartition observée à une répartition donnée ou test du $\chi^2$ d'ajustement
128	13.1.1	Les étapes de mise en œuvre
132	13.1.2	Cas particulier : variable à deux modalités
133	13.2	Comparaison de deux répartitions observées ou test du $\chi^2$ d'homogénéité
136	13.3	Test d'indépendance entre deux variables qualitatives
140		Résumé du chapitre

## 141    **Chapitre 14 : Liaison entre deux variables continues : notion de corrélation**

141	14.1	Introduction
142	14.2	Abord du problème
144	14.3	Un indicateur de covariation : le coefficient de corrélation
148	14.4	Le coefficient de corrélation « vrai »
149	14.5	Test de comparaison du coefficient de corrélation « vrai » $\rho$ à 0
151		Résumé du chapitre

## 153    **Chapitre 15 : A propos des tests d'hypothèses**

153	15.1	Rappels et précisions
155	15.2	Jugement d'interprétation - La causalité

**157 Chapitre 16 : Analyse des durées de survie**

157	16.1	Contexte
158	16.2	Estimation : courbe de survie
158	16.2.1	Définitions
158	16.2.2	Estimation par la méthode de Kaplan-Meier
161	16.2.3	Estimation par la méthode actuarielle
161	16.3	Test de comparaison de deux courbes de survie : test du log-rank approché

**165 Annexe A : Tables statistiques**

166	A.1	TABLE DE LA VARIABLE NORMALE REDUITE Z
167	A.2	TABLE DU TEST DE WILCOXON
168	A.3	TABLE DU TEST DE MANN-WHITNEY-WILCOXON
169	A.4	TABLE DE $\chi^2$
170	A.5	TABLE DU COEFFICIENT DE CORRELATION
171	A.6	TABLE DU t DE STUDENT



# Avant-propos

Ce polycopié contient le cours de PCEM1 commun aux deux CHU Pitié-Salpêtrière et Saint-Antoine.

Il est très utile de le compléter par le livre de A. J. Valleron :

**A.J. Valleron.** *Probabilités et statistiques*. Masson (collection Abrégés, Cours+exos)

Ce livre reprend le cours sous la forme de 24 fiches complétées de 140 exercices et 100 QCM corrigés.



# Introduction

La statistique constitue, 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 :

$$\text{variabilité totale} = \text{variabilité biologique} + \text{variabilité métrologique}$$

- 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.

$$\text{variabilité biologique} = \text{variabilité intra-individuelle} + \text{variabilité inter-individuelle}$$

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.

$$\text{variabilité métrologique} = \text{variabilité expérimentale} + \text{variabilité appareil de mesure}$$

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 fournie 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'aider à 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 *statistiques* (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*. Pour qu'un échantillon permette l'étude de la variabilité des caractéristiques d'intérêt de la popu-

lation, il faut qu'il soit convenablement sélectionné. On parlera d'*échantillon représentatif* si les individus le constituant ont été tirés au sort<sup>1</sup> dans la population. Si par exemple on souhaite déterminer les caractéristiques « moyennes » du poids et de la taille des prématurés masculins on tirera au hasard 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 expérimentaux

---

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, éléments

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  $x \notin A$
- $\emptyset$  est l'ensemble vide
- $E$  est l'ensemble universel.

### 2.2 Opérations sur les ensembles

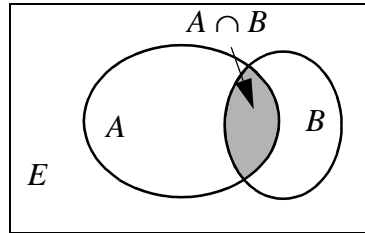
Soient  $A$  et  $B$  deux ensembles quelconques.

#### 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  $B$



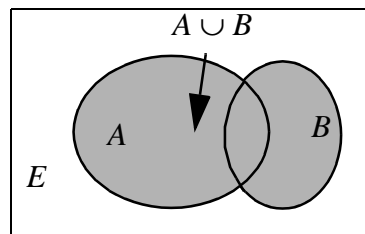
Cas particulier : si  $A \cap B = \emptyset$ , on dit que  $A$  et  $B$  sont **disjoints**.

### 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 :

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

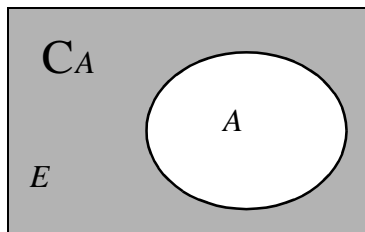
Le terme « ou » est employé au sens  $x \in A$  ou  $B$  si  $x$  appartient à  $A$ , ou à  $B$ , ou à  $A$  et  $B$  (car  $x \in A$  et  $B$  signifie  $x \in A$  et  $x \in B$ ).



### Complémentaire

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

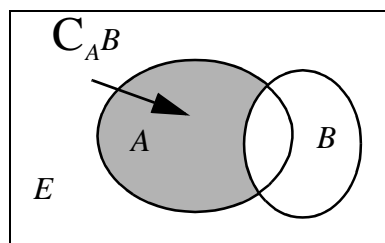
$$\bar{C}_A = \bar{A} = \{x : x \notin A\}$$



### 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  $B$ .

$$A - B = \bar{C}_A B = \{x : x \notin B \text{ et } x \in A\}$$



## Algèbre des ensembles

$$A \cup A = A$$

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$A \cup B = B \cup A$$

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

$$A \cup \emptyset = A$$

$$A \cup E = E$$

$$A \cup C A = E$$

$$C C A = A$$

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

$$A \cap A = A$$

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

$$A \cap B = B \cap A$$

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

$$A \cap E = A$$

$$A \cap \emptyset = \emptyset$$

$$A \cap C A = \emptyset$$

$$C E = \emptyset, C \emptyset = E$$

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

## 2.3 Ensembles finis, dénombrables, non dénombrables

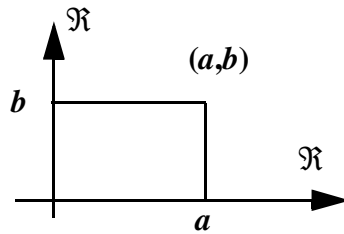
- Un ensemble est **fini** s'il est vide ( $\emptyset$ ) ou s'il contient un nombre fini d'éléments ; sinon, il est infini :  
 $A = \{a_1, a_2, a_3\}$  est fini ;  
 $I = \{x \in [0,1]\}$  est infini.
- Un ensemble infini est dit **dénombrable** si on peut faire correspondre de façon unique chaque élément de l'ensemble à un entier naturel et un seul :  
 $A = \{n : n \text{ est un entier pair}\}$  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  $\mathbb{R}$  :  $\{x \in [a,b]\}$  ou des intervalles de  $\mathbb{R}^2$  :  $\{(x,y) : x \in [a,b], y \in [c,d]\}$ .

## 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 :  $A = \{1, 2\}$

$$\mathbf{P}(A) = \{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$$

### 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  $\{A_i\} = \{A_1, A_2, \dots, A_n, \dots\}$  qui peut être finie ou non. On note :

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

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

## 2.6 Autres rappels mathématiques

### 2.6.1 Rappel sur les sommes

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

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

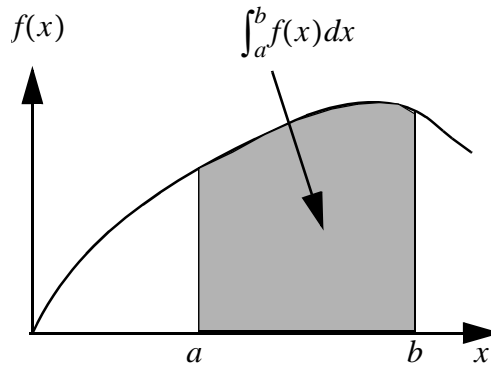
Si  $k$  est une constante (indépendante de  $i$ ), elle peut être sortie de la somme.

## 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.  $\int_a^b kf(x) dx = k \int_a^b f(x) dx$
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  $\tau$ . 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.  $\int_a^b f(x) dx = F(b) - F(a)$



# Chapitre 3

## Eléments de calcul des Probabilités

### 3.1 Introduction

Le calcul des probabilités est la théorie mathématique, 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 Expérience aléatoire, ensemble fondamental et événements

### Expérience aléatoire

On s'intéresse ici aux seules expériences dont le résultat n'est pas prévisible, les expériences aléatoires. Une expérience aléatoire est aussi appelée une *épreuve*.

### Ensemble fondamental

Pour une expérience aléatoire donnée, l'ensemble des résultats possibles est appelé l'ensemble fondamental, que nous noterons  $E$  dans la suite du cours. Chaque **résultat** d'expérience est un point de  $E$  ou un élément de  $E$ .

### Événement

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

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

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

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

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

On note parfois  $\Omega$  l'ensemble de tous les événements.

### Exemples

- On jette un dé et on observe le résultat obtenu. L'ensemble fondamental est formé par les 6 résultats possibles :  
 $E = \{1, 2, 3, 4, 5, 6\}$   
 L'événement correspondant à l'apparition d'un nombre pair est  $A = \{2, 4, 6\}$ , qui est bien un sous ensemble de  $E$ .  
 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  $C = \{3\}$ .
- Dans l'exemple précédent  $E$  était fini et donc dénombrable ;  $E$  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  $E = \{1, 2, 3, \dots, n, \dots\}$  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.
- 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.



### 3.3 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  $A + B$  ou  $A$  ou  $B$ .
2.  $A \cap B$  est l'événement qui se produit si  $A$  et  $B$  sont réalisés tous les deux.  
Il est parfois noté  $A \cdot B$  ou  $A$  et  $B$ .
3.  $\bar{A}$  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  $\bar{A}$ .

#### Événements incompatibles

Quand deux événements  $A$  et  $B$  sont tels que  $A \cap B = \emptyset$ , 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  $A_1, A_2, \dots, A_n$  forment une famille complète si les  $A_i$  constituent une partition de  $E$ , c'est-à-dire si :

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

#### Exemple

Reprenons l'exemple précédent du jeu de dés :

$E = \{1, 2, 3, 4, 5, 6\}$ ,  $A = \{2, 4, 6\}$ ,  $B = \{1, 2, 3, 5\}$ ,  $C = \{3\}$ .

$A \cup B = \{1, 2, 3, 4, 5, 6\}$  = apparition d'un nombre pair ou premier

$A \cap B = \{2\}$  = apparition d'un nombre pair et premier

$\bar{C} = \{1, 2, 4, 5, 6\}$  = apparition d'un nombre autre que 3

$A \cap C = \emptyset$  :  $A$  et  $C$  s'excluent mutuellement.

### 3.4 Règles du calcul des probabilités

Soit un ensemble fondamental  $E$ . Nous introduisons une fonction  $Pr$  qui, à tout événement  $A$ , associe un nombre réel positif ou nul.

$Pr$  est dite fonction de probabilité, et  $Pr(A)$  est appelée probabilité de l'événement  $A$ , si les conditions ou règles suivantes sont satisfaites :

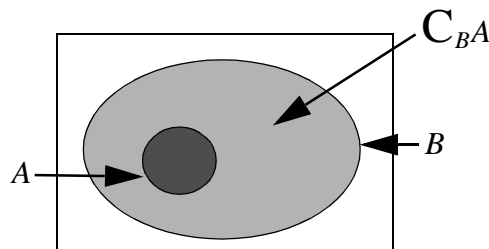
1.  $Pr(A) \geq 0$  pour tout événement  $A$  : une probabilité est positive ou nulle
2.  $Pr(S) = 1$  : la probabilité de l'événement certain est 1
3.  $(A \cap B = \emptyset) \Rightarrow (Pr(A \cup B) = Pr(A) + Pr(B))$  : permet le calcul de la probabilité de la réunion de deux événements **disjoints**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

$(A_i \cap A_j = \emptyset)$ , alors  $Pr(A_1 \cup A_2 \cup \dots) = Pr(A_1) + Pr(A_2) + \dots$ .

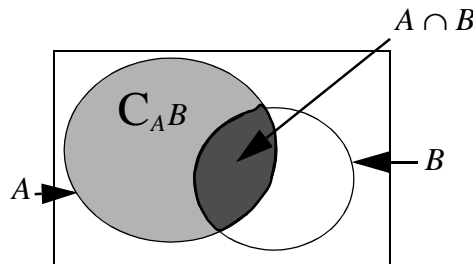
Cette quatrième condition est proche de la troisième. Elle ne peut cependant pas s'en déduire dans le cas d'un ensemble d'événements infini dénombrable.

Propriétés déduites des quatre conditions précédentes :

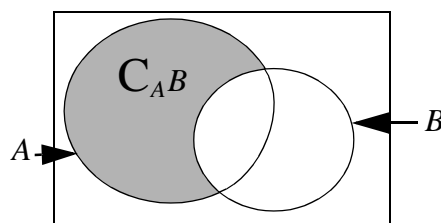
1.  $Pr(\emptyset) = 0$   
Soit  $A$  un événement quelconque.  $A$  et  $\emptyset$  sont évidemment disjoints puisque  $A \cap \emptyset = \emptyset$  ; donc  $Pr(A \cup \emptyset) = Pr(A) + Pr(\emptyset)$ . Or  $A \cup \emptyset = A$  ; donc  $Pr(A \cup \emptyset) = Pr(A)$ . D'où  $Pr(\emptyset) = 0$ .
2.  $Pr(A) \leq 1$   
 $A$  et son complémentaire  $\bar{A}$  sont disjoints, et leur réunion forme  $E$ , de probabilité 1. Donc  $Pr(S) = 1 = Pr(A \cup \bar{A}) = Pr(A) + Pr(\bar{A})$ . Toute probabilité étant positive ou nulle, on obtient bien  $Pr(A) \leq 1$ .
3.  $Pr(\bar{A}) = 1 - Pr(A)$   
A démontrer en exercice, en notant que  $S = A \cup \bar{A}$ .
4. Si  $A \subset B$ , alors  $Pr(A) \leq Pr(B)$ .  
A démontrer en exercice, en notant que  $B = A \cup \bar{A}B$ .



5.  $Pr(\bar{A}B) = Pr(A) - Pr(A \cap B)$   
A démontrer en exercice, en remarquant que  $A = \bar{A}B \cup (A \cap B)$ .



6.  $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$   
A démontrer en exercice, en remarquant que  $(A \cup B) = \bar{A}B \cup B$ .



## 3.5 Remarque

Alors que  $Pr(\emptyset) = 0$ , 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.

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  $E = \{a_1, a_2, \dots, a_n\}$  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.  $p_1 + p_2 + \dots + p_n = 1$

La probabilité d'un événement quelconque  $A$  est la somme des probabilités des  $a_i$  qu'il contient :

$$Pr(A) = \sum_{a_i \in A} p_i$$

### Exemple

On jette 3 pièces de monnaie et on compte le nombre de « face » obtenu. L'ensemble fondamental correspondant à cette expérience est  $E = \{0, 1, 2, 3\}$  puisqu'on peut obtenir comme résultat de l'expérience : 0 fois « face » (3 fois « pile »), 1 fois « face » (2 fois « pile »), 2 fois « face », ou 3 fois « face ».

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 fois « face »,  $A = \{a_2, a_3\}$  :

$$Pr(A) = p_2 + p_3 = 3/8 + 1/8 = 4/8 = 1/2$$

### 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.

$E = \{a_1, a_2, \dots, a_n\}$  et  $Pr(\{a_1\}) = p_1, Pr(\{a_2\}) = p_2, \dots, Pr(\{a_n\}) = p_n$

avec  $p_1 = p_2 = \dots = p_n = 1/n$

Les jeux de hasard - dés, cartes, loto, etc. - entrent précisément dans cette catégorie :

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

#### 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  $E$ .

#### 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 de tirer un trèfle ?

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

Quelle est la probabilité de tirer un roi ?

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

Quelle est la probabilité de tirer un roi de trèfle ?

$$Pr(\text{tirer un roi de trèfle}) = \frac{1}{52}$$

#### 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  $E = \{a_1, a_2, \dots, a_n, \dots\}$  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

$$A = \{a_{25}, a_{31}, a_{43}\}$$

$$Pr(A) = p_{25} + p_{31} + p_{43}$$

### 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 :

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

### 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  $Pr(\text{pile}) = Pr(\text{face}) = 1/2$  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

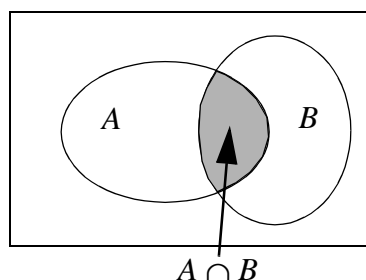
Soient  $A$  et  $B$  deux événements quelconques d'un ensemble fondamental  $E$  muni d'une loi de probabilité  $Pr$ . On s'intéresse à ce que devient la probabilité de  $A$  lorsqu'on apprend que  $B$  est déjà réalisé, c'est-à-dire lorsqu'on restreint l'ensemble des résultats possibles  $E$  à  $B$ .

La probabilité conditionnelle de  $A$ , sachant que l'événement  $B$  est réalisé, est notée  $Pr(A/B)$  et est définie par la relation suivante :

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

**Equation 1 : probabilité conditionnelle**

Dans cette équation, les probabilités des événements  $A \cap B$  et  $B$  doivent être calculées sur tout l'ensemble fondamental  $E$ , comme si on ne savait pas que  $B$  s'est déjà réalisé. Sinon, on obtient évidemment  $Pr(B) = 1$ .



**Figure 1 : probabilité conditionnelle**

Cette relation générale pour tout espace probabilisé s'interprète facilement dans le cas où  $E$  est un

espace équiprobable (mais cette relation est vraie pour un espace non-équiprobable !). En notant  $|A|$  le nombre d'éléments de  $A$  :

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

$Pr(A/B)$  traduit le rapport de la surface de  $A \cap B$  sur la surface de  $B$  dans la figure 1.

Toujours dans le cas où  $E$  est équiprobable, on a

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

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 ?

$B = \{\text{somme des deux dés} = 6\}$

$A = \{\text{au moins un des deux dés donne 2}\}$

$B = \{(2, 4), (4, 2), (1, 5), (5, 1), (3, 3)\}$

Nombre de réalisations de  $A \cap B = \{(2, 4), (4, 2)\} = 2$

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

## 4.2 Théorème de la multiplication

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

On en tire immédiatement

$$Pr(A \cap B) = Pr(A/B)Pr(B) = Pr(B/A)Pr(A)$$

### Equation 2 : théorème de la multiplication

L'équation 2 peut se généraliser facilement. Soient  $A_1, \dots, A_n$  des événements quelconques d'un espace probabilisé ; à partir de l'équation 2, on montre :

$$Pr(A_1 \cap A_2 \cap \dots \cap A_n) = Pr(A_1)Pr(A_2/A_1)Pr(A_3/(A_1 \cap A_2))\dots Pr(A_n/(A_1 \cap A_2 \dots \cap A_{n-1}))$$

### 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.

$Pr(1^{\text{er}} \text{ défectueux}) = 4/10$

$Pr(2^{\text{ème}} \text{ défectueux} / 1^{\text{er}} \text{ défectueux}) = 3/9$



$$Pr(3^{\text{ème}} \text{ défectueux} / 1^{\text{er}} \text{ et } 2^{\text{ème}} \text{ défectueux}) = 2/8$$

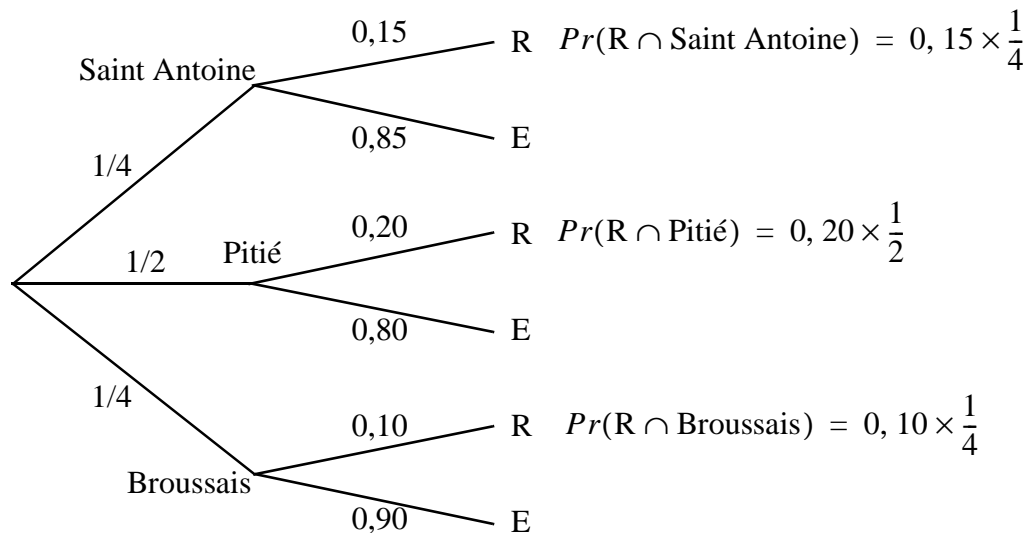
$$Pr(1^{\text{er}} \text{ et } 2^{\text{ème}} \text{ et } 3^{\text{ème}} \text{ défectueux}) = 4/10 \times 3/9 \times 2/8 = 1/30.$$

## 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 (l'université Pierre et Marie Curie a longtemps comporté ces 3 CHU) étaient respectivement (données arbitraires) de 0,20 ; 0,15 ; et 0,10 ( $0,20 = Pr(\text{Réussite}/\text{Pitié})$ ) ; on sait que  $1/4$  des étudiants de Paris VI étaient à Saint Antoine,  $1/4$  à Broussais et  $1/2$  à la Pitié. Quelle était la probabilité qu'un étudiant de Paris VI soit reçu au concours ?



R signifie réussite et E échec.

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

$$Pr(R) = 0,15 \times 1/4 + 0,20 \times 1/2 + 0,10 \times 1/4 = 0,1625$$

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 32 (section 4.2), on obtient la formule de Bayes :

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

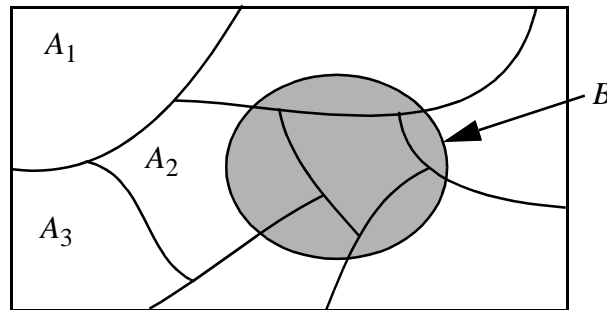
**Equation 3 : formule de Bayes**

Le théorème est une forme développée de cette formule que nous introduisons maintenant. Considérons des événements  $A_1, \dots, A_n$  tels qu'ils forment une **partition** de l'ensemble fondamental  $E$ .

Par définition, les  $A_i$  s'excluent mutuellement et leur union est  $E$  :

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

Soit  $B$  un événement quelconque



De  $E = A_1 \cup A_2 \cup \dots \cup A_n$  et de  $B \cap E = B$ , on tire  $B = B \cap (A_1 \cup A_2 \cup \dots \cup A_n)$ .

Soit, par distributivité,  $B = (B \cap A_1) \cup (B \cap A_2) \cup \dots \cup (B \cap A_n)$ .

En remarquant que les  $B \cap A_i$  sont exclusifs, puisque les  $A_i$  le sont, et en appliquant la 3<sup>ème</sup> règle du calcul des probabilités on obtient la formule dite des « probabilités totales » :

$$Pr(B) = Pr(B \cap A_1) + Pr(B \cap A_2) + \dots + Pr(B \cap A_n)$$

**Equation 4 : probabilités totales**

En appliquant le théorème de la multiplication :

$$Pr(B) = Pr(B/A_1)Pr(A_1) + Pr(B/A_2)Pr(A_2) + \dots + Pr(B/A_n)Pr(A_n)$$

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

$$Pr(A_i/B) = \frac{Pr(B/A_i)Pr(A_i)}{Pr(B/A_1)Pr(A_1) + Pr(B/A_2)Pr(A_2) + \dots + Pr(B/A_n)Pr(A_n)}$$

**Equation 5 : théorème 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 anciens CHU Saint Antoine, Pitié et Broussais respectivement :

$$Pr(R) = Pr(R/C_1)Pr(C_1) + Pr(R/C_2)Pr(C_2) + Pr(R/C_3)Pr(C_3)$$

[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 venait d'un CHU donné, on cherche la probabilité qu'un étudiant ait été inscrit à un CHU donné sachant qu'il a été reçu (probabilité des causes).

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

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

Avec  $Pr(C_1) = 0,25$  ;  $Pr(C_2) = 0,50$  ;  $Pr(C_3) = 0,25$  ;

et  $Pr(R/C_1) = 0,15$  ;  $Pr(R/C_2) = 0,20$  ;  $Pr(R/C_3) = 0,10$ .

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

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  $Pr(C_2) = 0,50$ ).

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  $Pr(s_1/D_1)$ ,  $Pr(s_1/D_2)$ , et  $Pr(s_1/D_3)$ .

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

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  $Pr(D_1)$ ,  $Pr(D_2)$  et  $Pr(D_3)$ .

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  $Pr(D_1/s_1)$ ,  $Pr(D_2/s_1)$ ,  $Pr(D_3/s_1)$  et on retient le diagnostic qui a la plus grande probabilité : c'est l'application de l'approche bayésienne 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  $Pr(A / B) = Pr(A)$ .

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

$A$  et  $B$  sont indépendants si et seulement si  $Pr(A \cap B) = Pr(A)Pr(B)$ .

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

### Note

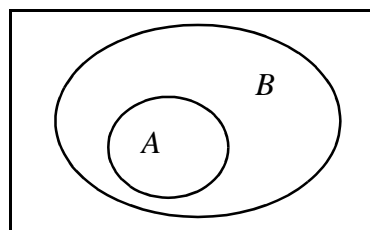
Ce qui est défini précédemment est l'indépendance de deux événements. Si on considère maintenant 3 événements  $A, B, C$ , on dira que ces 3 événements sont indépendants :

1. s'ils sont indépendants 2 à 2 :  $A$  indépendant de  $B$  ;  $A$  indépendant de  $C$  ; et  $B$  indépendant de  $C$
2. et si  $Pr(A \cap B \cap C) = Pr(A)Pr(B)Pr(C)$ . 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.

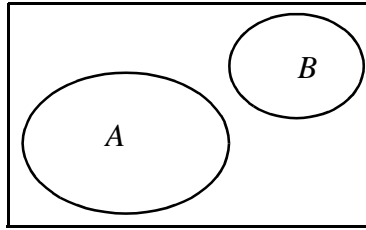


Alors  $Pr(A \cap B) = Pr(A)$ .

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

$A$  et  $B$  ne sont **pas indépendants**.

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



Alors  $Pr(A \cap B) = Pr(\emptyset) = 0$ .

$$\text{D'où } Pr(A/B) = \frac{Pr(A \cap B)}{Pr(B)} = \frac{0}{Pr(B)} = 0.$$

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



# Chapitre 5

## Evaluation de l'intérêt diagnostique des informations médicales

### 5.1 Introduction

La tâche essentielle des médecins est de traiter les patients. Pour prescrire un traitement, il faut savoir, plus ou moins précisément selon les cas, ce dont souffre le malade. Pour résumer en un seul terme un processus physiopathologique complexe, les médecins ont créé des concepts : les diagnostics.

La recherche « du » diagnostic est donc la première étape de la consultation clinique. Pour parvenir au diagnostic, le médecin accumule des informations, dont certaines lui sont spontanément livrées par le patient (le motif de la consultation, les symptômes), d'autres doivent être recherchées mais sont disponibles immédiatement (les signes physiques), d'autres enfin sont d'obtention plus ou moins difficile et coûteuse (les résultats d'examens complémentaires). De nouvelles procédures diagnostiques apparaissent fréquemment : on a vu, par exemple, l'apparition des échographies, de la tomodensitométrie (scanner), de l'IRM, pour ne citer que le domaine de l'imagerie. Il n'est bien sûr pas question d'effectuer tous les examens complémentaires sur tous les malades : il faut donc préciser les indications de ces examens, ce qui repose sur l'évaluation de leur intérêt diagnostique. Avant d'aborder la méthodologie de l'évaluation, nous reviendrons sur certains concepts utilisés dans ce paragraphe.

#### 5.1.1 Le diagnostic

On peut définir un diagnostic comme un concept résumant l'état d'un individu. Le terme de « diagnostic » est donc beaucoup moins précis qu'on pourrait le penser à première vue : on peut en général fournir plusieurs diagnostics pour un même état physiopathologique, les termes diagnostiques utilisés dépendant de l'aspect privilégié. Parmi ces aspects, on peut citer :

— la symptomatologie

- la physiopathologie et l'étiologie
- la conduite thérapeutique

En pratique, la précision du diagnostic dépendra souvent des possibilités thérapeutiques : par exemple, on ne recherchera pas, en général, le virus responsable d'un syndrome grippal, surtout si on s'attend à ce que la maladie guérisse spontanément.

D'un point de vue statistique, le diagnostic sera souvent considéré comme une variable aléatoire binaire : le patient souffre ou ne souffre pas de l'affection considérée, ou, exprimé autrement, le diagnostic est vrai ou faux chez ce patient. Les valeurs possibles de la variable seront notées  $M$  et  $\bar{M}$  (maladie présente ou absente), ou  $D$  et  $\bar{D}$  (diagnostic vrai ou faux).

## 5.1.2 Les informations médicales

On divise l'ensemble des informations médicales en signes cliniques et signes complémentaires. Les signes cliniques sont divisés en signes fonctionnels ou symptômes, décrits par le malade (spontanément ou par l'interrogatoire) et signes physiques, recherchés par le médecin. Les signes complémentaires peuvent être biologiques ou radiologiques. Leur intérêt peut être :

- diagnostique (caractère malin ou bénin d'une tumeur)
- thérapeutique (localisation précise d'une tumeur)
- pronostique (extension ganglionnaire)

D'un point de vue statistique, ces signes peuvent être représentés par des variables binaires (présence ou absence d'un nodule sur une image) ou continues (cholestérolémie).

Nous considérons ici le seul cas d'un signe binaire, présent (noté  $S$ ) ou absent (noté  $\bar{S}$ ). Si l'information est de type continu, on se ramène au cas binaire par l'introduction d'un seuil : d'un côté du seuil, les valeurs sont dites normales, et le signe binaire est absent ; de l'autre côté du seuil, les valeurs sont dites pathologiques, et le signe binaire est présent.

## 5.1.3 Situation expérimentale et estimation

Quand on cherche à évaluer l'intérêt diagnostique d'un signe pour une affection, on recherche le signe chez des individus présentant ou non l'affection considérée. Deux situations expérimentales sont à envisager :

- **un échantillon** représentatif d'une population est constitué. On pourra estimer, à partir de cet échantillon, toutes les probabilités d'événements par les fréquences observées correspondantes ;
- **deux échantillons** sont constitués, l'un représentatif des individus pour lesquels le diagnostic est vrai, l'autre représentatif des individus pour lesquels il est faux. Cette manière de procéder est souvent la seule possible en pratique, surtout quand la maladie considérée est rare. Il faut remarquer, cependant, qu'on ne peut plus estimer n'importe quelle probabilité par la fréquence observée correspondante ; ce point sera développé plus loin dans ce chapitre.



**Remarque :** nous utilisons actuellement le mot *estimation* dans le sens d'*approximation* de la vraie valeur. Nous donnerons des définitions plus rigoureuses dans le chapitre 10 page 91.

## 5.2 Les paramètres de l'évaluation

### 5.2.1 Sensibilité et spécificité

La **sensibilité** d'un signe pour une maladie est la probabilité que le signe soit présent si le sujet est atteint de la maladie considérée.

Il s'agit donc de la probabilité conditionnelle qu'on peut noter :

$$\text{Sensibilité} = Se = Pr(S / M)$$

Un test diagnostic est donc d'autant plus sensible que les sujets atteints de la maladie présentent plus souvent le signe S.

La **spécificité** d'un signe pour une maladie est la probabilité que le signe soit absent si le sujet n'est pas atteint de la maladie.

De manière similaire, on a :

$$\text{Spécificité} = Sp = Pr(\bar{S} / \bar{M})$$

Un test diagnostic est donc d'autant plus spécifique que les sujets indemnes de la maladie présentent moins souvent le signe S.

Pour un examen « parfait », c'est-à-dire n'effectuant aucune erreur, les valeurs de la sensibilité et de la spécificité sont égales à 1.

Si la présence du signe est définie par un « seuil de positivité », on observe que ces deux paramètres varient en sens inverse lorsqu'on fait varier ce seuil. Ceci explique qu'un seul de ces deux paramètres ne suffise pas à évaluer un examen. Supposons par exemple qu'on s'intéresse au signe température vis à vis de la grippe. On considère que le signe est présent si la température dépasse un certain seuil, par exemple 39°C. Si on augmente le seuil pour le porter à 40°C, la probabilité de dépasser le seuil (chez les sujets grippés) va diminuer, donc la sensibilité diminue. En revanche, la probabilité d'être en dessous du seuil (chez les sujets non grippés) va augmenter, donc la spécificité augmente.

Un test diagnostic de bonne sensibilité conduit à un résultat positif chez presque tous les malades. Il est donc utilisable pour un dépistage. Si le test possède une bonne spécificité, il conduit à un résultat négatif chez presque tous les non-malades. Il pourrait donc être utilisé en tant qu'examen de confirmation du diagnostic.

Ces considérations sont bien sûr schématiques, d'autres éléments intervenant dans l'évaluation, comme la fréquence de la maladie (prévalence), les risques liés à la maladie, à l'examen, l'existen-

ce et les performances d'autres examens concurrents...

## 5.2.2 Valeurs prédictives

En pratique, quand un médecin reçoit le résultat d'un examen complémentaire, positif ou négatif, il ne sait pas si le patient souffre de l'affection qu'il cherche à diagnostiquer ou non, et les probabilités qui l'intéressent s'expriment de la manière suivante : quelle est la probabilité de présence de la maladie  $M$  chez ce patient, sachant que l'examen a donné un résultat positif (ou négatif) ? Ces probabilités sont appelées valeurs prédictives. Plus précisément, on a :

- la **valeur prédictive positive** d'un signe pour une maladie est la probabilité que le sujet soit atteint de la maladie si le signe est présent ;
- la **valeur prédictive négative** d'un signe pour une maladie est la probabilité que le sujet soit indemne de la maladie si le signe est absent.

On peut noter ces paramètres :

$$VPP = Pr(M/S)$$

$$VPN = Pr(\bar{M}/\bar{S})$$

Comme les sensibilité et spécificité, les valeurs prédictives positive et négative varient en sens inverse, et doivent donc être considérées simultanément.

Les valeurs prédictives peuvent s'exprimer en fonction du couple sensibilité - spécificité, et de la fréquence de la maladie dans la population (cette probabilité  $Pr(M)$  s'appelle la **prévalence** de la maladie). Il suffit d'utiliser le théorème de Bayes :

$$\begin{aligned} VPP = Pr(M/S) &= \frac{Pr(S/M)Pr(M)}{Pr(S/M)Pr(M) + Pr(S/\bar{M})Pr(\bar{M})} \\ &= \frac{Se \times Pr(M)}{Se \times Pr(M) + (1 - Sp) \times (1 - Pr(M))} \end{aligned}$$

$$\begin{aligned} VPN = Pr(\bar{M}/\bar{S}) &= \frac{Pr(\bar{S}/\bar{M})Pr(\bar{M})}{Pr(\bar{S}/M)Pr(M) + Pr(\bar{S}/\bar{M})Pr(\bar{M})} \\ &= \frac{Sp \times (1 - Pr(M))}{(1 - Se) \times Pr(M) + Sp \times (1 - Pr(M))} \end{aligned}$$

## 5.2.3 Comparaison des deux couples de paramètres

En situation clinique, on a vu que les valeurs prédictives correspondent aux préoccupations des médecins, et elles pourraient sembler les « meilleurs » paramètres d'évaluation. Pourtant, en réalité, c'est la sensibilité et la spécificité qui sont le plus souvent utilisées pour évaluer les examens complémentaires. La raison en est la suivante :

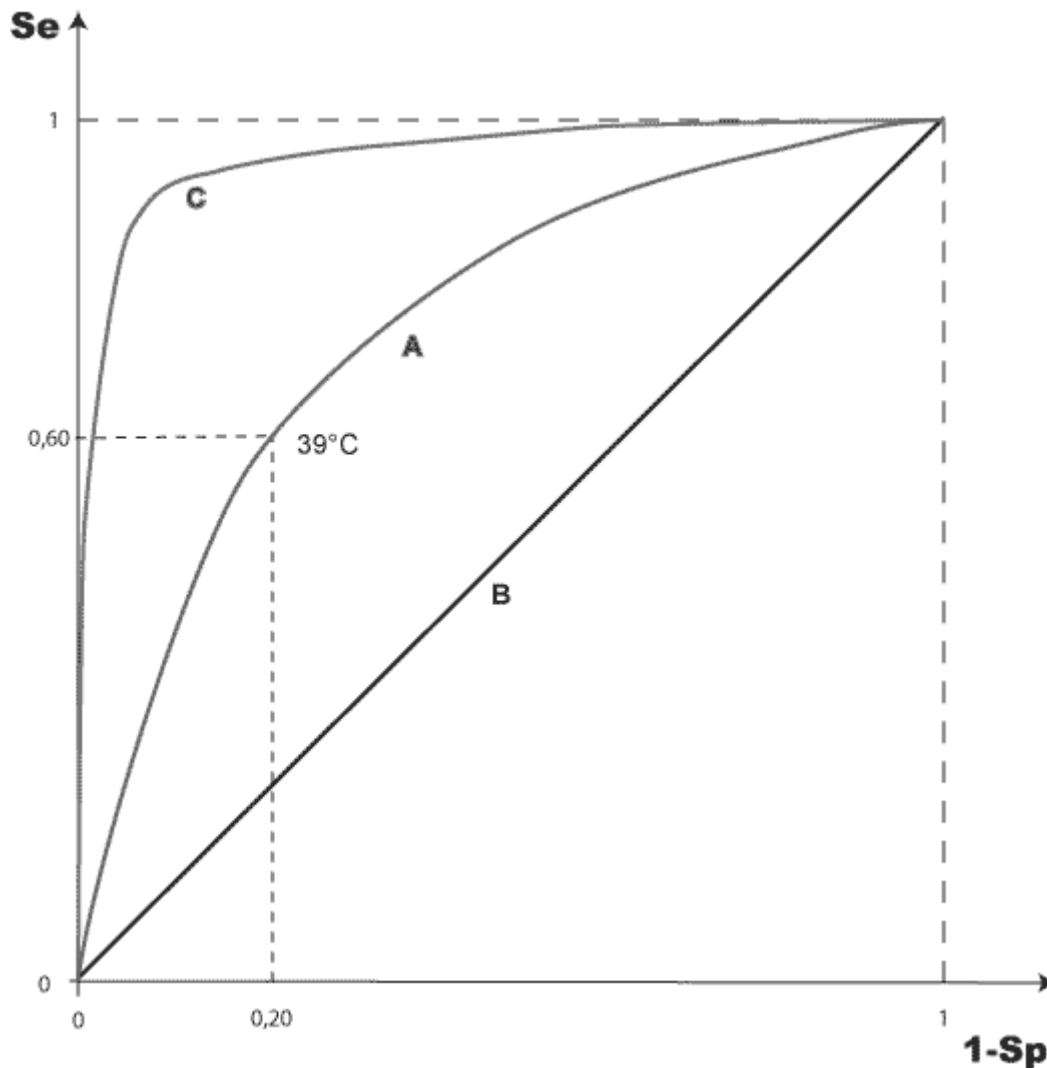
la sensibilité d'un examen pour une affection repose sur la définition de la population des « malades », et est donc caractéristique de la maladie et du signe. En particulier, elle n'est pas susceptible de varier d'un centre à l'autre (d'un service hospitalier spécialisé à une consultation de médecin généraliste, par exemple). Le même raisonnement peut s'appliquer à la spécificité, si on considère qu'elle repose aussi sur la définition de la maladie.

Les valeurs prédictives, au contraire, sont fonctions des proportions respectives de malades et de non-malades dans la population (de la prévalence de la maladie). Or ces proportions sont dépendantes des centres considérés ; les valeurs prédictives des examens varient donc d'un centre à l'autre pour une même maladie, ce qui explique qu'elles sont moins utilisées comme paramètre d'évaluation, même si elles sont intéressantes à connaître pour un centre donné.

## 5.2.4 Choix d'un seuil : courbes ROC

Lorsqu'un examen fournit des résultats de type continu, il faut déterminer le meilleur seuil entre les valeurs pathologiques et les valeurs normales. L'idéal serait d'obtenir une sensibilité et une spécificité égales à 1. Ce n'est généralement pas possible, et il faut tenter d'obtenir les plus fortes valeurs pour ces deux paramètres, sachant qu'ils varient en sens inverse.

On s'aide pour ce choix d'un outil graphique, la courbe ROC (*Receiver Operating Characteristics*). Une courbe ROC est le tracé des valeurs de la sensibilité  $Se$  en fonction de  $1-Sp$ .



Cet exemple (tiré du livre de A.J. Valleron) montre 3 courbes ROC correspondant à 3 examens différents.

La courbe A est celle obtenue pour l'exemple précédent de la température et de la grippe. Le point de la courbe le plus proche du coin supérieur gauche du carré contenant la courbe (ici  $Se = 0,6$ ,  $Sp = 0,8$ , et température =  $39^{\circ}\text{C}$ ) est celui qui permet d'obtenir le meilleur compromis entre sensibilité et spécificité (le coin supérieur gauche correspond à  $Se = Sp = 1$ ). En réalité, on ne choisira pas toujours ce point, car il faut aussi tenir compte des coûts des erreurs diagnostiques : il peut par exemple être beaucoup plus grave de ne pas détecter une maladie, que de traiter à tort.

La courbe B correspond à un examen qui n'apporte rien au diagnostic, puisque les variables signe et maladie sont ici indépendantes :  $Se = Pr(S/M) = 1 - Sp = Pr(S/M)$

La courbe C correspond à un bon critère diagnostique pour lequel on peut obtenir simultanément des valeurs élevées de sensibilité et de spécificité.

## 5.3 Estimation des paramètres de l'évaluation

### 5.3.1 Un échantillon représentatif

#### 5.3.1.1 Les données

Quand on a un échantillon représentatif d'une population, on peut résumer les données de l'expérience par un tableau de contingence  $2 \times 2$ , sur lequel sont indiqués les effectifs suivants :

- VP (Vrais Positifs) : ce sont les individus malades ( $M$ ) et chez lesquels le signe est présent  $\{S\}$  ;
- FP (Faux Positifs) : la maladie est absente  $\{\bar{M}\}$  et le signe est présent  $\{S\}$  ;
- FN (Faux Négatifs) : la maladie est présente  $\{M\}$  et le signe est absent  $\{\bar{S}\}$  ;
- VN (Vrais Négatifs) : la maladie est absente  $\{\bar{M}\}$  et le signe est absent  $\{\bar{S}\}$ .

**Tableau 1**

	$M$	$\bar{M}$
$S$	VP	FP
$\bar{S}$	FN	VN

#### 5.3.1.2 Estimation de la sensibilité et de la spécificité

Par définition, sensibilité =  $Se = Pr(S / M)$

On estime cette probabilité conditionnelle par le rapport des effectifs correspondant sur le tableau de contingence observé :

$$Se \approx \frac{VP}{VP + FN}$$

**Note** : On notera de manière identique, suivant un usage établi, les paramètres vrais, qui sont des probabilités conditionnelles, et leurs estimateurs, qui sont des rapports d'effectifs observés.

$$\text{Spécificité} = Sp = Pr(\bar{S} / \bar{M}) \approx \frac{VN}{VN + FP}$$

Par exemple, calculons les estimateurs de ces paramètres dans le cas où on cherche à diagnostiquer un diabète à partir d'un signe de la forme « la glycémie mesurée à jeun est supérieure à ... ». Pour deux seuils donnés  $S_1$  et  $S_2$ , on obtient les tableaux de contingence ci-dessous :

a. Seuil  $S_1$ 

Tableau 2

	$M$	$\bar{M}$
$S$	90	200
$\bar{S}$	10	300

 b. Seuil  $S_2$ 

Tableau 3

	$M$	$\bar{M}$
$S$	50	25
$\bar{S}$	50	475

On peut estimer les sensibilités et spécificités correspondant aux deux seuils par :

$$Se_1 \approx 90 / 100 = 0,90 ; Sp_1 \approx 300 / 500 = 0,60$$

$$Se_2 \approx 50 / 100 = 0,50 ; Sp_2 \approx 475 / 500 = 0,95.$$

On retrouve ici le fait que sensibilité et spécificité varient en sens inverse.

On constate d'autre part que le seuil  $S_1$  correspond à une bonne sensibilité (l'examen est positif chez 90 % des malades), mais à une spécificité médiocre (l'examen est positif chez 40 % des « non-malades ») ; il peut donc être utilisé pour un examen de dépistage, le diagnostic devant être confirmé ultérieurement par un examen plus spécifique.

Le seuil  $S_2$ , en revanche, induit un test d'une sensibilité qui pourrait être jugée trop faible pour un examen de dépistage. En revanche, sa spécificité peut être acceptable pour un examen de confirmation.

### 5.3.1.3 Estimation des valeurs prédictives

Les estimations s'obtiennent à partir du même tableau des données :

$$VPP = Pr(M/S) \approx \frac{VP}{VP + FP}$$

$$VPN = Pr(\bar{M}/\bar{S}) \approx \frac{VN}{VN + FN}$$

Par exemple, pour les tableaux de contingence vus ci-dessus, on a :

$$VPP_1 \approx 90 / 290 = 0,31 ; VPN_1 \approx 300 / 310 = 0,97$$

$$VPP_2 \approx 50 / 75 = 0,67 ; VPN_2 \approx 475 / 525 = 0,90$$

Ces résultats peuvent s'interpréter ainsi : en affirmant le diagnostic sur la base de la positivité de l'examen, on se trompe dans 69 % des cas avec le seuil  $S_1$  et 33 % des cas avec le seuil  $S_2$  ; et en éliminant le diagnostic en constatant la négativité de l'examen, on se trompe dans 3 % des cas avec le seuil  $S_1$  et 10 % des cas avec le seuil  $S_2$ .

### 5.3.2 Deux échantillons représentatifs

L'inconvénient du schéma expérimental ci-dessus (un seul échantillon) est que, si la maladie est peu fréquente ou rare, il faut constituer un échantillon de très grande taille pour obtenir un nombre suffisant de malades. Les non-malades, au contraire, seront « trop » nombreux. C'est pourquoi on constituera souvent, en pratique, deux échantillons, un échantillon de malades et un échantillon de non-malades. On peut encore résumer les résultats par un tableau comme celui du tableau 1 page 45, mais ce tableau doit être interprété différemment, les proportions respectives des malades et non-malades ne correspondant plus à la réalité : le rapport entre le nombre de malades et le nombre de non-malades du tableau dépend des tailles respectives choisies pour les deux échantillons, et n'a aucun lien avec la fréquence de la maladie dans la population (**la prévalence**).

On peut toujours estimer la sensibilité et la spécificité comme ci-dessus. En effet, la sensibilité par exemple est estimée uniquement à partir de VP et FN, donc de la répartition des malades entre ceux qui présentent le signe et les autres. Or l'échantillon des malades respecte cette répartition.

En revanche, l'estimation précédente des valeurs prédictives utilisait la répartition entre malades et non malades, que le tableau actuel ne représente pas correctement.

L'estimation des valeurs prédictives reste cependant possible à condition de connaître la prévalence de la maladie  $Pr(M)$ . On utilisera les formules introduites section 5.2.2 page 42 :

$$VPP = \frac{Se \times Pr(M)}{Se \times Pr(M) + (1 - Sp) \times (1 - Pr(M))}$$

$$VPN = \frac{Sp \times (1 - Pr(M))}{(1 - Se) \times Pr(M) + Sp \times (1 - Pr(M))}$$

On remplacera dans ces formules la sensibilité et la spécificité par leurs estimations.





# Chapitre 6

## Variables aléatoires

### 6.1 Définition d'une variable aléatoire

Considérons un ensemble fondamental  $E$  correspondant à une certaine expérience. Les éléments de  $E$ , résultats possibles de l'expérience, ne sont généralement pas des nombres. Il est cependant utile de faire correspondre un nombre à chaque élément de  $E$ , 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  $E$ , est une application de  $E$  dans  $\mathfrak{R}$  : à tout résultat possible de l'expérience (à tout élément de  $E$ ), la variable aléatoire  $X$  fait correspondre un nombre.

Lorsque  $E$  est fini ou infini dénombrable, toute application de  $E$  dans  $\mathfrak{R}$  est une variable aléatoire. Lorsque  $E$  est non dénombrable, il existe certaines applications de  $E$  dans  $\mathfrak{R}$  qui ne sont pas des variables aléatoires. En effet, la définition rigoureuse d'une variable aléatoire  $X$  impose que tout intervalle de  $\mathfrak{R}$  soit l'image d'un événement de  $E$  par l'application  $X$ . Cette condition est vérifiée pour toute application  $X$  si  $E$  est fini ou dénombrable, puisque toute partie de  $E$  est un événement. Ce n'est plus vrai si  $E$  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  $E$  dans un sous-ensemble discret de  $\mathfrak{R}$ , le plus souvent  $\mathbf{N}$  ou une partie de  $\mathbf{N}$ . On parle sinon de variable aléatoire **continue**.

Pour un nombre réel  $a$  donné, l'événement constitué de tous les résultats  $\xi$  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  $\xi$  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  $E$ , et si  $k$  est une constante, on peut montrer que les fonctions suivantes sont aussi des variables aléatoires :

$$\begin{aligned} (X + Y)(r) &= X(r) + Y(r) & (X + k)(r) &= X(r) + k \\ (kX)(r) &= kX(r) & (XY)(r) &= X(r) Y(r) \end{aligned}$$

pour tout élément  $r$  de  $E$ .

## 6.2 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  $E$  à valeurs finies :

$$X(E) = \{x_1, x_2, \dots, x_n\}.$$

$X(E)$  devient un ensemble probabilisé si l'on définit la probabilité  $Pr(X = x_i)$  pour chaque  $x_i$ , que l'on note  $p_i$ . L'ensemble des valeurs  $p_i = Pr(X = x_i)$  est appelé distribution ou loi de probabilité de  $X$ .

Puisque les  $p_i$  sont des probabilités sur les événements  $\{X=x_1, X=x_2, \dots, X=x_n\}$ , on a :

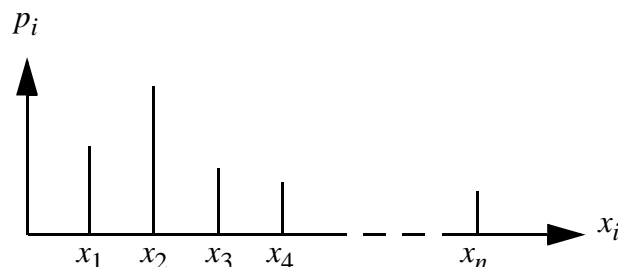
$$(\forall i), p_i \geq 0 \text{ et } \sum_{i=1}^n p_i = 1.$$

### 6.2.1 Représentation d'une loi de probabilité finie

On peut représenter la loi de probabilité  $p_i$  par une table :

$x_1$	$x_2$	.....	$x_n$
$p_1$	$p_2$	.....	$p_n$

Ou par un diagramme en bâtons :



où la hauteur du bâton positionné en  $x_i$  a pour valeur  $p_i$ .

### 6.2.2 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** (parfois aussi appelée **vraie**), ou **espérance mathématique** d'une loi  $p_i$  par

$$\mu_X = E(X) = \sum_{i=1}^n x_i p_i = x_1 p_1 + x_2 p_2 + \dots + x_n p_n.$$

$\mu_X$  peut être notée  $\mu$  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

$$E = \{(1, 1), (1, 2), (1, 3), \dots, (6, 6)\}$$

C'est un espace équiprobable (tous les couples résultats élémentaires du tirage sont équiprobables).

Considérons la variable aléatoire définie comme suit : soit  $r = (a, b)$  un élément quelconque de  $E$  ; on pose  $X(r) = X(a, b) = \max(a, b)$

(la valeur de  $X(r)$  est égale à  $a$  si  $a > b$  et à  $b$  dans le cas contraire).

$X$  est une variable aléatoire sur  $E$  avec  $X(E) = \{1, 2, 3, 4, 5, 6\}$ ,

et la loi de probabilité

$$p_1 = Pr(X = 1) = Pr(\{(1, 1)\}) = 1/36 ;$$

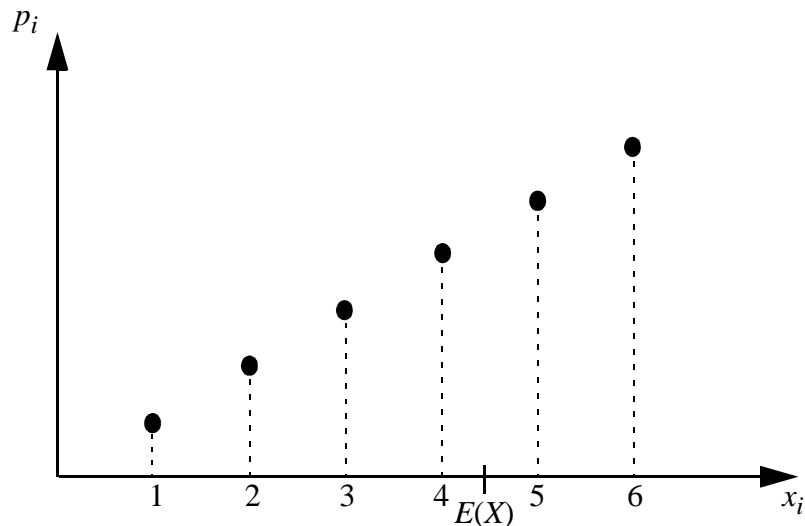
$$p_2 = Pr(X = 2) = Pr(\{(1, 2), (2, 1), (2, 2)\}) = 3/36 ;$$

$$p_3 = 5/36 ; p_4 = 7/36 ; p_5 = 9/36 ; p_6 = 11/36.$$

Soit :

$x_i$	1	2	3	4	5	6
$p_i$	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  $E$ .  
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  $E = \{1, 2, 3, 4, 5, 6\}$  uniforme (équiprobable). Soit  $X(E)$  une première variable aléatoire [noter que l'application définissant  $X$  est l'application identité].

On a  $X(E) = \{1, 2, 3, 4, 5, 6\}$

et  $p_{X1} = p_{X2} = p_{X3} = p_{X4} = p_{X5} = p_{X6} = 1/6$

$E(X) = (1 + 2 + 3 + 4 + 5 + 6) / 6 = 21/6$

Soit  $Y(E)$  une seconde variable aléatoire telle que

$Y(E) = 1$  si le chiffre tiré est impair

$Y(E) = 2$  si le chiffre tiré est pair.

Donc  $Y(E) = \{1, 2\}$

$p_{Y1} = Pr(\{1, 3, 5\}) = 1/2$

$p_{Y2} = Pr(\{2, 4, 6\}) = 1/2$

$E(Y) = 1/2 + 1 = 1,5$

Calculons maintenant la loi de  $(X + Y)(E)$

$(X + Y)(r) = X(r) + Y(r)$

Pour  $r = 1$ ,  $(X + Y)(1) = X(1) + Y(1) = 1 + 1 = 2$

Pour  $r = 2$ ,  $(X + Y)(2) = X(2) + Y(2) = 2 + 2 = 4$

Pour  $r = 3$ ,  $(X + Y)(3) = X(3) + Y(3) = 3 + 1 = 4$

Pour  $r = 4$ ,  $(X + Y)(4) = X(4) + Y(4) = 4 + 2 = 6$

Pour  $r = 5$ ,  $(X + Y)(5) = X(5) + Y(5) = 5 + 1 = 6$

Pour  $r = 6$ ,  $(X + Y)(6) = X(6) + Y(6) = 6 + 2 = 8$

On a donc  $(X + Y)(E) = \{2, 4, 6, 8\}$  et

$Pr(X + Y) = 2) = 1/6$ ,  $Pr(X + Y) = 4) = 2/6$ ,  $Pr(X + Y) = 6) = 2/6$ ,  $Pr(X + Y) = 8) = 1/6$

$E(X + Y) = 2/6 + 8/6 + 12/6 + 8/6 = 30/6$

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 :

$Pr(Y = y_i) = Pr[g(X) = g(x_i)]$

Si  $g$  est une fonction monotone, on a  $g(X) = g(x_i) \Leftrightarrow X = x_i$

D'où  $Pr(Y = y_i) = Pr(X = x_i) = p_i$

Donc :

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

On montre que ce résultat reste valide, même si  $g$  n'est pas monotone.

Par exemple, si l'on doit calculer  $E(X^2)$ , on considère la variable  $Y = X^2$  dont les valeurs

sont  $y_1 = x_1^2, y_2 = x_2^2, \dots, y_n = x_n^2$ . Alors :

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

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

### 6.2.3 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  $\text{var}(X)$  ou  $\sigma_X^2$ , est définie par :

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

L'écart-type de  $X$ , noté  $\sigma(X)$  ou  $\sigma_X$ , est défini par  $\sigma(X) = \sigma_X = \sqrt{\text{var}(X)}$ .

$\sigma_X$  peut être notée  $\sigma$  s'il n'y a pas de confusion possible.

**Remarques :**

- On démontre facilement que  $\text{var}(X) = E(X^2) - \mu_X^2$   
 En effet :  

$$E((X - \mu_X)^2) = \sum_{i=1}^n (x_i - \mu_X)^2 p_i = \sum_{i=1}^n (x_i^2 - 2\mu_X x_i + \mu_X^2) p_i$$

$$E((X - \mu_X)^2) = \sum_{i=1}^n x_i^2 p_i - 2\mu_X \sum_{i=1}^n x_i p_i + \mu_X^2 \sum_{i=1}^n p_i$$

$$E((X - \mu_X)^2) = \sum_{i=1}^n x_i^2 p_i - 2\mu_X^2 + \mu_X^2 = E(X^2) - \mu_X^2$$
- $\sigma_X^2 \geq 0$ , par définition
- Soit  $X$  une variable aléatoire de moyenne  $\mu$  et de variance  $\sigma^2$ .

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

On peut montrer facilement (faites l'exercice) que  $E(Y) = 0$  et  $\text{var}(Y) = E(Y^2) = 1$ .

- Si  $a$  est une constante, on montre que  $\text{var}(X + a) = \text{var}(X)$  et  $\text{var}(aX) = a^2 \text{var}(X)$ .

### 6.2.4 Loi de probabilité produit

Soient  $X$  et  $Y$  deux variables aléatoires finies sur le même espace fondamental  $E$  ayant pour image respective :

$$X(E) = \{x_1, x_2, \dots, x_n\}$$

$$Y(E) = \{y_1, y_2, \dots, y_m\}.$$

Considérons l'ensemble produit

$$X(E) \times Y(E) = \{(x_1, y_1), (x_1, y_2), \dots, (x_n, y_m)\}$$

(ensemble des couples  $(x_i, y_j)$  pour  $i = 1, \dots, n$  et  $j = 1, \dots, m$ )

Cet ensemble produit peut être transformé en ensemble probabilisé si on définit la probabilité du couple ordonné  $(x_i, y_j)$  par  $Pr([X = x_i] \cap [Y = y_j])$  que l'on note  $p_{xi,yj}$ . Cette loi de probabilité de  $X, Y$  est appelée distribution jointe de  $X$  et  $Y$ .

$\begin{array}{c} X \\ \diagdown \\ Y \end{array}$	$x_1$	$x_2$	$x_3$	.....	$x_n$	$\sum_{i=1, n} x_i$
$y_1$	$p_{x1,y1}$	$p_{x2,y1}$				$p_{y1}$
$y_2$	$p_{x1,y2}$					$p_{y2}$
.....						
$y_m$	$p_{x1,ym}$					
$\sum_{j=1, m} y_j$	$p_{x1}$	$p_{x2}$				1
	$m$	$n$				

Les probabilités  $p_{xi} = \sum_{j=1} p_{xi,yj}$  et  $p_{yj} = \sum_{i=1} p_{xi,yj}$

sont souvent appelées lois de probabilité marginales de  $X$  et de  $Y$ . Il s'agit simplement de leurs distributions.

La loi de probabilité  $p_{xi,yj}$  possède, bien entendu, les propriétés d'une loi :

1.  $p_{xi,yj} \geq 0, \forall i, j$
2.  $\sum_{i=1} \sum_{j=1} p_{xi,yj} = 1$

Soient  $\mu_X$  et  $\mu_Y$  les espérances de  $X$  et de  $Y$ ,  $\sigma_X$  et  $\sigma_Y$  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 :

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

De même que pour la variance (voir section 6.2.3), on a :

$$cov(X, Y) = E(XY) - \mu_X \mu_Y$$

La covariance de  $X$  et  $Y$  se note aussi  $\sigma_{XY}$ .

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

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

On peut vérifier que

$$\rho(X, Y) = \rho(Y, X)$$

$$-1 \leq \rho(X, Y) \leq 1$$

$$\rho(X, X) = 1$$

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

## 6.2.5 Variables aléatoires indépendantes

Soient  $X$  et  $Y$  deux variables aléatoires sur un même espace fondamental  $E$ .  $X$  et  $Y$  sont indépendantes si tous les événements  $X = x_i$  et  $Y = y_j$  sont indépendants :

$$Pr([X = x_i] \cap [Y = y_j]) = Pr(X = x_i) \cdot Pr(Y = y_j) \text{ pour tous les couples } (i, j).$$

Autrement dit, si  $p_{xi}$  et  $p_{yj}$  sont les distributions respectives de  $X$  et  $Y$ , les variables sont indépendantes si et seulement si on a

$$p_{xi,yj} = p_{xi}p_{yj}$$

(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.  $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$
3.  $\text{cov}(X, Y) = 0$  et  $\rho(X, Y) = 0$

## 6.2.6 Fonction de répartition

Si  $X$  est une variable aléatoire, on définit sa fonction de répartition  $F(x)$  par

$$F(x) = Pr(X \leq x) \text{ pour tout } x \in \Re$$

$$\text{Si } X \text{ est une variable aléatoire discrète on a } F(x) = \sum_{x_i \leq x} Pr(X = x_i) = \sum_{x_i \leq x} p_i$$

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

$$\lim_{x \rightarrow -\infty} F(x) = 0 \text{ et } \lim_{x \rightarrow \infty} F(x) = 1$$

## 6.3 Variables infinies dénombrables

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

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

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

## 6.4 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  $X(E)$  est un intervalle de  $\mathfrak{R}$  est une variable aléatoire continue (continue par opposition à discrète, cf supra).

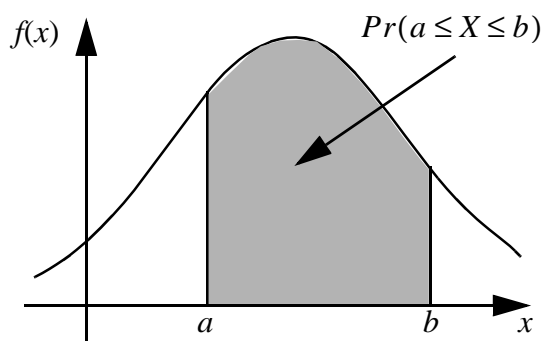
Rappelons que, par définition d'une variable aléatoire,  $a \leq X \leq b$  est un événement de  $E$  dont la probabilité est bien définie.

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

$$\int_a^b f(x)dx = Pr(a \leq X \leq b)$$

Remarques

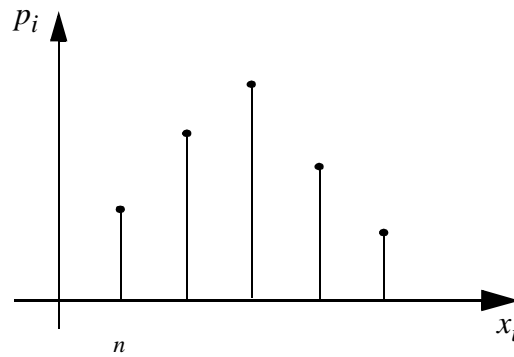
1. Si  $f$  est donnée, la probabilité  $Pr(a \leq X \leq b)$  est la surface sous la courbe entre  $a$  et  $b$



2. Le passage du discret au continu transforme les sommes  $\Sigma$  en intégrales  $\int$  et  $p_i$  en  $f(x)dx$ .



Ainsi, soit  $X$  une variable aléatoire discrète et  $p_i$  sa distribution



La formule  $Pr(x_k \leq X \leq x_n) = \sum_{i=k}^n p_i$  est analogue à  $Pr(a \leq X \leq b) = \int_a^b f(x)dx$

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

1.  $f(x) \geq 0$  (analogue à  $p_i \geq 0$ )
2.  $\int_{\Re} f(x)dx = 1$  (analogue à  $\sum_i p_i = 1$ )
3.  $\mu_X = E(X) = \int_{\Re} xf(x)dx$  (analogue à  $\sum_i x_i p_i$ )
4.  $\sigma_X^2 = var(X) = \int_{\Re} (x - \mu_X)^2 f(x)dx$  (analogue à  $\sum_i (x_i - \mu_X)^2 p_i$ )
5.  $\sigma_X^2 = var(X) = \int_{\Re} x^2 f(x)dx - \mu_X^2$  (analogue à  $\sum_i x_i^2 p_i - \mu_X^2$ )
6.  $\sigma(X) = \sigma_X = \sqrt{var(X)}$
7.  $F(x) = Pr(X \leq x) = \int_{-\infty}^x f(\tau)d\tau$  (analogue à  $\sum_{x_i \leq x} p_i$ )

Les propriétés de la fonction de répartition données section 6.2.6 page 55 sont conservées.

$$8. \quad Pr(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

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 7

## Exemples de distributions

### 7.1 Lois discrètes

#### 7.1.1 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  $\Pi$  la probabilité de l'événement succès :

$$Pr(\{\text{succès}\}) = Pr(X = 1) = \Pi$$

d'où

$$Pr(\{\text{échec}\}) = Pr(X = 0) = 1 - \Pi$$

##### Espérance de $X$

$$\mu_X = E(X) = \sum x_i Pr(X = x_i) = 1 \times Pr(X = 1) + 0 \times Pr(X = 0) = \Pi$$

##### Variance de $X$

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

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

$$\sigma_X^2 = \Pi - \Pi^2 = \Pi(1 - \Pi)$$

#### 7.1.2 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  $\Pi$  la probabilité de l'événement élémentaire succès. A cette expérience multiple on associe une variable aléatoire  $X$  qui mesure le nombre de succès obtenus.

**Distribution de  $X$** 

On montre aisément que la probabilité d'avoir  $k$  succès lors de  $n$  épreuves répétées est

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

**Rappel**

$n! = 1 \times 2 \times \dots \times n$  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  $(1-\Pi)^n$  ; la probabilité d'avoir au moins un succès est donc  $1 - (1-\Pi)^n$  (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 :

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

**Exemples**

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

- a. Probabilité que l'on ait exactement 2 faces

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

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

- b. Probabilité d'avoir 4 faces ou plus (au moins 4 faces)

C'est aussi la probabilité d'avoir au plus 2 piles (0, 1 ou 2 piles)

$$p_4 = Pr(4 \text{ faces}) = \frac{6!}{2!4!} \cdot \left(\frac{1}{2}\right)^4 \cdot \left(\frac{1}{2}\right)^2 = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6}{1 \times 2 \times 1 \times 2 \times 3 \times 4} \cdot \frac{1}{16} \cdot \frac{1}{4} = \frac{15}{64}$$

$$p_5 = Pr(5 \text{ faces}) = \frac{6!}{1!5!} \cdot \left(\frac{1}{2}\right)^5 \cdot \frac{1}{2} = \frac{1 \times 2 \times 3 \times 4 \times 5 \times 6}{1 \times 2 \times 3 \times 4 \times 5} \cdot \frac{1}{32} \cdot \frac{1}{2} = \frac{6}{64}$$

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

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

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

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

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

- b. la probabilité de n'avoir aucun succès

$$Pr(\text{aucun succès}) = (1 - \Pi)^7 = \left(\frac{2}{3}\right)^7 = \frac{128}{2187}$$

### Propriétés

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

$k$	0	1	2	.....	$n$
$Pr(X=k)$	$(1-\Pi)^n$	$\binom{n}{1} (1-\Pi)^{n-1} \Pi$	$\binom{n}{2} (1-\Pi)^{n-2} \Pi^2$		$\Pi^n$

On dit que  $X$  est distribuée selon une loi binomiale  $B(n, \Pi)$ .

On peut montrer que

Distribution binomiale $B(n, \Pi)$	
Espérance	$\mu = n\Pi$
Variance	$\sigma^2 = n\Pi(1 - \Pi)$
Ecart-type	$\sigma = \sqrt{n\Pi(1 - \Pi)}$

## 7.1.3 Loi de Poisson

La loi de Poisson (due à Siméon Denis Poisson en 1837) est la loi du nombre d'événements observé dans le cas où ces **événements** sont **indépendants et faiblement probables**. Elle peut s'appliquer

au nombre d'accidents, à l'apparition d'anomalies diverses, à la gestion des files d'attentes, au nombre de colonies bactériennes dans une boîte de Pétri, etc.

### Définition

Soit  $X$  la variable aléatoire représentant le nombre d'apparitions indépendantes d'un événement faiblement probable. La probabilité d'avoir  $k$  apparitions de l'événement est

$$Pr(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$$

Cette loi dépend d'un paramètre  $\lambda$ , nombre réel strictement positif.

Les nombres  $k$  possibles sont toutes les valeurs entières 0, 1, 2, etc. Cependant, lorsque  $k$  est suffisamment grand, la probabilité correspondante devient extrêmement faible.

### Propriétés

- On peut montrer que

Loi de Poisson $P(\lambda)$	
Espérance	$\mu = \lambda$
Variance	$\sigma^2 = \lambda$
Ecart-type	$\sigma = \sqrt{\lambda}$

La démonstration utilise le fait que  $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$

- Si deux variables aléatoires indépendantes  $X_1$  et  $X_2$  sont distribuées selon des lois de Poisson de paramètres  $\lambda_1$  et  $\lambda_2$ , alors la variable  $X_1 + X_2$  est distribuée selon une loi de Poisson de paramètre  $\lambda_1 + \lambda_2$ .

### Remarques

Si on connaît la probabilité de n'observer aucun événement  $Pr(X=0) = p$  :

- D'après la formule,  $p = e^{-\lambda} \frac{\lambda^0}{0!} = e^{-\lambda}$

On en déduit :

$$\lambda = -\ln p$$

- $$Pr(X = 1) = e^{-\lambda} \frac{\lambda^1}{1!} = p\lambda,$$

$$Pr(X = 2) = e^{-\lambda} \frac{\lambda^2}{2!} = Pr(X = 1) \frac{\lambda}{2},$$

$$Pr(X = 3) = e^{-\lambda} \frac{\lambda^3}{3!} = Pr(X = 2) \frac{\lambda}{3},$$

.....

$$Pr(X = k) = Pr(X = k - 1) \frac{\lambda}{k}$$

On peut ainsi calculer facilement de proche en proche les probabilités des diverses valeurs de  $k$ .

**Lien avec la loi binomiale**

Si une variable aléatoire  $X$  est distribuée selon une loi binomiale  $B(n, \Pi)$ , on montre que si  $\Pi$  est petit (en pratique inférieur à 0,1) et  $n$  assez grand (supérieur à 50), la loi binomiale peut être approximée par une loi de Poisson de paramètre  $\lambda = n\Pi$ .

Les calculs sont plus simples avec la loi de Poisson qu'avec la binomiale.

Notons que puisque  $X$  est distribuée selon une loi binomiale, ses valeurs possibles ne peuvent dépasser  $n$ , alors que l'approximation par la loi de Poisson autorise des valeurs supérieures. Cependant le calcul fournit des probabilités très faibles pour ces valeurs aberrantes.

## 7.2 Lois continues

### 7.2.1 Loi normale

#### 7.2.1.1 Définition

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

#### 7.2.1.2 Propriétés

**Allure de la courbe**

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

Exemples :

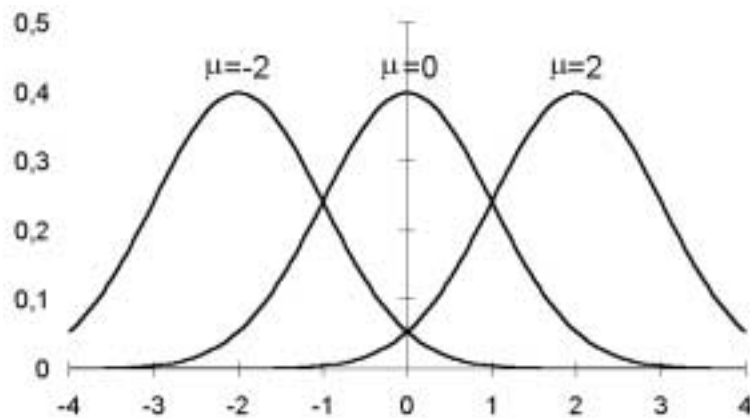


Figure 2 :  $N(\mu, 1)$  pour les valeurs de  $\mu$  -2 ; 0 et 2

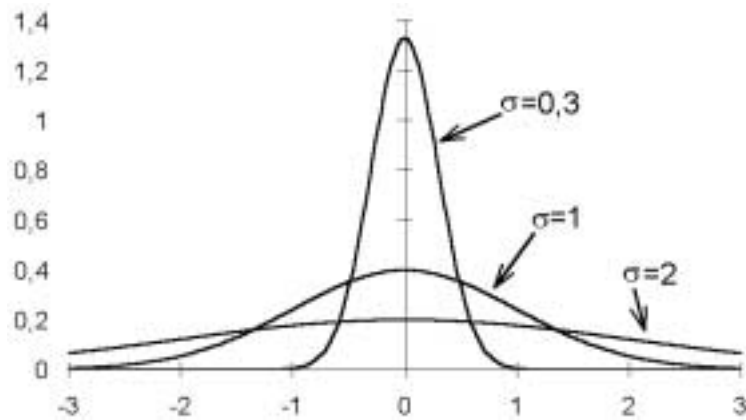


Figure 3 :  $N(0, \sigma^2)$  pour les valeurs de  $\sigma$  0,3 ; 1 et 2

### Caractéristiques

Loi normale $N(\mu, \sigma^2)$	
Espérance	$\mu$
Variance	$\sigma^2$
Ecart-type	$\sigma$

### La distribution normale centrée réduite

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



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

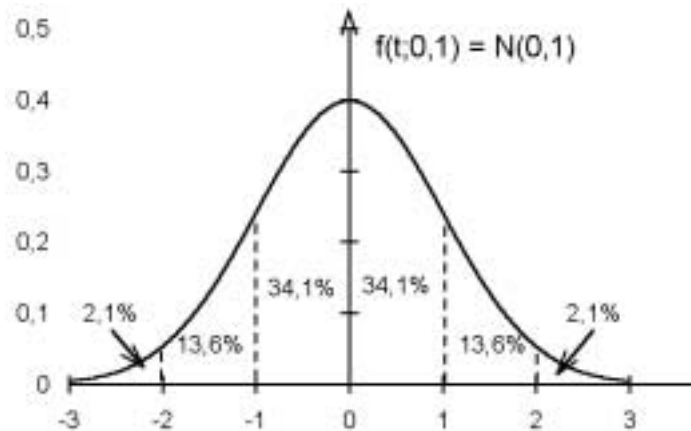


Figure 4 : loi normale centrée réduite  $N(0, 1)$

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  $\alpha$  donnée, de trouver les bornes  $-z_\alpha, +z_\alpha$  d'un intervalle symétrique autour de 0, tel que

$$Pr(t \notin [-z_\alpha; +z_\alpha]) = \alpha$$

ou encore, à partir de  $z_\alpha$ , de trouver  $\alpha$ .

D'où par exemple :

$$Pr(t \in [-z_\alpha; +z_\alpha]) = 1 - \alpha$$

$$Pr(t > z_\alpha) = Pr(t < -z_\alpha) = \alpha/2$$

On observe ainsi que environ 68 % de la surface est comprise entre (-1 et +1), 95 % entre (-2 et +2) et 99 % entre (-3 et +3) (la table ne permet pas de trouver des valeurs plus précises).

### Transformation d'une loi normale quelconque en loi normale centrée réduite

Soit une variable  $X$  distribuée selon une loi normale d'espérance  $\mu$  et d'écart-type  $\sigma$ .

Alors la variable  $t = \frac{X - \mu}{\sigma}$  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  $Pr(a \leq X \leq b)$ . Par la transformation, on a  $Pr(a \leq X \leq b) = Pr(c \leq t \leq d)$  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  $d$ .

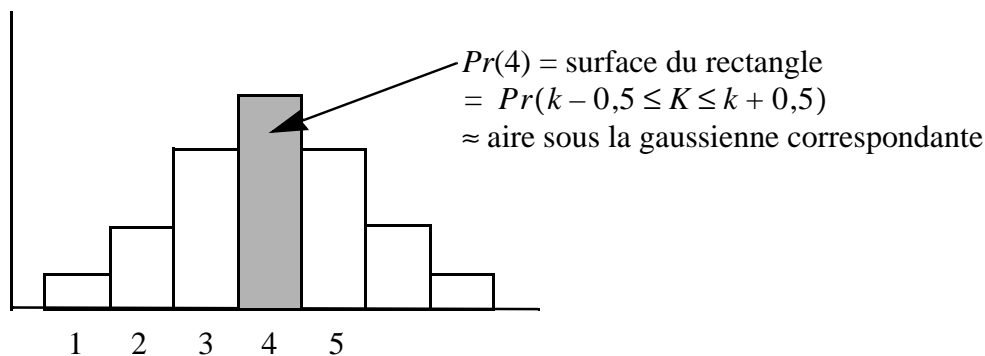
On remarque en particulier que  $Pr(-2 \leq t \leq 2) = Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0,95$

### Approximation de la distribution binomiale par la loi normale

Lorsque  $n$  est grand, et que  $\Pi$  et  $1-\Pi$  ne sont pas trop proches de 0 (en pratique si  $n\Pi \geq 5$  et  $n(1-\Pi) \geq 5$ ), alors on constate que la distribution binomiale tend vers la distribution normale de moyenne  $n\Pi$  et de variance  $n\Pi(1-\Pi)$  ; plus précisément, pour une variable  $K$  distribuée selon une loi binomiale  $B(n, \Pi)$  et une variable  $X$  distribuée selon une loi normale  $N(\mu = n\Pi, \sigma^2 = n\Pi(1-\Pi))$ , on a :

$$Pr(K = k) = Pr(k) = Pr(k - 0,5 \leq K \leq k + 0,5) \approx Pr(k - 0,5 \leq X \leq k + 0,5)$$

On choisit l'artifice de représenter graphiquement  $Pr(k)$  par un rectangle dont la base est  $[k - 0,5, k + 0,5]$  et la surface est  $Pr(k)$  pour comparer la loi discrète  $Pr(k)$  et la loi normale continue.



### Approximation de la loi de Poisson par la loi normale

Lorsque son paramètre  $\lambda$  est grand (en pratique supérieur à 25), une loi de Poisson peut être approchée par une loi normale d'espérance  $\lambda$  et de variance  $\lambda$ .

Le principe est analogue à celui utilisé pour l'approximation de la loi binomiale par la loi normale.

## 7.2.2 Loi du $\chi^2$ (chi-2)

### 7.2.2.1 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  $X_1, \dots, X_n$  des variables aléatoires indépendantes, chacune étant distribuée selon une loi normale centrée réduite :

$$\forall i, X_i \sim N(0, 1)$$

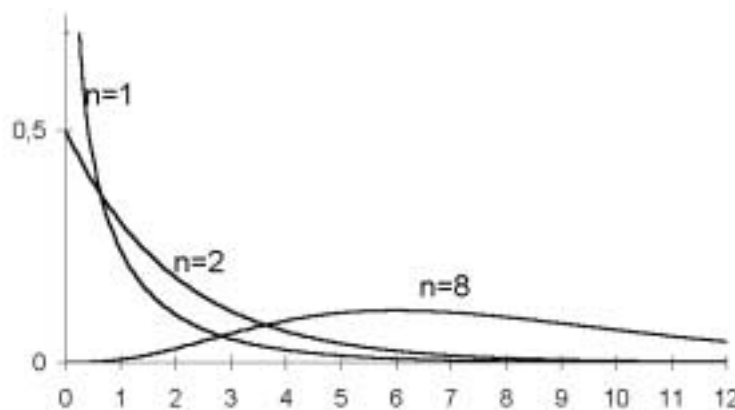
La distribution de  $S = X_1^2 + X_2^2 + \dots + X_n^2$  (somme des carrés des  $X_i$ ) est appelée loi de  $\chi^2$  à  $n$  degrés de liberté (en abrégé d. d. l.), que l'on note  $\chi^2(n)$  où  $n$  est le nombre de d. d. l., seul paramètre

de la loi.

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

### 7.2.2.2 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  $n \rightarrow +\infty$ . On a donc, après avoir centré et réduit cette variable :

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

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  $\chi^2$  (voir en fin de polycopié, table A.4). Cette table indique pour une probabilité  $\alpha$  donnée, et un degré de liberté  $n$  donné, la valeur  $K_{n,\alpha}$  telle que  $Pr(X > K_{n,\alpha}) = \alpha$ .

## 7.2.3 Loi de Student (hors programme)

Il s'agit encore d'une loi dérivée de la loi normale, très utilisée dans les tests statistiques. On con-

sidère une première variable aléatoire  $X$ , distribuée selon une loi normale centrée réduite, puis une seconde variable  $Y$ , indépendante de  $X$ , distribuée selon un  $\chi^2$  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  $t(n)$ .

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

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 assimilée à celle d'une loi normale centrée réduite.

## 7.2.4 Loi exponentielle (hors programme)

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	$1/\lambda$

# Chapitre 8

## Statistiques descriptives

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

### 8.1 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, sans pouvoir l'anticiper, on l'appellera **variable aléatoire**. Le dispositif permettant d'obtenir une valeur de la variable est l'**expérience aléatoire**. Cette définition imagée est compatible avec la définition du chapitre 6.

#### 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** qui produisent 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 **ordinaire**. Exemple : intensité d'une douleur (faible, moyenne, forte).

Si les valeurs ne peuvent pas être ordonnées, il s'agit d'une variable **catégorielle** (ou **nominale**).

#### 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.

#### 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 valeur 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 de 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 aléatoire produit un résultat  $x_i$  ; on disposera donc de  $x_1, \dots, x_n$ , ensemble appelé **échantillon de valeurs** de la variable étudiée  $X$ .

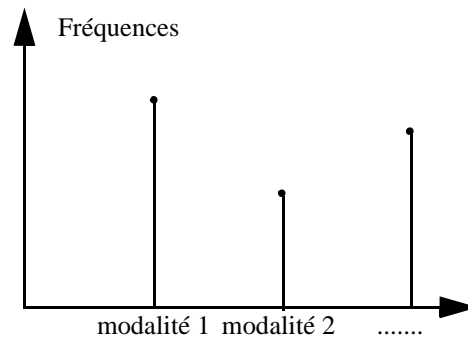
- De façon plus formelle, on définit un échantillon de la manière suivante :  
Un **échantillon** de taille  $n$  **d'une variable aléatoire**  $X$  est un ensemble  $X_1, X_2, \dots, X_n$  de  $n$  variables aléatoires, indépendantes entre elles, et ayant chacune la même distribution que  $X$ . On peut donc dire qu'un échantillon de valeurs de  $X$  est **une** réalisation de l'échantillon de la variable  $X$  tel qu'il vient d'être défini.

## 8.2 Représentation complète d'une série d'expériences

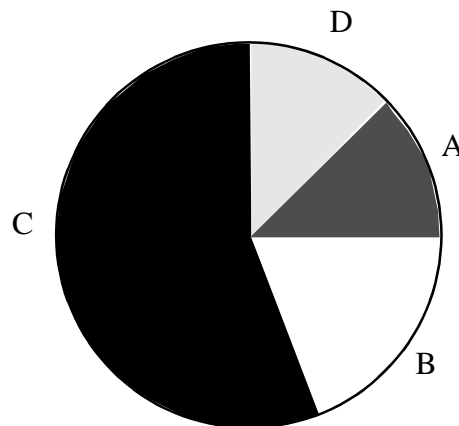
### 8.2.1 Cas d'une variable qualitative

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 observée** par substitution aux probabilités inconnues des fréquences observées. Si la variable est ordinale, on respectera cet ordre dans l'énumération des modalités por-

tées en abscisses.

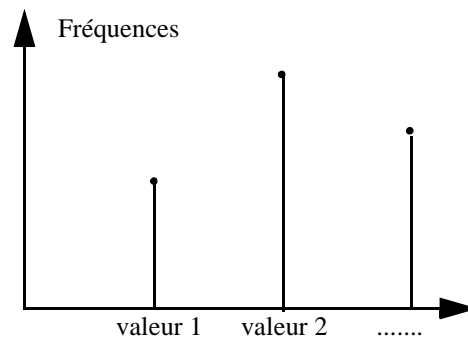


Un autre type de représentation est utilisé : la représentation en camembert où les différentes modalités sont représentées par secteurs angulaires d'angles au centre proportionnels aux fréquences observées.



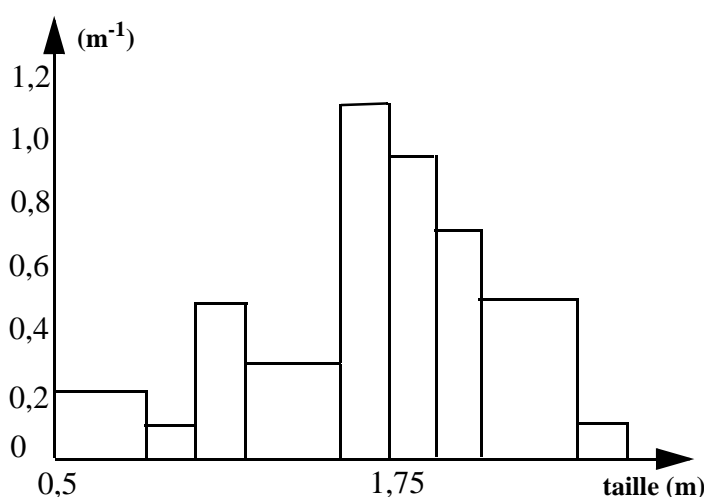
## 8.2.2 Cas d'une variable quantitative discrète

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



### 8.2.3 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 observée 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.



## 8.3 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.

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

### 8.3.1 Indicateurs de localisation des valeurs

#### Médiane observée

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 observée

C'est l'indicateur de localisation le plus fréquemment utilisé. La moyenne observée d'un échantillon de  $n$  valeurs  $x_1, \dots, x_n$  est définie comme la moyenne arithmétique de ces valeurs ; on la note souvent  $m_x$ , ou simplement  $m$  s'il n'y a pas de confusion possible :

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

Avec la série précédente, qui comporte  $n = 7$  valeurs, on obtient :

$$m = \frac{1}{7} \sum_{i=1}^7 x_i = \frac{12 + 3 + 24 + 1 + 5 + 8 + 7}{7} = 8,57$$

### 8.3.2 Indicateurs de dispersion des valeurs

#### Variance observée

La variance observée d'un échantillon  $\{x_i\} \ i = 1, \dots, n$  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  $n$  pour que la variance observée 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 observé

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

## 8.4 Reformulation de la moyenne et de la variance observées

### 8.4.1 Reformulation de la moyenne observée

Prenons le cas d'une variable quantitative discrète.

Les données sont notées  $x_1, \dots, x_n$ .

Les  $k$  valeurs possibles de la variable sont notées  $\text{val}_1, \text{val}_2, \dots, \text{val}_k$ .

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

Chaque donnée  $x_i$  coïncide avec une certaine valeur  $\text{val}_j$

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 ;  $x_9 = 6 = \text{val}_6$

$$\text{Alors : } \sum_{i=1}^n x_i = \sum_{j=1}^k n_j \text{val}_j$$

où  $n_j$  est le nombre de fois où une observation coïncide avec  $\text{val}_j$

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

$$\text{Finalement } m = \frac{1}{n} \sum_{i=1}^n x_i = \sum_{j=1}^k \frac{n_j}{n} \text{val}_j$$

Mais  $\frac{n_j}{n}$  est une approximation de  $\Pr(\text{face marquée} = \text{val}_j)$

Ainsi  $m$  est une estimation - une appréciation - de :

$$\sum_j \text{val}_j \Pr(\text{valeur de la variable} = \text{val}_j)$$

c'est-à-dire une appréciation de l'espérance mathématique de la variable.

On raccorde ainsi une moyenne observée 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  $Pr(\text{valeur de la variable} = \text{certain niveau})$ . La moyenne observée permet d'estimer quelque chose de plus grossier, une combinaison de toutes ces probabilités, précisément l'espérance mathématique

$$\mu = \sum_j \text{val}_j Pr(\text{valeur de la variable} = \text{val}_j)$$

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 observée approxime là encore l'espérance mathématique.

## 8.4.2 Reformulation de la variance observée

De la même façon on peut obtenir le résultat suivant :  $s^2$  est une approximation de la grandeur  $\sigma^2 = \sum_j (\text{val}_j - \mu)^2 Pr(\text{valeur de la variable} = \text{val}_j)$

Cette expression, introduite dans le chapitre 6 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 observée  $s^2$  approxime :

$$\sigma^2 = \int_{\mathfrak{R}} (x - \mu)^2 f(x) dx$$

LES DIFFERENCES ENTRE CES NOTIONS DE MOYENNE ET VARIANCE « VRAIES », ET DE MOYENNE ET VARIANCE OBSERVEES SONT **ESSENTIELLES** ; NOUS ENGAGEONS LE LECTEUR A BIEN LES COMPRENDRE AVANT DE POURSUIVRE.

## 8.5 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  $Pr(\text{valeur de la variable} = 1)$  car

$Pr(\text{valeur de la variable} = 1) + Pr(\text{valeur de la variable} = 0) = 1$ .

On utilise la notation conventionnelle suivante :  $Pr(\text{valeur de la variable} = 1)$  SE NOTE  $\Pi$ .

### 8.5.1 Expression de l'espérance mathématique de $X$

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

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

Ainsi,  $\mu = \Pi = Pr(\text{valeur de la variable} = 1) = \text{probabilité de la modalité codée 1} = \text{PROPORTION VRAIE des individus présentant la modalité 1}$ .

### 8.5.2 Expression de la variance de $X$

$$\sigma^2 = \sum_j (\text{val}_j - \mu)^2 Pr(\text{valeur de la v.a.} = \text{val}_j) = (0 - \Pi)^2 (1 - \Pi) + (1 - \Pi)^2 \Pi = \Pi(1 - \Pi)$$

### 8.5.3 Interprétation de la moyenne observée

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

Ainsi,  $m$  coïncide avec la fréquence observée de la modalité codée 1. Cette fréquence sera notée  $p$  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,  $m$  (donc  $p$ ) sera la fréquence observée 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 » =  $\Pi$
  - sa variance « vraie » =  $\Pi(1 - \Pi)$
- UNE PROPORTION OBSERVEE EST UNE MOYENNE OBSERVEE.

## 8.6 Conclusion : la variable aléatoire moyenne arithmétique

On a jusqu'ici associé une valeur de moyenne observée 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 observée. Autrement dit, la moyenne observée doit être vue comme une nouvelle variable aléatoire que nous appellerons moyenne arithmétique ; on la notera  $M$ . Dans certains cas, afin de rappeler que cette variable dépend de  $n$ , on notera  $M_n$  la variable construite à partir de  $n$  réalisations de  $X$ . On utilisera la terminologie suivante :

on dira que  $M$  (ou  $M_n$  si nécessaire) est la VARIABLE ALEATOIRE MOYENNE ARITHMETIQUE DEDUITE DE LA VARIABLE ALEATOIRE  $X$ , FONDEE SUR  $n$  REPETITIONS

ou, de façon équivalente que  $M$  (ou  $M_n$  si nécessaire) est la VARIABLE ALEATOIRE MOYENNE ARITHMETIQUE ASSOCIEE A LA VARIABLE ALEATOIRE  $X$ , FONDEE SUR  $n$  REPETITIONS

#### Remarque

Dans le cas où  $X$  est une variable de Bernoulli,  $M_n$  sera notée  $P_n$  (et  $M$  simplement  $P$ ). Il s'agit d'une variable aléatoire proportion dont on connaît déjà pratiquement la distribution puisque  $nP_n \sim B(n, \Pi)$  (voir section 7.1.2 page 59).

# 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
<b>variable qualitative</b>	répartition observée	répartition vraie
<b>variable quantitative discrète</b>	histogramme en bâtons	répartition vraie
<b>variable quantitative continue</b>	histogramme	densité de probabilité

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

	moyenne observée	espérance, ou moyenne « vraie »
<b>variable discrète</b>	$m = \frac{1}{n} \sum_{i=1}^n x_i$	$\mu = \sum_{j=1}^k \text{val}_j Pr(\text{variable} = \text{val}_j)$
<b>variable continue</b>	$m = \frac{1}{n} \sum_{i=1}^n x_i$	$\mu = \int_{\mathfrak{R}} xf(x)dx$
<b>variable de Bernoulli</b>	$m$ est notée $p$	$\mu = Pr(\text{variable} = 1)$ est notée $\Pi$

4. **Variances (variables quantitatives)**

	variances observées	variances « vraies »
<b>variable discrète</b>	$s^2 = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^n x_i^2 - m^2 \right]$	$\sigma^2 = \sum_{j=1}^k (\text{val}_j - \mu)^2 Pr(\text{variable} = \text{val}_j)$
<b>variable continue</b>	$s^2 = \frac{n}{n-1} \left[ \frac{1}{n} \sum_{i=1}^n x_i^2 - m^2 \right]$	$\sigma^2 = \int_{\mathfrak{R}} (x - \mu)^2 f(x)dx$

5. **Variables centrée et centrée réduite associées à une variable  $X$**

Si  $X$  est une variable aléatoire de moyenne  $\mu$  et de variance  $\sigma^2$ ,

- 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 9

## Fluctuations de la moyenne observée : la variable aléatoire moyenne arithmétique

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 arithmétique ».

### 9.1 Première propriété de la variable aléatoire moyenne arithmétique

#### 9.1.1 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  $Pr(X = 0) = Pr(X = 1) = 1/2$ . On a vu qu'une telle variable a pour espérance  $1/2$ , 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 arithmétique calculée sur un échantillon de deux individus », notée  $M_2$ , ici deux lancers de pièce.

On isole cette variable. Quelles valeurs peut-elle prendre, avec quelles probabilités ?

jet 1 : résultats	Proba jet 1	jet 2 : résultats	Proba jet 2	Proba jet1, jet2	$M_2$
0	1/2	0	1/2	1/4	$1/2(0+0) = 0$
0	1/2	1	1/2	1/4	$1/2(0+1) = 1/2$

jet 1 : résultats	Proba jet 1	jet 2 : résultats	Proba jet 2	Proba jet1, jet2	$M_2$
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,  $Pr(M_2 = 0) = \frac{1}{4}$ ,  $Pr(M_2 = \frac{1}{2}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$ ,  $Pr(M_2 = 1) = \frac{1}{4}$   
 Alors :

- moyenne vraie de  $M_2 = 0 \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{4} = \frac{1}{2} =$  moyenne vraie de  $X$
- variance vraie de  $M_2 = \left(0 - \frac{1}{2}\right)^2 \times \frac{1}{4} + \left(\frac{1}{2} - \frac{1}{2}\right)^2 \times \frac{1}{2} + \left(1 - \frac{1}{2}\right)^2 \times \frac{1}{4} = \frac{1}{8} = \frac{1}{2} \times \frac{1}{4}$

Ainsi la variance « vraie » de la moyenne arithmétique est plus faible que la variance « vraie » de la variable d'origine (la moitié ici). L'espérance reste inchangée. Et ainsi vont les choses si la taille des échantillons (ici 2) qui constituent les unités statistiques augmente. La dispersion de  $M$  diminue au fur et à mesure que  $M$  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 arithmétique deviennent de plus en plus probables dans un voisinage de l'espérance car le nombre de situations pouvant donner une valeur observée proche de l'espérance augmente dans ce voisinage. Cela est dû au fait que l'espérance mathématique est « au milieu » des valeurs possibles. On le voit sur l'exemple ci-dessus où l'espérance 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 observée 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.

## 9.1.2 Généralisation

- L'espérance mathématique, ou moyenne « vraie », de la variable aléatoire moyenne arithmétique calculée sur un échantillon de taille  $n$  coïncide avec la moyenne « vraie » de la variable étudiée, ce que l'on peut résumer par :

$$E(M_n) = E(X)$$

- La variance « vraie » de la variable aléatoire moyenne arithmétique calculée sur un échantillon de taille  $n$  est égale à la variance « vraie » de la variable **DIVISÉE PAR  $n$** , ce que l'on peut résumer par :

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

d'où la relation entre écarts-types :

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

- iii. Dans le cas où  $X$  est une variable de Bernoulli de paramètre  $\Pi$  ( $Pr(X = 1) = \Pi$ ), les relations précédentes deviennent :

$$\mu(P_n) = \Pi$$

$$\sigma^2(P_n) = \frac{\Pi(1 - \Pi)}{n}$$

## 9.2 Seconde propriété de la variable aléatoire moyenne arithmétique : le théorème central limite

On souhaiterait comparer, par curiosité, les distributions de plusieurs moyennes arithmétiques, 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 :

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

Maintenant toutes ces variables ont en commun leur espérance (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 arithmétiques 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  $\mu$ , de variance « vraie »  $\sigma^2$ . Soit  $M_n$  la variable aléatoire moyenne arithmétique associée à  $X$  construite sur  $n$  répétitions.

La distribution limite de la variable aléatoire  $\frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$  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 variable aléatoire moyenne arithmétique calculée sur un échantillon de taille suffisante. Nous reviendrons plus loin, au paragraphe résumé et précisions (voir page 85), 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.

## 9.3 Etude de la distribution normale (rappel)

La distribution limite que l'on a mise en évidence dépeint une variable aléatoire d'espérance mathématique 0 et 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  $\alpha$  et  $z_\alpha$ , où ces valeurs ont le sens suivant (voir figure 5) :

$$Pr(X \notin [-z_\alpha ; +z_\alpha]) = \alpha$$

En particulier, pour  $\alpha = 0,05$ , la valeur  $z_\alpha$  lue dans la table est 1,96, d'où  $z_{0,05} = 1,96$

On peut voir facilement que toute probabilité  $Pr(X \in [a,b])$  s'obtient à partir d'une telle table, quelles que soient les valeurs de  $a$  et  $b$ .

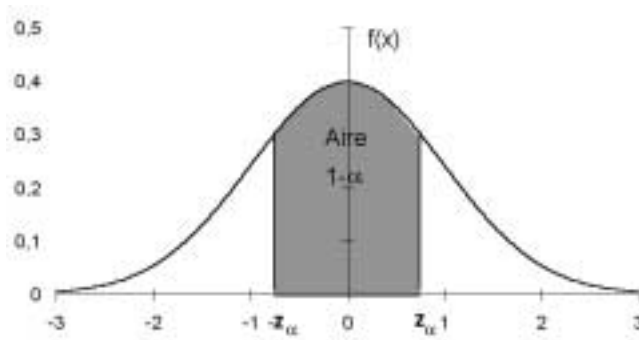


Figure 5 : loi normale centrée réduite

### 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 l'espérance est  $\mu$  et la variance  $\sigma^2$ , est distribuée selon une loi normale d'espérance  $\mu$  et de variance  $\sigma^2$ . 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  $N(2, 9, 4)$ .

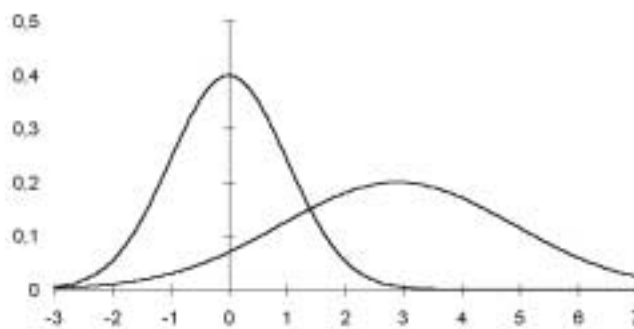


Figure 6 : exemple de lois normales

### Résumé et précisions (théorème central limite)

Si  $n$  est suffisamment grand,  $X$  ayant pour moyenne « vraie »  $\mu$ , pour variance « vraie »  $\sigma^2$ , alors :

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

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

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

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

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

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

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

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

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

### 9.4.1 Définition de l'intervalle de pari (I. P.) d'une moyenne observée

On considère une variable aléatoire de moyenne « vraie »  $\mu$  et de variance « vraie »  $\sigma^2$ .

On sait que pour  $n$  grand ( $n \geq 30$ , ou  $n\Pi$  et  $n(1 - \Pi) \geq 5$ ) :

la variable  $Z = \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$  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 observée 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 proba-

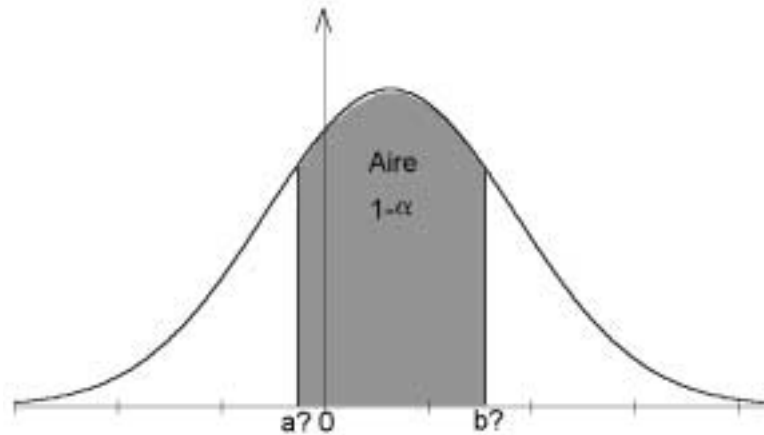
bilité fixée la valeur observée que l'on va obtenir.

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

**Exemple :**  $Pr(a \leq M_n \leq b) = 0,95$

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

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  $\mu$ .

Résolution :  $a = \mu - \lambda \frac{\sigma}{\sqrt{n}}$  et  $b = \mu + \lambda \frac{\sigma}{\sqrt{n}}$

La valeur  $\lambda$  inconnue doit vérifier :

$$Pr\left(\mu - \lambda \frac{\sigma}{\sqrt{n}} \leq M_n \leq \mu + \lambda \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$Pr\left(-\lambda \frac{\sigma}{\sqrt{n}} \leq M_n - \mu \leq \lambda \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$Pr\left(-\lambda \leq \frac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \lambda\right) = 1 - \alpha$$

Si le théorème central limite s'applique, l'expression ci-dessus suit une loi  $N(0, 1)$  ; notons-la  $Z$ . Alors  $\lambda$  doit vérifier  $Pr(-\lambda \leq Z \leq \lambda) = 1 - \alpha$ . C'est le  $z_\alpha$  de la table.

Finalement :  $\lambda = z_\alpha$

$$Pr\left(\mu - z_\alpha \frac{\sigma}{\sqrt{n}} \leq M_n \leq \mu + z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \text{ et}$$

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

Intervalle de Pari (I. P.) de la moyenne observée d'une variable de moyenne « vraie »  $\mu$ , de variance « vraie »  $\sigma^2$  construite sur un échantillon de taille  $n$

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

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  $n\Pi, n(1 - \Pi) \geq 5$  pour les variables de Bernoulli.

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

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

L'interprétation de l'intervalle de pari est fondamentale. Si cet intervalle est bien calculé, on est quasi sûr, avec une probabilité  $1 - \alpha$  (ici 0,95), d'obtenir une valeur de la moyenne observée comprise dans cet intervalle. En pariant que la valeur va tomber dans cet intervalle, on se trompera (en moyenne) dans cinq pour cent des expériences.

**Exemple :**

On a des raisons de penser que la fréquence d'une maladie dans la population est  $\Pi = 0,2$ . L'intervalle de pari de la moyenne observée (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.

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

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

- la longueur dépend de  $\alpha$   
Si  $\alpha' < \alpha$ , la longueur de  $IP_{1-\alpha'}$  est supérieure à la longueur de  $IP_{1-\alpha}$ .

**Exemple**

$$\alpha = 0,05 \Rightarrow z_{0,05} = 1,96$$

$$\alpha = 0,01 \Rightarrow z_{0,01} = 2,57$$



- 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 ( $2^2$ ).

### 9.4.3 L'intervalle de pari d'une variable aléatoire

Ce que l'on a dit pour une moyenne observée peut s'envisager pour une variable  $X$  quelconque dont on connaît la distribution.

L'IP de niveau  $1 - \alpha$  est l'intervalle  $[a, b]$  tel que  $Pr(a \leq X \leq b) = 1 - \alpha$ .

Exemple :

$X \sim N(0, 1)$

$IP_{1-\alpha} = [-z_\alpha ; z_\alpha]$

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 arithmétique  $M_n$  d'une variable aléatoire  $X$ , moyenne calculée sur  $n$  unités statistiques :

moyenne « vraie » de  $M_n$  = moyenne « vraie » de  $X$

$$\text{variance « vraie » de } M_n = \frac{\text{variance « vraie » de } X}{n}$$

2. **Théorème central limite**

Si  $X$  a pour moyenne « vraie »  $\mu$ , pour variance « vraie »  $\sigma^2$ ,  $M_n$  est, lorsque  $n$  est suffisamment grand ( $n \geq 30$ , ou  $n\Pi$  et  $n(1 - \Pi) \geq 5$ ), à peu près distribuée comme une variable normale de moyenne « vraie »  $\mu$  et de variance « vraie »  $\sigma^2/n$ , ce que l'on écrit :

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

3. **Intervalle de pari (I. P.)**

Lorsque les conditions ci-dessus sont satisfaites, l'intervalle

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

a la propriété suivante :

$$Pr(M_n \in IP_{1-\alpha}) = 1 - \alpha$$

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

# Chapitre 10

## Estimation - Intervalle de confiance

### 10.1 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 notion d'**intervalle de confiance** ou d'**estimation par intervalle**.

## 10.2 Estimation ponctuelle

### 10.2.1 Définition

A partir des données observées, 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  $\theta$  la grandeur à estimer et  $T$  ou  $T(\theta)$  son estimateur.

### 10.2.2 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.

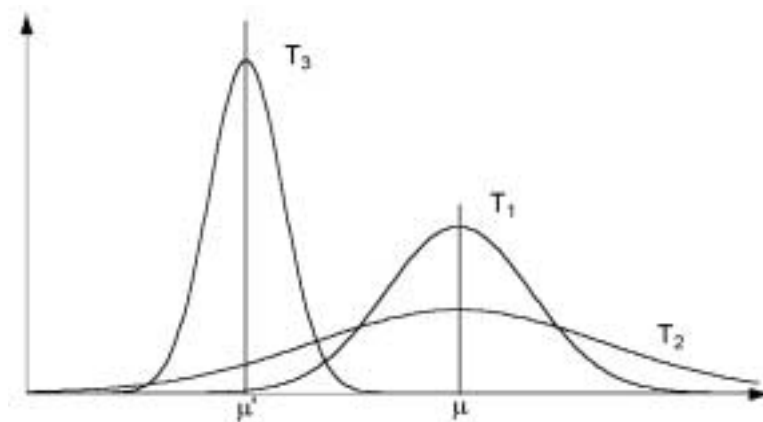


Figure 8 : densité de probabilité de 3 estimateurs  $T_1$ ,  $T_2$  et  $T_3$

La figure 8 représente les densités de probabilité de 3 estimateurs  $T_1$ ,  $T_2$  et  $T_3$  d'une moyenne  $\mu$ .

#### 10.2.2.1 Biais

On voit sur la figure 8 que  $T_1$  et  $T_2$  sont centrés autour de  $\mu$ , tandis que  $T_3$  a pour moyenne  $\mu'$  inférieure à  $\mu$ . 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  $\mu$ , et que  $T_3$  est un estimateur biaisé de  $\mu$ .

### 10.2.2.2 Variance

La variance d'un estimateur est définie de la manière usuelle :

$$\text{var}(T) = E[T - E(T)]^2$$

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  $\text{var}(T_1) < \text{var}(T_2)$ . On peut donc conclure que  $T_1$  est un meilleur estimateur de  $\mu$  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.

### 10.2.2.3 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 :

$$EQM(T) = \text{var}(T) + [E(T) - \theta]^2 = \text{var}(T) + B(T)^2$$

En particulier, l'erreur quadratique moyenne des estimateurs sans biais est égale à leur variance.

## 10.2.3 Exemple

On a souvent utilisé, dans ce cours, les quantités  $m$ , moyenne observée, et  $s^2$ , variance observée. La variable aléatoire moyenne arithmétique, notée  $M_n$ , a été étudiée au chapitre 8. De la même manière, on peut considérer la variable aléatoire variance  $S_n^2$ , définie par :

$$S_n^2 = \frac{n}{n-1} [M_{2,n} - M_n^2]$$

où  $M_{2,n}$  est la variable aléatoire « moyenne arithmétique de  $X^2$  ».

On va calculer  $E(S_n^2)$ . On rappelle que si  $U$  est une variable aléatoire, la variable moyenne arithmétique définie sur  $U$  a les propriétés suivantes :

$$E(M_{U,n}) = E(U) \quad (1) \quad \text{et} \quad \text{var}(M_{U,n}) = \frac{1}{n} \text{var}(U) \quad (2)$$

On a par ailleurs :

$$\text{var}(U) = E(U^2) - [E(U)]^2 \text{ et donc } E(U^2) = \text{var}(U) + [E(U)]^2 \quad (3).$$

On peut maintenant calculer  $E(S_n^2)$ . Soit  $X$  une variable aléatoire d'espérance  $E(X) = \mu$  et de variance  $\text{var}(X) = \sigma^2$ . On a :

$$E(S_n^2) = \frac{n}{n-1} [E(M_{2,n}) - E(M_n^2)]$$

Mais  $E(M_{2,n}) = E(X^2) = \sigma^2 + \mu^2$  d'après (1) et (3),

$$\text{et } E(M_n^2) = \text{var}(M_n) + [E(M_n)]^2 = \frac{\sigma^2}{n} + \mu^2 \text{ d'après (3), (2) et (1),}$$

$$\text{et finalement : } E(S_n^2) = \frac{n}{n-1} \left[ \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 \right] = \sigma^2.$$

$S_n^2$  est donc un estimateur sans biais de  $\sigma^2$ .

## 10.3 Estimation par intervalle - 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 (et proportions) « vraies ».

### 10.3.1 Exemple d'une proportion

L'idée directrice est la suivante : on souhaite associer à une valeur observée  $p$  un intervalle appelé INTERVALLE DE CONFIANCE qui ait « de bonnes chances » de contenir la valeur « vraie »  $\Pi$  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  $p$  - on construit autant d'intervalles de confiance. On voudrait qu'un grand nombre de ces intervalles contienne la valeur « vraie »  $\Pi$ . Par exemple que 95 % des intervalles en gros contiennent  $\Pi$ . 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  $1-\alpha$ . La valeur  $\alpha$  sera alors le risque - ou la probabilité - pour qu'un intervalle de confiance ne contienne pas la proportion « vraie »  $\Pi$ .

DE FACON GENERALE, L'INTERVALLE DE CONFIANCE AU RISQUE  $\alpha$  D'UNE VALEUR QUE L'ON CHERCHE A ESTIMER EST UN INTERVALLE QUI CONTIENT AVEC UNE PROBABILITE  $1 - \alpha$  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.

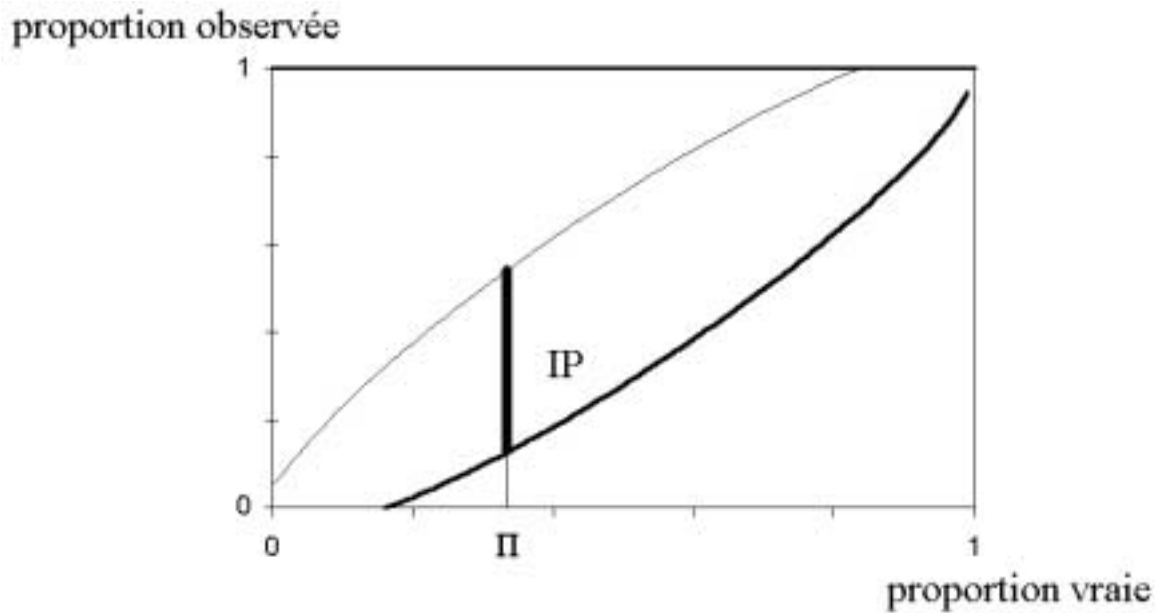


Figure 9

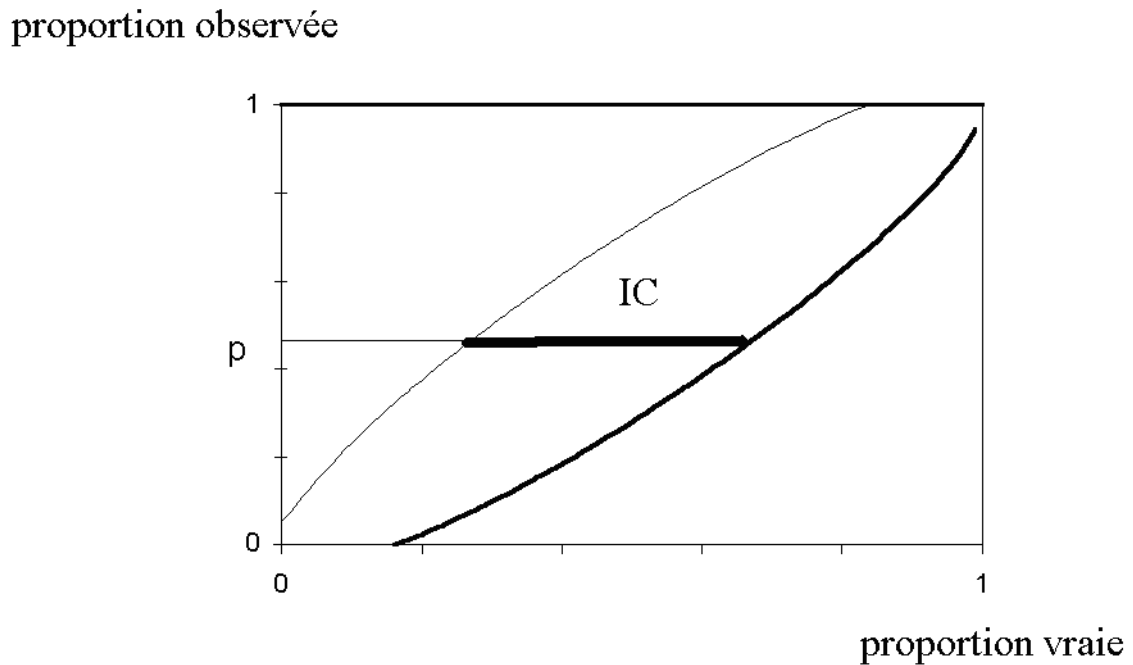


Figure 10

Considérons la figure 9. 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 observée 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 »,  $\Pi$ . Supposons que nous fassions un ensemble d'expériences, chaque expérience portant sur  $n$  individus étant productive d'une valeur de proportion observée  $p$ . On peut associer à chacune de ces expériences un point de coordonnées  $(\Pi, p)$  sur la figure 9. 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  $\Pi$ .

Maintenant supposons qu'une expérience unique ait été réalisée, produisant une valeur de proportion,  $p$ . 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 10. 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  $p$  un intervalle qui contiendra  $\Pi$  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.

### 10.3.2 Intervalle de confiance approché d'une proportion « vraie »

On montre qu'une bonne approximation de l'intervalle de confiance de niveau  $1 - \alpha$  de  $\Pi$ , fondé sur la valeur observée  $p$ ,  $p$  étant calculée sur  $n$  individus, est donnée par l'intervalle ci-dessous :

$$IC_{1-\alpha} = \left[ p - z_{\alpha} \sqrt{\frac{p(1-p)}{n}} ; p + z_{\alpha} \sqrt{\frac{p(1-p)}{n}} \right]$$

Notons  $\Pi_{\min}$  et  $\Pi_{\max}$  les bornes de cet intervalle.

Cette approximation n'est jugée satisfaisante que sous les CONDITIONS DE VALIDITE suivantes :  $n\Pi_{\min} \geq 5$ ,  $n(1-\Pi_{\max}) \geq 5$

LORSQUE LES CONDITIONS DE VALIDITE NE SONT PAS REMPLIES, IL FAUT AVOIR RECOURS A DES TABLES (hors programme).

**Exemple :**  $n = 100$ ,  $\alpha = 0,05$ ,  $p = 0,12$

$$IC_{0,95} = \left[ 0,12 - 1,96 \sqrt{\frac{0,12 \times 0,88}{100}} ; 0,12 + 1,96 \sqrt{\frac{0,12 \times 0,88}{100}} \right] = [0,06 ; 0,18]$$

conditions de validité

$$100 \times 0,06 = 6 \geq 5.$$

$$100 \times (1 - 0,18) = 82 \geq 5.$$



### 10.3.3 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 »  $\mu$ , intervalle fondé sur la valeur observée  $m$  obtenue après une expérience portant sur  $n$  individus. Le calcul de cet intervalle suppose en outre le calcul de la variance observée  $s^2$ . L'expression est la suivante :

$$IC_{1-\alpha} = \left[ m - z_{\alpha} \frac{s}{\sqrt{n}} ; m + z_{\alpha} \frac{s}{\sqrt{n}} \right]$$

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 :

$$IC_{1-\alpha} = \left[ m - t_{\alpha} \frac{s}{\sqrt{n}} ; m + t_{\alpha} \frac{s}{\sqrt{n}} \right]$$

où  $t_{\alpha}$  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,  $\sigma^2$ , est connue, l'intervalle de confiance a la forme suivante :

$$IC_{1-\alpha} = \left[ m - z_{\alpha} \frac{\sigma}{\sqrt{n}} ; m + z_{\alpha} \frac{\sigma}{\sqrt{n}} \right]$$

### 10.3.4 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  $p$  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.

### 10.3.4.1 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 :

$$2 \times 1,96 \sqrt{\frac{p(1-p)}{n}} \text{ qu'on arrondit ici à } 4 \sqrt{\frac{p(1-p)}{n}}$$

L'effectif de l'échantillon devra donc être au moins  $10000 p(1-p)$ .

Toutefois, cet effectif dépend de  $p$ , 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 :

$$4 \frac{p(1-p)}{i^2} \text{ au risque } 0,05 \text{ (ou } z_{\alpha}^2 \frac{p(1-p)}{i^2} \text{ au risque } \alpha)$$

#### 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.

### 10.3.4.2 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 nombre d'individus au moins égal à :

$$n = z_{\alpha/2}^2 \frac{s^2}{i^2}$$

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  $\sigma^2$ .

**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,  $\sigma = 10 \text{ mg.l}^{-1}$ . Supposons en outre que les résultats expérimentaux soient distribués normalement.

Si on effectue un dosage donnant  $90 \text{ mg.l}^{-1}$ , on a pour intervalle de confiance approché ( $\sigma$  étant connu) :

$$IC_{0,95} = [90 - 2\sigma ; 90 + 2\sigma] = [70 ; 110] \text{ soit un intervalle de longueur 40.}$$

Si on effectue deux dosages donnant  $90$  et  $96 \text{ mg.l}^{-1}$ , 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  $93 \text{ mg.l}^{-1}$  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.



# Chapitre 11

## 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 15.

### 11.1 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,  $\phi = 20\%$ . 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  $p$  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 observé (moyenne observée) 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  $\phi_1$ , 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 observé 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  $\phi = 0,2$  (fréquence supposé « vraie »), 95 % des valeurs des moyennes observées calculées sur 100 individus appartiendront à l'intervalle  $[0,12 ; 0,28]$ .

On adopte alors la stratégie suivante : si la valeur observée 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 observée 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 observée 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  $p$  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  $p \notin$  ou  $\in IP_{1-\alpha}$  c'est-à-dire :

$$p \notin \text{ ou } \in \left[ \phi_0 - z_{\alpha} \sqrt{\frac{\phi_0(1-\phi_0)}{n}} ; \phi_0 + z_{\alpha} \sqrt{\frac{\phi_0(1-\phi_0)}{n}} \right]$$

où  $\phi_0$  est la proportion hypothétique (0,2 dans l'exemple) et  $z_{\alpha}$  la borne de l'intervalle de pari au risque  $\alpha$  de  $p$ .

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

que

$$p - \varphi_0 \notin \text{ou} \in \left[ -z_\alpha \sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}} ; z_\alpha \sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}} \right]$$

ou encore selon que

$$\frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}} \notin \text{ou} \in [-z_\alpha ; z_\alpha]$$

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 arithmétique selon :

$$Z = \frac{P_n - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}}$$

avec  $\varphi_0 = 0,2$  (transcription de l'hypothèse).

Une telle variable aléatoire s'appelle usuellement « paramètre » du test et est notée conventionnellement  $Z$ . Ici on sait que  $Z \sim N(0, 1)$  et l'on construit l'intervalle de pari de niveau  $1 - \alpha$  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  $p$ , valeur observée de  $P_n$ , donc une valeur observée de  $Z$ , notée  $z$  :

$$z = \frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}}$$

On pourrait alors s'exprimer comme ceci (une terminologie plus précise sera indiquée plus loin) :

- si  $z \in IP_{0,95}$  on ne peut dire que le traitement est actif
- si  $z \notin IP_{0,95}$  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  $\alpha$  par exemple) avant de donner d'autres exemples de tests usuels

## 11.2 Principe général des tests d'hypothèses

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

### 11.2.1 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.

Le plus souvent, une des hypothèses est 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  $\varphi = 0,2$  (plus généralement  $\varphi = \varphi_0$ ).

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  $\varphi \neq 0,2$  (généralement  $\varphi \neq \varphi_0$ ).

**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 connaît alors la distribution. En d'autres termes, on cherche à construire une fonction des données à venir dont on connaît la distribution si  $H_0$  est vraie. Soit  $Z$  cette variable aléatoire.

**Exemple** : 
$$Z = \frac{P_n - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}} \sim N(0, 1)$$



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  $1 - \alpha$ , 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  $\alpha$** .

### 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  $z$ .

**Exemple :**  $z = \frac{p_{\text{réellement observé}} - \Phi_0}{\sqrt{\frac{\Phi_0(1 - \Phi_0)}{n}}}$

Alors décider que :

- si  $z$  appartient à la région critique, remettre en cause  $H_0$ , la **rejeter**, et conclure  **$H_1$  est vraie**, ou dire : « au risque  $\alpha$ ,  $H_0$  est rejetée ».
- si  $z$  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  $x_1, \dots, x_n$  ; calculer  $z$  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 15.

**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.

## 11.2.2 Justification de la règle de décision. Choix de $\alpha$

### 11.2.2.1 Interprétation de $\alpha$

On a déjà vu une interprétation de  $\alpha$  avec l'exemple des souris. De façon générale,  $\alpha$  est la probabilité pour que la valeur observée - ou calculée -  $z$  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  $\alpha$  est petit car alors on dit : voilà un événement qui avait  $100 \times \alpha$  % 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  $\Rightarrow \alpha$  doit être petit. Une autre interprétation de  $\alpha$  montre encore mieux que  $\alpha$  doit être petit. A nouveau, lorsque  $H_0$  est vraie, la probabilité d'obtenir un résultat  $z$  dans la région critique est  $\alpha$ . Mais alors on dit «  $H_1$  est vraie ». Donc

$\Rightarrow \alpha$  = « probabilité » de conclure  $H_1$  alors que  $H_0$  est vraie

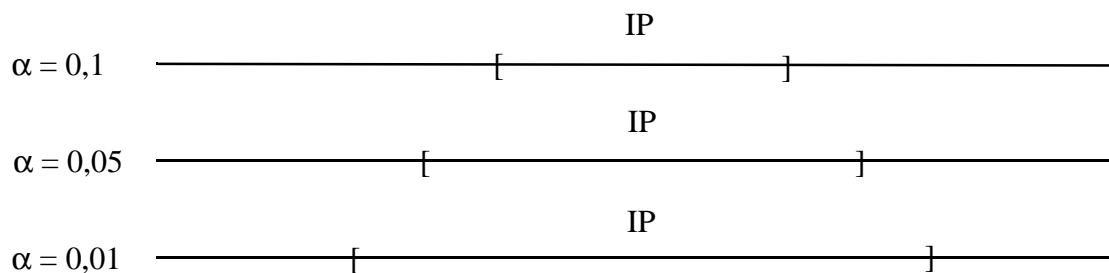
C'est un risque d'erreur qu'il convient de situer dans des valeurs acceptables (petites).

Cette valeur  $\alpha$  s'appelle **RISQUE DE PREMIERE ESPECE**.

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

### 11.2.2.2 Effet d'un changement de valeur de $\alpha$

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



Donc, toutes choses égales par ailleurs, la région critique diminue lorsque  $\alpha$  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  $\alpha = 0$ , on ne conclut jamais,  $H_0$  n'est jamais rejetée.

**Prendre une décision, c'est accepter un risque.**

Pour finir avec ce problème de  $\alpha$  il faut retenir :

- La valeur de  $\alpha$  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.

### 11.2.3 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 :

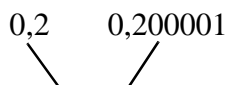
$$H_0 : \varphi = 0,2$$

$$H_1 : \varphi \neq 0,2$$

$$H_0 : \varphi = 0,200001$$

$$H_1 : \varphi \neq 0,200001$$

Les paramètres calculés, soit

$$z = \frac{p_{\text{observée}} - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}}$$


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

Considérons alors une expérience au cours de laquelle  $z \in \text{IP}_{0,95}$  pour les deux valeurs calculées. Peut-on conclure à la fois  $\varphi = 0,2$  et  $\varphi = 0,200001$  ? 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  $\varphi = \text{quelque chose}$ .

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,  $\Rightarrow H_0 : \varphi = 0,2$ . 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.

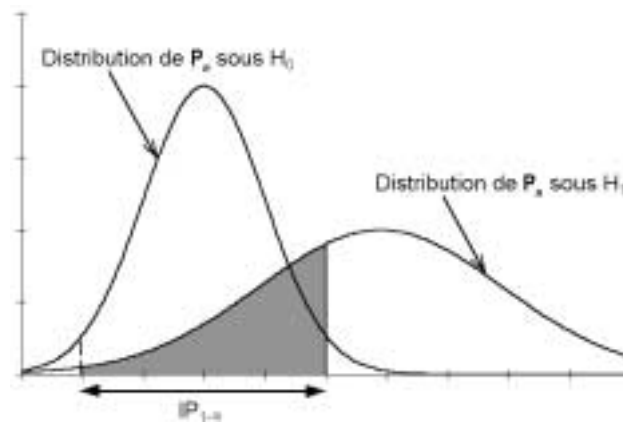
C'est le cas, en effet. L'aptitude d'un test à rejeter l'hypothèse nulle alors qu'elle est fausse est limitée. Précisément :

On appelle **PUISSANCE D'UN TEST  $P$**  la probabilité de rejeter l'hypothèse nulle, face à une hypothèse alternative, 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  $\beta$  :  $\beta = 1 - P$ .

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  $\varphi = 0,3$ , l'hypothèse  $H_0$  restant inchangée, c'est-à-dire  $\varphi = 0,2$ . 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 :

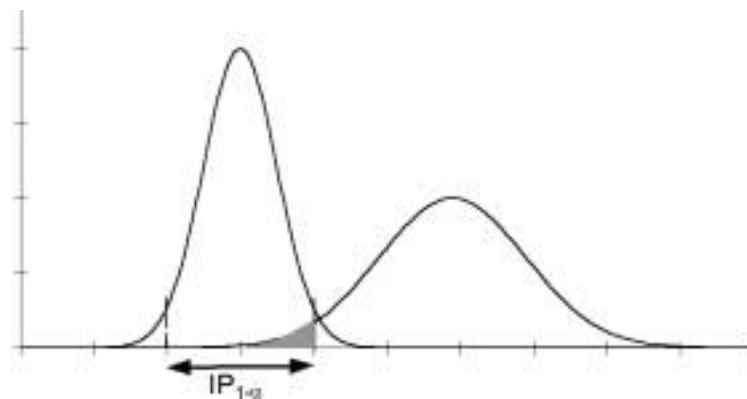
- sous l'hypothèse nulle ( $\varphi = 0,2$ ) :  $P_n \sim N\left(0,2, \frac{0,2(1-0,2)}{n}\right)$
- sous l'hypothèse alternative ( $\varphi = 0,3$ ) :  $P_n \sim N\left(0,3, \frac{0,3(1-0,3)}{n}\right)$



**Figure 11 : risque de deuxième espèce d'un test**

La figure 11 présente les deux distributions correspondantes, pour une certaine valeur de  $n$ . Supposons alors juste l'hypothèse  $H_1$  ; la valeur observée  $p$  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 observée appartienne à l'intervalle de pari associé au test, sachant que cette valeur observée est gouvernée par la distribution associée à  $H_1$ . Ainsi la valeur de cette aire grisée exprime le risque de deuxième espèce  $\beta$ , 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  $\varphi = 0,2$  et  $\varphi = 0,3$ , on aura « 6 chances sur dix » seulement de rejeter l'hypothèse  $\varphi = 0,2$  lorsque  $\varphi$  sera égal à 0,3. Autrement dit, 4 fois sur dix, on sera incapable de détecter que  $\varphi$  vaut 0,3 et non 0,2.



**Figure 12 : 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  $\varphi = 0,2$ ,  $\varphi = 0,4$  sont plus contrastées que les hypothèses  $\varphi = 0,2$ ,  $\varphi = 0,3$ ), plus les distributions de  $P_n$  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 12 reproduit les conditions de la figure 11, 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  $\phi = 0,2$  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.

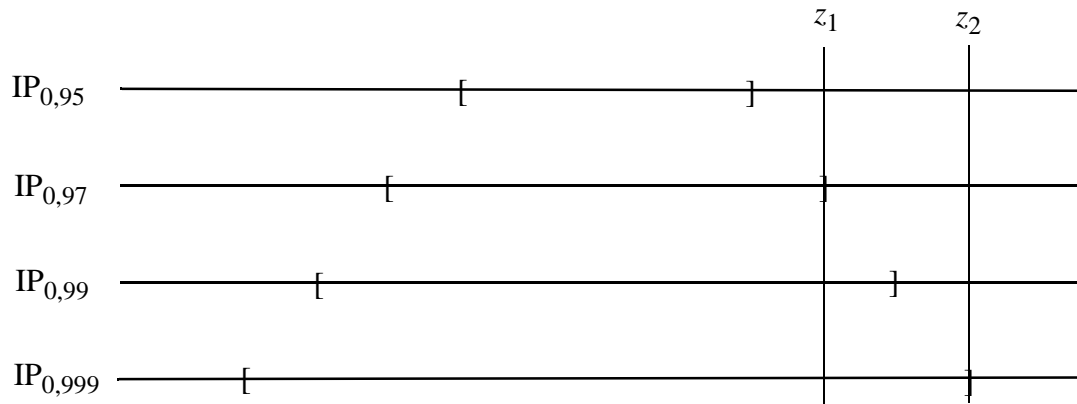
## 11.2.4 Amélioration de l'interprétation du rejet de $H_0$

### 11.2.4.1 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  $z_1$  et  $z_2$  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  $z_1$  est à l'extérieur des intervalles de pari jusqu'au niveau 0,97, que  $z_2$  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  $z$ .

#### Calcul pratique du degré de signification

On cherche dans la table la valeur de  $p$  pour laquelle  $z_p = z$ ,  $z_p$  étant du type  $z_\alpha$

Exemple:  $z = 2,43$ .

On trouve dans la table  $z_{0,02} = 2,32$  et  $z_{0,01} = 2,57$

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  $z$  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  $p$ .

#### 11.2.4.2 Orientation du rejet

Le rejet de  $H_0$  correspond généralement à l'une des deux situations :

- rejet car  $z$  est trop petit (inférieur à la borne inférieure de l'intervalle de pari)
- rejet car  $z$  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 :

$z$  est trop petit  $\Leftrightarrow$  le traitement est efficace

$z$  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  $\alpha$  ;  $\alpha = 0,05$
4. Mettre en place la règle de décision sur la base d'un intervalle de pari au risque  $\alpha$ .
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 15).

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,  $\alpha$ , chiffrant la probabilité de rejeter  $H_0$  alors qu'elle est vraie,
- le risque de deuxième espèce,  $\beta$ , 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 12

## Quelques tests usuels

### 12.1 Tests concernant des variables de Bernoulli

#### 12.1.1 Test d'égalité d'une proportion « vraie » à une valeur donnée (ou test de comparaison d'une proportion observée à une valeur donnée)

##### 12.1.1.1 Mise en place du test

**Exemple** : les souris du chapitre précédent

##### 1. Les hypothèses en présence

**H<sub>0</sub>** (hypothèse nulle) : la proportion « vraie » (dans la population des souris traitées) est égale à  $\varphi_0$  (proportion hypothétique ou supposée qu'on se donne pour le test).

**H<sub>1</sub>** (hypothèse alternative) : la proportion « vraie » est différente de  $\varphi_0$ .

Notations :

$H_0 : \varphi = \varphi_0$

$H_1 : \varphi \neq \varphi_0$

##### 2. Définition du paramètre

$$Z = \frac{P_n - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}}$$

où  $P_n$  représente la variable aléatoire proportion.

Sous  $H_0$ ,  $Z$  est à peu près distribuée selon  $N(0, 1)$

[conditions de validité :  $n\varphi_0 \geq 5$  et  $n(1 - \varphi_0) \geq 5$ ]

### 3. Choix d'un seuil de signification $\alpha$

Construction de l'intervalle de pari de niveau  $1 - \alpha$  :  $IP_{1-\alpha}$

**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 :

$$z = \frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}}$$

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

Si  $z \in IP_{1-\alpha}$  on ne rejette pas  $H_0$  ou « on ne conclut pas ».

### 5. Recueil des données. Conclusion

Rappelons les conditions de validité :  $n\varphi_0 \geq 5$  et  $n(1 - \varphi_0) \geq 5$

#### 12.1.1.2 Autre interprétation du paramètre $z$

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

$$|p - \varphi_0| > z_\alpha \sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}$$

c'est-à-dire si la proportion observée  $p$  est suffisamment différente de  $\varphi_0$ . Voilà pourquoi on dit que l'on compare  $p$  et  $\varphi_0$ . C'est pourquoi on dit aussi, lorsque  $H_0$  est rejetée :

La proportion observée est **significativement** différente de la valeur donnée, au risque  $\alpha$  (0,05), ou encore : la différence entre  $p$  et  $\varphi_0$  est significative. Ce qui indique une différence entre la valeur donnée et la proportion « vraie »  $\varphi$ .

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  $|p - \varphi_0|$  peut être ou non significative selon la valeur de  $n$ . Si l'on vous demande :  $p = 0,25$  et  $0,2$ , sont-elles significativement différentes, ne répondez pas ; demandez : quelle est la taille de l'échantillon sur lequel  $p$  a été calculé, à quel risque ?

## 12.1.2 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,  $\varphi_A$  et  $\varphi_B$  les fréquences « vraies » de souris cancéreuses dans ces deux populations,  $n_A$  et  $n_B$  les tailles des échantillons sur lesquels on calculera  $p_A$  et  $p_B$ , les fréquences observées correspondantes. Mettons en place le test.

### 1. Les hypothèses en présence

$H_0$  hypothèse nulle : les fréquences « vraies » coïncident  $\varphi_A = \varphi_B$

$H_1$  hypothèse alternative : les fréquences « vraies » sont différentes  $\varphi_A \neq \varphi_B$

### 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 :

$H_0 : \varphi = \varphi_0$

$H_1 : \varphi \neq \varphi_0$

On y parvient en reformulant les hypothèses

$H_0 : \varphi_A - \varphi_B = 0$

$H_1 : \varphi_A - \varphi_B \neq 0$

Il s'agit donc de comparer à 0 la différence  $\varphi_A - \varphi_B$ .

Auparavant on formait le paramètre

$$\frac{P_n - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}}$$

qui peut s'interpréter comme  $\frac{\text{v.a. proportion} - \text{valeur théorique}}{\text{écart-type de la v.a. proportion}}$

Alors on va former  $\frac{\text{différence des v.a. proportions} - \text{valeur théorique}}{\text{écart-type des différences des v.a. proportions}}$

soit  $\frac{P_{nA} - P_{nB}}{\text{écart-type des différences des v.a. proportions}}$

La difficulté est de former l'expression de l'écart type des différences des % expérimentaux. Remarquons d'abord que les variables aléatoires  $P_{nA}$  et  $P_{nB}$  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 can-

céreuse ou non dans la population non traitée.

Alors :  $\text{var}(P_{nA} - P_{nB}) = \text{var}(P_{nA}) + \text{var}(-P_{nB}) = \text{var}(P_{nA}) + \text{var}(P_{nB})$  (voir chapitre 6)

Par ailleurs, sous l'hypothèse nulle, les moyennes « vraies » de  $P_{nA}$  et  $P_{nB}$  coïncident avec une valeur  $\Pi$  - inconnue. D'où :

$$\text{var}(P_{nA}) = \frac{\Pi(1-\Pi)}{n_A} \text{ et } \text{var}(P_{nB}) = \frac{\Pi(1-\Pi)}{n_B}$$

si  $n_A$  et  $n_B$  sont les tailles des échantillons sur lesquels  $P_{nA}$  et  $P_{nB}$  sont calculées.

$$\text{Donc : } \text{var}(P_{nA} - P_{nB}) = \frac{\Pi(1-\Pi)}{n_A} + \frac{\Pi(1-\Pi)}{n_B}$$

Maintenant,  $\Pi$  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 :

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

$$\text{soit : } \hat{\Pi} = \frac{n_A p_A + n_B p_B}{n_A + n_B}$$

Finalement on adopte le paramètre suivant :

$$Z = \frac{P_{nA} - P_{nB}}{\sqrt{\frac{\hat{\Pi}(1-\hat{\Pi})}{n_A} + \frac{\hat{\Pi}(1-\hat{\Pi})}{n_B}}}$$

$$\text{avec } \hat{\Pi} = \frac{n_A p_A + n_B p_B}{n_A + n_B}$$

Sous l'hypothèse nulle  $Z$  est à peu près distribuée selon  $N(0, 1)$ .

Conditions de validité :

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

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

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  $z$ , dont on connaît 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  $z = \frac{P_A - P_B}{\sqrt{\frac{\hat{\Pi}(1-\hat{\Pi})}{n_A} + \frac{\hat{\Pi}(1-\hat{\Pi})}{n_B}}}$ , conclusion.

## 12.2 Tests concernant des variables quantitatives

### 12.2.1 Tests de comparaison à une valeur donnée

Ces tests concernent les variables quantitatives continues et permettent de traiter les types de questions suivantes :

1. 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.
2. la distribution de la taille des individus dans cette sous population est-elle dissymétrique par rapport à cette moyenne « vraie », c'est-à-dire témoigne-t-elle d'une inégalité de fréquences entre les « petites » tailles et les « grandes tailles », ce qui est le cas par exemple si la fréquence des « 20-25 cms de moins que la moyenne » est différente de celle des « 20-25 cms de plus que la moyenne » ?

Ces deux tests sont apparentés dans la mesure où le premier met à l'épreuve  $E(X) = \mu_0$ , l'autre le fait que  $X - \mu_0$  et  $\mu_0 - X$  ont la même densité de probabilité. Cette dernière condition, qui entraîne alors  $E(X) - \mu_0 = \mu_0 - E(X)$  et donc  $E(X) = \mu_0$ , étant plus contraignante que la première.

#### 12.2.1.1 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 et **n'est valide que lorsque  $n \geq 30$** .

1. Les hypothèses en présence :

$H_0$  : la moyenne « vraie » coïncide avec la valeur donnée  $\mu_0$  :  $\mu = \mu_0$

$H_1$  :  $\mu \neq \mu_0$

2. Construction du paramètre

$$Z = \frac{M_n - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

$Z$  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  $\sigma^2$ . 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  $IP_{1-\alpha}$

$$IP_{1-\alpha} = [-z_{\alpha} ; z_{\alpha}] ; z_{0,05} = 1,96$$

#### 4. Définition de la règle de décision

La règle de décision est tout à fait similaire au cas des proportions.

Si  $z \notin IP_{1-\alpha}$ , rejet de  $H_0$ . On dit alors : au risque  $\alpha$  la moyenne « vraie » diffère de la valeur donnée ou, pour les mêmes raisons que pour les proportions : la moyenne observée est significativement différente, au risque  $\alpha$ , de la valeur donnée ; ou encore : la moyenne observée et la valeur donnée sont significativement différentes, au risque  $\alpha$ .

Si  $z \in IP_{1-\alpha}$ , on ne conclut pas. La moyenne observée n'est pas significativement différente de la valeur donnée.

5. Recueil des données. Calcul de  $z = \frac{m - \mu_0}{\sqrt{\frac{s^2}{n}}}$ . Conclusion.

### 12.2.1.2 Test de symétrie d'une variable (X) par rapport à une valeur donnée ( $\mu_0$ ) : test de Wilcoxon

#### 1. Les hypothèses en présence :

$H_0$  : les variables  $X - \mu_0$  et  $\mu_0 - X$  ont même densité de probabilité

$H_1$  : les variables  $X - \mu_0$  et  $\mu_0 - X$  n'ont pas la même densité de probabilité

#### 2. Construction du paramètre

Le paramètre est construit à partir des valeurs ordonnées par ordre croissant des valeurs absolues des  $x_i - \mu_0$  où les  $x_i$  sont les valeurs de  $X$  observées dans l'échantillon ; à chaque valeur on associe son rang de classement et l'on garde la mémoire de son signe. On attribue aux éventuels ex-æquo un rang commun égal à la moyenne des rangs qu'ils occupent.

#### Exemple

Si les valeurs observées (qui ne seront disponibles qu'après réalisation de l'expérience) sont :

-2,3 ; 4 ; 1 ; 5,6 ; -1,2

Le classement sera : 1 (+) ; 1,2 (-) ; 2,3 (-) ; 4 (+) ; 5,6 (+)

On s'intéresse alors à la somme des rangs des places occupées par les valeurs positives, appelée  $T^+$ . Ici la valeur de  $T^+$  serait  $1+4+5 = 10$ .

Le paramètre du test est :

$$Z = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$

La variable  $Z$  a une distribution connue :

- Lorsque  $n > 15$  cette distribution est à peu près  $N(0, 1)$ .
- Pour  $n \leq 15$ , il s'agit d'une distribution faisant l'objet d'une table spécifique, la table du test de Wilcoxon.



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

Construction de l'intervalle de pari centré  $IP_{1-\alpha}$   
 $IP_{1-\alpha} = [-W_\alpha ; W_\alpha]$  ; lorsque  $n > 15$ ,  $W_\alpha = z_\alpha$

4. Définition de la règle de décision

Si  $z \notin IP_{1-\alpha}$ , rejet de  $H_0$ . On dit alors : au risque  $\alpha$  la densité de probabilité de  $X$  n'est pas symétrique par rapport à  $\mu_0$  ; selon le signe de  $z$ , on conclura que  $X$  est « plutôt plus grand que  $\mu_0$  », ou que  $X$  est « plutôt plus petit que  $\mu_0$  ».

Si  $z \in IP_{1-\alpha}$ , on ne conclut pas ; on ne rejette pas  $H_0$ .

5. Recueil des données, calcul de  $z$ , conclusion.

Remarque : si  $n < 6$  ce test ne permet jamais de rejeter  $H_0$

## 12.2.2 Tests de comparaison de variables quantitatives

Ces tests concernent les variables quantitatives continues et permettent de traiter les types de questions suivantes :

1. 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.
2. la distribution de la variable aléatoire taille des individus dans la population  $A$  coïncide-t-elle avec la distribution de la variable aléatoire taille des individus dans la population  $B$ .

Ces deux tests sont apparentés, l'hypothèse d'égalité des distributions étant plus contraignante que l'hypothèse d'égalité des moyennes « vraies » seules. Dans les deux cas 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  $n_B$ ).

### 12.2.2.1 Test d'égalité de deux moyennes « vraies » (ou test de comparaison de deux moyennes observées)

**Ce test n'est valide que lorsque  $n_A$  et  $n_B$  sont  $\geq 30$** , cas dit des grands échantillons.

Il s'agit d'un problème très proche du problème traité en 12.2.1.1

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

$$Z = \frac{M_{nA} - M_{nB}}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

$Z$  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.

Calculs :

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

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

les  $x_{iA}$  et  $x_{iB}$  étant les valeurs de tailles observées dans les échantillons des populations  $A$  et  $B$  respectivement.

$$z = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$

Conclusion.

### 12.2.2.2 Test d'égalité de deux distributions (ou test de comparaison de deux distributions) : test de Mann-Whitney-Wilcoxon

1. Les hypothèses en présence

$H_0$  les densités de probabilité coïncident dans les deux populations :  $f_A = f_B$

$H_1$  les densités de probabilité ne coïncident pas :  $f_A \neq f_B$

2. Construction du paramètre : cette construction suit les mêmes lignes que celles du test de Wilcoxon décrit section 12.2.1.2.

Par convention, on considère que  $n_A \leq n_B$ .

On ordonne par valeurs croissantes l'ensemble des données observées (dont on disposera

après réalisation de l'expérience). On attribue aux éventuels ex-æquo un rang commun égal à la moyenne des rangs qu'ils occupent. Puis on calcule la somme des rangs de classement occupés par les données issues de l'échantillon de la population  $A$ , soit  $T_A$ .

On calcule également  $\delta = T_A - \frac{n_A(n_A + n_B + 1)}{2}$ .

Puis  $T'_A$  de la façon suivante :

- $T'_A = T_A + 0,5$  si  $\delta > 0$
- $T'_A = T_A - 0,5$  si  $\delta < 0$

### Exemple

Si les valeurs observées sont :

- Echantillon de population  $A$  : 1,7 ; 6,1 ; 3,2 ; 1,5
- Echantillon de population  $B$  : 4,3 ; 0,5 ; 1,1 ; 2,7 ; 5,4

Le classement conduit à 0,5 ( $B$ ) ; 1,1 ( $B$ ) ; 1,5 ( $A$ ) ; 1,7 ( $A$ ) ; 2,7 ( $B$ ) ; 3,2 ( $A$ ) ; 4,3 ( $B$ ) ; 5,4 ( $B$ ) ; 6,1 ( $A$ ) et à  $T_A = 3+4+6+9 = 22$ .

Enfin  $\delta = 22 - 4 \times 10 / 2 = 2$ . Donc  $T'_A = 22,5$ .

Le paramètre du test est :

- $Z = \frac{T_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B (n_A + n_B + 1)/12}}$  lorsque  $n_A$  et  $n_B \leq 10$
- $Z = \frac{T'_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B (n_A + n_B + 1)/12}}$  lorsque  $n_A$  ou  $n_B > 10$

$Z$  a une distribution connue :

- Lorsque  $n_A$  ou  $n_B > 10$  cette distribution est à peu près  $N(0,1)$ .
- Lorsque  $n_A$  et  $n_B \leq 10$ , il s'agit d'une distribution faisant l'objet d'une table spécifique, la table du test de Mann-Whitney-Wilcoxon.

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

Construction de l'intervalle de pari  $IP_{1-\alpha}$

Cet intervalle est du type  $IP_{1-\alpha} = [-M_\alpha ; M_\alpha]$

Exemple : si  $n_A = 3$  et  $n_B = 5$ , on a  $M_{0,05} = 2,20$

### 4. Règle de décision

Si  $z \notin IP_{1-\alpha}$ , rejet de  $H_0$ . On dit alors : au risque  $\alpha$  la densité de probabilité de la variable étudiée n'est pas la même dans les populations  $A$  et  $B$  ; selon le signe de  $z$ , on conclura que la variable est « plutôt plus grande dans  $A$  que dans  $B$  », ou que la variable est « plutôt plus petite dans  $A$  que dans  $B$  ».

Si  $z \in IP_{1-\alpha}$ , on ne conclut pas ; on ne rejette pas  $H_0$ .

5. Mise en œuvre de l'expérience ; calcul de  $z$  ; conclusion.

Remarque : si  $n_A < 3$  ou  $n_B < 4$ , ce test ne permet jamais de rejeter  $H_0$

### 12.2.3 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 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 observées. 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 observées 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 des paires de valeurs obtenues par unité statistique (différence des notes, différence des glycémies par individu). Cela revient au problème de la comparaison d'une moyenne (moyenne des différences) à zéro ou à la question de la symétrie d'une distribution (celle des différences) par rapport à zéro. On se ramène ainsi à des tests que l'on connaît (cf. section 12.2.1).

On note  $d$  la **variable aléatoire différence** entre résultats pour un même sujet.

#### 12.2.3.1 Test de comparaison de deux moyennes sur séries appariées

**Ce test n'est valide que si  $n \geq 30$**

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

$$Z = \frac{M_{nd}}{\sqrt{\frac{s^2}{n}}}$$

où  $s^2$  est la variance observée des différences, soit  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - m_d)^2$   
 $n$  est le nombre de paires  
 $M_{nd}$  est la variable aléatoire moyenne arithmétique des différences  
et  $m_d$  est la moyenne observée des différences.

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

Les étapes se succèdent alors de façon ordinaire :

choix de  $\alpha$ , construction de l'IP, définition de la règle de décision, calcul de  $z = \frac{m_d}{\sqrt{\frac{s^2}{n}}}$ , conclusion.

### 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.

### 12.2.3.2 Test de symétrie de la distribution des différences

Ce test est un cas particulier du test vu au paragraphe 12.2.1.2. car les hypothèses considérées dans ce cas sont les suivantes :

1. Hypothèses en présence  
 $H_0$  : La densité de probabilité de la variable aléatoire  $d$  est symétrique par rapport à zéro.  
 $H_1$  : La densité de probabilité de la variable  $d$  n'est pas symétrique par rapport à zéro ; il existe des domaines de valeurs de  $d$  plus probables que leur opposé (par exemple si le domaine  $[2,1 ; 2,4]$  est plus probable que le domaine  $[-2,4 ; -2,1]$ ).
2. Construction du paramètre  
Le paramètre se construit comme en 12.2.1.2 : on range dans l'ordre croissant de leurs valeurs et sans tenir compte de leur signe les  $n$  différences  $d_i$ .

La suite se déroule comme en 12.2.1.2.

# Résumé du chapitre

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

$$z = \frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1 - \varphi_0)}{n}}} ; \text{v.a.} \sim N(0, 1) ; \text{validité } n\varphi_0 \geq 5 \text{ et } n(1 - \varphi_0) \geq 5$$

2. Comparaison de deux proportions observées

$$z = \frac{p_A - p_B}{\sqrt{\frac{\hat{\Pi}(1 - \hat{\Pi})}{n_A} + \frac{\hat{\Pi}(1 - \hat{\Pi})}{n_B}}} ; \text{v.a.} \sim N(0, 1) ; \hat{\Pi} = \frac{n_A p_A + n_B p_B}{n_A + n_B}$$

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

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

$$z = \frac{m - \mu_0}{\sqrt{\frac{s^2}{n}}} ; \text{v.a.} \sim N(0, 1) ; \text{validité } n \geq 30$$

4. Test de symétrie d'une variable par rapport à une valeur donnée

Ordonner les valeurs absolues des écarts à la valeur donnée et calculer  $T^+$ , somme des rangs des écarts positifs.

$$z = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} ; \text{v.a.} \sim N(0, 1) \text{ si } n > 15 ; \text{v.a.} \sim \text{Wilcoxon sinon.}$$

5. Comparaison de deux moyennes observées

$$z = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} ; \text{v.a.} \sim N(0, 1) ; \text{validité } n_A \text{ et } n_B \geq 30$$

6. Test d'égalité de deux distributions (on suppose  $n_A \leq n_B$ )

Ordonner les valeurs.  $T_A$  = somme des rangs des données A.  $\delta = T_A - \frac{n_A(n_A + n_B + 1)}{2} \cdot T'_A$   
 $= T_A + 0,5$  si  $\delta > 0$ ,  $T'_A = T_A - 0,5$  sinon

$$z = \frac{T'_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B (n_A + n_B + 1)/12}} \sim N(0, 1) \text{ lorsque } n_A \text{ ou } n_B > 10$$

$$z = \frac{T_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B (n_A + n_B + 1)/12}} \sim \text{Mann-Whitney-Wilcoxon si } n_A \text{ et } n_B \leq 10$$

7. Comparaison de deux moyennes sur séries appariées

On utilise le test 3 en comparant la moyenne de la variable différence  $d$  à 0

8. Test de symétrie des différences (séries appariées)

On utilise le test 4 de symétrie de la variable  $d$  par rapport à 0.

# Chapitre 13

## 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 diffère-t-elle 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.

### 13.1 Comparaison d'une répartition observée à une répartition donnée ou test du $\chi^2$ d'ajustement

Supposons que l'on souhaite savoir si la répartition de la couleur des cheveux dans la population des habitants du département  $A$  diffère de 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  $\varphi_i$  la probabilité de survenue de l'événement « la  $i^{\text{ème}}$  modalité est observée ».

#### Exemple :

$\varphi_1$  = probabilité qu'un individu tiré au hasard dans le département  $A$  ait les cheveux blonds

$\varphi_2$  = probabilité qu'un individu tiré au hasard dans le département  $A$  ait les cheveux bruns

etc...

Notons par ailleurs  $\varphi_{hi}$  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).

### 13.1.1 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 :  $\varphi_i = \varphi_{hi}$  pour tous les  $i$  de 1 à  $k$ .

$H_1$  : hypothèse alternative :  $\varphi_i \neq \varphi_{hi}$  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

$H_0 : \varphi = \varphi_{h1}$  et  $1 - \varphi = \varphi_{h2} = 1 - \varphi_{h1}$

ce qui s'écrit avec les nouvelles notations :

$\varphi_1 = \varphi_{h1}$  et  $\varphi_2 = 1 - \varphi_{h1}$

Mais on n'avait retenu que la condition  $\varphi = \varphi_{h1}$  (en fait  $\varphi = \varphi_0$ ) car dans ce cas les deux conditions ci-dessus sont redondantes.

Le paramètre calculé retenu était :

$$z = \frac{p - \varphi_{h1}}{\sqrt{\frac{\varphi_{h1}(1 - \varphi_{h1})}{n}}}$$

Calculons son carré

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

$$z^2 = \frac{(np - n\varphi_{h1})^2}{n\varphi_{h1}} + \frac{(n(1 - p) - n(1 - \varphi_{h1}))^2}{n(1 - \varphi_{h1})} = \frac{(np - n\varphi_{h1})^2}{n\varphi_{h1}} + \frac{(n(1 - p) - n\varphi_{h2})^2}{n\varphi_{h2}}$$



Or  $np$  = 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  $\varphi_{h1}$ . On s'attend donc à observer  $n\varphi_{h1}$  individus présentant cette valeur. Ce nombre d'individus attendu s'appellera effectif attendu ou calculé de la première modalité et sera noté  $A_1$ .

De la même façon,  $n(1 - p)$  = 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  $\varphi_{h2} = 1 - \varphi_{h1}$ . On s'attend donc à observer  $n\varphi_{h2}$  individus présentant cette valeur. Ce nombre d'individus attendu s'appellera effectif attendu ou calculé de la seconde modalité et sera noté  $A_2$ .

$$D'où \chi^2 = \frac{(O_1 - A_1)^2}{A_1} + \frac{(O_2 - A_2)^2}{A_2}$$

où les  $O_i$  représentent les effectifs observés dans les différentes modalités, les  $A_i$  représentent les effectifs  $n\varphi_{hi}$  dits prévus ou calculés ou **ATTENDUS** 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 :

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

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  $A_i$  valent  $n\varphi_{hi}$ .

On remarque que  $Q$  chiffre l'écart entre ce qui est prévu par l'hypothèse  $H_0$  et ce qui est obtenu ; cet écart se fonde naturellement sur les différences  $O_i - n\varphi_{hi}$  car  $n\varphi_{hi}$  est le nombre attendu d'individus présentant la modalité  $i$ .

**Exemple** : si  $\varphi_{hi} = 0,4$ , 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 hypothétique.

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

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  $A_i$  DOIVENT ETRE AU MOINS EGAUX A 5

### 3. Intervalle de pari

$\alpha$  étant choisi (0,05), construction de l'intervalle de pari  $IP_{1-\alpha}$

La variable  $\chi^2$  a l'allure présentée figure 13. On remarque qu'il serait stupide de choisir l'in-

tervalle de pari centré dessiné sur cette figure car alors des valeurs numériques voisines de zéro pour la valeur  $Q_c$  du paramètre  $Q$  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 14) :

$$IP_{1-\alpha} = [0 ; K_{ddl,\alpha}]$$

C'est cette valeur, notée  $K_{ddl,\alpha}$  qui est lisible directement dans une table.

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

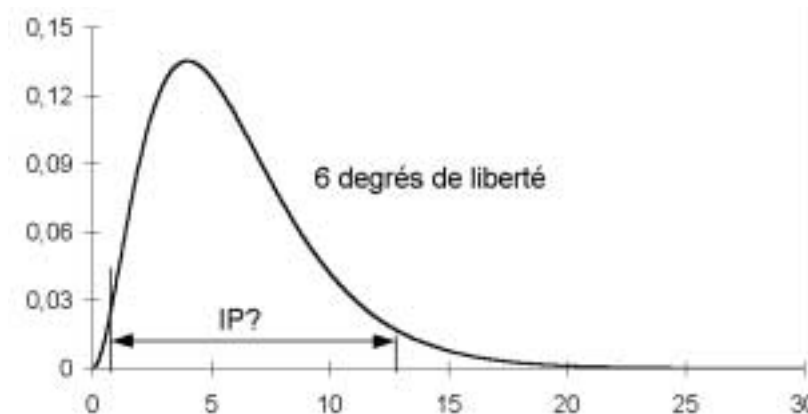


Figure 13 : distribution de  $\chi^2$

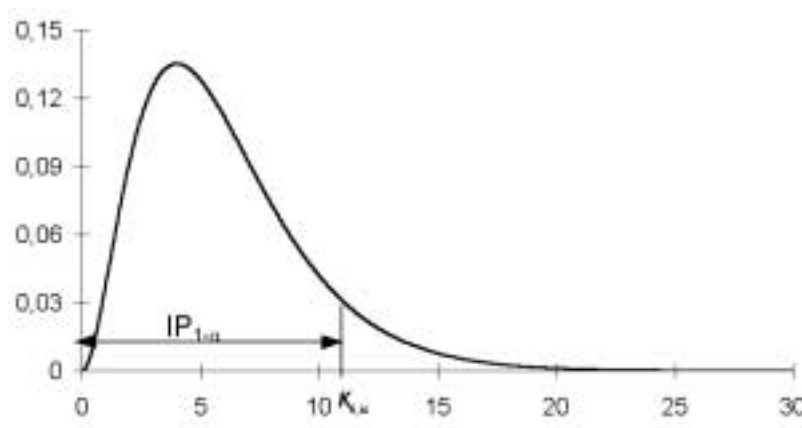


Figure 14 : distribution de  $\chi^2$

### Usage de la table

Cette table comporte - comme celle du  $t$  de Student - une entrée entière appelée nombre de degrés de liberté (ddl). On montre que pour le test envisagé ici

$\text{nombre de degrés de liberté} = \text{nombre de modalités} - 1$
---

**Exemple** :  $K_{5;0,05}$  (5 ddl, si 6 modalités) = 11,07

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

4. Règle de décision

Si  $Q_c \leq K_{ddl,\alpha}$  on ne conclut pas

Si  $Q_c > K_{ddl,\alpha}$   $H_0$  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  $\alpha$ .

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
<b>effectifs observés</b> ( $O_i$ )	25	9	3	37 ( $n$ )
<b>effectifs attendus</b> ( $A_i = n \varphi_{hi}$ )	14,8	11,1	11,1	37
<b>répartition donnée</b> ( $\varphi_{hi}$ )	0,4	0,3	0,3	1

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

On obtient ici :

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

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

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.

### 13.1.2 Cas particulier : variable à deux modalités

On a vu que le paramètre du test  $Q$  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 :  $Q = Z^2$ .

En outre, et sinon il y aurait incohérence, on peut vérifier l'égalité suivante :

$$K_{1;\alpha} = z_{\alpha}^2$$

**Exemple :** pour  $\alpha = 0,05$   $K_{1;0,05} = 3,84 = (1,96)^2$

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  $\chi^2$  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 11

Une race de souris présente un taux de cancers spontanés de 0,2. Sur 100 souris traitées on observe 34 cancers soit  $p = 0,34$ . La différence est elle significative ?

- test de comparaison :

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

- test du  $\chi^2$  :

	cancer	absence de cancer	
répartition théorique	0,2	0,8	
effectifs attendus	20	80	
effectifs observés	34	66	100 (effectif total)

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

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

## 13.2 Comparaison de deux répartitions observées ou test du $\chi^2$ 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 :

$H_0$  :  $\varphi_{1i} = \varphi_{2i}$  pour toutes les modalités  $i$ .

$H_1$  :  $\varphi_{1i} \neq \varphi_{2i}$  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.

- 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 4 : effectifs observés ( $O_{1i}$  et  $O_{2i}$ )**

	blonds	bruns	roux	nombre total
<b>échantillon 1</b>	25	9	3	$37 = n_1$
<b>échantillon 2</b>	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 5 : répartition de « référence »**

	blonds	bruns	roux	nombre total
<b>mélange</b>	38	26	13	77
<b>fréquences</b>	$38/77 = 0,49$	$26/77 = 0,34$	$13/77 = 0,17$	

Ces trois fréquences, 0,49, 0,34, 0,17, vont jouer maintenant le rôle des probabilités hypothétiques  $\phi_{hi}$  de la section 13.1. Pour la commodité de l'écriture, on les note respectivement  $p_1, p_2, p_3$ .

- iii. On forme le tableau des effectifs attendus.  
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 ATTENDUS**.

**Tableau 6 : effectifs attendus ( $A_{1i}$  et  $A_{2i}$ )**

	blonds	bruns	roux
<b>échantillon 1</b>	18,1 ( $n_1 p_1$ )	12,6 ( $n_1 p_2$ )	6,3 ( $n_1 p_3$ )
<b>échantillon 2</b>	19,6 ( $n_2 p_1$ )	13,6 ( $n_2 p_2$ )	6,8 ( $n_2 p_3$ )

- iv. On calcule finalement le paramètre du test

On montre que le paramètre adapté à ce test est :

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

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 :

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

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,  $Q$  est distribué comme un  $\chi^2$  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 attendus  $A_j$  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  $Q$  à calculer a alors la forme ci-dessus, où la somme comprend  $k \times m$  termes.

La distribution de  $Q$ , sous  $H_0$  est alors un  $\chi^2$  à  $(k - 1) \times (m - 1)$  degrés de liberté.

Les conditions de validité du test sont :  $A_j \geq 5, 1 \leq j \leq km$

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

La valeur observée de  $Q$ , notée  $Q_c$ , sera comparée à la valeur  $K_{ddl;0,05}$  :

- si  $Q_c \leq K_{ddl;0,05}$  on ne conclut pas. Il n'est pas démontré que les deux répartitions « vraies » diffèrent.
- si  $Q_c > K_{ddl;0,05}$  on conclut que les deux répartitions observées diffèrent significativement.

**Suite de l'exemple** : on obtient :

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

soit :  $Q_c = 9,96$

Or :  $K_{2;0,05} = 5,99 \Rightarrow$  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  $\chi^2$  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  $Q$  coïncide avec le carré de la valeur de  $Z$ ,  $Z$  étant le paramètre formé pour comparer directement ces proportions (voir chapitre 12).

## 13.3 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 6 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 :

$Pr(\text{la modalité de la couleur des cheveux est } l \text{ et la modalité de la couleur des yeux est } c) = Pr(\text{la modalité de la couleur des cheveux est } l) \times Pr(\text{la modalité de la couleur des yeux est } c)$ , 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  $\times$  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)  $\times$  (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  $Q$ , et s'exprime exactement comme précédemment, c'est-à-dire :

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

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 attendus 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 (D. Schwartz, *Méthodes statistiques à l'usage des médecins et des biologistes*, Flammarion (collection statistique en biologie et médecine), 3<sup>e</sup> 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 13.2 page 133. 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 attendus s'obtiennent comme précédemment. Ainsi, l'effectif attendu relatif au couple « blonds, marrons » sera :  $45/124 \times 33/124 \times 124 = 11,9$ .

### REMARQUES

- Pour alléger les calculs, on peut remarquer que l'effectif attendu 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.
- La somme des effectifs attendus, soit en ligne, soit en colonne, coïncide avec les mêmes sommes sur les effectifs observés. Cette remarque permet une vérification partielle des calculs.
- 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,  $Q$  EST  
DISTRIBUE SELON un  $\chi^2$  à :

(nombre de modalités de la première variable - 1)  $\times$  (nombre de modalités de la seconde variable - 1)

DEGRES DE LIBERTE.

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

3. La suite des étapes est habituelle

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

- si la valeur calculée de  $Q$ , notée  $Q_c$ , est inférieure à  $K_{ddl,\alpha}$ , on ne rejette pas l'hypothèse d'indépendance des deux variables.
- si la valeur calculée  $Q_c$  est supérieure à  $K_{ddl,\alpha}$ , on rejette l'hypothèse d'indépendance des deux variables. On dira alors que les deux variables sont liées, au risque  $\alpha$ .

**Exemple :**

Dans l'exemple ci-dessus, la valeur de  $Q_c$ , 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  $K_{6;0,05}$  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 observées on a :

$$Pr(\text{yeux bleus}) = 44/124 = 0,35$$

$$Pr(\text{yeux bleus} / \text{cheveux blonds}) = 25/45 = 0,56$$

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  $Q_c$  ci-dessus chiffre dans leur ensemble les différences entre  $Pr(A / B)$  et  $Pr(A)$ , c'est-à-dire les écarts de  $Pr(A \text{ et } B)$  par rapport au produit  $Pr(A)Pr(B)$ , 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  $\chi^2$ . Effectifs observés  $O_j$ , effectifs attendus  $A_j$ .

Conditions de validité générales :  $A_j \geq 5$

Paramètre général :

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

## 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

$Q \sim \chi^2(\text{nombre de modalités} - 1)$

## Comparaison de plusieurs répartitions observées (homogénéité)

$H_0$  : Les répartitions coïncident

$H_1$  : Les répartitions diffèrent

Nombre de cases = nombre de modalités  $\times$  nombre de répartitions

$Q \sim \chi^2((\text{nombre de modalités} - 1) \times (\text{nombre de répartitions} - 1))$

## Test d'indépendance de deux variables qualitatives

$H_0$  : Les deux variables sont indépendantes

$H_1$  : Les deux variables sont liées

$Q \sim \chi^2((\text{nb de modalités de 1}^{\text{ère}} \text{ variable} - 1) \times (\text{nb de modalités de 2}^{\text{ème}} \text{ variable} - 1))$

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  $\chi^2$  est  $(l - 1)(c - 1)$ .

# Chapitre 14

## Liaison entre deux variables continues : notion de corrélation

### 14.1 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 12, 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 13) 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 :

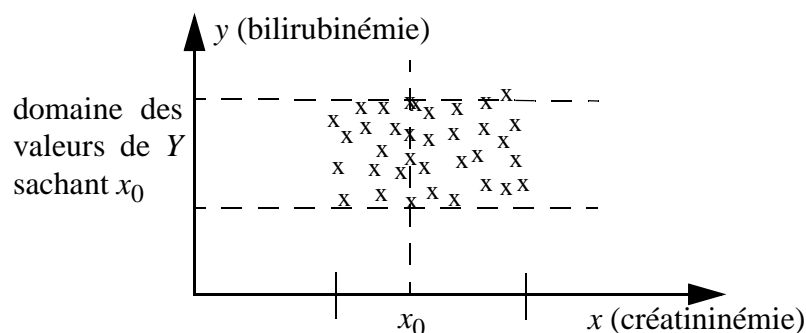
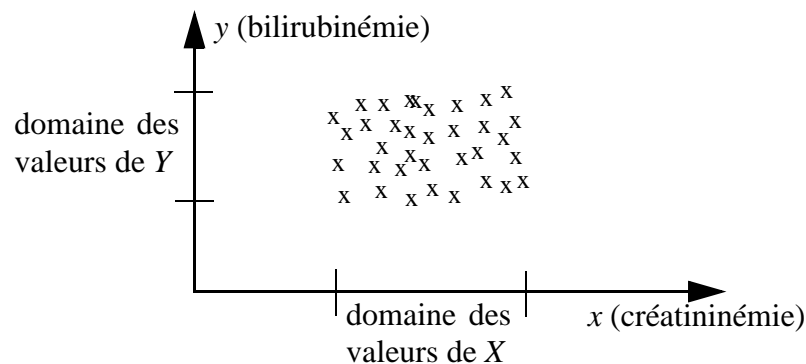
- la consommation de cigarettes (quotidienne ou cumulée) et la capacité respiratoire sont-elles liées ?
- la gastrinémie et la quantité de cellules ECL sont-elles liées ?
- les valeurs de glycémie obtenues selon deux méthodes de dosage sur les mêmes échantillons sanguins sont-elles liées [ici, il faut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

## 14.2 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  $Y$ .

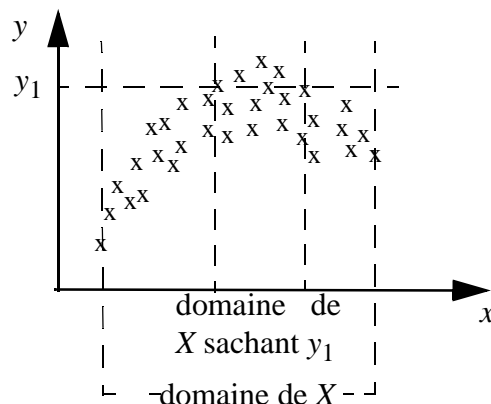
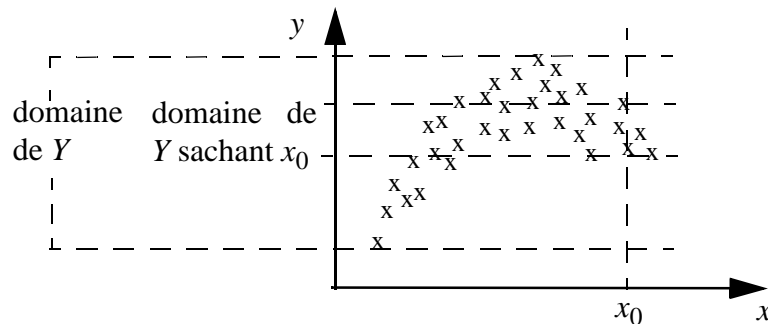
Intéressons nous à un nouvel individu ; ne mesurons chez lui que la valeur de la créatininémie,  $x_0$ . Que peut-on dire alors, sur la base de cette connaissance et sur la base de l'expérience ci-dessus concernant le domaine des valeurs possibles de  $Y$  **pour ce même individu** ? On peut proposer 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.



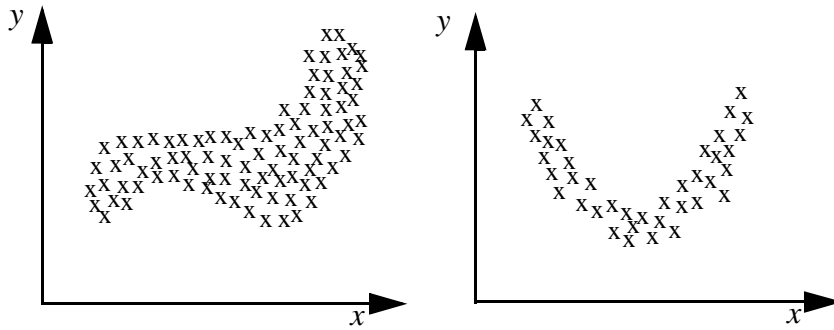
La modification ici concerne aussi bien l'amplitude du domaine que sa localisation en termes de valeurs.

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-

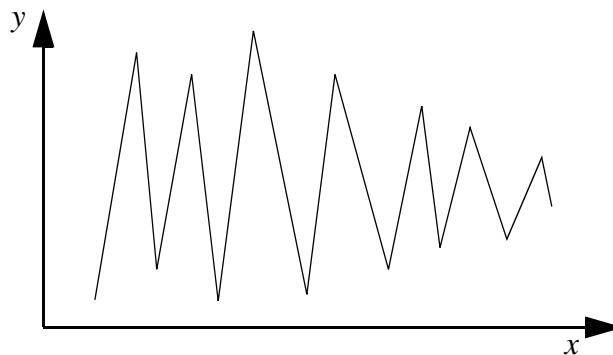
dessous.



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 ?



## 14.3 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 inchangés s'ils produisent  $X + a$ ,  $Y + b$ . Ainsi l'indicateur se fondera-t-il sur les valeurs  $x_i - m_x$  et  $y_i - m_y$ . Par ailleurs, on souhaite que l'indicateur ne dépende pas des unités exprimant  $X$  et  $Y$  ; alors on travaillera sur



$$x_{ri} = \frac{x_i - m_x}{s_X} \text{ et } y_{ri} = \frac{y_i - m_y}{s_Y}$$

Maintenant si  $X$  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.

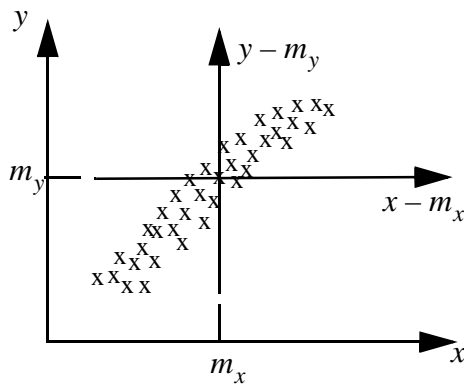
Retenons, exprimé sur la base des valeurs observées :

$$r = \frac{\frac{1}{n-1} \sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y}$$

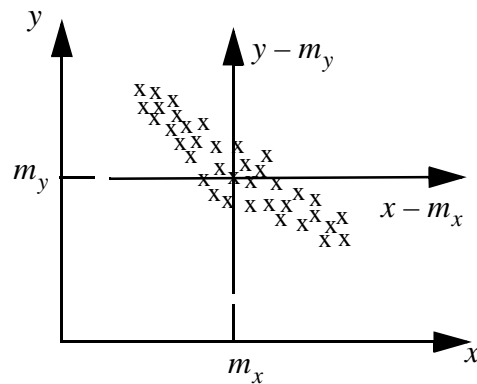
Le numérateur de cette expression est appelé la covariance observée des deux variables  $X$  et  $Y$ , notée  $cov_0(X, Y)$ , dont on montre qu'elle s'exprime aussi sous la forme

$$cov_0(X, Y) = \frac{n}{n-1} \left( \frac{1}{n} \sum_i x_i y_i - \bar{x} \bar{y} \right)$$

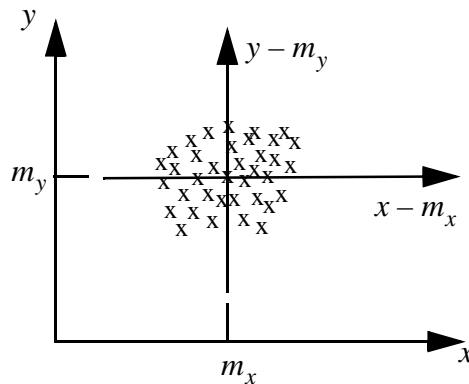
Les figures ci-dessous présentent diverses situations relativement au coefficient de corrélation observé.



$r > 0$ , grand



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



$r$  voisin de zéro

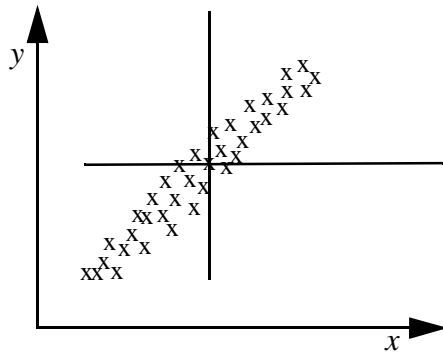
Propriétés numériques fondamentales de  $r$  :

- $r$  a toujours une valeur comprise entre -1 et 1 ;
- $r$  prend la valeur -1 (respectivement 1) si et seulement si il existe des valeurs  $a$  et  $b$  telles qu'on ait pour tout  $i$   $y_i = ax_i + b$  avec  $a$  négatif (respectivement  $a > 0$ ).

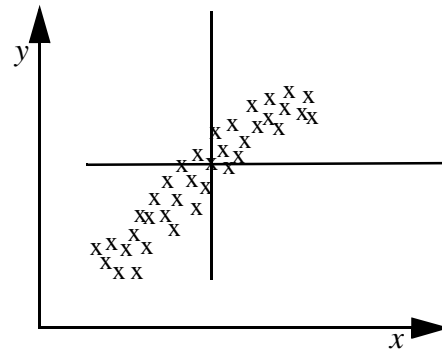
Remarques :

- 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 valeurs observées,
  - 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 observé très faible en valeur absolue.

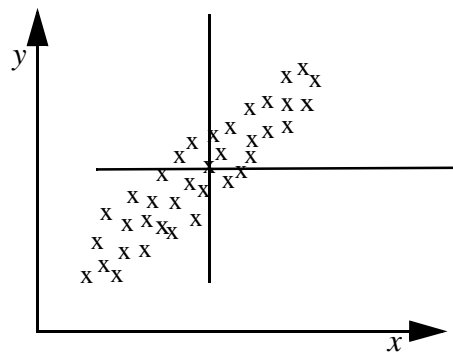
Quelques exemples sont présentés ci-dessous pour fixer les idées.



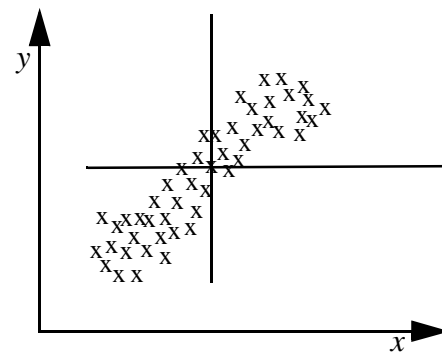
$$r \approx 0,9$$



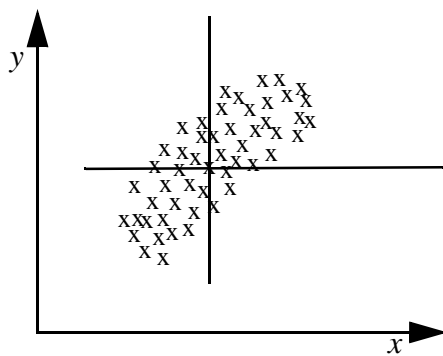
$$r \approx 0,7$$



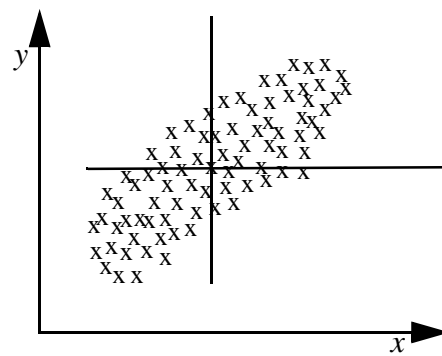
$$r \approx 0,7$$



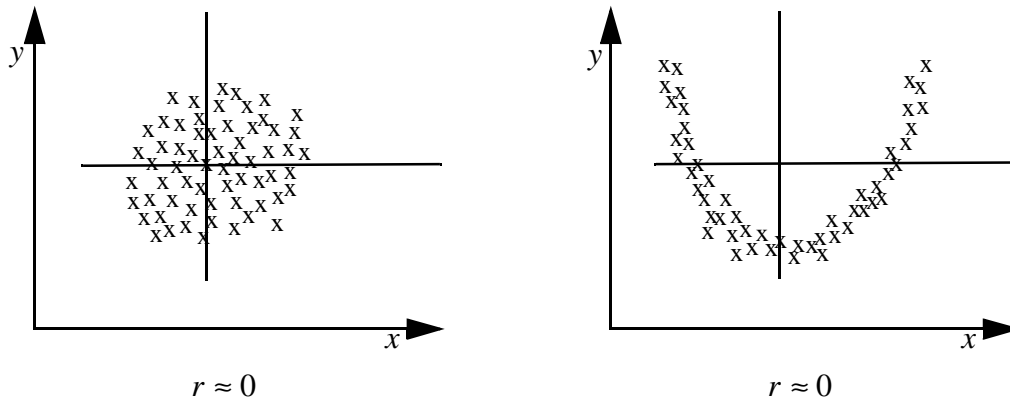
$$r \approx 0,6$$



$$r \approx 0,5$$



$$r \approx 0,5$$



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 observé  $r$ .

## 14.4 Le coefficient de corrélation « vrai »

Cherchons à substituer de la façon la plus naturelle possible des grandeurs « vraies » aux grandeurs observées constitutives de  $r$ . On note l'apparition au dénominateur de  $s_X$  et  $s_Y$  auxquelles on substitue naturellement  $\sigma_X$  et  $\sigma_Y$ , les écarts types « vrais » de  $X$  et  $Y$ . Au numérateur on remarque  $m_x$  et  $m_y$  auxquels on substitue  $E(X)$  et  $E(Y)$  les moyennes « vraies » de  $X$  et  $Y$ . Reste au numérateur une moyenne observée (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  $\rho$  :

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

Remarque : à propos des notions d'espérance, de covariance « vraie », de coefficient de corrélation « vrai », voir le chapitre 6.

## 14.5 Test de comparaison du coefficient de corrélation « vrai » $\rho$ à 0

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 »  $\rho$ . 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 observé

$$r = \frac{\frac{1}{n-1} \sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y}$$

- 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  $(x_i, y_i)$  expérimentaux. L'intervalle de pari pour  $r$  est de la forme  
 $IP_{1-\alpha} = [-\text{corr}_\alpha(n-2) ; \text{corr}_\alpha(n-2)]$ ,  $\text{corr}_\alpha(n-2)$  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 observé doit être supérieur à 0,444, ou inférieur à -0,444.

### Autre formulation du test

On peut montrer que  $t = r \sqrt{\frac{n-2}{1-r^2}}$  est, sous  $H_0$ , distribué selon une loi de Student à  $n-2$  ddl.

Si on préfère utiliser ce paramètre plutôt que  $r$ , il faut lire la table de Student pour construire l'intervalle de pari.

# 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 observé :

$$r = \frac{\frac{1}{n-1} \sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y} = \frac{\frac{n}{n-1} \left( \frac{1}{n} \sum_i x_i y_i - m_x m_y \right)}{s_X s_Y}$$

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 15

## A propos des tests d'hypothèses

### 15.1 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 11. 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  $\alpha$  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 :  $\varphi_1 = \varphi_0$  ; puis service 2 :  $\varphi_2 = \varphi_0$ .

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 :

$Pr(\text{conclure } H_1 \text{ au moins une fois si } H_0 \text{ est vraie}) = 1 - Pr(\text{ne rejeter } H_0 \text{ aucune des deux fois si } H_0 \text{ est vraie})$

Or  $Pr(\text{ne pas rejeter } H_0 \text{ si } H_0 \text{ est vraie}) = 1 - \alpha$

Donc  $Pr(\text{ne rejeter } H_0 \text{ aucune des deux fois si } H_0 \text{ est vraie}) = (1 - \alpha)^2$

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 observée 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 observée 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  $\alpha$  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 CLAI-  
REMENT AVANT LA REALISATION DE L'EXPERIENCE

## 15.2 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 dimi-

nuer 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 plirai » 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 l'ouvrage de D. Schwartz, R. Flamant et J. Lellouch : « L'essai thérapeutique chez l'homme » (Flammarion Médecine Sciences).

# Chapitre 16

## Analyse des durées de survie

### 16.1 Contexte

On cherche à évaluer l'efficacité d'un nouveau traitement chez des personnes atteintes de maladies graves, qui peuvent donner lieu à des complications, à des rechutes, ou même aboutir au décès. Dans tous les cas, le critère d'intérêt peut s'exprimer comme la durée entre l'instauration du traitement et l'apparition de l'événement témoignant de l'échec du traitement. Pour simplifier l'écriture, on supposera par la suite que l'événement considéré est le décès, et on s'intéressera à la durée de survie des individus, comptée à partir d'une date définie ; à partir de l'instauration du traitement dans l'exemple précédent. Il faut noter qu'en réalité ce chapitre concerne toutes les données censurées, telles qu'elles seront définies ci-dessous.

Au moment où on effectue l'analyse des données, on dispose en général d'un échantillon composé de deux types d'information :

- Certains individus sont décédés, et on dispose alors d'une information complète, de la forme  $(T = t)$ , où  $T$  est la variable aléatoire « durée de survie ».
- D'autres individus, au contraire, sont vivants au moment de l'analyse. Tout ce qu'on connaît de leur durée de survie est que celle-ci est supérieure au temps qui s'est écoulé entre l'instauration du traitement et le moment de l'analyse. On a donc une information de la forme  $(T > t)$ .

**Définition** : une donnée dont on ne connaît qu'une borne est dite **censurée**.

Un échantillon comportant, en plus des données « habituelles », c'est à dire dont la valeur est observée, des données censurées est appelé un échantillon censuré.

Dans la suite de ce chapitre, on va chercher à résoudre deux types de problèmes :

- un problème d'estimation : comment « décrire » au mieux l'information contenue dans un échantillon censuré ? On verra que la réponse consiste en une courbe, souvent appelée « courbe de survie ».
- un problème de test : comment comparer deux courbes de survie ?

## 16.2 Estimation : courbe de survie

### 16.2.1 Définitions

On appelle **fonction de survie** de la variable  $T$ , et on note  $S(t)$ , la fonction qui à tout  $t > 0$  associe la valeur de  $Pr(T > t)$ , la probabilité que  $T$  soit strictement supérieur à  $t$ .

On appelle **risque instantané** de décès au temps  $t$ , et on note  $h(t)$ , la limite suivante :

$$\lim_{dt \rightarrow 0} \frac{Pr[(T < t + dt)/(T \geq t)]}{dt}$$

En français, le numérateur exprime la probabilité pour qu'un individu décède juste après le temps  $t$  sachant qu'il était vivant à cet instant. Le risque instantané est donc une probabilité conditionnelle par unité de temps ; il exprime à chaque instant la « force de mortalité ».

Connaître la variable aléatoire  $T$ , c'est connaître sa fonction de survie, puisqu'elle n'est autre que  $1 -$  sa fonction de répartition.

Pour estimer cette fonction  $S$ , on utilise une courbe de survie  $\hat{S}$ , construite à partir d'un échantillon censuré (remarque : les calculs qui suivent peuvent aussi s'appliquer à un échantillon ordinaire, c'est-à-dire sans données censurées).

Il existe deux méthodes principales d'estimation des fonctions de survie à partir d'un échantillon : la méthode de Kaplan-Meier et la méthode actuarielle. Ces deux méthodes reposent sur des principes voisins.

### 16.2.2 Estimation par la méthode de Kaplan-Meier

Dans ce paragraphe, on utilisera l'exemple suivant : on dispose de l'échantillon censuré de durées de survie suivant :

$$t_1=2, t_2=3^*, t_3=4, t_4=4, t_5=4^*, t_6=5, t_7=8^*, t_8=10, t_9=12^*$$

Les valeurs suivies d'un astérisque ( $t^*$ ) désignent les données censurées.

Comment représenter au mieux cet échantillon ? La description la plus complète utilise la courbe de survie.

Bien que la durée de survie soit par nature une variable continue, on considère ici qu'elle ne prend que des valeurs discrètes, multiples entiers d'une unité de temps choisie arbitrairement.

La construction de la courbe de survie correspondant à l'exemple donné ci-dessus repose sur les calculs suivants :

- On montre qu'il faut qu'une courbe de survie reste constante entre deux temps de décès consécutifs observés ;
- Pour  $t=0$ , on a toujours  $\hat{S}(0) = 1$  ;
- Entre  $t=0$  et  $t=2$ ,  $S(t) = 1$  car aucun décès n'est survenu entre  $T=0$  et  $T=2$  ;

- En  $T=t_1=2$ , on observe un décès. Celui-ci est pris en compte de la manière suivante : C'est à partir de l'estimation des risques instantanés qu'on construit la courbe de survie. Ici, puisque la durée de survie est supposée discrète, **le risque instantané** de décès à  $t_1$  s'exprime par  $h(t_1) = Pr(T=t_1 / T \geq t_1)$ . Il peut donc s'estimer par la proportion correspondante

$$\hat{h}(t_1) = D_1/N_1, \text{ où}$$

- $D_1$  est le nombre d'individus décédés à  $t_1$  (ici  $D_1=1$ ),
- $N_1$  est le nombre d'individus susceptibles de décéder (on dit « à risque ») à  $t_1$ . Ici, on a  $N_1 = 9$ .

On a donc :  $h(t_1) = Pr(T = t_1 / T \geq t_1) \approx 1/9 = 0,111 = \hat{h}(t_1)$ .

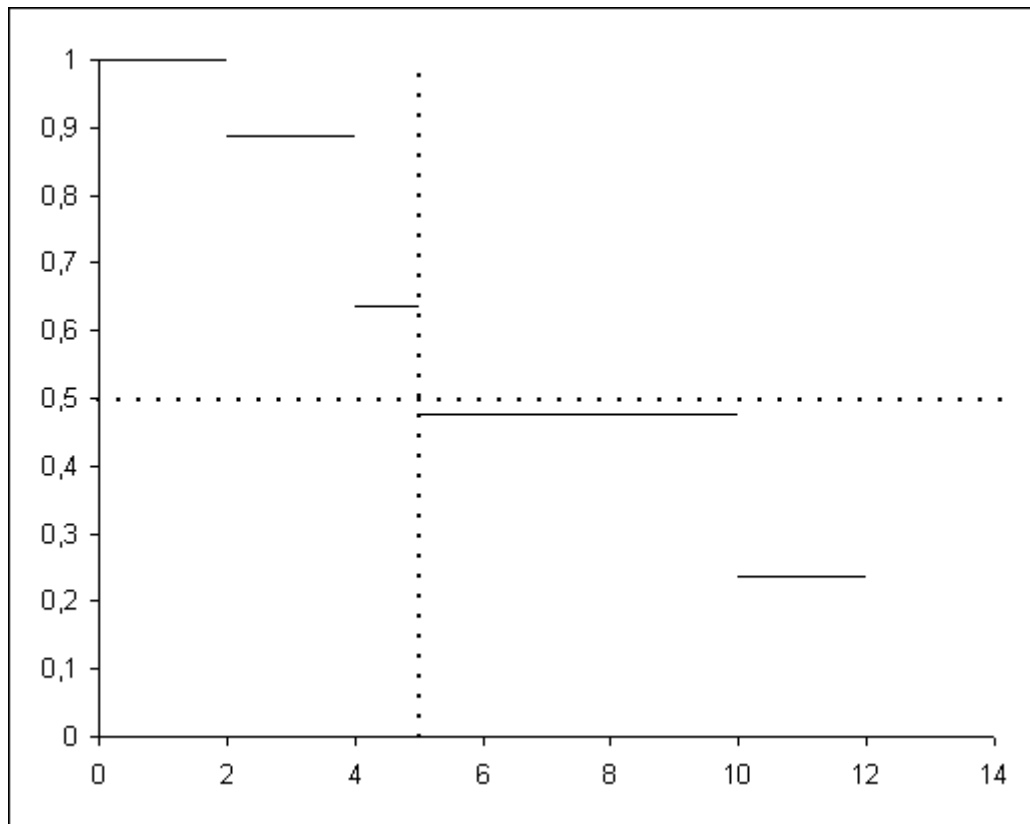
La probabilité de survivre après  $t_1$  est donc  $S(t_1) = 1 - h(t_1) \approx 0,889 = \hat{S}(t_1)$ .

- $S(t)$  conserve cette valeur tant qu'aucun décès n'est observé, c'est-à-dire jusqu'à  $T=t_3=4$ . (Notons que l'observation censurée en  $t_2=3$  ne modifie pas localement la fonction de survie  $S(t)$ ).
- En  $T = t_3=4$ , on a :  $D_3 = 2, N_3 = 7$ , donc  $h(t_3) \approx 2/7 = 0,286$   
Donc  $Pr(T > t_3 / T \geq t_3) = 1 - Pr(T = t_3 / T \geq t_3) = 1 - h(t_3) \approx 1 - 0,286 = 0,714$ .  
Mais  $S(t_3) = Pr(T > t_3) = Pr(T > t_3 / T \geq t_3) \times Pr(T \geq t_3)$ .  
Or  $Pr(T \geq t_3) = S(t_1) = 1 - h(t_1)$ .  
Finalement,  $S(t_3) = (1 - h(t_1)) \times (1 - h(t_3)) \approx 0,889 \times 0,714 = 0,635$ .  
 $S(t)$  conserve cette valeur jusqu'à  $T = t_6 = 5$ .
- En  $T = t_6 = 5$ , on trouve par le même raisonnement :  
 $S(t_6) = (1 - h(t_6)) \times S(t_3) \approx (1 - 0,25) \times 0,635 = 0,75 \times 0,635 = 0,476$ .
- On trouve de même  $S(t_8) \approx 0,476 \times 0,5 = 0,238$ .

Les calculs nécessaires à la construction d'une courbe de survie se résument sous la forme d'un tableau :

$t_i$	$D_i$	$N_i$	$\hat{h}(t_i)$	$1-\hat{h}(t_i)$	$\hat{S}(t_i)$
2	1	9	0,111	0,889	0,889
4	2	7	0,286	0,714	0,635
5	1	4	0,250	0,750	0,476
10	1	2	0,500	0,500	0,238

Et la courbe de survie correspondante peut être dessinée :



Cette courbe résume l'évolution d'un échantillon de patients après l'événement définissant le temps 0, événement qui est souvent en pratique la date de début d'un traitement, mais peut aussi être la date de diagnostic d'une maladie, ou la date de naissance.

A partir de la courbe, on peut estimer la probabilité de survivre pour des durées données ; on s'intéresse par exemple souvent à la survie à 5 ans en cancérologie.

On peut aussi estimer la durée correspondant à une probabilité de survie donnée. La probabilité de survie la plus souvent utilisée est égale à 0,5, et on parle alors de **médiane de survie**, notée  $t_m$  : c'est la durée pour laquelle la fonction de survie est égale à 0,5. On l'estime sur la courbe de survie par le temps correspondant à l'intersection de la courbe avec la droite horizontale d'équation  $S(t) = 0,5$ . Sur l'exemple,  $t_m = 5$ .

### Remarque 1

La courbe de survie se termine de manière différente selon que le ou les événements correspondant à la durée maximale observée sont des décès ou des censures. Si le dernier événement est un décès isolé ou plusieurs décès sans censure, le risque instantané au temps correspondant vaut 1, et la courbe rejoint l'axe des abscisses. La fonction de survie est nulle au-delà du temps maximum observé sur l'échantillon. A l'inverse, si le dernier événement est une censure ou comporte au moins une censure, la courbe de survie ne rejoint pas l'axe des abscisses après le temps maximum. La courbe de survie n'est pas définie après ce temps.



**Remarque 2**

La courbe de survie est une estimation, réalisée à partir des données, de la fonction de survie « vraie ». Comme toute estimation, elle est aléatoire, et a une espérance et une variance. On montre qu'elle est sans biais. Sa variance est calculable, mais la formule est hors programme.

**16.2.3 Estimation par la méthode actuarielle**

La méthode décrite ci-dessus est adaptée aux échantillons de faible taille, comme on en rencontre fréquemment dans les essais cliniques. Lorsque les échantillons sont de grande taille (par exemple l'échantillon des délais de survenue du premier accident de voiture après obtention du permis de conduire chez les clients d'une compagnie d'assurances), une autre méthode d'estimation de fonction de survie est proposée, la méthode actuarielle.

La variable délai de survie est ici considérée comme continue. Le principe de la méthode est d'estimer la fonction de survie en des temps définis à l'avance, par exemple tous les mois. On note  $0, b_1, b_2, \dots, b_r$ , les différents temps retenus.

On utilise le même principe de récurrence que plus haut pour exprimer la fonction de survie au temps  $b_i$  ( $S(b_i)$ ) en fonction de  $S(b_{i-1})$  :

$$S(b_i) = S(b_{i-1}) \times \Pr(T > b_i / T > b_{i-1})$$

Cette dernière probabilité conditionnelle s'estime par le complémentaire à 1 de la proportion d'individus de l'échantillon vivants au temps  $b_{i-1}$  qui décèdent dans l'intervalle  $[b_{i-1}, b_i]$  (pour ne pas compliquer les notions et notations, on ne distingue pas ici entre intervalles ouverts ou fermés).

Soient  $D_i$  le nombre d'individus dont le décès a été observé dans l'intervalle  $[b_{i-1}, b_i]$ ,  $N_i$  le nombre de sujets qu'on sait vivants au temps  $b_{i-1}$ , et  $C_i$  le nombre d'individus censurés dans l'intervalle  $[b_{i-1}, b_i]$ .

On estime que dans l'intervalle de délais  $[b_{i-1}, b_i]$ , un nombre « moyen » de  $N_i - C_i/2$  patients étaient à risque de décéder (ils sont  $N_i$  au temps  $b_{i-1}$  et  $N_i - C_i$  au temps  $b_i$ ).

Finalement la probabilité conditionnelle ci-dessus est estimée par :  $1 - \frac{D_i}{N_i - \frac{C_i}{2}}$

De proche en proche on obtient des estimations des  $S(b_i)$  aux différents temps d'intérêt. Les valeurs de la courbe de survie à d'autres temps, jusqu'à  $b_r$ , s'obtiennent par interpolation linéaire.

**16.3 Test de comparaison de deux courbes de survie : test du log-rank approché**

Dans un essai thérapeutique, on compare le plus souvent deux traitements : un nouveau traitement et un traitement de référence, qui peut être le meilleur traitement actuellement connu de la maladie, ou un traitement placebo. Quand la réponse aux traitements est décrite par une courbe de survie,

on est amené naturellement à se poser le problème de la comparaison statistique de deux courbes de survie. De manière plus précise, on cherche à effectuer un test visant à démontrer que les fonctions de survie « vraies » correspondant aux deux traitements sont différentes.

Le test que nous allons décrire ici est celui du log-rank approché. Ce test est construit à partir des deux fonctions de survie  $S_A(t)$  et  $S_B(t)$ , correspondant aux traitements  $A$  et  $B$ , respectivement. Ces fonctions représentent les probabilités  $Pr(T > t)$  avec les traitements  $A$  et  $B$ , et sont définies dans une même population de malades. Le test se déroule selon les étapes habituelles :

1. L'**hypothèse nulle** est l'égalité des deux courbes de survie. L'hypothèse alternative est le complémentaire de l'hypothèse nulle. En écriture abrégée, on note :  
 $H_0 : S_A(t) = S_B(t)$ , pour  $t > 0$ ,  
 $H_1$  : il existe au moins un  $t$  tel que  $S_A(t) \neq S_B(t)$
2. Le **paramètre du test** est basé sur une distance entre les deux courbes, distance qui ressemble beaucoup à la distance qui est utilisée pour le test du  $\chi^2$ . Sous l'hypothèse nulle, on calcule en chaque temps où il y a un décès dans au moins une des deux courbes un effectif attendu de décès pour chaque groupe. Cet effectif se calcule simplement de la manière suivante :
  - Soit, au temps  $i$ , les *données* :
    - pour le groupe  $A$  :  $D_{iA}$  décès parmi  $N_{iA}$  individus à risque
    - pour le groupe  $B$ ,  $D_{iB}$  décès parmi  $N_{iB}$  individus à risque
 ( $D_{iA}$  ou  $D_{iB}$  peut être nul)
  - Sous  $H_0$ , les risques instantanés de décès sont égaux dans les deux groupes, et on peut estimer le *risque commun* par :  
 $\hat{h}(t_i) = D_i / N_i$ , où  $D_i = D_{iA} + D_{iB}$  et  $N_i = N_{iA} + N_{iB}$
  - et les *effectifs attendus* de décès dans les groupes  $A$  et  $B$  au temps  $i$  sont :  
 $E_{iA} = N_{iA} \times \hat{h}(t_i)$   
 $E_{iB} = N_{iB} \times \hat{h}(t_i)$
  - Par *sommation* des effectifs observés et calculés de décès sur tous les temps, pour chaque groupe, on obtient :  
 $D_A = \sum D_{iA}$ ,  $D_B = \sum D_{iB}$ ,  $E_A = \sum E_{iA}$ ,  $E_B = \sum E_{iB}$
  - Le *paramètre du test* est finalement :  $Q = \frac{(D_A - E_A)^2}{E_A} + \frac{(D_B - E_B)^2}{E_B}$
  - Sous l'hypothèse nulle, cette quantité suit approximativement la loi d'un  $\chi^2$  à un *degré de liberté*, d'où la suite de la construction du test
3. On choisit  $\alpha = 0,05$  et on déduit l'**intervalle de pari** de  $Q$  :  $[0 \ 3,84]$
4. **Règle de décision** :
  - Si  $Q_c > 3,84$  on rejette l'hypothèse nulle. On conclut que les courbes sont significativement différentes au risque  $\alpha = 0,05$ .
  - Si  $Q_c < 3,84$  on ne rejette pas l'hypothèse nulle.
5. Les **conditions de validité** du test sont :  $E_A \geq 5$  et  $E_B \geq 5$ .

En pratique, on effectue les calculs en résumant les données et les effectifs dans un tableau. Par

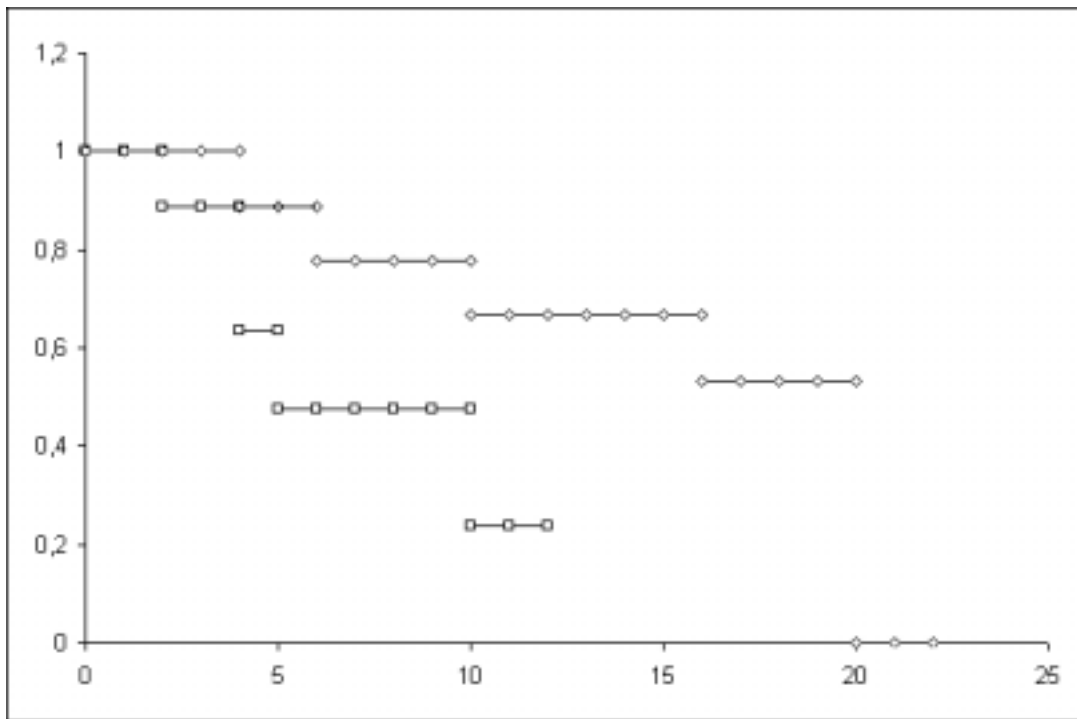
exemple, soient les données suivantes :

Groupe A	$t_{A1}=2, t_{A2}=2, t_{A3}=3^*, t_{A4}=3^*, t_{A5}=4, t_{A6}=4, t_{A7}=4, t_{A8}=4, t_{A9}=4^*, t_{A10}=4^*, t_{A11}=5, t_{A12}=5, t_{A13}=8^*, t_{A14}=8^*, t_{A15}=10, t_{A16}=10, t_{A17}=12^*, t_{A18}=12^*$
Groupe B	$t_{B1}=4, t_{B2}=6, t_{B3}=10, t_{B4}=14^*, t_{B5}=16, t_{B6}=16, t_{B7}=16^*, t_{B8}=20, t_{B9}=20$

On peut construire le tableau suivant :

$t_i$	$D_{iA}$	$N_{iA}$	$D_{iB}$	$N_{iB}$	$\hat{h}_i$	$E_{iA}$	$E_{iB}$
2	2	18	0	9	0,074	1,333	0,667
4	4	14	1	9	0,217	3,043	1,957
5	2	8	0	8	0,125	1,000	1,000
6	0	6	1	8	0,071	0,429	0,571
10	2	4	1	7	0,273	1,091	1,909
16	0	0	2	5	0,400	0	2,000
20	0	0	2	2	1,000	0	2,000
<b>Total</b>	10		7			6,896	10,104

On trouve  $Q_c = 2,350$ . On ne rejette pas l'hypothèse nulle d'égalité des fonctions de survie entre les groupes A et B.



### Les courbes de survie de l'exemple

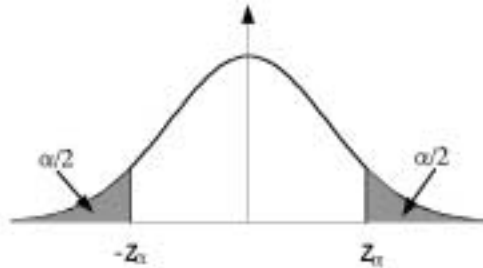
Les traits portant des petits carrés représentent la courbe A

Ceux portant de petits losanges représentent la courbe B

# Annexe A

## Tables statistiques

# A.1 TABLE DE LA VARIABLE NORMALE REDUITE Z



$\alpha$	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
<b>0,00</b>	$\infty$	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
<b>0,10</b>	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
<b>0,20</b>	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
<b>0,30</b>	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
<b>0,40</b>	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
<b>0,50</b>	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
<b>0,60</b>	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
<b>0,70</b>	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
<b>0,80</b>	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
<b>0,90</b>	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

La probabilité  $\alpha$  s'obtient par addition des nombres inscrits en marge  
 exemple : pour  $z_\alpha = 0,994$ , la probabilité est  $\alpha = 0,30 + 0,02 = 0,32$

TABLE POUR LES PETITES VALEURS DE LA PROBABILITÉ

$\alpha$	0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
$z_\alpha$	3,29053	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

(d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)

## A.2 TABLE DU TEST DE WILCOXON

*Table adaptée de Siegel*

$n$	$\alpha$		
	0,05	0,02	0,01
<b>6</b>	2,118		
<b>7</b>	1,961	2,299	
<b>8</b>	2,044	2,324	2,464
<b>9</b>	2,026	2,263	2,381
<b>10</b>	1,947	2,253	2,456
<b>11</b>	2,009	2,276	2,454
<b>12</b>	2,008	2,322	2,479
<b>13</b>	1,964	2,313	2,523
<b>14</b>	1,952	2,329	2,517
<b>15</b>	1,965	2,306	2,533

Indique, pour  $n \leq 15$  les valeurs de  $W_\alpha$  pour  $\alpha = 0,05, 0,02$  et  $0,01$ .

## A.3 TABLE DU TEST DE MANN-WHITNEY-WILCOXON

Table adaptée de Siegel

$n_A$	$n_A$								
	$\alpha$	3	4	5	6	7	8	9	10
<b>4</b>	<b>0,05</b>	2,333	1,905						
	<b>0,01</b>	2,687	2,483						
<b>5</b>	<b>0,05</b>	2,117	2,107	2,110					
	<b>0,01</b>	2,415	2,596	2,528					
<b>6</b>	<b>0,05</b>	1,962	2,047	2,118	2,018				
	<b>0,01</b>	2,479	2,473	2,483	2,498				
<b>7</b>	<b>0,05</b>	2,074	2,003	1,965	2,086	2,057			
	<b>0,01</b>	2,530	2,570	2,615	2,514	2,568			
<b>8</b>	<b>0,05</b>	1,960	1,970	1,991	2,014	2,037	1,953		
	<b>0,01</b>	2,572	2,480	2,576	2,530	2,500	2,584		
<b>9</b>	<b>0,05</b>	2,052	2,099	2,013	1,956	2,022	1,982	2,040	
	<b>0,01</b>	2,422	2,561	2,680	2,546	2,551	2,560	2,570	
<b>10</b>	<b>0,05</b>	1,961	2,065	2,033	2,017	2,010	2,008	2,009	2,011
	<b>0,01</b>	2,366	2,489	2,523	2,560	2,498	2,541	2,580	2,540

Indique, pour  $n_A \leq 10$  et  $n_B \leq 10$ ,  $n_A \leq n_B$ , les valeurs de  $M_\alpha$ , pour  $\alpha=0,05$  et  $\alpha=0,01$ .

Exemple :  $n_A=5$ ,  $n_B=8$  :  $M_{0,05}=1,991$



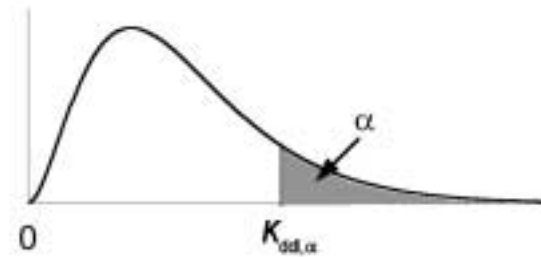
# A.4 TABLE DE $\chi^2$

La table donne la probabilité  $\alpha$  pour que  $\chi^2$  égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d. d. l.)

Quand le nombre de degrés de liberté est élevé,

$\sqrt{2\chi^2}$  est à peu près distribué normalement

autour de  $\sqrt{2(\text{d.d.l.}) - 1}$  avec une variance égale à 1



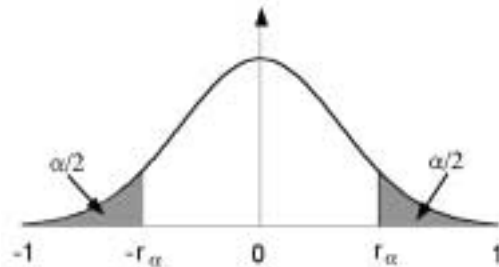
$\alpha$	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
ddl									
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

Exemple : avec d. d. l. = 3, pour  $K_{3,\alpha} = 0,584$  la probabilité est  $\alpha = 0,90$

(d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)

## A.5 TABLE DU COEFFICIENT DE CORRELATION

La table indique la probabilité  $\alpha$  pour que le coefficient de corrélation égale ou dépasse, en valeur absolue, une valeur donnée  $r_\alpha$ , c'est-à-dire la probabilité extérieure à l'intervalle  $(-r_\alpha, +r_\alpha)$ , en fonction du nombre de degrés de liberté (d. d. l.)

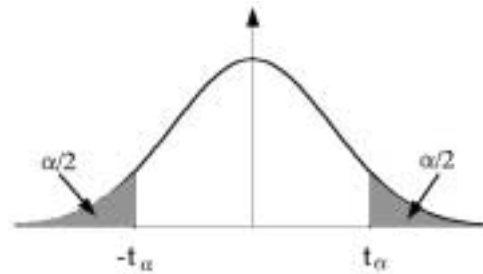


ddl \ $\alpha$	0,10	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

Exemple : avec d. d. l. = 30, pour  $r_\alpha = 0,3494$  la probabilité est  $\alpha = 0,05$

(d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)

# A.6 TABLE DU $t$ DE STUDENT



$\alpha$	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
ddl									
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
$\infty$	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Exemple : avec d. d. l. = 10, pour  $t = 2,228$ , la probabilité est  $\alpha = 0,05$

(d'après Fisher et Yates, Statistical tables for biological, agricultural, and medical research (Oliver and Boyd, Edinburgh) avec l'aimable autorisation des auteurs et des éditeurs)