Collection *Sciences fondamentales*

STATISTIQUE ÉPIDÉMIOLOGIE

3e édition

T. Ancelle

MALOINE

STATISTIQUE ÉPIDÉMIOLOGIE

STATISTIQUE ÉPIDÉMIOLOGIE

Thierry Ancelle

Collection dirigée par J.-F. d'Ivernois

3e édition

MALOINE
23, RUE DE L'ÉCOLE-DE-MÉDECINE, 75006 PARIS
2011

Auteur :
Thierry ANCELLE
Maître de Conférences des Universités – Praticien hospitalier
Faculté de médecine, Université Paris-Descartes

Collection « Sciences fondamentales »

Harry M., *Génétique moléculaire et évolutive*, 2001
Campbell P.N., Smith A.D., *Biochimie illustrée*, 2002
Fawcett W., Jensh R., *Histologie. L'essentiel*, 2002
Mc Geown J.G., *Physiologie. L'essentiel*, 2003
Roitt I., Rabson A., *Immunologie. L'essentiel*, 2002
Bassaglia Y., *Biologie cellulaire*, 2e édition, 2004
Meyer S., Reeb C., Bosdeveix R., *Botanique. Biologie et physiologie végétales*, 2004

Dépôt légal : janvier 2011 – ISBN : 978-2-224-03042-1
Imprimé en Italie

Remerciements

À Laure Beaujouan
pour la révision critique du manuscrit original,
et à l'ensemble des lecteurs qui nous ont
fait part de leurs remarques
sur les précédentes éditions.

Table des matières

Chapitre 20
ANALYSE DE SURVIE

Chapitre 21
PERFORMANCES D'UNE TECHNIQUE

ANNEXES

Avertissement

Cet ouvrage n'est pas un traité de statistique. Il n'aborde les méthodes statistiques que sous un angle pragmatique. Il s'agit donc plutôt d'un manuel d'utilisation.

De nombreux praticiens des sciences de la vie, médecins, infirmières, professions paramédicales, psychologues, vétérinaires, biologistes, doivent utiliser les outils statistiques pour analyser des données, estimer des paramètres, tester une hypothèse.

Dans un domaine limité à leur application courante, la plupart de ces outils sont simples à manier. Les calculs rebutants sont maintenant effectués par des tableurs ou des logiciels bureautiques disponibles pour le grand public. Mais cette simplicité comporte le danger majeur d'erreur méthodologique. Les logiciels effectuent des calculs mais ne garantissent pas leur pertinence.

Toute opération statistique effectuée sur une série de données suppose que ces données suivent un modèle mathématique standard, une loi. La théorie statistique est fondée sur les lois du hasard. Lois purement abstraites qui n'ont de sens que si les variables étudiées sont strictement aléatoires, autrement dit engendrées seulement par le hasard.

Or les variables étudiées dans les sciences de la vie ne sont pas gouvernées par le hasard. Elles sont déterminées à la fois par les lois de la génétique et les contraintes de l'environnement. C'est seulement la multitude du nombre de combinaisons possibles des messages du code génétique, soumis à l'immensité des contraintes environnementales externes, qui donne aux êtres vivants des caractéristiques dont la variabilité s'apparente à celle qui serait fournie par le seul hasard.

Ainsi, la **variabilité biologique** ne fait que donner une illusion, une image du hasard. Dans les sciences du vivant, l'utilisation de la statistique ne va donc pas de soi. Elle est fondée sur une approximation qui assimile la distribution des valeurs observées à des modèles théoriques. Le grand danger pour l'utilisateur est de finir par oublier cette approximation, de persister à utiliser des modèles sans vérifier leur pertinence et finalement d'obtenir des résultats justes d'un point de vue mathématique mais totalement faux dans leur interprétation statistique. Comme en pratique on ne peut pas toujours vérifier l'adéquation des données aux modèles théoriques, les statisticiens ont défini des **conditions d'application** à leur utilisation. Ces conditions, qui dans les manuels figurent souvent en petites lettres, comme les clauses des contrats d'assurance, doivent être impérativement respectées. Et si les conditions d'application d'une méthode ne sont pas remplies, il ne faut pas utiliser la méthode.

Ce point est fondamental et on se gardera d'oublier qu'en biologie, les lois statistiques sont des modèles, mais que les modèles ne font pas loi.

La statistique appliquée aux sciences de la vie permet :

- d'**organiser** les données disparates provenant des observations individuelles ;
- de **décrire** clairement les phénomènes par des paramètres résumant ces observations ;

- d'**estimer** les valeurs de ces paramètres dans les populations d'où proviennent les échantillons observés ;
- de **comparer** ces paramètres entre plusieurs populations ;
- de **prédire** la probabilité de survenue d'événements.

Nous avons pris le parti dans cet ouvrage de présenter les méthodes statistiques usuelles dans un but volontairement utilitaire. Le formalisme mathématique inévitable a été réduit au minimum. Chaque fois que cela a été possible, nous avons préféré illustrer les concepts par des images plutôt que de développer les démonstrations, tout en assumant les dangers d'une illustration réductrice.

Nous avons particulièrement insisté sur les conditions d'application. Nous avons tenté de privilégier la réflexion sur la marche à suivre en utilisant de nombreux exemples. La majeure partie du temps consacré à une étude statistique doit porter sur le choix et la pertinence de la méthode employée, puis sur la signification et l'interprétation des résultats.

Le livre est divisé en quatre parties. La première étudie les outils servant à décrire des données. La deuxième aborde les méthodes d'estimation d'un paramètre inconnu mesuré sur un échantillon. La troisième partie, après un rappel des principes des tests statistiques, présente une série de « fiches pratiques » de 18 tests ; ce chapitre débute par des tableaux d'aide au choix d'un test en fonction de la nature des problèmes, des paramètres à comparer et des conditions d'application (chapitre 13). La quatrième partie est orientée vers les concepts statistiques utilisés en épidémiologie de terrain.

Notre seule ambition serait que ce manuel fournisse au lecteur une aide pratique et lui communique l'envie d'approfondir les notions qu'il aura entrevues.

Avertissement 3e édition

Est-il encore utile de consulter un manuel de statistiques alors que la généralisation de l'informatique et d'Internet permet à chacun d'accéder à des logiciels de statistiques ? Aujourd'hui, bien rares sont ceux qui s'évertuent encore à calculer une variance, un test de chi2 ou une corrélation avec une calculette. Pourtant, l'emploi des logiciels ne résout pas tous les problèmes. L'abondance des tests proposés, le dédale des menus déroulants, les listes d'outils aux noms ésotériques égarent l'utilisateur novice qui n'a pas de stratégie pour se repérer dans cette jungle. Autre difficulté pour le néophyte : les incontournables questions présentées dans les fenêtres des logiciels : risque alpha consenti, hypothèse uni ou bilatérale, homoscédasticité de la distribution, etc. Les réponses à toutes ces questions nécessitent une maîtrise du langage et des concepts statistiques que seule une formation élaborée peut offrir. Nous avons donc choisi de conserver toutes les formules présentées dans les éditions précédentes ainsi que le détail des calculs dans les exemples, non pour inciter le lecteur à les refaire à la main, mais pour bien montrer ce qui est exécuté dans la « boîte noire » des logiciels. En ce qui concerne les tests, nous invitons le lecteur à consulter l'introduction du chapitre 13 qui fournit des éléments de stratégie de leur utilisation.

Première partie

STATISTIQUE DESCRIPTIVE

STATISTIQUE DESCRIPTIVE

VARIABLES

I. VARIABLES QUANTITATIVES
II. VARIABLES QUALITATIVES

ORGANISATION DES DONNÉES

I. TRI DES DONNÉES
II. REGROUPEMENT EN CLASSES
III. TRANSFORMATION DE VARIABLE
IV. EFFECTIFS ET FRÉQUENCES
V. DISTRIBUTION

DESCRIPTION DES DONNÉES

I. TABLEAUX
II. GRAPHIQUES

MESURES EN STATISTIQUE

I. PARAMÈTRES DE POSITION
II. PARAMÈTRES DE DISPERSION

REPRÉSENTATION D'UNE DISTRIBUTION

I. VARIABLE DISCRÈTE : FRÉQUENCE RELATIVE DES CLASSES
II. VARIABLE CONTINUE : DENSITÉ DE PROBABILITÉ
III. CAS D'UNE VARIABLE QUALITATIVE BINAIRE

LOIS DE DISTRIBUTION

I. LOI BINOMIALE
II. LOI DE POISSON
III. LOI NORMALE

VARIABLES

En statistique, on appelle « variable » une caractéristique ou un facteur susceptible de prendre une valeur différente selon les individus (ou les unités statistiques) étudiés. La taille d'un individu est une variable. La couleur des cheveux est une variable. La durée d'incubation d'une maladie est une variable, elle change selon les sujets. Votre performance au 100 mètres est une variable qui change de jour en jour.

On distingue plusieurs types de variables, selon les valeurs qu'elles sont susceptibles de prendre : on peut les différencier en deux groupes : les variables quantitatives qu'on mesure et les variables qualitatives qu'on observe et qu'on dénombre.

I. VARIABLES QUANTITATIVES

Ces variables sont caractérisées par des valeurs numériques. Elles sont exploitables arithmétiquement.

1. Variables quantitatives continues

Ce sont des variables qui peuvent prendre n'importe quelle valeur numérique dans l'intervalle des observations. L'ensemble des valeurs possibles appartient à l'ensemble des nombres réels. Il existe donc une infinité de valeurs théoriques possibles. Les unités décimales utilisées dépendent de leur utilité pratique et de la précision de l'instrument de mesure (exemple 1.1).

Exemple 1.1. VARIABLES QUANTITATIVES CONTINUES

VARIABLE	VALEUR	UNITÉ DE MESURE
poids	56,3	kg
taille	1,72	m
cholestérol	2,45	g/L
pression artérielle	14,5	cm de Hg

Ce type de variable est particulièrement utilisé en biologie médicale car il permet de quantifier objectivement les observations.
La précision est limitée par l'instrument de mesure et les valeurs des variables dites continues ne sont pas strictement continues. Elles sont séparées par des intervalles que cette imprécision ne permet pas d'explorer. Ainsi, la pression artérielle est rarement mesurée avec une précision supérieure au

demi-cm de Hg. Dans une enquête, les sujets se répartissent donc en groupes présentant des valeurs « sautant » de 0,5 en 0,5 cm.
Parfois, c'est l'observateur lui-même qui décide de regrouper des valeurs très précises et très nombreuses, en classes plus réduites. Par exemple, à partir d'une série de sujets dont le poids a été mesuré en kg avec 2 décimales, il décide de d'arrondir les résultats en dizaine de kg. Il transforme alors la variable quantitative continue en **variable quantitative discrète**. On appelle ce procédé ***discrétisation*** d'une variable continue.

2. Variables quantitatives discrètes

Les variables discrètes sont des variables numériques discontinues. Le plus souvent il s'agit de nombres entiers. Il n'existe aucune valeur intermédiaire possible. Une variable discrète est le résultat d'un dénombrement (exemple 1.2).

Exemple 1.2. VARIABLES QUANTITATIVES DISCRÈTES

VARIABLE	VALEUR	UNITÉ DE MESURE
rechute d'une maladie	2	rechutes par an
rappel vaccin	3	injections
parité	5	accouchements
dentition	32	dents

3. Variables temporelles

Ce sont des variables quantitatives particulières qui utilisent les unités de mesure du temps. Elles sont difficiles à exploiter car elles sont basées sur les moyens traditionnels de mesure du temps qui utilisent des divisions en système non-décimal et variable selon les échelles.
Elles se divisent en deux groupes selon qu'elles définissent une durée ou un instant donné.

- Les variables de durée : secondes, minutes, heures, jours, mois, ans… Elles sont de nature continue.
- Les variables servant à définir un instant donné (début ou fin d'un événement). Il en existe deux types principaux, le type date et le type horaire. Les logiciels statistiques modernes permettent de soustraire deux dates ou deux horaires afin de calculer une durée (exemple 1.3).

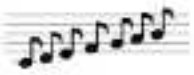
Exemple 1.3. VARIABLES TEMPORELLES

VARIABLE	VALEUR	UNITÉ	TYPE
âge de la grossesse	28	semaines	durée
date de début de la grossesse	15/12/2000	jour/mois/année	date
date de naissance	25/09/1947	jour/mois/année	date
heure de début d'une gastro-entérite	22 h 15	heure/minute	horaire
heure de consommation d'un plat	20 h	heure	horaire
incubation d'une maladie	2 h 15	heure + minute	durée

II. VARIABLES QUALITATIVES

Ce sont des variables qui n'ont pas de valeur numérique. Leurs valeurs sont des qualités réparties en classes. On dénombre les effectifs appartenant à chacune des classes.

1. Variables qualitatives ordinales



Elles s'expriment en classes qui peuvent être ordonnées selon une échelle de valeurs (d'où le nom de variables ordinales) (exemple 1.4).

Exemple 1.4. VARIABLES QUALITATIVES ORDINALES

VARIABLE	CLASSES
niveau d'étude	primaire, secondaire, supérieur
complication d'une maladie	modérée, moyenne, sévère
score d'appréciation de ce livre	convenable, médiocre, exécrable

Ces variables peuvent être recodées pour simplifier leur traitement informatique. Elles prennent alors l'allure de variables qualitatives discrètes (1, 2, 3...). Mais il s'agit d'un simple artifice et rien n'autorise à les manipuler arithmétiquement. Par exemple, un niveau d'études supérieures codé 3 n'a pas une valeur 3 fois supérieure à un niveau d'études primaires codé 1.

2. Variables qualitatives nominales

Il s'agit de variables dont les classes ne peuvent pas être hiérarchisées : elles sont nommées, mais pas ordonnées. L'ordre de présentation est donc arbitraire (exemple 1.5).

Exemple 1.5. VARIABLES QUALITATIVES NOMINALES

VARIABLE	CLASSES
groupe sanguin	A, B, O, AB
état civil	célibataire, marié, divorcé...
nationalité	allemande, française, monégasque...
religion	chrétienne, juive, musulmane...
accident	voie publique, sport, jeux, vie domestique

3. Variables binaires

Il s'agit d'un type particulier de variable qualitative nominale qui ne peut prendre que deux valeurs. On les appelle aussi :

- variables ***dichotomiques***, car elles partagent la population en 2 parties ;
- variables ***booléennes*** pouvant prendre la valeur « VRAI » ou la valeur « FAUX » ;
- variables de ***Bernouilli*** codées respectivement 1 et 0.

Ce type de variable est extrêmement utilisé dans les sciences de la vie et notamment en épidémiologie (exemple 1.6).

Exemple 1.6. VARIABLES BINAIRES

VARIABLE	CLASSES	
état de santé	malade	sain
survie	vivant	décédé
sexe	homme	femme
tabagisme	fumeur	non-fumeur
statut dans une étude épidémiologique	cas	témoin
statut vaccinal	vacciné	non vacciné
exposition à un aliment	consommé	non consommé
caractère génétique	HLA B27	absence de ce gène
réussite à l'examen de statistique	reçu	collé

Exercice

Soit le tableau de données brutes suivant :

A	B	C	D	E	F	G	H
N°	Identification	Sexe	Date de naissance	Taille en cm	Nationalité	Couleur des yeux	Niveau d'études
1	Aurélien	M	24/04/1965	170	F	marron	primaire
2	Hadrien	M	25/02/1956	163	F	bleu	secondaire
3	Julien	M	12/03/1982	162	B	noir	supérieur
4	Émilie	F	30/12/1981	165	F	vert	primaire
5	Steve	M	23/05/1974	182	IRL	marron	primaire
6	Marco	M	12/01/1978	178	E	noir	secondaire

De quel type sont les données des colonnes B à H ?

Résumé

On distingue plusieurs types de variables selon les valeurs qu'elles sont susceptibles de prendre : variables quantitatives qu'on mesure, variables qualitatives qu'on dénombre.

TYPES DE VARIABLES

QUANTITATIVES	QUALITATIVES
continues	ordinales
discrètes	nominales
temporelles	binaires

ORGANISATION DES DONNÉES

Après avoir recueilli des données dans une étude statistique, l'étape suivante consiste à décrire l'ensemble des données sous une forme synthétique.

Par définition, une étude statistique recueille des données portant sur une série de sujets. Chacun de ces sujets est appelé ***unité statistique***. Chaque observation ou mesure effectuée sur chaque unité statistique est une des valeurs de la variable étudiée.

En général, une étude statistique comprend le recueil de plusieurs variables. On aboutit ainsi à une matrice de données dont chaque ligne représente une unité statistique et chaque colonne une variable.

La première étape de la description consiste à trier les données, à les regrouper et éventuellement à les transformer afin de les visualiser sous une forme condensée qui permet d'appréhender globalement leur **distribution**.

I. TRI DES DONNÉES

Cette opération consiste à organiser de façon cohérente la masse des données d'une variable quantitative ou qualitative ordinale. Le tri consiste à ranger les unités statistiques par ordre croissant ou décroissant des valeurs de la variable. Tous les logiciels de saisie de données possèdent une fonction de tri qui rend cette opération aisée. Lorsqu'on étudie une variable qualitative, les unités statistiques sont regroupées selon les différentes classes de cette variable.

II. REGROUPEMENT EN CLASSES

Lorsqu'on étudie une variable quantitative sur un nombre important d'individus, il est nécessaire de regrouper les données pour les présenter clairement. Cette opération aboutit à transformer une variable quantitative continue en variable quantitative discrète. On appelle ce procédé ***discrétisation*** (exemple 2.1).
On peut également transformer une variable quantitative discrète en variable qualitative ordinale.
On construit une échelle de classification en divisant la série en classes. On définit ainsi des bornes entre lesquelles, on compte les individus. Le choix des bornes est une opération délicate. En effet, ce regroupement fait perdre une certaine quantité d'informations. Les classes sont choisies en fonction de la distribution générale de l'ensemble des valeurs observées et selon des critères de pertinence fixés par l'observateur (exemple 2.2).

On peut ainsi choisir :

- Une échelle par *amplitude*, en divisant les valeurs de la série en intervalles égaux. La variable obtenue est de type quantitatif discret. Dans ce cas, le nombre de sujets par classe est irrégulier.

Exemple 2.1. DISCRÉTISATION D'UNE VARIABLE CONTINUE

La série suivante représente les résultats d'un test ELISA pratiqué chez 60 sujets. Les résultats sont exprimés en densité optique (DO) et triés par ordre croissant.
Tableau brut de données (DO) :

0,058	0,116	0,211	0,228	0,284	0,288	0,289	0,370	0,412	0,476	0,484	0,495	0,495	0,551	0,629
0,655	0,669	0,670	0,683	0,694	0,727	0,729	0,734	0,739	0,750	0,806	0,816	0,824	0,826	0,886
0,941	0,945	0,946	0,957	0,976	0,980	0,999	1,001	1,003	1,004	1,029	1,042	1,069	1,075	1,077
1,077	1,078	1,127	1,131	1,133	1,134	1,134	1,135	1,145	1,161	1,172	1,203	1,215	1,242	1,246

Variable discrétisée: on divise la série en 13 classes avec des intervalles de DO de valeur 0,1, s'échelonnant de 0,0 à 1,3. La ligne du bas (n) représente le nombre d'individus présentant une valeur de DO située entre les bornes de chaque classe.

Classe	1	2	3	4	5	6	7	8	9	10	11	12	13
bornes de DO	0,0 - 0,099	0,1 - 0,199	0,2 - 0,299	0,3 - 0,399	0,4 - 0,499	0,5 - 0,599	0,6 - 0,699	0,7 - 0,799	0,8 - 0,899	0,9 - 0,999	1,0 - 1,099	1,1 - 1,199	1,2 - 1,299
n	1	1	5	1	5	1	6	5	5	7	10	9	4

- Une échelle par *fréquence*, en répartissant la série observée en groupes d'effectifs égaux. Les bornes, sont alors déterminées et les intervalles sont irréguliers. La variable obtenue est de type qualitatif ordinal.
- Une échelle de *convenance*, choisie par l'opérateur en fonction de la pertinence des bornes. Ce choix dépend de ce que l'on veut montrer. La variable obtenue est de type qualitatif ordinal.

Quel que soit le choix effectué, il est important de diviser la série en groupes exclusifs qui ne se chevauchent pas. Il faut notamment définir avec précision dans quelle classe ranger les individus présentant une valeur égale à la valeur d'une borne. Par exemple, si l'on divise une série d'âge de 10 ans en 10 ans, il faut préciser qu'un sujet ayant 10 ans révolus est classé dans la tranche 10-19 ans. Un sujet de 9 ans et 11 mois sera classé dans la tranche 0-9.
Lorsqu'on réalise une étude, il est judicieux d'effectuer les regroupements en classes selon des critères habituellement utilisés par d'autres auteurs afin de pouvoir comparer les résultats.

III. TRANSFORMATION DE VARIABLE

Parfois, les valeurs d'une variable quantitative sont exprimées par des nombres difficiles à manipuler simplement. Parfois aussi elles sont dispersées de façon inhomogène. Elles sont alors difficiles à représenter. Il peut être utile de transformer la valeur brute de la variable x en une nouvelle valeur x' (exemple 2.3).

- $\mathbf{x' = ax}$: lorsque les valeurs brutes sont toutes des multiples d'un nombre de grande ou de faible taille.
- $\mathbf{x' = x + b}$: lorsque certaines valeurs sont négatives et d'autres positives.
- $\mathbf{x' = x - b}$: lorsque toutes les valeurs de la variable sont comprises dans un intervalle de petite ou de grande taille par rapport à leur valeur.
- $\mathbf{x' = ax + b}$: combinaison des deux précédentes.

Exemple 2.2. REGROUPEMENT EN CLASSES

La série suivante représente le poids en kg d'une série de 80 sujets adultes classés en ordre croissant

45	50	55	58	60	63	64	64	65	66	67	67	67	67	68	68	68	68	68	68
70	70	71	71	71	71	72	72	72	72	73	73	73	73	73	73	73	73	73	73
74	74	74	74	74	74	74	74	74	75	75	75	75	76	76	76	76	77	77	77
78	78	79	79	79	79	80	80	80	80	80	81	81	81	82	82	83	84	84	86

Regroupement des données :

- Par échelle d'*amplitude* de 10 kg :

classe de poids (kg)	40-49	50-59	60-69	70-79	80-89
nombre de sujets	1	3	16	46	14

- Par échelle de *fréquence* en quatre groupes comprenant chacun 25 % des sujets :

classe de poids (kg)	45-68	70-73	74-77	78-86
nombre de sujets	20	20	20	20

- Par échelle de *convenance* : faible poids < 65 kg, poids moyen 65-79 kg, poids élevé > 79 kg :

classe de poids (kg)	< 65	65-79	> 79
nombre de sujets	8	58	14

- L'échelle par amplitude a permis de créer une variable quantitative discrète à 5 classes. La distribution montre un pic dans la classe 70-79 kg.
- Le regroupement par échelle de fréquence décrit mieux la distribution, mais a peu d'intérêt démonstratif : les bornes ne représentent aucune valeur pertinente dans le poids d'un individu adulte.
- Le regroupement par une échelle de convenance en trois classes est intéressant si l'on désire simplement connaître les fréquences relatives des faibles poids et des poids élevés dans l'étude.

Dans cet exemple, l'échelle par amplitude est un bon compromis entre la nécessité de décrire correctement la distribution et le choix de bornes pertinentes.

- $\mathbf{x' = 1/x}$: lorsque la valeur brute est une fraction, comme c'est souvent le cas des résultats biologiques exprimés en dilution. On en revient alors à un nombre simple. On appelle parfois titre l'inverse d'une dilution.
- $\mathbf{x' = \log(x)}$: lorsque la distribution de la variable s'étire de façon exponentielle vers une de ses extrémités. Ce type de transformation dite logarithmique est très fréquent en biologie. Elle permet de présenter des résultats sur une même échelle (exemple 2.4).
  Ce type d'échelle est familier : la fameuse échelle de Richter pour les tremblements de terre est une échelle de type logarithmique. Pour chaque augmentation d'un degré de l'échelle, l'énergie libérée est multipliée par 30.
- De très nombreuses autres transformations peuvent être effectuées. $x' = x^2$, $x' = \sqrt{x}$, $x' = e^x$, ainsi que toutes sortes de combinaisons. Quelle que soit la transformation effectuée, il faudra bien sûr en tenir compte en effectuant les calculs sur cette variable et en exprimant les résultats.

Exemple 2.3. TRANSFORMATION D'UNE VARIABLE X EN VARIABLE X'

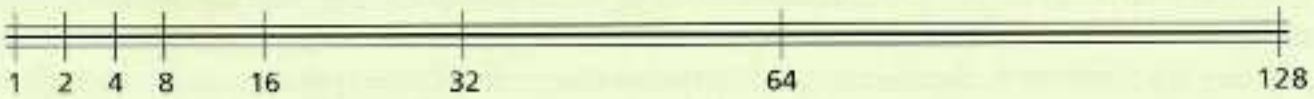
Valeur de x	Transformation	Transformée x'
27 000 à 53 000	$x' = x/1\,000$	27 à 53
0,0015 à 0,0092	$x' = 1\,000\ x$	1,5 à 9,2
10 000 à 10 010	$x' = x - 10\,000$	0 à 10
– 2 à – 1	$x' = x + 2$	0 à 1
– 0,004 à – 0,002	$x' = 1\,000x + 4$	0 à 2
1/2, 1/3, 1/4, 1/8, *etc.*	$x' = 1/x$	2, 3, 4, 8, *etc.*

Exemple 2.4. TRANSFORMATION LOGARITHMIQUE

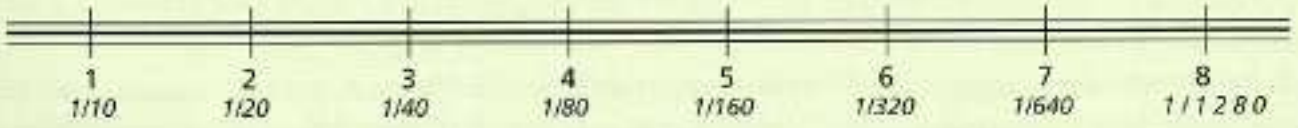
De nombreux résultats biologiques sont souvent exprimés en dilution de raison 2, c'est-à-dire selon une progression dite géométrique.
Le résultat noté est la dernière dilution du prélèvement qui présente une réaction positive au test effectué. Un résultat noté 1/40 signifie que le prélèvement est positif pour cette dilution et pour toutes les dilutions précédentes et qu'il est négatif pour toutes les dilutions suivantes.
Le tableau ci-dessous exprime les résultats (x) d'un sérodiagnostic.

x	$x' = 1/10x$	$x'' = \log_2 x' + 1$
1/10	1	1
1/20	2	2
1/40	4	3
1/80	8	4
1/160	16	5
1/320	32	6
1/640	64	7
1/1 280	128	8

Après la transformation $x' = 1/10x$, la série est simplifiée, mais la distribution est toujours difficile à représenter sur une même échelle.



Après la seconde transformation $x'' = \log_2 x' + 1$, tous les résultats peuvent être présentés sur une même échelle.



Dans cet exemple, les valeurs qu'on étudie sont des facteurs de dilution (1, 2, 3, etc.) plutôt que les valeurs exactes des dilutions (1/10, 1/20, *etc.*). Il est donc logique d'étudier comme variable ce facteur qui se distribue selon une progression arithmétique simple.

IV. EFFECTIFS ET FRÉQUENCES

Lorsqu'une variable est divisée en classes, on obtient après regroupement et dénombrement un nombre de sujets dans chaque classe. C'est l'**effectif** de la classe. Il s'exprime par un nombre entier (n). Ce chiffre n'a de sens que si le total général (N) des effectifs de toutes les classes étudiées est présenté.
Un deuxième indicateur permet de mesurer le poids relatif de l'effectif d'une classe par rapport au total général de la série étudiée. C'est la fréquence.
La **fréquence** d'un effectif (que certains auteurs préfèrent appeler fréquence relative) est le rapport de l'effectif (n) de la classe sur le total (N) de la série étudiée.
La fréquence (n/N) est exprimée le plus souvent en pourcentage (%).
Le total des fréquences de chaque classe est de 100 % (à condition que les classes soient exclusives).
Effectifs et fréquences sont deux indicateurs équivalents (exemple 2.5).

Exemple 2.5. RÉPARTITION DU POIDS EN KG D'UNE SÉRIE DE 80 SUJETS ADULTES

Classe de poids (kg)	< 65	65-79	> 79	Total
Nombre de sujets	8	58	14	80

Les effectifs des 3 classes définies ci-dessus sont respectivement 8, 58 et 14 sujets.
La fréquence de la classe < 65 kg est 8/80, soit 10 %.
La fréquence de la classe 65-79 kg est 58/80, soit 72,5 %.
La fréquence de la classe > 79 kg est 14/80, soit 17,5 %.

Effectifs et fréquences cumulés

Lorsque les classes d'une variable sont ordonnées (variable quantitative discrète ou qualitative ordinale) on peut ajouter à l'effectif de chaque classe le total des effectifs des classes inférieures. On obtient ainsi les **effectifs cumulés**.
Les **fréquences cumulées** sont obtenues en divisant les effectifs cumulés par le total de la série (exemple 2.6).

V. DISTRIBUTION

La **distribution** d'une série de données est constituée par l'ensemble des effectifs répartis entre les classes de la variable étudiée. On apprécie une distribution, en examinant les fréquences des effectifs dans toutes les classes.
Ce terme de distribution sera très utilisé par la suite. Nous verrons en effet, qu'une des bases des analyses statistiques consiste à regarder si une distribution ***observée*** ressemble ou non à des **modèles** de distribution théorique connus. Si la distribution observée « colle » avec un modèle, on peut alors utiliser toutes les propriétés mathématiques du modèle pour étudier la distribution et la comparer à d'autres du même type.

Exemple 2.6. EFFECTIFS ET FRÉQUENCES CUMULÉS

Répartition d'une série de 30 sujets selon la composition de leur fratrie (nombre de frères et sœurs dans la famille)

FRATRIE	EFFECTIFS	FRÉQUENCE %	EFFECTIF CUMULÉ	FRÉQUENCE CUMULÉE (%)
1	11	*36,7*	11	*36,7*
2	6	*20,0*	17	*56,7*
3	5	*16,7*	22	*73,4*
4	4	*13,3*	26	*86,7*
> 4	4	*13,3*	30	*100,0*
Total	30	*100,0*	–	–

On lit dans ce tableau que :
a) 5 sujets appartiennent à une fratrie de 3 personnes ;
b) la fréquence des sujets appartenant à une fratrie de 3 personnes est de 16,7 % ;
c) l'effectif cumulé des sujets appartenant à une fratrie de 3 personnes au plus, est de 22 (11 + 6 + 5) ;
d) la fréquence (cumulée) des sujets appartenant à une fratrie de 3 personnes au plus, est de 73,4 % (36,7 % + 20,0 % + 16,7 %).

Exercice

Dans le cadre de la prévention du paludisme transfusionnel, on a examiné 121 sérums de sujets suspects de paludisme. La technique de dépistage s'exprime en dilution. La dilution au 1/8 est le seuil de détection et la dilution au 1/512 correspond à un paludisme évolutif, Les résultats figurent dans le tableau ci-dessous :

Dilution	n
1/2	4
1/4	5
1/8	8
1/16	22
1/32	25
1/64	16
1/128	11
1/256	7
1/512	9
1/1024	6
1/2048	5
1/4096	3

1) Calculer en pourcentage, les fréquences relatives des sujets pour chaque dilution.
2) Transformer les dilutions en variable arithmétique simple.
3) Calculer les fréquences de la distribution regroupée en :
   - 3 classes de fréquences relatives équivalentes ;
   - en classes d'amplitude égale à 2 dilutions ;
   - en 3 classes (non significatif, suspect, évolutif).

DESCRIPTION DES DONNÉES

Il existe trois procédés pour décrire un ensemble de données statistiques ou une distribution : les tableaux, les diagrammes et le calcul de paramètres simples résumant à eux seuls l'ensemble de la distribution.

I. TABLEAUX

1. Tableau brut de données

Le tableau brut est le tableau élémentaire de travail. Toutes les données y figurent, unité par unité et variable par variable. Les individus ou unités statistiques sont en ligne, les variables en colonnes.
Un tel tableau comprend en général deux sortes de variables : les variables permettant d'identifier chaque unité statistique et les variables mesurées pour l'étude.
Les variables d'identification, peu nombreuses, permettent de retrouver un individu donné : nom, prénom, adresse, numéro d'identification anonyme, numéro d'ordre dans l'étude, numéro de prélèvement, *etc.* Par définition, ces variables ne se prêtent pas à des regroupements, et elles disparaissent donc des tableaux agrégés.
Dans les études en pathologie humaine et en épidémiologie, la saisie informatique de ces données d'identification est encadrée par une législation stricte supervisée par la Commission nationale informatique et liberté (CNIL).

Un tableau brut est rarement présentable tel quel, à moins que les données soient peu nombreuses comme dans le tableau ci-dessous (exemple 3.1).

Exemple 3.1. TABLEAU BRUT DE DONNÉES STATISTIQUES

A	B	C	D	E	F	G	H
N°	Identification	Sexe	Date de naissance	Taille en cm	Nationalité	Couleur des yeux	Niveau d'études
1	Antonin	M	30/06/1978	185	GB	bleu	supérieur
2	Aurélien	M	24/04/1965	170	F	marron	primaire
3	Hadrien	M	25/02/1956	163	F	marron	secondaire
4	Peggy	F	08/04/1977	177	GB	bleu	supérieur
5	Julien	M	12/03/1982	162	F	noir	supérieur
6	Marie	F	?	175	F	marron	supérieur
7	Émilie	F	30/12/1981	165	F	marron	*refus réponse*
8	Julie	F	25/11/1978	*350*	F	noir	secondaire
9	Steve	M	23/05/1974	182	IRL	marron	primaire
10	Marco	M	12/01/1978	178	E	noir	secondaire

Ce tableau brut comporte (volontairement) quelques données aberrantes ou manquantes.

2. Tableaux de fréquences

Ils servent à présenter un ensemble de données sous forme **agrégée**. Un tableau est par définition une matrice comportant au moins deux entrées, l'une horizontale (lignes), l'autre verticale (colonnes). L'une des entrées est constituée par les classes de la variable.
La seconde entrée est constituée par les effectifs des sujets dans chaque classe de la variable étudiée ou par leurs fréquences. Bien que l'information soit redondante, on peut rendre service au lecteur d'un tableau en y faisant figurer à la fois effectifs et fréquences.
Un tableau correct doit présenter le total des effectifs de la série étudiée et le total des fréquences pour bien montrer que les classes sont exclusives (exemple 3.2).

Exemple 3.2.

VARIABLE : SEXE	EFFECTIF n*	FRÉQUENCE %
Hommes	28 420	*48,6*
Femmes	30 101	*51,4*
Total	58 521	*100,0*

Population française : recensement de 1999. Source INSEE 2000 Répartition par sexe.

* en milliers d'individus.

Un tableau peut être enrichi d'une deuxième entrée portant sur une deuxième variable. Ce type de tableau plus touffu, est plus difficile à construire. Les pourcentages qui y figurent peuvent se rapporter soit au total des lignes, soit au total des colonnes, soit au total général. Il vaut mieux utiliser une police de caractère différente pour exprimer les pourcentages. Il est impératif dans ce type de tableaux de faire figurer le total des pourcentages utilisés afin de faire comprendre au lecteur, quelle distribution est étudiée (exemple 3.3).

Exemple 3.3.

Population française : recensement de 1999. Source INSEE 2000
Répartition par âge et par sexe. (Effectifs donnés en milliers d'individus)

Tableau 1

SEXE / ÂGE	HOMMES n	HOMMES %	FEMMES n	FEMMES %	TOTAL n	TOTAL %
0-19	7 355	*25,9*	7 013	*23,3*	14 368	*24,5*
10-39	8 236	*29,0*	8 248	*27,4*	16 484	*28,2*
40-59	7 554	*26,6*	7 646	*25,4*	15 200	*26,0*
60-74	3 661	*12,9*	4 304	*14,3*	7 965	*13,6*
> 74	1 614	*5,6*	2 890	*9,6*	4 504	*7,7*
Total	28 420	*100,0*	30 101	*100,0*	58 521	*100,0*

Tableau 2

SEXE / ÂGE	HOMMES n	HOMMES %	FEMMES n	FEMMES %	TOTAL n	TOTAL %
0-19	7 355	*51,2*	7 013	*48,8*	14 368	*100,0*
10-39	8 236	*50,0*	8 248	*50,0*	16 484	*100,0*
40-59	7 554	*49,7*	7 646	*50,3*	15 200	*100,0*
60-74	3 661	*46,0*	4 304	*54,0*	7 965	*100,0*
> 74	1 614	*35,8*	2 890	*64,2*	4 504	*100,0*
Total	28 420	*48,6*	30 101	*51,4*	58 521	*100,0*

Exemple 3.3. (*Suite*)

Tableau 3

SEXE / ÂGE	HOMMES n	HOMMES %	FEMMES n	FEMMES %	TOTAL N	TOTAL %
0-19	7 355	*12,6*	7 013	*12,0*	14 368	*24,6*
10-39	8 236	*14,0*	8 248	*14,1*	16 484	*28,1*
40-59	7 554	*12,9*	7 646	*13,1*	15 200	*26,0*
60-74	3 661	*6,3*	4 304	*7,3*	7 965	*13,6*
> 74	1 614	*2,8*	2 890	*4,9*	4 504	*7,7*
Total	28 420	*48,6*	30 101	*51,4*	58 521	*100,0*

Les trois tableaux ci-dessus étudient la même distribution d'effectifs. Cependant, les pourcentages qui figurent dans les colonnes ne sont pas identiques.

- Le tableau 1 représente la fréquence relative des classes d'âge, pour chaque sexe. Les effectifs sont rapportés au total du sexe étudié. On y lit que la classe d'âge la plus fréquente chez les hommes est celle de 10 à 39 ans. C'est également la classe la plus représentée chez les femmes.
- Le tableau 2 représente la fréquence relative des sexes, à l'intérieur de chaque classe d'âge. Les effectifs sont rapportés au total de la classe d'âge étudiée. On y lit que dans la classe des plus de 74 ans, les femmes sont presque 2 fois plus nombreuses.
- Le tableau 3 représente la fréquence relative de chaque classe d'âge et de sexe. Les effectifs sont rapportés au total général de la population. Cette présentation n'est pas très parlante.

On constate sur ces 3 exemples l'importance de représenter le total des pourcentages en lignes et en colonnes, surtout si les effectifs ne figurent pas sur le tableau.
Au-delà de deux variables, un tableau devient illisible. Il vaut mieux dans ce cas représenter les données sur une série de plusieurs tableaux de structure identique.

3. Problème des données manquantes

Ce type de données est inhérent à toute étude de grande ampleur. Il faut évidemment s'efforcer de les éviter. Il existe deux sortes de données perturbant une analyse statistique : les données manifestement aberrantes qui sont dues à des erreurs de mesures, de recopiage ou de saisie. Les données manquantes qui sont dues à des refus de réponse, à des mesures non pratiquées ou à des oublis lors de la saisie.

Dans tous les cas, il faut :

- Tenter de récupérer le maximum de données manquantes.
- Effectuer si possible une double saisie par deux opérateurs différents afin de détecter les erreurs de saisie. Certains logiciels de saisie permettent de détecter automatiquement les discordances.
- Prévoir un code spécial pour les données aberrantes, afin qu'elles ne soient pas comptabilisées dans les calculs.
- Prévoir un code spécial pour les données manquantes. De nombreux logiciels le font automatiquement.
- Prévoir une règle de décision sur les données manquantes ou aberrantes. La règle la plus rigoureuse (et la plus honnête) consiste à exprimer ces données dans les tableaux, y compris dans les tableaux agrégés. Les résultats y gagneront toujours en crédibilité. Les fréquences et autres paramètres seront alors calculés en prenant comme dénominateur le sous-total des données disponibles (exemple 3.4).

Exemple 3.4. ÉTUDE DU POIDS DES INDIVIDUS SUR UNE SÉRIE DE 85 SUJETS

		N	% (1)	% (2)
Poids en kg	< 65	8	*9,4*	*10,0*
	65 - 79	58	*68,2*	*72,5*
	> 79	14	*16,5*	*17,5*
Total données disponibles		80	*94,1*	*100,0*
Données manquantes		5	*5,9*	
Total des individus		85	*100,0*	

1 : % du total des individus ; 2 : % du total des données disponibles.

Les pourcentages exprimés dans la colonne 1, laissent entendre que la distribution des poids est strictement identique entre les données disponibles et les données manquantes. L'interprétation correcte des données de cette étude est de signaler qu'elle porte sur 94,1 % des individus prévus, et que parmi ceux-ci, la distribution des poids est celle qui figure sur la colonne 2. On ne fait ainsi aucune hypothèse sur les données manquantes, qui sont inconnues.

II. GRAPHIQUES

Les graphiques sont les images des études statistiques, comme les tableaux en étaient l'écriture. Les tableaux avaient pour objet de présenter des données de façon exacte. Les graphiques ont pour objet de faire ressortir au lecteur une vision synthétique du phénomène étudié. Les tableaux sont précis, mais c'est souvent au prix d'une abondance de nombres qui rendent leur lecture incommode. Les graphiques à l'inverse illustrent une tendance générale, ils donnent une image globale des résultats de l'étude.
Tout graphique est le résultat d'un choix par son concepteur. Il faut accepter une nouvelle perte d'information au bénéfice de la clarté. Toute la difficulté réside dans l'opposition entre cette nécessaire clarté et le maintien d'une rigueur scientifique. Un graphique doit être le plus simple possible, et n'utiliser que le minimum de moyens. Il faut bannir tous les gadgets bureautiques qui ne sont pas directement utiles.

- Ainsi, nous déconseillons :
  - l'utilisation de graphes 3D qui n'apportent aucun élément supplémentaire à la démonstration et brouillent le message ;
  - les superpositions de multiples graphes sur un même graphique ; des graphiques juxtaposés sont plus clairs à analyser ;
  - la colorisation abusive. Un dégradé de gris ou un simple camaïeu sont plus facile à lire qu'un manteau d'arlequin.
- En résumé, un graphique doit être :
  - simple ;
  - compréhensible par lui-même ;
  - légendé (titre, axes, unités) ;
  - honnête.

Nous nous limiterons aux graphiques visant à décrire des distributions. D'autres graphiques illustrant des tendances temporelles de paramètres ou d'indicateurs épidémiologiques seront vus dans les chapitres correspondants.
Un graphique de distribution a pour but de décrire des effectifs ou des fréquences en fonction d'une variable. Le graphique élémentaire comporte donc en ordonnée les effectifs ou les fréquences, et, en abscisses les valeurs de la variable.

1. Polygone de fréquence

Ce graphe linéaire est adapté à la représentation de la distribution d'une variable quantitative continue. En ordonnée figurent les effectifs ou les fréquences (figure 3-1). En abscisse figurent les valeurs de la variable quantitative discrétisée. Chaque point du polygone représente l'effectif ou la fréquence pour le point central de la classe de la variable. Le trait reliant deux points suggère les effectifs ou les fréquences possibles entre deux valeurs centrales. Les deux extrémités du polygone de fréquence doivent rejoindre l'axe des abscisses. La surface comprise sous le polygone représente 100 % des observations.
L'intérêt principal de ce type de graphe est de pouvoir représenter sur un même graphique plusieurs distributions.

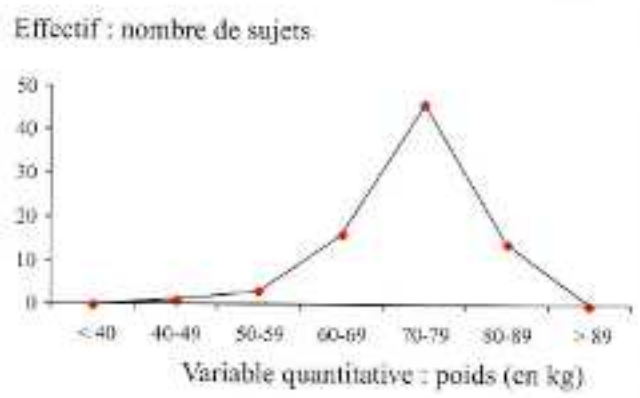


Figure 3-1. Polygone de fréquence
Distribution du poids chez 80 sujets adultes

2. Histogramme

L'histogramme est le graphe adapté pour représenter la distribution d'une variable quantitative discrète.
Il est aussi utilisé pour les variables continues. En effet, il est souvent difficile, voire impossible de représenter la distribution d'une variable continue, puisqu'il existe une infinité de valeurs possibles. Si par exemple on désire représenter la distribution de la taille de 100 sujets au mm près, il y a de grandes chances d'obtenir au maximum un seul individu pour chaque observation. Il est donc nécessaire de « discrétiser » la variable.
Un histogramme est un diagramme constitué de barres verticales contiguës. Les ordonnées représentent les effectifs de la distribution. Les abscisses représentent les classes de la variable. L'échelle des abscisses désigne :

- soit les montants de la barre s'il s'agit d'une variable continue regroupée en classes (figure 3-2). Dans ce cas les abscisses désignent les bornes des classes, la largeur de la barre représente l'intervalle de la classe (son amplitude) ;
- soit le centre de la barre s'il s'agit d'une variable discrète (figure 3-3).

Dans un histogramme vrai, la surface de chaque barre est proportionnelle à la fréquence relative des effectifs (figure 3-2). La surface totale des barres représente 100 % des observations. Si (et seulement si) les classes sont d'amplitude égale, les ordonnées peuvent représenter les fréquences relatives de la distribution.
Le choix du nombre de classes est une opération délicate. Trop de classes aboutit à un effet « colonnes de Buren », sans effet démonstratif. Trop peu de classes aboutit à un effet « podium olympique » en gommant les détails (figure 3-4).
Il est malaisé de représenter par histogramme plusieurs distributions sur un même graphique (barres de couleurs différentes, superposées ou juxtaposées). Cela est possible lorsque l'une des deux distributions présente des effectifs constamment inférieurs à l'autre.

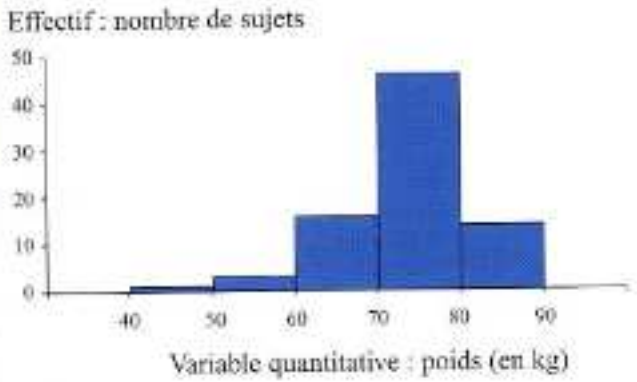


Figure 3-2. Histogramme
Distribution du poids chez 80 sujets adultes

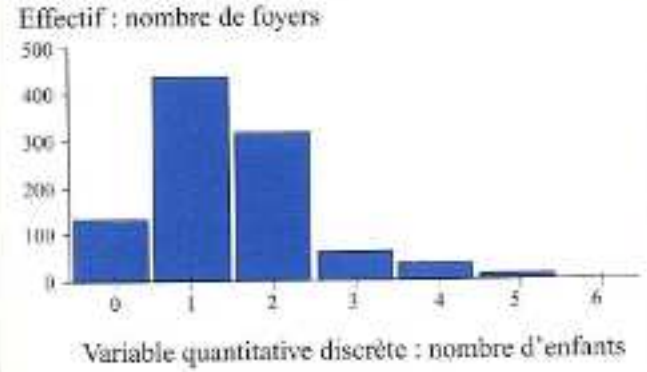


Figure 3-3. Histogramme
Distribution du nombre d'enfants par foyer dans une population de 1 000 foyers

Dans les autres cas, pour éviter une représentation confuse, il est conseillé :

- soit de réaliser plusieurs histogrammes empilés utilisant les mêmes échelles ;
- soit d'utiliser les polygones de fréquences dans lesquels chaque point correspond au milieu de chaque barre.

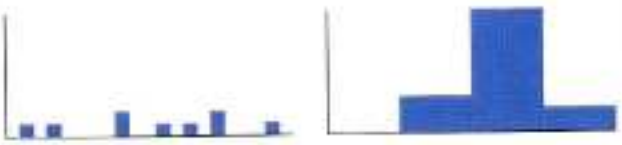


Figure 3-4.

3. Diagramme en barres

Ce type de graphe est adapté à la représentation d'une variable qualitative ordinale. Il peut aussi être utilisé pour représenter une variable qualitative nominale.
Il est constitué de barres verticales disjointes pour montrer qu'il ne s'agit pas d'une variable continue (figure 3-5). Les ordonnées représentent les effectifs ou les fréquences de la distribution. La barre des abscisses, sans échelle numérique, n'est là que pour servir de base au graphe. La largeur des barres, identique, n'est choisie que sur des critères visuels. Les libellés de chaque barre sont les modalités de la variable.
Dans le cas d'une variable ordinale, les libellés doivent être ordonnés par valeurs croissantes.

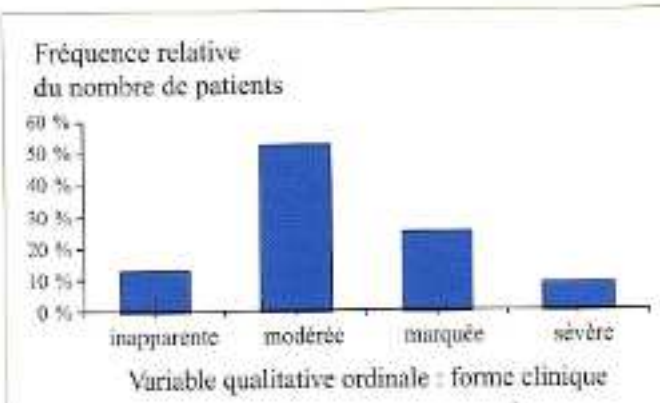


Figure 3-5. Diagramme en barres
Distribution des formes cliniques d'une maladie dans une série de 941 malades

Dans le cas d'une variable nominale, l'ordre importe peu et dépend de l'effet recherché ; il peut alors être pertinent d'ordonner les barres par taille croissante ou décroissante.

4. Diagramme en barres horizontales

Ce type de graphe est adapté aux variables qualitatives nominales.

À l'inverse du diagramme précédent, les effectifs ou les fréquences sont portés sur l'axe horizontal et les libellés des classes sur l'axe vertical (figure 3-6). Cela permet d'utiliser des libellés de grande taille. L'axe horizontal comporte une échelle numérique. Les barres sont de largeur égale, choisie sur des critères visuels. L'ordre de classes importe peu. Il sera volontiers choisi par ordre croissant ou décroissant de taille pour montrer la tendance.
La superposition de plusieurs distributions est possible en juxtaposant les barres de tonalité différente pour chaque valeur de la variable.

5. Camembert

Ce type de graphe (*pie chart* en anglais) est adapté à la représentation d'une seule distribution d'une variable qualitative nominale.
Le camembert est un cercle divisé en secteurs. Chaque secteur représente une classe de la variable. La surface du secteur est proportionnelle à la fréquence de l'effectif de la classe (figure 3-7).
Le camembert peut être choisi à la place d'un diagramme en barres à condition que le nombre de classes soit faible. Un maximum de six classes nous paraît raisonnable. Ce type de graphe est peu précis. Il doit être réservé pour montrer un effet relatif dans une distribution, par exemple un ou deux secteurs prédominants vis-à-vis d'autres secteurs moindres.

Figure 3-6. Diagramme en barres horizontales
Distribution des demandes de sérodiagnostic de toxoplasmose en fonction du type de patient

Figure 3-7. Camembert
Distribution des sérodiagnostics de la toxoplasmose dans un laboratoire en fonction du type de patient

6. Pyramide

Ce type de graphe particulier est utilisé pour montrer la distribution par âge et par sexe d'une population (figure 3-8). C'est l'équivalent d'un double histogramme inversé et juxtaposé.
Sur l'axe horizontal figure les effectifs, les hommes d'un coté, les femmes de l'autre. L'axe vertical médian comporte les classes d'âge étudiées.
Ce type de graphique est utilisé en démographie. Il permet d'appréhender d'un coup d'œil la structure d'une population et de la comparer à d'autres.

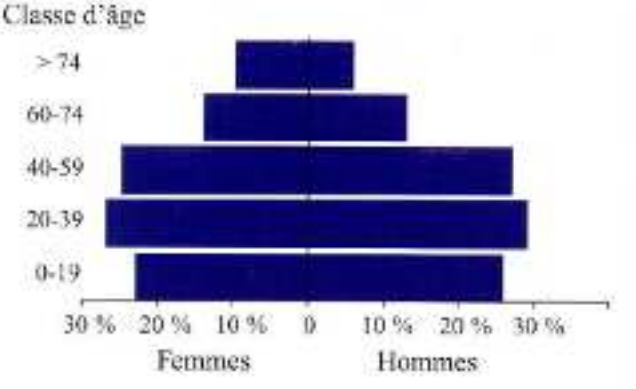


Figure 3-8. Pyramide des âges
Population française, 1999

MESURES EN STATISTIQUE

Nous avons vu dans les chapitres précédents comment synthétiser un tableau brut de données par le regroupement en classes et l'expression de distributions au moyen de tableaux de données agrégées et de diagrammes.

La difficulté réside maintenant dans la nécessité de résumer ces données afin de les exprimer et éventuellement de les comparer à d'autres données du même type provenant d'une série différente.

Il existe des méthodes permettant de résumer en quelques nombres l'ensemble d'une distribution. Ces nombres sont appelés des **paramètres**.

Il existe deux types de paramètres : les paramètres de **position** et les paramètres de **dispersion** (figure 4-1).

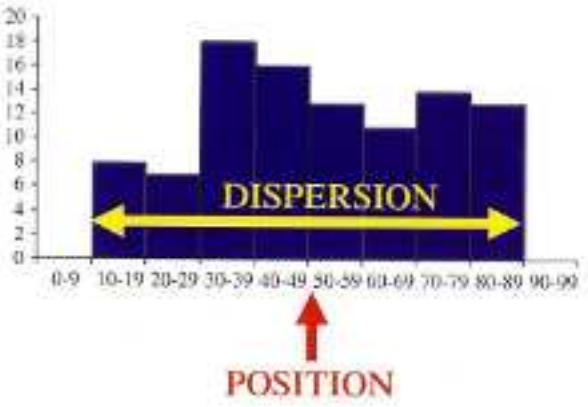


Figure 4-1. Paramètres

Dans tous les cas, il faut être conscient que l'utilisation de ces paramètres se fait au prix d'une perte d'information. Quelle que soit l'utilisation statistique plus ou moins sophistiquée qu'on en fera, il faudra toujours conserver l'ensemble des données et savoir retourner au tableau brut lorsqu'on voudra « faire parler » les résultats d'une étude.

I. PARAMÈTRES DE POSITION

Ils permettent de résumer en quelques valeurs la position d'une distribution en fonction des valeurs possibles de la variable étudiée.

1. Médiane

La médiane est un paramètre de tendance centrale qui sert à résumer une série de données d'une variable quantitative.

Définition : la médiane est la valeur qui partage la série des individus en deux groupes d'effectifs égaux (figure 4-2). Ainsi, la moitié

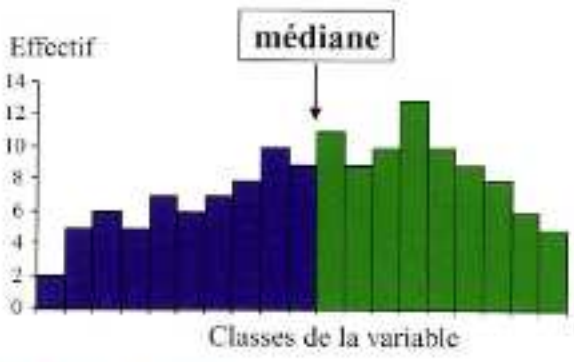


Figure 4-2. Médiane

des sujets présente une valeur inférieure à la médiane, l'autre moitié présente une valeur supérieure à la médiane.

Calcul : il s'agit en fait d'une estimation. Elle nécessite de classer les sujets de l'étude par ordre de valeur croissante de la variable. Si le nombre de sujets est impair, la médiane de la série est la valeur de la variable observée chez le sujet médian. Si le nombre de sujet est pair, la médiane est située entre les deux valeurs qui partage la série en deux ; dans ce cas, en pratique, on prend la moyenne des deux valeurs centrales (exemple 4.1).

Exemple 4.1. ESTIMATION DE LA MÉDIANE

La série suivante représente le poids en kg d'une série de 80 sujets.

86	76	80	81	74	73	72	79	55	66	50	73	73	68	67	74	73	67	71	79
74	74	77	74	71	80	72	74	77	75	71	73	75	76	76	77	71	68	65	73
72	80	76	72	60	63	70	75	79	81	58	68	67	68	68	68	70	80	82	73
64	74	75	64	73	67	73	80	82	84	45	83	84	74	74	78	81	77	73	79

Après classement par ordre croissant, on obtient :

45	50	55	58	60	63	64	64	65	66	67	67	67	67	68	68	68	68	68	68
70	70	71	71	71	71	72	72	72	72	73	73	73	73	73	73	73	73	73	73
74	74	74	74	74	74	74	74	74	75	75	75	75	76	76	76	76	77	77	77
77	78	79	79	79	79	80	80	80	80	80	81	81	81	82	82	83	84	84	86

La médiane qui partage la série en deux groupes de taille identique a pour valeur (73 + 74) / 2 = 73,5 kg.

Excel® : fonction MEDIANE (86 ; 76 ; 80.....73 ; 79) = 73,5.

Propriétés : la médiane est un paramètre essentiellement descriptif. Elle ne nécessite pas de connaître la totalité des valeurs ; on peut en effet la calculer en ne connaissant pas les valeurs extrêmes ; il suffit de connaître le nombre de sujets inférieurs et supérieurs à deux bornes extrêmes. Dans l'exemple 4.1, on aurait pu calculer la même médiane en sachant que 20 sujets avaient un poids inférieur à 70 kg et 9 sujets un poids supérieur à 80 kg.
Utilisée en complément avec les extrêmes et l'étendue (*cf.* IV.2) elle permet de résumer une distribution. En revanche, elle ne se prête pas aux calculs statistiques usuels permettant d'estimer des paramètres dans une population ou de comparer des distributions entre elles.

2. Quartiles

Les quartiles sont les 3 valeurs qui partagent la distribution en quatre (figure 4-3).

- Le premier quartile est la valeur qui partage, d'un côté de la distribution, un quart des valeurs les plus faibles et de l'autre 75 % des valeurs les plus élevées.
- Le deuxième quartile est la médiane.
- Le troisième quartile est la valeur qui partage, d'un côté de la distribution, 75 % des valeurs les plus faibles et de l'autre un quart des valeurs les plus élevées (exemple 4.2).

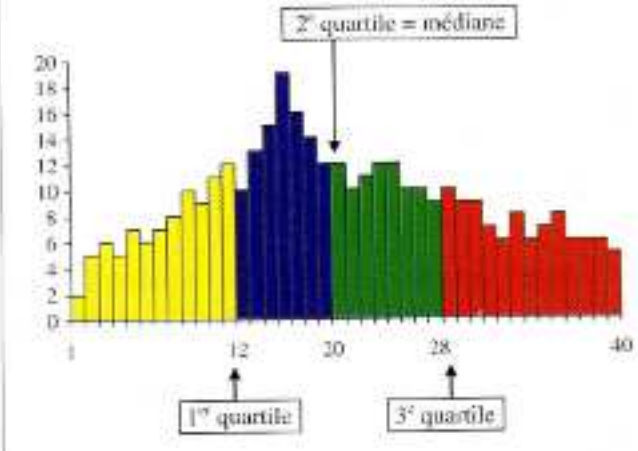


Figure 4-3. Quartiles

Exemple 4.2. ESTIMATION DES QUARTILES

En reprenant les données de l'exemple 4.1 :

45	50	55	58	60	63	64	64	65	66	67	67	67	67	68	68	68	68	68	68
70	70	71	71	71	71	72	72	72	72	73	73	73	73	73	73	73	73	73	73
74	74	74	74	74	74	74	74	74	75	75	75	75	76	76	76	76	77	77	77
77	78	79	79	79	79	80	80	80	80	80	81	81	81	82	82	83	84	84	86

Le premier quartile, qui partage la distribution en 1/4 et 3/4 est la valeur 69 kg.
Le troisième quartile qui partage la distribution en 3/4 et 1/4 est la valeur 77 kg.

Excel® : fonction QUARTILE.

3. Déciles et percentiles

Les déciles sont les 9 valeurs qui partagent la distribution en 10 groupes de tailles égales (figure 4-4). Chaque groupe comprend 10 % des effectifs. De la même manière, les percentiles sont les valeurs qui partagent la distribution en 100 groupes de taille égale.

Le percentile 10 % (ou 1er décile) est la valeur qui partage d'un côté de la distribution, 10 % des valeurs les plus faibles, et de l'autre 90 % des valeurs les plus élevées.

Le percentile 25 % est le 1er quartile.

Le percentile 50 % est la médiane.

Le percentile 90 % (ou 9e décile) est la valeur qui partage, d'un côté de la distribution, 90 % des valeurs les plus faibles, et de l'autre 10 % des valeurs les plus élevées (exemple 4.3).

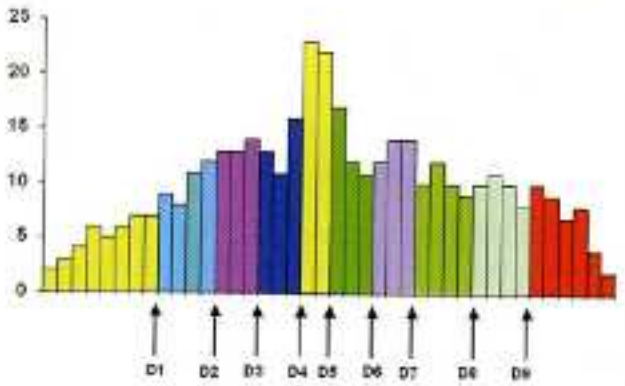


Figure 4-4. Déciles

Exemple 4.3. ESTIMATION DES PERCENTILES

En reprenant les données de l'exemple 4.1, la distribution peut être partagée en 20 groupes de 4 sujets ; chaque groupe représente 5 % de la distribution.

45	50	55	58	60	63	64	64	65	66	67	67	67	67	68	68	68	68	68	68
70	70	71	71	71	71	72	72	72	72	73	73	73	73	73	73	73	73	73	73
74	74	74	74	74	74	74	74	74	75	75	75	75	76	76	76	76	77	77	77
77	78	79	79	79	79	80	80	80	80	80	81	81	81	82	82	83	84	84	86

Le percentile 5 % est la valeur 59 kg qui sépare 5 % des sujets de faible poids de 95 % des sujets de poids plus élevé.
Le percentile 25 % est la valeur 69 kg (premier quartile).
Le percentile 95 % est la valeur 82,5 kg.
Le percentile 97,5 % est la valeur 84 kg.

Excel® : fonction CENTILE.

4. Mode

Définition : dans une distribution comportant de nombreuses données, le mode est la valeur qui revient le plus souvent. On l'appelle souvent pic de la distribution.

Propriétés : le mode est un paramètre purement descriptif. Il n'est utilisé que pour définir l'allure générale de la distribution. Le mode n'est pas utilisé dans les calculs statistiques. Lorsqu'il n'existe qu'un seul mode avec un pic très accentué, on dit que la distribution est *unimodale*. Il peut exister un deuxième pic de valeurs dans une autre partie de la distribution. On dit alors qu'elle est *bimodale* (figure 4-5). Cette situation se produit notamment lorsqu'on recueille des données correspondant à deux groupes de sujets différents, comportant par exemple des sujets sains et malades (exemple 4.4).

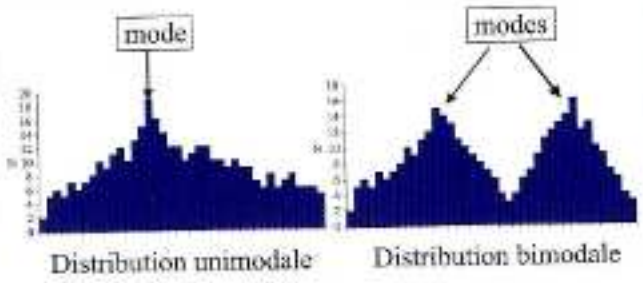


Figure 4-5. Mode

Exemple 4.4. ESTIMATION DU MODE

En reprenant les données de l'exemple 4.1, le mode est la valeur 73 kg qui est la valeur la plus fréquente. Elle représente 12,5 % des valeurs (10 sur 80).

Excel® : fonction MODE (86 ; 76 ; 80... 73 ; 79) = 73.

5. Moyenne

Comme la médiane, la moyenne est un paramètre de tendance centrale qui sert à résumer une série de données d'une variable quantitative (exemple 4.5).

Définition : la moyenne est une valeur calculée résultant de la somme algébrique des valeurs observées dans la série, divisée par le nombre de sujets.

Calcul : si l'on appelle $\mathbf{x}$ les différentes valeurs de la variable quantitative étudiée, $\mathbf{\Sigma x}$ la somme de ces valeurs et N le nombre de sujets, la moyenne $\boldsymbol{\mu}$ d'une série est égale à

$$\mu = \frac{\Sigma x}{N}$$

Propriétés

Contrairement à la médiane, le calcul de la moyenne nécessite d'utiliser toutes les données.
La moyenne possède des propriétés arithmétiques permettant d'effectuer des estimations et des comparaisons.

Exemple 4.5. CALCUL D'UNE MOYENNE

En reprenant les données de l'exemple 4.1, on obtient :

$\Sigma x = 5824$, $N = 80$ $\mu = 5824/80 = 72{,}8$

Le poids moyen de cette série est de 72,8 kg.

Excel® : fonction MOYENNE (86 ; 76 ; 80... 73 ; 79) = 72,8.

La moyenne est sensible au poids des valeurs extrêmes. Si les valeurs sont dispersées de façon homogène autour d'une valeur centrale, la moyenne est un bon indicateur de la distribution. Si au contraire, il existe des valeurs très élevées (ou très basses) à l'une des extrémités de la distribution, ces valeurs pèsent de façon importante sur la valeur moyenne. Le nombre obtenu sera un mauvais indice de valeur centrale ; dans ce cas il sera plus adapté soit d'utiliser la médiane pour décrire une distribution, soit de transformer la variable. Il est donc nécessaire, avant de calculer une moyenne de considérer l'aspect général de la distribution des valeurs à l'aide d'un histogramme.

Calcul de la moyenne sur des variables transformées

Nous avons examiné les procédures de transformations de variables au chapitre 2.III.
La moyenne μ' calculée sur une variable transformée x' peut elle-même subir la transformation inverse pour obtenir la moyenne μ exprimée dans les mêmes unités que la variable x initiale.

Transformée	Moyenne
$x' = ax$	$\mu = \mu'/a$
$x' = x/a$	$\mu = a\mu'$
$x' = x + b$	$\mu = \mu' - b$
$x' = x - b$	$\mu = \mu' + b$
$x' = ax + b$	$\mu = (\mu' - b)/a$
$x' = \log_a(x)$	$\mu_g = a^{\mu'}$
$x' = \ln(x)$	$\mu_g = e^{\mu'}$

Moyenne géométrique

Lorsque la variable x a subi une transformation non arithmétique tel que : $x' = \log_a(x)$, la moyenne μ' est la moyenne des logarithmes :

$$\mu' = \frac{\Sigma \log_a(x)}{n}$$

On peut revenir aux unités initiales en calculant la ***moyenne géométrique*** μ_g

$$\mu_g = a^{\mu'}$$

On remarque que la moyenne géométrique calculée après une transformation logarithmique n'est pas identique à la moyenne qui aurait été calculée sur les valeurs brutes de la variable. Le choix dépend de l'allure de la distribution initiale. Si les valeurs de la variable d'origine suivent une progression géométrique, c'est la moyenne géométrique qui a un sens (exemple 4.6).
On peut calculer directement une moyenne géométrique μ_g d'une série de N valeurs en posant :

$$\mu_g = (x_1 x_2 \ldots x_n)^{1/N}$$

Le log de zéro n'est pas calculable par définition. Si une valeur $x = 0$, il faut décider une règle : soit supprimer cette valeur de la série et ne faire une analyse que sur des données supérieures à 0, soit décider que $\log(x) = 0$. Ceci revient à transformer la valeur initiale 0 en valeur 1.

Exemple 4.6. CALCUL D'UNE MOYENNE GÉOMÉTRIQUE

Soit la série de valeurs suivantes et les effectifs correspondants. On note que les valeurs de la variable augmentent selon une raison 2.

Valeurs x	2	4	8	16	32
$x' = \log_2(x)$	1	2	3	4	5
Effectif	2	2	2	2	2

Les valeurs **x'** sont obtenues en prenant le logarithme de base 2 (ce qui est logique puisque la progression des valeurs **x** est de raison 2).
La moyenne **μ'** des valeurs de **x'** est égale à (1 + 1 + 2 + 2 + 3 + 3 + 4 + 4 + 5 + 5)/10 = 3
La moyenne géométrique $\mu_g = 2^{\mu'} = 2^3 = 8$. Cette moyenne correspond bien à la valeur centrale de la distribution (médiane = 8).

Si l'on avait calculé la moyenne arithmétique de la distribution, celle-ci aurait été de (2 + 2 + 4 + 4 + 8 + 8 + 16 + 16 + 32 + 32)/10 = 12,4. Ce nombre ne représente pas la moyenne réelle de la distribution. Ceci vient du fait que la dispersion arithmétique des valeurs est fortement asymétrique. La moyenne arithmétique n'était donc pas un bon indicateur de cette distribution.
Puisque la distribution est à l'évidence une distribution à progression géométrique, il est logique d'étudier le logarithme des valeurs et de calculer une moyenne géométrique.

Excel® : fonction MOYENNE.GEOMETRIQUE (2 ; 4 ; 8 ; 16 ; 32) = 8.

6. Fréquence relative

Lorsqu'une variable est de nature qualitative ou lorsqu'une variable quantitative a été divisée en classes, les paramètres de position précédents ne sont pas utilisables. Le moyen le plus simple pour résumer une distribution est de calculer les fréquences relatives de sujets porteurs de chaque modalité de la variable. Ces proportions, comprises entre 0 et 1, s'expriment habituellement en pourcentage (%) en les multipliant par 100.

Calcul : si **N** est le nombre total de sujets d'une distribution et $\mathbf{n_i}$ le nombre de sujets présentant la modalité **i** de la variable étudiée, le pourcentage $\mathbf{P_i}$ de sujets présentant cette modalité est :

$$P_i = \frac{n_i}{N}$$

Propriétés

Si les modalités de la variable quantitative sont exclusives, la somme des pourcentages obtenus est égale à 100 % (exemple 4.7).

Exemple 4.7. DISTRIBUTION DE LA CATÉGORIE SOCIO-PROFESSIONNELLE D'UN ÉCHANTILLON DE 13 459 PARTURIENTES. FRANCE, JANVIER 1995

PROFESSION	N	P_i (%)
Sans profession	2 337	*17,4*
Agriculteur	96	*0,7*
Commerçant-artisan	304	*2,2*
Cadre supérieur	967	*7,2*
Profession intermédiaire	2 046	*15,2*
Employé	4 587	*34,1*
Ouvrier, service	2 662	*19,8*
Inconnu	460	*3,4*
Total N	13 459	*100,0*

7. Pourcentage

Lorsqu'une variable est de type qualitatif binaire, on peut la considérer comme une variable dite de Bernouilli, prenant la valeur 1 pour l'une des modalités et la valeur zéro pour l'autre. On peut, par exemple, attribuer la valeur 1 si la caractéristique étudiée est présente et 0 si elle est absente.
On a ainsi transformé une variable qualitative binaire en variable quantitative.
La moyenne ($\Sigma x/N$) n'est autre que la proportion **P** des **n** sujets possédant la valeur 1.

$$P = \frac{n}{N}$$

Cette proportion, comprise entre 0 et 1 s'exprime habituellement par un pourcentage (%) en la multipliant par 100.

II. PARAMÈTRES DE DISPERSION

Les paramètres de position d'une variable quantitative résument en un seul nombre la distribution. Ce nombre n'est pas suffisant. Les trois distributions de la figure 4-6, ont toutes le même paramètre central. Elles sont pourtant différentes par leur étalement. Il nous manque un paramètre qui résumerait la dispersion des valeurs autour de la valeur centrale.

1. Extrêmes

Ce sont les deux valeurs extrêmes de la distribution, valeurs minimum et maximum (Excel® : fonction MIN et MAX). Les extrêmes donnent une idée brute de la dispersion de la distribution de part et d'autre de la médiane.

2. Étendue

C'est la *différence* (*range* en anglais) entre les deux valeurs extrêmes. L'étendue donne en un seul chiffre une idée de la distribution autour de la médiane. Ce paramètre est utile si les valeurs extrêmes ne s'éloignent pas trop des valeurs voisines. Si les deux valeurs extrêmes sont aberrantes, l'étendue donne une fausse idée de la dispersion.

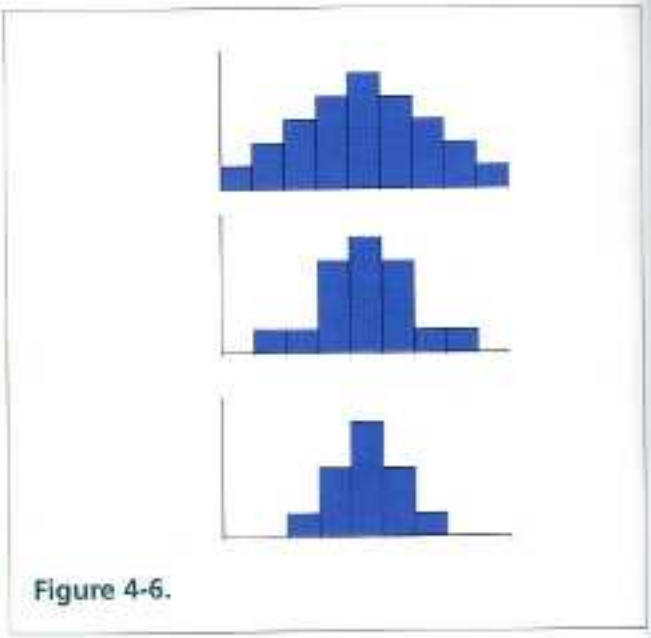


Figure 4-6.

3. Intervalle interquartile et semi-interquartile

L'intervalle interquartile est la différence entre les valeurs du premier et du troisième quartile (chap. 4.1.2). Cet intervalle est de même nature que l'étendue, mais s'affranchit des valeurs extrêmes. C'est donc un meilleur paramètre de dispersion.
L'intervalle semi-interquartile est la moitié de l'intervalle interquartile. Il offre une valeur plus adaptée de la dispersion lorsque la distribution est dissymétrique.

Les extrêmes, l'étendue et l'intervalle interquartile sont des paramètres de dispersion qui sont associés à la médiane lorsqu'on désire décrire simplement une distribution d'une variable quantitative. Mais ils ne permettent pas d'effectuer des estimations ou des comparaisons statistiques (exemple 4.8).

Exemple 4.8. PARAMÈTRES DE DISPERSION

En reprenant les données de l'exemple 4.1, la distribution a les paramètres de dispersion suivants :

- Extrêmes : 45 et 86 kg.
- Étendue : 86 – 45 = 41 kg.
- Intervalle interquartile : 77 – 69 = 8 kg.
- Intervalle semi-interquartile : 8/2 = 4 kg.

4. Variance

C'est le paramètre de dispersion le plus utilisé. Son principe est de résumer l'ensemble des écarts de chaque valeur d'une distribution par rapport à la moyenne.

Définition : la variance d'une distribution est la moyenne des carrés des écarts à la moyenne de chacune des valeurs. De façon plus exacte, on l'appelle aussi ***écart quadratique moyen***.

Calcul : si l'on appelle **x** chaque valeur de la distribution d'une variable quantitative, **μ** la moyenne, et N le nombre de sujets, la variance σ^2 est :

$$\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$$

On remarque que le dénominateur ici est **N**, car on considère la série étudiée comme une population exhaustive et non comme un échantillon. Nous verrons au chapitre 8.II une formule légèrement différente avec le terme **n – 1** au dénominateur utilisée lorsqu'on « estime » une variance inconnue en travaillant sur un échantillon de taille **n**.

Calcul pratique de la variance

Avec une simple calculette, il est fastidieux de calculer les carrés de chaque écart. La formule suivante est strictement analogue, mais plus pratique à utiliser (exemple 4.9) :

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}$$

Exemple 4.9. CALCUL D'UNE VARIANCE

En reprenant les données de l'exemple 4.1,

On a $\Sigma x^2 = 428\,126$ $\quad \Sigma x = 5\,824$, $N = 80$ $\quad \sigma^2 = \dfrac{428\,126 - \frac{(5\,824)^2}{80}}{80} = 51{,}7$

Excel® : fonction VAR.P (86 ; 76 ; 80... 73 ; 79) = 51,7.

Calcul de la variance sur des variables transformées

Nous avons étudié les procédures de transformations de variables au chapitre 2.III.
La variance σ'^2 calculée sur une variable transformée **x'**, peut elle-même subir la transformation inverse pour obtenir la variance σ^2 de la variable **x** initiale. La transformation de la variance σ'^2 calculée sur une variable transformée x' s'effectue de la façon suivante :

TRANSFORMÉE	VARIANCE
$x' = ax$	$\sigma^2 = \sigma'^2/a^2$
$x' = x/a$	$\sigma^2 = a^2\sigma'^2$
$x' = x + b$	$\sigma^2 = \sigma'^2$
$x' = x - b$	$\sigma^2 = \sigma'^2$
$x' = ax + b$	$\sigma^2 = \sigma'^2/a^2$

Lorsqu'une variable x a subi une transformation plus complexe, on ne manipule la variance que sur les données transformées.

Propriétés : comme pour le calcul de la moyenne, celui de la variance nécessite d'utiliser toutes les valeurs de la distribution. La variance est le meilleur indicateur de la dispersion d'une variable autour de sa moyenne. Plus la variance est faible, plus la distribution est resserrée. Plus la variance est élevée, plus la distribution est étalée.
Son inconvénient est de s'exprimer par une unité élevée au carré, qui n'a pas le même ordre de grandeur que les valeurs de la distribution.
Afin d'utiliser un paramètre de dispersion plus explicite, on utilise sa racine carrée, qu'on appelle écart type.

5. Écart type

L'écart type (*standard deviation* en anglais) est la racine carrée de la variance.

Calcul :

$$\sigma = \sqrt{\sigma^2} \quad \text{ou de façon équivalente} \quad \sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$$

Propriétés : de même que la variance, l'écart type est un indicateur de dispersion extrêmement utilisé. Plus il est élevé, plus la dispersion est élevée. Plus il est faible, plus la dispersion est resserrée. Il a en outre l'avantage de s'exprimer dans les mêmes unités que la moyenne (exemple 4.10).

Exemple 4.10. CALCUL DE L'ÉCART TYPE

En reprenant les données de l'exemple 4.9,

$\sigma^2 = 51{,}7$ $\qquad \sigma = \sqrt{51{,}7} = 7{,}2$ kg

Excel® : fonction ECARTYPEP (86 ; 76 ; 80... 73 ; 79) = 7,2.

6. Coefficient de variation

Le coefficient de variation est un indicateur combinant la moyenne et l'écart type.

Définition : on appelle coefficient de variation (CV), le rapport de l'écart type sur la moyenne.

Calcul :

$$CV = \frac{\sigma}{\mu} \times 100$$

L'unité du CV est un nombre sans dimension. Il est exprimé en pourcentage.

Propriété : le coefficient de variation exprime le degré de dispersion d'une distribution en fonction de la valeur moyenne. Il est utile pour comparer la dispersion de deux variables quantitatives de nature différente (unités différentes) (exemple 4.11).

Exemple 4.11. CALCUL D'UN COEFFICIENT DE VARIATION

En reprenant les données des exemples 4.5 et 4.10, on a :

$\mu = 72{,}8$ kg et $\sigma = 7{,}2$ kg

$CV = \frac{7{,}2}{72{,}8} \times 100 = 9{,}9\ \%$

Si dans la même série on avait observé une taille moyenne de 165 cm avec un écart type de 16 cm, on aurait $CV = \frac{16}{165} \times 100 = 9{,}7\ \%$. Bien que les deux valeurs mesurées, poids et taille, soient de nature différente, on aurait pu tout de même observer que la dispersion de ces deux variables autour de leur moyenne respective était semblable.

7. Variance et écart type d'une variable qualitative binaire

Nous avons vu (chap. 1.II.3) qu'une variable qualitative binaire peut être considérée comme une variable de Bernouilli (0 et 1), analogue à une variable quantitative. Sa moyenne s'exprime par le pourcentage P. On peut en calculer aisément la variance σ^2 et l'écart type σ.

$$\sigma^2 = P(1-P) \qquad \sigma = \sqrt{\sigma^2} = \sqrt{P(1-P)}$$

On note que pour P = 50 %, on a $\sigma^2 = 0{,}25$.
Pour P = 10 % et P = 90 %, $\sigma^2 = 0{,}09$.

On constate donc que la variance d'une variable binaire est maxima pour un pourcentage de 50 % (figure 4-7).

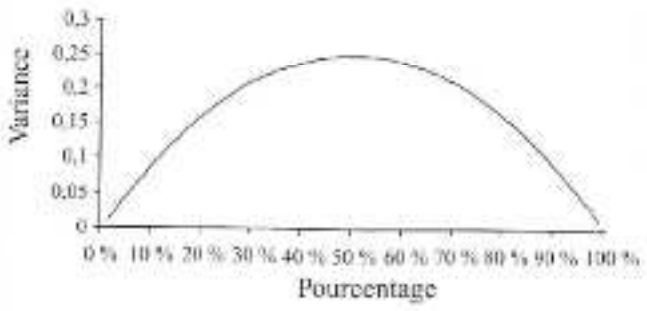


Figure 4-7. Variance d'une variable binaire en fonction du pourcentage

Exercices

Exercice 4.1

On a noté le poids d'une série de 11 nouveau-nés.

poids (g)	3 250	3 482	3 122	3 498	3 743	3 854	3 359	2 985	3 043	3 634	3 507

Estimez la médiane.

Calculez la moyenne.

Exercice 4.2

Estimez la médiane de la série suivante :

poids (g)	2512	2876	2956	3128	3359	3482	3546	3678	3720	3987

Exercice 4.3

Soit la série de valeurs suivante (n = 53) :

9,7	5,8	11,9	16,1	15,7	17,9	2,2	10	15,3	6,6	8,2	4,2
3,6	7	7,9	2,5	8,7	9,3	11,5	9,5	9,6	9,5	16,3	10,6
10,2	8,9	18,8	14,4	20,5	8,3	17,6	4,5	13,1	14,6	18,6	10,6
8,9	13,7	9,4	14	5,2	7,6	4,9	9,5	6,8	10,8	11,1	9,7
19,7	4	8	0,6	16,7							

Calculez :

- la médiane ;
- le premier quartile ;
- le troisième quartile ;
- le mode ;
- l'étendue ;
- l'espace inter-quartile ;
- la moyenne ;
- la variance ;
- l'écart type ;
- le coefficient de variation.

Résumé

PARAMÈTRES DE POSITION ET DE DISPERSION

VARIABLE QUANTITATIVE

Paramètres de position

médiane	valeur qui partage la distribution en deux parts égales
premier quartile	valeur qui partage la distribution en 1/4 et 3/4
troisième quartile	valeur qui partage la distribution en 3/4 et 1/4
mode(s)	valeur(s) plus fréquemment observée(s)
déciles	valeurs qui partagent la distribution en 10 parts égales
percentiles	valeurs qui partagent la distribution en 100 parts égales
moyenne	somme des valeurs divisée par le nombre des observations $\mu = \frac{\Sigma x}{N}$

Paramètres de dispersion

extrêmes	les deux valeurs basses et hautes d'une distribution
étendue	distance entre les deux extrêmes
intervalle interquartile	distance entre 1er et 3e quartile
variance	moyenne de la somme des carrés des écarts à la moyenne : $\sigma^2 = \frac{\Sigma(x-\mu)^2}{N}$
écart type	racine carrée de la variance : $\sigma = \sqrt{\sigma^2}$
coefficient de variation	rapport entre l'écart type et la moyenne : $CV = \frac{\sigma}{\mu} \times 100$

VARIABLE QUALITATIVE BINAIRE

pourcentage	proportion d'une des deux modalités multipliée par 100 (%) : $P = \frac{n}{N} \times 100$
variance	produit de la proportion par son complément à 1 : $\sigma^2 = P(1-P)$
écart type	racine carrée de la variance : $\sigma = \sqrt{P(1-P)}$

Médiane, quartiles, percentiles, mode, extrêmes, étendue, intervalles interquartiles sont des paramètres descriptifs (figure 4-8). On peut les résumer de façon concise par un « diagramme en moustaches » (figure 4-9).

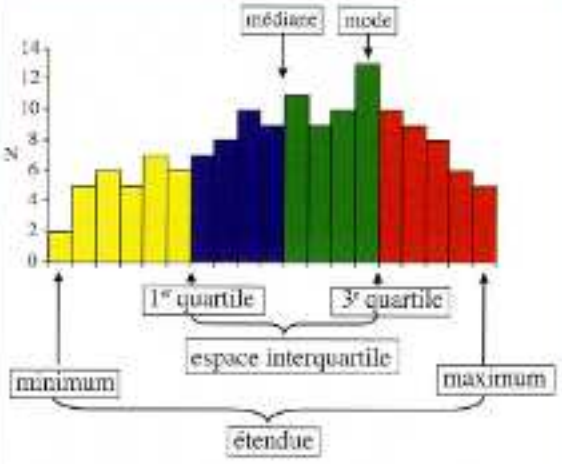


Figure 4-8.

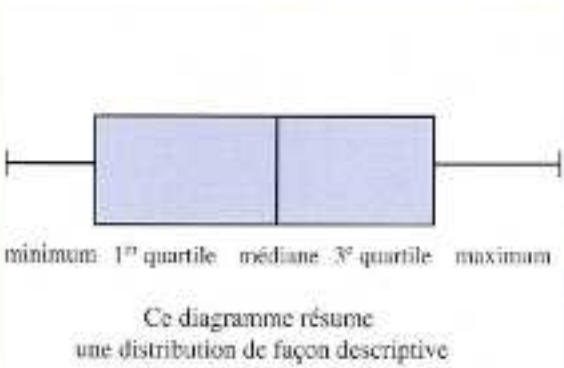


Figure 4-9. Diagramme en moustaches ou *box plot*

Moyenne, variance, écart type, coefficient de variation sont des paramètres arithmétiques (figure 4-10).

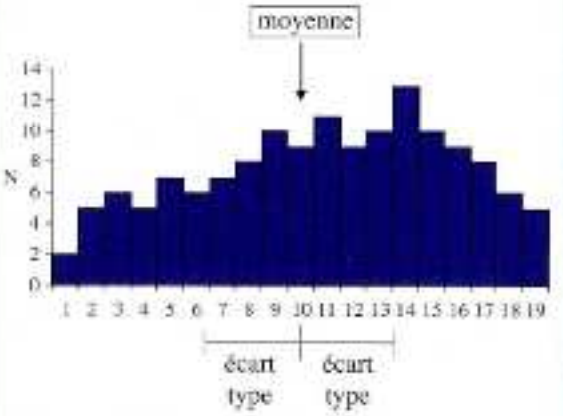


Figure 4-10.

5

REPRÉSENTATION D'UNE DISTRIBUTION

Lorsque les valeurs d'une variable ont été recueillies, ordonnées et classées, on examine leur distribution, c'est-à-dire la répartition des fréquences des individus pour chaque classe. Une des premières phases d'une analyse consiste à décrire cette distribution.

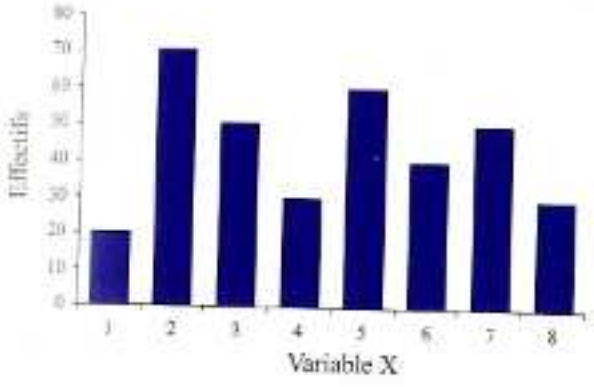


Figure 5-1. Les valeurs de la distribution se répartissent un peu n'importe comment. Il est difficile de leur superposer un modèle simple.

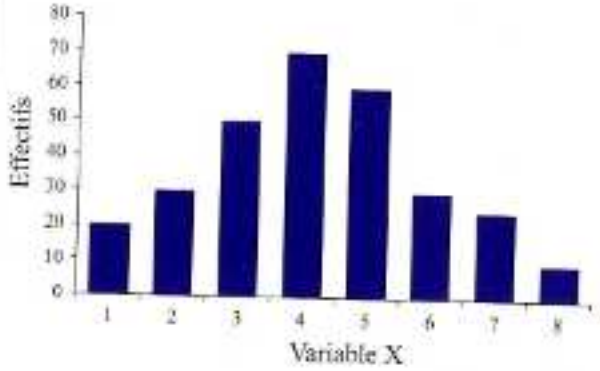


Figure 5-2. Les valeurs de la distribution se répartissent selon un modèle fréquemment observé en biologie. La plus grande partie des valeurs se trouvent regroupées autour d'une valeur centrale. Puis, au fur et à mesure qu'on s'éloigne de cette valeur centrale, les fréquences diminuent de façon à peu près symétrique.

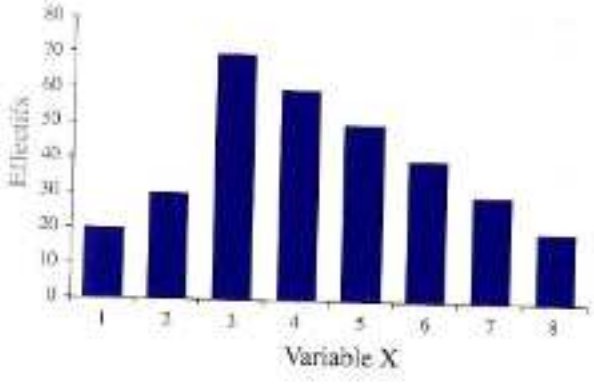


Figure 5-3. La dispersion de la distribution est asymétrique. Elle est plus importante pour les valeurs élevées que pour les valeurs basses. Ce type de distribution est aussi très souvent rencontré en biologie. Une transformation adéquate de la variable X permet souvent de revenir à la distribution 2, plus facile à analyser.

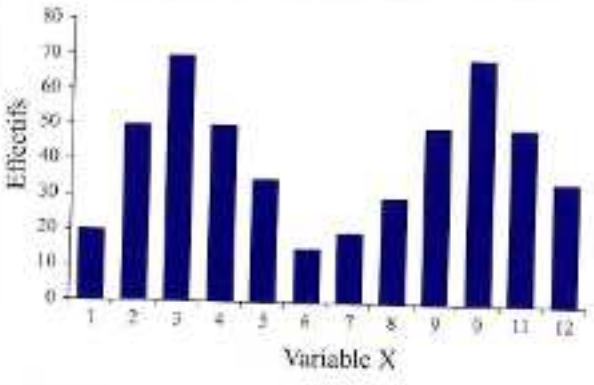


Figure 5-4. Les valeurs de la distribution sont réparties en deux sous-populations. On observe deux pics. Cette distribution est appelée bimodale. Elle s'observe très souvent en médecine, lorsque la valeur d'une mesure se distribue de façon différente parmi les sujets sains et les sujets malades.

I. VARIABLE DISCRÈTE : FRÉQUENCES RELATIVES DES CLASSES

Si, en ordonnées, on remplace les effectifs n_i de chaque classe par leurs rapports au total des effectifs n_i/N, on obtient le diagramme des fréquences relatives de chaque classe p(x). Le graphe a une allure identique. Mais maintenant la somme des barres qui représente la somme de toutes les fréquences relatives de chaque classe est égale à 1 (100 %).
On peut ainsi lire sur la figure 5-5 que 9,7 % des valeurs de X sont égales à 6.
On peut comparer maintenant la distribution observée de notre variable ***statistique*** à des modèles théoriques de distribution d'une variable ***aléatoire***. Ces modèles sont des fonctions mathématiques qui définissent quelle est la probabilité de chaque valeur de X.
Si on place à présent en ordonnées, non pas la fréquence relative de chaque classe, mais la somme des fréquences relatives de toutes les classes ***inférieures*** à x, on obtient le diagramme en escalier des fréquences cumulées (figure 5-6). Les valeurs des classes sont représentées par le montant de la marche. Chaque plateau représente l'intervalle entre deux classes. La valeur lue pour un plateau est la fréquence de toutes les valeurs inférieures à la valeur du montant à droite.
On peut ainsi lire sur la figure 5-6 que 37 % des valeurs de X sont inférieures à 7.
On peut comparer maintenant la distribution cumulée observée de notre variable ***statistique*** à des modèles théoriques de distribution cumulée d'une variable ***aléatoire***. Ces modèles, nommés ***fonction de répartition***, sont des fonctions mathématiques qui définissent quelle est la probabilité que chaque valeur de X soit inférieure à une valeur donnée x.

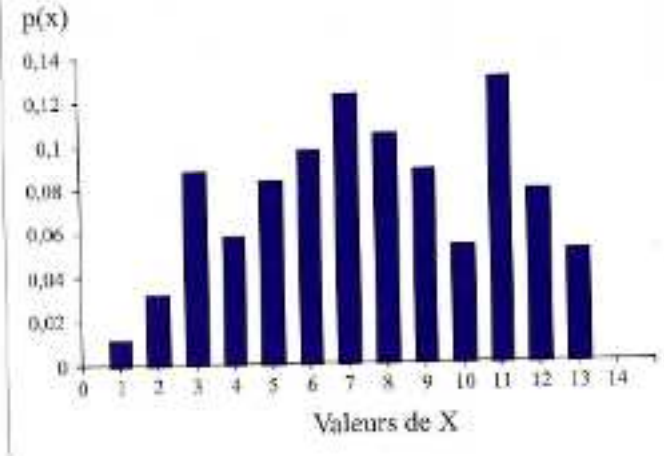


Figure 5-5. Distribution des fréquences relatives d'une variable X

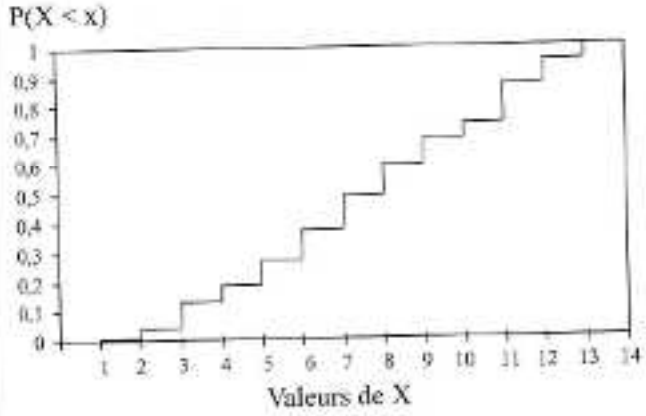


Figure 5-6. Diagramme des fréquences cumulées d'une variable X

II. VARIABLE CONTINUE : DENSITÉ DE PROBABILITÉ

Nous avons vu que le mode de représentation adéquat d'une variable continue regroupée en classes était l'histogramme (chap. 3.II.2). Nous avons accepté une certaine perte d'information par ce regroupement en classes (figure 5-7).
Si on désirait affiner l'observation, il serait possible de diminuer l'amplitude des classes, donc d'augmenter leur nombre. En conséquence, l'effectif de chaque classe ainsi que leur fréquence relative diminueraient également. Cela se traduirait graphiquement par une diminution de la hauteur et de la largeur des barres (figure 5-8). En revanche, la somme des fréquences relatives, donc de l'aire

totale de l'histogramme, serait évidemment toujours égale à 1 (100 % des individus sont toujours compris entre les valeurs extrêmes). Si l'on poursuivait l'opération indéfiniment on aboutirait à un graphe illisible. On observerait tout au plus une densité de points plus élevée pour les valeurs correspondant aux pics initiaux de la distribution.

Une manière d'y remédier consiste à porter en ordonnées, non pas la fréquence relative de chaque classe p(x), mais le rapport $f(x) = \frac{p(x)}{a}$, où **a** représente l'amplitude de la classe (largeur du rectangle). L'aire de chaque rectangle représente toujours la fréquence relative de la classe et f(x) est appelé **densité de fréquence relative** (figure 5-9).

Lorsque la taille de la classe diminue, le rapport f(x) demeure à une échelle constante. Lorsque l'amplitude **a** de la classe tend vers zéro, la fréquence relative p(x) tend aussi vers zéro, mais le rapport f(x) tend vers une valeur limite, appelée **densité de probabilité**.

Quelle que soit la taille de la classe, la distribution est toujours représentée à la même échelle grâce à cette transformation (figure 5-10).

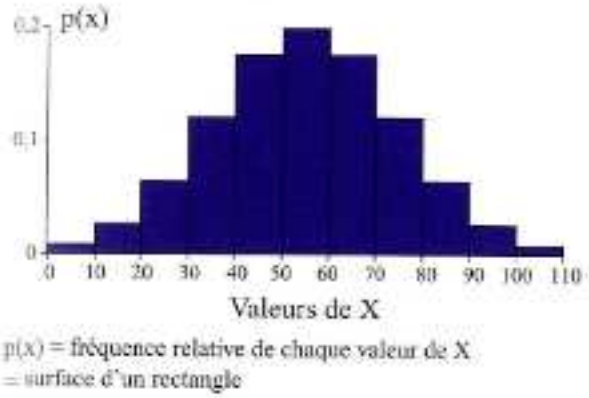


Figure 5-7. Distribution des fréquences relatives d'une variable continue X

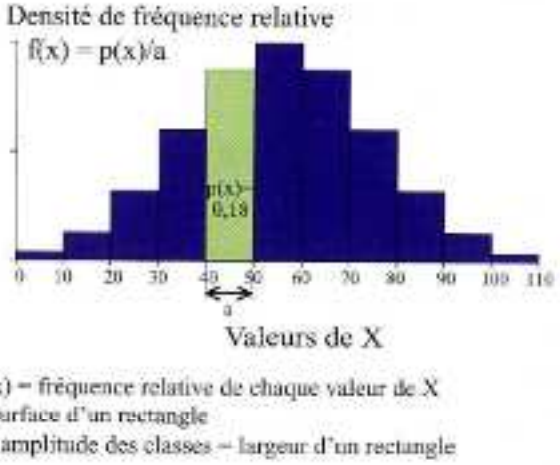


Figure 5-9. Courbe de la densité des fréquences relatives d'une variable continue X

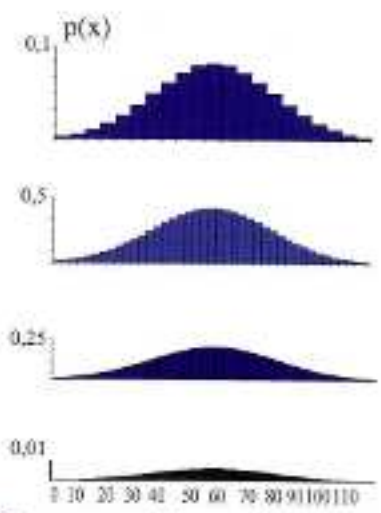


Figure 5-8.

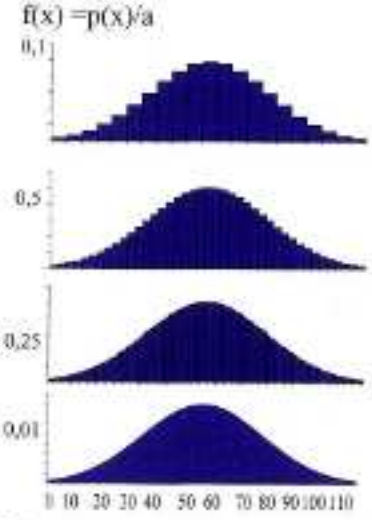


Figure 5-10.

On aboutit finalement au modèle théorique de la distribution de la variable continue. On l'appelle **loi de distribution** de la variable. Dans certains cas simples, on peut en calculer une expression mathématique qui est la fonction de densité de probabilité **f(x)** et en tirer des propriétés intéressantes (figure 5-11).

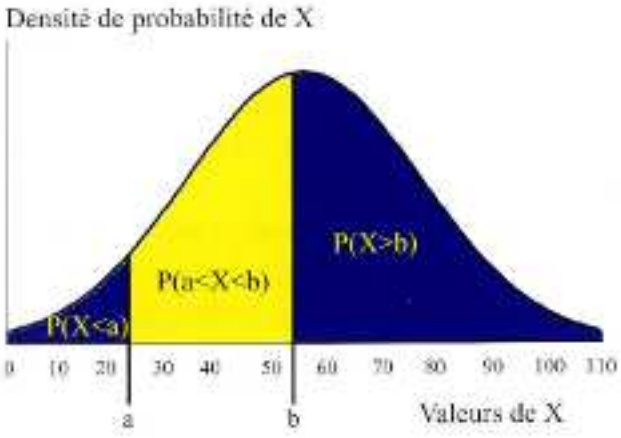


Figure 5-11. Courbe de la densité de probabilité d'une variable continue X

- L'aire contenue sous la courbe entre deux valeurs **a** et **b**, représente la probabilité que X soit comprise entre **a** et **b**.
- L'aire contenue sous la courbe avant une valeur **a**, représente la probabilité que X soit inférieure à **a**.
- L'aire contenue sous la courbe après une valeur **b**, représente la probabilité que X soit supérieure à **b**.
- L'aire sous la courbe représente la somme totale de toutes les probabilités de chaque valeur de la variable X. Elle est égale à 1 (100 %).

III. SYMÉTRIE ET ÉTALEMENT D'UNE DISTRIBUTION

Les distributions observées dans le domaine de la médecine et de la biologie sont souvent disposées de façon à peu près symétrique autour des valeurs centrales (moyenne et médiane) et se caractérisent par un certain degré de dispersion déterminant le degré d'étalement de la courbe. Symétrie et étalement d'une courbe de distribution peuvent être appréciés respectivement par le calcul des coefficients de dissymétrie et d'aplatissement. Ces coefficients sont produits par la plupart des logiciels statistiques dans les paramètres descriptifs d'une distribution et il est donc utile de connaître leur signification.

1. Coefficient de dissymétrie
γ_1 ou *skewness*

Il peut être simplement calculé avec Excel® grâce à la fonction : COEFFICIENT.ASYMETRIE (x_1 ; x_2 ; ..x_n). Voir formules de calcul en Annexes, Formulaire 24.

- Lorsque γ_1 est égal à zéro, la distribution est symétrique (figure 5.12).
- Lorsque γ_1 est supérieur à zéro, la distribution est étalée vers la droite : il y a plus de sujets présentant des valeurs élevées.
- Lorsque γ_1 est inférieur à zéro, la distribution est étalée vers la gauche : il y a plus de sujets présentant des valeurs basses.

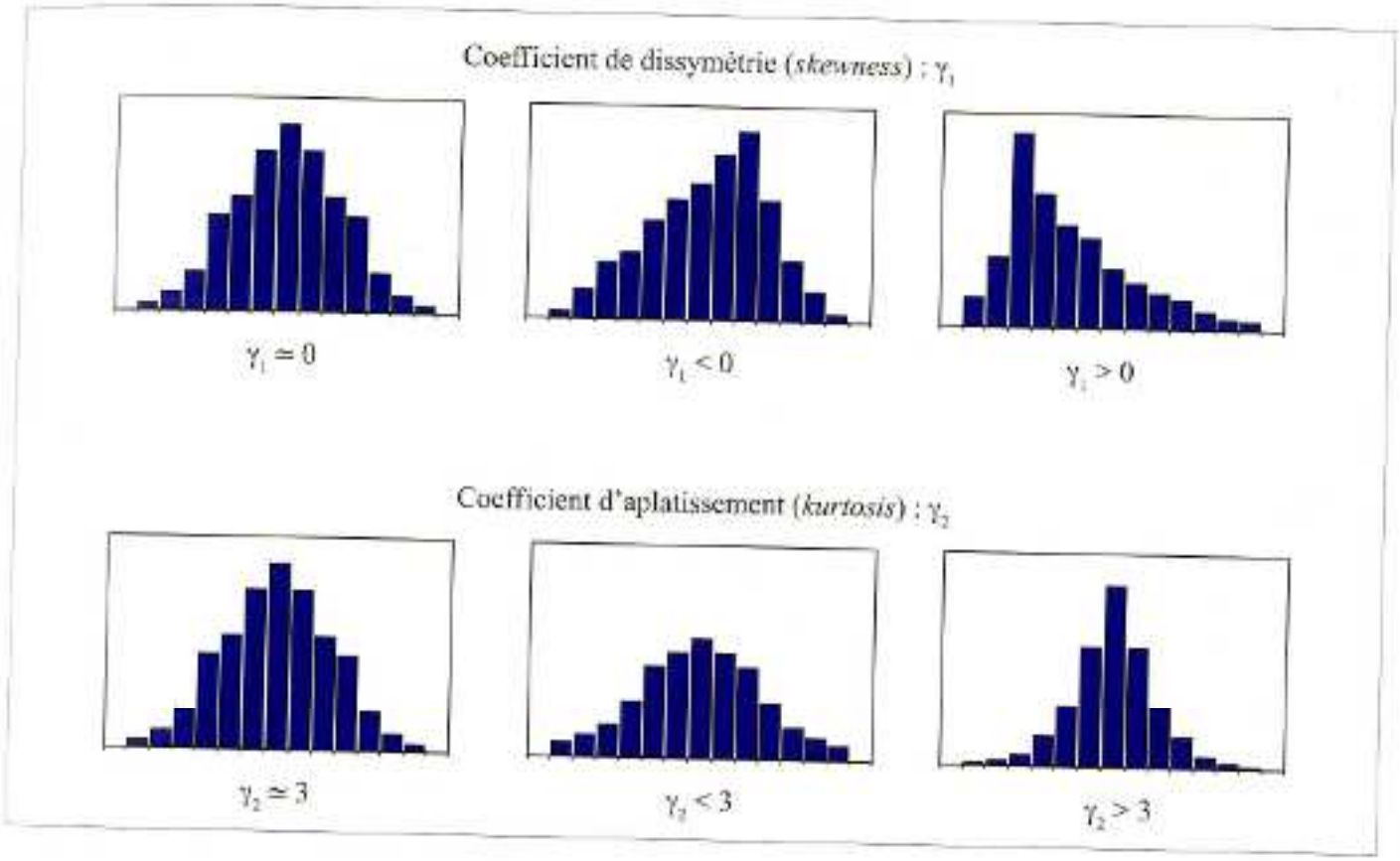
2. Coefficient d'aplatissement
γ_2 ou *kurtosis*

Ce coefficient est surtout utilisé pour apprécier si une distribution suit la loi dite « normale ». Cette loi très utilisée en biologie et médecine, est étudiée au chapitre suivant (6.III).

- Lorsque γ_2 est égal à 3, la distribution est normale (figure 5.12).

- Lorsque γ_2 est supérieur à 3, la distribution est pointue : il y a de nombreux sujets présentant des valeurs proches de la moyenne et peu de sujets à valeurs extrêmes.
- Lorsque γ_2 est inférieur à 3, la distribution est aplatie : la dispersion des valeurs est large, il y a de nombreux sujets présentant des valeurs extrêmes et peu de sujets autour de la moyenne.

Attention : certains logiciels appellent *kurtosis*, l'*excès* d'aplatissement en retranchant la valeur 3 au calcul (*kurtosis excess*). Dans ce cas, l'interprétation se fait par rapport à la valeur zéro : courbe pointue si $\gamma_2 > 0$ et aplatie si $\gamma_2 < 0$. C'est le cas de Excel® avec la fonction : KURTOSIS (x_1 ; x_2 ; ..x_n). *Cf.* formules de calcul en Annexes formulaire 25.



IV. CAS D'UNE VARIABLE QUALITATIVE BINAIRE

La distribution se résume à deux effectifs. Les fréquences relatives des deux classes sont complémentaires à 1. On s'intéresse donc la plupart du temps à une seule des deux classes qui est la caractéristique de la variable. La distribution de la variable se résume donc au pourcentage de la caractéristique étudiée.

LOIS DE DISTRIBUTION

Les modèles de distribution les plus importants sont les distributions régies par la loi binomiale, la loi de Poisson et la loi normale.

I. LOI BINOMIALE

Cette loi est née du jeu. La loi binomiale a été inventée par des parieurs qui voulaient prévoir leurs chances de gagner aux dés en sortant des as. En cessant de jouer, ils se sont mis à spéculer.

1. À quoi sert la loi binomiale ?

On utilise la loi binomiale dans deux situations ; lorsqu'on désire connaître :

- la probabilité de **k** succès au bout de **n** tentatives sachant la probabilité **P** de gagner à chacune des tentatives. C'est la situation de jeu de hasard ;
  ou bien, de façon plus pratique,
- la probabilité d'observer **k** individus possédant une caractéristique donnée dans un échantillon de **n** individus tirés d'une population où la proportion **P** de la caractéristique est connue (exemple 6.1).

Exemple 6.1.

Voici deux questions faisant appel à la loi binomiale :

1) Quelle est la probabilité de tirer 3 as en jetant 10 fois un dé ?
2) Quelle est la probabilité d'observer 3 malades dans un échantillon de 10 sujets choisis au hasard dans une population où la fréquence de la maladie est de 17 % ?

En pratique biologique et médicale on utilise la loi binomiale pour étudier la distribution d'une variable qualitative binaire dans un échantillon de sujets. La plus utilisée en épidémiologie est la variable malade/non-malade.

En fait, la difficulté de la loi binomiale n'est pas d'effectuer les calculs, mais de savoir poser le problème. Il faut donc bien connaître la définition de ses termes.

2. Définition des termes de la loi binomiale

- **Jeu de hasard**
- ***Succès :*** c'est l'une des valeurs d'une variable binaire qu'on appelle « événement ». L'événement peut, soit se produire et c'est un succès, soit ne pas se produire et c'est un échec.
  Le terme **k** désigne le nombre de succès après une série de tentatives.

- ***Tentative :*** c'est l'épreuve qui consiste à tenter que l'événement se produise. On la dénomme aussi tirage.
  Le terme **n** désigne le nombre d'épreuves que l'on tente.
- ***Probabilité de survenue*** de l'événement : c'est la probabilité d'obtenir un succès à chaque tentative ; elle est comprise entre 0 et 1.
  Le terme **P** la désigne.

■ Étude d'un échantillon

- ***Caractéristique :*** c'est l'une des valeurs d'une variable binaire dont on veut étudier la distribution. Chaque sujet de l'échantillon en est porteur ou non. La caractéristique est l'équivalent du succès dans la situation de jeu.
  Le terme **k** désigne le nombre d'individus porteurs de la caractéristique dans un échantillon.
- ***Échantillon :*** c'est un groupe de sujets choisis au hasard dans une population. La taille de l'échantillon est l'équivalent du nombre de tentatives dans la situation de jeu.
  Le terme **n** désigne la taille de l'échantillon.
- ***Proportion de la caractéristique :*** c'est la proportion de la caractéristique dans la population d'où est tiré l'échantillon.
  Le terme **P** désigne cette proportion.

La loi binomiale

JEU DE HASARD		ÉTUDE D'UN ÉCHANTILLON	
Nombre d'événements gagnants (succès)	k	Nombre de sujets porteurs de la caractéristique dans l'échantillon	k
Nombre de tentatives	n	Taille de l'échantillon	n
Probabilité de gagner à chaque tentative	P	Proportion de sujets porteurs de la caractéristique dans la population	P

On appelle X la variable qui peut prendre la valeur k.
On appelle P(X = k) : la probabilité d'observer la valeur k.

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

Exemple 6.2. LOI BINOMIALE

Quelle est la probabilité de tirer 3 as en jetant 10 fois un dé ?

k = nombre d'as à sortir = 3
n = nombre de tentatives = 10
P = probabilité de sortir un as = une face sur 6 = 0,17

$$P(3 \text{ as}) = \frac{10!}{3!(10-3)!} 0{,}17^3 \, 0{,}83^{10-3} = 0{,}16$$

La probabilité d'obtenir 3 as est de 16 %

Excel® : fonction LOI.BINOMIALE (3 ; 10 ; 0,17 ; faux) = 0,16.

Exemple 6.3. LOI BINOMIALE

Quelle est la probabilité d'observer 3 malades dans un échantillon de 10 sujets choisis au hasard dans une population où la fréquence de la maladie est de 17 % ?

La caractéristique étudiée est : « être malade ».
La proportion de sujets porteurs de la caractéristique est donnée par la fréquence de la maladie dans la population.
k : nombre de malades = 3
n : taille de l'échantillon = 10
P : fréquence de la maladie dans la population = 0,17

$$P(3 \text{ malades}) = \frac{10!}{3!(10-3)!}\ 0{,}17^{3}\ 0{,}83^{10-3} = 0{,}16$$

La probabilité d'observer 3 malades est de 16 %.

On aurait pu poser la même question pour zéro malade, 1 malade, 2 malades,...10 malades. Il suffit de refaire les mêmes opérations en calculant P (0), P (1), P (2), *etc.*

La probabilité de n'observer aucun malade :	P(0) = 0,155
La probabilité d'observer 1 malade :	P(1) = 0,318
La probabilité d'observer 2 malades :	P(2) = 0,293
La probabilité d'observer 3 malades :	P(3) = 0,160
La probabilité d'observer 4 malades :	P(4) = 0,057
La probabilité d'observer 5 malades :	P(5) = 0,014
La probabilité d'observer 6 malades :	P(6) = 0,0024
La probabilité d'observer 7 malades :	P(7) = 0,00028
La probabilité d'observer 8 malades :	P(8) = 0,00002
La probabilité d'observer 9 malades :	P(9) = 0,000001
La probabilité d'observer 10 malades :	P(10) = 0,00000002

La somme des probabilités pour tout l'échantillon est égale à 1.

On peut dessiner le graphe de cette distribution.

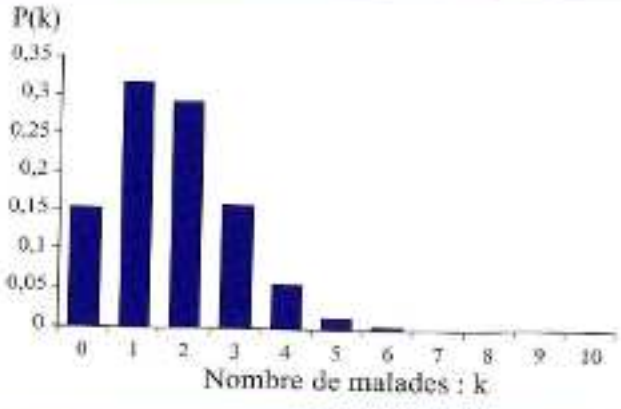


Figure 6-1. Distribution de la probabilité d'obtenir un nombre k de malades dans un échantillon de 10 sujets issus d'une population où la fréquence de la maladie est de 17 %

3. Conditions d'application de la loi binomiale

Pour utiliser la loi binomiale, il faut que :

- la variable étudiée soit de type binaire ;
- les tentatives soient indépendantes les unes des autres (situation de jeu) ou bien que l'échantillon soit tiré au sort ;

- chaque événement ait la même probabilité de succès au cours de chaque tentative ou bien que tous les individus de la population étudiée aient la même chance d'être tiré au sort ;
- la taille n de l'échantillon soit négligeable par rapport à la taille N de la population ($n/N < 10\ \%$). Lorsque cette condition n'est pas respectée, il faut utiliser une autre loi appelée loi hypergéométrique.

Exemples de distributions binomiales en fonction de la probabilité P et de la taille de l'échantillon n

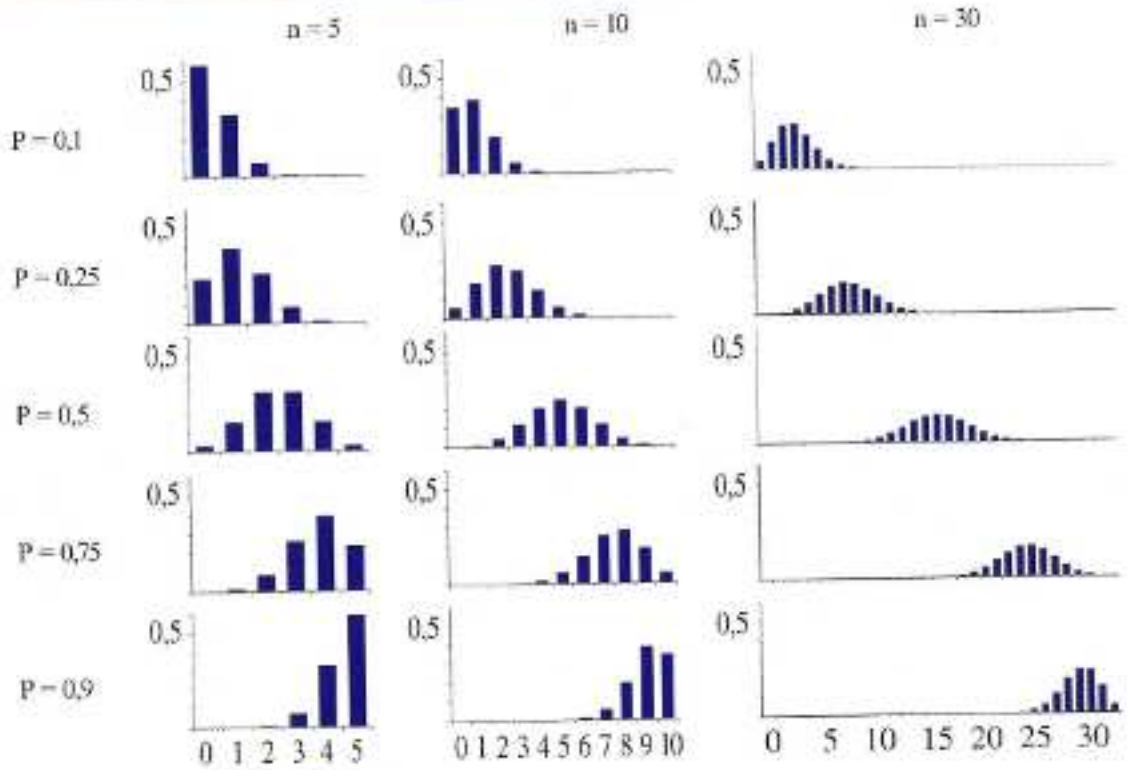


Figure 6-2.

- P représente la probabilité théorique de l'événement (situation de jeu) ou bien la fréquence de la caractéristique étudiée dans la population (étude d'un échantillon).
- L'ordonnée de chaque courbe représente la probabilité P(k) d'observer k sujets porteurs de la caractéristique dans un échantillon de n individus.

Les distributions binomiales présentées sur la figure 6-2 varient en fonction de la valeur de la proportion P et de la taille de l'échantillon.

On remarquera que dans la case centrale et dans les trois cases centrales de la colonne de droite, les distributions prennent une forme symétrique. Dans ces cases, les produits nP et n(1 – P) sont supérieurs ou égaux à 5. Nous verrons plus tard que lorsque ces deux conditions sont respectées ($nP \geq 5$ et $n(1 - P) \geq 5$) la loi binomiale peut être remplacée par la loi normale.

4. Propriétés additives de la loi binomiale (exemple 6.4)

Souvent la question posée n'est pas simplement la probabilité d'observer une seule valeur k, mais la probabilité d'observer :

- une valeur inférieure à **k** : $P(X < k)$;
- une valeur au plus égale à **k** : $P(X \leq k)$;
- une valeur supérieure à **k** : $P(X > k)$;
- une valeur au moins égale à **k** : $P(X \geq k)$.

Fonctions de répartition de la loi binomiale

- $P(X < k) = P(0) + P(1) + \ldots\ldots\ldots + P(k-1) = 1 - P(X \geq k)$
- $P(X \leq k) = P(0) + P(1) + \ldots\ldots + P(k-1) + P(k) = 1 - P(X > k)$
- $P(X > k) = P(k+1) + \ldots\ldots\ldots + P(k_n) = 1 - P(X \leq k)$
- $P(X \geq k) = P(k) + P(k+1) + \ldots\ldots\ldots + P(k_n) = 1 - P(X < k)$

Exemple 6.4. LOI BINOMIALE CUMULATIVE

1) Quelle est la probabilité d'observer **moins de** 4 malades dans un échantillon de 10 sujets choisis au hasard dans une population où la fréquence de la maladie est de 17 %.
En reprenant les résultats de l'exemple 6.3 :

$$P(X < 4) = P(0) + P(1) + P(2) + P(3) = 0{,}155 + 0{,}318 + 0{,}293 + 0{,}160 = 0{,}926 = 92{,}6\ \%$$

Excel® : fonction LOI.BINOMIALE (3 ; 10 ; 0,17 ; vrai) = 0,926.

2) Quelle est la probabilité d'observer **plus de** 3 malades dans un échantillon de 10 sujets choisis au hasard dans une population où la fréquence de la maladie est de 17 %.

$$P(X > 3) = 1 - P(X < 4) = 1 - 0{,}926 = 0{,}074 = 7{,}4\ \%$$

5. À quoi sert la loi binomiale ? (bis)

En fait, il est peu intéressant en pratique de calculer les probabilités d'observer des événements. En général, la question qui se pose est :
l'échantillon sur lequel j'observe un certain nombre de sujets possédant une caractéristique provient-il bien de la population pour laquelle je connais la proportion P de la caractéristique étudiée ?

En d'autres termes, est-ce que la probabilité d'avoir observé ce que j'observe est suffisamment élevée pour admettre que mon échantillon provienne de la population ?

Ou bien, est-ce que la probabilité d'avoir observé ce que j'observe est vraiment trop faible ? Alors je rejetterai l'idée que mon échantillon est représentatif.

Le seuil de probabilité au-dessous duquel on rejette une hypothèse est arbitraire. Il est très souvent choisi à 5 %. Nous reviendrons plus tard sur ces notions fondamentales (exemple 6.5).

Exemple 6.5.

Si dans notre échantillon de 10 sujets provenant d'une population où la fréquence supposée d'une maladie est de 17 %, nous avions trouvé 5 malades, que pourrait-on en conclure ?
On observe que la proportion de malades dans cet échantillon (50 %) est beaucoup plus élevée que la proportion attendue (17 %).
La question pertinente est donc de se demander quelle était la probabilité d'en avoir observé AUTANT, c'est-à-dire de calculer la probabilité d'en avoir observé **au moins** 5.

$$P(X \geq 5) = P(5) + P(6) + P(7) + P(8) + P(9) + P(10) = 0{,}014 + 0{,}0024 + 0{,}00028 + \ldots = 0{,}0168$$

Il n'y a donc que 1,7 chance sur 100 d'observer un tel résultat.
On en conclut (avec un risque de se tromper de 1,7 sur cent) que l'échantillon ne provient pas de la population présumée. Il doit provenir d'une autre population où existent des facteurs favorisant la maladie.

Remarque : dans l'exemple ci-dessus, nous avons pu faire les calculs car nous avons choisi une taille d'échantillon très petite ($n = 10$).
Nous verrons plus tard que lorsque l'échantillon est grand et lorsque nP et $n(1 - P)$ sont ≥ 5, il existe d'autres moyens simplifiés pour faire ce type de travail.
Mais gardez en mémoire que la loi binomiale est le moyen exact pour calculer les probabilités de survenue d'événements de nature binaire. Si vous possédez un bon ordinateur et un bon tableur, préférez la loi binomiale !

II. LOI DE POISSON

Commençons par deux exemples (exemples 6.6 et 6.7).

Exemple 6.6.

Sachant que la fréquence annuelle de la trichinellose (une maladie parasitaire) est de 10 cas pour 50 millions d'habitants, quelle est la probabilité d'observer 3 cas pendant une année dans une région qui compte 10 millions d'habitants ?

Avec ce que nous savons de la loi binomiale, nous disposons des éléments pour résoudre la question.
Nous avons $P = 0{,}0000002$, $n = 10\,000\,000$ et $k = 3$.
Si vous avez le courage de vous atteler à la tâche, vous allez vous apercevoir qu'il est pratiquement impossible de calculer $P(X = k)$ même avec votre calculette en raison des factorielles et des puissances. Vous pourriez le faire seulement avec un tableur moderne possédant la fonction binomiale.

Excel® : fonction LOI.BINOMIALE (3 ; 10000000 ; 0,0000002 ; faux) = 0,86 soit 86 %.

Exemple 6.7.

Sachant que dans un service d'urgence, on accueille en moyenne 5 entorses par week-end, quelle est la probabilité d'observer 3 entorses au cours du prochain week-end ?

Cette question anodine nous paraît aisée à traiter avec ce que nous savons de la loi binomiale. Pourtant, il existe une difficulté insurmontable. Nous ne connaissons pas la proportion P de la caractéristique dans la population des consultants. Nous ne disposons ici que du numérateur qui permettrait de calculer cette proportion. Les sujets porteurs de la caractéristique inverse « ne pas avoir d'entorse » ne sont pas dénombrables.

Il existe donc des situations dans lesquelles il n'est pas possible d'utiliser la loi binomiale :

- soit parce que la caractéristique de la variable étudiée est très rare ; sa proportion **P** dans la population est donc très faible. La taille **n** de l'échantillon nécessaire pour observer quelques cas serait très élevée ;
- soit lorsque la caractéristique est un événement de type accidentel. Dénombrer les non-événements serait absurde. La caractéristique étudiée n'est donc pas connue sous forme d'une proportion. Elle est représentée par un nombre absolu. Il représente le nombre d'événements attendus en moyenne ***pendant une période déterminée***.

1. À quoi sert la loi de Poisson ?

La loi de Poisson s'applique, comme la loi binomiale, à des variables qualitatives. Elle permet de répondre à la question suivante :

« Connaissant le nombre moyen **μ** d'événements attendus pendant une période donnée, quelle est la probabilité d'observer **k** individus ayant subi cet événement pendant une période équivalente ? » (exemple 6.8.)

Loi de Poisson

- μ : nombre moyen d'événements observés dans la population pendant une période donnée
- e = 2,718.....
- X : la variable représentant le nombre d'individus ayant subi l'événement observé pendant la période donnée
- k : une valeur de cette variable X
- P(X = k) : la probabilité d'observer la valeur k

$$P(X = k) = \frac{e^{-\mu}\mu^{k}}{k!}$$

Exemple 6.8. LOI DE POISSON

Quelle est la probabilité d'observer 3 entorses au cours d'un week-end ordinaire de garde aux urgences, sachant qu'en moyenne 5 cas d'entorse sont admis par week-end ?

On a : $\mu = 5$ et $k = 3$

Probabilité d'observer 3 entorses : $P(3) = \frac{2{,}718^{-5}\ 5^3}{3!} = 0{,}140$

La probabilité d'observer 3 entorses est de 14 %.

Excel® : fonction LOI.POISSON (3 ; 5 ; faux) = 0,14.

On peut évidemment calculer l'ensemble de toutes les probabilités possibles

Probabilité de n'observer aucune entorse :	p (0) = 0,007
Probabilité d'observer une entorse :	p (1) = 0,034
Probabilité d'observer deux entorses :	p (2) = 0,084
Probabilité d'observer trois entorses :	p (3) = 0,140
Probabilité d'observer quatre entorses :	p (4) = 0,175
Probabilité d'observer cinq entorses :	p (5) = 0,175
Probabilité d'observer six entorses :	p (6) = 0,146
Probabilité d'observer sept entorses :	p (7) = 0,104
Probabilité d'observer huit entorses :	p (8) = 0,065
etc.	

Comme pour la loi binomiale, souvent la question posée est d'observer une valeur inférieure, supérieure, au plus égale ou au moins égale à **k** (exemple 6.9).

Fonctions de répartition de la loi de Poisson

- $P(X < k) = P(0) + P(1) + \ldots\ldots\ldots\ldots + P(k-1)$
- $P(X \leq k) = P(0) + P(1) + \ldots\ldots + P(k-1) + P(k)$
- $P(X > k) = 1 - P(X \leq k)$
- $P(X \geq k) = 1 - P(X < k)$

Exemple 6.9. LOI DE POISSON CUMULATIVE

a) La probabilité d'observer moins de 2 entorses si on en observe 5 en moyenne est de : $P(X < 2) = P(0) + P(1) = 0{,}04$ soit 4 % c'est-à-dire peu de chances de passer une garde tranquille.

Excel® : fonction LOI.POISSON (1 ; 5 ; vrai) = 0,04.

b) La probabilité d'observer plus de 6 entorses est de :
$P(X > 6) = 1 - P(X \leq 6) = 1 - [P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6)]$
$= 1 - (0{,}007 + 0{,}034 + 0{,}084 + 0{,}140 + 0{,}175 + 0{,}175 + 0{,}146) = 0{,}238$ soit 23,8 %.
Environ une chance sur quatre de passer une soirée agitée.

Excel® : fonction = 1 - LOI.POISSON (6 ; 5 ; vrai) = 0,238.

2. Conditions d'application de la loi de Poisson

- Les événements doivent être dénombrables.
- Les événements doivent être indépendants les uns des autres.
- Elle s'applique aux événements rares dont la probabilité de survenue est inférieure à 0,05. Si la probabilité est supérieure, il faut appliquer la loi binomiale.

Exemples de lois de Poisson pour différentes valeurs de la moyenne de survenue d'un événement pendant une période donnée

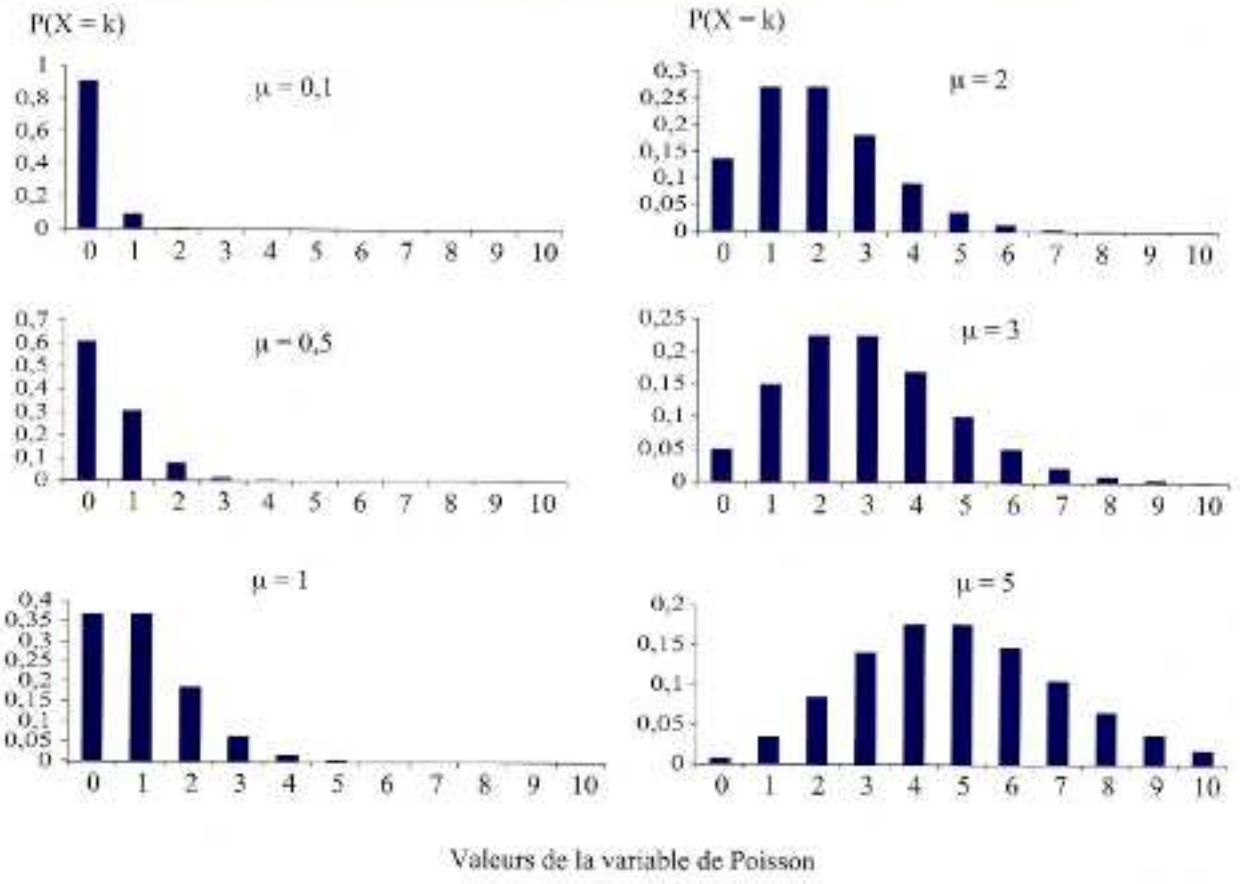


Figure 6-3.

L'abscisse de chaque graphe représente le nombre k d'événements possibles.
L'ordonnée de chaque graphe représente la probabilité P (k) que l'on observe ce nombre k.

III. LOI NORMALE

C'est la plus importante des lois utilisées en statistique. La loi normale s'applique aux variables quantitatives continues.

En biologie, on constate souvent que la distribution des valeurs d'une variable s'agglutine autour d'une valeur moyenne. Ensuite ces valeurs décroissent symétriquement de part et d'autre de cette moyenne. C'est le cas par exemple de la distribution de la taille des individus dans une population.

Nous avons vu au chapitre 5.II qu'on pouvait représenter la distribution d'une variable continue par sa densité de probabilité.

Lorsqu'une variable aléatoire X suit une loi normale, sa courbe de densité de probabilité a une forme tout à fait particulière appelée ***courbe en cloche*** (figure 6-4).

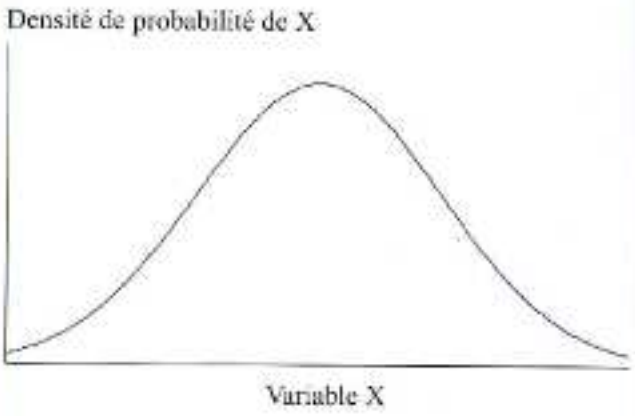


Figure 6-4. Courbe en cloche suivant une loi normale

L'expression mathématique de cette courbe figure en Annexes § 24.5.

1. Propriétés de la loi normale

- La loi normale est centrée autour de la moyenne (figure 6-5). La médiane d'une distribution normale est égale à sa moyenne.
- L'aire contenue entre les deux points d'inflexion de la courbe mesure la probabilité que les valeurs de X soient comprises entre – 1 écart type et + 1 écart type autour de la moyenne. Cette probabilité est de 68 %.
- L'aire comprise entre – 1,96 écart type et + 1,96 écart type autour de la moyenne représente 95 % de l'aire totale. En d'autres termes, 95 % des valeurs de X sont comprises à peu près entre 2 écarts type de part et d'autre de la moyenne.
- Inversement, 5 % des valeurs de X sont extérieures à l'intervalle de 2 écarts type autour de la moyenne : 2,5 % à gauche dans les valeurs basses et 2,5 % à droite dans les valeurs hautes.
- L'aire contenue entre – 3 écarts type et + 3 écarts type représente 99,7 % des valeurs de X. En d'autres termes, la quasi-totalité des valeurs de la distribution normale est contenue dans cet intervalle.

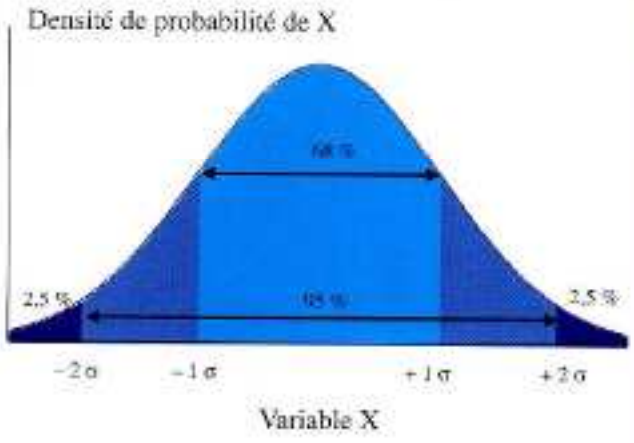


Figure 6-5. Loi normale

2. Loi normale cumulée

La fonction de répartition d'une loi normale cumulée dessine une courbe en S (figure 6-6). Cette courbe représente la surface de la courbe en cloche. Le point d'inflexion de cette courbe correspond à la moyenne.
Chaque point de la courbe donne la probabilité que x soit inférieure à une valeur donnée.

- 2,5 % des valeurs sont inférieures à -2σ.
- 97,5 % des valeurs sont supérieures à $+2\sigma$.

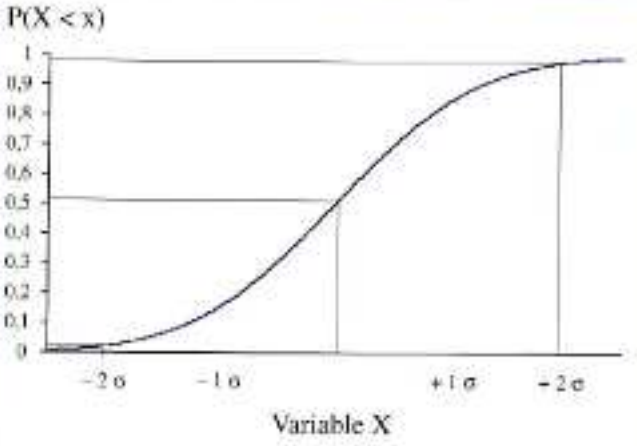


Figure 6-6. Fonction de répartition de la loi normale

3. Loi normale centrée réduite

Toute variable suivant une loi normale est représentée par sa courbe en cloche dont la position et la forme dépendent de sa moyenne et de son écart type (figure 6-7).
On ne peut pas établir une table de toutes les distributions normales possibles. Mais par une transformation de variable adéquate, toutes les lois normales peuvent se résumer en une seule distribution.

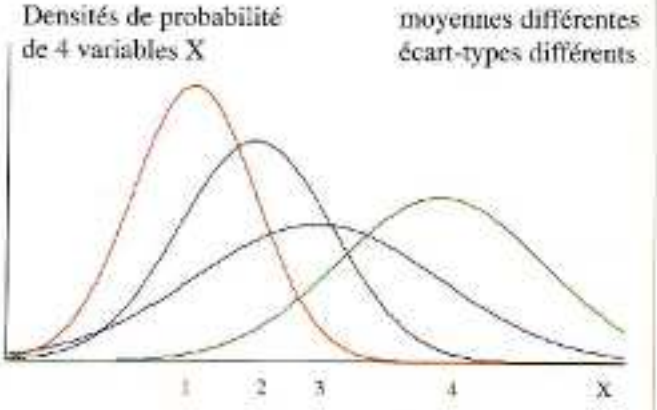


Figure 6-7. Plusieurs types de loi normale

1) Posons dans un premier temps une nouvelle variable X' égale à X moins la moyenne de la distribution étudiée ($X' = X - \mu$). Cela aboutit à **centrer** la distribution autour de zéro. Quelle que soit la variable X étudiée, sa transformée X' est centrée autour d'une moyenne égale à 0 (figure 6-8).

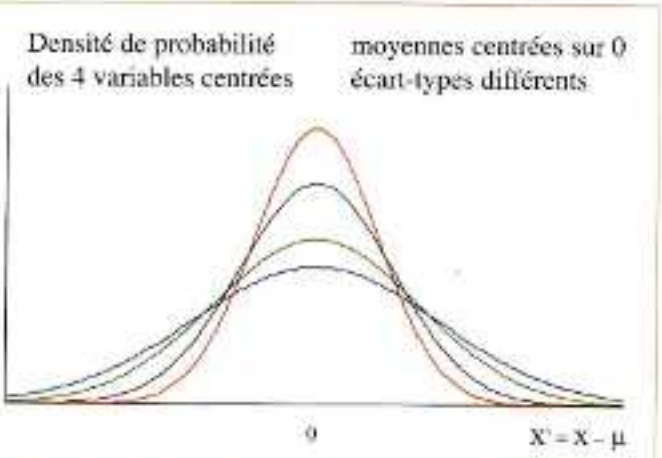


Figure 6-8. Plusieurs types de loi normale centrée

2) Dans un deuxième temps, posons une nouvelle variable, qu'on appellera Z, en divisant X' par l'écart type de la distribution étudiée soit $Z = \frac{X - \mu}{\sigma}$. La nouvelle variable Z a ainsi un écart type **réduit** à 1.

Quelle que soit la variable d'origine, sa transformée, la variable centrée réduite Z est centrée autour de 0 et a pour écart type 1. Nous aboutissons ainsi à une seule courbe (figure 6-9).

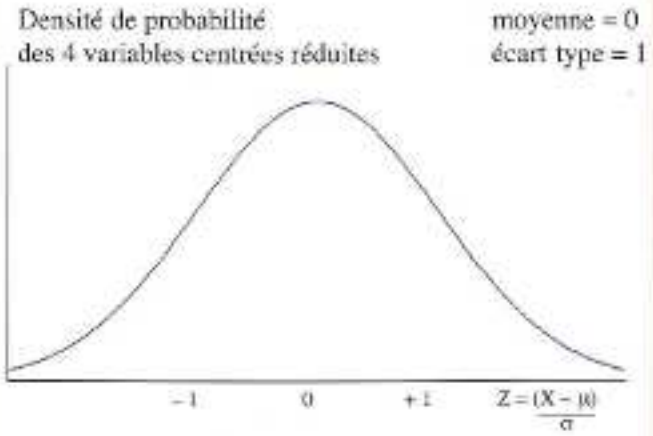


Figure 6-9. Loi normale centrée réduite Z

La variable centrée réduite Z

- X : variable normale
- μ : moyenne de la variable X
- σ : écart type de la variable X

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

4. Propriétés de la loi de Z normale centrée réduite

Formule en Annexes, § 24.7

- La loi de Z est centrée autour de la valeur zéro (figure 6-10).
- La loi de Z a pour écart type la valeur 1.
- 95 % des valeurs de Z sont comprises entre – 1,96 et + 1,96 ≈ (– 2 et + 2).
- 2,5 % des valeurs de Z sont inférieures à – 1,96.
- 2,5 % des valeurs de Z sont supérieures à + 1,96.

Toutes les propriétés de cette loi Z dite *normale centrée réduite* sont fondamentales à connaître. Elles conditionnent tout le raisonnement concernant l'estimation d'une distribution à partir d'un échantillon et la plupart des tests statistiques usuels.

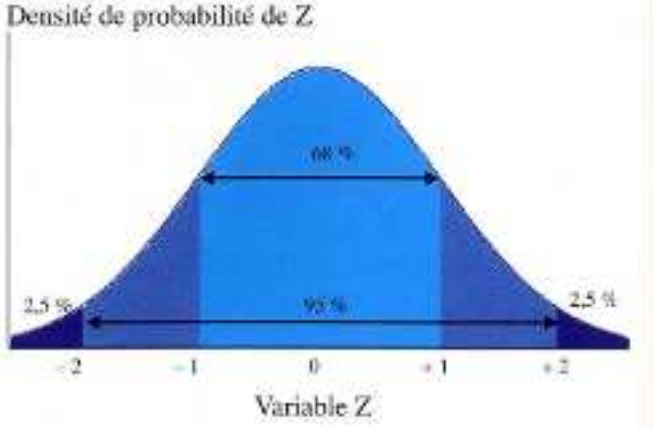


Figure 6-10. Loi normale centrée réduite Z

Exercices

Exercice 6.1

Si un vaccin produit un effet indésirable chez 15 % des sujets vaccinés, quelle est la probabilité d'observer au moins 4 sujets présentant un tel effet au cours d'une séance de vaccination de 10 sujets ?

Exercice 6.2

On sait que la probabilité d'incident médical au cours d'un vol chez les passagers des avions de ligne est de 1 pour 11 000 passagers. Quelle est la probabilité d'observer la survenue d'un incident lors d'un vol au cours d'un voyage de durée moyenne dans un appareil de 300 places ?

Exercice 6.3

Lors d'une épidémie de gale dans d'un établissement pour personnes âgées, on a constaté la survenue de 25 cas de gale en un mois parmi 80 pensionnaires logeant dans 40 chambres doubles. On a compté 24 chambres sans cas de gale, 7 chambres avec 1 seul cas par chambre et 9 chambres avec 2 cas par chambre.

Peut-on dire qu'il existe un risque de transmission secondaire de personne à personne ?

Exercice 6.4

On observe en moyenne 1 accident mortel par week-end sur les routes d'un département.

1) Calculez les probabilités d'observer 0, 1, 2, 3, 4 accidents mortels.

2) Quelle est la probabilité d'observer au moins un accident mortel ?

3) Quelle est la probabilité d'observer moins de deux accidents mortels ?

Exercice 6.5

On a observé dans un arrondissement d'un département, la survenue de 4 cas de leucémie pendant une année. Depuis 20 ans, la moyenne annuelle du nombre de cas de leucémie dans cet arrondissement est de 1,4 cas. Peut-on conclure à un risque accru de leucémie cette année là ?

Exercice 6.6

Si une variable suit une loi normale :

1) Quelle est la probabilité d'observer une valeur inférieure à la moyenne μ ?

2) Quelle est la probabilité d'observer une valeur inférieure à $(\mu - 1\sigma)$?

3) Quelle est la probabilité d'observer une valeur contenue entre $(\mu - 2\sigma)$ et $(\mu + 1\sigma)$?

Résumé

LOIS DE DISTRIBUTION

LOI BINOMIALE

- Probabilité d'observer **k** événements parmi **n** tentatives, sachant que la probabilité de l'événement est **P**.
- Probabilité d'observer **k** individus porteurs d'une caractéristique dans un échantillon de **n** individus, sachant que la fréquence de la caractéristique dans la population est **P**.

LOI DE POISSON

- Probabilité d'observer **k** événements pendant une période donnée, sachant que le nombre moyen d'événements pendant cette période est **μ**.

LOI NORMALE

Elle s'applique aux variables quantitatives continues.

- La probabilité d'observer une valeur comprise entre – 1,96 écart type et + 1,96 écart type autour de la moyenne est de 95 %.
- La probabilité d'observer une valeur comprise entre – 1 écart type et + 1 écart type autour de la moyenne est de 68 %.

LOI NORMALE CENTRÉE RÉDUITE Z

À partir d'une variable quantitative X de moyenne μ et d'écart type σ, on construit une variable centrée réduite Z en posant

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

- La moyenne d'une variable normale centrée réduite est égale à 0.
- L'écart type d'une variable normale centrée réduite est égal à 1.
- La probabilité d'observer une valeur comprise entre – 1,96 et + 1,96 est de 95 %.

Deuxième partie

ESTIMATION

ESTIMATION

SONDAGE

I. BIAIS DE SÉLECTION
II. TIRAGE AU SORT : LE HASARD
III. SONDAGES ALÉATOIRES
IV. SONDAGES EMPIRIQUES

MESURES STATISTIQUES SUR UN ÉCHANTILLON

I. PARAMÈTRES DE POSITION
II. PARAMÈTRES DE DISPERSION

ESTIMATION D'UN PARAMÈTRE

I. ESTIMATION D'UNE MOYENNE INCONNUE
II. ESTIMATION D'UN POURCENTAGE INCONNU
III. RISQUE D'ERREUR CONSENTIE α
IV. TAILLE D'UN ÉCHANTILLON

SONDAGE

Nous avons abordé précédemment la manière de classer, de résumer et de présenter des données statistiques. Ces données concernaient l'ensemble des populations d'étude.

Lorsqu'on travaille sur l'ensemble d'une population, on appelle **recensement** l'opération consistant à dénombrer les sujets afin de recueillir des données statistiques. On dit que la collecte des données est **exhaustive**.

Lorsqu'on travaille dans le domaine des sciences de la vie, il est exceptionnel de recueillir des données sur des populations entières, à moins que la population d'étude soit très limitée en raison de ses caractéristiques (par exemple la population des malades atteints du virus Ébola pendant une période d'épidémie, dans une région donnée).

En pratique, on recueille la plupart du temps des données sur un groupe limité, sélectionné à l'intérieur d'une population.

On appelle ***échantillonnage*** l'opération consistant à identifier un sous-groupe d'individus dans une population afin d'y recueillir des données statistiques.

On appelle ***échantillon*** le groupe d'individus qui a été sélectionné.

On appelle ***sondage*** la méthode utilisée pour échantillonner.

L'avantage de l'échantillonnage est de permettre une énorme économie de moyens. Un échantillon permet de connaître les paramètres de mesure d'une variable dans l'ensemble de la population, sans être obligé d'étudier toute cette population. Nous verrons même qu'il suffit d'une très faible partie de cette population pour être capable de la caractériser. Bien évidemment, le prix à payer est un certain flou dans la mesure, une certaine imprécision.

Le travail sur échantillon n'a qu'un but : extrapoler les données observées à l'ensemble de la population. Les paramètres mesurés sur un échantillon (moyenne, variance, écart type, pourcentage) sont des ***estimateurs*** des vraies valeurs inconnues dans la population.

La qualité primordiale d'un échantillon est donc d'être ***représentatif*** de la population qu'il est censé décrire. L'échantillon doit être l'image, réduite, mais fidèle de cette population.

I. BIAIS DE SÉLECTION

Lorsqu'un échantillon n'est pas représentatif, il fournit des données et des paramètres ***biaisés***.
Le processus de sélection des individus ne doit pas procéder d'un choix subjectif. Ce processus doit être indépendant de toutes les caractéristiques des individus. S'il existe la moindre liaison entre une particularité des individus et le processus de sélection, l'échantillon sera biaisé. En d'autres termes, on introduit des biais dès que le processus de sélection influe sur le résultat (exemple 7.1).

Exemple 7.1. ÉCHANTILLONS BIAISÉS

- Interroger des passants dans la rue. On sélectionne la fraction de la population valide, passant dans ce quartier et ne travaillant pas à l'heure du sondage. On élimine, les sujets impotents, les enfants, les gens ne fréquentant pas ce quartier, *etc.*
- Interroger des abonnés au téléphone. Ce type de tirage dans l'annuaire du téléphone est assez souvent utilisé. On accepte de négliger les sujets ne possédant pas le téléphone (populations marginales, défavorisées, passagères, étrangères) ou utilisant exclusivement un téléphone portable. En outre, on élimine également, si l'on n'y prend pas garde, les personnes ne répondant pas directement au téléphone (enfants, impotents).
- Interroger des malades hospitalisés. Ces malades n'ont pas les mêmes caractéristiques que les malades du même type admis en clinique ou traités en ambulatoire.
- Interroger les lycéens d'un seul établissement pour connaître certains comportements. On sait que la localisation géographique d'un établissement scolaire est fortement liée aux conditions sociales du lieu.
- Tester les prélèvements conservés dans un laboratoire hospitalo-universitaire. On sélectionne des prélèvements issus de maladies plus graves, plus ciblées, plus évoluées que les prélèvements du même type conservés dans des laboratoires d'analyse de ville.

Dans tous ces exemples, le processus de choix associe une caractéristique importante des individus qui vont être sélectionnés. Les résultats des mesures seront biaisés et les extrapolations qu'on en tirera sur la population d'étude seront fausses.

II. TIRAGE AU SORT : LE HASARD

Pour que la sélection des individus n'aboutisse pas à un échantillon biaisé, il n'existe qu'une seule méthode : faire confiance au hasard, tirer au sort dans l'ensemble de la population d'étude les individus de l'échantillon.
Le tirage au sort s'effectue à l'aide de tables ou de générateurs de nombres aléatoires. Ces générateurs sont fournis par la plupart des logiciels statistiques par une fonction généralement appelée ***random***. On appelle cette opération ***randomisation*** (du vieux français « aller à randon », aller dans tous les sens, randonner). Excel® : fonction ALEA().
Si l'on admet que le ***hasard*** est la rencontre de 2 événements totalement indépendants, on admettra que l'attribution d'un numéro tiré au sort est indépendante des caractéristiques d'un individu donné. Sa sélection n'est due qu'au hasard.

On distingue deux types de sondage.
- Les sondages aléatoires dans lesquels la probabilité de sélection pour chaque individu est définie dès la constitution du plan de sondage.
- Les sondages empiriques dans lesquels un choix s'exerce sur le terrain en fonction de règles préalables.

III. SONDAGES ALÉATOIRES

Ce sont les plus utilisés en pratique scientifique. Ils ne laissent en effet aucune liberté de choix à l'expérimentateur et seul le hasard détermine la sélection de l'échantillon.

De nombreuses méthodes de sondage aléatoire existent : elles ont été mises au point en vue de simplifier certains problèmes logistiques, de réduire les coûts et les temps de travail, tout en conservant des résultats extrapolables.

1. Sondage élémentaire

C'est la méthode de référence et elle fonctionne comme une loterie. Elle consiste à numéroter chaque sujet de la population d'étude. Cette liste numérotée constitue la ***base de sondage***. Après avoir fixé la taille **n** de l'échantillon, on tire au sort les numéros des individus qui constitueront l'échantillon. Dans ce type de sondage **tous** les individus, **toutes** les unités statistiques constituant la population ont *a priori* la même probabilité d'être sélectionnés. Si l'on appelle N la taille de la population et n la taille désirée de l'échantillon, cette probabilité *a priori* est pour chaque individu de **n/N**.
On appelle fraction de sondage ou taux de sondage ce rapport n/N.

a) Tirage avec remise

Dans une vraie loterie, le même numéro peut sortir plusieurs fois. À chaque tirage, l'ensemble des numéros a la même probabilité de sortir Pour que chaque sujet de l'échantillon ait la même chance d'être tiré, il faudrait donc, en toute rigueur, « remettre » le sujet tiré dans la base de sondage, afin qu'à chaque tirage, **tous** les individus de la population aient la même probabilité d'être tirés (figure 7-1 en haut).

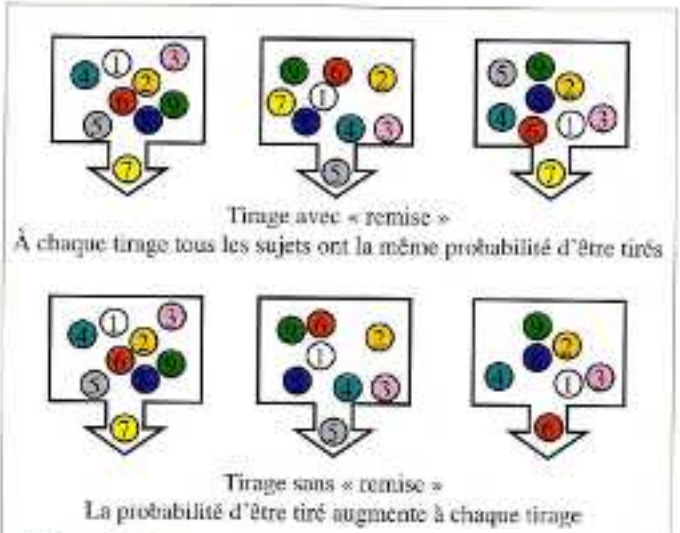


Figure 7-1.

b) Tirage sans remise

Évidemment, c'est rarement le cas et on se contente en pratique de tirer l'ensemble de l'échantillon « sans remise ». Prenons l'exemple, d'une population de 10 000 sujets dans lequel on tire 10 individus sans remise. Le premier individu de l'échantillon avait une chance sur 10 000 d'être tiré. Le second n'a plus que une chance sur 9 999, le troisième une chance sur 9 998 et le dixième 1 chance sur 9 991.
En pratique, on néglige ce problème tant que l'échantillon est petit vis-à-vis de la population.
La méthode du sondage aléatoire élémentaire nécessite une base de sondage, c'est-à-dire la liste nominale de tous les individus ou unités statistiques composant une population. Lorsque la population d'étude est de très grande taille ou lorsqu'elle n'est pas définie par avance, cette méthode n'est pas applicable.

2. Sondage systématique

Ce type de sondage est utilisé lorsqu'on dispose d'une base de grande taille ordonnée, mais non numérotée. La base peut être considérée comme une pile d'individus ou d'unités statistiques, classée selon n'importe quel ordre. Pour économiser l'étape lourde de numérotage, on détermine un **pas de sondage** qui est le rapport entre la taille de la population **N** et la taille désirée de l'échantillon **n**. Le pas est donc égal à **N/n**. C'est l'inverse du taux de sondage. Après avoir tiré au sort un premier individu entre **1** et **N/n**, on tire ensuite tous les suivants de façon systématique en balayant la pile de pas en pas (figure 7-2). À la fin de la pile, on aura tiré **n** individus (exemple 7.2).

Exemple 7.2. SONDAGE SYSTÉMATIQUE

Soit une population de 5 000 individus classée par ordre alphabétique. On désire obtenir un échantillon de 100 sujets. Le pas de sondage est 5 000/100 = 50. On tire au sort un premier individu entre le premier et le cinquantième de la pile, par exemple le numéro 27. À partir de ce sujet, on en tire un tous les 50. Le deuxième individu tiré sera le 77e de la pile, le troisième sera le 127e, *etc.* Le 100e sujet de l'échantillon sera le 4 927e de la pile.

Dans un sondage systématique, le seul tirage au sort concerne le premier individu. Le reste de l'échantillon est déterminé par le pas de sondage. On constate donc que tous les individus de la pile avaient *a priori* la même probabilité d'être tirés. Dès que le premier sujet a été tiré, la sélection s'opère ensuite mécaniquement.

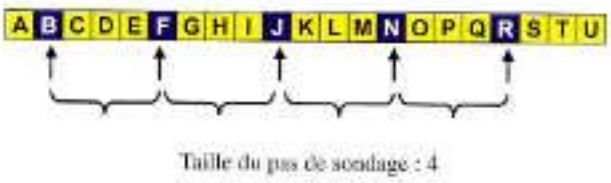


Figure 7-2. Sondage systématique

Cette méthode du sondage systématique est aussi utilisée lorsqu'on ne dispose pas de base de sondage, notamment lorsque l'échantillon doit être constitué à partir d'une population qui augmente au cours du temps (exemple 7.3).

Exemple 7.3. SONDAGE SYSTÉMATIQUE

L'exemple le plus fréquent est celui de la sélection de patients se présentant à une consultation. Si l'on voulait réaliser un sondage élémentaire, il faudrait recueillir les données chez ***tous*** les patients pendant une période déterminée et sélectionner ensuite l'échantillon pour analyser les données. On conviendra qu'il est plus économique, de recueillir des données seulement dans l'échantillon. On détermine donc un pas de sondage en estimant la taille **N** de la population qu'on s'attend à recevoir pendant la période de l'étude.

Si l'on s'attend à recevoir environ 1 000 patients pendant la durée de l'étude et qu'on désire travailler sur un échantillon de 100 sujets, le pas de sondage sera 1 000/100 = 10. On tire au sort un chiffre entre 1 et 10 pour déterminer qui sera le premier sujet de l'échantillon par exemple 7. Le premier sujet sera le 7e consultant à partir du début de l'étude. On recueillera ensuite les données de façon systématique toutes les 10 consultations. Le deuxième sujet sera le 17e consultant, le troisième sujet sera le 27e consultant, *etc.* jusqu'à ce qu'on ait obtenu 100 sujets.

Inconvénient : il peut arriver que l'échantillon issu d'un sondage systématique soit biaisé lorsque l'ordre des individus de la population dans la pile correspond à une caractéristique dont la présence revienne exactement avec la même périodicité que le pas de sondage. On peut imaginer, par exemple, que la pile comporte de façon alternée des sujets de sexe opposé. Si le pas de sondage est un nombre pair, l'échantillon sera composé uniquement d'individus du même sexe. Si le sexe est un facteur important de l'étude, l'échantillon sera donc complètement biaisé (figure 7-3).

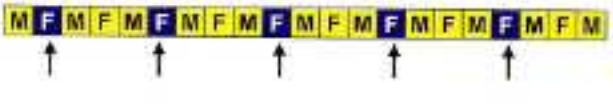


Figure 7-3.

Cette éventualité est rare, mais il faut s'assurer, lors d'un sondage systématique qu'il n'y a pas de relation entre le pas de sondage et la périodicité d'une caractéristique influant sur les variables mesurées.
Une autre difficulté survient lorsque le pas de sondage n'est pas un diviseur entier de la population. Il faut alors prendre comme pas de sondage la valeur entière la plus proche. Dans ce cas la probabilité *a priori* d'être tiré n'est plus identique pour chaque individu. Des formules complexes permettent de corriger cet effet dans le calcul des estimateurs.

3. Sondage à plusieurs degrés

Lorsque la population est de très grande taille, il est souvent impossible de réaliser un sondage élémentaire ou systématique. On peut alors pratiquer une partition de la population en groupes. La liste des groupes représente les **unités primaires** (UP). Celle liste constitue la première base de sondage. On pratique un premier sondage élémentaire ou systématique sur cette liste d'UP.
On pratique ensuite un deuxième sondage élémentaire ou systématique sur les individus des groupes qui ont été tirés. Il s'agit là d'un sondage à deux degrés (figure 7-4).
On peut également partager les unités primaires en sous-groupes d'**unités secondaires** (US), et pratiquer à nouveau un sondage élémentaire ou systématique sur ces US. Enfin, on pratique un sondage élémentaire ou systématique sur les individus composant les US. Il s'agit là d'un sondage à trois degrés (figure 7-5).
On peut imaginer de procéder avec 4, 5, ... degrés de sondage.

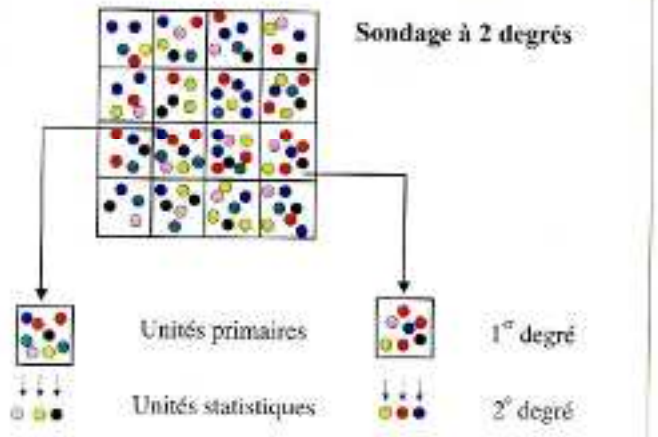


Figure 7-4.

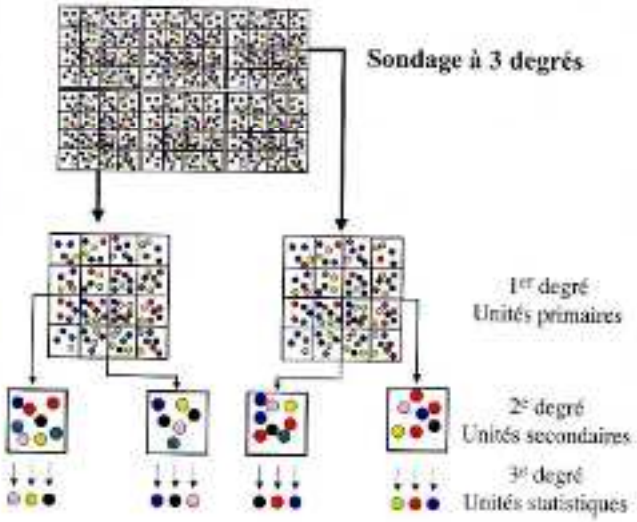


Figure 7-5.

Cette méthode permet une grande économie de moyens. Il existe néanmoins un certain nombre de difficultés.

- On perd en précision dans le calcul des estimateurs. Intuitivement, il est clair que la précision des résultats dans un sondage à deux degrés ne sera pas identique selon qu'on décide de sélectionner peu d'unités primaires et beaucoup d'individus dans chacune d'elles ou à l'inverse beaucoup d'unités primaires et peu d'individus dans chacune d'elles.

La difficulté provient du fait que la dispersion de la variable étudiée n'est pas obligatoirement homogène dans les groupes sélectionnés. Des individus peuvent être semblables dans chaque groupe, mais très différents d'un groupe à l'autre. Autrement dit, la *variance intra-groupe* peut être faible et la *variance inter-groupe* élevée. C'est ce que l'on nomme l'**effet de grappe**. Plus l'effet de grappe est élevé, plus la précision de l'estimation diminue. Pour éviter cela, il faut s'arranger pour réaliser un sondage tel que la dispersion (ou variance) soit maximale à l'intérieur des groupes et minimale entre les groupes. Cela nécessite une connaissance *a priori* de la dispersion de la variable dans la population qu'on étudie.
Il existe des ouvrages spécialisés dans le traitement de ces problèmes (*cf.* Ardilly, Annexes, § 25).

- En outre, le calcul des estimateurs devient aussi très complexe dans ce type de sondage. En effet, pour chaque degré de sondage, il faut déterminer le type de tirage, avec ou sans remise, puis le taux de sondage des unités. Les formules de calcul des estimateurs prennent en compte toutes ces procédures.

4. Sondage en grappes

C'est une variante du sondage à plusieurs degrés, lorsqu'on ne dispose pas de base de sondage permettant de faire l'ultime sondage dans le dernier degré. On décide alors de prendre tous les individus du dernier degré. Cet ensemble constitue une ***grappe*** (figure 7-6). En fin de compte, le travail sur ce dernier degré est équivalent à un recensement exhaustif.
Ici encore les formules des estimateurs sont complexes et dépendent de la taille relative des grappes et de leur probabilité d'être tirées.
Dans un sondage en grappe, on obtiendra une meilleure précision en tirant des grappes, nombreuses, de taille voisine et petite, et avec une forte variabilité interne (variance intra-grappe élevée). Il faut donc que les grappes se ressemblent et que les individus qui les composent soient les plus dissemblables possibles. Ceci est souvent difficile à obtenir en pratique car « qui se ressemble s'assemble ».

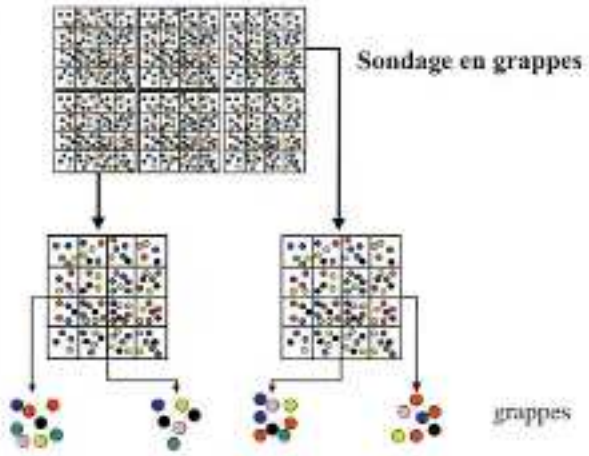


Figure 7-6.

5. Sondage stratifié

Ce type de sondage permet d'améliorer la précision des estimateurs.
Dans un sondage élémentaire, on conçoit intuitivement que la précision obtenue dépend de la taille de l'échantillon. Si la variance est élevée, il faudra que l'échantillon soit de grande taille pour obtenir une précision acceptable. Or la variance

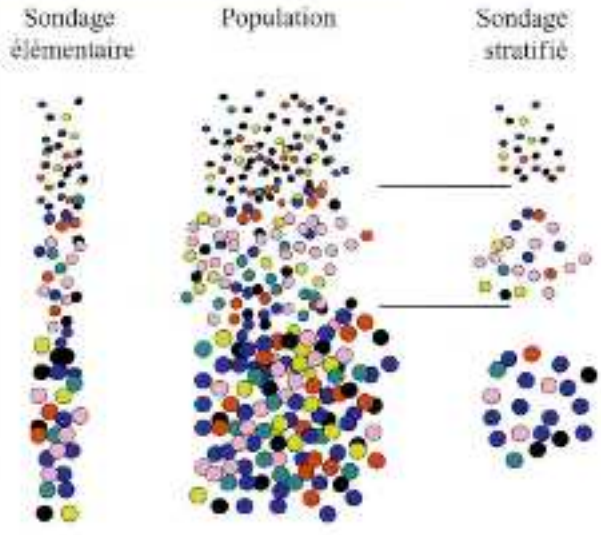


Figure 7-7.

de la variable étudiée dépend parfois d'un caractère particulier de la population. Si cette liaison est connue, il peut être judicieux de diviser la population en **strates** correspondantes aux classes de ce caractère. À l'intérieur de chaque strate, la variance devient alors homogène. Si on réalise un sondage, non pas global, mais à l'intérieur de chaque strate, on obtient une précision plus grande. Le sondage à l'intérieur de chaque strate peut être soit élémentaire, soit systématique.

Le taux de sondage à l'intérieur de chaque strate peut être :
- soit égal ;
- soit proportionnel aux écarts type (si on en a une idée *a priori*) ;
- soit choisi de façon raisonnée si l'on veut augmenter la précision dans une strate particulière.

Les formules de calcul sont plus complexes que pour un sondage élémentaire.

Exemple 7.4. SONDAGE STRATIFIÉ

On décide de mesurer la fréquence de la bilharziose dans une île peuplée de 300 000 habitants. On s'attend à une fréquence de la maladie d'environ 4 %. La population est répartie à 50 % en zone très urbanisée, 25 % dans des bourgs et 25 % en zone rurale. On sait que la bilharziose se contracte presque exclusivement en zone rurale, mais il n'est pas exclu que les citadins puissent se contaminer en se rendant dans les campagnes.
On dispose de moyens financiers permettant d'étudier un échantillon de 3 000 sujets.
Si l'on décide un sondage élémentaire, sur l'ensemble de la population avec une fraction de sondage de 1/100, l'échantillon sera composé d'environ 1 500 citadins, de 750 habitants des bourgs et de seulement 750 ruraux. On détectera peu de cas de bilharziose chez les citadins avec une charge de travail très lourde. À l'inverse, le nombre de cas attendus chez les ruraux sera proportionnellement élevé, mais au total on disposera de peu de cas. La précision de la mesure sera faible.

On peut alors opter pour un sondage stratifié en trois strates : urbaine, semi-urbaine, rurale. Puis, on pratiquera un sondage élémentaire dans chaque strate avec une fraction de sondage de 1/200 dans les villes, de 1/100 dans les bourgs et de 1/50 dans les campagnes. Au total, l'échantillon comprendra environ 750 citadins, 750 habitants des bourgs, et 1 500 ruraux. En favorisant la strate dans laquelle se situe a *priori* le plus grand nombre de cas, on améliore la précision sur la fréquence générale de la bilharziose.

6. Sondages stratifiés à plusieurs degrés

Toutes les combinaisons des différents types de sondage précédents peuvent être effectuées. Les formules de calcul des estimateurs deviennent alors extrêmement compliquées.

IV. SONDAGES EMPIRIQUES

On les utilise lorsqu'on ne dispose pas de base de sondage. Ils sont rapides à effectuer et ne coûtent pas cher. En revanche, ils sortent du cadre probabiliste et ne permettent pas de calculer la précision des estimateurs (chap. 9.IV). Ils sont particulièrement utilisés dans les sondages d'opinion et les enquêtes de consommation ou de comportement.

1. Méthode des quotas

Lorsqu'on connaît la structure d'une population selon certaines variables préalablement choisies (par exemple âge, sexe, profession, niveau d'études, quartier d'habitation), on définit l'échantillon avec la même structure. On pose l'hypothèse que les paramètres mesurés sur cet échantillon seront identiques dans la population. Cela suppose que les variables choisies pour structurer l'échantillon soient bien celles qui expliquent la variable étudiée. Si tel n'était pas le cas, s'il manquait dans la structure une caractéristique très liée à la variable étudiée, les résultats seraient complètement biaisés. La méthode des quotas est donc utilisée dans des enquêtes concernant des thèmes dont on connaît bien les déterminants (enquêtes d'opinion et de consommation).
L'enquêteur, malgré certaines contraintes garantissant un choix relativement aléatoire, dispose cependant d'une certaine liberté de choisir des sujets (risque de biais).

2. Méthode des itinéraires

Variante de la méthode des quotas, elle impose à l'enquêteur des trajets à respecter, pour remplir les quotas. Elle réduit la liberté de choix de l'enquêteur.

3. Méthode des transects

Méthode utilisée dans des études de terrain, notamment en écologie animale. Elle consiste à tracer plusieurs lignes parallèles à travers une aire de terrain et à prélever tous les individus sur une étroite bande de part et d'autre de cette ligne. Connaissant la surface de la bande, on extrapole à la surface de l'aire d'étude les résultats observés.

4. Méthode des unités-types

Elle consiste à définir des individus moyens pour toutes les variables de l'enquête. On fait l'hypothèse que les mesures de la variable étudiée sur ces individus moyens seront elles-mêmes centrées sur la valeur moyenne de la population.
Cette méthode est encore plus risquée que les précédentes, mais elle fournit des résultats très rapidement.

Exercice

Choisissez le type de sondage qui vous paraît le plus approprié pour déterminer :

1) la fréquence de la gale chez les personnes âgées résidant en maison de retraite et dans les centres de long et moyen séjour en France ;
2) la fréquence de la consommation de tabac chez les élèves d'un lycée de 2 000 élèves ;
3) le nombre moyen d'habitants par logement d'une ville comportant 70 000 logements répertoriés ;
4) la couverture vaccinale contre la rougeole chez les enfants de moins de 5 ans d'un département français ;
5) la fréquence d'utilisation d'une marque de dentifrice en France ;
6) la fréquence d'une maladie rare, à prédominance rurale (mais pas exclusivement) chez les travailleurs inscrits à une caisse de mutualité sociale ;
7) le degré de malnutrition chez les enfants de moins de cinq ans d'une région d'un pays en voie de développement.

Résumé

Lorsqu'on désire connaître les paramètres d'une population (moyenne, pourcentage), il n'est pas nécessaire de travailler sur l'ensemble des données concernant cette population. On peut se limiter à l'étude d'une petite partie de la population sélectionnée par sondage de telle manière que l'échantillon obtenu soit représentatif, c'est-à-dire sans biais. Il existe de nombreuses méthodes de sondage. Elles ont pour but d'obtenir l'échantillon le plus représentatif possible en économisant au maximum la charge de travail.
La meilleure méthode pour obtenir un échantillon représentatif sans biais de sélection est de s'en remettre au hasard pour opérer le processus de sélection. Les sondages effectués de cette manière sont appelés sondages aléatoires.

TYPES DE SONDAGE

ALÉATOIRE	EMPIRIQUE
élémentaire	quotas
systématique	itinéraires
à plusieurs degrés	transects
en grappes	unités type
stratifié	
stratifié à plusieurs degrés	

MESURES STATISTIQUES SUR UN ÉCHANTILLON

Sur un échantillon, on peut mesurer les mêmes types de paramètres de position et de dispersion que sur une population. Mais ces paramètres que l'on mesure sur un échantillon n'ont pas d'intérêt pour eux-mêmes. Ils n'ont de valeur qu'en tant qu'***estimateur*** des vrais paramètres inconnus dans la population.

I. PARAMÈTRES DE POSITION

1. Moyenne

On appelle **m** la moyenne d'une variable quantitative calculée sur un échantillon, par opposition à **μ** la moyenne inconnue dans la population. La moyenne d'un échantillon est souvent aussi notée $\bar{x}$.

Soit x : les valeurs de la variable
Σx : la somme de ces valeurs
n : la taille de l'échantillon

$$m = \frac{\Sigma x}{n}$$

La moyenne **m** est l'estimateur de la moyenne **μ** inconnue.

2. Pourcentage

On appelle **p** un pourcentage observé sur un échantillon, par opposition à P le pourcentage inconnu dans la population.

Soit n : la taille de l'échantillon
k : le nombre d'individus présentant la caractéristique étudiée.

$$p = \frac{k}{n}$$

Le pourcentage **p** est l'estimateur du pourcentage P inconnu.

II. PARAMÈTRES DE DISPERSION

1. Variance

On appelle s^2 la variance d'une variable quantitative calculée sur un échantillon, par opposition à σ^2 la variance inconnue dans la population.

Soit **x** : les valeurs de la variable
n : la taille de l'échantillon
m : la moyenne de l'échantillon

$$s^2 = \frac{\Sigma(x-m)^2}{n-1} = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

Remarque

On note que le dénominateur de la variance de l'échantillon est ici $\mathbf{n-1}$ au lieu de **N** au dénominateur de la variance d'une distribution donnée au chapitre 4.2.4. Cela provient du fait que la variance d'un échantillon est utilisée comme un estimateur de la variance inconnue $\boldsymbol{\sigma^2}$. On démontre que pour que cet estimateur soit non biaisé, la somme des carrés des écarts à la moyenne doit être divisée par $\mathbf{n-1}$.

2. Écart type (*standard deviation* en anglais, sd en abrégé)

On appelle **s** l'écart type calculé sur les valeurs de l'échantillon, par opposition à σ l'écart type inconnu de la population. L'écart type est la racine carrée de la variance : $s = \sqrt{s^2}$. L'écart type **s** est l'estimateur de l'écart type $\boldsymbol{\sigma}$ inconnu.

Exemple 8.1.

On a noté les valeurs suivantes en unités internationales (UI) d'une variable dans un échantillon de 10 individus.

12 10 15 8 7 9 10 14 12 10

- moyenne des valeurs :
  $m = (12 + 10 + \ldots\ldots + 10)/10 = 107/10 = \mathbf{10,7\ UI}$
- variance des valeurs :
  $s^2 = [(12 - 10,7)^2 + (10 - 10,7)^2 + \ldots\ldots (10 - 10,7)^2]/(10 - 1) = 58,1/9 = \mathbf{6,46}$ (sans unité)
  Excel® : fonction VAR (12 ; 10 ; 15... 12 ; 10) = 6,46.
- écart type des valeurs :
  $s = \sqrt{6,46} = \mathbf{2,54\ UI}$
  Excel® : fonction ECARTYPE (12 ; 10 ; 15... 12 ; 10) = 2,54.

ESTIMATION D'UN PARAMÈTRE

Faire une estimation, c'est tenter de définir les paramètres d'une population à partir des paramètres observés sur un échantillon.

Lorsqu'on observe un paramètre sur un échantillon, on pressent :

1) que la valeur observée a fort peu de chances d'être exactement la valeur inconnue de la population ;
2) que cette valeur est néanmoins assez proche de la valeur inconnue si notre échantillon est représentatif ;
3) qu'en répétant l'échantillonnage, on trouverait d'autres valeurs, toutes assez proches les unes des autres.

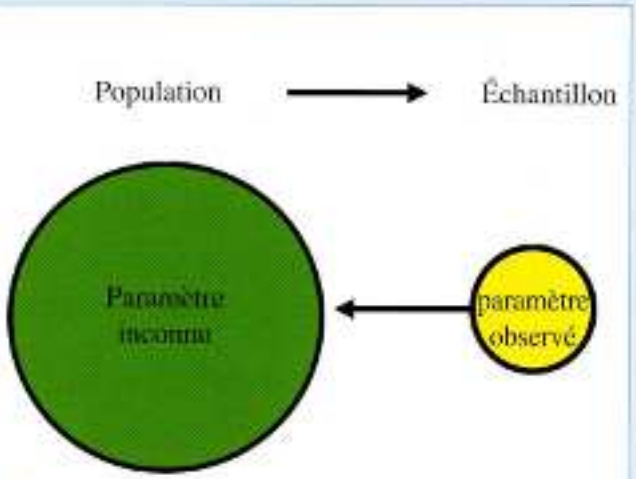


Figure 9-1.

Ces trois hypothèses sont une sorte de pari. Nous parions que la valeur observée est proche de la valeur exacte. Mais il faut préciser ce que l'on entend par « proche ». Le but de l'estimation en statistique est de calculer des bornes qui permettent de situer avec une confiance suffisamment grande où se trouve la valeur inconnue du paramètre dans la population. Une estimation aboutit donc à calculer ce qu'on nomme un « ***intervalle de confiance*** ». Ce terme est parfois appelé trivialement « fourchette d'estimation ».

Le statisticien se sait donc incapable de connaître la vraie valeur, mais il en fournit modestement une estimation à l'aide de deux bornes.

Rappelez-vous le génie d'Archimède. Avant lui, depuis des siècles, les mathématiciens s'acharnaient à trouver par des algorithmes complexes, la vraie valeur de π : (16/9)², 142/45, 62832/20000.... Archimède a mis tout le monde d'accord, en décrétant qu'on ne pourrait jamais calculer la vraie valeur de π, mais qu'elle se trouvait à coup sûr entre le périmètre du polygone extérieur au cercle 3 + 10/71 et celui du polygone intérieur 3 + 1/7. Ce renoncement a fait faire aux mathématiques un pas capital qui conduira à la découverte des nombres irrationnels.

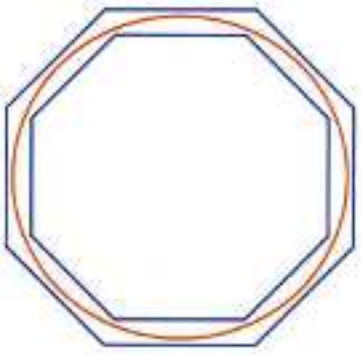


Figure 9-2.

Seule différence avec Archimède, l'intervalle des statisticiens n'est lui-même pas certain. Il est toujours affecté d'un risque d'erreur. L'essentiel de leur travail va être de réduire au minimum ce risque d'erreur.

Les paramètres que l'on peut estimer simplement se limitent à la moyenne et au pourcentage.

Dans la suite du chapitre, nous n'envisagerons que les cas où l'échantillon est issu d'un sondage aléatoire élémentaire.

I. ESTIMATION D'UNE MOYENNE INCONNUE

Lorsqu'on a observé la moyenne d'une variable quantitative sur un échantillon, le problème est d'estimer la véritable moyenne μ inconnue de la population d'où est extrait l'échantillon.
Cette estimation nécessite de savoir comment *fluctue* une moyenne observée sur un échantillon.

1. Fluctuation d'échantillonnage d'une moyenne

A priori, si l'échantillon est bien choisi, si les sujets ont été tirés au sort, en un mot, si l'échantillon est ***représentatif*** de la population, nous espérons que sa valeur $\mathbf{m_1}$ observée est assez proche de la valeur μ inconnue. Mais nous ne savons pas à quelle distance et de quel coté de μ cette valeur $\mathbf{m_1}$ se trouve.
Imaginons maintenant que nous avons la chance de disposer d'un deuxième échantillon de même taille. Nous obtiendrons alors une deuxième valeur moyenne $\mathbf{m_2}$, sans doute différente de $\mathbf{m_1}$, et que nous espérons toujours assez proche de μ ; mais on ignore encore à quelle distance et de quel coté de μ cette valeur $\mathbf{m_2}$ se trouve.
Il en sera de même si nous disposons d'un troisième échantillon. La seule chose que nous pouvons espérer, c'est peut-être de mieux cerner la valeur μ mais là encore sans aucune certitude (figure 9-3).

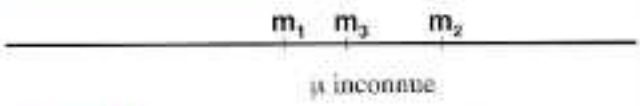


Figure 9-3.

Imaginons maintenant (bien que cela soit difficilement réalisable en pratique) disposer de la **totalité** des échantillons possibles tirés dans la population. Pour chaque échantillon, nous obtiendrions une moyenne $\mathbf{m}$.
Si nous classions ces valeurs et si nous disposions leurs fréquences sur un graphique, nous obtiendrions l'allure de la figure 9-4. On constate que l'ensemble de toutes les valeurs de $\mathbf{m}$ se répartit selon une courbe en forme de cloche. Cette courbe illustre les fluctuations de la moyenne. Et si on connaissait la moyenne μ de la population, on constaterait que cette courbe est centrée sur elle.

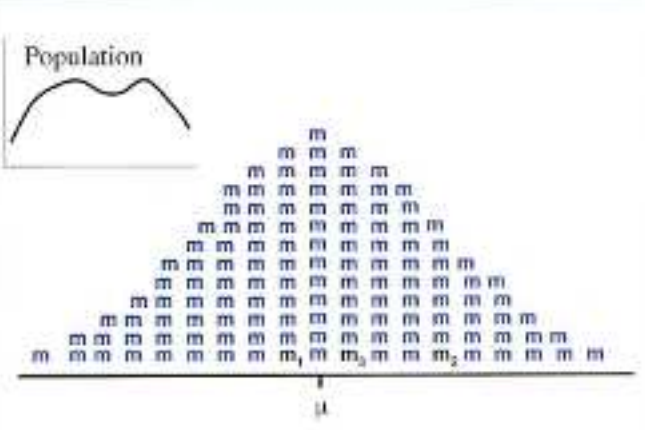


Figure 9-4. Fluctuation d'échantillonnage d'une moyenne

Cette image est l'illustration d'un théorème fondamental, sur lequel repose une grande part du raisonnement en statistique : le **théorème central limite**. Ce théorème énonce que :

1. La moyenne d'une variable quantitative calculée sur un échantillon est elle-même une variable aléatoire. Elle varie selon les échantillons.
2. Cette variable suit une **loi normale***.
3. Cette loi normale est centrée sur la moyenne μ de la population.

* à condition que les effectifs des échantillons soient égaux et suffisamment grands.

Ainsi, nous savons à présent que les moyennes des échantillons d'une distribution quelconque suivent une loi normale.

2. Écart type de la moyenne

Puisque la moyenne d'un échantillon est elle-même une variable aléatoire, on peut en calculer son écart type. On démontre que l'écart type de la moyenne **m** peut être estimé par la valeur :

$$s_m = \frac{s}{\sqrt{n}}$$

Notation **s** : écart type des valeurs de l'échantillon.
n : taille de l'échantillon.

Condition d'application : cette formule n'est valide que si la taille de l'échantillon est négligeable par rapport à la taille de la population (**n** inférieure à 10 % de la taille de la population). Si tel n'est pas le cas, il faut utiliser un facteur correctif « d'exhaustivité » (*cf.* Annexes, § 24.10).

Remarque importante : il ne faut pas confondre l'écart type des valeurs de l'échantillon **s** (§ 8.2.2) avec l'écart type de la moyenne s_m. Pour éviter cette confusion, on appelle parfois l'écart type de la moyenne (s_m) **erreur standard** (*standard error* en anglais).

3. Intervalle de confiance d'une moyenne

N'oublions pas que le but de notre démarche était de tenter d'estimer la valeur de la moyenne inconnue de la population à partir d'*une* observation sur *un seul* échantillon.
Il nous faut donc estimer un intervalle dans lequel la moyenne inconnue **μ** a la plus grande probabilité de se trouver.
On démontre (grâce au théorème central limite) qu'il y a 95 % de chances que la moyenne **μ** de la population se trouve comprise dans l'intervalle compris entre

$$m - 1{,}96s_m \qquad \text{et} \qquad m + 1{,}96s_m$$

On appelle cet intervalle, ***intervalle de confiance*** à 95 % de la moyenne μ.

On peut exprimer l'intervalle de confiance à 95 % par ces deux formules de signification équivalente :

$$m - 1{,}96 s_m < \mu < m + 1{,}96 s_m \qquad \text{ou bien :} \qquad \mu = m \pm 1{,}96 s_m$$

Notations μ : la moyenne inconnue de la population.
m : la moyenne calculée sur l'échantillon.
s_m : l'écart type *de la moyenne* (chap. 9.I.2).

Condition d'application : le calcul de l'intervalle de confiance par ces formules nécessite que la taille de l'échantillon soit supérieure ou égale à 30. Si tel n'est pas le cas, le terme 1,96 devrait être remplacé par une valeur choisie dans la table T de Student (*cf.* Annexes, § 24.11).

4. Signification de l'intervalle de confiance d'une moyenne

L'intervalle de confiance à 95 % d'une moyenne μ nous indique les bornes entre lesquelles on estime sa position. On ne connaît pas avec exactitude sa vraie valeur, mais on peut dire qu'elle a 95 chances sur 100 d'être comprise dans cet intervalle.
On peut dire en complément qu'il y a quand même 5 chances sur 100 pour que μ soit à l'extérieur de cet intervalle.

Exemple 9.1. CALCUL DE L'INTERVALLE DE CONFIANCE D'UNE MOYENNE

Lors d'une enquête sur la durée de sommeil des enfants de 2 à 3 ans effectuée sur un échantillon de 540 enfants d'un département français, on a trouvé une moyenne du temps de sommeil par nuit de 11,7 heures. L'écart type est 1,3 heure. On veut connaître la moyenne générale du temps de sommeil chez tous les enfants du département.

L'écart type de la moyenne est $s_m = \dfrac{1{,}3}{\sqrt{540}} = 0{,}056$ heure.

L'intervalle de confiance à 95 % est $11{,}7 \pm 1{,}96 \times 0{,}056 = 11{,}7 \pm 0{,}11$ heure.
La moyenne du temps de sommeil est donc comprise entre 11,6 et 11,8 heures.

II. ESTIMATION D'UN POURCENTAGE INCONNU

Lorsqu'on a observé un pourcentage sur un échantillon, le problème est d'estimer le véritable pourcentage **P** inconnu de la population d'où est extrait l'échantillon.

1. Fluctuation d'échantillonnage d'un pourcentage

Le raisonnement sur les fluctuations d'échantillonnage d'une moyenne (chap. 9.I.1) s'applique de la même manière pour un pourcentage. On démontre que :

1. Un pourcentage observé sur un échantillon est lui-même une variable aléatoire. Il varie selon les échantillons.
2. Cette variable suit **une loi normale***.
3. Cette loi normale est centrée sur le pourcentage **P** de la population.

* À condition que les effectifs des échantillons soient égaux et suffisamment grands.

2. Écart type d'un pourcentage

Puisqu'un pourcentage calculé sur un échantillon est lui-même une variable aléatoire, on peut en calculer son écart type. On démontre que l'écart type du pourcentage **p** peut être estimé par la valeur suivante :

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

Cette formule n'est valide que si la taille n de l'échantillon est négligeable par rapport à la taille de la population (**n** inférieure à 10 % de la taille de la population). Si tel n'est pas le cas, il faut utiliser un facteur correctif « d'exhaustivité » (*cf.* Annexes, § 24.10).

3. Intervalle de confiance d'un pourcentage

N'oublions pas que le but de notre démarche était de tenter d'estimer la valeur du pourcentage inconnu de la population à partir d'*une* observation sur un *seul* échantillon.
Il nous faut donc estimer un intervalle dans lequel le pourcentage inconnu **P** a la plus grande probabilité de se trouver.
On démontre (grâce au théorème central limite) qu'il y a 95 % de chances que le pourcentage **P** de la population se trouve compris dans l'intervalle compris entre :

$$p - 1{,}96s_p \qquad \text{et} \qquad p + 1{,}96s_p$$

On appelle cet intervalle, ***intervalle de confiance*** à 95 % du pourcentage **P**.
On peut exprimer l'intervalle de confiance à 95 % par ces deux formules de signification équivalente :

$$p - 1{,}96s_p < P < p + 1{,}96s_p \qquad \text{ou bien} \qquad P = p \pm 1{,}96s_p$$

Notations **P** : le pourcentage inconnu de la population.
p : le pourcentage calculé sur l'échantillon.
$\mathbf{s_p}$: l'écart type du pourcentage (chap. 9.II.2).

Condition d'application

Ces formules nécessitent que l'effectif de l'échantillon soit suffisamment grand. Si on appelle p_i et p_s les bornes inférieures et supérieures de l'intervalle de confiance (calculées comme si les conditions étaient remplies), il faut que les termes np_i, np_s, $n(1 - p_i)$, $n(1 - p_s)$ soient supérieurs ou égaux à 5. Si l'un de ces termes est inférieur à 5, l'intervalle de confiance ne serait pas valide. Il faudrait renoncer à ce résultat et recourir à la loi binomiale. Les calculs sont complexes et peuvent être réalisés à l'aide de tableur. Certains ouvrages de statistiques présentent des tables fournissant directement les intervalles de confiance d'un pourcentage en fonction des effectifs de l'échantillon.

4. Signification de l'intervalle de confiance d'un pourcentage

L'intervalle de confiance à 95 % d'un pourcentage **P** nous indique les bornes entre lesquelles on estime sa position. On ne connaît pas avec exactitude sa vraie valeur, mais on peut dire qu'il a 95 chances sur 100 d'être compris dans cet intervalle.
On peut dire en complément qu'il y a quand même 5 chances sur 100 pour que **P** soit à l'extérieur de cet intervalle.

Exemple 9.2. CALCUL DE L'INTERVALLE DE CONFIANCE D'UN POURCENTAGE

Lors d'une enquête sur la durée de sommeil des enfants de 2 à 3 ans effectuée sur un échantillon de 540 enfants d'un département français, on a trouvé 86 enfants présentant des troubles du sommeil. On veut connaître la proportion de troubles du sommeil chez tous les enfants du département. La proportion d'enfants présentant des troubles du sommeil dans l'échantillon est de 86/540 = 15,9 %.

L'écart type s_p est : $\sqrt{\frac{0{,}159\,(1-0{,}159)}{540}} = 0{,}016$

L'intervalle de confiance à 95 % est : $0{,}159 \pm 1{,}96 \times 0{,}016 = 0{,}159 \pm 0{,}031$.
La proportion d'enfants présentant des troubles dans ce département est donc comprise entre 12,8 % et 19,0 %.

III. RISQUE D'ERREUR CONSENTIE α

Nous avons jusqu'à présent estimé une moyenne ou un pourcentage inconnu avec un intervalle de confiance à 95 %, c'est-à-dire avec un risque d'erreur de 5 %. On appelle ce risque d'erreur, **risque α**.
Ce risque était déterminé par notre choix d'une valeur de 1,96 dans les formules.
Il ne serait pas raisonnable de choisir un risque d'erreur plus élevé, mais rien ne nous empêche de choisir un risque moindre.
Il faudrait alors remplacer le nombre 1,96 par une autre valeur.

α	$\lvert Z_\alpha \rvert$
• 20 %	1,28
• 10 %	1,65
• 5 %	1,96
• 2 %	2,33
• 1 %	2,58
• 0,1 %	3,3

Figure 9-5. Valeurs de Zα pour quelques risques usuels

La correspondance entre le risque **α** consenti et ces valeurs sont fournies par la table de Z (loi normale centrée réduite) figurant à la fin de l'ouvrage (Table 1). Pour chaque valeur du risque α, il existe une valeur Z_α. La figure 9-5 en montre un extrait condensé.

Les formules d'intervalle de confiance d'une moyenne et d'un pourcentage peuvent être généralisées ainsi :

Moyenne : $\mu = m \pm Z_\alpha s_m$

Pourcentage : $P = p \pm Z_\alpha s_p$

Exemple 9.3. CHOIX D'UN RISQUE α

Un enquêteur prudent serait tenté de choisir un risque α faible. Prenons, par exemple, 1 % au lieu de 5 %. Il voudrait donc obtenir un intervalle de confiance à 99 %.
Pour un risque α de 1 %, la valeur de Z lue dans la table est 2,58. Le calcul de l'intervalle de confiance à 99 % d'une moyenne ou d'un pourcentage donnerait respectivement :

$$\mu = m \pm 2{,}58\, s_m \quad \text{ou} \quad P = p \pm 2{,}58\, s_p$$

Cet intervalle de confiance à 99 % est plus large que l'intervalle de confiance à 95 %. Cet enquêteur prudent a donc moins de chances de se tromper, mais il fournit une estimation moins précise. En reprenant l'exemple 9.2, la proportion d'enfants avec des troubles du sommeil serait estimée entre 11,8 % et 20,0 %.

Ainsi, le choix d'un risque d'erreur plus faible se paye au prix d'un intervalle de confiance plus large, donc d'une estimation moins précise. Le consensus général adopté par l'ensemble de la communauté scientifique est de présenter des intervalles de confiance d'au moins 95 %.

IV. TAILLE D'UN ÉCHANTILLON

1. Précision d'une estimation

La précision d'une estimation dépend de deux facteurs que l'on peut contrôler lorsqu'on bâtit une étude.

- **Le choix du risque d'erreur.** Ce choix détermine la valeur Z_α qui entre dans les formules générales des intervalles de confiance (chap. 9.III). Plus α est petit, plus Z_α est grand. Nous avons vu que pour un risque α de 5 %, la valeur Z_α était de 1,96.
- **La taille n de l'échantillon.** Ce facteur intervient dans les formules qui déterminent l'écart type de la moyenne ou du pourcentage (chap. 9.I.2 et 9.II.2). Dans ces deux formules, la taille **n** figure au dénominateur. On en déduit que :
  - plus la taille de l'échantillon est grande ;
  - plus l'écart type $\mathbf{s_m}$ ou $\mathbf{s_p}$ est petit ;
  - plus l'intervalle de confiance est resserré ;
  - et donc plus grande est la précision.

2. Calcul de la taille d'un échantillon

Lorsqu'on bâtit une enquête, il faut définir une taille d'échantillon compatible avec les moyens dont on dispose, mais suffisamment grande pour que l'intervalle de confiance ne soit pas démesuré. Il serait désastreux en effet, après un long travail d'échantillonnage et de recueil de données d'aboutir à un intervalle de confiance d'une moyenne de 15,3 ± 14 ou d'un pourcentage variant de 15 % à 85 %.

Il existe des formules permettant de calculer la taille minimale d'un échantillon pour obtenir une précision désirée. Ces formules sont valables uniquement pour des échantillons provenant de sondage aléatoire élémentaire.

pour une moyenne	pour un pourcentage
$n = \sigma^2 \frac{Z_\alpha^2}{i^2}$	$n = P(1-P) \frac{Z_\alpha^2}{i^2}$

Ces formules sont assez délicates à utiliser car elles nécessitent de faire des choix et elles comportent des termes que nous ne connaissons pas.

- $\mathbf{Z_\alpha}$. On prend en général la valeur 1,96. Si on désire un risque α plus faible, cette valeur sera plus élevée et la taille de l'échantillon aussi.
- $\boldsymbol{\sigma^2}$ est la variance de la variable quantitative étudiée *dans la population*. Mais on ne la connaît pas *a priori*. On l'estime d'après des études antérieures sur le même sujet, ou au besoin par une étude pilote.
- **P** est le pourcentage de la variable qualitative étudiée *dans la population*. Par définition ce pourcentage est inconnu, puisqu'on réalise l'enquête dans le but de le connaître ! Là aussi on l'estime sur des études antérieures ou par une étude pilote.
- **i** est la précision désirée, c'est-à-dire la moitié de l'intervalle de confiance. Par exemple, si on veut estimer la moyenne du poids des individus dans une population on peut exiger une précision de ± 3 kg. Si on veut estimer le pourcentage de sujets malades, on peut exiger une précision de ± 4 % (± 0,04). On constate dans les formules que la précision **i** se trouve au dénominateur. Si l'on désire une précision trop élevée, **i** est choisi très petit et la taille de l'échantillon sera très grande.

Cette relation entre précision et taille de l'échantillon n'est pas linéaire. Comme le montre la figure 9-6, l'intervalle de confiance, diminue fortement lorsque la taille de l'échantillon augmente dans les valeurs faibles. Mais lorsque la taille de l'échantillon devient élevée, le gain en précision est dérisoire. Il y a donc un seuil optimal de taille d'échantillon qui représente le meilleur compromis entre une précision souhaitable et une taille d'échantillon compatible avec les moyens dont on dispose.

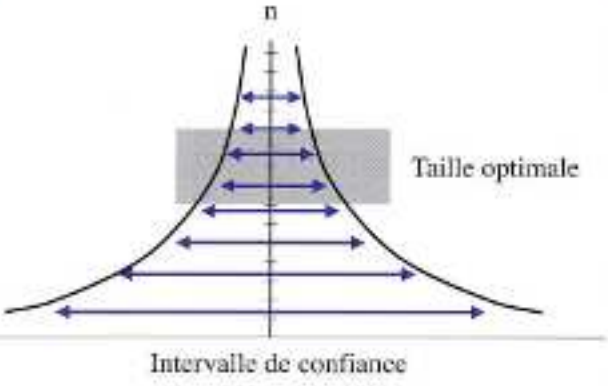


Figure 9-6. Précision obtenue en fonction de la taille de l'échantillon

Exemple 9.4. CALCULS DE TAILLE D'ÉCHANTILLON

On désire estimer la proportion de troubles du sommeil chez les enfants de 2 à 3 ans d'un département français. Des études antérieures pratiquées dans d'autres régions montrent que la proportion de ces troubles est d'environ 16 %. On désire une précision de ± 3 % et on choisit un risque α de 5 %.

La taille de l'échantillon nécessaire est $n = 0{,}16(1-0{,}16)\dfrac{1{,}96^2}{0{,}03^2} = 574$

Les commanditaires de l'enquête jugent que la précision de ± 3 % est insuffisante, et exigent une précision de ± 2 %.

La taille de l'échantillon serait $n = 0{,}16(1-0{,}16)\dfrac{1{,}96^2}{0{,}02^2} = 1\,291$

On voit que, pour gagner 1 % de précision, la charge de travail sera doublée !

Exercices

Exercice 9.1

On a mesuré la glycémie d'un échantillon de 25 sujets représentatifs d'une population d'étude. On trouve une moyenne de 1,52 g/L et un écart type de 0,40 g/L.

1) Calculez l'intervalle de confiance à 95 % de cette moyenne.

2) Calculez l'intervalle de confiance à 99 % de cette moyenne.

Exercice 9.2

Pour connaître la fréquence des parasitoses dans un département d'outre-mer de 350 000 habitants, on pratique une enquête sur un échantillon de 3 500 personnes. On dépiste, parmi eux, 1 050 sujets atteints d'une parasitose.

Calculez la fréquence estimée des parasitoses dans ce département et son intervalle de confiance à 95 %.

Exercice 9.3

Parmi 12 malades ayant subi une greffe de moelle dans un service d'hématologie, 8 ont contracté une infection fungique (mycose).

Calculez le pourcentage des complications d'infection fongique et son intervalle de confiance à 95 %.

Exercice 9.4

Parmi les 200 élèves d'une école, 70 ont été tirés au sort pour subir un examen dentaire. Dix cas de carie ont été détectés. Quel est le pourcentage estimé de caries dentaires dans l'école et son intervalle de confiance à 95 % ?

Exercice 9.5

Dans une région d'endémie du paludisme, on désire mesurer la fréquence des enfants impaludés résistants au traitement par la chloroquine. On décide donc de compter la proportion d'échecs au traitement sur un échantillon d'enfants impaludés et traités par ce produit.

De quelles données doit-on disposer pour calculer la taille de l'échantillon ?

Résumé

ESTIMATION D'UNE MOYENNE ET D'UN POURCENTAGE

Le calcul d'une moyenne ou d'un pourcentage observé sur un échantillon représentatif d'une population a pour but d'estimer la moyenne ou le pourcentage inconnu dans cette population.
On estime ces paramètres inconnus en calculant un ***intervalle de confiance***. Cet intervalle est composé de deux bornes entre lesquelles la valeur inconnue du paramètre a la plus grande probabilité de se situer. Le degré de confiance qu'on exige est déterminé dans la formule de calcul.
On se fixe en général un intervalle de confiance à 95 %.

Pour une moyenne m, l'intervalle de confiance à 95 % = $m \pm 1{,}96 \frac{s}{\sqrt{n}}$

Pour un pourcentage p, l'intervalle de confiance à 95 % = $p \pm 1{,}96 \sqrt{\frac{p(1-p)}{n}}$

Cela signifie que le paramètre a 95 chances sur 100 d'être situé dans cet intervalle. En corollaire, cela signifie qu'il existe un risque d'erreur (α) de 5 %.
Si on désire une plus grande certitude d'encadrer la valeur du paramètre inconnu, on choisit un risque d'erreur α plus faible, en regardant dans la table Z, la valeur Z correspondant à ce risque. Cette valeur est plus élevée que 1,96, l'intervalle proposé sera donc plus large et la précision de l'estimation sera moins bonne.
La largeur de l'intervalle de confiance dépend aussi de la taille de l'échantillon. Plus la taille de l'échantillon est élevée, meilleure est la précision.

Troisième partie

TESTS STATISTIQUES

TESTS STATISTIQUES

Introduction

PRINCIPE DES TESTS

I. PRINCIPE DES TESTS DE COMPARAISON
II. PRINCIPE DES TESTS DE LIAISON

TESTS DE COMPARAISON

I. TEST Z OU TEST DE L'ÉCART RÉDUIT
II. TEST T DE STUDENT
III. TEST F DE FISHER-SNEDECOR
IV. TESTS DE χ^2
V. TEST EXACT DE FISHER
VI. TESTS NON-PARAMÉTRIQUES OU TESTS DE RANGS

TESTS DE LIAISON

I. TEST DU χ^2 D'INDÉPENDANCE
II. TEST DU χ^2 DE TENDANCE
III. TESTS DE CORRÉLATION
IV. RÉGRESSION

UTILISATION PRATIQUE DES TESTS STATISTIQUES

I. CRITÈRES DE CHOIX D'UN TEST STATISTIQUE
II. STRATÉGIE D'UTILISATION DES TESTS STATISTIQUES
III. TEST Z POUR COMPARER UNE MOYENNE OBSERVÉE À UNE MOYENNE THÉORIQUE
IV. TEST Z POUR COMPARER DEUX MOYENNES
V. TEST Z POUR COMPARER DEUX MOYENNES SUR DEUX SÉRIES APPARIÉES
VI. TEST T POUR COMPARER UNE MOYENNE OBSERVÉE À UNE MOYENNE THÉORIQUE
VII. TEST T DE STUDENT POUR COMPARER DEUX MOYENNES
VIII. TEST T POUR COMPARER 2 MOYENNES SUR 2 SÉRIES APPARIÉES
IX. TEST F POUR COMPARER DEUX VARIANCES
X. TEST F POUR COMPARER PLUSIEURS MOYENNES
XI. TEST DE WILCOXON
XII. TEST DE WILCOXON POUR SÉRIES APPARIÉES
XIII. TEST DE KRUSKAL-WALLIS (KW)
XIV. TEST DE χ^2 DE CONFORMITÉ OU D'AJUSTEMENT
XV. TEST DE χ^2 D'HOMOGÉNÉITÉ
XVI. TEST DE χ^2 À 4 CASES POUR COMPARER DEUX POURCENTAGES
XVII. TEST DE χ^2 DE McNEMAR POUR SÉRIES APPARIÉES
XVIII. TEST DE χ^2 D'INDÉPENDANCE
XIX. TEST DE χ^2 DE TENDANCE
XX. TEST DU COEFFICIENT DE CORRÉLATION
XXI. TEST DU COEFFICIENT DE CORRÉLATION DES RANGS DE SPEARMAN

TESTS STATISTIQUES DIVERS

I. ÉPREUVE DE NORMALITÉ
II. TEST DE BARTLETT
III. TEST DE LEVENE
IV. CORRÉLATION LINÉAIRE MULTIPLE
V. RÉGRESSION LINÉAIRE MULTIPLE

Introduction

Le test statistique est l'outil de la comparaison, de même que le calcul de l'intervalle de confiance était l'outil statistique de l'estimation.
Une comparaison statistique porte sur des séries de données qui sont résumées en moyenne, pourcentage, distribution par classes, indicateur de liaison entre 2 variables, *etc.*
Nous pouvons assimiler un test à une pesée (figure 10.1).

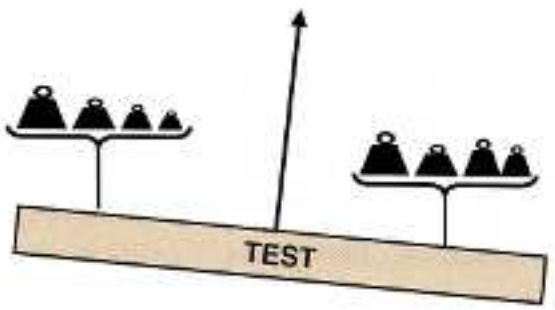


Figure 10-1.

Lorsqu'on effectue une comparaison entre deux ou plusieurs séries de données, on observe toujours une différence, plus ou moins grande entre les paramètres mesurés. Le but du test est de déterminer si la différence observée est simplement due au hasard, c'est-à-dire aux fluctuations d'échantillonnage, ou si au contraire la différence observée est bien réelle.
Au total, les tests servent à extrapoler les résultats observés sur des échantillons à l'ensemble des populations dont ils sont issus. L'échantillon n'est qu'une image ponctuelle, les observations qu'on en tire n'ont aucun intérêt en tant que telles, et n'ont de valeur que si on les extrapole à la population d'où est issu l'échantillon.
L'intérêt majeur des tests statistiques, est donc de réaliser une économie énorme de moyens, en permettant de déceler des différences sur un nombre réduit d'observations.
En contrepartie, il faut admettre l'existence d'un certain flou dans les conclusions, il faut assumer un risque d'erreur. Les tests statistiques sont conçus pour déterminer ce risque d'erreur.
La réalisation d'un test suppose quelques réflexions préalables concernant ses conditions d'utilisation et d'application.

Conditions d'utilisation d'un test

Un test statistique doit être réalisé dans le cadre d'une réflexion scientifique qui consiste à bâtir des hypothèses à partir de faits antérieurs observés. Ensuite, ces hypothèses sont testées et selon les résultats des tests, elles sont soit rejetées, soit acceptées. Puis de nouvelles hypothèses peuvent ensuite être bâties et à nouveau testées.
Les résultats d'un test n'ont de valeur que s'ils sont inscrits dans cette démarche logique.
Prenons un contre-exemple : on décide de comparer la taille des individus qui passent le samedi après-midi sur les trottoirs de droite et de gauche du boulevard Saint-Michel à Paris. Malgré, le côté surréaliste de cette enquête, il n'est pas impossible de mettre en évidence une différence, peut-être même « statistiquement significative » après les calculs appropriés. Mais ce résultat n'aurait aucune signification, aucun sens réel. La démarche, qui consisterait, à rechercher *a posteriori* une explication à cette observation serait absurde.

Un test statistique n'a de sens que s'il teste une hypothèse ***préalablement posée*** afin de répondre à une question. Tout test statistique doit donc avoir pour objectif de vérifier une hypothèse justifiée, soit par la connaissance acquise à partir d'autres études, soit à partir de conjectures produites par des observations.

OBSERVATION ⟶ HYPOTHÈSE ⟶ TEST

Conditions d'application

Tous les tests sont basés sur des modèles « stochastiques », c'est-à-dire sur des lois de distributions théoriques issues de la théorie des probabilités. Ces lois sont strictes. Elles sont donc accompagnées de ***conditions d'application***, tout aussi strictes. Si ces conditions ne sont pas remplies, le modèle n'est pas applicable, et le résultat du test même s'il est juste arithmétiquement, n'a plus aucun sens statistique. La première condition est donc celle qui suppose qu'il existe une adéquation entre la distribution étudiée et la distribution théorique sur laquelle est basé le test. Parmi les autres conditions d'application, figurent la comparabilité des échantillons, leur taille, la forme supposée de la distribution dans les populations d'où ils sont issus, et par-dessus tout le seul rôle du hasard dans la sélection des échantillons comparés.

Si dans cet ouvrage je vous fais grâce de la modélisation et de la théorie, il faut impérativement qu'en échange, vous prêtiez une attention rigoureuse aux conditions d'application. Elles sont plus importantes que les calculs, qui du reste sont effectués instantanément par les logiciels. Utilisez votre temps et votre jugement à vérifier les conditions d'application des tests et à interpréter leurs résultats plutôt qu'à faire des calculs.

- Nous avons vu qu'une série d'observations portant sur une variable peut être décrite :
  - soit par des ***paramètres*** résumant la distribution : moyenne, pourcentage, variance ;
  - soit par la ***distribution*** des effectifs sous forme de tableau ou de diagramme.

- Il existe parallèlement deux familles de tests (figure 10.2) :
  - les tests ***paramétriques*** qui comparent des paramètres ;
  - les tests ***semi-paramétriques*** (test de χ^2) et les tests de ***rang*** qui comparent des distributions.

Il existe deux situations d'utilisation des tests statistiques : les tests de comparaison entre des séries d'individus (ou d'unités statistiques) et les tests de liaison entre deux variables.

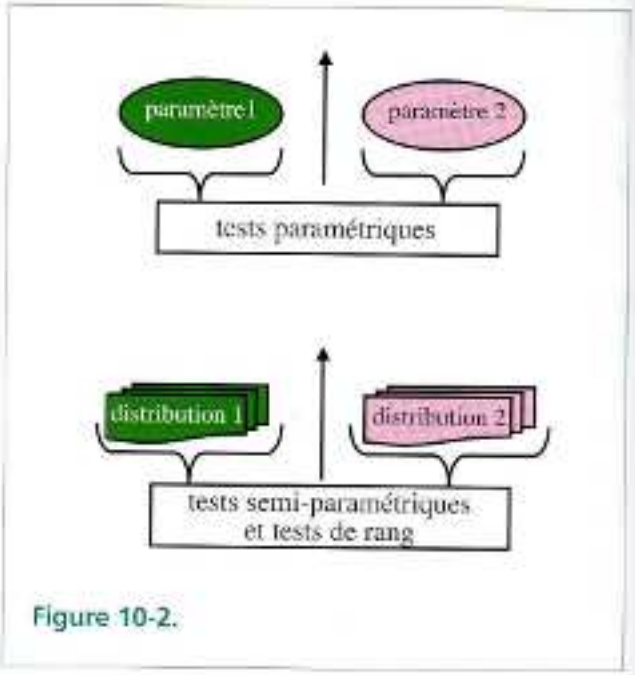


Figure 10-2.

PRINCIPE DES TESTS

I. PRINCIPE DES TESTS DE COMPARAISON

Les tests de comparaison servent à comparer des séries de données entre elles. Il existe schématiquement deux situations de comparaison.

- Comparer un échantillon observé à une population de référence (figure 10.3).
  On se demande si la distribution de **la population** dont est issu l'échantillon est identique à la distribution théorique, ou bien si elle est différente.
- Comparer deux ou plusieurs échantillons entre eux (figure 10.4).
  On se demande si les distributions **des populations** dont sont issus les échantillons sont identiques ou différentes.

Ainsi, dans ces deux situations, l'objet du test est de comparer des populations.
Nous ne développons ici que les principes très succincts sans entrer dans la théorie des tests.
Le principe général d'un test est de regarder si la différence qu'on observe est due au hasard ou si au contraire cette différence est telle qu'il est fort peu probable de l'observer par hasard.
Quelle que soit la nature d'un test, son principe et son déroulement sont toujours les mêmes.

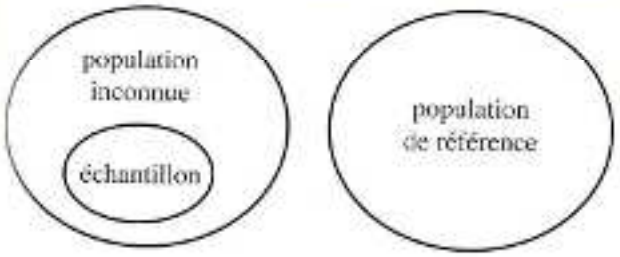


Figure 10-3.

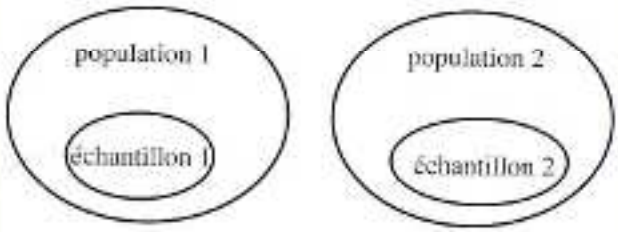


Figure 10-4.

1. Établir l'hypothèse nulle (H_0)

Cela consiste à poser *a priori* l'hypothèse que les paramètres ou les distributions des populations d'où sont issus les échantillons étudiés sont identiques (figure 10.5).

À moins d'un fort coup de chance, on observera toujours une différence entre les paramètres ou les distributions comparés. Cette différence peut être plus ou moins grande. Proposer

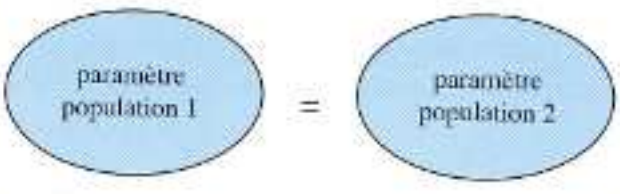


Figure 10-5. Hypothèse nulle H_0

l'hypothèse nulle, c'est supposer que la différence observée provient seulement des fluctuations d'échantillonnage.

2. Proposer une hypothèse alternative (H_1) (exemple 10.1)

On appelle hypothèse alternative H_1, l'hypothèse qui sera retenue au cas où les résultats du test aboutiraient à rejeter l'hypothèse nulle H_0. Rejeter H_0 c'est dire que H_0 est fausse. C'est dire que la différence observée est trop grande pour qu'on l'attribue à une simple fluctuation d'échantillonnage. On suppose donc dans ce cas que les paramètres ou les distributions des populations d'où sont issus les échantillons étudiés sont différents.
C'est ce qu'on nomme **l'hypothèse alternative H_1**.
Selon le type du problème posé, on propose une hypothèse alternative bilatérale ou unilatérale.

- **H_1 bilatérale.** L'hypothèse alternative est bilatérale lorsqu'on ne cherche pas à connaître le sens de la différence. On se contente de postuler que les deux paramètres ou les deux distributions sont différents (figure 10.6).
- **H_1 unilatérale.** L'hypothèse alternative est unilatérale lorsqu'on s'intéresse à un sens particulier de l'inégalité de 2 paramètres tel que :

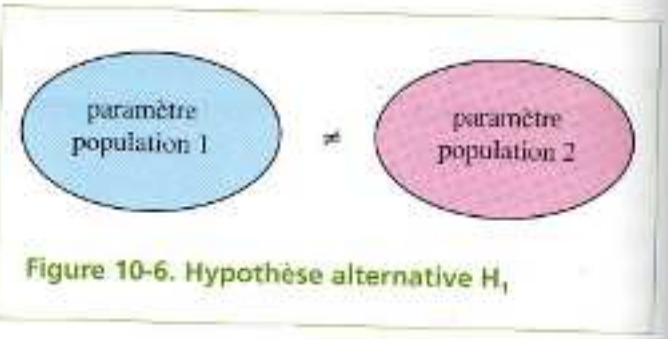


Figure 10-6. Hypothèse alternative H_1

paramètre 1 > paramètre 2 ou paramètre 1 < paramètre 2

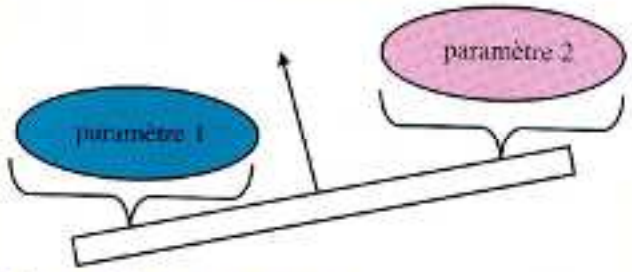


Figure 10-7. H_1 unilatérale

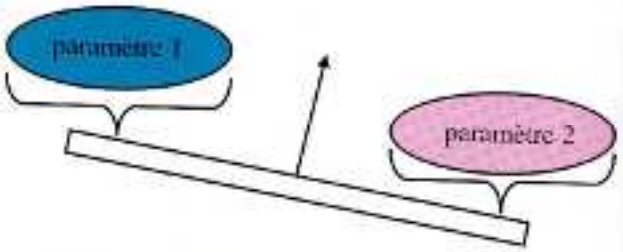


Figure 10-8. H_1 unilatérale

3. Calcul d'un test de comparaison

Une fois que les hypothèses sont clairement posées, le test est appliqué. Tous les tests statistiques de comparaison consistent :
- à calculer une quantité mathématique exprimant l'écart entre les paramètres ou les distributions ;
- à confronter cette quantité à un modèle de distribution théorique.

L'écart peut être exprimé :
- soit par une différence (différence entre 2 moyennes ou 2 pourcentages) ;
- soit par un rapport (par exemple s^2_1/s^2_2 lorsqu'on compare 2 variances) ;
- soit par un indicateur plus complexe lorsqu'on compare des distributions.

Exemple 10.1. PROPOSITION D'HYPOTHÈSES

On veut comparer la fréquence du paludisme dans deux régions d'Afrique. Si l'on appelle respectivement P_1 et P_2 les fréquences des individus infectés dans ces deux régions, on pose :
- hypothèse nulle : les deux fréquences sont identiques. $H_0 : P_1 = P_2$;
- hypothèse alternative : les deux fréquences sont différentes : $H_1 : P_1 \neq P_2$.

Il s'agit ici d'une hypothèse alternative bilatérale, car on ignore a priori dans quelle région la fréquence est la plus élevée.

On désire tester un vaccin contre le paludisme en comparant la survenue du paludisme entre un groupe vacciné et un groupe témoin non vacciné.
Si l'on appelle respectivement P_1 et P_2 les fréquences des individus infectés dans chacune des 2 populations représentées par les 2 groupes, on pose :
- hypothèse nulle : les deux fréquences sont identiques. $H_0 : P_1 = P_2$. Ceci revient à dire que le vaccin n'a aucune efficacité ;
- hypothèse alternative : la fréquence des individus infectés dans le groupe vacciné est inférieure à la fréquence dans le groupe non vacciné. $H_1 : P_1 < P_2$.

Il s'agit ici d'une hypothèse alternative unilatérale car on s'intéresse dans ce cas exclusivement aux effets bénéfiques attendus du vaccin. Une différence observée dans l'autre sens (fréquence du paludisme plus élevée dans le groupe vacciné) n'aurait évidemment aucun intérêt.

Quelle que soit la quantité calculée, elle peut être considérée elle-même comme une valeur particulière d'une variable aléatoire, puisqu'elle est calculée à partir d'observations faites sur des échantillons. Sous l'hypothèse nulle H_0, cette variable suit une loi de probabilité. Une grande partie des travaux fondamentaux des statisticiens a été de rechercher pour chaque test possible sa loi de distribution sous H_0.
Appelons $\mathbf{u_e}$ l'expression mathématique de l'écart observé et **U** le modèle théorique de sa distribution sous H_0.
Le test statistique consiste à examiner la position de la valeur $\mathbf{u_e}$ par rapport au modèle théorique **U** dont on connaît la loi de distribution, la forme, et la formulation mathématique.
Ainsi on examine si $\mathbf{u_e}$ est une valeur « assez » probable de la loi **U** ou au contraire si $\mathbf{u_e}$ a une valeur « trop » éloignée pour admettre qu'il s'agit d'une simple fluctuation d'échantillonnage. Prenons à titre d'exemple une variable aléatoire **U** ayant une loi de distribution représentée par la courbe de la figure 10.9. Sa forme, même si elle vous rappelle quelque chose, est ici indifférente.

Nous savons que **l'aire** située sous la courbe entre deux valeurs données de U représente la **probabilité** que U soit comprise entre ces deux valeurs (chap. 5.II). Nous pouvons donc fixer deux seuils, u_1 et u_2, qui délimiteraient trois zones (figure 10.9) : la zone centrale, comprise entre les deux seuils, zone dont l'aire représente les valeurs « assez » probables de U et les deux zones extérieures à cet intervalle dont l'aire totale représente les valeurs « peu » probables de U.

Appelons α, la somme $\alpha_1 + \alpha_2$ des aires des deux zones extérieures peu probables. (figure 10.10).
On en déduit que :
- la probabilité que U soit inférieur à u_1 est égale à α_1 ;
- la probabilité que U soit supérieur à u_2 est égale à α_2 ;
- la probabilité que U soit extérieur à l'intervalle $[u_1,u_2]$ est égale α.

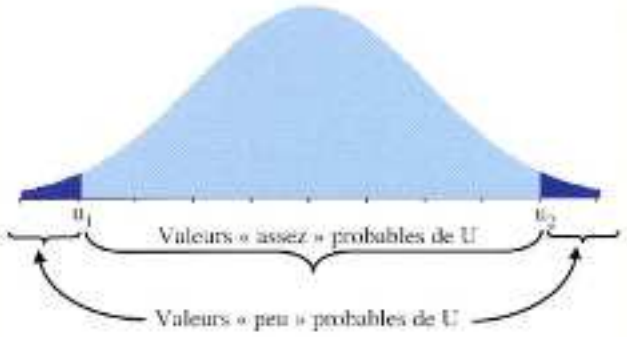


Figure 10-9. Loi de distribution de la variable U

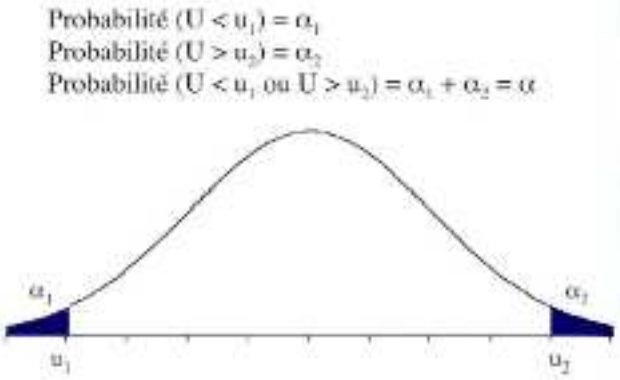


Figure 10-10. Loi de distribution de U

4. Résultats d'un test de comparaison

Selon les résultats du test, on peut donc se trouver dans deux situations.

- La valeur u_0 est comprise dans l'intervalle $[u_1\text{-}u_2]$ (figure 10.11). On en conclut que la différence observée entre les paramètres ou les distributions étudiés n'est **pas significative**. La différence peut s'expliquer par les seules fluctuations d'échantillonnage. **On ne peut pas rejeter H_0.**
- La valeur u_0 est extérieure à l'intervalle des valeurs limite $[u_1\text{-}u_2]$ (figure 10.12). Il est encore possible que ce résultat soit du à une simple fluctuation d'échantillonnage (il reste en effet α chances que cela soit vrai). Mais on décide de ne pas tenir compte de cette faible probabilité. **On rejette l'hypothèse nulle H_0 et on accepte l'hypothèse H_1** d'une différence réelle entre les paramètres ou les distributions étudiés. On dit que cette différence est significative.

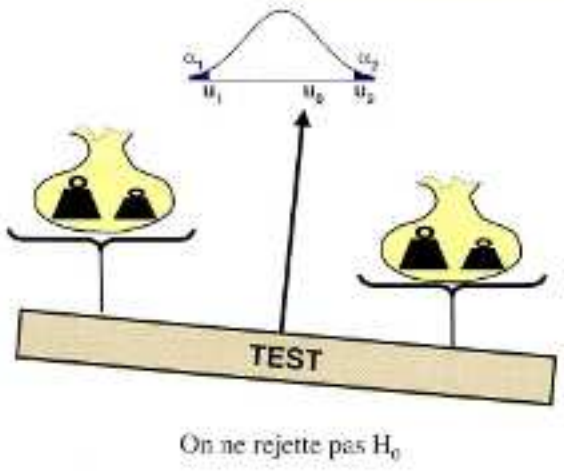


Figure 10-11.

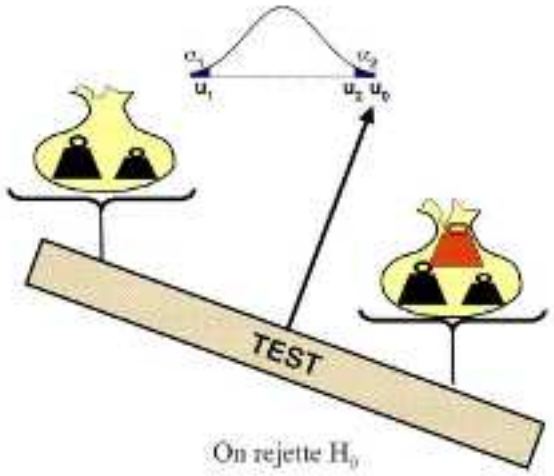


Figure 10-12.

5. Choix du risque d'erreur

a) Le risque α

En rejetant H_0, on prend un certain risque : c'est la probabilité α d'observer des valeurs rares de la variable U, si H_0 est vraie. En d'autres termes, c'est le risque de se tromper en rejetant H_0, si par malheur H_0 était vraie. Ce prix à payer lorsqu'on travaille sur des échantillons est le risque d'affirmer une différence, alors qu'elle n'existe pas.

On l'appelle **risque de première espèce ou risque α.**

α = probabilité de rejeter H_0, si H_0 est vraie

Lorsqu'on réalise un test on se fixe donc un seuil au-delà duquel on accepte de prendre ce risque. Ce seuil est fixé *a priori*.
On assigne communément à α, la valeur 5 %. Ce choix de 5 % est universellement admis par tous les statisticiens, mais libre à vous d'en choisir un autre.
Les prudents choisiront un risque plus bas. Ils auront donc moins de probabilité de se tromper en rejetant H_0, si H_0 est vraie. Mais en échange, ils concluront moins souvent. S'ils choisissent un risque trop bas, ils ne démontreront jamais rien.
À l'inverse, les téméraires qui voudraient conclure à tout prix opteraient pour un risque d'erreur plus élevé. Mais les conséquences d'une fausse différence peuvent être désastreuses, notamment dans les essais thérapeutiques.
La quasi-totalité des publications scientifiques n'admettent pas le choix d'un risque supérieur à 5 % pour affirmer une différence significative.
Certains suggèrent que l'origine de ce risque maximum provient d'une forte tradition des parieurs britanniques qui admettaient de risquer un shilling pour un pari d'une livre (qui autrefois valait 20 shillings, soit un rapport de 5 %). Cette origine pourrait tout autant être française, puisque jadis un franc valait 20 sous. Le français près de ses sous, tout autant qu'un anglais, ne devait pas être prêt à en risquer plus d'un pour gagner un franc hypothétique.
En résumé, le risque α est le risque de conclure à une différence qui n'existe pas.

b) Risque β et puissance d'un test

Un deuxième risque d'erreur existe : celui de ne pas avoir rejeté H_0 alors que H_1 était vrai. Ceci arrive lorsqu'il existe bel et bien une différence entre les paramètres étudiés, mais la valeur observée u_o se situe néanmoins dans l'intervalle comprenant 95 % des valeurs probables de U. L'observateur, myope en quelque sorte, ne voit pas la différence et ne rejette pas H_0.

Ce risque est appelé **risque de deuxième espèce ou risque β.**

β = probabilité de ne pas rejeter H_0, si H_1 est vraie

Le risque **β** est aussi appelé « manque de puissance ». Par opposition, on appelle puissance d'un test la valeur **1 – β**.
La puissance d'un test est liée à la taille des effectifs des échantillons. Plus la taille des échantillons comparés augmente, plus la puissance augmente et plus le risque **β** diminue.
La valeur du risque **β** n'intervient pas dans l'interprétation d'un test car on ne sait pas la calculer. Il faut cependant toujours tenir compte de ce risque lorsqu'on ne rejette pas H_0. C'est une des raisons pour laquelle on ne peut jamais vérifier qu'une hypothèse nulle est vraie, car si la taille de l'échantillon avait été beaucoup plus grande on aurait peut-être pu la rejeter.
En pratique, on utilise le risque **β** dans le calcul de la taille des échantillons lorsqu'on bâtit une étude (chap. 11.1.4).

6. Interprétation finale d'un test de comparaison

a) L'hypothèse nulle n'est pas rejetée

Cela signifie que rien ne permet d'affirmer que les paramètres ou les distributions comparés sont différents. C'est d'ailleurs la seule chose que l'on peut dire. On n'affirme jamais qu'une hypothèse nulle est vraie, car :

- elle aurait peut-être pu être rejetée si la puissance du test avait été plus élevée (*cf.* risque **β**) ;
- il existe de nombreuses raisons qui peuvent expliquer que les distributions ou paramètres mesurés sont identiques bien que les populations soient complètement distinctes.

b) On rejette H_0

Le corollaire de ce rejet est l'acceptation de l'hypothèse H_1.

- Si on avait choisi l'hypothèse alternative H_1 ***bilatérale***, on l'accepte en affirmant que les distributions ou paramètres étudiés sont ***différents***.
- Si on avait choisi l'hypothèse alternative H_1 **unilatérale** on l'accepte en affirmant que l'un des paramètres est **inférieur (ou bien supérieur)** à l'autre. En effet, dans ce cas on s'intéresse au sens de la différence. On ne s'intéresse qu'à une seule extrémité de la distribution de U. Le risque de se tromper est donc deux fois moindre que dans l'hypothèse bilatérale.

c) Le degré de signification *p*

Le risque α est un seuil fixé *a priori* (le plus souvent à 5 %). Lorsque le calcul du test montre que le seuil U_α a été franchi, on rejette H_0 avec un risque égal à α. Mais on désire aller plus loin et préciser la limite du risque pris. On appelle **degré de signification *p***, cette limite *a posteriori*. C'est la probabilité, si H_0 était vraie, d'observer la valeur u_0 avec les valeurs obtenues à partir les échantillons étudiés. En d'autres termes le degré de signification indique la probabilité d'avoir rejeté H_0 si on avait fixé le risque de 1re espèce égal à *p* au lieu de α.
On dit alors que la différence observée est **significative** au risque *p*.

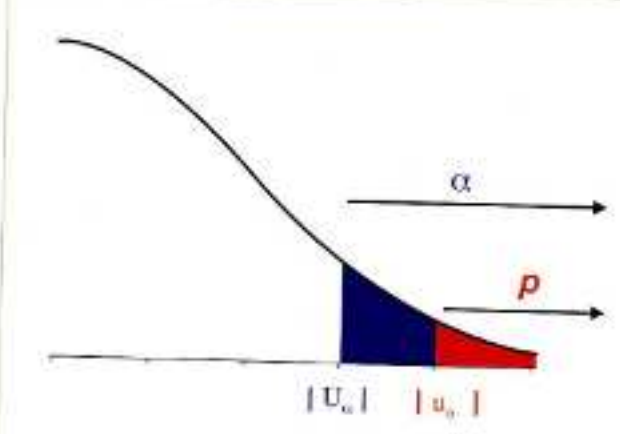


Figure 10-13.

Le risque β est aussi appelé « manque de puissance ». Par opposition, on appelle puissance d'un test la valeur $1 - \beta$.
La puissance d'un test est liée à la taille des effectifs des échantillons. Plus la taille des échantillons comparés augmente, plus la puissance augmente et plus le risque β diminue.
La valeur du risque β n'intervient pas dans l'interprétation d'un test car on ne sait pas la calculer. Il faut cependant toujours tenir compte de ce risque lorsqu'on ne rejette pas H_0. C'est une des raisons pour laquelle on ne peut jamais vérifier qu'une hypothèse nulle est vraie, car si la taille de l'échantillon avait été beaucoup plus grande on aurait peut-être pu la rejeter.
En pratique, on utilise le risque β dans le calcul de la taille des échantillons lorsqu'on bâtit une étude (chap. 11.I.4).

6. Interprétation finale d'un test de comparaison

a) L'hypothèse nulle n'est pas rejetée

Cela signifie que rien ne permet d'affirmer que les paramètres ou les distributions comparés sont différents. C'est d'ailleurs la seule chose que l'on peut dire. On n'affirme jamais qu'une hypothèse nulle est vraie, car :

- elle aurait peut-être pu être rejetée si la puissance du test avait été plus élevée (*cf.* risque β) ;
- il existe de nombreuses raisons qui peuvent expliquer que les distributions ou paramètres mesurés sont identiques bien que les populations soient complètement distinctes.

b) On rejette H_0

Le corollaire de ce rejet est l'acceptation de l'hypothèse H_1.

- Si on avait choisi l'hypothèse alternative H_1 ***bilatérale***, on l'accepte en affirmant que les distributions ou paramètres étudiés sont ***différents***.
- Si on avait choisi l'hypothèse alternative H_1 **unilatérale** on l'accepte en affirmant que l'un des paramètres est **inférieur (ou bien supérieur)** à l'autre. En effet, dans ce cas on s'intéresse au sens de la différence. On ne s'intéresse qu'à une seule extrémité de la distribution de U. Le risque de se tromper est donc deux fois moindre que dans l'hypothèse bilatérale.

c) Le degré de signification *p*

Le risque α est un seuil fixé *a priori* (le plus souvent à 5 %). Lorsque le calcul du test montre que le seuil U_α a été franchi, on rejette H_0 avec un risque égal à α. Mais on désire aller plus loin et préciser la limite du risque pris. On appelle **degré de signification *p***, cette limite *a posteriori*. C'est la probabilité, si H_0 était vraie, d'observer la valeur u_0 avec les valeurs obtenues à partir les échantillons étudiés. En d'autres termes le degré de signification indique la probabilité d'avoir rejeté H_0 si on avait fixé le risque de 1re espèce égal à p au lieu de α.
On dit alors que la différence observée est **significative** au risque p.

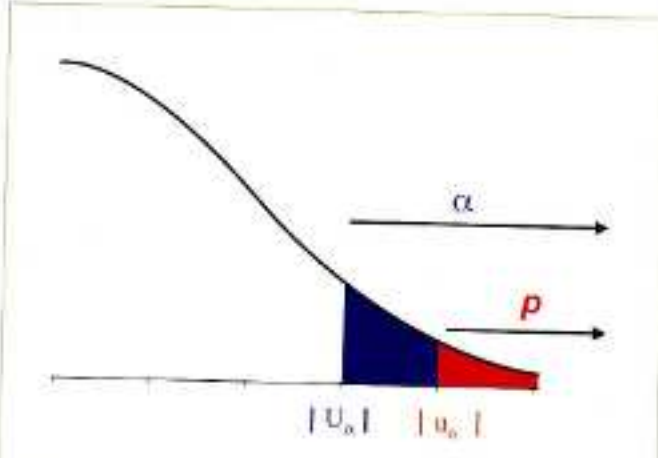


Figure 10-13.

d) Conclusion : la différence est-elle significative ?

Tous les tests comparatifs aboutissent à évaluer une différence en la qualifiant de significative ou non significative.

- Une différence déclarée **non significative** implique qu'on ne peut pas conclure. Soit il n'existe pas de différence entre les paramètres ou les distributions comparées, soit il existe peut-être une différence, mais rien ne permet de l'affirmer avec les données dont on dispose.
- Une différence déclarée **significative** veut dire qu'on affirme qu'il existe une différence entre les paramètres ou les distributions comparées. En déclarant cette affirmation, on prend un certain risque.

Si ce risque n'est pas précisé, cela veut dire implicitement qu'il est inférieur au seuil de 5 %.
Si on déclare une différence significative avec un risque p inférieur à une certaine valeur, on précise par là le risque maximum qu'on prend en faisant cette affirmation. Bien entendu la valeur de p est toujours inférieure à 5 % (exemple 10.2).

Exemple 10.2. INTERPRÉTATION D'UN TEST

Le résultat d'un test bilatéral montre que $u_o = 4{,}5$. La table suivante donne les probabilités α pour que $|U|$ soit supérieure à 2, 3, 4 et 5 :

U	2	3	4	5
α	0,05	0,02	0,01	0,001

La valeur $u_o = 4{,}5$ est supérieure à la valeur seuil $U_{5\,\%} = 2$. On rejette donc H_0. La valeur immédiatement inférieure à u_o est 4 soit $U_{1\,\%}$. On conclut donc en acceptant H_1 avec un degré de signification $p < 0{,}01$.

II. PRINCIPE DES TESTS DE LIAISON

Dans ce type de problème, on se demande s'il existe une liaison entre une ou plusieurs variables étudiées sur un échantillon. Nous nous limiterons à l'étude de la liaison entre 2 variables.
Tester l'existence d'une liaison entre deux variables X et Y, c'est vérifier qu'il existe une relation d'ordre statistique entre ces variables. On dit que deux variables sont liées, lorsque la variation de l'une entraîne une variation de l'autre. En science physique, une telle relation s'exprimerait par une fonction de type $Y = f(X)$. Pour toute valeur de X, on en déduit exactement la valeur de Y. On dit dans ce cas que la relation est déterministe.
Dans les sciences de la vie, la variabilité biologique est responsable de fluctuations dans les mesures des variables. Pour une valeur observée X, on peut observer plusieurs valeurs Y et inversement. En raison de ces fluctuations, il est difficile de visualiser sur un tableau ou un graphique l'existence d'une relation. La démarche consiste à supposer que la liaison étudiée suit un modèle mathématique théorique et l'objet du test sera de vérifier si la relation observée se rapproche suffisamment du modèle théorique.
La démarche consiste donc à bâtir une hypothèse nulle H_0 qui suppose qu'il n'existe pas d'adéquation avec le modèle proposé. Si H_0 est rejetée, on accepte l'hypothèse alternative H_1 qui affirme qu'il existe une relation statistique significative entre les deux variables. Mais il ne s'agira que d'une relation statistique. Il est important de garder à l'esprit, lorsqu'on teste une liaison, que cette relation n'est pas déterministe et qu'elle ne démontre en aucun cas une relation de causalité.

Exercices

Exercice 10.1

Posez H_0 et H_1 dans les situations suivantes :

1) comparaison de deux traitements nouveaux A et B ;
2) comparaison de quatre traitements A, B, C et D ;
3) comparaison d'un traitement A à un placebo (produit inactif) ;
4) variation de la hauteur des arbres en fonction de leur altitude.

Exercice 10.2

Dans une étude comparant les performances psychomotrices de deux groupes de candidats A et B à une qualification professionnelle, les auteurs concluent que les performances du groupe A sont supérieures à celles du groupe B avec un risque d'erreur de moins de 2 %.

Ce chiffre de 2 % correspond :
1) à un risque α
2) à un risque β
3) à un degré de signification p

Exercice 10.3

Vous participez à la mise au point d'un nouveau traitement supposé efficace sur une maladie mortelle, mais dangereux en cas d'utilisation erronée. L'efficacité du produit est testée sur des groupes d'animaux malades et sains.

Vous choisissez un risque α :
1) de 10 %
2) de 5 %
3) de 1 %

Exercice 10.4

Vous participez à la mise au point d'un nouveau vaccin potentiellement efficace dans la prévention d'une maladie grave, et par ailleurs anodin en ce qui concerne les effets secondaires. L'efficacité du vaccin est testée en comparant un échantillon de sujets vaccinés par le nouveau vaccin et un échantillon vacciné par un vaccin placebo (sans effet sur la maladie étudiée).

Vous choisissez prioritairement de ***diminuer*** :
1) le risque α
2) le risque β
3) la puissance
4) la taille des échantillons

Résumé

On distingue deux grands types de tests statistiques : les tests de comparaison qui comparent des distributions ou des paramètres (moyennes, pourcentages, *etc.*) et les tests de liaison qui étudient la variation d'une variable par rapport à une autre. Les tests statistiques s'appuient sur des valeurs recueillies sur des échantillons. Ils visent à étudier si une différence ou une liaison entre variables existe dans les populations d'où sont issus ces échantillons.
Leur principe général consiste à poser une hypothèse et à examiner si cette hypothèse est vérifiée.
L'hypothèse dont on part est toujours l'hypothèse nulle : elle consiste à supposer que les paramètres ou distributions comparées sont identiques, ou bien qu'il n'existe pas de liaison entre deux variables.
Le calcul du test aboutit :

- soit à ne pas rejeter cette hypothèse nulle H_0, et dans cette situation on ne peut conclure ;
- soit à la rejeter. On accepte alors une hypothèse dite alternative H_1 et on affirme que les paramètres ou distributions comparés sont différents ou qu'il existe une liaison entre les deux variables étudiées.

TYPE DE TEST

HYPOTHÈSE		COMPARAISON DE PARAMÈTRES OU DE DISTRIBUTIONS	LIAISON ENTRE 2 VARIABLES
nulle	H_0	Les paramètres ou les distributions sont identiques	Absence de liaison
alternative	H_1		
• bilatérale		Les paramètres ou les distributions sont différents	Présence d'une liaison
• unilatérale		Un paramètre est supérieur à l'autre	

La conclusion d'un test suppose la prise de certains risques :

- rejeter à tort H_0 alors que H_0 est vraie, c'est le risque α ;
- ne pas rejeter H_0 alors que H_1 est vraie, c'est le risque β.

Le risque α peut se contrôler lorsqu'on effectue un test. On fixe son seuil en général à 5 %.
Le risque β ne peut pas se mesurer. Cependant, il est lié à la taille des échantillons et on peut donc le maîtriser en construisant l'étude.

La conclusion d'un test aboutit à affirmer :

- soit qu'une différence ou une liaison n'est pas significative. On ne peut conclure ;
- soit qu'une différence ou une liaison est significative. On peut se contenter de cette affirmation et le risque d'erreur est inférieur à 5 %. On peut aussi préciser le risque maximum que l'on prend. Sa valeur est appelée degré de signification p ($p < 0{,}00\ldots$).

TESTS DE COMPARAISON

Les principales lois de distribution théoriques utilisées pour les tests statistiques de comparaison courants sont :

- la loi Z normale centrée réduite ;
- la loi T de Student ;
- la loi F de Fisher ;
- la loi du χ^2.

Ce chapitre n'aborde que le principe général des tests de comparaison. Le chapitre 13 détaille le processus de calcul de chaque test.

I. TEST Z OU TEST DE L'ÉCART RÉDUIT

Le test Z sert à comparer des paramètres en testant leur différence. On utilise ce test pour comparer :
- la moyenne d'un échantillon à une moyenne théorique (chap. 13.III) ;
- deux moyennes (chap. 13.IV) ;
- deux moyennes de deux séries « appariées » (chap. 13.V) ;
- les rangs de deux distributions (*cf.* test de Wilcoxon chap. 11.VI, 13.XI et 13.XII).

1. Principe du test Z

Prenons l'exemple de deux paramètres mesurés sur deux échantillons que l'on désire comparer.
- Sous H_0 les paramètres des populations d'où sont issus les 2 échantillons sont identiques.
- Sous H_1 bilatérale, les paramètres sont différents.
- Sous H_1 unilatérale, un des paramètres est supérieur à l'autre.

On compare les 2 paramètres observés par leur différence arithmétique Δ. Cette différence est une variable aléatoire et si H_0 est vraie elle est proche de zéro. Si les échantillons sont de taille suffisante et si H_0 est vraie, la division de cette différence par son écart type (s_d) suit une loi Z normale centrée réduite, de moyenne zéro et d'écart type 1 (*cf.* chap. 6.III.3).
Le test consiste à calculer l'écart réduit $z_o = |\Delta| / s_d$, puis à comparer cette valeur à la distribution théorique de la loi Z.

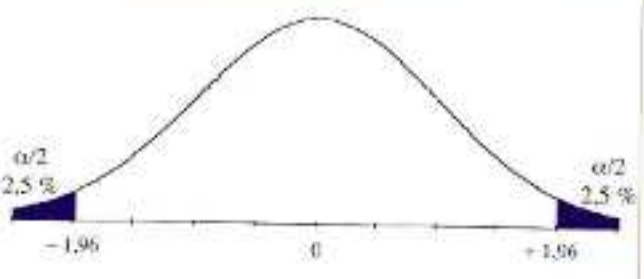


Figure 11-1. Variable normale centrée réduite Z

On utilise pour cela la table de la loi Z (*cf.* utilisation de la table Z, chap. 11.I.3). Si H_0 est vraie, la valeur absolue $|z_0|$ n'a que 5 chances sur 100 d'être supérieure à 1,96.

Condition d'application : les effectifs de chaque échantillon doivent être supérieurs ou égaux à 30. Le calcul des différents types de test Z sont détaillés chapitre 13.III à 13.V.

2. Interprétation du test Z avec un risque α fixé à 5 %

Hypothèse H_1 bilatérale

- Lorsque la valeur observée z_0 est inférieure à 1,96, on formule qu'**on ne rejette pas l'hypothèse nulle**. On ne peut pas affirmer que les échantillons proviennent de populations différentes. On dit que la différence entre les paramètres n'est pas significative.
- Lorsque la valeur observée z_0 est supérieure à 1,96, on formule le **rejet** de H_0. On accepte H_1 en affirmant que les échantillons proviennent de populations différentes. On affirme que la différence entre les paramètres est significative. On recherche le degré de signification p (*cf.* utilisation de la table Z au chap. 11.I.3).

Hypothèse H_1 unilatérale

Dans ce cas on s'intéresse au sens de la différence. On postule que l'un des paramètres est supérieur ou inférieur à l'autre. On ne considère qu'une seule extrémité de la distribution de Z (figure 11.2). Le risque d'erreur est donc deux fois moindre que dans l'hypothèse bilatérale. La valeur seuil pour un risque de 5 % est donné par la valeur $z_{10\,\%} = 1{,}645$.

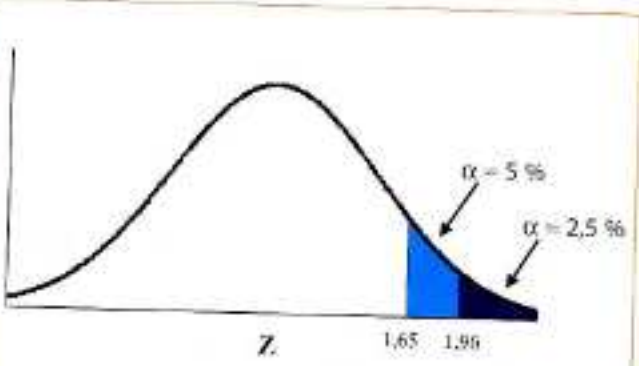


Figure 11-2. Hypothèse H_1 unilatérale

- Lorsque la valeur observée z_0 est inférieure à 1,645, on formule qu'**on ne rejette pas l'hypothèse nulle**. La différence entre les paramètres n'est pas significative.
- Lorsque la valeur observée z_0 est supérieure à 1,645, on formule le **rejet** de H_0. On accepte H_1 en affirmant non seulement que la différence entre les paramètres est significative, mais en outre que l'un des paramètres est **inférieur ou supérieur** à l'autre. On recherche le degré de signification p (*cf.* utilisation de la table Z, chap. 11.I.3).

3. Utilisation pratique de la table Z

La table de Z figure en Annexes, chapitre 27, Table 1. Pour mesurer le risque de signification p, lorsque z_0 est supérieur à 1,96, on cherche dans la colonne Z la valeur immédiatement inférieure à z_0.
Si H_1 est bilatérale, la valeur lue dans la colonne α donne le degré de signification p.
Si H_1 est unilatérale, on divise la valeur α par 2 (exemples 11.1 et 11.2).

4. Calcul du nombre de sujets nécessaires à un test Z

Il arrive qu'après avoir réalisé des tests sur des données recueillies antérieurement aux analyses, l'auteur s'aperçoive avec dépit qu'il n'obtient aucune différence significative. Jugeant qu'il s'agit d'un

Exemple 11.1. INTERPRÉTATION DU RÉSULTAT D'UN TEST Z

On admet qu'un test Z a donné la valeur $z_o = 2,4$.
Si l'hypothèse H_1 est bilatérale, on rejette H_0 puisque z_o est supérieure à 1,96. On constate dans la colonne Z que la valeur immédiatement inférieure à 2,4 est 2,326. Cette valeur correspond à un degré de signification p égal à 0,02. On rejette donc H_0 avec un degré de signification $p < 0,02$.
Si H_1 est unilatérale, on rejette H_0 puisque z_o est supérieure à 1,645. Le degré de signification est divisé par deux soit $p < 0,01$.

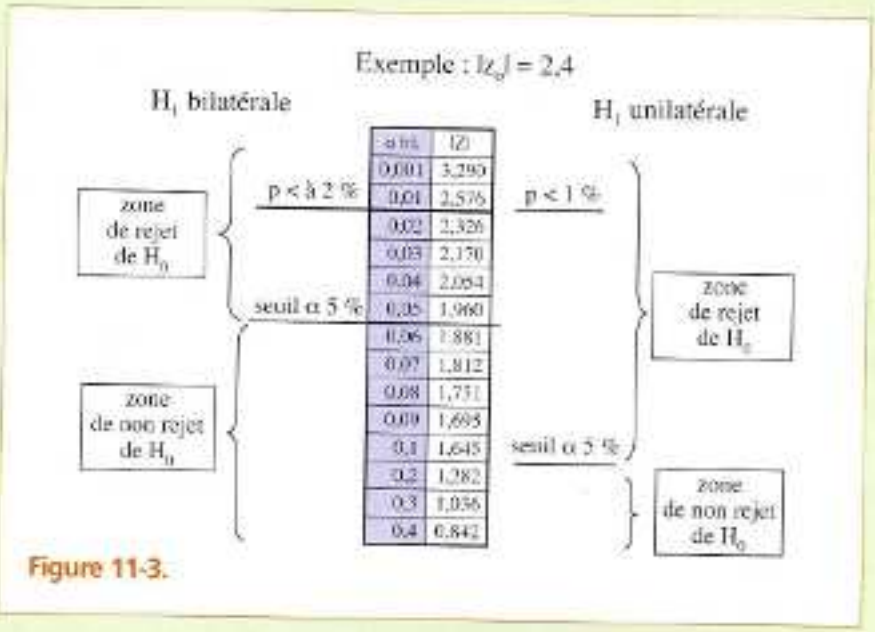


Figure 11-3.

Exemple 11.2. INTERPRÉTATION DU RÉSULTAT D'UN TEST Z

On veut comparer la fréquence du paludisme dans deux régions d'Afrique. La fréquence de la maladie a été mesurée dans deux échantillons d'individus tirés au sort dans chacune des deux régions. Le calcul du test a montré une valeur z_o égale à 2,6.
La valeur z_o est supérieure à 1,96 (risque α **bilatéral** 5 %). On rejette donc l'hypothèse nulle d'égalité des fréquences du paludisme dans les deux régions. La valeur de Z immédiatement inférieure à 2,6 correspond à une valeur α de 0,01.
On affirme donc que la fréquence du paludisme est significativement différente entre les deux régions avec un degré de signification $p < 1$ %.

On désire étudier la supériorité d'un nouveau médicament visant à réduire l'hypertension artérielle. Le nouveau produit A est administré à un groupe de malades hypertendus. L'effet sur la pression artérielle est mesuré dans ce groupe et comparée à un groupe témoin de sujets hypertendus traités par un produit classique B.
On pose H_0 : effet de A similaire à l'effet de B ;
H_1 *unilatérale* : A est supérieur à B (réduction de l'hypertension).

Le résultat du test montre une valeur $|z_o|$ égale à 1,82. Cette valeur étant supérieure à 1,645 (risque α **unilatéral** 5 %) on décide de rejeter H_0.
Pour une valeur de Z égale à 1,82, on trouve une valeur α proche de 0,07. Comme H_1 est unilatérale, la valeur de p est divisée par 2 : $p = 0,07/2 = 0,035$.
On déclare que A possède un effet significativement supérieur à B avec un degré de signification $p < 4$ %.
On voit ici l'intérêt de poser dès le début une hypothèse unilatérale lorsque la nature du problème s'y prête. Si l'hypothèse avait été bilatérale, le résultat du test n'aurait pas été concluant.

manque de puissance dû à la faiblesse des effectifs, il est alors tenté d'augmenter la taille des échantillons. Le rajout *a posteriori* de nouveaux sujets dans une étude est très risqué : biais de sélection, absence de représentativité des populations étudiées, sujets moins bien définis que les sujets initialement prévus, *etc.*

Une attitude rigoureuse consiste à définir dès le début de l'enquête la différence minimale qu'on souhaite observer pour rejeter H_0 et de calculer le nombre de sujets nécessaires.

Il existe des formules qui permettent de calculer le nombre minimum de sujets nécessaires à un test Z (*cf.* Annexes, Formulaire 12). Ces formules dépendent du type de comparaison.

Quelle que soit la formule, il faudra dans tous les cas :

- préciser l'hypothèse H_1 bilatérale ou unilatérale (le nombre de sujets exigé est moins élevé lorsque H_1 est unilatérale) ;
- choisir un seuil de risque α : en général 5 % ;
- choisir un risque β : en général 20 % ;
- se fixer une différence minimale Δ entre les paramètres à comparer. De façon générale, plus la différence escomptée Δ est faible, plus la taille de l'échantillon sera élevée.
- estimer la variance de cette différence. Ceci est la partie la plus délicate car on ne la connaît pas. Il faut alors utiliser les résultats d'études antérieures ou ceux d'une étude pilote (exemples 11.3 et 11.4).

Exemple 11.3. CALCUL DE TAILLE D'ÉCHANTILLON POUR UN TEST Z

On veut comparer le poids d'un groupe de nouveau-nés supposés hypotrophiques par rapport au poids des nouveau-nés de la population générale. Combien faut-il de sujets pour espérer observer une différence moyenne de 0,3 kg avec un risque α de 5 % et une puissance de 80 % ? On estime que l'écart type dans la population générale des nouveau-nés est de 0,5 kg.

Pour une puissance de 80 %, β = 20 %. La problématique posée ici est celle d'un test *unilatéral* puisqu'on s'intéresse à des nouveau-nés qui auraient un poids plus faible que la moyenne générale. On prend donc la valeur $z_{2\alpha} = 1{,}645$.

On a $n \geq 0{,}5^2\,(1{,}645 + 0{,}842)^2/0{,}3^2$ soit en pratique $n \geq 18$.

(*cf.* formules de calcul de taille d'échantillon, Annexes, Formulaire 12)

Exemple 11.4. CALCUL DE TAILLE D'ÉCHANTILLON POUR UN TEST Z

On veut comparer le poids de 2 groupes de nouveau-nés d'effectifs égaux. Combien faut-il de sujets dans chaque groupe pour espérer observer une différence moyenne de 0,3 kg avec un risque α de 5 % et une puissance de 80 % ? On estime que l'écart type dans la population générale des nouveau-nés est de 0,5 kg.

Pour une puissance de 80 %, β = 20 %. La problématique posée ici est celle d'un test *bilatéral*. On prend donc la valeur $z_\alpha = 1{,}96$.

On a $n_1 \geq 2 \times 0{,}5^2 \times (1{,}96 + 0{,}842)^2/0{,}3^2$ soit en pratique $n \geq 44$ dans chaque groupe.

(*cf.* formule de calcul de taille d'échantillon, Annexes, Formulaire 12)

II. TEST T DE STUDENT

Lorsque la taille des échantillons est faible ($n < 30$), le rapport entre les différences de leurs moyennes et l'écart type ne suit pas une loi normale centrée réduite Z. On ne peut donc pas utiliser ce test. On démontre que dans ces conditions le rapport suit une loi T de Student. La loi de Student ressemble à une loi normale, mais elle est plus étalée. Ce qui signifie que l'espace comprenant 95 % des valeurs de T est plus large. Pour qu'un test T soit significatif, la valeur de T est donc plus élevée que la valeur de Z. Au-delà d'un effectif de 30, la loi de Student tend à se superposer à la loi normale centrée réduite.

Le test T sert à comparer :
- la moyenne d'un petit échantillon à une moyenne théorique (chap. 13.VI) ;
- les moyennes de 2 petits échantillons (chap. 13.VII) ;
- les moyennes de 2 petites séries « appariées » (chap. 13.VIII) ;
- tester un coefficient de corrélation (chap. 13.XXI, 13.XXII).

1. Principe du test T

Il est en tout point identique à celui du test Z. Le test consiste à estimer l'écart type s_d de la différence Δ, à calculer la valeur $t_o = |\Delta|/s_d$, puis à comparer cette valeur à la distribution théorique de la loi T de Student. Si H_0 est vraie t_o n'a que 5 chances sur 100 d'être supérieure à $T_{5\%}$. On utilise pour cela la table de la loi T (Annexes, Tables statistiques 2).

2. Interprétation du test T

Elle dépend de l'application du test. En ce qui concerne la comparaison de moyennes, l'interprétation du test T est identique à celle du test Z.

Hypothèse H_1 bilatérale

- Lorsque la valeur observée t_o est inférieure à $T_{5\%}$, la différence entre les paramètres n'est pas significative.
- Lorsque la valeur observée t_o est supérieure à $T_{5\%}$, la différence entre les paramètres est significative. On recherche le degré de signification p (*cf.* chap. 11.II.3 : utilisation de la table T).

Hypothèse H_1 unilatérale

La valeur seuil pour un risque de 5 % est donnée par la valeur $T_{10\%}$.
- Lorsque la valeur observée t_o est inférieure à $T_{10\%}$, la différence entre les paramètres n'est pas significative.
- Lorsque la valeur observée t_o est supérieure à $T_{10\%}$, la différence entre les paramètres est significative, mais en outre on affirme que l'un des paramètres est **inférieur (ou supérieur)** à l'autre. On recherche le degré de signification p (*cf.* chap. 11.II.3 : utilisation de la table T).

Conditions d'application : le test T a l'avantage de pouvoir étudier des échantillons de petite taille. En revanche, il exige que les variances des populations soient identiques. Il importe donc soit de faire l'hypothèse que ces variances sont identiques, soit de le vérifier par un test approprié (*cf.* test F, chap. 11.III).

Le calcul des différents tests T sont détaillés chapitre 13.VI à VIII.

3. Utilisation de la table T

La table T est plus difficile à utiliser que la loi Z, car il y a autant de distribution T de Student que de **degré de liberté (ddl)**. Dans la situation présente, on appelle degré de liberté la taille de chaque échantillon diminuée de la valeur 1.

- Pour un échantillon de taille n, ddl = n – 1.
- Pour deux échantillons de taille n_1 et n_2, ddl = $(n_1 - 1) + (n_2 - 1)$.

La table T comporte sur chaque ligne, les valeurs possibles des **ddl** et en colonne les valeurs de α. On commence donc par repérer dans la table T (Table 2), à la ligne correspondante au nombre de **ddl**, la valeur $T_{5\%}$. Si le calcul du test montre une valeur observée t_o supérieure à $T_{5\%}$, on cherche ensuite, dans la même ligne, la valeur de T immédiatement inférieure à t_o. La valeur correspondante lue dans la ligne α donne le degré de signification p.
Si l'hypothèse H_1 est unilatérale, on rejette H_0 pour une valeur de t_o supérieure à $T_{10\%}$. Le degré de signification p, obtenu comme précédemment, est divisé par 2 (exemple 11.5).

Exemple 11.5. INTERPRÉTATION DU RÉSULTAT D'UN TEST T

Un test T bilatéral de comparaison de moyennes a été effectué sur deux échantillons comportant chacun 3 sujets. Le calcul du test aboutit à une valeur $t_o = 3,9$.
Le nombre de ddl est 3 + 3 – 2 = 4.
Pour ddl = 4, $T_{5\%} = 2,776$. On observe que t_o est supérieur à $T_{5\%}$. On rejette donc H_0. La valeur de T immédiatement inférieure à 3,9 lue dans la table est 3,747. Elle correspond à un risque de 0,02. On conclut donc à une différence significative entre les paramètres étudiés avec $p < 0,02$.
Si le test avait été unilatéral, on aurait rejeté aussi H_0 puisque 3,9 est supérieur à $T_{10\%} = 2,132$. La valeur de p aurait été de 0,02/2 soit $p < 0,01$.

α	0,0001	0,001	0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5	0,9
$T_{ddl=4}$	15,534	8,610	4,604	3,747	3,298	2,999	2,776	2,132	1,533	1,190	0,741	0,134

zone de rejet H_0 — zone de non-rejet de H_0

III. TEST F DE FISHER-SNEDECOR

Le test F sert à comparer deux variances par leur rapport. On l'utilise principalement dans deux situations :

- lorsque l'on veut vérifier la condition d'application d'égalité des variances dans un test T de comparaison de moyennes ;
- lorsqu'on réalise une analyse de variance pour comparer plusieurs moyennes (chap. 11.III.3).

1. Principe du test F de comparaison de deux variances

Sous l'hypothèse nulle H_0, les 2 variances sont égales et leur rapport est égal à 1. Lorsque le rapport est significativement différent de 1, on rejette H_0 et on accepte l'hypothèse alternative H_1 qui suppose que les 2 variances sont différentes (sans préciser le sens de cette différence) : H_1 est bilatérale.
On démontre que sous H_0 le rapport de 2 variances suit une loi, dite loi F de Fisher-Snedecor. Le principe du test F de comparaison de deux variances consiste donc à calculer le rapport des variances de

deux échantillons $F_o = s_1^2/s_2^2$ (en plaçant la plus élevée au numérateur) puis à comparer cette valeur à la distribution théorique de la loi F (Annexes, Tables statistiques 3).

Condition d'application : les distributions doivent être supposées normales dans les deux populations d'où proviennent les deux échantillons.
Le calcul du test est détaillé au chapitre 13.IX.

2. Interprétation d'un test F de comparaison de 2 variances

Attention : les tables de la loi F ont pour particularité d'être construite pour des tests *unilatéraux*. Or, lorsqu'on compare simplement les variances de deux échantillons, l'hypothèse alternative est bilatérale. La valeur seuil pour un risque de 5 % est donc donnée par la valeur $F_{2,5\%}$.

- Lorsque la valeur observée F_o est inférieure à $F_{2,5\%}$, la différence entre les variances n'est pas significative. Si le test servait à vérifier la condition d'égalité de variances d'un test T, en toute rigueur, le non-rejet de H_0 ne permet pas d'affirmer une égalité des variances mais seulement de dire qu'il n'y a pas d'argument permettant d'affirmer que les variances sont différentes.
- Lorsque la valeur observée F_o est supérieure à $F_{2,5\%}$, la différence entre les variances est significative. On recherche le degré de signification *p* (*cf.* chap. 11.III.4 : utilisation de la table F, H_1 bilatérale).

3. Analyse de la variance pour comparer plusieurs moyennes

Appellations équivalentes : analyse de la variance, ANOVA.
Pour comparer plus de deux moyennes entre elles, on ne peut ni tester leurs différences, ni leur rapport. Le test ANOVA permet de comparer les moyennes de plusieurs échantillons.

Principe du test ANOVA

On considère les échantillons comme des groupes d'individus.
Le principe du test ANOVA consiste à scinder la variation totale de l'ensemble des observations en deux termes :

- la variation entre les groupes. Cette variation est mesurée par l'écart moyen entre chaque moyenne et la moyenne générale. On l'appelle **variance entre groupes** (s_g^2) ;
- la variation moyenne des individus à l'intérieur des groupes. Elle est mesurée par la moyenne pondérée des variances de chaque groupe. Cette variance est appelée **variance résiduelle** (s_r^2) (exemple 11.6).

Exemple 11.6.

Plaquons 3 accords de 3 notes sur une guitare (figure 11.4) : la variation totale entendue est représentée par l'écart entre la note la plus basse et la note plus aiguë de l'ensemble des notes jouées. La variation entre accords (entre groupes) est représentée par l'écart entre les valeurs moyennes de l'accord le plus bas et de l'accord le plus haut. La variation résiduelle est représentée par l'écart moyen entre la note la plus basse et la note la plus aiguë de chaque accord.

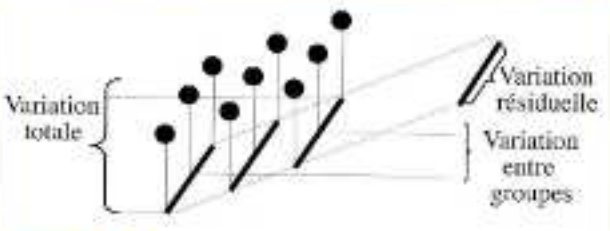


Figure 11-4.

Sous l'hypothèse nulle H_0, les moyennes des échantillons sont identiques. La variation entre les groupes est proche de zéro. La variation totale observée est donc due principalement à la variation des individus indépendamment du groupe auquel ils appartiennent. La variance entre groupes est inférieure à la variance résiduelle.

Sous H_1, les moyennes des échantillons sont différentes : la variation totale est due principalement à la variation des groupes entre eux (figure 11.5). La variance entre groupes est supérieure à la variance résiduelle. Il s'agit donc d'une hypothèse H_1 *unilatérale*.
Le test consiste à comparer la variation entre les groupes et la variation résiduelle en calculant le rapport des 2 variances correspondantes $F_o = s_g^2/s_r^2$.
On se retrouve donc dans la situation d'une comparaison de deux variances par leur rapport. Nous avons vu au paragraphe précédent qu'on utilisait le test F.

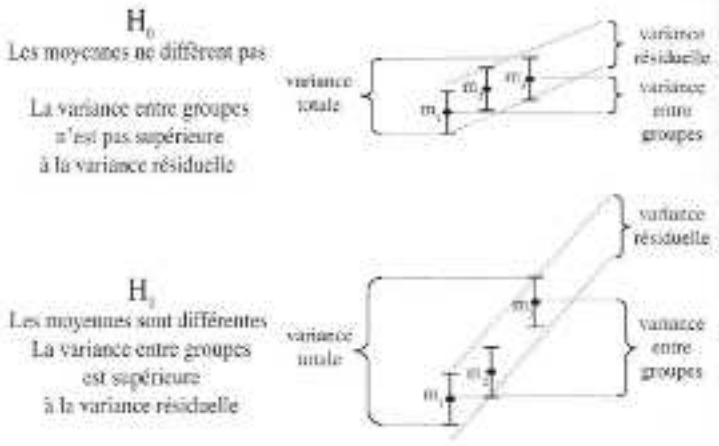


Figure 11-5. Comparaison de plusieurs moyennes par décomposition de la variance

- Lorsque la valeur observée F_o est inférieure à $F_{5\%}$, la différence entre les variances n'est pas significative. Cela signifie que les moyennes ne diffèrent pas significativement.
- Si la valeur observée F_o est supérieure à $F_{5\%}$ on rejette H_0 et on accepte l'hypothèse H_1. Cela signifie que les moyennes des groupes étudiés diffèrent entre elles de façon significative. On recherche le degré de signification p (*cf.* chap. 11.III.4 : utilisation de la table F, H_1 unilatérale).

Lorsque H_0 a été rejetée, et seulement dans ce cas, on peut réaliser des comparaisons de moyennes 2 à 2 par des tests T en utilisant la variance résiduelle s_r^2 comme variance commune.

Condition d'application : les distributions des populations d'où proviennent les échantillons doivent être normales et de même variance.
Le calcul du test est détaillé au chapitre 13.X.

4. Utilisation des tables de F

Il y a autant de distribution F que de **degré de liberté (ddl)** pour chacun des deux échantillons. On appelle ici degré de liberté k_1 et k_2 la taille de chaque échantillon diminuée de la valeur 1.

On a $k_1 = n_1 - 1$ et $k_2 = n_2 - 1$

Les tables de la loi F sont donc à 3 dimensions, avec une table pour chaque valeur possible de α. Chacune des tables comporte en colonne les valeurs possibles de k_1 et en ligne les valeurs possibles de k_2. En Annexes, les Tables statistiques 3 donnent les valeurs de F correspondant aux risques α de 5 %, 2,5 %, 1 % et 1 pour mille.

Attention : les tables de F sont données pour une hypothèse H_1 ***unilatérale****, car en pratique le plus souvent (et notamment dans le test l'analyse de la variance, chap. 11.III.3) on pose l'hypothèse qu'une des 2 variances est supérieure à l'autre.*

- Si l'hypothèse H1 est *unilatérale* (ANOVA), le seuil de rejet de H_0 est donné pour une valeur F au risque $\alpha = 5\ \%$. On commence donc par repérer dans la table de $F_{5\ \%}$, à la ligne et à la colonne correspondante au nombre de ddl, la valeur $F_{5\ \%}$. Par exemple, pour $k_1 = 4$ et $k_2 = 8$, la valeur seuil de F au risque $\alpha = 5\ \%$ est de 3,84. Si le calcul du test montre une valeur observée F_o supérieure à $F_{5\ \%}$, on rejette H_0 ; on cherche ensuite, dans les tables suivantes aux mêmes lignes et colonnes, la valeur de F immédiatement inférieure à F_o. Le risque α correspondant à cette valeur donne le degré de signification *p*.
- Si l'hypothèse H1 est *bilatérale* (comparaison de 2 variances), le seuil de rejet de H_0 au risque 5 % est donné pour une valeur $F_{2,5\ \%}$. Si le calcul du test montre une valeur observée F_o supérieure à $F_{2,5\ \%}$, on rejette H_0 ; on cherche ensuite, dans les tables suivantes aux mêmes lignes et colonnes, la valeur de F immédiatement inférieure à F_0. Pour obtenir p, on *multiplie* par 2 la valeur de α correspondante à cette valeur de F.

Exemple 11.7. COMPARAISON DE 2 VARIANCES PAR UN TEST F

La comparaison des variances de 2 échantillons de 10 et 7 individus par un test F *bilatéral* a montré un résultat $F_o = 8$. On a donc $k_1 = 9$ et $k_2 = 6$. On note pour ces 2 ddl dans la table que $F_{2,5\ \%} = 5{,}52$. On rejette donc H_0. On note que 8 est encore supérieur à $F_{1\ \%}$. Les deux variances diffèrent donc significativement avec $p < 0{,}02$ (0,01 × 2).

$\alpha = 0{,}025$

k2 \ k1	5	6	7	8	9
5	7,15	6,98	6,85	6,76	6,68
6	5,99	5,82	5,70	5,60	5,52

$\alpha = 0{,}01$

k2 \ k1	5	6	7	8	9
5	10,95	10,67	10,46	10,29	10,16
6	8,75	8,47	8,26	8,10	7,98

IV. TESTS DE χ^2

Formulations équivalentes : tests de chi-deux, tests de chi-carré, test de χ^2 de Pearson.
Les tests de χ^2 servent à comparer des distributions.
Les tests de χ^2 peuvent être appliqués sur tous types de variables : qualitative nominale, ordinale, qualitative binaire, quantitative discrète ou continue discrétisée.

Selon la situation on distingue :

- test de χ^2 de **conformité (ou d'ajustement)**. Il sert à comparer une distribution observée sur un échantillon à une distribution connue dans une population ou à une distribution théorique : binomiale, Poisson, normale, *etc.* ;
- le test de χ^2 d'**homogénéité**. Il sert à comparer **deux** ou **plusieurs** distributions observées sur des échantillons ;

- le test de χ^2 d'**indépendance**. Il sert à étudier sur un même échantillon la liaison entre les distributions de 2 variables (nominales ou binaires). Ce n'est donc pas à proprement parler un test de comparaison et il sera étudié au chapitre 12.I.

Quelle que soit la situation, le principe et le calcul du test sont identiques.

1. Principe du χ^2

Dans tous les cas, le principe consiste à comparer les **effectifs** des classes des distributions.

a) Hypothèses

- Test de conformité ou d'ajustement (figure 11.6) : sous H_0, l'échantillon observé provient de la population dont on connaît la distribution théorique. La distribution observée devrait lui être identique. Si on observe une différence, H_0 est rejetée et on accepte H_1 : la distribution observée est différente de la distribution théorique, l'échantillon étudié n'appartient pas à cette population.
- Test d'homogénéité (figure 11.7) : sous H_0, les échantillons étudiés proviennent de la même population. Leurs distributions devraient être identiques entre elles et identiques à la distribution observée sur le total des échantillons. Si elles sont différentes, H_0 est rejetée et on accepte l'hypothèse H_1, d'hétérogénéité : les distributions sont différentes entre elles, les échantillons proviennent donc de populations différentes.

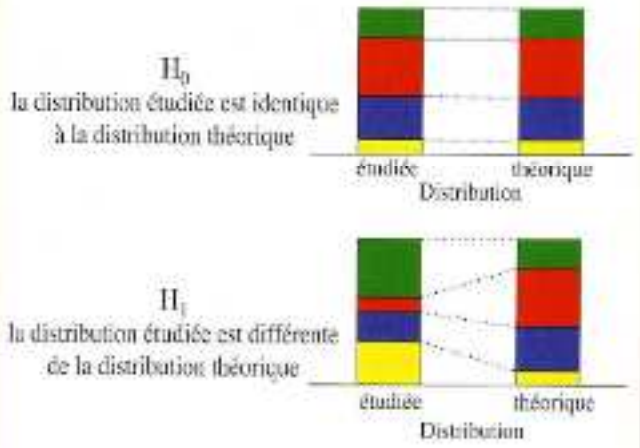


Figure 11-6. Test de χ^2 de conformité

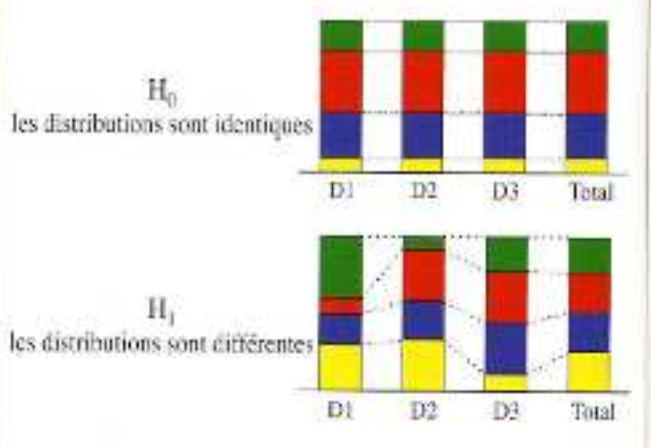


Figure 11-7. Test de χ^2 d'homogénéité

Le test de χ^2 est un test très général qui sert à comparer toutes sortes de distribution d'effectifs. Ainsi, on peut comparer la distribution :

- d'une variable qualitative à plusieurs classes :
  - à une distribution théorique,
  - entre deux échantillons,
  - entre plusieurs échantillons ;
- d'une variable qualitative binaire à 2 classes :
  - à une distribution théorique. Cela revient à comparer un pourcentage observé à un pourcentage théorique,
  - entre deux échantillons. Cela revient à comparer deux pourcentages.

– entre deux échantillons dont chaque sujet de l'un est « apparié » à un sujet de l'autre échantillon (*cf.* chap. 11.IV.1.d),
– entre plusieurs échantillons. Cela revient à comparer plusieurs pourcentages.

b) Tableau de contingence

Le test de χ^2 s'applique à des effectifs regroupés sur un tableau qu'on appelle ***tableau de contingence***. Un tableau de contingence est un tableau comportant des effectifs dans ses cases, et, les totaux de chaque ligne et de chaque colonne dans ses marges.

Le tableau de contingence a une forme qui dépend de la nature de la comparaison (figure 11.8).

La forme la plus générale est celui de la comparaison de la distribution d'une variable qualitative à plusieurs classes entre plusieurs échantillons (*cf.* exemple 11.8).

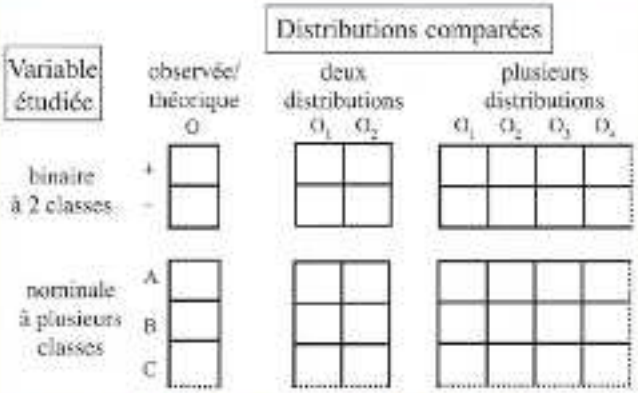


Figure 11-8. Types de tableaux de contingences

Exemple 11.8. TABLEAU DE CONTINGENCE D'EFFECTIFS OBSERVÉS

Distribution des patients selon l'évolution de leur maladie parmi 3 échantillons de malades ayant subi un traitement différent.

		TRAITEMENTS E_1	E_2	E_3	TOTAL N
Évolution de la maladie	guérison	5	6	16	**27**
	rechute	9	9	10	**28**
	décès	15	4	7	**26**
	Total	**29**	**19**	**33**	**81**

Ce tableau est un tableau de contingence 3 × 3 d'effectifs observés.

c) Calcul du test de χ^2 : cas général

Soit le tableau d'effectifs observés (Tableau A). Dans chaque case du tableau A, figurent les effectifs observés (o_{ij}) pour chaque classe de la variable et chaque distribution.

Tableau A : Effectifs observés

Classes de la variable A	Échantillons E_1	E_2	…	E_j	Total
A_1	o_{11}				t_1
A_2					t_2
…					…
A_i				o_{ij}	t_i
Total	n_1	n_2	…	n_j	N

Tableau B : Effectifs théoriques

	Échantillons E_1	E_2	…	E_j	Total
A_1	c_{11}				t_1
A_2					t_2
…					…
A_i				c_{ij}	t_i
Total	n_1	n_2	…	n_j	N

Le principe du test consiste d'abord à calculer pour chaque case du tableau l'***effectif théorique*** qui serait observé si les distributions étaient identiques.
Dans ce cas, les effectifs de chaque distribution devraient être répartis en proportion égale. Par exemple, l'effectif de la case A_1E_1 devrait être égal à n_1 fois t_1/N.
Dans chacune des cases, l'effectif théorique est égal au produit du total de la ligne par le total de la colonne et divisé par le total général : $c_{ij} = n_i t_j / N$.
On dispose ainsi d'un second tableau (Tableau B) composé d'effectifs théoriques dont les totaux marginaux (lignes et colonnes) sont identiques au tableau des effectifs observés.
Sous l'hypothèse nulle, la somme des différences entre effectifs observés et théoriques correspondants devrait être proche de zéro. Le principe général du test consiste donc à examiner si l'ensemble de toutes ces différences n'excède pas une certaine valeur.
En fait, on teste une expression un peu plus compliquée (on élève la différence au carré et on divise par l'effectif théorique), mais le principe reste le même.
On démontre que, sous l'hypothèse nulle, la somme de toutes ces expressions suit une loi dite loi du χ^2.

$$\chi^2_o = \Sigma \frac{(o_{ij} - c_{ij})^2}{c_{ij}}$$

Si les différences $o_{ij} - c_{ij}$ sont très faibles, la valeur de χ^2_o n'excède pas une certaine valeur seuil $\chi^2_{5\%}$.
Le test consiste à comparer la valeur trouvée χ^2_o à la distribution théorique de la loi de χ^2.
On utilise pour cela la table de la loi du χ^2 (*cf.* chap. 11.IV.3). Si H_0 est vraie, la valeur χ^2_o n'a que 5 chances sur 100 d'être supérieure à $\chi^2_{5\%}$. Si la somme des différences dépasse la valeur seuil $\chi^2_{5\%}$, on rejette alors H_0.

Condition d'application : les effectifs théoriques calculés dans chaque case du tableau doivent être supérieurs ou égaux à 5. Sinon, il faut regrouper les effectifs de certaines classes.
Les calculs des différents types de test χ^2 de comparaison sont détaillés au chapitre 13.XIV à XVII.

d) Cas particuliers

■ Comparaison de deux pourcentages

Lorsqu'on désire comparer une variable binaire entre deux échantillons, le test revient à comparer deux pourcentages. Le tableau de contingence comporte seulement 4 cases (*cf.* formule de calcul simplifiée, chap. 13.XVI).
Les conditions d'application sont les mêmes que pour le χ^2 général (effectifs théoriques supérieurs ou égaux à 5). Lorsque les conditions d'application ne sont pas remplies (au moins un effectif théorique inférieur à 5), on peut utiliser le test exact de Fisher (chap. 11.V) qui est valide quelle que soit la taille des effectifs.
Dans de nombreux ouvrages, figure le test du χ^2 corrigé de Yates (*cf.* Annexes, Formulaire 21) applicable lorsque l'un au moins un des effectifs théoriques est inférieur à 5 et supérieur ou égal à 3. Cette correction, qui n'est pas admise par tous, n'apporte qu'une valeur approximative. Elle est inutile lorsqu'on peut utiliser le test exact de Fisher.
Le nombre de sujets nécessaires pour comparer deux pourcentages peut être calculé au moyen d'une formule assez complexe (Annexes, Formulaire 23, ou plus rapidement avec EpiInfo/Epitable/ Echantillon/Taille échantillon/Deux proportions).

■ **Séries appariées**

On dit que deux séries sont appariées lorsque chaque sujet de l'une est en relation avec un sujet de l'autre. Dans deux séries appariées sur l'âge, chaque sujet de l'une a été sélectionné avec un « jumeau » de même âge dans l'autre série. L'ensemble des deux séries est donc composé d'un ensemble de paires. Lorsqu'on désire comparer la présence ou l'absence d'une caractéristique (variable binaire) entre ces deux séries, on obtient quatre sortes de paires de sujets :

- les paires concordantes lorsque les deux sujets de la paire possèdent la caractéristique (++), ou bien, lorsque les deux sujets de la paire ne la possèdent pas (−−) ;
- les paires discordantes lorsqu'un sujet seulement présente la caractéristique (+ − ou − +).

Le test du χ^2 de Mc Nemar permet d'effectuer la comparaison de deux pourcentages sur deux séries appariées.

Le mode de calcul et l'interprétation sont détaillés au chapitre 13.XVII.

2. Interprétation du test χ^2 avec un risque α fixé à 5 %

a) Hypothèse H_1 bilatérale

- Lorsque la valeur observée χ^2_o est inférieure à $\chi^2_{5\%}$, on formule qu'**on ne rejette pas l'hypothèse nulle**. On ne peut pas affirmer que les échantillons proviennent de populations différentes. On dit que la différence entre les distributions n'est pas significative.
- Lorsque la valeur observée χ^2_o est supérieure à $\chi^2_{5\%}$, on formule le rejet de H_0. On accepte H_1 en affirmant que les échantillons proviennent de populations différentes. On affirme que la différence entre les distributions est significative. On recherche le degré de signification p (*cf.* chap. 11.IV.3 : utilisation de la table χ^2).

b) Hypothèse H_1 unilatérale

Ce type d'hypothèse n'est proposé que lorsqu'on compare deux pourcentages (tableau à 4 cases). Dans ce cas on peut s'intéresser au sens de la différence. On postule que l'un des pourcentages est supérieur ou inférieur à l'autre. Le risque d'erreur est donc deux fois moindre que dans l'hypothèse bilatérale. La valeur seuil pour un risque de 5 % est de $\chi^2_{10\%}$.

- Lorsque la valeur observée χ^2_o est inférieure à $\chi^2_{10\%}$, on formule qu'**on ne rejette pas l'hypothèse nulle**. La différence entre les pourcentages n'est pas significative.
- Lorsque la valeur observée χ^2_o est supérieure à $\chi^2_{10\%}$, on formule le **rejet** de H_0. On accepte H_1 en affirmant non seulement que la différence entre les pourcentages est significative, mais en outre que l'un des pourcentages est **inférieur (ou supérieur)** à l'autre.

3. Utilisation de la table de χ^2

■ **Degré de liberté**

Dans un tableau de contingence, on appelle degrés de liberté (ddl), le nombre de cases qu'on peut remplir librement lorsque les totaux des lignes et des colonnes sont fixés. Par exemple, dans un tableau à 4 cases dont les totaux sont fixés, on est libre de ne choisir que l'effectif d'une seule case. Les autres effectifs sont alors automatiquement déduits.

De façon générale, dans un tableau de contingence à r lignes et k colonnes, le nombre de degrés de liberté est :

$$ddl = (r - 1) \times (k - 1)$$

Dans l'exemple 11.8, on a ddl = (3 – 1) (3 – 1) = 4.

■ **Il existe autant de distributions de χ^2 que de degrés de liberté (ddl)**

On commence donc par repérer dans la table χ^2 (Table 5), à la ligne correspondante au nombre de ddl, la valeur $\chi^2_{5\%}$. Si le calcul du test montre une valeur observée χ^2_o supérieure à $\chi^2_{5\%}$, on rejette H_0; on cherche ensuite, dans la même ligne, la valeur de χ^2 immédiatement inférieure à χ^2_o. Le risque α correspondant à cette valeur donne le degré de signification p.

Si l'hypothèse H_1 est unilatérale, on rejette H_0 pour une valeur de χ^2_o supérieure à $\chi^2_{10\%}$. Le degré de signification p, obtenu comme précédemment, est divisé par 2.

Exemple 11.9. INTERPRÉTATION D'UN TEST DE χ^2

Un test de χ^2 à 4 ddl a montré la valeur $\chi^2_o = 12$. Pour ddl = 4, $\chi^2_{5\%} = 9{,}49$. On rejette donc H_0. La valeur de χ^2 immédiatement inférieure à 12 est 11,67 qui correspond à un risque de 0,02. On accepte donc la différence entre les distributions étudiées avec $p < 0{,}02$.

α	0,0001	0,001	0,01	0,02	0,03	0,04	0,05	0,10	0,20	0,30	0,50	0,90
$\chi^2_{ddl=4}$	23,51	18,47	13,28	11,67	10,71	10,03	9,49	7,78	5,99	4,88	3,36	1,06
	zone de rejet H_0							zone de non-rejet de H_0				

V. TEST EXACT DE FISHER

Lorsqu'un tableau de contingence comprend 4 cases, le test revient à comparer deux pourcentages. Nous avons vu que ce test pouvait être réalisé par un test de χ^2 à 4 cases à condition que les effectifs théoriques soient supérieurs ou égaux à 5. Lorsque cette condition n'est pas remplie, il existe une façon exacte de tester l'homogénéité de 2 distributions de 2 variables binaires : le test exact de Fisher.

1. Principe du test exact de Fisher

La méthode exacte de Fisher consiste à calculer, si H_0 est vraie, la probabilité d'avoir observé une configuration donnant un écart au moins aussi grand que l'écart observé entre les 2 pourcentages que l'on compare. Si cette probabilité p est inférieure à 5 %, on rejette l'hypothèse H_0 d'homogénéité entre les deux distributions. En d'autres termes, on juge peu probable que la configuration observée soit due au hasard. On rejette donc l'hypothèse d'égalité entre les 2 pourcentages.

2. Calcul du test de Fisher

Pour pratiquer un test exact de Fisher, il faut écrire les tableaux de chacune des configurations donnant une différence entre pourcentages au moins aussi grande que la différence observée et en calculer les probabilités respectives.
La formule de calcul des probabilités de chaque configuration est donnée en Annexes, Formulaire 20.

Lorsque les effectifs sont élevés, les configurations sont très nombreuses et les calculs très fastidieux. En pratique, on utilisera les résultats fournis par des logiciels. Le test exact de Fisher, qui était surtout utilisé lorsque les effectifs d'un tableau de contingence étaient inférieurs à 5, devrait maintenant être utilisé de façon systématique puisqu'il fournit des valeurs exactes. Lorsqu'un logiciel statistique fournit les résultats de plusieurs tests de comparaison de pourcentages (χ^2 simple, χ^2 corrigé, χ^2 corrigé de Yates, test de Fisher), il faut prendre le résultat du test de Fisher qui est exact (exemple 11.10).

Exemple 11.10. TEST EXACT DE FISHER

Soit le tableau de résultats suivant :

6	2	8
1	8	9
7	10	17

$p_1 = 6/7 = 85{,}7\ \%$ $\quad$ $p_2 = 2/10 = 20{,}0\ \%$

Sous H_0, les 2 pourcentages observés devraient être identiques. Leur différence devrait être nulle. Dans cet exemple, leur différence est de 65,7 %. Le principe du test est de calculer la probabilité d'observer une différence au moins aussi grande, si H_0 est vraie.
En faisant varier les effectifs de chaque case, tout en maintenant les totaux des lignes et des colonnes, on obtient l'ensemble des configurations possibles du tableau de contingence. Sous H_0, on peut calculer la probabilité de chacune de ces configurations.

La figure 11.9 donne l'ensemble des configurations possibles du tableau. L'histogramme représente leurs probabilités respectives sous H_0. On peut vérifier que seules les configurations 1, 2 et 8 présentent une différence $|p_1 - p_2|$ supérieure ou égale à 65,7 %. Les probabilités respectives de ces 3 configurations sont 0,0004, 0,013 et 0,002. Le résultat du test exact de Fisher est la somme p de ces probabilités.
$p = 0{,}0004 + 0{,}013 + 0{,}002 = 0{,}0154$. La probabilité d'avoir observé une configuration donnant un écart de pourcentage au moins aussi grand que 65,7 % est donc inférieure à 5 % ; on conclut donc que la distribution observée n'est pas due au hasard. Il existe une différence significative entre les 2 pourcentages comparés avec $p < 0{,}016$. Si H_1 avait été unilatérale ($p_1 > p_2$), on n'aurait seulement considéré que les 2 premières configurations et on aurait $p = 0{,}0004 + 0{,}013 = 0{,}0134$. On conclurait à une différence significative avec $p < 0{,}014$.

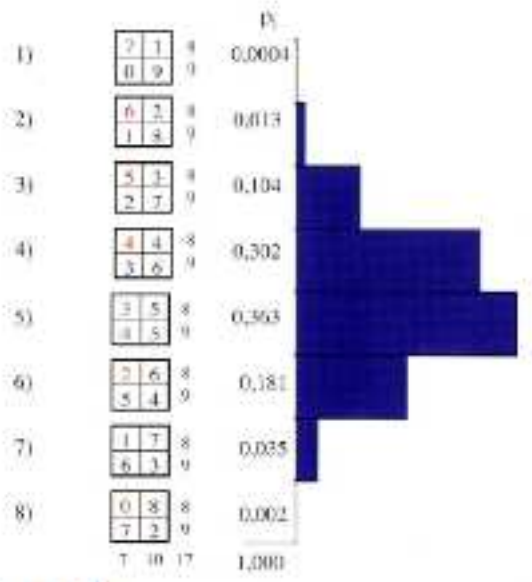


Figure 11-9.

VI. TESTS NON-PARAMÉTRIQUES OU TESTS DE RANGS

Cette famille de tests a les mêmes applications que les tests paramétriques de comparaison de variables quantitatives (moyennes, variances). Ils s'en distinguent de façon fondamentale en étant basés non pas sur la comparaison des ***valeurs*** des variables étudiées, mais sur les ***rangs*** des individus classés selon la valeur des variables après avoir mélangé les séries à comparer.
On appelle rang, le numéro d'ordre d'une valeur après classement de la variable par ordre croissant. Sur la série 12, 41, 53, 82, la valeur 53 a pour rang 3 et la valeur 82 a pour rang 4.

Sous H_0, les individus devraient être rangés de façon aléatoire, les valeurs d'une série alternant avec celle de l'autre, comme les couleurs dans un jeu de cartes très bien battues (figure 11.10). Sous H_1, si les valeurs d'une des séries à comparer sont en moyenne plus élevées, leurs rangs après classement sont donc en moyenne plus élevés.

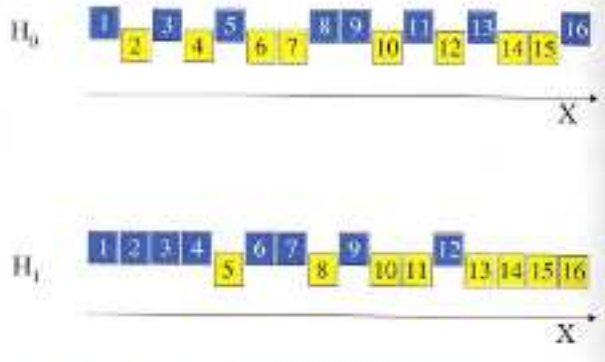


Figure 11-10. Test de rang non-paramétrique

On compare les rangs des observations classées selon leurs valeurs. Sous l'hypothèse nulle, les valeurs des 2 séries devraient être mélangées de façon homogène. La somme des rangs des 2 séries réunies est égale à la somme des N premiers entiers : $N(N+1)/2$. Si les valeurs sont mélangées de façon homogène, on démontre que la somme attendue des rangs de chaque série est respectivement $n_1(N+1)/2$ et $n_2(N+1)/2$. Sous H_0, la différence entre la somme des rangs d'une des séries et sa valeur attendue fluctue autour de zéro. Le rapport de cette différence sur son écart type suit une loi de Z normale centrée réduite.

L'interprétation finale du test est identique à celle d'un test paramétrique.

L'avantage majeur d'un test non paramétrique est de s'affranchir des conditions d'application usuelles des tests paramétriques (normalité des distributions, égalité des variances, *etc.*). Sa difficulté résidait dans la nécessité de classer les individus lorsque les séries étaient de grande taille. Ce travail était long et fastidieux. Depuis la généralisation des ordinateurs et des tableurs modernes, cette opération est quasi instantanée.
Lorsqu'il existe un doute sur les conditions d'application, il ne faut pas hésiter à utiliser un test non-paramétrique.

Principaux tests non-paramétriques

TEST	APPLICATION	TEST PARAMÉTRIQUE ÉQUIVALENT
Wilcoxon	Comparaison de 2 moyennes	T de Student, test Z
Kruskal-Wallis	Comparaison de plusieurs moyennes	F (ANOVA)

Les calculs de ces tests sont détaillés au chapitre 13.XI, XII et XIII.

Exercices

Exercice 11.1

Quels tests utiliser pour :

1) comparer les moyennes des poids d'un échantillon de 35 animaux alimentés par un produit A et de 25 animaux alimentés par un produit B ;
2) comparer les moyennes de temps réalisés dans une course contre la montre de 7 équipes de 10 coureurs cyclistes ;
3) comparer les pourcentages de conifères observés dans des échantillons pris dans 6 parcelles de forêt ;
4) vérifier si la structure d'âge d'un échantillon d'individus est différente de la structure d'âge de la population dont ils sont issus ;
5) comparer la proportion d'infections nosocomiales parmi 8 malades d'une salle A et 9 malades d'une salle B ;
6) comparer la fréquence de répartition des 4 groupes sanguins parmi 5 groupes de nationalité différente ?

Exercice 11.2

La comparaison de 2 traitements a montré une différence significative ($p < 0,04$) lors d'une première étude et un résultat non significatif lors d'une deuxième étude. Les deux études de protocole identique ont porté sur des échantillons différents mais de tailles équivalentes.

Comment peut-on expliquer ces résultats discordants ?

Exercice 11.3

On désire comparer dans 3 groupes de malades E_1, E_2 et E_3, la distribution de l'évolution d'une maladie divisée en 3 classes : guérison, rechute, décès ; les échantillons ont pour taille respective, 29, 19 et 33 individus. Le tableau ci-dessous donne les effectifs observés de chaque classe de la variable pour chaque échantillon. Y a-t-il une différence dans l'évolution de la maladie entre les 3 groupes ?

Tableau de contingence des effectifs observés

		Échantillons			Total	
		E_1	E_2	E_3	N	%
Évolution de la maladie	guérison	5	6	16	27	*33,3*
	rechute	9	9	10	28	*34,6*
	décès	15	4	7	26	*32,1*
	Total	29	19	33	81	*100,0*

Résumé

Les tests de comparaison servent à comparer des différences entre des moyennes, des variances, des distributions d'effectifs ou des pourcentages.
Leur principe est toujours le même : il consiste à poser une hypothèse nulle H_0 d'égalité entre les paramètres ou les distributions étudiées et à rejeter cette hypothèse si elle n'est pas vérifiée, au profit d'une hypothèse alternative.
Le calcul du test consiste à calculer une expression mathématique de la différence qui, sous H_0, suit un modèle théorique d'une loi de probabilité connue. Si l'expression calculée est peu probable (inférieure à 5 %), on rejette H_0.
Selon la nature de la comparaison à effectuer, on utilise la loi Z normale centrée réduite, la loi T de Student, la loi F de Fisher ou la loi du χ^2.

PRINCIPAUX TESTS DE COMPARAISON

LE TEST	SERT À COMPARER...
Z de l'écart réduit	1 moyenne observée à une moyenne théorique 2 moyennes 2 moyennes de 2 séries appariées
T de Student	1 moyenne observée à une moyenne théorique 2 moyennes 2 moyennes de 2 séries appariées
Wilcoxon	2 moyennes
F de Fisher-Snedecor	2 variances plusieurs moyennes
Kruskal-Wallis	plusieurs moyennes
χ^2 de conformité χ^2 d'homogénéité χ^2 à 4 cases χ^2 de McNemar	1 distribution observée à une distribution théorique plusieurs distributions, plusieurs pourcentages 2 pourcentages 2 pourcentages sur 2 séries appariées
Fisher exact	2 pourcentages

12

TESTS DE LIAISON

Dans ce chapitre, on utilise tour à tour les termes d'« indépendance » et de « liaison » qui parfois prêtent à confusion. Comme on l'a vu brièvement au chapitre 10.II, le principe des tests de liaison consiste à tester « l'indépendance » entre 2 variables (hypothèse nulle) et en cas de rejet à accepter qu'il existe une « liaison » entre ces variables (hypothèse alternative). C'est ainsi qu'il faut comprendre ces deux termes.

Ce chapitre n'aborde que le principe général des tests de liaison. Le chapitre 13 détaille le processus de calcul de chaque test.

Les tests utilisés dépendent de la nature des variables étudiées. On distingue l'étude de la liaison entre deux variables qualitatives, entre une variable binaire et une variable qualitative ordinale, et entre deux variables quantitatives.

I. TEST DU χ^2 D'INDÉPENDANCE

Le test de χ^2 d'indépendance s'applique à l'étude de la liaison entre deux variables qualitatives.

1. Principe

Si A et B sont des variables qualitatives nominales, on obtient un tableau de contingence (chap. 11.IV.1.b). Les lignes du tableau représentent les classes de A, les colonnes les classes de B.

Sous H_0, la distribution de A devrait être indépendante de la distribution de B. En d'autres termes, la fréquence des effectifs dans chaque classe de A devrait être identique, quelle que soit la classe de B. Si les distributions de A sont différentes selon B, on rejette H_0, et on accepte H_1 (bilatérale) : il y a une liaison entre A et B.

Sur la figure 12.1, on observe dans le tableau supérieur que la distribution de la variable A est indépendante de la variable B. Quelle que soit B, la classe A1 est plus fréquente que la classe A3.
Dans le tableau inférieur, il existe une liaison entre A et B. La classe A1 est plus fréquente si B = 1. À l'inverse, la classe A3 est plus fréquente si B = 3.
Le mode de calcul et les conditions d'application du test et de ses variantes sont les mêmes que celles exposées au chapitre 11.IV. Les calculs sont détaillés au chapitre 13.XVIII.

2. Interprétation du test du χ^2 d'indépendance

Si χ^2 observé est supérieur à $\chi^2_{5\%}$, on rejette H_0. On accepte donc l'hypothèse H_1 d'une liaison entre les deux variables A et B. On dit que la liaison est significative. Comme dans le cas général d'un test de χ^2 (chap. 11.IV.3), on recherche dans la table de χ^2, la valeur du degré de signification p.

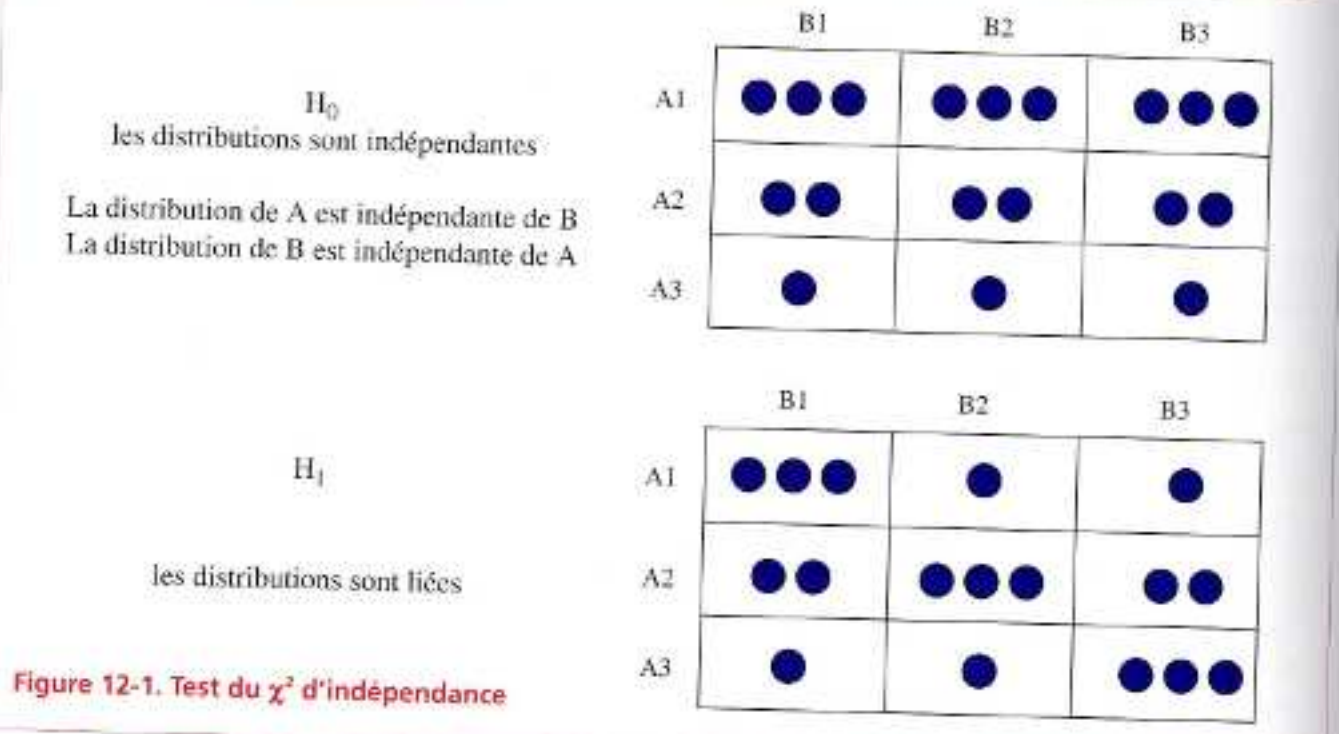


Figure 12-1. Test du χ^2 d'indépendance

II. TEST DU χ^2 DE TENDANCE

Ce test s'applique à l'étude de la liaison entre des pourcentages (variable qualitative binaire) et une variable qualitative de type ordinal (variable qualitative ordonnée selon une hiérarchie).
Le test est appelé ***test de tendance***. Une des conditions d'application suppose qu'il existe une liaison linéaire entre les pourcentages et les valeurs de la variable ordinale. Sous H_0, les pourcentages sont identiques quelle que soit la classe de la variable X. Sous H_1, on affirme : 1) que les pourcentages diffèrent et 2) qu'ils augmentent ou diminuent en fonction de l'**ordre** de la classe de X (figure 12.2).
Comme on le constate, ce test permet non seulement d'affirmer que les pourcentages diffèrent entre eux (comme le ferait un simple test de χ^2 de comparaison de plusieurs pourcentages), mais qu'ils diffèrent proportionnellement aux valeurs de la variable ordinale.

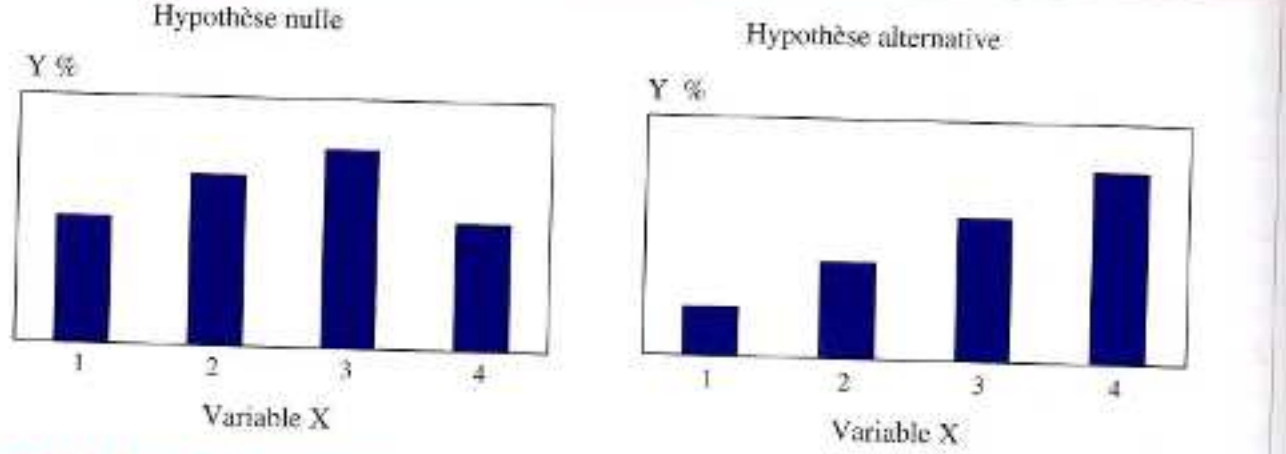


Figure 12-2. Test du χ^2 de tendance

Le test aboutit à calculer une valeur qui suit une loi de χ^2. Comme dans le cas général d'un test de χ^2 (chap. 11.IV.3), on recherche dans la table de χ^2, la valeur du degré de signification p.
Les calculs détaillés du test sont exposés au chapitre 13.XIX.

III. TESTS DE CORRÉLATION

Le test de corrélation sert à étudier la liaison entre 2 variables quantitatives. Ce test est utilisé lorsque les deux variables X et Y sont aléatoires et jouent un rôle symétrique (les deux variables peuvent être placées indifféremment en abscisses ou en ordonnées). On cherche simplement à savoir s'il existe une liaison entre ces deux variables et à quantifier l'intensité de la liaison.

Si X et Y sont des variables quantitatives, on obtient sur un graphe un nuage de points (figure 12.3). Chaque point représente une valeur de X pour une valeur de Y chez le même individu. Les variables X et Y peuvent être de nature différente, exprimées avec des unités différentes.
Dire que deux variables sont corrélées, c'est affirmer qu'il existe une liaison entre ces deux variables. Plus X varie dans un sens, plus Y varie. Si Y varie dans le même sens, on dit que la corrélation est positive, si Y varie dans le sens opposé on dit que la corrélation est négative. À l'inverse, on dira que les variables ne sont pas corrélées lorsqu'elles varient indépendamment l'une de l'autre.

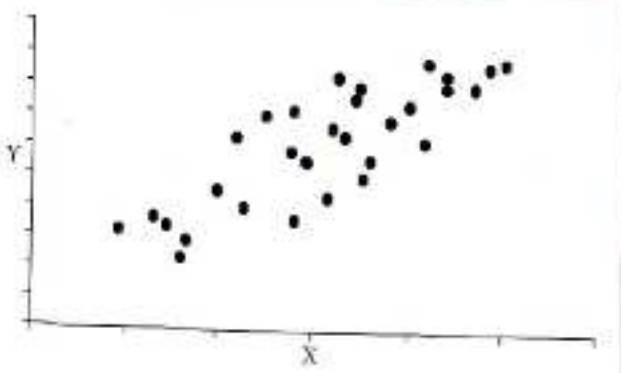


Figure 12-3.

1. Covariance

La covariance est un indicateur qui mesure la liaison entre deux variables X et Y. On appelle ***covariance***, la moyenne des produits des écarts de X et Y à leur moyenne respective m_x et m_y (*cf.* formule de la covariance, Annexes, Formulaire 13).
Les points de la figure 12.4 représentent 4 couples de valeurs **x** et **y**. Les distances de chaque point à l'axe des moyennes $\mathbf{m_x}$ et $\mathbf{m_y}$ représentent les écarts aux moyennes. Les aires des rectangles représentent donc les produits des écarts $(x - m_x)(y - m_y)$ de chaque couple. Les rectangles en mauve représentent les produits positifs et les rectangles verts les produits négatifs. La covariance peut être illustrée par la somme des aires des rectangles.
Sur le diagramme 12.4.a, on constate que la somme de ces rectangles est proche de zéro. La covariance est nulle. Il n'y a apparemment pas de liaison entre X et Y.

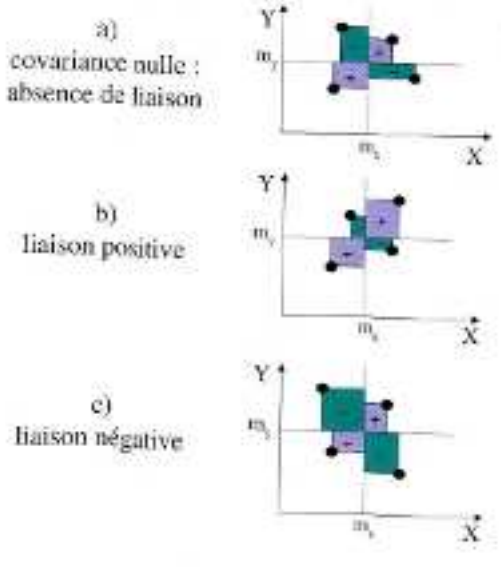


Figure 12-4.

Sur le diagramme 12.4.b, on constate que la somme des rectangles est positive. Il semble exister une liaison positive entre X et Y ; plus X est élevé, plus Y est élevé.
Sur le diagramme 12.4.c, on constate que la somme des rectangles est négative. Il semble exister une liaison négative entre X et Y ; plus X est élevé, plus Y est bas.

2. Coefficient de corrélation

La covariance est le produit de deux termes exprimés en unités qui peuvent être différentes. Elle ne se prête donc pas à l'analyse statistique. On réduit alors la covariance en la divisant par le produit des écarts types s_x et s_y de chaque distribution. On obtient ainsi un coefficient sans unité appelé ***coefficient de corrélation*** **r** (*cf.* formule du coefficient de corrélation, Annexes, 24.13).
On démontre que ce coefficient de corrélation varie entre – 1 (corrélation négative) et + 1 (corrélation positive) en passant par 0 (absence de corrélation). Plus ce coefficient en valeur absolue est proche de 1, plus la liaison est forte.
Le coefficient de corrélation est un indicateur de la force de la liaison.
On peut aussi calculer le ***coefficient de détermination*** R^2 qui est le carré du coefficient de corrélation. Il varie donc entre 0 et 1 et s'exprime souvent en pourcentage. Il exprime la part de la dispersion des valeurs due à la corrélation. Par exemple, un coefficient de détermination R^2 de 0,49 ($r = 0,70$) signifie que 49 % de la dispersion des valeurs de X est due à la corrélation et 51 % est due aux aléas.

3. Test du coefficient de corrélation

Lorsqu'on a calculé un coefficient de corrélation **r** entre 2 variables X et Y, la question qui se pose est de savoir si ce coefficient est suffisamment éloigné de zéro, pour affirmer la liaison. On propose comme hypothèse nulle qu'il n'existe aucune liaison entre X et Y dans la population d'où est issu l'échantillon étudié : sous H_0 le coefficient **r** est donc nul.
Sous H_0, le rapport $|r - 0|$ sur son écart type suit une loi T de Student.
Si le test aboutit à rejeter H_0, **r** est significativement différent de 0, c'est-à-dire qu'il existe une liaison significative entre X et Y.
Le calcul du test et ses conditions d'application sont détaillés au chapitre 13.XX.

Interprétation

Un test de corrélation qui permet de rejeter H_0 aboutit à deux valeurs.

1) Une valeur de *p* qui mesure son degré de signification.
2) Une valeur de **r** qui mesure la force de l'association. Plus sa valeur absolue est élevée, plus la liaison est forte.

- Un test de corrélation ne peut s'interpréter que pour la plage des valeurs étudiées. En deçà et au-delà de cette plage, la relation entre X et Y peut très bien s'écarter d'une liaison linéaire (figure 12.5a). L'extrapolation des résultats à des valeurs extrêmes non testées est donc dangereuse.
- Un coefficient de corrélation élevé, n'implique pas obligatoirement une relation déterministe linéaire entre X et Y (figure 12.5b).

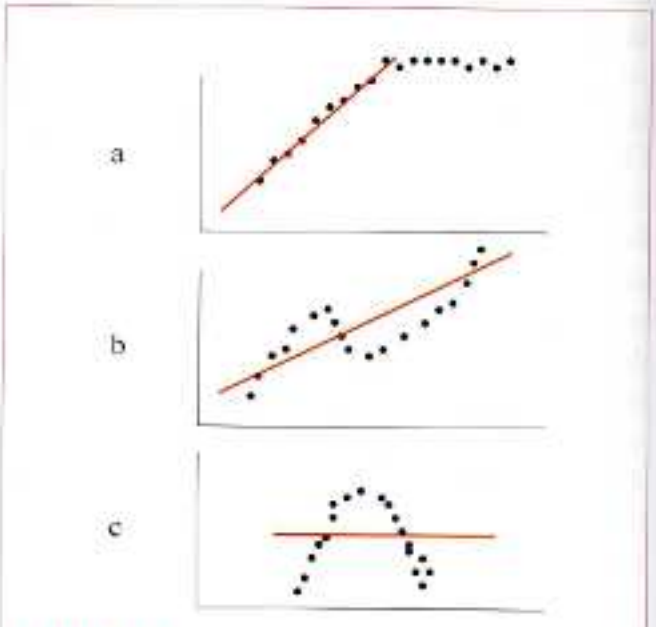


Figure 12-5.

- Un coefficient de corrélation proche de zéro, n'implique pas obligatoirement l'absence de liaison (figure 12.5c).
- Enfin et surtout, un coefficient de corrélation, même proche de 1, n'est en aucun cas un critère de causalité. Il permet seulement d'affirmer une liaison statistique entre deux variables et de générer des hypothèses.

4. Test de corrélation des rangs de Spearman

Ce test fait partie de la catégorie des tests non-paramétriques. Son principe consiste à classer les valeurs des deux séries et à tester la corrélation, non pas entre les valeurs, mais entre leurs **rangs**. Son interprétation est identique à celle du test de corrélation **r**. Son calcul pratique est exposé chapitre 13.XXI.

IV. RÉGRESSION

La corrélation servait à étudier la liaison de deux variables jouant un rôle symétrique (X et Y interchangeables). Il arrive souvent que les deux variables ne jouent pas un rôle symétrique. L'une est une conséquence de l'autre. Si on étudie par exemple la liaison entre le poids des enfants et leur âge, la variable « poids » est une variable ***dépendante*** de la variable « âge », qui est la variable ***explicative***. Le poids est fonction de l'âge, mais l'inverse n'est pas vrai. On appelle alors X la variable explicative et Y la variable dépendante de X.
On utilise la régression lorsqu'on désire analyser ce type de problème. On cherche à savoir quelle est la relation de dépendance de Y par rapport à X. On dit que l'on recherche une ***régression*** de Y en fonction de X.

La régression a trois fonctions essentielles :
- décrire la façon dont Y est liée à X ;
- tester l'existence de la liaison ;
- estimer une valeur de Y pour une valeur donnée de X.

1. Description

Nous nous limiterons ici au cas le plus simple dans lequel le modèle utilisé pour résumer la variation de Y en fonction de X suit une droite.
Lorsqu'on dispose d'un tableau de données de 2 variables quantitatives appariées, nous avons vu qu'on pouvait représenter ces données par un graphe de points, ayant pour abscisse les valeurs x et comme ordonnées les valeurs y.
Le premier problème consiste à tracer la droite qui résume au mieux la relation entre X et Y. Cette droite se représente graphiquement selon la forme $y = a + bx$ où **b** est la pente de la droite.
Sur la figure 12.6, on a représenté trois fois la même série de points. Les trois droites ont été tracées pour tenter de représenter au mieux la distribution de Y en fonction de X. On pourrait en tracer encore plus.

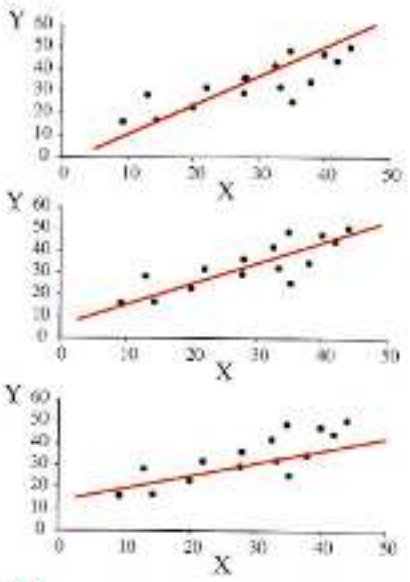


Figure 12-6.

Intuitivement, la meilleure estimation du modèle, donc la meilleure droite, est celle pour laquelle la somm des distances verticales de chaque point à cette droite serait la plus faible. Comme certaines valeurs so positives et d'autres négatives, on calcule plutôt la somme des carrés des distances.
On cherche donc les paramètres a et b d'une droite telle que la somme des carrés des distances so minimale.

Cette droite des moindres carrés est appelée droite de régression (figure 12.7).

- Elle passe par le point correspondant aux deux moyennes m_x et m_y.
- Sa pente **b** est donnée par le rapport de la covariance XY sur la variance de X (*cf.* Annexes, For mulaire 14).
- Lorsque les deux variables X et Y sont indépendantes ou non liées, la droite de régression est hori zontale et sa pente est nulle.

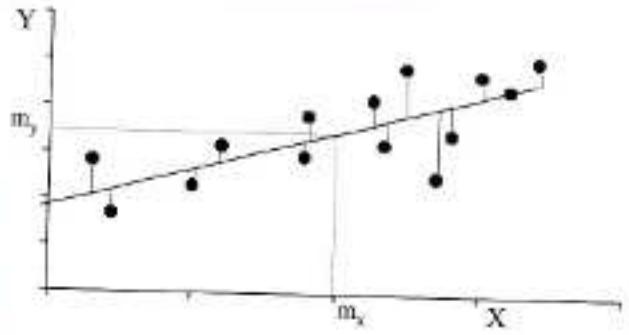


Figure 12-7. Droite des moindres carrés

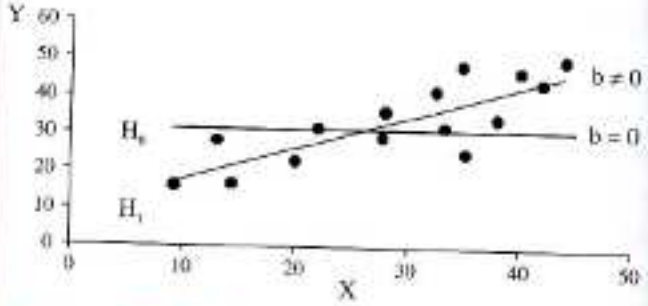


Figure 12-8. Test de la pente de la droite de régression

2. Test de la pente de la droite de régression

Lorsqu'on a calculé la pente d'une droite de régression, la question qui se pose est de savoir si cette pente est suffisamment éloignée de zéro, pour affirmer la liaison entre les variables X et Y. On propose comme hypothèse nulle que la pente de la droite est horizontale dans la population d'où est issu l'échantillon étudié (figure 12.8). Sous H_0, le rapport |b-0| sur son écart type suit une loi T de Student.
Les éléments du calcul du test de la pente figurent en Annexes, Formulaire 14.
Son interprétation est identique à celle du test de corrélation.
Lorsqu'on rejette H_0, on admet que la pente de la droite de régression est oblique. On accepte donc l'hypothèse alternative d'une liaison entre les deux variables. Lorsqu'on ne rejette pas H_0, cela signifie soit qu'il n'existe aucune liaison détectable, soit que la liaison n'est pas linéaire.

3. Estimations

L'intérêt principal d'une droite de régression est de pouvoir estimer une valeur de **Y** connaissant une valeur **x**. On quitte ici la problématique des tests pour revenir à celle de l'estimation d'un paramètre.
On peut ainsi estimer :

- la valeur moyenne de Y pour une valeur donnée de X ;
- la valeur y pour un individu présentant une valeur x.

Les formules correspondantes figurent en Annexes, Formulaire 14.

Exercice

Quels tests choisir pour :

1) tester la liaison entre la hauteur des arbres et leur altitude ;
2) étudier l'effet de la dose d'un traitement sur le nombre de leucocytes sanguins ;
3) tester la liaison entre le poids et la taille ;
4) tester l'effet de la dose d'un traitement sur la fréquence des effets secondaires ?

Résumé

Les tests de liaison servent à vérifier qu'il existe une relation de dépendance entre deux variables observées sur un échantillon. H_0 = indépendance, H_1 = liaison.

- Si les variables sont qualitatives, on utilise le test du χ^2. Si H_0 est rejetée, cela signifie que les variables sont significativement liées.
- Si les variables sont quantitatives, on utilise les tests de corrélation ou de régression. Les tests de corrélation (coefficient r et test des rangs de Spearman) servent à tester la liaison entre deux variables jouant un rôle symétrique dans leur dépendance. Le test de la pente de la droite de régression sert à tester la liaison entre une variable X explicative, et une variable Y dépendante de X. En outre, les paramètres de la régression permettent de procéder à des estimations de la variable dépendante Y en fonction d'une valeur donnée de X.

PRINCIPAUX TESTS DE LIAISON

LE TEST	SERT À TESTER...
χ^2 d'indépendance	l'indépendance entre 2 variables qualitatives
χ^2 de tendance	l'indépendance entre 1 variable ordinale et plusieurs pourcentages
T de Student	un coefficient de corrélation
	un coefficient de corrélation de Spearman
	la pente d'une droite de régression
F de Fisher-Snedecor	la pente d'une droite de régression

POUR TESTER LA LIAISON ENTRE	ON UTILISE...
deux variables qualitatives une variable qualitative ordinale et plusieurs pourcentages	test χ^2 d'indépendance test χ^2 de tendance
deux variables quantitatives symétriques	tests de corrélation
deux variables quantitatives, 1 dépendante et 1 explicative	test de la droite de régression

13

UTILISATION PRATIQUE DES TESTS STATISTIQUES

Ce chapitre aborde le processus détaillé du calcul de chaque test statistique.

I. CRITÈRES DE CHOIX D'UN TEST STATISTIQUE

Le choix d'un test dépend de plusieurs facteurs qu'il importe d'identifier au préalable.

- ***La nature des variables à comparer :***
  - quantitative continue ou discrète ;
  - qualitative binaire ;
  - qualitative nominale à plusieurs classes ;
  - qualitative ordinale.
- ***Les grandeurs étudiées :***
  - moyennes ;
  - pourcentages ;
  - variances ;
  - effectifs ;
  - rangs.
- ***La nature du problème :***
  - comparaison d'un échantillon à une population de référence ;
  - comparaison de deux échantillons ;
  - comparaison de plusieurs échantillons ;
  - liaison entre deux variables.
- ***Le type de séries comparées :***
  - indépendantes ;
  - appariées.
- ***La taille des échantillons.***
- ***Les conditions d'applications des tests.*** Selon les cas :
  - normalité des distributions dans la population d'où est issu l'échantillon ;
  - égalité des variances ;
  - taille minimum des échantillons.

II. STRATÉGIE D'UTILISATION DES TESTS STATISTIQUES

1. Domaines d'application

LE TEST	SERT À COMPARER OU À TESTER...	CHAPITRE 13
Z de l'écart réduit	1 moyenne observée à une moyenne théorique	III
	2 moyennes	IV
	2 moyennes de 2 séries appariées	V
T de Student	1 moyenne observée à une moyenne théorique	VI
	2 moyennes	VII
	2 moyennes de 2 séries appariées	VIII
	un coefficient de corrélation	XX
	un coefficient de corrélation de Spearman	XXI
Wilcoxon	2 moyennes de rangs de 2 séries ordonnées	XI
	2 séries appariées ordonnées	XII
F de Fisher-Snedecor	2 variances	IX
	plusieurs moyennes	X
Kruskal-Wallis	plusieurs moyennes de rangs de séries ordonnées	XIII
χ^2 de conformité	1 distribution observée à une distribution théorique	XIV
χ^2 d'homogénéité	plusieurs distributions, plusieurs pourcentages	XV
χ^2 à 4 cases	2 pourcentages	XVI
χ^2 de McNemar	2 pourcentages sur 2 séries appariées	XVII
χ^2 d'indépendance	la liaison entre 2 variables qualitatives	XVIII
χ^2 de tendance	liaison entre variable ordinale et plusieurs pourcentages	XIX
Fisher exact	2 pourcentages (*cf.* chap. 11.V)	

2. Choix d'un test en fonction de la nature du problème

■ Comparer un échantillon à une population de référence ou théorique

GRANDEUR ÉTUDIÉE	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
moyenne	Z de l'écart réduit	effectif échantillon $n \geq 30$	III
	t Student	normalité de la distribution	VI
pourcentage	χ^2 de conformité	effectifs théoriques ≥ 5	XIV
distribution	χ^2 de conformité	effectifs théoriques ≥ 5	XIV

■ Comparer deux échantillons entre eux

GRANDEUR ÉTUDIÉE	SÉRIES COMPARÉES	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
moyennes	Indépendantes	Z de l'écart réduit	effectif des 2 échantillons n_1 et $n_2 \geq 30$	IV
		T Student	normalité des 2 distributions	VII
moy. rangs		Wilcoxon		XI
	Appariées	Z de l'écart réduit	nombre de paires ≥ 30	V
		T Student	normalité des différences	VIII
moy. rangs		Wilcoxon	nombre de paires > 20	XII
variances	Indépendantes	F Fisher-Snedecor	distributions normales et de même variance	IX
pourcentages	Indépendantes	χ^2 4 cases	effectifs théoriques ≥ 5	XVI
		Fisher	test exact (*cf.* § 11.5)	
	Appariées	χ^2 McNemar	nombre paires discordantes ≥ 10	XVII
distributions	Indépendantes	χ^2 d'homogénéité	effectifs théoriques ≥ 5	XV

■ Comparer plusieurs échantillons

GRANDEUR ÉTUDIÉE	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
moyennes	F Fisher-Snedecor	distributions normales et de même variance	X
moyennes des rangs	Kruskal-Wallis	effectifs de chaque échantillon > 10	XIII
pourcentages	χ^2 d'homogénéité	effectifs théoriques ≥ 5	XV
distributions	χ^2 d'homogénéité	effectifs théoriques ≥ 5	XV

■ Tester la liaison entre deux variables

GRANDEUR ÉTUDIÉE	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
Distributions	χ^2 d'indépendance	effectifs théoriques ≥ 5	XVIII
Pourcentages	χ^2 de tendance	variable explicative de type qualitatif ordinal	XIX
Coefficient de corrélation	T de corrélation	conditions d'un test T de Student	XX
Coef. corr. entre rangs	test de Spearman	nombre de paires > 10	XXI

3. Choix d'un test en fonction des paramètres à comparer

■ Moyennes

TYPE DE COMPARAISON	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
observée/théorique	Z de l'écart réduit	effectif échantillon $n \geq 30$	III
	T Student	normalité de la distribution	VI
deux moyennes	Z de l'écart réduit	effectifs des 2 échantillons n_1 et $n_2 \geq 30$	IV
sur séries indépendantes	T Student	normalité des 2 distributions	VII
2 moyennes rangs	Wilcoxon	nombre de paires > 20	XI
deux moyennes	Z de l'écart réduit	nombre de paires ≥ 30	V
sur séries appariées	T Student	normalité des différences	VIII
rangs séries appariées	Wilcoxon	nombre de paires > 20	XII
plusieurs moyennes	F Fisher-Snedecor	distributions normales et de même variance	X
plusieurs moy. rangs	Kruskal-Wallis	effectif de chaque échantillon > 10	XIII

■ Pourcentages

TYPE DE COMPARAISON	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
observé/théorique	χ^2 de conformité	effectifs théoriques ≥ 5	XIV
deux pourcentages sur	χ^2 4 cases	effectifs théoriques ≥ 5	XVI
2 séries indépendantes	Fisher	test exact (*cf.* § 11.5)	
sur 2 séries appariées	χ^2 McNemar	nombre de paires discordantes ≥ 10	XVII
plusieurs pourcentages	χ^2 d'homogénéité	effectifs théoriques ≥ 5	XV

■ Distributions

TYPE DE COMPARAISON	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
observée/théorique	χ^2 de conformité	effectifs théoriques ≥ 5	XIV
2 ou plusieurs distributions	χ^2 d'homogénéité	effectifs théoriques ≥ 5	XV

■ Variances

TYPE DE COMPARAISON	TEST	CONDITIONS D'APPLICATION	CHAPITRE 13
2 variances	F Fisher-Snedecor	distributions normales et de même variance	IX

4. Conditions d'application des tests

TEST	POUR COMPARER OU TESTER	CONDITIONS D'APPLICATIONS
Z de l'écart réduit	moyenne observée/théorique	effectif échantillon $n \geq 30$
	2 moyennes	effectifs échantillons n_1 et $n_2 \geq 30$
	2 moyennes de 2 séries appariées	nombre de paires ≥ 30
T de Student	moyenne observée/théorique	normalité de la distribution dans la population
	2 moyennes	normalité des distributions dans les 2 populations et (variances égales ou bien $n_1 = n_2$)
	2 moyennes de 2 séries appariées.	normalité des différences
	un coefficient de corrélation	normalité des distributions et variance constante
	un coefficient de corrélation de Spearman	effectif échantillon > 10
Wilcoxon	2 moyennes de rangs	effectifs échantillons n_1 et $n_2 \geq 10$
	moyennes de rangs de 2 séries appariées	effectifs des paires ≥ 10
F de Fisher-S.	2 variances	normalité des distributions dans les 2 populations
	plusieurs moyennes	normalité des distributions dans les populations
Kruskal-Wallis	plusieurs moyennes de rangs	effectifs échantillons $n_i \geq 10$
χ^2 de conformité	distribution observée/théorique	effectifs théoriques ≥ 5
χ^2 d'homogénéité	plusieurs distributions ou pourcentages	effectifs théoriques dans chaque case ≥ 5
χ^2 à 4 cases	2 pourcentages	effectifs théoriques dans chaque case ≥ 5
χ^2 de McNemar	2 pourcentages sur 2 séries appariées	nombre de paires discordantes ≥ 10
χ^2 de tendance	liaison entre % et variable ordinale	effectifs théoriques $c_j \geq 5$ et liaison linéaire
χ^2 d'indépendance	liaison entre 2 variables qualitatives	effectifs théoriques dans chaque case ≥ 5
Fisher exact	2 pourcentages	aucune

III. TEST Z POUR COMPARER UNE MOYENNE OBSERVÉE À UNE MOYENNE THÉORIQUE

Quand choisir ce test ?

Lorsqu'on désire comparer une moyenne observée dans un échantillon à une moyenne connue dans une population de référence.

Variable	quantitative
Paramètre étudié	moyenne
Taille de l'échantillon	supérieure ou égale à 30
Hypothèse nulle	$M = \mu$
H_1 bilatérale	$M \neq \mu$
H_1 unilatérale	$M > \mu$ ou bien $M < \mu$

Formulations

μ : la moyenne théorique connue de la population de référence.
M : la moyenne inconnue de la population d'où est issu l'échantillon.
m : la moyenne observée dans un échantillon.
s : l'écart type de l'échantillon.
n : l'effectif de l'échantillon.

Conditions d'application

La taille de l'échantillon doit être supérieure ou égale à 30.
Si cette condition n'est pas réalisée, il faut utiliser le test T de Student.

Principe du test (*cf.* détails § 11.1)

Si l'hypothèse nulle est vraie,

- **m** est l'une des valeurs possibles d'une variable normale centrée autour de **μ (=M)** ;
- la différence entre cette variable et **μ** suit une loi normale de moyenne 0 ;
- le rapport de cette différence sur l'écart type de **μ** suit une loi de Z normale centrée réduite.

Calcul intermédiaire

L'écart type de **μ** peut être estimé par l'écart type de la moyenne de l'échantillon : $\frac{s}{\sqrt{n}}$

Test Z :

$$z = \frac{|m - \mu|}{\frac{s}{\sqrt{n}}}$$

Résultats

H_1	z	Rejet H_0	Interprétation
bilatérale	< 1,96	Non	m n'est pas significativement différent de μ
	≥ 1,96	Oui	m diffère significativement de μ
unilatérale	< 1,65	Non	m n'est pas significativement supérieur (ou inférieur) à μ
	≥ 1,65	Oui	m est significativement supérieur (ou inférieur) à μ

Exemple 13.3.

Lors d'une enquête sur la durée de sommeil des enfants de 2 à 3 ans dans un département français, on a trouvé une moyenne du temps de sommeil par nuit de 10,2 heures dans un groupe de 40 enfants. L'écart type est 2,1 heures. La moyenne attendue du temps de sommeil est de 11,7 heures chez les enfants de cet âge.
H_0 : les enfants de l'échantillon dorment autant que ceux de la population.
H_1 bilatérale : la durée de sommeil des enfants de l'échantillon est différente.

$z = (11,7 - 10,2)/(2,1/\sqrt{40}) = 4,5$. On rejette donc H_0.

Z est encore inférieur à 4,5, pour un risque $\alpha = 0,00001$.

Conclusion : la population des enfants examinés présente un temps de sommeil significativement plus court que la population générale ($p < 10^{-5}$).

IV. TEST Z POUR COMPARER DEUX MOYENNES

Quand choisir ce test ?

Lorsqu'on veut comparer les moyennes observées dans deux échantillons.
Autre test équivalent : test de Wilcoxon non paramétrique.

Variables	quantitatives
Paramètres étudiés	moyennes
Taille des échantillons	au moins de 30 par échantillon
Séries comparées	indépendantes
Hypothèse nulle	$\mu_1 = \mu_2$
H_1 bilatérale	$\mu_1 \neq \mu_2$
H_1 unilatérale	$\mu_1 > \mu_2$ ou bien $\mu_1 < \mu_2$

Formulations

μ_1 et μ_2 : les moyennes inconnues des deux populations d'où sont issus les échantillons.
m_1 et m_2 : les moyennes des deux échantillons.
s^2_1 et s^2_2 : les variances des deux échantillons
n_1 et n_2 : les effectifs des deux échantillons.

Conditions d'application

Les effectifs de chaque échantillon doivent être supérieurs ou égaux à 30.
Lorsque cette condition n'est pas remplie, il faut utiliser le test T de Student.

Principe du test (*cf.* principes généraux, chap. 11.I)

Si l'hypothèse nulle est vraie,

- m_1 et m_2 sont deux valeurs possibles d'une variable normale centrée autour de la moyenne commune aux deux populations ;
- la différence $m_1 - m_2$ suit une loi normale de moyenne 0 ;
- le rapport de cette différence sur son écart type suit une loi de Z.

Calculs intermédiaires

Les variances des moyennes des deux populations d'où sont issus les deux échantillons peuvent être estimées par s^2_1/n_1 et s^2_2/n_2.
La variance de la différence est égale à la somme des variances.

Écart type s_d de la différence $m_1 - m_2$: $s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Test Z :

$$z = \frac{|m_1 - m_2|}{s_d}$$

Résultat

H_1	z	REJET H_0	INTERPRÉTATION
bilatérale	< 1,96	Non	m_1 n'est pas significativement différente de m_2
	≥ 1,96	Oui	m_1 diffère significativement de m_2
unilatérale	< 1,65	Non	m_1 n'est pas significativement supérieure (ou <) à m_2
	≥ 1,65	Oui	m_1 est significativement supérieure (ou inférieure) à m_2

Exemple 13.4.

On désire comparer la pression artérielle diastolique d'un groupe de sujets sains et d'un groupe de sujets atteints de drépanocytose (hémoglobinopathie SS). Une étude donne les résultats suivants.

	Effectif n	Pression artérielle diastolique moyenne (mmHg)	Variance s^2
sujets sains	88	70,1	10,8
sujets drépanocytaires	85	61,8	6,9

H_0 : les pressions artérielles sont identiques.
H_1 bilatérale : la pression artérielle est différente chez les sujets drépanocytaires.

$$s_d = \sqrt{\frac{10,8}{88} + \frac{6,9}{85}} = 0,45 \qquad z = \frac{|70,1 - 61,8|}{0,45} = 18,4$$ On rejette H_0.

Z est encore inférieur à 18,4 pour un risque $\alpha = 0,00001$.

Il existe donc une différence significative de la pression diastolique moyenne entre les 2 groupes étudiés. La pression diastolique est significativement plus basse chez les sujets atteints de drépanocytose ($p < 10^{-5}$).

V. TEST Z POUR COMPARER DEUX MOYENNES SUR DEUX SÉRIES APPARIÉES

Quand choisir ce test ?

Lorsqu'on veut comparer deux séries d'une variable quantitative provenant d'échantillons de même taille et lorsque chaque observation d'un échantillon est liée à une observation homologue de l'autre échantillon. Chaque couple de valeur constitue une paire.

Ce type de test est particulièrement adapté lorsqu'on désire comparer deux valeurs de même type observées chez un même individu. Dans ce cas, il n'y a qu'un seul échantillon, mais deux séries de valeurs observées.

Variables	quantitatives
Paramètre étudié	moyenne des différences entre sujets appariés
Taille des échantillons	supérieure ou égale à 30
Séries comparées	appariées
Hypothèse nulle	$m_d = 0$
H_1 bilatérale	$m_d \neq 0$
H_1 unilatérale	$m_d > 0$ ou bien $m_d < 0$

Formulations

$\mathbf{x_i}$ et $\mathbf{y_i}$: valeurs observées dans chaque série.
$\mathbf{d_i}$: différence observée entre deux valeurs appariées.
$\mathbf{s_d^2}$: variance des différences.
$\mathbf{m_d}$: moyenne des différences entre sujets appariés.
$\mathbf{s_{md}}$: écart type de la moyenne des différences.
n : nombre de couples appariés.

Conditions d'application

Le nombre de paires doit être supérieur ou égal à 30. Si cette condition n'est pas remplie, il faut utiliser le test T de Student pour séries appariées.

Principe du test (*cf.* principes généraux, chap. 11.I)

On teste l'hypothèse que les différences individuelles entre sujets appariés sont nulles. Si les conditions d'application sont vérifiées, la moyenne des différences suit une loi normale de moyenne 0. Le rapport de cette différence sur son écart type suit une loi de Z normale centrée réduite de moyenne 0 et d'écart type 1.

L'intérêt d'un test apparié est d'éliminer la variabilité entre individus de la même série. On ne prend en compte que la variabilité des différences entre paires. Un test apparié est donc plus puissant qu'un simple test de comparaison de deux moyennes.

Calculs intermédiaires

Différence entre paires : $d_i = x_i - y_i$. Moyenne des différences : $m_d = \dfrac{\Sigma d_i}{n}$

Variance des différences : $s_d^2 = \dfrac{\Sigma d_i^2 - \dfrac{(\Sigma d_i)^2}{n}}{n-1}$

Écart type de la moyenne des différences : $s_{md} = \sqrt{\frac{s_d^2}{n}}$

Test Z :

$$z = \frac{|m_d - 0|}{s_{md}}$$

Résultat

H_1	z	REJET H_0	INTERPRÉTATION
bilatérale	< 1,96	Non	les moyennes des séries ne diffèrent pas significativement
	≥ 1,96	Oui	les moyennes des 2 séries diffèrent significativement
unilatérale	< 1,65	Non	les moyennes des 2 séries ne diffèrent pas
	≥ 1,65	Oui	la moyenne d'une des séries est significativement supérieure (ou inférieure) à l'autre

Exemple 13.5.

On désire étudier le volume globulaire moyen (VGM exprimé en μ^3) chez les ouvriers embauchés dans une entreprise de produits chimiques. On dose le VGM chez 30 sujets avant embauche (série A) et 3 mois après la prise de poste (série B).

A	94,4	92,4	95,5	97,6	97,4	98,5	90,6	94,5	97,2	92,8	93,6	91,6	91,3	93,0	93,5
B	94,4	89,4	95,5	98,1	97,4	93,7	90,0	92,5	100,3	92,3	92,6	91,1	90,1	89,3	89,7
A	90,8	93,9	94,2	95,3	94,3	94,3	94,3	97,7	94,8	94,6	94,0	98,9	96,7	99,8	93,7
B	89,4	94,1	92,5	92,2	93,4	90,0	93,8	97,1	96,1	93,7	90,8	98,2	96,8	98,9	93,6

On pose : H_0 : les VGM sont identiques avant et après embauche ;
H_1 bilatérale : les VGM sont différents.

Les deux séries sont appariées puisque les mesures sont effectuées sur les mêmes individus. On calcule les différences entre individus $d_i = A - B$ et leurs carrés (d_i^2) :

d_i	0	3	0	-0,5	0	4,8	0,6	2	-3,1	0,5	1	0,5	1,2	3,7	3,8
d_i^2	0	9	0	0,25	0	23,04	0,36	4	9,61	0,25	1	0,25	1,44	13,69	14,44
d_i	1,4	-0,2	1,7	3,1	0,9	4,3	0,5	0,6	-1,3	0,9	3,2	0,7	-0,1	0,9	0,1
d_i^2	1,96	0,04	2,89	9,61	0,81	18,49	0,25	0,36	1,69	0,81	10,24	0,49	0,01	0,81	0,01

$\Sigma d_i = 34,2$ $\Sigma d_i^2 = 125,8$ $m_d = 34,2/30 = 1,14$

$s_d^2 = (125,8 - 34,2^2/30)/29 = 2,994$ $s_{md} = \sqrt{2,994/30} = 0,316$ $z = 1,14/0,316 = 3,61$

Cette valeur est supérieure à 1,96. On rejette donc H_0.
z est encore supérieur à la valeur de $Z_{0,001} = 3,29$.

On conclut à une différence significative des VGM avant et trois mois après embauche avec $p < 0,001$. On constate qu'en moyenne les VGM avant embauche étaient de $m_A = 94,7\ \mu^3$ et trois mois après embauche $m_B = 93,6\ \mu^3$.

Cette différence semble faible, mais elle est hautement significative. On peut vérifier que si on avait utilisé le test Z de comparaison de 2 moyennes sans tenir compte de l'appariement, on aurait abouti à une valeur $Z = (94,71 - 93,57)/0,723 = 1,58$. On n'aurait donc pas rejeté H_0.
On constate donc la plus grande puissance d'un test apparié qui gomme la variabilité entre individus.

VI. TEST T POUR COMPARER UNE MOYENNE OBSERVÉE À UNE MOYENNE THÉORIQUE

Quand choisir ce test ?

Lorsqu'on désire comparer une moyenne observée à une moyenne connue dans une population de référence. Ce test, robuste, peut être utilisé à la place du test Z (*cf.* III) lorsque l'échantillon est trop petit ($n < 30$).

Variable	quantitative
Paramètre étudié	moyenne
Taille de l'échantillon	indifférente
Hypothèse nulle	$M = \mu$
H_1 bilatérale	$M \neq \mu$
H_1 unilatérale	$M > \mu$ ou bien $M < \mu$

Formulations

μ : la moyenne théorique connue de la population de référence.
M : la moyenne inconnue de la population d'où est issu l'échantillon.
m : la moyenne observée dans un échantillon.
s : l'écart type de l'échantillon.
n : l'effectif de l'échantillon.
ddl : nombre de degré de liberté.

Conditions d'application

La distribution de la variable doit être supposée normale dans la population d'où est issu l'échantillon.

Principe du test (*cf.* principes généraux, chap. 11.II)

Si l'hypothèse nulle est vraie, le rapport de la différence $|m - \mu|$ sur l'écart type de μ suit une loi T de Student à $n - 1$ ddl.

Calcul intermédiaire

L'écart type de la moyenne μ peut être estimé par l'écart type de la moyenne de l'échantillon : $\frac{s}{\sqrt{n}}$

Test T de Student :

$$t = \frac{|m - \mu|}{\frac{s}{\sqrt{n}}} \qquad ddl = n - 1$$

Résultats

H_1	t	REJET H_0	INTERPRÉTATION
bilatérale	$< T_{5\%}$	Non	m n'est pas significativement différent de μ
	$\geq T_{5\%}$	Oui	m diffère significativement de μ
unilatérale	$< T_{10\%}$	Non	m n'est pas significativement supérieure (ou <) à μ
	$\geq T_{10\%}$	Oui	m est significativement supérieure (ou inférieure) à μ

Exemple 13.6.

Dans un échantillon de 18 sujets suspects d'être atteints de trypanosomiase (maladie du sommeil), on a mesuré la quantité de protéines dans le liquide céphalorachidien (protéinorachie). On trouve dans ce groupe une protéinorachie moyenne de 460 mg/L avec un écart type de 280 mg/L. Dans la population générale, la protéinorachie est en moyenne de 300 mg/L. On se demande si ce groupe de sujet présente une protéinorachie différente de la normale.

H_0 : la protéinorachie observée ne diffère pas de la moyenne générale.

H_1 bilatérale : la protéinorachie observée est différente de celle de la population.

Condition d'application : on suppose que la protéinorachie est distribuée normalement dans la population.

$$t = (460 - 300)/(280/\sqrt{18}) = 2{,}4$$

Dans la table T de Student, pour ddl = 18 – 1 = 17, on observe la valeur 2,11 correspondant à un risque α de 5 %. On rejette donc l'hypothèse nulle. La valeur 2,4 est encore inférieure à $t_{3\,\%}$ (2,368).

On conclut que la protéinorachie des sujets étudiés est différente de la protéinorachie normale ($p < 0{,}03$).

Remarque : dans cet exemple précis, sachant que la maladie entraîne une protéinorachie élevée, on aurait pu poser une hypothèse alternative *unilatérale* : les sujets suspects de la maladie présentent une protéinorachie *plus élevée* que la normale. On aurait alors rejeté H_0 (avec un risque de 5 %) pour une valeur de t supérieure à $T_{10\,\%}$, c'est-à-dire supérieure à 1,74. On aurait conclu avec $p < 0{,}015$.

VII. TEST T DE STUDENT POUR COMPARER DEUX MOYENNES

Quand choisir ce test ?

Lorsqu'on veut comparer les moyennes observées dans deux échantillons. Ce test peut être utilisé à la place du test Z (*cf.* IV) lorsqu'au moins un des 2 échantillons est trop petit ($n < 30$).

Variables	quantitatives
Paramètre étudié	moyennes
Taille des échantillons	indifférente
Séries comparées	indépendantes
Hypothèse nulle	$\mu_1 = \mu_2$
H_1 bilatérale	$\mu_1 \neq \mu_2$
H_1 unilatérale	$\mu_1 > \mu_2$ ou bien $\mu_1 < \mu_2$

Formulations

μ_1 et μ_2 : les moyennes inconnues des deux populations d'où sont issus les échantillons.
m_1 et m_2 : les moyennes des deux échantillons.
s_1 et s_2 : les écarts types des deux échantillons
n_1 et n_2 : les effectifs des deux échantillons.
ddl : nombre de degrés de liberté.

Conditions d'application

Les distributions des populations d'où sont issus les échantillons doivent être normales ou tout au moins « à peu près » normales, car le test est « robuste ».
Les variances des deux populations d'où sont issus les échantillons doivent être supposées égales (leur rapport ne devrait pas dépasser 3). Sinon, la taille des échantillons doit être équivalente.

Principe du test (*cf.* principes généraux, chap. 11.II)

Si l'hypothèse nulle est vraie, le rapport de la différence $\mu_1 - \mu_2$ sur son écart type suit une loi T de Student lorsque les effectifs sont faibles.

Calcul intermédiaire

Estimation de la variance commune aux deux échantillons : $s^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

On estime l'écart type s_d de la différence $\mu_1 - \mu_2$ par : $s_d = \sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}$

Test T de Student : $t = \dfrac{|m_1 - m_2|}{s_d}$ $\quad ddl = n_1 + n_2 - 2$

Résultats

H_1	t	REJET H_0	INTERPRÉTATION
bilatérale	$< T_{5\%}$	Non	m_1 n'est pas significativement différent de m_2
	$\geq T_{5\%}$	Oui	m_1 diffère significativement de m_2
unilatérale	$< T_{10\%}$	Non	m_1 n'est pas significativement supérieur ou inférieur à m_2
	$\geq T_{10\%}$	Oui	m_1 est significativement supérieur ou inférieur à m_2

Exemple 13.7.

On a mesuré un marqueur biologique chez 2 séries de sujets, l'une composée de sujets sains, l'autre de sujets atteints d'hépatite alcoolique. L'étude a retrouvé les résultats suivants :

	Effectif (n)	Moyenne du marqueur (g/L)	Écart type
sujets sains	15	1,6	0,19
sujets alcooliques	12	1,4	0,21

H_0 : la valeur moyenne du marqueur est identique dans les 2 populations.
H_1 bilatérale : la valeur moyenne du marqueur est différente chez les sujets atteints d'hépatite alcoolique.
Condition d'application : on suppose que le marqueur se distribue normalement dans les 2 populations.

$$s^2 = \frac{(15-1)0{,}19^2 + (12-1)0{,}21^2}{15+12-2} = 0{,}04 \qquad s_d = \sqrt{\frac{0{,}04}{15} + \frac{0{,}04}{12}} = 0{,}077$$

$t = (1{,}6 - 1{,}4)/0{,}077 = 2{,}60$ ddl = 15 + 12 − 2 = 25

Pour ddl = 25, $T_{5\%} = 2{,}06$. On rejette donc H_0. On constate que pour un risque $\alpha = 0{,}02$, $T_{2\%} = 2{,}485$ est encore inférieure à la valeur t observée.

Conclusion : les malades atteints d'hépatite alcoolique présente une valeur du marqueur significativement différente de celle des sujets sains ($p < 2$ %).

Excel® : fonction TEST.STUDENT. Dans la boite de dialogue, entrer dans la case « type » la valeur 2 si les deux variances sont supposées équivalentes (homoscédastiques), ou la valeur 3 si les 2 variances sont très inégales (hétéroscédastiques).
Dans l'exemple 13.7, TEST.STUDENT (série 1 ; série 2 ; 2 ; 2) donne directement la valeur de p.

VIII. TEST T POUR COMPARER 2 MOYENNES SUR 2 SÉRIES APPARIÉES

Quand choisir ce test ?

Lorsqu'on veut comparer deux séries d'une variable quantitative provenant de deux échantillons de même taille et lorsque chaque observation d'un échantillon est liée à une observation homologue de l'autre échantillon. Chaque couple de valeur constitue une paire.

Ce type de test est particulièrement adapté lorsqu'on désire comparer deux valeurs de même type observées chez un même individu. Dans ce cas, il n'y a qu'un seul échantillon, mais deux séries de valeurs observées. Ce test peut être utilisé à la place du test Z (*cf.* IV) lorsque la taille de chaque série est trop petite ($n < 30$).

Variables	quantitatives
Paramètre étudié	moyenne des différences entre sujets appariés
Nombre d'échantillons	deux ou un seul avec deux séries
Taille des échantillons	indifférente
Séries comparées	appariées
Hypothèse nulle	$m_d = 0$
H_1 bilatérale	$m_d \neq 0$
H_1 unilatérale	$m_d > 0$ ou bien $m_d < 0$

Formulations

$\mathbf{x_i}$ et $\mathbf{y_i}$: valeurs observées dans chaque série.
$\mathbf{d_i}$: différence observée entre deux sujets appariés.
$\mathbf{s^2_d}$: variance des différences.
$\mathbf{m_d}$: moyenne des différences entre sujets appariés.
$\mathbf{s_{md}}$: écart type de la moyenne des différences.
$\mathbf{n}$: nombre de couples appariés.

Conditions d'application

Les différences doivent être distribuées de façon normale.

Principe du test (*cf.* principes généraux, chap. 11.II)

On teste l'hypothèse que les différences individuelles entre individus appariés sont nulles. Si les conditions d'application sont vérifiées, la moyenne des différences divisée par son écart type suit une loi T de Student à $n - 1$ ddl.

L'intérêt d'un test apparié est d'éliminer la variabilité entre individus de la même série. On ne prend en compte que la variabilité des différences entre paires. Un test apparié est donc plus puissant qu'un simple test de comparaison de deux moyennes.

Calculs intermédiaires

Différence entre paires : $d_i = x_i - y_i$

Moyenne des différences : $m_d = \dfrac{\Sigma d_i}{n}$

Variance des différences : $s_d^2 = \dfrac{\Sigma d_i^2 - \dfrac{(\Sigma d_i)^2}{n}}{n-1}$

Écart type de la moyenne des différences : $s_{md} = \sqrt{\dfrac{s_d^2}{n}}$

Test T :

$$t = \frac{|m_d - 0|}{s_{md}} \qquad ddl = n - 1$$

Résultats

H_1	t	REJET H_0	INTERPRÉTATION
bilatérale	$< T_{5\%}$	Non	les moyennes des séries ne diffèrent pas significativement
	$\geq T_{5\%}$	Oui	les moyennes des 2 séries diffèrent significativement
unilatérale	$< T_{10\%}$	Non	les moyennes des 2 séries ne diffèrent pas
	$\geq T_{10\%}$	Oui	la moyenne d'une des séries est significativement supérieure ou inférieure à l'autre

Exemple 13.8.

On désire étudier l'effet d'une nouvelle stratégie de traitement du diabète en mesurant l'effet sur la glycémie. On dose la glycémie (en g/L) chez 15 sujets avant le début du nouveau protocole (série A) et 3 mois après (série B).

A	2,47	3,09	2,14	2,47	3,06	2,72	2,29	1,90	2,34	2,75	2,67	2,80	2,51	2,23	2,20
B	2,30	2,96	2,23	2,34	2,84	2,59	2,15	1,88	2,32	2,65	2,68	2,58	2,43	2,02	2,17

On pose H_0 : les glycémies sont identiques avant et après le nouveau protocole
H_1 unilatérale : la glycémie est abaissée grâce au nouveau protocole

Les deux séries sont appariées puisque les mesures sont effectuées sur les mêmes individus. On calcule les différences entre individus :

d_i	0,17	0,13	-0,09	0,13	0,22	0,13	0,14	0,02	0,02	0,1	-0,01	0,22	0,08	0,21	0,03

$\Sigma d_i = 1,5$ $\qquad \Sigma d_i^2 = 0,266$ $\qquad m_d = 1,5/15 = 0,10$

$s_d^2 = (0,266 - 1,5^2/15)/14 = 0,0083$ $\qquad s_{md} = \sqrt{0,0083/15} = 0,0235$

L'effectif de l'échantillon étant inférieur à 30 on utilise un test T en supposant que les différences sont distribuées normalement.

$t = 0,10/0,0235 = 4,25$ $\qquad ddl = 15 - 1 = 14$

Pour ddl = 14, $T_{10\%} = 1,761$. On rejette donc H_0.
t est supérieur à la valeur de $T_{0,1\%} = 4,14$.

On conclut que la glycémie est abaissée significativement après administration de la nouvelle stratégie avec $p < 0,0005$ (H_1 unilatérale). On constate qu'en moyenne les glycémies avant protocole étaient de $m_A = 2,51$ g/L et 3 mois après le début du protocole $m_B = 2,41$ g/L. Le test apparié permet de détecter une faible différence, qui n'aurait pas été détectée par un test T de comparaison des 2 moyennes (t = (2,51-2,41)/0,118 = 0,85).
La conclusion générale de cette étude, montre un résultat peu satisfaisant d'un point de vue thérapeutique (baisse moyenne de la glycémie de 0,1 g/L). Néanmoins cette différence est hautement significative d'un point de vue statistique.

Excel® : fonction TEST.STUDENT fournit directement la valeur de p.
Dans la boîte de dialogue, entrer dans la case « type » la valeur 1 (série appariée).
Dans l'exemple 13.8, TEST.STUDENT (série1 ; série2 ; 1 ; 1) = *0,0004*.

IX. TEST F POUR COMPARER DEUX VARIANCES

Test de Fisher-Snedecor.

Quand choisir ce test ?

Lorsqu'on veut comparer les variances de 2 séries de variables quantitatives.
Lorsqu'on veut vérifier les conditions d'application de certains tests paramétriques qui exigent que les variances soient identiques (test T de Student).

Variables	quantitatives
Paramètres étudiés	variances
Taille des échantillons	indifférente
Séries étudiées	indépendantes
Hypothèse nulle	$\sigma_1^2 = \sigma_2^2$
H_1 bilatérale	$\sigma_1^2 \neq \sigma_2^2$
H_1 unilatérale	$\sigma_1^2 > \sigma_2^2$

Formulations

$\boldsymbol{\sigma_1^2}$ et $\boldsymbol{\sigma_2^2}$: les variances inconnues des deux populations d'où sont issus les échantillons.
$\mathbf{s_1^2}$ et $\mathbf{s_2^2}$: les variances des deux échantillons à comparer.
$\mathbf{n_1}$ et $\mathbf{n_2}$: les effectifs de chaque échantillon.
$\mathbf{k_1}$ et $\mathbf{k_2}$: degrés de libertés pour chaque échantillon.

Conditions d'application

Les distributions doivent être normales dans les deux populations d'où proviennent les deux échantillons.

Principe du test (*cf.* principes généraux, chap. 11.III)

On teste le rapport des deux variances s_1^2 et s_2^2 en nommant s_1^2 la variance la plus élevée. Sous l'hypothèse nulle, ce rapport est peu différent de 1 et les fluctuations d'échantillonnage suivent une loi dite de Fisher.

Test F :

$$F = \frac{s_1^2}{s_2^2}$$

$k_1 = n_1 - 1$ et $k_2 = n_2 - 1$

Résultats

H_1	F	REJET H_0	INTERPRÉTATION
bilatérale	$< F_{2,5\,\%}$	Non	s_1^2 ne diffère pas significativement de s_2^2
	$\geq F_{2,5\,\%}$	Oui	s_1^2 diffère significativement de s_2^2
unilatérale	$< F_{5\,\%}$	Non	s_1^2 ne diffère pas significativement de s_2^2
	$\geq F_{5\,\%}$	Oui	s_1^2 est significativement supérieure à s_2^2

Exemple 13.9.

On désire comparer la pression artérielle diastolique (PAD) d'un groupe de sujets sains (m = 70,1) et d'un groupe de sujets atteints de drépanocytose (m = 61,8). On ne dispose que de 20 individus par groupe. La variance de la PAD est respectivement de 116,7 et de 47,6.

En raison de la faiblesse de l'effectif, il faut réaliser un test T en supposant que les variances sont identiques dans les populations des 2 groupes. On commence donc par comparer les 2 variances des échantillons par un test F, en posant :

- H_0 : les deux variances ne sont pas différentes ;
- H_1 bilatérale : les deux variances diffèrent.

$F = 116,7/47,6 = 2,45$ avec $k_1 = k_2 = 20 - 1 = 19$

Comme le test est bilatéral, on regarde dans la table $F_{2,5\ \%}$. (*cf.* chap. 11, III.2 et 3).

Excel® : fonction TEST.F : fournit directement la valeur *p* pour un test bilatéral.
Dans l'exemple 13.9, TEST.F (série sujets sains ; série drépanocytaires) = 0,065.

La table de $F_{2,5\ \%}$ ne donne pas la valeur seuil pour des ddl égaux à 19, mais pour des ddl égaux à 20, $F_{2,5\ \%} = 2,46$. La valeur trouvée F est donc inférieure à la valeur seuil. On ne rejette pas H_0.

Le fait de ne pas rejeter H_0 ne signifie pas pour autant que les deux variances soient identiques dans les 2 populations. On se contente de constater qu'il n'y a pas d'arguments pour déclarer que les variances sont différentes. On peut donc raisonnablement comparer les deux moyennes de la PAD par un test T en supposant que la condition d'égalité des variances est remplie dans les deux populations.

X. TEST F POUR COMPARER PLUSIEURS MOYENNES

Test de Fisher-Snedecor, analyse de la variance, ANOVA.

Quand choisir ce test ?

Lorsqu'on désire comparer les moyennes observées sur plusieurs échantillons.

Variables	quantitatives
Paramètres étudiés	variances
Séries comparées	indépendantes
Hypothèse nulle	$s_g^2 = s_r^2$: les moyennes sont identiques
H_1 unilatérale	$s_g^2 > s_r^2$: les moyennes sont différentes

Formulations

s_g^2 : la variance entre séries étudiées.
s_r^2 : la variance résiduelle ou variance entre individus de chaque série.
k_1 et k_2 : degrés de liberté.

Conditions d'application

Les distributions des populations d'où proviennent les échantillons doivent être normales et de même variance.

Principe du test (*cf.* principes généraux, chap. 11.III.3)

On teste le rapport de la variance entre séries et de la variance résiduelle (ou variance entre individus indépendamment de la série). Sous l'hypothèse nulle, les individus proviennent de la même population. La variabilité est donc identique entre séries et entre individus de chaque série. Sous H_0, le rapport de s_g^2 et de s_r^2 est donc proche de 1 et suit une loi F de Fisher-Snedecor.

Calculs intermédiaires

n_i : les effectifs de chaque série.
N : nombre total des observations.
c : nombre de séries à comparer.
T_i : totaux des observations de chaque série.
T_g : Total général des observations.
Σx^2 : Total des carrés des observations.

VARIANCE	NUMÉRATEUR (NUM)	DÉNOMINATEUR (DDL)	VARIANCE = NUM/DDL
entre séries	$\Sigma(T_i^2/n_i) - T_g^2/N$	$c - 1$	s_g^2
résiduelle	$\Sigma x^2 - \Sigma(T_i^2/n_i)$	$N - c$	s_r^2

Test F :

$$F = \frac{s_g^2}{s_r^2}$$

$k_1 = c - 1$ et $k_2 = N - c$

Résultats

H_1	F	REJET H_0	INTERPRÉTATION
unilatérale	$< F_{5\%}$	Non	les moyennes ne diffèrent pas significativement
	$\geq F_{5\%}$	Oui	les moyennes diffèrent significativement

Si H_0 a été rejeté, on peut comparer les moyennes deux à deux par un test T en utilisant la variance résiduelle comme variance commune.

Exemple 13.10.

On désire comparer un indicateur mesuré en g/L entre trois groupes de patients atteints de 3 formes cliniques d'une maladie notées A, B, C.

A	19,5	38,0	33,2	52,2	77,4	92,6	77,4	48,3	95,6	51,5	34,3		
B	44,1	29,5	68,8	34,9	51,2	42,5	32,0	92,2	84,5	61,2	65,8	53,1	87,2
C	76,6	88,3	83,4	88,2	87,3	96,0	85,6	100,2	84,8	93,5	134,6	124,7	

- H_0 : les trois moyennes sont identiques.
- H_1 : les trois moyennes sont différentes.

	n_i	T_i	m_i	Σx^2	T_i^2	T_i^2/n_i	T_0^2/N
A	11	620,0	56,36	41 509	384 400	34 945	
B	13	747,0	57,46	48 358	558 009	42 924	
C	12	1 143,2	95,27	112 204	1 306 906	108 909	
Σ	36	2 510,2		202 071		186 778	175 031

Variance	Numérateur	Dénominateur	Variance	F
entre séries	186 778 – 175 031 = 11 747	3 –1 = 2	5 873,5	12,67
résiduelle	202 071 – 186 778 = 15 293	36 – 3 = 33	463,4	

$F = 5\,873{,}5/463{,}4 = 12{,}67$ $k_1 = 2$ $k_2 = 33$

On ne dispose pas dans les tables des valeurs de F pour $k_2 = 33$. Mais on constate que pour $k_2 = 30$, la valeur F trouvée est encore supérieure à $F_{0,001}$ (8,77). On rejette donc H_0.

On conclut que les trois moyennes diffèrent significativement ($p < 0{,}001$).
Les comparaisons 2 à 2 en utilisant la variance résiduelle 463,4 comme variance commune donnent :
- A *versus* C : $s_d = 8{,}99$; ddl = 21; t = 4,33. On rejette H_0 avec $p < 0{,}001$;
- B *versus* C : $s_d = 8{,}62$; ddl = 23; t = 4,39. On rejette H_0 avec $p < 0{,}001$;
- A *versus* B : $s_d = 8{,}82$; ddl = 22; t = 0,12. On ne rejette pas H_0.

On conclut donc en affirmant que l'indicateur diffère significativement dans la forme clinique C. En revanche, il n'y a pas de différence significative entre les formes A et B.

Excel® : Outils/Utilitaire d'analyse/Analyse de variance à un facteur.
Sélectionner les plages des valeurs. On obtient le tableau complet d'analyse avec les valeurs de F et *p*.
Dans l'exemple 13.10, *p = 0,00008*.

XI. TEST DE WILCOXON

Quand choisir ce test ?

Lorsqu'on veut comparer deux séries d'une variable quantitative.

Le test de Wilcoxon est un test non paramétrique qui s'intéresse non pas aux valeurs des variables comme les tests Z ou T, mais aux rangs des valeurs après les avoir ordonné (*cf.* chap. 11.VI). Sa difficulté résidait auparavant dans la nécessité de classer les individus lorsque les séries étaient de grande taille. Depuis la généralisation des ordinateurs et des tableurs modernes, cette opération est quasi instantanée.

Test équivalent : tests paramétriques Z ou test T pour comparer deux moyennes.

Variables	quantitatives
Grandeur étudiée	somme des rangs d'une série
Taille des échantillons	au moins 10
Séries comparées	indépendantes
Hypothèse nulle	distributions superposées
H_1 bilatérale	distributions décalées
H_1 unilatérale	distributions décalées dans un sens ou dans l'autre

Formulations

$\mathbf{n_1}$ et $\mathbf{n_2}$: les effectifs des deux échantillons.

$\mathbf{N} = n_1 + n_2$.

Conditions d'application

Les effectifs des deux échantillons doivent être supérieurs à 10 et ne pas comporter trop de valeurs *ex æquo*.

Le test ne nécessite aucune condition sur la distribution des valeurs dans la population. Notamment, dans le cas de séries de petite taille, il n'exige pas une distribution normale.

Principe du test

On compare les rangs des observations classées selon leurs valeurs. Sous H_0, les valeurs sont mélangées de façon homogène et on démontre que la somme attendue des rangs de chaque série est respectivement $n_1(N + 1)/2$ et $n_2(N + 1)/2$. Sous H_0, la différence entre la somme des rangs d'une des séries et sa valeur attendue fluctue autour de zéro. Le rapport de cette différence sur son écart type suit une loi de Z normale centrée réduite.

Calculs intermédiaires

- Il faut classer toutes les observations des deux séries selon leurs valeurs de la première à la Nième et numéroter leur ordre. Cela définit le rang de chaque observation. Lorsque deux valeurs sont identiques, on calcule leur rang moyen *ex æquo*.
- Somme des rangs d'une des séries : $\mathbf{w_1}$
- Somme attendue : $\mathbf{w_a} = n_1(N + 1)/2$
- Variance de w_1 : $\mathbf{s_{w1}^2} = n_1 n_2(N + 1)/12$

S' il existe de nombreux *ex aequo*, on devra utiliser une formule de variance corrigée.

Test W :

$$z = \frac{|w_1 - w_a|}{\sqrt{s_{w1}^2}}$$

Résultats

H_1	z	REJET H_0	INTERPRÉTATION
bilatérale	< 1,96	Non	les distributions ne sont pas significativement décalées
	≥ 1,96	Oui	les distributions sont décalées
unilatérale	< 1,65	Non	les distributions ne sont pas décalées dans un sens donné
	≥ 1,65	Oui	les distributions sont décalées dans un sens donné

Exemple 13.11.

On veut comparer deux séries de patients dont on a étudié la cytorachie (présence de cellules dans le liquide cérébrospinal). La cytorachie est mesurée en nombre de cellule par µL (c/µL). Les valeurs des 2 séries sont présentées ici de façon ordonnée par valeur croissante.

• Série 1 : n = 18

c/µL	8	9	18	21	26	34	41	45	49	84	85	154	160	173	348	480	560	612

• Série 2 : n = 20

c/µL	5	6	7	8	10	10	11	14	16	17	19	20	22	23	27	35	40	92	100	200

On constate que la distribution de ces deux séries de faible taille est loin d'être normale. Un test non paramétrique est donc adapté. On mélange les 2 séries ordonnées et on numérote les rangs des valeurs.

c/µL	5	6	7	8	8	9	10	10	11	14	16	17	18	19	20	21	22	23	26
rang	*1*	*2*	*3*	*4,5*	*4,5*	*6*	*7,5*	*7,5*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*

c/µL	27	34	35	40	41	45	49	84	85	92	100	154	160	173	200	348	480	560	612
rang	*20*	*21*	*22*	*23*	*24*	*25*	*26*	*27*	*28*	*29*	*30*	*31*	*32*	*33*	*34*	*35*	*36*	*37*	*38*

Lorsque des valeurs sont identiques, on leur attribue des rangs *ex aequo* en calculant la moyenne des rangs qu'elles occupent : ainsi les valeurs 10 qui occupent respectivement les cases 7 et 8 auront pour rang (7 + 8)/2 = 7,5. La valeur suivante 11 prend son rang normal à la 9e place.

• H_0 : Les distributions des deux séries sont superposées.
• H_1 bilatérale : les distributions sont décalées.

On calcule la somme des rangs de la série 1 (la plus courte) :

$W_1 = 4,5 + 6 + 13 + 16 + 19 + 21 + 24 + 25 + 26 + 27 + 28 + 31 + 32 + 33 + 35 + 36 + 37 + 38 = 451,5$

$W_0 = 18 \times (38 + 1)/2 = 351$

$s^2_{w1} = 18 \times 20 \times (38 + 1)/12 = 1\,170$ $\qquad s_{w1} = \sqrt{1\,170} = 34,2$

$z = (451,5 - 351)/34,2 = 2,94$ (> 1,96). On rejette donc H_0 avec $p < 0,01$

On en conclut qu'en moyenne les valeurs observées dans la série de patients n° 1 sont plus élevées que dans la série n° 2 avec $p < 1\%$.

Excel® : utiliser la fonction RANG en sélectionnant les 2 séries pour obtenir le rang de chaque valeur. Calculer les sommes W_1 et W_0 et appliquer ensuite les formules.

XII. TEST DE WILCOXON POUR SÉRIES APPARIÉES

Quand choisir ce test ?

Lorsqu'on veut comparer deux séries d'une variable quantitative et lorsque chaque observation d'un échantillon est liée à une observation homologue de l'autre échantillon (paires).

Le test de Wilcoxon pour séries appariées est un test non paramétrique qui s'intéresse non pas aux valeurs des différences comme les tests Z ou T appariés, mais aux rangs des différences après les avoir ordonnées.

Test équivalent : tests paramétriques Z ou test T pour comparer deux moyennes sur deux séries appariées.

Variables	quantitatives
Grandeur étudiée	somme des rangs des différences
Taille des échantillons	au moins 20 paires à différence non nulle
Séries comparées	appariées
Hypothèse nulle	distribution des différences centrée autour de 0
H_1 bilatérale	distribution des différences décalée par rapport à 0
H_1 unilatérale	distribution des différences décalée positivement ou négativement

Formulations **n** : nombre de couples appariés dont la différence n'est pas nulle.
$\mathbf{w_p}$ et $\mathbf{w_n}$: somme des rangs des différences positives et négatives.

Conditions d'application

Le nombre de paires de différence non nulle doit être supérieur à 20 et ne pas comporter trop d'ex-æquo. Le test ne nécessite aucune condition sur la distribution des valeurs des différences. Notamment, dans le cas de séries de petite taille, il n'exige pas une distribution normale.

Principe du test

On compare les rangs des différences non nulles classées selon leurs valeurs absolues. Sous H_0, la somme des rangs des différences négatives doit être égale à la somme des rangs des différences positives (valeur attendue $n(n+1)/4$). Sous H_0, la différence entre la somme des rangs des différences positives (ou négatives) et sa valeur attendue fluctue autour de zéro. Le rapport de cette différence sur son écart type suit une loi de Z normale centrée réduite.

Calculs intermédiaires

- Calculer les différences entre les deux séries.
- Classer toutes les différences non nulles selon leurs valeurs absolues, de la première à la Nième et numéroter leur ordre. Cela définit le rang de chaque différence absolue. Lorsque deux valeurs sont identiques, on calcule leur rang moyen.
- Somme des rangs des différences positives : w_p
- Somme attendue : $w_a = n(n+1)/4$
- Variance de w_p : $s^2_{wp} = n(n+1)(2n+1)/24$

Test W :

$$z = \frac{|w_p - w_u|}{\sqrt{s^2_{wp}}}$$

Résultats

H_1	z	REJET H_0	INTERPRÉTATION
bilatérale	< 1,96	Non	les distributions ne sont pas significativement décalées
	≥ 1,96	Oui	les distributions sont décalées
unilatérale	< 1,65	Non	les distributions ne sont pas décalées dans un sens donné
	≥ 1,65	Oui	les distributions sont décalées dans un sens donné

Exemple 13.12.

On veut comparer deux séries de patients atteints de maladie du sommeil (trypanosomiase). Dans la série T^+, l'agent pathogène a été détecté ; dans la série T^-, l'agent n'a pas été détecté. Dans les 2 séries, on a étudié un indicateur (dosé en mg/L). Les patients des 2 séries sont appariés selon des critères cliniques de sévérité de la maladie : chaque patient d'une série correspond à un autre patient de l'autre série présentant les mêmes critères. On veut savoir si la valeur de l'indicateur diffère selon la présence ou l'absence du trypanosome.

T^+	88,3	44,1	286	85,6	56,4	93,6	122	53,5	78,6	318	58,7	123,7	296,2	235,1
T^-	43,7	56,8	153	34,5	41,9	90,3	34,8	67,9	231,7	301,8	287,1	45,1	185	292,6

On calcule les différences entre paires d_i :

d_i	44,6	−12,7	133	51,1	14,5	3,3	87,2	−14,4	−153,1	16,2	−228,4	78,6	111,2	−57,5

On classe les différences par ordre *croissant en valeur absolue* et on repère les rangs des valeurs positives :

di	3,3	−12,7	−14,4	14,5	16,2	44,6	51,1	−57,5	78,6	87,2	111,2	133	−153,1	−228,4
rangs	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*

La somme des rangs des différences positives (en bleu) est égale à :

$w_p = 1 + 4 + 5 + 6 + 7 + 9 + 10 + 11 + 12 = 65$

$w_u = 14\,(14 + 1)/4 = 52,5$

$s^2_{wp} = 14\,(14 + 1)\,(2 \times 14 + 1)/24 = 253,75$

$z = (65 - 52,5)/\sqrt{253,75} = 0,78$

z est inférieur à 1,96 : on ne rejette donc pas H_0.
En conclusion, la valeur moyenne de l'indicateur n'est pas différente entre le groupe T^+ et le groupe T^-.

Excel® : Éliminer les différences nulles.
Appliquer la fonction ABS sur la série des différences non nulles.
Appliquer la fonction RANG sur la série des valeurs absolues.
Éliminer les rangs des différences négatives.
Calculer la somme des rangs des différences positives et appliquer ensuite les formules.

XIII. TEST DE KRUSKAL-WALLIS (KW)

Quand choisir ce test ?

Lorsqu'on désire comparer les moyennes observées sur plusieurs échantillons. Le test de Kruskal-Wallis est un test non paramétrique qui s'intéresse non pas aux valeurs des variables, mais aux rangs des valeurs après les avoir ordonné (*cf.* chap. 11.VI).
Test équivalent : test F pour comparer plusieurs moyennes.

Variables	quantitatives
Paramètres étudiés	rangs moyens des valeurs
Taille des échantillons	supérieure à 10
Séries comparées	indépendantes
Hypothèse nulle	distributions superposées
H_1 bilatérale	distribution décalées

Formulations

$\mathbf{n_i}$: les effectifs de chaque échantillon.
$\mathbf{c}$: nombre de séries à comparer.
$\overline{W}_i$: rang moyen de chaque série.
$\mathbf{N}$: nombre total des observations.
$\overline{W}$: rang moyen général : $(N + 1)/2$.

Conditions d'application

Les effectifs de chaque échantillon doivent être supérieurs à 10 et ne pas comporter trop d'*ex æquo*.
Contrairement au test F, le test KW ne nécessite aucune condition de normalité et d'égalité de variance des distributions des valeurs de la variable.

Principe du test

Le test KW est un test de rangs qui compare les rangs moyens de chaque série. Sous l'hypothèse nulle H_0, les différences entre les rangs moyens de chacune des séries et la moyenne générale des rangs est nulle. On montre que sous H_0 la quantité calculée ci-dessous suit une loi de χ^2 à $c - 1$ ddl.

Calculs intermédiaires

On mélange toutes les valeurs des séries à comparer.
On classe ces valeurs par ordre croissant. On numérote leur ordre. Cela aboutit à donner un rang à chacune des valeurs. Si des valeurs sont *ex æquo* on leur attribue un rang moyen (*cf.* exemple).

Test de Kruskal-Wallis :

$$\chi^2 = \frac{12}{N} \frac{\Sigma n_i (\overline{W}_i - \overline{W})^2}{(N+1)} \qquad ddl = c - 1$$

Résultats

H_1	χ^2	REJET H_0	INTERPRÉTATION
unilatérale	$< \chi^2_{5\%}$	Non	les distributions ne sont pas significativement décalées
	$\geq \chi^2_{5\%}$	Oui	les distributions sont décalées

L'interprétation finale du test lorsque H_0 est rejetée, signifie que les moyennes des séries comparées diffèrent significativement entre elles.

Exemple 13.13.

On veut comparer le taux d'IgM dans le LCS entre 3 groupes de malades atteints de 3 formes cliniques de maladie du sommeil (stade 1, stade 2A et stade 2B). On obtient les résultats suivants chez 43 sujets : taux d'IgM en mg/mL

stade 1	3,3	3,6	3,9	3,9	4,7	4,8	5,8	8,9	11,8	12,0	12,1	21	31,9		
stade 2A	2,3	3,8	4,3	4,9	6,1	7,2	7,7	9	9,1	9,4	9,4	12,2	22,1	61	63
stade 2B	2,4	2,8	7	12,2	32,1	34,1	45,2	60	62,8	125	133	140	229	263	354

On constate que les variances des 3 séries sont très différentes (60, 975 et 11435). De plus, les distributions des échantillons s'écartent manifestement de la normale. Il est donc risqué de supposer que les populations d'où sont issus ces échantillons sont de variances égales et ont une distribution normale. On décide donc de ne pas faire un test F d'analyse de la variance. On choisit le test de Kruskal-Wallis. On mélange et on classe l'ensemble des valeurs en leur attribuant un rang.

mg/mL	2,3	2,4	2,8	3,3	3,6	3,8	3,9	3,9	4,3	4,7	4,8	4,9	5,8	6,1	7
rang	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
mg/mL	7,2	7,7	8,9	9	9,1	9,4	9,4	11,8	12,0	12,1	12,2	12,2	21	22,1	31,9
rang	16	17	18	19	20	21	22	23	24	25	26,5	26,5	28	29	30
mg/mL	32,1	34,1	45,2	60	61	62,8	63	125	133	140	229	263	354		
rang	31	32	33	34	35	36	37	38	39	40	41	42	43		

Lorsqu'il existe des valeurs *ex æquo*, on calcule un rang moyen pour ces valeurs (les valeurs 12,2 sont affectées du rang (26 + 27)/2 = 26,5).

On calcule le rang moyen de chaque série :

Stade 1 : $\overline{W}_1 = (4 + 5 + 7 + 8 + 10 + 11 + 13 + 18 + 23 + 24 + 25 + 28 + 30)/13 = 15,85$

Stade 2A : $\overline{W}_2 = (1 + 6 + 9 + 12 + 14 + 16 + 17 + 19 + 20 + 21 + 22 + 26,5 + 29 + 35 + 37)/15 = 18,97$

Stade 2B : $\overline{W}_3 = (2 + 3 + 15 + 26,5 + 31 + 32 + 33 + 34 + 36 + 38 + 39 + 40 + 41 + 42 + 43)/15 = 30,4$

$\overline{W} = (43 + 1)/2 = 22$ ddl = 3 – 1

Calcul du test $\chi^2 = \frac{12}{43 \times 44}[13(15,85 - 22)^2 + 15(18,97 - 22)^2 + 15(30,4 - 22)^2] = 10,7$

Pour ddl = 2, la valeur seuil de $\chi^2_{5\%}$ est de 5,99. On rejette donc H_0. La valeur immédiatement inférieure à 10,7 correspond à un risque de 0,01.

Conclusion : les distributions des taux d'IgM dans le LCS sont significativement différentes entre les 3 groupes de malades ($p < 0,01$). La moyenne du taux d'IgM chez les malades au stade 1 est de 9,8 mg/mL, au stade 2A de 15,4 mg/mL et au stade 2B elle est de 100,2 mg/mL. Le taux d'IgM apparaît donc comme nettement plus élevé chez les malades au stade 2B.

Excel® : utiliser la fonction RANG en sélectionnant les séries pour obtenir le rang de chaque valeur. Calculer les sommes des rangs W_i et appliquer ensuite les formules.

XIV. TEST DE χ^2 DE CONFORMITÉ OU D'AJUSTEMENT

Quand choisir ce test ?

Lorsqu'on désire comparer une distribution observée sur un échantillon :
- soit à une distribution connue dans une population : test de conformité ;
- soit à une distribution théorique (binomiale, normale, Poisson..) : test d'ajustement.

Variable	qualitative nominale
Paramètre étudié	effectifs observés et attendus
Taille des échantillons	effectifs attendus supérieurs ou égaux à 5
Hypothèse nulle	distribution de l'échantillon = distribution théorique
H_1 bilatérale	distribution de l'échantillon ≠ distribution théorique

Formulations

$\mathbf{o_i}$: effectifs observés dans l'échantillon.
N : total de l'échantillon.
$\mathbf{f_i}$: fréquences de chaque classe de la variable dans la population ou dans la distribution théorique.
$\mathbf{c_i}$: effectifs théoriques = $N \times f_i$.
r : nombre de lignes.
ddl : degré de liberté.

VARIABLE	DISTRIBUTION THÉORIQUE (%)	ÉCHANTILLON EFFECTIFS OBSERVÉS	ÉCHANTILLON EFFECTIFS THÉORIQUES
A_1	f_1	o_1	$c_1 = f_1 N$
...	...	...	
A_i	f_i	o_i	$c_i = f_i N$
Total	100 %	N	N

Conditions d'application

Tous les effectifs théoriques c_i doivent être supérieurs ou égaux à 5.
Si cette condition n'est pas réalisée, il faut regrouper certaines classes de la variable.

Principe du test (*cf.* principes généraux, chap. 11.IV)

Dans ce type de test de χ^2, les effectifs théoriques sont les effectifs attendus, que l'on calcule en connaissant les fréquences des classes de la variable dans la population ou selon la distribution théorique.

Calculs intermédiaires

On calcule l'effectif attendu pour chaque classe de la variable en multipliant la taille de l'échantillon par la fréquence de la classe dans la population ou par la fréquence donnée par la loi de distribution théorique. $c_i = N \times f_i$

Test du χ^2 de conformité ou d'ajustement :

$$\chi^2 = \Sigma \frac{(o_i - c_i)^2}{c_i} \qquad ddl = r - 1$$

Résultats

H_1	χ^2	REJET H_0	INTERPRÉTATION
bilatérale	$< \chi^2_{5\%}$	Non	on ne peut affirmer que la distribution étudiée est différente de la distribution théorique
	$\geq \chi^2_{5\%}$	Oui	la distribution étudiée diffère significativement de la distribution théorique

Exemple 13.14.

Sur un échantillon de 284 sujets, on a observé la structure par âge ci-dessous (o_i). On veut vérifier si cet échantillon diffère de la structure par âge de la population française (% pop).

Âge	o_i	% pop. = f_i	$c_i = 284 \times f_i$
0-19	73	*24,6*	69,9
20-39	82	*28,1*	79,8
40-59	75	*26,0*	73,8
60-74	36	*13,6*	38,6
> 74	18	*7,7*	21,9
Total	284	*100,0*	284,0

Sous H_0, la distribution par âge de l'échantillon est identique à celle de la population.
Sous H_1, la distribution observée diffère de celle de la population générale.

On calcule les effectifs théoriques (c_i) en multipliant l'effectif total de l'échantillon 284 par la fréquence de chaque classe dans la population générale.

$$\chi^2_o = \frac{(73-69{,}9)^2}{69{,}9} + \frac{(82-79{,}8)^2}{79{,}8} + \frac{(75-73{,}8)^2}{73{,}8} + \frac{(36-38{,}6)^2}{38{,}6} + \frac{(16-21{,}9)^2}{21{,}9} = 1{,}09$$

Pour ddl = 5 – 1 = 4, la valeur lue dans la table de $\chi^2_{5\%}$ est de 9,49. La valeur 1,09 observée est bien inférieure à cette valeur. On ne rejette donc pas H_0.

Le résultat du test montre qu'il n'existe aucun argument pour dire que l'échantillon est différent de la structure par âge de la population d'origine. Bien qu'on ne puisse jamais affirmer qu'une hypothèse nulle est vraie, on peut cependant considérer que l'échantillon est représentatif de la population en ce qui concerne sa structure par âge.

Excel® : calculer la série des c_i. La fonction TEST.KHIDEUX fournit directement la valeur p.
Dans l'exemple 13.14 TEST.KHIDEUX (série o_i ; série c_i) = 0,90.

XV. TEST DE χ^2 D'HOMOGÉNÉITÉ

Test de Pearson, test de chi-carré.

Quand choisir ce test ?

Lorsqu'on désire comparer les distributions observées entre **plusieurs** échantillons d'une variable qualitative nominale à ***plusieurs*** classes. Si la variable est binaire, le test revient à comparer plusieurs pourcentages.

Variables	qualitatives nominales ou binaires
Paramètres étudiés	effectifs des classes et des échantillons
Taille des échantillons	effectifs théoriques supérieurs ou égaux à 5
Séries comparées	indépendantes
Hypothèse nulle	les distributions ou les pourcentages sont identiques
H_1 bilatérale	les distributions ou les pourcentages sont différents

Formulations

$\mathbf{o_{ij}}$: effectifs observés.
$\mathbf{t_i}$: les totaux des lignes.
$\mathbf{n_j}$: totaux des colonnes.
N : total général.
$\mathbf{c_{ij}}$: effectifs théoriques = $n_j\, t_i/N$.
r : nombre de lignes.
k : nombre de colonnes.
ddl : degré de liberté.

Variable	Échantillons E_1	E_2	…	E_j	Total
A_1	o_{11} c_{11}	o_{12} c_{12}	…	o_{1j} c_{1j}	t_1
…	…	…	…	…	…
A_i	o_{i1} c_{i1}	o_{i2} c_{i2}	… …	o_{ij} c_{ij}	t_i
Total	n_1	n_2	…	n_j	N

Conditions d'application

Tous les effectifs théoriques c_{ij} doivent être supérieurs ou égaux à 5. Si ces conditions ne sont pas réalisées, il faut regrouper certaines classes de la variable.

Principe du test (*cf.* principes généraux, chap. 11.IV)

Sous H_0, les différences entre les effectifs observés et les effectifs théoriques de chaque case devraient être nulles. Le principe du test de χ^2 consiste à regarder si l'ensemble de ces différences est proche de zéro, ou si au contraire l'ensemble des différences est trop éloigné d'une valeur seuil, auquel cas, on rejettera H_0.

Test du χ^2 d'homogénéité :

$$\chi^2 = \Sigma \frac{(o_{ij} - c_{ij})^2}{c_{ij}} \qquad ddl = (r - 1)(k - 1)$$

Résultats

H_1	χ^2	REJET H_0	INTERPRÉTATION
bilatérale	$< \chi^2_{5\%}$	Non	on ne peut affirmer que les distributions sont ≠
	$\geq \chi^2_{5\%}$	Oui	les distributions diffèrent significativement

Exemple 13.15.

Le test de dépistage pour le virus VIH est proposé systématiquement lors d'une grossesse. On désire savoir si la fréquence d'acceptation de ce test varie selon la religion de la femme enceinte. Un échantillon de 3608 femmes est étudié.

Religion :	A	B	C	D	Total
Test effectué	477	1 746	248	135	2 606
Test non fait	135	582	218	67	1 002
Total tests	612	2 328	466	202	3 608
% tests effectués	*77,9*	*75,0*	*53,2*	*66,8*	

H_0 : la fréquence d'acceptation du test est identique quelle que soit la religion.
H_1 bilatérale : la fréquence d'acceptation du test est différente selon les religions.

ddl = (4 – 1) (2 – 1) = 3

$c_{1,1} = 2\,606 \times 612/3\,608 = 442,0$
$c_{1,2} = 2\,606 \times 2\,328/3\,608 = 1681,5$
$c_{1,3} = 2\,606 \times 466/3\,608 = 336,6$
$c_{1,4} = 2\,606 \times 202/3\,608 = 145,9$

$c_{2,1} = 1\,002 \times 612/3\,608 = 170,0$
$c_{2,2} = 1\,002 \times 2\,328/3\,608 = 646,5$
$c_{2,3} = 1\,002 \times 466/3\,608 = 129,4$
$c_{2,4} = 1\,002 \times 202/3\,608 = 56,1$

$$\chi^2 = \frac{(477 - 442,0)^2}{442,0} + \frac{(135 - 170,0)^2}{170,0} + \frac{(1\,746 - 1\,681,5)^2}{1\,681,5} + \frac{(582 - 646,5)^2}{646,5} + \ldots$$
$$+ \frac{(248 - 336,6)^2}{336,6} + \frac{(218 - 129,4)^2}{129,4} + \frac{(135 - 145,9)^2}{145,9} + \frac{(67 - 56,1)^2}{56,1}$$

$$\chi^2 = 2,77 + 7,21 + 2,47 + 6,44 + 23,3 + 60,7 + 0,81 + 2,12 = 105,8$$

Pour ddl = 3, la valeur seuil $\chi^2_{5\%}$ est de 7,81. La valeur trouvée 105,8 est très supérieure. On rejette donc H_0. La valeur trouvée est en outre supérieure à la dernière valeur 21,1 fournie par la table correspondant à $p = 0,0001$.

On peut donc rejeter H_0 avec un risque infime d'erreur ($p < 0,0001$). La fréquence de réalisation du test VIH n'est pas identique selon la religion des femmes enceintes. Le test VIH est moins fréquemment réalisé chez les femmes de confession C (53,2 %).

Excel® : la fonction TEST.KHIDEUX fournit directement la valeur p. Cependant, il faut faire les calcul des $c_{ij} = t_i n_j / N$ en créant une matrice des c_{ij} de même format que les o_{ij}. Ensuite, saisir les 2 matrices avec la fonction.
Dans l'exemple 13.15 TEST.KHIDEUX (série o_{ij} ; série c_{ij}) = 9.10^{-23}!!

XVI. TEST DE χ^2 À 4 CASES POUR COMPARER DEUX POURCENTAGES

Test de χ^2 2 × 2, test de chi-carré, test de χ^2 d'homogénéité, test de Pearson. Ce test peut remplacer le test exact de Fisher.

Quand choisir ce test ?

Lorsqu'on désire comparer deux pourcentages observés sur deux échantillons.

Variable	qualitative binaire
Paramètre étudié	effectifs des classes et des échantillons
Séries comparées	indépendantes
Hypothèse nulle	$P_1 = P_2$
H_1 bilatérale	$P_1 \neq P_2$
H_1 unilatérale	$P_1 > P_2$ ou $P_1 < P_2$

Formulations

a, b, c, d : les effectifs observés (o_i) dans chaque case du tableau.
$\mathbf{n_1}$ et $\mathbf{n_2}$: les effectifs des deux échantillons $n_1 = a + c$ et $n_2 = b + d$.
$\mathbf{p_1}$ et $\mathbf{p_2}$: les deux pourcentages observés: $p_1 = a/n_1$ et $p_2 = b/n_2$.
$\mathbf{P_1}$ et $\mathbf{P_2}$: les pourcentages inconnus des deux populations d'où sont issus les échantillons.
$\mathbf{t_1}$ et $\mathbf{t_2}$: les totaux des effectifs observés pour les 2 classes de la variable.
N : le total général des effectifs observés dans toutes les cases.
ddl : nombre de degré de liberté = 1.

CLASSES DE LA VARIABLE	ÉCHANTILLONS 1	2	TOTAUX
caractère présent	a	b	t_1
caractère absent	c	d	t_2
totaux : effectifs des échantillons	n_1	n_2	N
pourcentage	$p_1 = a/n_1$	$p_2 = b/n_2$	

Conditions d'application

Tous les effectifs théoriques n_1t_1/N, n_1t_2/N, n_2t_1/N, n_2t_2/N doivent être supérieurs ou égaux à 5.
Si cette condition n'est pas réalisée, il faut appliquer le test exact de Fisher (chap. 11.V).

Principe du test (*cf.* principes généraux, chap. 11.IV)

On teste la distribution observée par rapport à une distribution théorique où les effectifs des 2 échantillons seraient uniformément répartis. Le principe est strictement identique au test de χ^2 d'homogénéité. La taille du tableau (4 cases) permet d'utiliser une formule simplifiée dans le calcul du χ^2.

Test du χ^2 à 4 cases : formule simplifiée strictement équivalente à la formule générale

$$\chi^2 = \frac{N(ad - bc)^2}{n_1 n_2 t_1 t_2} \qquad \text{ddl} = 1$$

Résultats

H_1	χ^2	REJET H_0	INTERPRÉTATION
bilatérale	< 3,84	Non	p1 ne diffère pas significativement de p_2
	≥ 3,84	Oui	p1 diffère significativement de p_2
unilatérale	< 2,71	Non	p1 ne diffère pas significativement de p_2
	≥ 2,71	Oui	p1 est significativement supérieur ou inférieur à p_2

Pour obtenir une valeur plus précise de p on peut utiliser la propriété du χ^2 à 4 cases qui n'est autre que le carré d'une loi normale centrée réduite Z ($\sqrt{\chi^2} = Z$). Il suffit donc de calculer la racine de χ^2 et de rechercher p dans la table de Z.

Exemple 13.16.

On désire étudier le risque de complications après traitement des fractures, en fonction de l'existence d'une ouverture cutanée (fracture ouverte). On étudie une série de 165 fractures opérées dans un centre chirurgical.

Fracture ouverte	Complications	Absence de complications	Total	% *complications*
Non	23	113	136	*16,9 %*
Oui	10	19	29	*34,5 %*
Total	33	132	165	

H_0 : la fréquence des complications est identique qu'il y ait ou non une fracture ouverte.
H_1 unilatérale : il existe une fréquence de complications postopératoires plus élevée chez les sujets présentant une fracture ouverte.
On vérifie que l'effectif théorique le plus petit 33 × 29/165 est > 5,

$$\chi^2 = 165\frac{(23 \times 19 - 113 \times 10)^2}{33 \times 132 \times 136 \times 29} = 4,6$$

On rejette H_0 ($\chi^2 > 3,84$). La valeur de χ^2 observée est supérieure à $\chi^2_{4\%}$ (4,22) correspondant à un risque α unilatéral de 2%

Conclusion : la fréquence des complications postopératoires est significativement plus élevée chez les sujets présentant une fracture ouverte ($p < 0,02$).

Excel® : la fonction TEST.KHIDEUX fournit directement la valeur p. Mais il faut calculer la matrice des c_{ij} (*cf.* XV). Si on dispose d'un logiciel statistique, on doit préférer le test exact de Fisher qui donne directement la valeur de p . *Cf.* Chapitre 11.V et Annexes, Formulaire 20. Avec EpiInfo6, utiliser un tableau d'analyse d'enquêtes (EpiInfo6/Epitable/Probabilité/Test de Fisher) et lire le résultat du test de Fisher.

XVII. TEST DE χ^2 DE McNEMAR POUR SÉRIES APPARIÉES

Quand choisir ce test ?

Lorsqu'on désire comparer deux pourcentages observés sur deux échantillons dont chaque individu de l'un est **apparié** à un individu de l'autre. On dispose dans ce cas d'un double échantillon composé de paires. Les paires peuvent être soit concordantes (les deux individus ont la même caractéristique), soit discordantes (un individu possède la caractéristique, l'autre pas).

Variable	qualitative binaire
Paramètre étudié	effectifs des paires discordantes
Taille des échantillons	nombre de paires discordantes ≥ 10
Séries comparées	appariées
Hypothèse nulle	$P_1 = P_2$
H_1 bilatérale	$P_1 \neq P_2$
H_1 unilatérale	$P_1 > P_2$ ou $P_1 < P_2$

Formulations

On note par le signe + la présence du caractère étudié, et par le signe – son absence. Pour chaque paire d'individus, on peut observer, selon la présence ou l'absence du caractère étudié, une des quatre configurations suivantes :
++, – +, + –, – –. Les résultats s'expriment selon un des 2 tableaux suivants.

	ÉCHANTILLON 1	ÉCHANTILLON 2	NOMBRE DE PAIRES
caractère	+	+	e
	–	+	f
	+	–	g
	–	–	h

PRÉSENTATION ÉQUIVALENTE

Échantillon 2 caractère	ÉCHANTILLON 1 caractère présent	ÉCHANTILLON 1 caractère absent
présent	e	f
absent	g	h

f et **g** : nombre de paires discordantes.
$\mathbf{p_1}$ et $\mathbf{p_2}$: les deux pourcentages observés.
$\mathbf{P_1}$ et $\mathbf{P_2}$: les pourcentages inconnus des deux populations d'où sont issus les échantillons.
ddl : nombre de degré de liberté = 1.

Conditions d'application

Le nombre de paires discordantes f + g doit être supérieur ou égal à 10.

Principe du test

Si H_0 est vraie, le nombre **f** de paires discordantes -/+ est égale au nombre **g** de paires discordantes +/-, soit **f = g**. Ce qui revient à dire que **f** et **g** doivent être égaux à la moitié du total des paires discordantes. Cette valeur (f + g)/2 représente l'effectif théorique **c**. Le test de χ^2 revient donc à calculer $(f-c)^2/c + (g-c)^2/c$, qui se simplifie en $(f-g)^2/(f+g)$. Le test de McNemar apparié est plus puissant qu'un simple test de χ^2 pour comparer deux pourcentages.

Test du χ^2 de McNemar :

$$\chi^2 = \frac{(f-g)^2}{f+g} \qquad ddl = 1$$

Résultats

Le résultat se lit et s'interprète comme celui d'un test de χ^2 à 4 cases à 1 ddl pour comparer 2 pourcentages.

Exemple 13.17.

On désire comparer 2 techniques biologiques, l'ELISA et l'hémagglutination (IHAT), dans le diagnostic de l'hydatidose (kyste hydatique). Un total de 56 malades a été testé simultanément par chacune des 2 techniques. La performance (sensibilité) d'une technique est jugée par le nombre de résultats positifs observés.

H_0 : les performances des 2 techniques sont équivalentes.

H_1 bilatérale : les performances des 2 techniques diffèrent.

Résultat ELISA	Résultat IHAT	Nombre de malades
+	+	43
−	+	2
+	−	10
−	−	1

L'ELISA a été rendu positif dans 53 cas sur 56 (94,3 %).
L'IHAT a été rendu positif dans 45 cas sur 56 (77,1 %).

$$\chi^2 \text{ McNemar} = \frac{(10-2)^2}{10+2} = \frac{64}{12} = 5{,}33$$

La valeur observée 5,33 est supérieure à la valeur seuil $\chi^2_{5\%}$ (3,84). On rejette donc H_0.
On lit que pour une valeur de χ^2 égale à 4,71 le risque d'erreur est de 0,03.

Conclusion : les performances des 2 techniques diffèrent significativement ($p < 0{,}03$). L'ELISA est plus sensible que l'hémagglutination.

XVIII. TEST DE χ^2 D'INDÉPENDANCE

Test de Pearson, test de chi-carré.

Quand choisir ce test ?

Lorsqu'on désire tester l'indépendance entre deux variables qualitatives à plusieurs classes observées sur un échantillon (*cf.* tests de liaison, chap. 12).

Variables	qualitatives nominales
Paramètre étudié	effectifs des classes de chaque variable
Taille des échantillons	effectifs théoriques supérieurs ou égaux à 5
Hypothèse nulle	les 2 variables sont indépendantes
H_1 bilatérale	les 2 variables sont liées

Formulations

$\mathbf{o_{ij}}$: effectifs observés.
$\mathbf{t_i}$: les totaux des lignes.
$\mathbf{n_j}$: totaux des colonnes.
$\mathbf{N}$: total général.
$\mathbf{c_{ij}}$: effectifs théoriques $= n_j\, t_i/N$.
$\mathbf{r}$: nombre de lignes.
$\mathbf{k}$: nombre de colonnes.
ddl : degrés de liberté.

	Variable B				
Variable A	B_1	B_2	...	B_j	Total
A_1	o_{11} c_{11}	o_{12} c_{12}	...	o_{1j} c_{1j}	t_1
...	...	...	...	...	...
A_i	o_{i1} c_{i1}	o_{i2} c_{i2}	... ...	o_{ij} c_{ij}	t_i
Total	n_1	n_2	...	n_j	N

Conditions d'application

Tous les effectifs théoriques c_{ij} doivent être supérieurs ou égaux à 5. Sinon, il faut regrouper certaines classes, ou utiliser le test exact de Fisher (chap. 13.XIX) si le tableau ne comporte que 4 cases.

Principe du test (*cf.* principes généraux, chap. 12.I)

Il est identique au test de χ^2 d'homogénéité. Seule, l'interprétation des résultats change.

Test du χ^2 :

$$\chi^2 = \Sigma \frac{(o_{ij} - c_{ij})^2}{c_{ij}} \qquad ddl = (r-1)(k-1)$$

Résultats

H_1	χ^2	Rejet H_0	Interprétation
bilatérale	$< \chi^2_{5\,\%}$	Non	les 2 variables ne sont pas significativement liées
	$\geq \chi^2_{5\,\%}$	Oui	les 2 variables sont liées

Exemple 13.18.

On désire étudier la distribution de l'évolution d'une maladie divisée en 3 classes (guérison, rechute, décès) en fonction de l'administration d'un médicament divisé en 3 classes : abstention (NT), voie orale (VO), voie parentérale (IV); l'échantillon est composé de 81 individus. Le tableau donne les effectifs de chaque classe de A dans chaque classe de B.

Tableau de contingence des effectifs observés :

		Médicament			Total
		NT	VO	IV	N
Évolution de la maladie	guérison	5	6	16	27
	rechute	9	9	10	28
	décès	15	4	7	26
	Total	29	19	33	81

Hypothèses : nous sommes dans la situation d'un test de liaison entre 2 variables.
Les hypothèses sont : H_0 : les deux variables sont indépendantes ;
H_1 : les deux variables sont liées.

$c_{1,1} = 27 \times 29/81 = 9,67$ $\quad c_{1,2} = 27 \times 19/81 = 6,33$ $\quad c_{1,3} = 27 \times 33/81 = 11$
$c_{2,1} = 28 \times 29/81 = 10,02$ $\quad c_{2,2} = 28 \times 19/81 = 6,57$ $\quad c_{2,3} = 28 \times 33/81 = 11,4$
$c_{3,1} = 26 \times 29/81 = 9,31$ $\quad c_{3,2} = 26 \times 19/81 = 6,10$ $\quad c_{3,3} = 26 \times 33/81 = 10,59$

$$\chi^2 = \frac{(5-9,67)^2}{9,67} + \frac{(6-6,33)^2}{6,33} + \frac{(16-11)^2}{11} + \ldots + \frac{(15-9,31)^2}{9,31} + \frac{(4-6,1)^2}{6,1} + \frac{(7-10,59)^2}{10,59}$$

$$\chi^2 = 2,26 + 0,02 + 2,27 + 0,10 + 0,9 + 0,17 + 3,48 + 0,72 + 1,22 = 11,14$$

On rejette donc H_0.
On trouve $\chi^2_{ob} = 11,14$ supérieure à $\chi^2_{3\,\%} = 10,71$ pour ddl = (3 – 1) (3 – 1) = 4.
On conclut donc qu'il existe une liaison entre ces deux variables ($p < 0,03$).

En examinant les fréquences d'évolution de la maladie (tableau a), on constate que le décès est plus fréquent lorsque le médicament n'a pas été administré (NT) et la guérison plus fréquente lorsque le médicament a été administré par voie IV.
Le raisonnement serait identique en examinant le mode d'administration du médicament selon l'évolution de la maladie (tableau b).

Il sera néanmoins plus logique dans la présentation finale des résultats d'exprimer que l'évolution de la maladie est dépendante du traitement choisi (tableau a). Les deux variables ne jouent pas un rôle symétrique.

a) Évolution de la maladie selon l'administration du médicament

Évolution	NT %	VO %	IV %	Total %
guérison	*17,3*	*31,6*	*48,5*	*33,3*
rechute	*31,0*	*47,4*	*30,3*	*34,6*
décès	*51,7*	*21,0*	*21,2*	*32,1*
Total	*100,0*	*100,0*	*100,0*	*100,0*

b) Administration des médicaments selon l'évolution de la maladie

Évolution	NT %	VO %	IV %	Total %
guérison	*18,5*	*22,2*	*59,3*	*100,0*
rechute	*32,1*	*32,1*	*35,8*	*100,0*
décès	*57,7*	*15,4*	*26,9*	*100,0*
Total	*35,8*	*23,5*	*40,7*	*100,0*

Excel® : la fonction TEST.KHIDEUX fournit directement la valeur p. Mais il faut calculer la matrice des c_{ij} (*cf.* XV).
Dans l'exemple 13.18 TEST.KHIDEUX (série o_{ij} ; série c_{ij}) = 0,025.

XIX. TEST DE χ^2 DE TENDANCE

Test de χ^2 d'Armitage.

Quand choisir ce test ?

Lorsqu'on désire mettre en évidence la tendance évolutive d'un pourcentage selon une variable ordinale (ou quantitative discrète). Cela revient à tester la liaison entre la distribution d'une variable binaire Y et une seconde variable X. Ce test permet de conclure non seulement à une différence entre pourcentages observés (comme un simple test de χ^2 comparant plusieurs pourcentages), mais en outre à mettre en évidence une relation entre l'augmentation (ou la diminution) du pourcentage en fonction de la variable X.

Variables	qualitative binaire et qualitative ordinale
Paramètre étudié	effectifs des classes et des échantillons
Taille des échantillons	effectifs théoriques supérieurs ou égaux à 5
Séries comparées	indépendantes
Hypothèse nulle	les pourcentages sont identiques
H_1 bilatérale	les pourcentages varient en fonction de X

Formulations

$\mathbf{x_i}$: les classes numérotées de la variable X. Si la variable est qualitative ordinale on attribue un numéro d'ordre à chacune de ses classes.
$\mathbf{o_{1j}}$ et $\mathbf{o_{2j}}$: effectifs observés.
$\mathbf{t_1}$ et $\mathbf{t_2}$: les totaux des 2 lignes.
$\mathbf{n_j}$: totaux des colonnes.
N : total général.
$\mathbf{c_{ij}}$: effectifs théoriques = $n_j\ t_i/N$.
k : nombre de colonnes.
ddl = degré de liberté.

Variable binaire Y	Variable ordinale B x_1	x_2	...	x_j	Total
+	o_{11} c_{11}	o_{12} c_{12}	...	o_{1j} c_{1j}	t_1
–	o_{21} c_{21}	o_{22} c_{22}	... ...	o_{2j} c_{2j}	t_2
Total	n_1	n_2	...	n_j	N

Conditions d'application

- Tous les effectifs théoriques c_{ij} doivent être supérieurs ou égaux à 5. Si cette condition n'est pas réalisée, il faut regrouper certaines classes de X et numéroter cette classe en lui donnant la valeur moyenne des classes regroupées.
- Il faut par ailleurs que la liaison supposée entre Y et X soit linéaire.

Principe du test (*cf.* principes généraux, chap. 12.II)

Sous H_0, les différences entre les effectifs observés et les effectifs théoriques de chaque case devraient être nulles. Le principe du test de χ^2 consiste à regarder si l'ensemble de ces différences est proche de zéro, ou si au contraire l'ensemble des différences est trop éloigné d'une valeur seuil, auquel cas, on rejettera H_0.

Test du χ^2 de tendance (*cf.* principes généraux, chap. 12.II)

$$\chi^2 = \frac{N^3[\Sigma x_i(o_{1i} - c_{1i})]^2}{t_1 t_2[N\Sigma(n_i x_i^2) - (\Sigma n_i x_i)^2]} \qquad ddl = 1$$

Résultats

H_1	χ^2	REJET H_0	INTERPRÉTATION
bilatérale	$< \chi^2_{5\%}$	Non	on ne peut affirmer que les pourcentages sont différents
	$\geq \chi^2_{5\%}$	Oui	les pourcentages augmentent (ou diminuent)

Exemple 13.19.

On désire étudier la plombémie (plomb dans le sang) d'un échantillon de 172 enfants, en fonction de l'exposition professionnelle des parents au plomb. Les enfants dont le dosage du plomb est supérieur à 70 microgrammes/L sont définis comme « hyperplombémiques ». Pour chaque enfant, on note l'exposition des parents : soit aucun parent n'est exposé (0), soit un seul est exposé (1), soit les 2 sont exposés (2). On se demande s'il existe une liaison entre le nombre de parents exposés et l'existence d'une hyperplombémie chez les enfants.

Plombémie chez les enfants	Parents exposés au plomb 0	1	2	Total
> 70 microg/L	6	51	12	69
< 70 microg/L	41	60	2	103
Total	47	111	14	172
% hyperpb > 70	*12,8*	*45,9*	*85,7*	*40,1*

On vérifie que les effectifs théoriques les plus faibles $14 \times 69/172$ sont > 5.

On pose H_0 : il n'existe pas de liaison entre le nombre de parents exposés et l'hyperplombémie. H_1 bilatérale : il existe une liaison entre l'exposition des parents et l'hyperplombémie.

On a $o_{11} = 47 \times 69/172 = 18{,}9$ $\quad o_{12} = 111 \times 69/172 = 44{,}5$ $\quad o_{13} = 14 \times 69/172 = 5{,}6$

$$\chi^2 = \frac{172^3[0(6-18{,}9) + 1(51-44{,}5) + 2(12-5{,}6)]^2}{69 \times 103[172(47 \times 0^2 + 111 \times 1^2 + 14 \times 2^2) - (47 \times 0 + 111 \times 1 + 14 \times 2)^2]} = 28{,}4$$

ddl = 3 – 1 = 2 On rejette H_0. La valeur de χ^2 est encore supérieure à $\chi^2_{0{,}0001}$.

On peut donc affirmer qu'il existe une relation entre le nombre de parents exposés professionnellement au plomb et l'existence d'une hyperplombémie chez les enfants ($p < 0{,}0001$). Cette relation hautement significative, n'apporte cependant aucun argument de causalité.

XX. TEST DU COEFFICIENT DE CORRÉLATION

Quand choisir ce test ?

Lorsqu'on désire tester l'existence d'une liaison entre 2 variables quantitatives.
Test équivalent : test du coefficient de corrélation des rangs de Spearman.

Variables	quantitatives
Paramètre étudié	coefficient de corrélation
Séries comparées	séries appariées
Hypothèse nulle	absence de liaison entre X et Y : $\rho = 0$
H_1 bilatérale	liaison entre X et Y : $\rho \neq 0$
H_1 unilatérale	liaison positive : $\rho > 0$ ou liaison négative : $\rho < 0$

Formulations

ρ : coefficient de corrélation dans la population d'où est extrait l'échantillon.
r : coefficient de corrélation estimé sur l'échantillon.
x_i et **y_i :** les valeurs x et y d'un couple de données.
n : nombre de couples de l'échantillon étudié.
ddl : degrés de liberté.

Conditions d'application

- Les variables X et Y doivent être aléatoires.
- L'association entre X et Y doit être linéaire.
- Dans la population d'où est extrait l'échantillon, il faut que les distributions de Y liées à chaque valeur de X soit normales et de variance constante, et que symétriquement les distributions de X liées à chaque valeur de Y soit aussi normales et de variance constante. Cette condition, impossible à vérifier en pratique est le plus souvent vérifiée en biologie.
- Les observations pour chaque variable doivent être indépendantes les unes des autres. Cette condition n'est pas remplie lorsque, par exemple, on compare des données Y en fonction du temps X. Les données de la veille ne sont pas indépendantes des données du lendemain. On dit qu'il y a autocorrélation, et l'analyse de ces problèmes se fait par des techniques d'analyse de ***séries chronologiques***.

Principe du test (*cf.* principes généraux, chap. 12.III)

Sous H_0, le coefficient de corrélation doit être nul. On teste donc la différence $\mathbf{r - 0}$. Sous H_0, le rapport de cette différence sur son écart type suit une loi T de Student.

Calcul intermédiaire

$$r = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\sqrt{\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right]\left[\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}\right]}}$$

écart type de r : $s_r = \sqrt{\frac{1 - r^2}{n - 2}}$

Test du coefficient de corrélation :

$$t = \frac{|r - 0|}{s_r} \qquad ddl = n - 2$$

Résultats

H_1	t	REJET H_0	INTERPRÉTATION
bilatérale	$< t_{5\%}$	Non	absence de liaison significative entre X et Y
	$\geq t_{5\%}$	Oui	liaison entre X et Y
unilatérale	$< t_{10\%}$	Non	absence de liaison positive ou négative entre X et Y
	$\geq t_{10\%}$	Oui	liaison positive (ou négative) entre X et Y

Exemple 13.20.

On désire vérifier la corrélation entre la taille (en cm) et le poids (en kg) des enfants de 2 ans sur un échantillon de 15 individus.

taille : X	82,9	83,4	82,4	82,1	84,8	86,7	84,0	89,0	85,0	85,4	87,7	87,7	86,4	86,4	86,9
poids : Y	8,7	9,2	9,5	10,1	10,4	10,5	10,8	11,0	11,5	11,6	12,4	13,6	13,8	13,9	14,6

H_0 : il n'existe aucune corrélation entre taille et poids.
H_1 bilatérale : il existe une relation entre taille et poids.
On fait l'hypothèse d'une distribution normale de la taille et du poids dans la population.
On calcule les produits xy :

xy	721,23	767,28	782,8	829,21	881,92	910,35	907,2	979	977,5	990,64	1 087,5	1 192,7	1 192,3	1 201	1 268,7

Σxy	Σx	Σx^2	Σy	Σy^2	$\Sigma xy - \Sigma x \Sigma y/15$	$\Sigma x^2 - (\Sigma x)^2/15$	$\Sigma y^2 - (\Sigma y)^2/15$
14 689,4	1 280,8	109 425,3	171,6	2 011,0	37	62,1	47,9

$r = 37/\sqrt{62,1 \times 47,9} = 0,68$

$s_r = \sqrt{(1 - 0,72^2)/(15 - 2)} = 0,2$ $\qquad t = 0,68/0,2 = 3,4$ $\qquad ddl = 13$

La valeur de t est supérieure à la valeur de $T_{5\%}$ avec ddl = 13. On rejette donc H_0. La valeur t est encore supérieure à $T_{1\%}$. On conclut donc qu'il existe une liaison positive significative entre la taille et le poids des enfants de 2 ans ($p < 0,01$).

Excel® : la fonction COEFFICIENT.CORRELATION fournit directement la valeur de p.
Appliquer ensuite les formules de calculs de s_r et t.
Dans l'exemple 13.20, COEFFICIENT.CORRELATION (série taille ; série poids) = 0,68.

XXI. TEST DU COEFFICIENT DE CORRÉLATION DES RANGS DE SPEARMAN

Quand choisir ce test ?

Lorsqu'on désire tester l'existence d'une liaison entre 2 variables quantitatives. Il s'applique dans les mêmes circonstances que le test de corrélation précédent. Ce test est un test non-paramétrique, qui s'intéresse non pas aux valeurs de la variable, mais à leurs **rangs**.
Test équivalent : test du coefficient de corrélation.

Variables	quantitatives
Paramètre étudié	rangs des valeurs
Taille de l'échantillon	> 10
Séries comparées	séries appariées
Hypothèse nulle	absence de liaison entre X et Y : $r_s = 0$
H_1 bilatérale	liaison entre X et Y : $r_s \neq 0$
H_1 unilatérale	liaison positive : $r_s > 0$ ou liaison négative $r_s < 0$

Formulations

r_s : coefficient de corrélation de Spearman.
x_i' et y_i' : les rangs des valeurs observées. On appelle rang, le numéro d'ordre d'une valeur après classement de la variable par ordre croissant. Sur la série 1,4,5,8, la valeur 5 a pour rang 3 et la valeur 8 a pour rang 4.
n : nombre de couples de l'échantillon étudié.
ddl : degrés de liberté.

Conditions d'application

- Le nombre de couples à tester doit être supérieur à 10 et ne pas comporter trop de valeurs ex-æquo.
- Les variables X et Y doivent être aléatoires et jouer un rôle symétrique.
- Les observations pour chaque variable doivent être indépendantes les unes des autres.
- Il n'y a pas d'exigences de normalité des distributions de X et Y dans la population, ni d'exigences de linéarité de la relation entre X et Y.

Principe du test (*cf.* principes généraux, chap. 12.III)

Il est identique au test de corrélation, à ceci près que l'on confronte les rangs des observations au lieu de leurs valeurs. Sous H_0, le coefficient de corrélation doit être nul. On teste donc la différence $r_s - 0$. Sous H_0, le rapport de cette différence sur son écart type suit une loi T de Student.

Calcul intermédiaire

- On classe chacune des 2 séries de données par ordre de valeur croissante.
- On note pour chaque individu (ou unité statistique) le numéro de rang x' et y' de chacune des valeurs x et y. S'il existe des *ex æquo*, on tire au sort les rangs respectifs de chaque *ex æquo*, ou on leur attribue un rang moyen.

$$r_s = 1 - \frac{6\Sigma(x'_i - y'_i)^2}{n(n^2-1)} \qquad \text{écart type } s_{r_s} = \sqrt{\frac{1-r_s^2}{n-2}}$$

Test de Spearman :

$$t = \frac{|r_s - 0|}{s_r} \qquad ddl = n - 2$$

Résultats

H_1	t	REJET H_0	INTERPRÉTATION
bilatérale	$< t_{5\,\%}$	Non	absence de liaison significative entre X et Y
	$\geq t_{5\,\%}$	Oui	liaison entre X et Y
unilatérale	$< t_{10\,\%}$	Non	absence de liaison positive ou négative entre X et Y
	$\geq t_{10\,\%}$	Oui	liaison positive ou négative entre X et Y

Exemple 13.21

On désire vérifier la corrélation entre la taille (en cm) et le poids (en kg) des enfants de 2 ans sur un échantillon de 15 individus.

taille : X	82,9	83,4	82,4	82,1	84,8	86,7	84,0	89,0	85,0	85,4	87,7	87,7	86,4	86,4	86,9
poids : Y	8,7	9,2	9,5	10,1	10,4	10,5	10,8	11,0	11,5	11,6	12,4	13,6	13,8	13,9	14,6

H_0 : il n'existe aucune corrélation entre taille et poids.
H_1 bilatérale : il existe une relation entre taille et poids.

On remplace les valeurs par leur rang dans chacune des séries :

x'	3	4	2	1	6	11	5	15	7	8	13,5	13,5	9,5	9,5	12
y'	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

On calcule les carrés des différences $(x - y)^2$:

$(x' - y')^2$	4	4	1	9	1	25	4	49	4	4	6,25	2,25	12,25	20,25	9

$\Sigma (x' - y')^2 = 155$ $\quad r_s = (1 - 6 \times 155)/15\,(15^2 - 1) = 0{,}72$ $\quad s_r = \sqrt{(1 - 0{,}72^2)/(15 - 2)} = 0{,}19$
$t = 0{,}72/0{,}19 = 3{,}79$ $\quad ddl = 13$
La valeur de t est supérieure à la valeur de $T_{5\,\%}$ avec ddl = 13. On rejette donc H_0. La valeur t est encore supérieure à $T_{1\,\%}$.

On conclut donc qu'il existe une liaison positive significative entre la taille et le poids des enfants de 2 ans ($\rho < 0{,}01$).

On constate qu'on observe un résultat proche du test de corrélation de l'exemple précédent.

Excel® : utiliser la fonction RANG appliquée aux 2 séries pour obtenir le rang de chaque valeur. Appliquer ensuite les formules de calcul de r_s, s_r et t.

14

TESTS STATISTIQUES DIVERS

I. ÉPREUVE DE NORMALITÉ

Nous avons vu que dans de nombreuses situations en statistiques, on faisait l'hypothèse qu'une distribution était normale. De nombreux tests exigent cette condition.
Par ailleurs, dans les sciences de la vie, il est possible que des variables ne soient pas distribuées de façon normale, même si ce modèle est le plus fréquent. Il ne faut pas oublier que la loi normale n'est qu'un modèle, parmi beaucoup d'autres.
Avant d'utiliser un test statistique, exigeant la normalité des distributions à comparer, il est donc prudent de vérifier cette condition avant d'affirmer un résultat.
La plupart du temps, lorsqu'on travaille sur des variables connues, cette condition est implicite. On sait que telle variable (poids, taille, durée d'incubation, *etc.*) suit une loi normale dans la population. On peut donc effectuer les tests en rappelant cette hypothèse admise.
Parfois, on sait que la variable brute, ne suit pas une loi normale, mais qu'une transformation adéquate (transformation logarithmique le plus souvent) permet de redresser la distribution selon une loi normale. Dans ce cas, le test statistique doit être appliqué en utilisant les valeurs transformées.
Parfois, lorsqu'on travaille sur une variable nouvelle, l'allure de la distribution peut être douteuse. Il est alors nécessaire de pratiquer une épreuve de normalité.

1. Méthode approchée

Elle consiste à examiner si la distribution observée possède les propriétés d'une distribution normale.

- Moyenne = médiane.
- Espace interquartile symétrique autour de la médiane.
- Intervalle contenu entre – 1 écart type et + 1 écart type de part et d'autre de la moyenne contenant environ 2/3 des valeurs.
- Intervalle contenu entre – 2 écarts type et + 2 écarts type de part et d'autre de la moyenne contenant environ 95 % des valeurs.
- Intervalle contenu entre – 3 écarts type et + 3 écarts type de part et d'autre de la moyenne contenant environ 99 % des valeurs.
- Coefficient de dissymétrie (*skewness*) égal à zéro.
- Coefficient d'aplatissement (*kurtosis*) égal à 3.

2. Méthode graphique

L'allure d'une distribution normale suit une courbe en cloche. L'ajustement de la courbe des fréquences observées à la courbe normale est d'autant plus fin que les effectifs sont nombreux.
Si on calcule les fréquences cumulées de la distribution, on obtient la fonction de répartition : $P(X \leq x)$. Cette fonction suit une courbe en S.
Une méthode plus fine consiste à représenter la fonction de répartition à l'aide d'une échelle à ordonnées gaussiennes (papier gausso-arithmétique). Si la distribution est normale, la courbe en S se transforme alors en droite, dite droite de Henry (figure 14.1).

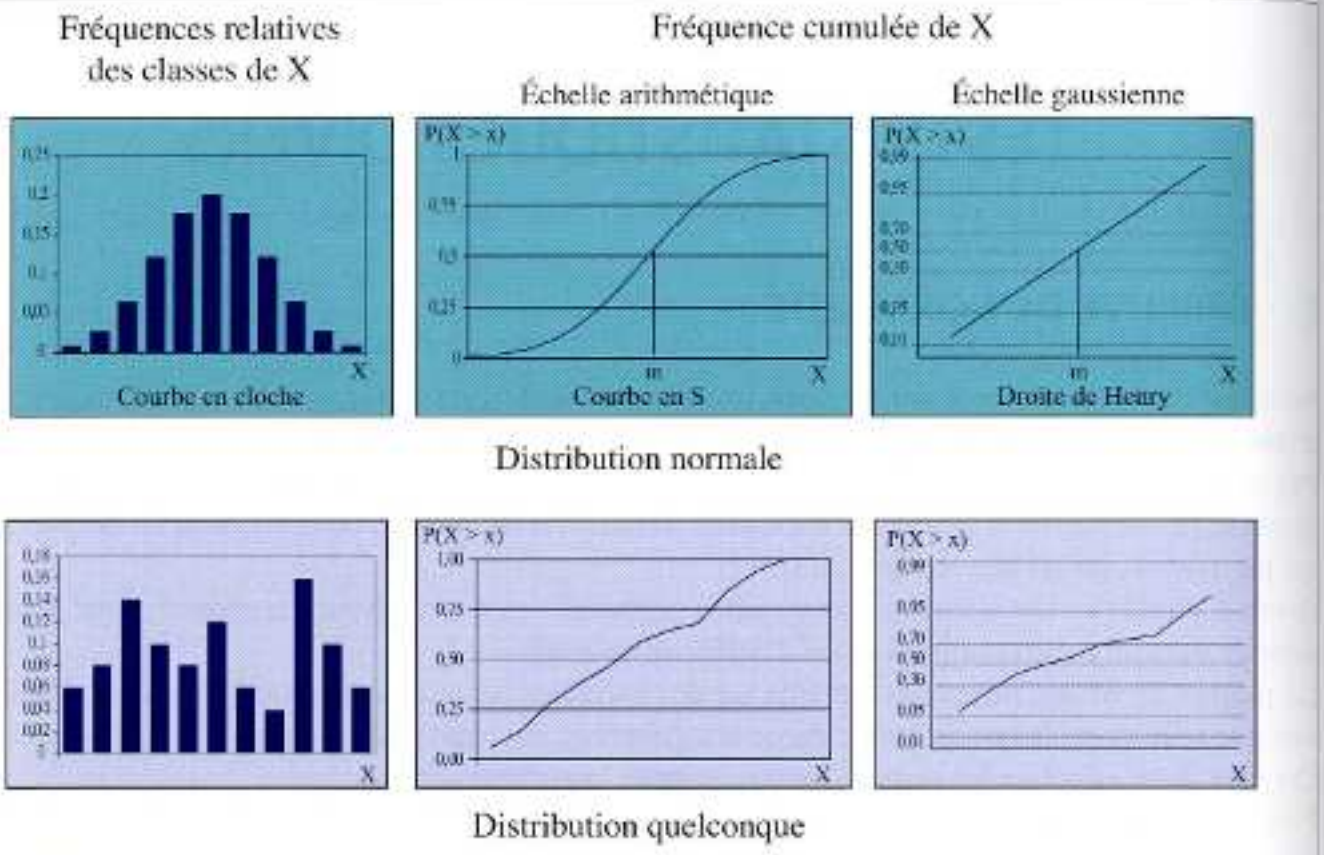


Figure 14-1. Épreuve graphique de normalité

La fonction de répartition de la distribution du haut se transforme en droite sur un graphe à ordonnées en échelle gaussienne : la distribution est proche de la normale.
La distribution du bas se transforme en ligne brisée. La distribution n'est pas normale.

3. Méthode analytique

a) Test de χ^2

Il s'agit d'une méthode plus formelle pour vérifier si une distribution s'écarte de la loi normale. La méthode consiste à construire un test statistique de comparaison d'une distribution observée à une distribution théorique. On peut utiliser un test de χ^2.

Principe

Il consiste :

- à rechercher les probabilités des valeurs de la distribution si celle-ci était normale ;
- à calculer les effectifs théoriques qui seraient observés si la distribution avait été normale ;
- à comparer les effectifs observés aux effectifs théoriques par un test de χ^2.

Le test de χ^2 s'effectue comme un classique test de conformité (chap. 13.XVI).
H_0 : la distribution observée ne diffère pas d'une distribution normale.
H_1 : la distribution observée diffère d'une distribution normale.

- Si la moyenne et l'écart type sont connus dans la population le nombre de ddl est c – 1 (c : nombre de classes de la distribution).
- Si la moyenne et l'écart type sont calculés à partir des données de l'échantillon, le nombre de ddl doit être diminué de 2 : ddl = c – 3.

Exemple 14.1.

Soit la distribution suivante ayant pour paramètre : moyenne m = 53 et écart type s = 4.

	Classes de distribution	< 47	[47-49[	[49-51[	[51-53[	[53-55[	[55-57[	[57-59[	[59-61[	[61-63[	≥ 63	Total
1	Effectifs observés : o	0	16	32	23	17	14	11	8	10	0	131
2	$p_i = P(x_{i1} < X < x_{i2})$	0,067	0,092	0,150	0,191	0,191	0,150	0,092	0,044	0,017	0,006	1,00
3	Effectifs théoriques : c	8,8	12,0	19,6	25,1	25,1	19,6	12,0	5,8	2,2	0,8	131

Le symbole **[49** signifie que la valeur 49 de la borne inférieure est incluse dans l'intervalle.
Le symbole **51[** signifie que la valeur 51 de la borne supérieure n'est pas incluse dans l'intervalle.

Pour chaque classe de la variable les probabilités p_i ont été obtenues en :
- calculant les variables centrées réduites z telle que z = (m – x)/s pour chaque borne x ;
- notant la probabilité que Z soit comprise entre les deux bornes de la classe si la distribution était normale (pour cela il faut disposer de la table de la distribution cumulée de Z).

Les effectifs théoriques c_i ont été obtenus en multipliant les probabilités p_i par l'effectif total :

$$c_i = p_i \times 131.$$

Ces effectifs théoriques sont les effectifs attendus si la distribution était normale.

- on calcule le χ^2

$$\chi^2 = \frac{(0-8,8)^2}{8,8} + \frac{(16-12,0)^2}{12,0} + \ldots\ldots\ldots + \frac{(10-2,2)^2}{2,2} + \frac{(0-0,8)^2}{0,8} = 52,3$$

$$\text{ddl} = 10 - 1 - 2 = 7 \qquad p < 0,0001$$

On rejette donc H_0. La distribution observée s'éloigne significativement d'une distribution normale. Ce résultat pouvait être prévu d'après l'allure de l'histogramme de la distribution (figure 14.2).

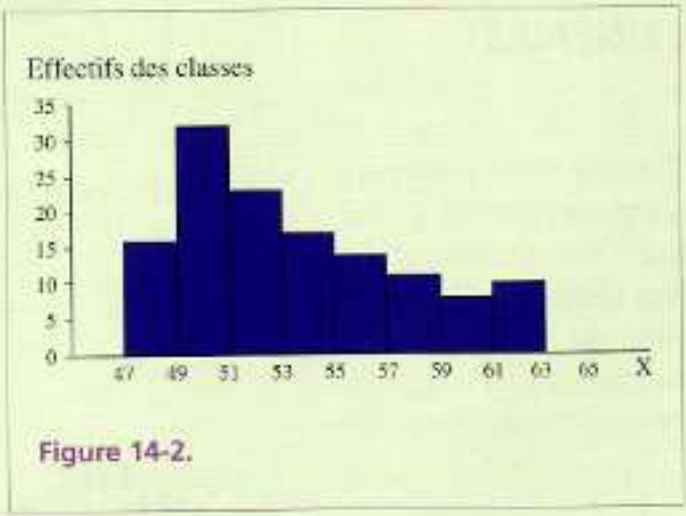


Figure 14-2.

Les paragraphes suivants énumèrent un certain nombre de tests statistiques parfois utilisés et cités dans des publications scientifiques. Afin d'aider le lecteur à comprendre leur utilisation nous avons juste exposé l'objectif de chacun de ces tests et la signification de leur résultat.

b) Test de Kolmogorov-Smirnov

Il s'agit d'un test non-paramétrique qui de façon générale permet de tester si une variable X suit une distribution définie par sa fonction de répartition. Il s'applique donc en particulier à la

fonction de répartition en S de la loi normale et permet de tester la normalité d'une distribution (figure 14.5).
Une valeur de $p < 0,05$ signifie que la distribution observée s'écarte d'une distribution normale.

c) Test de Shapiro-Wilk

Il est produit par certains logiciels de statistiques pour tester la normalité d'une distribution.

II. TEST DE BARTLETT

Il sert à tester la dispersion de plusieurs distributions, autrement dit à tester leurs variances (figure 14.3). On l'utilise pour vérifier l'homogénéité des variances exigée dans les conditions d'application de certains tests. Un résultat avec $p < 0,05$ signifie que les variances sont hétérogènes.

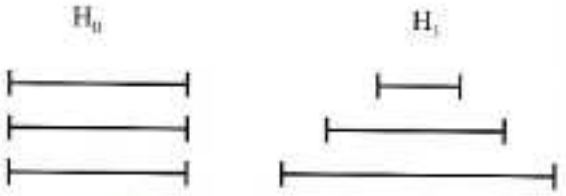


Figure 14-3. Comparaison de l'homogénéité de variances

III. TEST DE LEVENE

Il a le même objet que le test de Bartlett. Il a l'avantage d'être plus robuste et de ne pas exiger une normalité stricte des distributions.

IV. CORRÉLATION LINÉAIRE MULTIPLE

Cette analyse teste la liaison entre plusieurs variables (la corrélation vue au chapitre § 12.3 testait la liaison entre 2 variables seulement). Le principe en est exactement identique. L'analyse aboutit à un coefficient de corrélation **r** (figure 14.4). Un test du coefficient **r** significatif indique qu'il existe une liaison entre les variables.

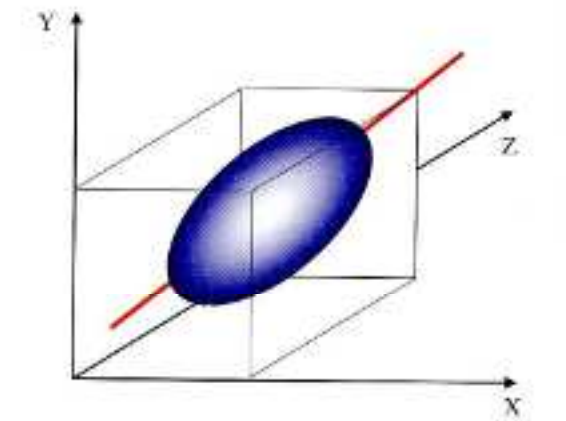


Figure 14-4. Corrélation multiple

V. RÉGRESSION LINÉAIRE MULTIPLE

On utilise cette analyse lorsqu'on veut étudier la liaison d'une variable Y en fonction de plusieurs variables X_i. Son principe est identique à celui de la régression entre une variable Y et une seule variable X.

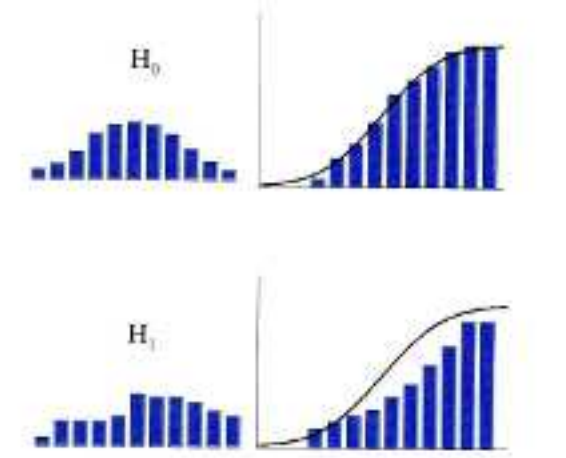


Figure 14-5. Test de Kolmogorov-Smirnov

Quatrième partie

ÉPIDÉMIOLOGIE

ÉPIDÉMIOLOGIE

Introduction

MESURES EN ÉPIDÉMIOLOGIE

I. MESURES DE BASE
II. INDICATEURS ÉPIDÉMIOLOGIQUES

ENQUÊTES ÉPIDÉMIOLOGIQUES

I. PROTOCOLE D'ENQUÊTE
II. TYPES D'ENQUÊTES
III. ENQUÊTES DE COHORTE
IV. ENQUÊTES CAS-TÉMOINS
V. ENQUÊTES TRANSVERSALES
VI. CRITÈRES DE CAUSALITÉ DANS UNE ENQUÊTE ÉTIOLOGIQUE
VII. BIAIS DANS LES ENQUÊTES ÉTIOLOGIQUES
VIII. PRISE EN COMPTE D'UN TIERS FACTEUR : ANALYSE STRATIFIÉE

INVESTIGATION D'UNE ÉPIDÉMIE

I. DÉFINITIONS
II. OBJECTIFS
III. CHRONOLOGIE
IV. ASPECTS OPÉRATIONNELS

MESURES D'IMPACT

I. FRACTION ÉTIOLOGIQUE DU RISQUE
II. FRACTION PRÉVENTIVE
III. INTERVALLE DE CONFIANCE

STANDARDISATION DES TAUX

I. POSITION DU PROBLÈME
II. PRINCIPE
III. MÉTHODE DIRECTE
IV. MÉTHODE INDIRECTE
V. CONDITIONS D'APPLICATION
VI. EXTENSION DE LA MÉTHODE

ANALYSE DE SURVIE

I. PRINCIPE
II. MÉTHODE DE KAPLAN-MEIER
III. LA MÉTHODE ACTUARIELLE
IV. COMPARAISON DE COURBES DE SURVIE : TEST DU LOG RANK

PERFORMANCES D'UNE TECHNIQUE

I. MESURE EXPÉRIMENTALE DES PERFORMANCES D'UN TEST
II. PERFORMANCES D'UN TEST EN SITUATION RÉELLE
III. REPRODUCTIBILITÉ ET CONCORDANCE

Introduction

L'épidémiologie est une science aux contours incertains. Le terme « d'épidémiologie » est lui-même ambigu et son sens a varié au cours des âges.

Dans son sens littéral, l'épidémiologie est la science des phénomènes qui concerne l'ensemble d'une population vivant sur un territoire (επι–δεμοσ–λογοσ). Le mot « demos » évoque autant la notion de « pays » que l'ensemble des individus qui le peuple.

À l'époque d'Hippocrate, le terme « épidémie » s'appliquait à tout événement affectant une communauté humaine : désastres, cataclysmes naturels, guerres. Mais peu à peu le sens s'est limité aux phénomènes pathologiques.

Au XIX^{e} siècle, de nombreux praticiens soucieux de santé communautaire, ont commencé à mettre en relation la fréquence de certaines maladies avec des caractéristiques d'hygiène dont on pressentait le rôle dans l'état de santé des populations : habitat, consommation d'aliments, approvisionnement en eau, *etc.* Cette approche s'est appliquée avec succès à l'étude des grandes épidémies de l'époque. L'étude du choléra à Londres par John Snow qui découvrit par une étude topologique la « source » de l'épidémie en fut une des plus brillantes illustrations.

À l'ère pastorienne, la recherche des causes des maladies infectieuses est devenue la préoccupation majeure. Les nombreuses découvertes des microbiologistes (agents pathogènes, modes de transmission, vecteurs, réservoirs de germes, vaccins, *etc.*) ont orienté l'épidémiologie dans la voie de recherche sur l'histoire naturelle des maladies infectieuses.

Cette réduction de l'épidémiologie à l'infectiologie a perduré dans les esprits et encore maintenant l'épidémiologie est souvent comprise comme étant l'étude des épidémies.

Au cours de la seconde moitié du XX^{e} siècle, le champ d'application de l'épidémiologie s'est élargi et a repris son sens originel. Toutes les pathologies humaines, et pas seulement les maladies infectieuses, peuvent faire l'objet d'études épidémiologiques : les maladies chroniques et génétiques, les accidents, les maladies dues aux comportements, à l'environnement, à l'alimentation, aux soins eux-mêmes, *etc.* L'épidémiologie s'est également développée dans le domaine des maladies animales.

Parallèlement à l'extension de son champ d'application, les méthodes utilisées en épidémiologie ont profondément changé. Les outils se sont enrichis. Aux microscopes et aux boîtes de Pétri, sont venus s'ajouter les techniques de biologie moléculaire, l'informatique, les perfectionnements de la statistique et de la modélisation mathématique. L'épidémiologie est devenue une science multidisciplinaire exercée par des spécialistes très divers. Les corps de métiers exerçant cette discipline se recrutent certes encore fréquemment parmi les médecins, mais aussi parmi les vétérinaires, pharmaciens, biologistes, infirmiers, ingénieurs sanitaires, statisticiens, démographes, géographes, économistes, responsables de santé publique, *etc.*

Dans son acception moderne, l'épidémiologie s'applique à l'étude des phénomènes de santé dans la communauté. Une définition très générale peut ainsi être proposée :

L'épidémiologie est la discipline qui a pour but
l'étude de la distribution des phénomènes de santé dans une population
et des facteurs qui conditionnent leurs fréquences.

Dans cette définition simple on constate que cette discipline se divise en deux grands axes :
- l'épidémiologie descriptive qui mesure les phénomènes de santé dans une population ;
- l'épidémiologie analytique qui recherche les causes éventuelles des maladies.

L'épidémiologie est donc une science qui participe directement aux actions de santé publique. Elle est une des disciplines qui permet d'étayer la prise de décision. Elle apporte aux responsables de la politique de santé, des mesures, des prévisions et des évaluations.

En pratique l'épidémiologie se décompose en activités :
- de surveillance ;
- d'investigation ;
- de recherche ;
- d'évaluation.

Les méthodes de surveillance et d'évaluation, qui représentent un vaste champ d'activité en épidémiologie sortent du cadre de cet ouvrage.

Nous aborderons dans les chapitres suivants les méthodes de base utilisées pour :
- mesurer la fréquence des maladies et leur distribution dans une population ;
- étudier la liaison entre des facteurs de risque et la survenue des maladies.

15

MESURES EN ÉPIDÉMIOLOGIE

Comme l'étymologie du mot « épidémiologie » le suggère, une mesure en épidémiologie est rapportée à la population qui est étudiée. En simplifiant, on pourrait dire que l'épidémiologie est la science mettant en « rapport » le clinicien, qui comptabilise des cas, avec le démographe qui fournit les données sur les populations.

Toute mesure en épidémiologie doit donc être précédée par une définition précise des deux termes du rapport.

- Définition d'un cas (numérateur).
- Définition de la population d'étude (dénominateur).

I. MESURES DE BASE

Les rapports calculés en épidémiologie peuvent être des proportions, des ratios, des cotes, des indices ou des taux.

1. Proportion

Dans une proportion, le numérateur est une part du dénominateur. $P = a/(a + b)$. Numérateur et dénominateur sont donc de même nature. Une proportion s'exprime sous forme d'un nombre compris entre 0 et 1, ou bien sous forme d'un pourcentage (ou pour mille, pour dix mille, pour cent mille, *etc.*) (exemple 15.1).

Exemple 15.1. PROPORTION

Dans une population de 7 500 enfants de moins de 5 ans, on constate que 5 300 sont correctement vaccinés contre la rougeole. La proportion d'enfants vaccinés est de 5 300/7 500 = 0,707 = 70,7 %. Cette proportion est communément appelée « couverture vaccinale ».

2. Ratio

Un ratio représente le rapport entre les effectifs de 2 classes d'une même variable. Ou de façon équivalente le rapport des fréquences de 2 classes. Le numérateur et le dénominateur sont donc de même nature, mais sont exclusifs l'un de l'autre. Un ratio s'exprime par un nombre sans unités (exemple 15.2).

Exemple 15.2. RATIO

Dans une population de 100 individus, on observe 49 hommes et 51 femmes. Le ratio H/F (sex ratio en anglais) = 49/51 = 0,96 ou 0,96 homme pour 1 femme.

3. Cote

La cote est le ratio de la probabilité de survenue d'un événement sur la probabilité de non-survenue de cet événement. Lorsque la variable est binaire (comme dans l'exemple ci-dessus) on utilise parfois le terme de cote pour exprimer un ratio. Le terme est principalement utilisé dans les enquêtes étiologiques : cote des exposés sur les non-exposés (*cf.* chap. 16.IV.3). Elle s'exprime en nombre d'unités du numérateur pour une unité du dénominateur (exemple 15.3).

Exemple 15.3. COTE

Lors d'une épidémie de 75 cas d'une maladie, on a observé 53 cas ayant consommé un aliment X et 22 cas n'en ayant pas consommé :
cote d'exposition chez les cas : 53/22 = 2,4 soit 2,4 cas exposés pour 1 cas non-exposé.

4. Indice

Un indice (appellé aussi communément ratio) est le rapport de deux effectifs qui sont de nature différente. On les utilise surtout comme indicateurs de fonctionnement, notamment en économie de la santé (exemple 15.4).

Exemple 15.4. INDICES

Indice : nombre de...	Exemple	Expression de l'indice
lits d'hôpital/médecin	850 lits, 10 médecins	85 lits pour 1 médecin
individus par foyer	1 200 personnes, 250 foyers	4,8 personnes par foyer
enfants par infirmière nutritionniste	1 000 enfants, 10 infirmières	100 enfants par infirmière
réfugiés par agent de santé	15 290 réfugiés, 15 agents	1 000 réfugiés par agent
réfugiés par latrines	15 000 réfugiés, 75 latrines	200 personnes par latrines
litres d'eau par personne par jour	100 000 pers., 1 000 t d'eau/j	10 L par personne par jour

5. Taux

En épidémiologie, un taux est un rapport qui prend en compte la notion de temps. Un taux mesure la probabilité de survenue d'un événement au cours du temps. Au numérateur, figurent des individus ayant subi un événement pendant une période de temps déterminé. Au dénominateur, figure l'ensemble des individus susceptibles de connaître l'événement pendant cette période.
La notion sera développée au chapitre traitant de l'incidence (chap. 15.II.2).

II. INDICATEURS ÉPIDÉMIOLOGIQUES

Selon le phénomène observé, maladie ou décès, on distingue :
- les indicateurs de morbidité qui décrivent la fréquence des maladies ;
- les indicateurs de mortalité qui décrivent la fréquence des décès.

Ces indicateurs ont pour objectif de répondre à deux types de questions très différentes.

■ Quelle est la ***fréquence*** du phénomène à un moment déterminé ?
Les indicateurs adéquats sont des indicateurs ***statiques*** : prévalence, mortalité proportionnelle, létalité. Ils donnent une image de la situation à un moment donné en mesurant son ampleur. Ces mesures sont obtenues par des enquêtes transversales.
Les indicateurs statiques ne donnent aucune information sur la vitesse de survenue des événements. On pourrait les comparer à l'altimètre d'un aéronef qui indique l'altitude, mais n'indique en rien si l'appareil est stable, en ascension ou en piqué. Pour connaître la dynamique du phénomène il faudrait réaliser plusieurs mesures successives par une succession d'enquêtes transversales.

■ Quelle est la ***vitesse*** de survenue du phénomène pendant une période déterminée ?
Les indicateurs adéquats sont des indicateurs ***dynamiques*** : taux d'incidence, taux de mortalité. Ils donnent une idée de la vitesse d'apparition des phénomènes. On pourrait les comparer au variomètre d'un aéronef. Ces indicateurs sont obtenus par des enquêtes longitudinales ou par des systèmes de surveillance continue.
Quels que soient les indicateurs utilisés, les résultats doivent être exprimés en précisant la date de l'étude, les limites géographiques et les caractéristiques de la population étudiée.

1. Prévalence

C'est un indicateur statique de morbidité. La prévalence d'une maladie se définit comme la proportion du nombre de cas d'une maladie observée à un instant donné sur la population dont sont issus ces cas.

$$\text{Prévalence} = \frac{\text{nombre de cas observés à un instant t}}{\text{population à risque à cet instant t}}$$

- La prévalence s'exprime sous forme d'un chiffre entre 0 et 1, ou d'un pourcentage : nombre de cas pour 100 (1 000,10 000, *etc.*).
- La prévalence mesure tous les cas inclus par la définition de cas, indépendamment de l'évolution de la maladie (cas récents et anciens confondus).
- La notion d'instantanéité est à prendre dans un sens large. L'instant correspond à l'ensemble de la période pendant laquelle est menée l'enquête. Pour obtenir une mesure pertinente, il faut que la durée de l'enquête soit négligeable par rapport à la durée de la maladie. Pour cette raison, la prévalence est un indicateur plus volontiers utilisé dans l'étude des maladies chroniques (exemple 15.5).

Exemple 15.5. PRÉVALENCE

La toxoplasmose est une infection qui persiste définitivement après la contamination. Lors d'une enquête exhaustive réalisée en France sur l'ensemble des 13 485 femmes enceintes ayant accouché pendant la dernière semaine de janvier 1995, on a observé que 7 322 d'entre elles possédaient des anticorps résiduels contre la toxoplasmose. La prévalence P = 7 322/13 485 = 0,543.

Résultat : la prévalence de la toxoplasmose chez les femmes enceintes, en France, fin janvier 1995, était de 54,3 %.

2. Incidence

L'incidence est un indicateur dynamique de morbidité. C'est un taux qui prend en compte la vitesse de survenue de la maladie dans une population. Au numérateur de l'incidence, figure le nombre de ***nouveaux cas*** d'une maladie apparus pendant une période de temps donnée. Selon la période de temps considérée, on distingue plusieurs façons de calculer l'incidence.

a) Incidence cumulée

L'incidence cumulée est le rapport du nombre de nouveaux cas d'une maladie survenue pendant une période de temps déterminée divisée par la population à risque de développer la maladie pendant cette période. Certains utilisent simplement le terme d'incidence pour l'incidence cumulée.

$$\text{Incidence cumulée} = \frac{\text{nombre de nouveaux cas pendant une période } \Delta t}{\text{population à risque pendant la période } \Delta t}$$

- L'incidence cumulée s'exprime sous forme d'un chiffre compris entre 0 et 1 (ou pourcentage) : x cas pour cent (ou pour 1 000, 10 000, *etc.*) personnes pendant la période Δt.
- La période d'étude est en général une période de temps systématique. Le choix de la période dépend de la dynamique générale de la maladie. On calcule des incidences cumulées annuelles, mensuelles, hebdomadaires, ou même journalières dans certains cas.
- L'incidence cumulée de la maladie, est calculée sur une population. Elle est équivalente au **risque** moyen de contracter la maladie pendant la période étudiée pour un individu quelconque de cette population.

Condition d'utilisation de l'incidence cumulée

- La période de mesure doit être impérativement précisée (sinon le chiffre fourni n'a aucun sens).
- En toute rigueur, la taille de la population à risque est celle du début de la période d'étude. Le calcul de l'incidence cumulée suppose, que la population reste ***stable*** pendant toute la période d'étude et que tous les sujets soient suivis de façon identique, sans perdus de vue. C'est le cas des cohortes (*cf.* chap. 16.III). Dans le cas de populations dynamiques ouvertes, lorsqu'on dispose des chiffres de population en début et en fin d'étude, on peut utiliser comme dénominateur la moyenne de ces deux chiffres à condition qu'ils soient de même ordre de grandeur. Lorsque la population d'étude est très instable pendant la période du temps d'étude ou lorsqu'il existe un nombre élevé de perdus de vue, il faut utiliser un autre mode de calcul du dénominateur (*cf.* densité d'incidence).

Exemple 15.6. INCIDENCE CUMULÉE

Au Kenya, on a enregistré en 1994 un total de 6 100 000 nouveaux cas de paludisme. La population était de 29 300 000 habitants. En supposant la population stable pendant cette année-là, l'incidence cumulée du paludisme a été de 6,1/29,3 = 0,208 soit 20,8 cas pour 100 habitants.

b) Taux d'attaque

Le taux d'attaque a la même signification qu'une incidence cumulée. Il s'utilise plus volontiers en cas d'épidémie. La différence avec l'incidence cumulée réside dans la période de mesure qui n'est pas une période systématique, mais qui dépend de la durée de l'épidémie. La population d'étude est le plus souvent restreinte à une population définie comme population à risque par l'étude descriptive des cas. Le taux d'attaque est l'incidence cumulée de la maladie pendant la durée de l'épidémie (exemple 15.7).

Exemple 15.7. TAUX D'ATTAQUE

Dans une maison d'arrêt contenant 300 prisonniers, on a observé la survenue de 21 cas de trichinellose entre le 13 et le 25 août 1985. Aucune admission, ni sortie n'a été enregistrée pendant cette période.

Taux d'attaque de la trichinellose : 21/300 = 0,07.

Résultat : le taux d'attaque de l'épidémie de trichinellose d'août 1985 parmi les prisonniers a été de 7 %.

c) Densité d'incidence

Termes équivalents : taux de densité d'incidence, taux d'incidence.

Lorsque la population est très instable (nombreuses arrivées et départs) ou lorsqu'il existe de nombreux perdus de vue pendant la période d'étude, le dénominateur utilisé pour l'incidence cumulée est impropre. Il est nécessaire de tenir compte des variations intermédiaires de ce dénominateur.

On utilise alors le concept de personnes-temps. On divise la période d'étude en sous-périodes pour lesquelles on dispose de données démographiques sur la population suivie.

Par exemple, si la période d'étude dure un an et si on possède des données démographiques mensuelles sur la présence des sujets dans la zone d'étude, on utilisera le terme de personnes-mois. Ainsi, un sujet suivi pendant une période de 12 mois comptera pour 12 personnes-mois. Un sujet suivi pendant 3 mois comptera au dénominateur pour 3 personnes-mois. Un sujet suivi 12 mois, mais tombé malade au sixième mois comptera pour 6 personnes-mois. Lorsqu'il guérit, il est à nouveau comptabilisé dans le dénominateur. S'il guérit et s'il est immunisé, il ne doit plus être comptabilisé.

Le dénominateur est la somme des personnes-temps. Dans ce mode de calcul, on constate que tous les individus ne « pèsent » pas le même poids. Un sujet suivi pendant une période longue a plus de chance de développer la maladie et d'être détecté qu'un sujet faisant une brève apparition dans la population suivie. Cette probabilité plus élevée est compensée par un accroissement proportionnel du dénominateur.

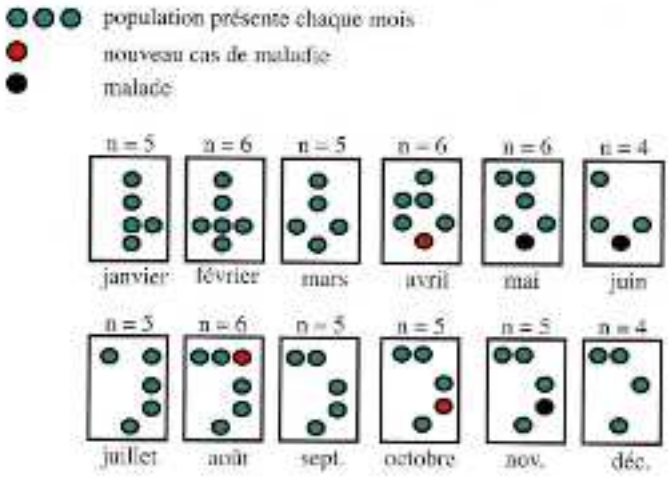
On appelle densité d'incidence le rapport du nombre de nouveaux cas divisé par la somme des personnes-temps, à risque de développer la maladie.

$$\text{Densité d'incidence} = \frac{\text{nombre de nouveaux cas pendant une période } \Delta t}{\text{somme des personnes – temps pendant la période } \Delta t}$$

- La densité d'incidence s'exprime par un nombre de nouveaux cas pour x personnes-temps.
- L'expression personnes-temps dépend de la durée de suivi et de la durée d'incubation des maladies :
  - personnes-années dans les longues enquêtes de pathologie professionnelle qui durent plusieurs années voire plusieurs dizaines d'années ;
  - personnes-mois, personnes-semaine ;
  - personnes-jour dans les crises aiguës : épidémies brutales et massives.
- La densité d'incidence permet une mesure plus précise que l'incidence cumulée lorsque la proportion de perdus de vue est élevée dans une population d'étude. Son calcul nécessite de disposer de données démographiques pour tous les segments de temps utilisés.
- La condition majeure d'utilisation de la densité d'incidence impose donc de connaître la population à risque et le nombre de cas survenus pour chaque période de temps.
- La densité d'incidence n'est pas une proportion, mais un taux. Si la maladie peut être contractée plusieurs fois par un même individu pendant la période d'étude, cet individu est comptabilisé en autant de « cas » au numérateur. Au dénominateur, il n'est comptabilisé en personnes-temps que pendant les périodes où il est « susceptible » c'est-à-dire à risque de développer la maladie.

Exemple 15.8. DENSITÉ D'INCIDENCE

La figure 15.1 représente la population d'une unité géographique recensée mois par mois. Ce recensement permet de calculer la densité d'incidence en utilisant au dénominateur des personnes-mois.



Dénominateur : 5 + 6 + 5 + 6 + (6 – 1) + (4 – 1) + 5 + 6 + 5 + 5 + (5 – 1) + 4 = 59 personnes-mois

Densité d'incidence = 3/59 = 0,051 soit 5,1 cas pour 100 personnes-mois

Figure 15-1.

3. Risque de maladie

Le risque de tomber malade pendant une période donnée est mathématiquement équivalent à l'incidence cumulée de la maladie pendant cette période.

$$R = I_c$$

On utilise donc souvent en pratique un terme pour l'autre. Notamment lors des enquêtes étiologiques. Il existe cependant une différence sémantique qu'il faut garder à l'esprit :

- une incidence est une mesure de fréquence réalisée dans le passé sur un groupe de sujet ;
- un risque est une mesure de probabilité de tomber malade pour un sujet donné.

Le risque est donc une probabilité qui projette sur l'avenir une mesure d'incidence effectuée dans le passé. Assimiler un risque probable à une incidence mesurée suppose donc :

- que les conditions de transmission de la maladie ne changent pas ;
- que l'individu auquel s'applique ce risque soit un individu théorique « moyen » représentatif de l'ensemble des individus de la population étudiée. Le risque réel d'un individu particulier qui dépend de toutes ses composantes personnelles n'est pas calculable.

L'emploi du terme « risque » est donc particulièrement délicat. Il faut distinguer le risque calculé (donc théorique), le risque réel pour un individu (non calculable) et le risque ressenti subjectivement. Ce risque ressenti (ou perçu) par le public est fort éloigné du risque calculé par un épidémiologiste.

4. Relation entre incidence et prévalence

Incidence et prévalence sont liées par la durée de la maladie.

Si on appelle Pr : la prévalence ;
I_d : la densité d'incidence ;
d : la durée de la maladie :

$$Pr = \frac{I_d d}{1 + I_d d}$$

5. Relation entre risque de maladie et densité d'incidence

Nous avons vu que le risque moyen de maladie pour un individu était assimilable à l'incidence cumulée. Si on dispose d'un taux de densité d'incidence, et si l'on suppose que ce taux est constant pendant la période d'étude, on peut estimer l'incidence cumulée, donc le risque par la relation suivante :

Si on appelle Δt : la période d'étude ;
I_c : l'incidence cumulée pendant la période Δt (risque) ;
I_d : la densité d'incidence ;

$e = 2{,}71828\ldots$:

$$I_c = 1 - e^{-Id\Delta t}$$

L'incidence cumulée ainsi estimée est une proportion comprise entre 0 et 1. Elle est égale au risque de contracter la maladie pendant la période d'étude.

6. Mortalité globale

La mortalité est un indicateur dynamique. Elle est similaire à une incidence dont l'événement étudié n'est plus la survenue d'une maladie, mais le décès. Comme l'incidence, la mortalité peut être mesurée de façon cumulative sur une population suivie pendant une période donnée (proportion) ou sous forme de taux dont le dénominateur est exprimé en personnes-temps.

$$\textbf{Mortalité globale} = \frac{\text{nombre de décès pendant une période } \Delta t}{\text{population étudiée pendant la période } \Delta t}$$

$$\textbf{Taux brut de mortalité} = \frac{\text{nombre de décès pendant une période } \Delta t}{\text{somme des personnes – temps pendant la période } \Delta t}$$

Un taux de mortalité s'exprime en nombre de décès pour x personnes-temps.
Évidemment, le taux de mortalité est dépendant de la structure par âge de la population considérée. Les données de mortalité sont le plus souvent stratifiées par classe d'âge.

7. Mortalité spécifique

La mortalité spécifique est :

- Soit un taux de mortalité dû à une pathologie particulière. Elle se distingue de la mortalité brute par son numérateur où ne figurent que les décès dus à une cause particulière.

Mortalité spécifique pour une cause x

$$\frac{\text{nombre de décès dus à cette cause pendant une période } \Delta t}{\text{population étudiée pendant la période } \Delta t}$$

Taux spécifique de mortalité pour une cause x

$$\frac{\text{nombre de décès dus à cette cause pendant une période } \Delta t}{\text{somme des personnes – temps pendant la période } \Delta t}$$

Si on dispose des taux spécifiques de mortalité pour toutes les causes de décès, leur somme est égale au taux brut de mortalité.

- Soit un taux de mortalité dans un sous-groupe particulier. Les sous-groupes particulièrement étudiés sont les tranches d'âge.

Mortalité spécifique pour une classe d'âge donnée

$$\frac{\text{nombre de décès dans une classe d'âge pendant une période } \Delta t}{\text{population de la classe d'âge étudiée pendant la période } \Delta t}$$

Taux spécifique de mortalité pour une classe d'âge donnée

$$\frac{\text{nombre de décès dans une classe d'âge pendant une période } \Delta t}{\text{somme des personnes – temps de la classe d'âge pendant la période } \Delta t}$$

On peut combiner les taux spécifiques dus à une cause donnée et selon les classes d'âge.

8. Risque de décès

Dénomination équivalente : probabilité de décès, risque moyen de décès.
Dans les tables de survie, pour calculer les espérances de vie, il est nécessaire de disposer des risques de décès pour chaque classe d'âge. Ces risques sont équivalents à la mortalité pour chaque classe.
À condition de choisir une classe d'âge suffisamment étroite pour supposer que le risque de décès ne varie pas dans cette classe d'âge, on peut calculer la relation entre risque de décès et taux brut de mortalité.

Δt : amplitude des classes d'âge.
TBM_x : taux brut de mortalité pour la classe d'âge x.
q_x : risque de décès pour la classe d'âge x.
$e = 2{,}71828\ldots$

$$q_x = 1 - e^{-TBM_x \Delta t}$$

9. Mortalité proportionnelle

La mortalité proportionnelle représente la part des décès dus à une cause donnée sur l'ensemble de tous les décès observés pendant une période donnée.

Mortalité proportionnelle liée à une cause x

$$\frac{\text{nombre de décès dus à une cause donnée}}{\text{nombre total de décès dans la population étudiée}}$$

La mortalité proportionnelle s'exprime par un nombre compris entre 0 et 1 ou par un pourcentage. Elle représente le poids relatif d'une maladie dans la mortalité globale (exemple 15.9).

Exemple 15.9. MORTALITÉ PROPORTIONNELLE

La mortalité proportionnelle due au paludisme dans la population illustrée sur la figure 15.2 est de 16 %.

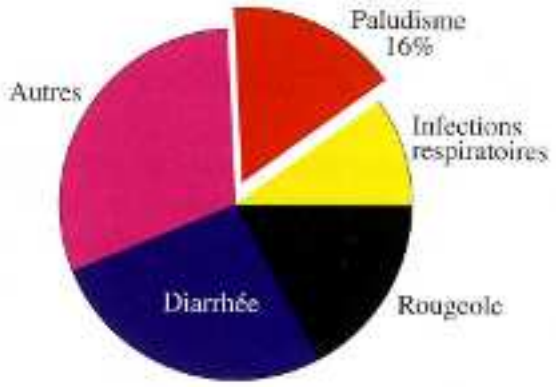


Figure 15-2. Mortalité proportionnelle due au paludisme chez les réfugiés mozambicains de moins de 5 ans, Malawi, 1987-1989

10. Létalité

La létalité est un indicateur statique. Elle représente la part des décès dus à une maladie donnée parmi les malades atteints de cette maladie.

$$\text{Létalité} = \frac{\text{nombre de décès dus à une maladie}}{\text{nombre de patients atteints par cette maladie}}$$

La létalité est donc une proportion qui s'exprime par un nombre compris entre 0 et 1 ou par un pourcentage.
La létalité est un indicateur témoignant de la gravité de la maladie et de la qualité des soins (exemple 15.10).

Exemple 15.10. MORTALITÉ ET LÉTALITÉ

Dans un camp de réfugiés de 18000 personnes, on a observé 184 décès pendant l'année 2000. Dans cette population, 12900 cas de paludisme ont été observés. Parmi les décès, 44 étaient dus au paludisme. On suppose que la population du camp est restée stable.

- La mortalité brute en 2000 est de 184/18000 = 0,0102 soit 10,2 décès pour 1000 personnes.
- La mortalité spécifique due au paludisme est de 44/18000 soit 2,4 décès dus au paludisme pour 1000 personnes.
- La mortalité proportionnelle due au paludisme est de 44/184 = 0,239 soit 23,9 %.
- La létalité due au paludisme est de 44/12900 = 0,0034 soit 0,34 % des cas.

Exercice

On dispose des données suivantes concernant le paludisme pour 4 régions obtenues sur une période d'une année.

	région 1	région 2	région 3	région 4
Population	125 254	15 987	25 789	31 313
Cas de paludisme	4 569	1 749	487	524
Nombre total de décès	2 453	556	900	1 025
Nombre de décès dus au paludisme	569	217	152	63

Calculer pour chaque année et pour chaque région :

1) l'incidence du paludisme pour 1 000 habitants ;

2) la mortalité brute pour 1 000 habitants ;

3) la mortalité spécifique pour 1 000 habitants ;

4) la mortalité proportionnelle du paludisme en % ;

5) la létalité du paludisme en %.

Comparez et commentez les résultats.

Résumé

PRINCIPAUX INDICATEURS ÉPIDÉMIOLOGIQUES	
Prévalence	nombre de cas/population
Incidence cumulée	nombre de nouveaux cas/population
Taux d'attaque	nombre de nouveaux cas (épidémie)/population
Densité d'incidence	nombre de nouveaux cas/personnes-temps
Mortalité globale	nombre de décès/population
Taux brut de mortalité	nombre de décès/personnes-temps
Mortalité spécifique	nombre de décès pour 1 maladie ou 1 strate/population
Taux spécifique de mortalité	nbre de décès pour 1 maladie ou 1 strate/personnes-temps
Mortalité proportionnelle	nombre de décès pour une maladie/nombre total de décès
Létalité	nombre de décès pour une maladie/nombre de cas

La notion de « nouveaux » cas ou de décès s'entend pendant une période donnée.
Le dénominateur représente la population à risque pendant cette même période.

16

ENQUÊTES ÉPIDÉMIOLOGIQUES

Par opposition au processus de surveillance des maladies qui implique un suivi permanent, les enquêtes épidémiologiques sont des études établies de façon ponctuelle afin de répondre à une question conjoncturelle.

I. PROTOCOLE D'ENQUÊTE

Quel que soit son type, une enquête doit faire l'objet d'un protocole initial décrivant les différentes étapes du travail. Le protocole doit comporter :

- Le contexte de l'étude et sa justification.
- La définition des objectifs généraux et spécifiques de l'étude.
- La méthodologie utilisée : type d'étude.
- La définition de la population d'étude.
- La définition des cas.
- La définition des variables étudiées.
- Le questionnaire.
- Le mode de sélection de l'échantillon étudié : plan de sondage.
- Le mode de collecte des données sur les individus.
- Les modes de saisie et d'analyse : support d'information, logiciels.
- Le plan d'analyse.
- Les méthodes statistiques utilisées.
- Les aspects éthiques, les modes d'information du public concerné.
- Les modes de communication des résultats.
- Le calendrier des tâches.
- Les références documentaires.
- Les institutions et personnels responsables.
- Le budget et les modes de financement.

1. Plan d'analyse

Parmi toutes les étapes précédentes, le plan d'analyse est une partie essentielle. Il consiste à prévoir dans le détail la façon dont les données et les résultats seront structurés et analysés. Cette pratique permet de circonscrire le travail aux thèmes choisis dans les objectifs et d'éviter le traitement de données inutiles. C'est la raison pour laquelle il est impératif de travailler sur le plan d'analyse avant de rédiger le questionnaire.

Un plan d'analyse doit :
- poser les différentes hypothèses soulevées par les objectifs de l'étude ;
- définir les indicateurs qui serviront à vérifier ces hypothèses ;
- définir les variables permettant de mesurer concrètement ces indicateurs ;
- estimer le nombre de sujets nécessaire pour atteindre les objectifs du travail ;
- planifier les analyses comportant :
  - la structure de l'échantillon et la mesure de sa représentativité,
  - les tris à plat et les mesures des fréquences des variables,
  - les croisements de variables,
  - les calculs statistiques.

Remarque importante

Par principe, les hypothèses doivent être posées *a priori* au moment de cette étape de réflexion. L'étude est réalisée pour les vérifier et doit s'en tenir là. Cette démarche implique qu'on ne doit pas tester de nouvelles hypothèses qui seraient générées au vu des résultats obtenus lors de l'exploitation de l'enquête. Si des résultats inattendus étaient constatés, il faudrait alors proposer une nouvelle enquête pour vérifier ces hypothèses.

2. Questionnaire

Son élaboration s'effectue au décours du plan d'analyse. Il est composé d'une série de questions visant à recueillir toutes les variables définies comme étant nécessaires à l'analyse. Quel que soit son type, un questionnaire doit toujours être testé avant d'être appliqué en situation réelle.

Le questionnaire comporte des éléments concernant :
- l'enquêteur, le mode de recueil des données, les lieux et heures de recueil ;
- le sujet étudié : identification, localisation, code d'anonymat, *etc.* ;
- la qualité de la réponse : notamment refus, absence, perte de prélèvement, *etc.* ;
- les instructions à l'enquêteur pour chaque question ;
- le recueil des données proprement dites, comportant :
  - des questions fermées à choix simple ou choix multiple déjà codées,
  - des questions ouvertes, codées plus tard par les responsables de la saisie,
  - les mesures réalisées par l'enquêteur ou notées dans des registres.

Le recueil des données se fait encore très souvent sur un support papier. Les données sont transcrites secondairement sur ordinateur. La pratique de questionnaire informatisé où la saisie s'opère directement en machine permet d'éviter l'étape de saisie secondaire, source d'erreurs supplémentaires.

3. Définition des cas

Un cas est un malade qui remplit des critères définis préalablement à l'enquête, au moment du protocole. Un cas représente l'unité statistique de l'étude. Certains malades, bien qu'ayant fait l'objet d'un diagnostic et justifiant d'un traitement médical, peuvent ne pas être considérés comme des cas

d'un point de vue épidémiologique, s'ils ne présentent pas les critères imposés par l'investigation. Ces critères sont appelés **critères d'inclusion**. Par opposition, on peut également définir des **critères d'exclusion** qui interdisent d'inclure un cas si ils sont présents.
La définition d'un cas consiste à dresser une liste de critères d'inclusion ou d'exclusion. Il s'agit d'une définition opérationnelle n'ayant pas la prétention de définir tout ce qui constitue un cas d'un point de vue clinique. Elle doit privilégier la simplicité sur l'exhaustivité.
Les critères de définition d'un cas doivent être impérativement fixés dès le début de l'enquête après une réflexion critique. En principe, ils ne doivent pas être changés en cours d'investigation.
Un cas est défini par deux types de critère : ceux qui déterminent l'appartenance des cas à une population donnée en terme de temps, lieux et personnes et ceux qui résument une symptomatologie clinique et biologique.

a) Critères temps-lieux-personnes

Toute définition de cas comprend :

- *des limites temporelles* : on sélectionne tous les cas ayant présenté des symptômes pendant une période déterminée ;
- *des limites territoriales* : il faut fixer aux enquêteurs la zone géographique de recherche les cas ;
- une définition de la *population d'étude* : population générale ou sous-population particulière : enfants, femmes enceintes, hommes adultes, *etc.*

b) Critères clinico-biologiques

Ces critères sont à l'usage des investigateurs de terrain qui vont les utiliser pour inclure ou exclure des malades. Les investigateurs sont rarement des spécialistes de la maladie en cause. Les critères comportent en général l'association de plusieurs signes ou symptômes. Ils doivent être établis de façon précise, claire et non biaisée.

- ***Précision :*** lorsqu'il s'agit de symptômes cliniques, on tente dans la mesure du possible de quantifier leur intensité. Lorsqu'il s'agit de signes biologiques, on fixe le seuil au-delà ou en deçà duquel on décide d'inclure le cas.
- ***Clarté :*** les associations de signes et symptômes doivent être libellées sous forme de liens utilisant des opérateurs logiques : ET, OU, absence de.

Si, par exemple, on définit comme cas, un malade présentant les signes A **et** B, cela signifie pour l'enquêteur qu'un malade sera inclus seulement s'il présente à la fois le signe A et le signe B. Un malade ne présentant qu'un seul des symptômes ne sera pas inclus.
À l'inverse, si on définit comme cas, un malade présentant les signes A **ou** B, cela signifie pour l'enquêteur qu'un malade sera inclus s'il présente soit le signe A, soit le signe B, soit les deux.
Lorsque plusieurs signes sont possibles, mais non suffisants, on choisit la solution de définir comme cas, un malade présentant au moins X signes parmi une liste de signes.

- ***Absence de biais :*** une définition ne doit contenir dans ses termes que des critères portant sur la maladie. Elle ne doit pas contenir de termes faisant référence à une éventuelle exposition à un facteur de risque. Même si un malade ne paraît pas devoir être rattaché à la cause que l'on recherche, il doit être inclus, s'il présente les critères clinico-biologiques d'inclusion.

	Définition incorrecte	Définition améliorée
imprécision :	fièvre	fièvre > 39°
	hyperéosinophilie	éosinophilie > 1 000 c/mm^3
	diarrhée	plus de 3 selles liquides par jour
ambiguïté :	œdème, myalgies	œdème ET myalgie
	œdème et/ou myalgie	œdème OU myalgies
biais :	syndrome + consommation d'un aliment suspect	syndrome exclusivement

c) Qualités d'une définition

Les conséquences d'une définition de cas doivent être évaluées avant de l'appliquer sur le terrain. Trois paramètres doivent être évoqués : la faisabilité, la sensibilité et la spécificité.

- **Faisabilité**. La recherche des cas s'effectue le plus souvent en collectivité humaine, en dehors du milieu hospitalier. Les examens exigés doivent donc être simples, faciles à exécuter, peu coûteux et acceptables par les individus. Il faut donc exclure d'une définition de cas les résultats d'examens trop sophistiqués, même s'ils sont indispensables au diagnostic individuel.
- **Sensibilité** d'une définition. Une définition sensible permet de retrouver le plus grand nombre possible de malades. Son inconvénient est d'inclure éventuellement des patients atteints d'autres pathologies ayant des signes communs avec ceux de la définition. Si ce nombre est trop élevé, la définition est trop sensible.
- **Spécificité** d'une définition. Une définition spécifique permet d'affirmer avec une plus grande certitude que le cas est un vrai malade. Une définition trop spécifique risque de limiter trop fortement le nombre de cas et de négliger un grand nombre de cas utiles.

Comme on le constate, ces 3 paramètres sont contradictoires. L'établissement d'une définition de cas doit donc peser les divers avantages et inconvénients des combinaisons envisagées.
Une façon pratique de combiner les 2 termes antagonistes de sensibilité et de spécificité est de créer plusieurs niveaux de définition. Par exemple :

- cas certain : définition très spécifique ;
- cas probable : définition intermédiaire ;
- cas suspect : définition très sensible.

II. TYPES D'ENQUÊTES

On distingue selon leur objet deux types d'enquête : les enquêtes descriptives et les enquêtes étiologiques.

1. Enquêtes descriptives

Les enquêtes descriptives sont mises en œuvre afin de recueillir des données sur l'état d'une maladie à un moment précis. L'indicateur mesuré dans ce type d'enquête est la prévalence. Une enquête descriptive peut être considérée comme une photographie instantanée de la situation épidémiologique. On pourrait lui opposer un système de surveillance continue, qui mesure l'incidence par tranche de temps. L'enquête descriptive peut être soit exhaustive (réalisée sur l'ensemble de la population d'intérêt), soit effectuée sur un échantillon représentatif de cette population (*cf.* sondage, chap. 7).

- La mise en œuvre d'une enquête descriptive s'effectue lorsqu'on désire mesurer l'amplitude d'un phénomène de santé. C'est le cas lorsqu'on désire :
  - décider la mise en place d'un programme de santé publique ;
  - évaluer l'impact d'un programme de santé publique ;
  - vérifier une rumeur avant d'entreprendre une investigation d'épidémie ;
  - surveiller une maladie de façon discontinue.

- Exemple d'enquêtes descriptives :
  - enquête de séroprévalence du paludisme ;
  - enquête nutritionnelle ;
  - enquête de couverture vaccinale.

2. Enquêtes étiologiques

Plus ambitieuse qu'une simple enquête descriptive, une enquête à visée étiologique cherche à établir une relation spécifique entre la survenue d'une maladie et des facteurs de risque.
Contrairement aux études expérimentales et aux essais thérapeutiques, les enquêtes étiologiques sont des enquêtes d'observation. Ici, l'investigateur n'a aucun rôle dans l'attribution des facteurs étudiés. Seul, les propriétés personnelles de chaque individu, les circonstances, les aléas de sa vie, le prédisposent à être ou non exposé et à en subir les éventuelles conséquences. L'investigateur n'intervient que dans le choix des groupes à comparer.
Pour cette raison majeure, une enquête étiologique ne peut jamais démontrer une relation causale. Tout au plus aboutit-elle à offrir un maximum de présomptions en faveur de cette relation.
Les enquêtes étiologiques sont très largement utilisées tant en santé publique qu'en médecine expérimentale.

Distribution des sujets dans une enquête étiologique

Une enquête étiologique a pour objectif de comparer des malades et des non-malades selon leur niveau d'exposition à un ou plusieurs facteur de risque. La population étudiée se partage donc entre sujets définis par leur statut vis-à-vis de la maladie, et par leur statut vis-à-vis de l'exposition.
Le terme de maladie est pris ici dans un sens très large d'événement pathologique. Cela peut être une maladie précise, le décès, une complication de maladie. Les cas sont les sujets présentant l'événement pathologique étudié. Les sujets témoins ou « sains » peuvent être tout à fait sains mais aussi des malades ne présentant pas l'événement : survivant, forme clinique simple, *etc.*
De façon schématique et par convention, nous conviendrons de présenter les résultats de toute enquête étiologique en deux colonnes, les malades à gauche, et les non-malades à droite, et, si l'exposition est dichotomique (Oui, Non) en deux lignes, la ligne supérieure comportant les sujets exposés, et la ligne inférieure les sujets non-exposés.

Présentation schématique des résultats d'une enquête étiologique

	Cas	Sains
Exposés	a	b
Non-exposés	c	d

On distingue plusieurs types de méthodologie d'enquête selon la chronologie du recueil des données et le type de comparaison effectuée :

- les enquêtes de cohorte ;
- les enquêtes cas-témoins ;
- les enquêtes transversales.

III. ENQUÊTES DE COHORTE

1. Principe

Une **cohorte** se définit comme un groupe de sujets **suivis dans le temps** (figure 16.1). Si l'événement observé est la survenue d'une maladie, on mesure, à la fin de l'étude, le nombre de sujets atteints par la maladie pendant la période d'étude. Ce nombre divisé par la taille du groupe est par définition ***l'incidence*** de la maladie pendant la période d'étude.

Une **étude de cohorte** est une étude comparant plusieurs cohortes (figure 16.2). Dans le schéma le plus simple, une étude de cohorte se résume à comparer une cohorte de sujets exposés à une cohorte de sujets non-exposés. À l'issue d'une étude de cohorte, on compare donc le taux d'incidence entre exposés (I_e) et non-exposés (I_{ne}). Une étude de cohorte est parfois appelée étude exposés/non-exposés.

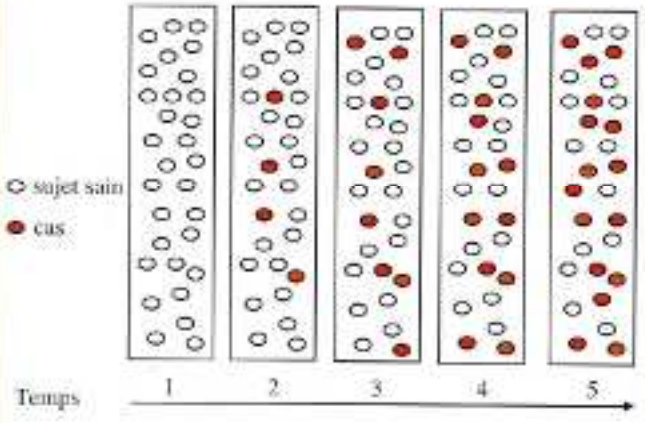


Figure 16-1. Cohorte

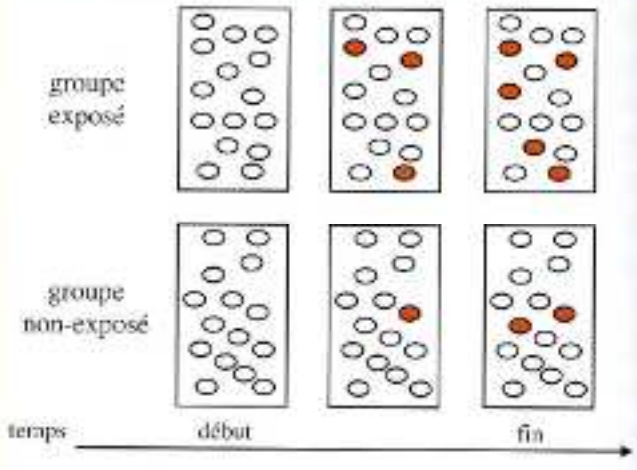


Figure 16-2. Étude de cohorte

2. Présentation des données

- **Cohorte à un facteur d'exposition :** figures 16.3 et 16.4

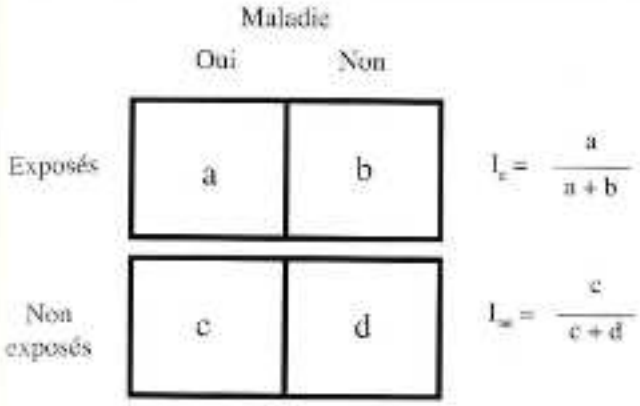


Figure 16-3. Étude de cohorte : présentation des données (1)

	Population à risque	Cas	Taux d'incidence
Exposés	N_e	a	$I_e = \frac{a}{N_e}$
Non exposés	N_{ne}	c	$I_{ne} = \frac{c}{N_{ne}}$

Figure 16-4. Étude de cohorte : présentation des données (2)

- **Cohorte à plusieurs niveaux d'exposition :** figure 16.5

Niveau d'exposition	Population à risque	Cas	Taux d'incidence
Elevé	N_1	a_1	$I_{e1} = a_1 / N_1$
Moyen	N_2	a_2	$I_{e2} = a_2 / N_2$
Bas	N_3	a_3	$I_{e3} = a_3 / N_3$
Non exposés	N_{ne}	c	$I_{ne} = c / N_{ne}$

Figure 16-5. Étude de cohorte : présentation des données (3)

3. Mesures dans une enquête de cohorte

Une étude de cohorte permet de calculer :

- Des ***taux d'incidence*** dans chaque groupe de comparaison. Ces taux d'incidence sont assimilés aux probabilités ou **risques** de survenue de la maladie.
- ***La différence de risque*** entre exposés et non exposés : $DR = I_e - I_{ne}$
- ***Le risque relatif***, en calculant le rapport entre l'incidence chez les exposés sur l'incidence chez les non-exposés. Cet indicateur est appelé **risque relatif ou ratio de risque ou rapport de risque (RR).** Il est utilisé très fréquemment en recherche étiologique.

4. Le risque relatif

$$RR = \frac{I_e}{I_{ne}}$$

Les risques I_e et I_{ne} étant des valeurs comprises entre 0 et 1, RR est un nombre sans unité compris entre 0 et l'infini. Un risque relatif « nul » a pour valeur 1. Plus RR est éloigné de 1 (supérieur ou inférieur) plus l'association entre la survenue de la maladie et la présence du facteur étudié est forte.

Intervalle de confiance du risque relatif

Une enquête de cohorte est rarement réalisée sur l'ensemble d'une population à risque d'une maladie donnée. Elle est effectuée sur un échantillon représentatif de cette population. Le RR est donc une variable aléatoire qui subit des fluctuations d'échantillonnage. On calcule donc un intervalle de confiance à 95 % du RR (IC 95 %).

Il existe des formules complexes pour calculer les intervalles de confiances du RR. En pratique, on utilise les logiciels statistiques (EpiInfo6/Epitable/Analyse/Étude de cohorte). Une méthode simple pour effectuer ce type de calcul (mais approchée) est la méthode de Miettinen : sa formule figure en Annexes, Formulaire § 15.

Interprétation d'un risque relatif

- Si RR = 1 (la valeur 1 est comprise entre les bornes de l'IC 95 %), cela signifie que l'on n'a pas détecté d'excès de risque dans le groupe exposé. Il n'y a pas de relation démontrée entre la maladie et l'exposition au facteur étudié.
- Si RR est significativement supérieur à 1 (borne inférieure de l'IC 95 % > 1), cela signifie qu'il existe un excès de risque dans le groupe exposé. Il y a donc une relation entre l'exposition au facteur étudié et la survenue de la maladie. Le facteur peut être considéré comme un **facteur de risque**. On conclut en affirmant que si un sujet est exposé, le risque de contracter la maladie est RR fois supérieur que s'il n'était pas exposé.

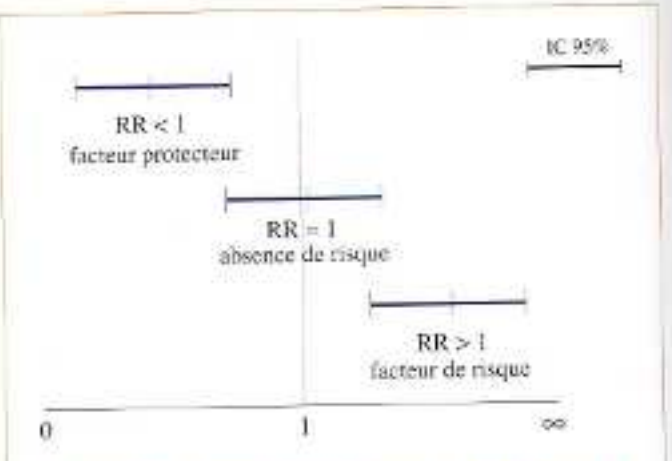


Figure 16-6. Interprétation d'un risque relatif

- Si RR est significativement inférieur à 1 (borne supérieure de l'IC 95 % < 1), cela signifie qu'il existe un risque moindre de contracter la maladie s'il y a exposition au facteur. Ce facteur peut être considéré comme un **facteur protecteur**. Un RR égal à 0,1 pour un facteur protecteur est l'équivalent d'un RR égal à 10 pour un facteur de risque.

Comparaison RR et χ^2

Dans une enquête de cohorte, on pourrait comparer les taux d'incidence par un simple test de χ^2. Ce test permettrait d'affirmer qu'il existe une différence significative entre les incidences

observées. L'intervalle de confiance du risque relatif a la même signification que le résultat p d'un test de χ^2.

- Si p est supérieur à 0,05 (test non significatif), alors, l'IC 95 % du RR contient la valeur 1. Le test n'est pas significatif, le facteur étudié ne ressort pas comme facteur de risque.
- Si p est inférieur à 0,05, alors l' IC 95 % exclut la valeur 1.

L'intérêt du calcul du RR et de son intervalle de confiance est de donner la **force, le sens et le degré de signification** de l'association, alors que le test de χ^2 ne donne que le **degré de signification** de l'association.

Valeur ponctuelle du RR	: force de l'association
Position par rapport à 1	: sens de l'association
Intervalle de confiance du RR	: degré de signification
Test du χ^2	: degré de signification

Cohorte à plusieurs niveaux d'exposition (*cf.* figure 16.5)

Dans ce cas, on calcule le risque relatif, pour chaque niveau d'exposition. Le dénominateur (niveau de référence) est soit le risque chez les non-exposés, soit le risque du plus faible niveau d'exposition.

Pour chaque niveau d'exposition on a :

$$RR_i = \frac{\text{Incidence de la strate i}}{\text{Incidence de la strate de référence}}$$

L'intervalle de confiance par la méthode de Miettinen est obtenu à l'aide du χ^2 calculé sur chaque tableau à 4 cases dont la ligne supérieure correspond à la strate i et la ligne inférieure à la strate de référence.

5. Choix du groupe de référence

Les sujets non-exposés doivent être choisis dans la même population d'où proviennent les sujets exposés. Cette obligation est parfois difficile à suivre en pratique. Il faut néanmoins tendre à respecter la comparabilité entre groupes d'exposition, sous peine d'introduire des biais (*cf.* VII).
Les définitions d'une exposition et d'une non-exposition doivent être précisées au début de l'étude. On devra prévoir les changements d'exposition en cours d'étude.

6. Nombre de sujets nécessaires à une enquête de cohorte

Il existe des formules permettant de calculer le nombre de sujets nécessaires à une étude de cohorte. La formule concernant une cohorte simple entre un groupe exposé et non-exposé figure en Annexes, Formulaire 16. En pratique, on utilise un logiciel (EpiInfo6/Epitable/Echantillonne/ Taille d'échantillon/ Etude de cohorte).
Ce nombre dépend de la puissance que l'on désire affecter à l'étude et de la différence escomptée entre les incidences dans les groupes exposés et non-exposés.

Pour aboutir à effectuer ce calcul, il faut donc faire plusieurs approximations.

- Estimer une valeur de l'incidence chez les sujets non-exposés sur des connaissances antérieures ou une enquête préalable.
- Choisir un RR minimum. Si on a une idée *a priori* de l'incidence de la maladie chez les exposés, on peut l'estimer par le rapport des incidences attendues chez les exposés et les non-exposés. Plus le RR choisi est petit, plus il faudra de sujets pour observer une valeur significative.
- Choisir un risque de première espèce α : en général 5 %.
- Choisir un risque de seconde espèce β : en général 20 %.
- Préciser, le cas échéant (EpiInfo), le ratio non-exposés/exposés.

7. Avantages et inconvénients d'une enquête de cohorte

AVANTAGES	INCONVÉNIENTS
bien adaptée pour étudier :	non adaptée pour étudier :
• les risques (incidences) • des expositions rares • plusieurs maladies • la séquence exposition-maladie	• des maladies rares • plusieurs expositions
• peu de biais de sélection • peu de biais de mémorisation	• coût élevé • longue période de latence • problèmes d'éthique

Exemple 16.1. ENQUÊTE DE COHORTE

Pour savoir si l'exposition au virus VIH est un facteur de risque de survenue de tuberculose (TB), on suit pendant 2 ans une cohorte de 215 sujets infectés par le virus du SIDA (VIH $^+$) et une cohorte de 298 sujets vivants dans les mêmes conditions, mais non infectés par le virus (VIH $^-$). Au bout de 2 ans, on note les résultats suivants :

EXPOSITION	COHORTE	CAS DE TB	INCIDENCE %	RR	IC 95 %
VIH $^+$	215	8	*3,72*	11	1,4-88
VIH $^-$	298	1	*0,34*		

L'incidence de la tuberculose dans le groupe VIH $^+$ est 11 fois plus élevée que dans le groupe VIH $^-$. La borne inférieure de l'intervalle de confiance à 95 % est supérieure à 1. On peut donc en conclure qu'il existe un lien entre l'exposition au virus VIH et la survenue de tuberculose. Le VIH est un facteur de risque de tuberculose.

IV. ENQUÊTES CAS-TÉMOINS

1. Principe

Une autre manière de mesurer une association entre présence d'un facteur de risque et survenue d'une maladie, consiste à comparer la fréquence d'exposition au facteur parmi un groupe de malades et parmi un groupe de sujets non-malades choisis comme témoins. Si la survenue de la maladie est liée à l'exposition, on doit observer un pourcentage d'exposition plus élevé chez les cas que chez les témoins.

2. Présentation des données

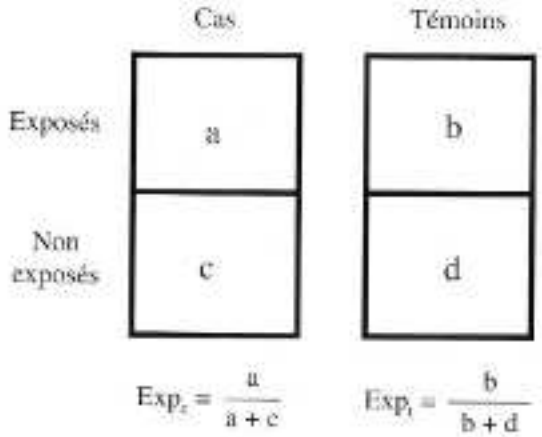


Figure 16-7. Étude cas-témoins : présentation des données

3. Mesures dans une enquête cas-témoins

Puisque c'est l'investigateur qui a choisi le nombre de témoins, il n'est pas possible dans ce type d'enquête de calculer des taux d'incidence. On peut seulement comparer les expositions au facteur de risque entre cas et témoins.

Une étude cas-témoins permet de calculer :

■ **Les fréquences d'exposition**

- La fréquence d'exposition chez les cas est égale à a/(a + c).
- La fréquence d'exposition chez les témoins est égale à b/(b + d).

La comparaison de ces fréquences est intuitivement compréhensible. Elle pourrait être effectuée par un simple test de χ^2.

■ **Les cotes d'exposition**

- La cote d'exposition chez les cas est égale à a/c.
- La cote d'exposition chez les témoins est égale à b/d.

Les cotes en tant que telles ne servent à rien. Mais leur rapport, permet d'évaluer la liaison entre l'exposition et la maladie. Ce rapport est appelé **rapport de cotes** (RC) ou **odds ratio** (OR). Ce dernier terme est le plus utilisé.

4. L'odds ratio (OR)

L'odds ration dans une enquête cas-témoins est le rapport de la cote d'exposition chez les cas sur la cote d'exposition chez les témoins.

$$OR = \frac{a/c}{b/d}$$

L'OR est un nombre sans unité compris entre 0 et l'infini. Un odds ratio « nul » a pour valeur 1. Plus l'OR est éloigné de 1 (supérieur ou inférieur), plus l'association entre la survenue de la maladie et la présence du facteur étudié est forte.

Intervalle de confiance de l'odds ratio

Par définition, les cas et les témoins sont issus d'une population. L' OR est donc une variable aléatoire qui subit des fluctuations d'échantillonnage. On calcule donc un intervalle de confiance à 95 % de l' OR (IC 95 %).

Il existe des formules complexes pour calculer les intervalles de confiances de l'OR. En pratique, on utilise les logiciels statistiques pour effectuer ce type de calcul (EpiInfo6/Epitable/Analyse/ Étude cas-Témoins). Une méthode simple (mais approchée) est la méthode de Miettinen : sa formule figure en Annexes, Formulaire § 15.

Interprétation d'un odds ratio

On démontre que lorsque la prévalence de la maladie dans la population d'origine est faible, l'odds ratio calculé dans une étude cas-témoins est un estimateur du risque relatif qui aurait pu être calculé si l'étude avait été construite comme une enquête de cohorte.

Un rapport de cote peut donc être assimilé à un risque relatif sous condition d'une prévalence faible de la maladie (inférieure à 10 %). Son interprétation est donc similaire à celle d'un risque relatif (figure 16.8).

- Si OR = 1 (la valeur 1 est comprise entre les bornes de l'IC 95 %), cela signifie que l'on n'a pas détecté une différence d'exposition entre cas et témoins. Il n'y a pas de relation démontrée entre la maladie et l'exposition au facteur étudié.
- Si OR est significativement supérieur à 1 (borne inférieure de l'IC 95 % > 1), cela signifie que la fréquence d'exposition est supérieure chez les cas que chez les témoins. Il y a donc une relation entre l'exposition au facteur étudié et la survenue de la maladie. Le facteur peut être considéré comme un **facteur de risque**. On conclut en affirmant que si le sujet est exposé, le risque de contracter la maladie est multiplié par la valeur de l'OR.

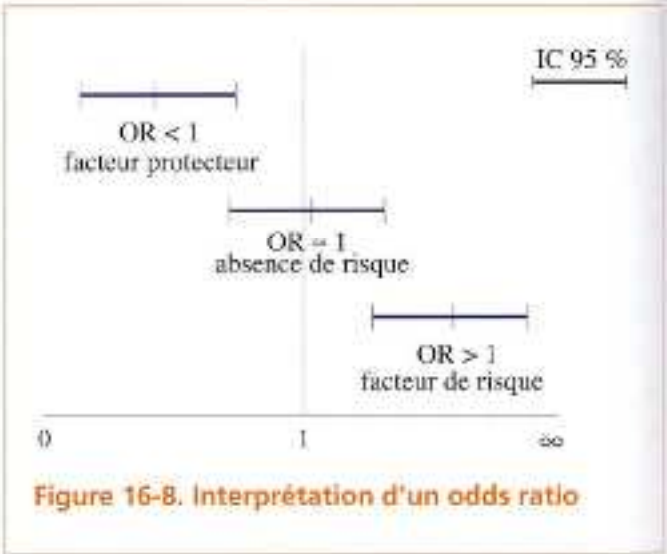


Figure 16-8. Interprétation d'un odds ratio

- Si OR est significativement inférieur à 1 (borne supérieure de l'IC 95 % < 1), cela signifie que la fréquence d'exposition est plus faible chez les cas que chez les témoins. Ce facteur peut être considéré comme un **facteur protecteur**. Un OR égal à 0,1 pour un facteur protecteur est l'équivalent d'un OR égal à 10 pour un facteur de risque.

Comparaison OR et χ^2

L'intervalle de confiance de l'odds ratio a la même signification que le résultat p d'un test de χ^2.

- Si p est supérieur à 0,05 (test non significatif), alors, l'IC 95 % du OR contient la valeur 1. Le test n'est pas significatif, le facteur étudié ne ressort pas comme facteur de risque.
- Si p est inférieur à 0,05, alors l' IC 95 % exclut la valeur 1.

L'intérêt du calcul du OR est de donner la **force, le sens et le degré de signification** de l'association, alors que le test de χ^2 ne donne que le degré de signification de l'association.

Valeur ponctuelle de l'OR	: force de l'association
Position par rapport à 1	: sens de l'association
Intervalle de confiance de l'OR	: degré de signification
Test du χ^2	: degré de signification

Enquête cas-témoins à plusieurs niveaux d'exposition

Niveau d'exposition	Cas	Témoins	OR
Élevé	a_1	b_1	a_1d/b_1c
Moyen	a_2	b_2	a_2d/b_2c
Bas	a_3	b_3	a_3d/b_3c
Non exposé	c	d	référence

On calcule l'odds ratio pour chaque niveau d'exposition. Le niveau de référence est composé des effectifs chez les non-exposés ou chez les exposés au plus faible niveau d'exposition.

Pour chaque niveau d'exposition on a : $$OR_i = \frac{a_i/c}{c_i/d}$$

L'intervalle de confiance par la méthode de Miettinen est obtenu à l'aide du χ^2 calculé sur chaque tableau à 4 cases dont la ligne supérieure correspond à la strate i et la ligne inférieure à la strate de référence.

5. Choix des témoins

Un témoin doit être un individu issu de la population d'où proviennent les cas. Il doit donc être le plus proche possible des cas, à l'exception de sa présentation clinique. Un témoin doit donc être défini avec des critères cliniques négatifs : absence de signes. Un mauvais choix du groupe témoin entraine un biais dans l'analyse (*cf.* VII).

Dans une enquête cas-témoins, les témoins peuvent être soit tirés au sort dans cette population, soit choisis dans des groupes spécifiques : famille, amis, voisins, collègues, patients du même service avec une pathologie différente.

Les informations sur l'exposition des témoins doivent être recueillies de la même manière chez les témoins et les cas, et doivent être identiques.

Lorsque à chaque cas est affecté un (ou plusieurs témoins) l'enquête est dite « appariée » (*cf.* chap. 16.IV.8).

6. Nombre de témoins par cas

Il n'est pas nécessaire que le nombre de témoins soit identique au nombre de cas. Les groupes doivent cependant être rationnellement équilibrés. Si le nombre de cas est élevé, il suffit de choisir un groupe témoin de la même taille que le groupe des cas. Si le nombre de cas est faible, on peut

augmenter la puissance de l'étude en choisissant plusieurs témoins par cas. Il n'est pas utile de dépasser cinq témoins par cas, car au-delà la puissance gagnée est dérisoire par rapport à l'augmentation de la charge de travail.

7. Nombre de sujets nécessaires à une enquête cas-témoins

Il existe des formules permettant de calculer le nombre de sujets nécessaires à une étude cas-témoins. La formule figure en Annexes, Formulaire 17. En pratique, on utilise un logiciel (EpiInfo6/ Epitable/Echantillonne/Taille d'échantillon/Etude cas-témoins).
Ce nombre dépend de la puissance que l'on désire affecter à l'étude et de la différence escomptée entre expositions. Pour aboutir à effectuer ce calcul, il faut faire plusieurs approximations.

- Choisir un OR minimum. Plus l'OR choisi est bas, plus il faudra de sujets pour observer une valeur significative.
- Choisir un risque de première espèce α : en général 5 %, soit $Z_\alpha = 1{,}96$.
- Choisir un risque de seconde espèce β : en général 20 %, soit $Z_{2\beta} = 0{,}84$.
- Estimer une valeur de la proportion de témoins exposés sur des connaissances antérieures ou une enquête préalable.
- Choisir un nombre de témoins par cas.
- Préciser, le ratio témoins/cas.

8. Enquête cas-témoins appariée

L'appariement consiste à recruter des témoins en fonction de leur similitude avec des cas (même groupe d'âge, même sexe, *etc.*). On réalise un appariement lorsqu'on désire éliminer un facteur de confusion (*cf.* chap. 16.VIII.2) ou augmenter la puissance d'une analyse.
Dans le cas d'un simple appariement (1 cas pour 1 témoin), on obtient le tableau suivant :

	TÉMOINS EXPOSÉS	TÉMOINS NON EXPOSÉS
cas exposés	e	f
cas non exposés	g	h

Dans cette situation on a : $$OR = \frac{f}{g}$$

Les données peuvent être analysées par le test de χ^2 de McNemar (chap. 13.XVII).

9. Avantages et inconvénients d'une enquête cas-témoins

AVANTAGES	INCONVÉNIENTS
bien adaptée pour étudier : • des maladies rares • plusieurs facteurs de risque	non adaptée pour étudier : • des expositions rares • plusieurs maladies • la séquence temporelle exposition-maladie
• coût faible • rapidité d'exécution • échantillons de taille modérée	• pas de calcul de taux d'incidence • biais de mémorisation et de sélection • OR biaisé si prévalence élevée

Exemple 16.2. ENQUÊTE CAS-TÉMOINS

On veut vérifier le rôle de la consommation de viande de mouton dans la survenue de la toxoplasmose chez les femmes enceintes non prémunies. On dispose d'un groupe de 80 femmes ayant contracté la toxoplasmose au cours de leur grossesse (cas) et de 80 femmes enceintes séronégatives n'ayant pas contracté la toxoplasmose (témoins). La consommation de mouton pendant la grossesse a été notée et les résultats sont exprimés sur le tableau ci-dessous :

CONSOMMATION DE MOUTON	CAS	TÉMOINS	OR	IC 95 %
Oui	55	28	4,1	2,1-7,9
Non	25	52		
% exposition	*68,7*	*35,0*		

L'exposition à la consommation de viande de mouton est près de 2 fois plus élevée chez les cas. On aurait pu vérifier si cette différence était significative par un test de χ^2 ($\chi^2 = 18,2$, $p < 0,0001$). L'odds ratio, estimateur du risque relatif, signifie que le risque de contracter la toxoplasmose est 4 fois plus élevé chez les consommatrices de mouton. La borne inférieure de l'intervalle de confiance est supérieure à 1. La consommation de viande mouton est donc un facteur de risque de toxoplasmose chez les femmes enceintes.

10. Variantes des enquêtes cas-témoins

a) Enquête cas croisés (*case cross-over study*)

Dans ce type d'enquête, on choisit comme témoins, les cas eux-mêmes. Ce procédé est utilisé notamment lors d'épidémies de toxi-infections alimentaires collectives. En tant que cas, le sujet est interrogé sur sa consommation d'un aliment suspect pendant une période présumée à risque. En tant que témoin, il est interrogé sur sa consommation pendant une période antérieure considérée comme non à risque lorsqu'il n'était pas malade et qu'il n'existait encore aucun cas. Le laps de temps entre les deux périodes étudiées doit être supérieur à la durée d'incubation de la maladie. On retrouve donc les 4 combinaisons possibles entre cas/« témoin » et exposition/non-exposition. L'analyse se fait comme celle d'une enquête cas-témoins appariée. Ce type d'étude permet de réduire considérablement le temps d'investigation en évitant la recherche de témoins.

b) Enquête cas pour cas (*case-to-case study*)

Dans ce type d'enquête, utilisé également dans l'investigation rapide d'épidémie de maladie infectieuse, on utilise comme témoins des cas considérés comme « non-épidémiques ». Il s'agit de cas dit « sporadiques » représentant le « bruit de fond » du système de surveillance et identifiés le plus souvent dans les centres nationaux de référence. Ce sont des malades qui présentent la même maladie que les cas, mais pour lesquels on a la certitude qu'ils n'ont pas été contaminés par la souche responsable de l'épidémie. Ils sont identifiés par des techniques de typage biologique (PCR, champ pulsé, *etc.*) capables de distinguer les isolats sporadiques différents de la souche commune aux cas « épidémiques ». Ce type d'étude permet, comme la précédente, d'éviter la fastidieuse recherche de témoins, puisque les « témoins-cas » sont déjà identifiés dans la base du système de surveillance.

c) Enquête cas-témoin intra-cohorte (*nested case-control study*)

Comme son nom l'indique, l'enquête cas-témoins est construite à partir des sujets inclus dans une cohorte. Le procédé consiste à tirer au sort, pour chaque survenue d'un cas, un ou plusieurs témoins parmi les sujets encore sains au moment de la survenue d'un cas. Cas et témoins sont donc appariés en fonction de la durée d'exposition. Ce type d'étude, utilisé pour répondre à des objectifs secondaires à ceux de l'étude de cohorte, a donc pour avantage d'équilibrer parfaitement les durées d'exposition entre les deux groupes.

V. ENQUÊTES TRANSVERSALES

Une enquête transversale est bâtie comme une enquête descriptive dans laquelle des données sur les éventuels facteurs de risque ont été recueillies en même temps que l'information sur le statut vis-à-vis de la maladie. Une enquête transversale aboutit donc à mesurer des différences de prévalence entre groupes exposés P_e et non exposés P_{ne}. Ces différences s'expriment en général par un rapport de prévalence (RP).
Une enquête transversale peut donc s'analyser mathématiquement comme une enquête de cohorte.

Facteur étudié	Cas	Sains	Prévalence	Rapport de prévalence
exposés	a	b	$P_e = a/(a + b)$	P_e/P_{ne}
non exposés	c	d	$P_{ne} = c (c + d)$	

Cependant, l'interprétation en est beaucoup plus limitée. Une enquête transversale ne permet de mesurer qu'une simple liaison entre facteur de risque et *présence* (et non pas *survenue*) de la maladie. Une telle enquête, ne permet pas de savoir si la maladie est apparue ***après*** ou ***avant*** exposition. Au terme d'une telle enquête on ne sait donc pas si la maladie est une conséquence ou une cause de l'exposition. Une enquête transversale n'apporte donc qu'une simple indication sur une ***éventuelle*** liaison entre le facteur étudié et la maladie. Elle permet, cependant, d'apporter des arguments lorsqu'une liaison est suspectée, afin de bâtir une nouvelle enquête de type cohorte ou cas-témoins (exemple 16.3).

Exemple 16.3. ENQUÊTE TRANSVERSALE

On veut connaître les facteurs de risque de contracter une giardiase parmi les enfants d'une crèche. Sur les 64 enfants de la crèche, 22 étaient porteurs de *Giardia intestinalis* au moment de l'enquête. Parmi les divers facteurs testés, on examine le mode de consommation hydrique.

BOISSON	CAS	SAINS	PRÉVALENCE %	RP	IC 95 %
eau du robinet	20	26	*43,5*	3,9	1,02–15,1
eau minérale	2	16	*11,1*		

La prévalence de la giardiase est 3,9 fois plus élevée chez les enfants consommant l'eau du robinet. Ce résultat n'est pas suffisant à lui seul pour affirmer le lien de causalité. Il permet néanmoins de générer une hypothèse afin de bâtir une enquête étiologique orientée vers cette cause.
Le calcul de l'IC 95 % est effectué comme pour une enquête de cohorte (EpiInfo6/Epitable/Analyse/Etude de Cohorte).

VI. CRITÈRES DE CAUSALITÉ DANS UNE ENQUÊTE ÉTIOLOGIQUE

Les enquêtes étiologiques aboutissent à mettre en cause un facteur, en affirmant que le risque de contracter la maladie est plus élevé dans le groupe exposé que chez les sujets non-exposés.
La mise en cause d'un facteur de risque doit s'entourer d'un certain nombre de précautions. Une association mesurée par le calcul peut en effet être :

- réelle et causale : c'est la situation idéale ;
- réelle et non causale : l'association observée en terme mathématique peut très bien ne refléter aucun lien de causalité. C'est l'inconvénient des enquêtes d'observation, par opposition aux essais contrôlés ;
- due au hasard : n'oublions pas qu'il existe toujours un risque α de première espèce, c'est-à-dire un risque de conclure à tort à une différence qui n'existe pas ;
- due à un biais : ces biais sont des imperfections inhérentes à toutes les études d'observations.

En épidémiologie analytique, on s'intéresse aux causes des maladies. Il existe cependant une difficulté liée au fait qu'il existe assez rarement une seule cause spécifique entraînant une maladie. Il existe une seconde difficulté d'ordre plus philosophique portant sur la définition d'une cause.
D'un point de vue pragmatique, on adopte la position des responsables de santé publique dont le problème est la prise de décision. Plutôt que de rechercher la cause *princeps* d'une maladie, on privilégie la mise en évidence d'une relation probabiliste entre la fréquence d'une maladie Y et un facteur X. Cette approche permet de distinguer une liste de 8 critères de causalité permettant d'associer la survenue d'une maladie à l'action d'un facteur.
Ces critères sont résumés dans la liste ci-après.

VII. BIAIS DANS LES ENQUÊTES ÉTIOLOGIQUES

Un biais est une erreur systématique qui se glisse dans une enquête lorsque le choix des groupes exposés ou non exposés n'a pas été réalisé indépendamment de la survenue de la maladie (cohorte), ou bien, lorsque le choix des cas ou des témoins n'a pas été réalisé indépendamment de l'exposition). Les conséquences en sont une distorsion dans la mesure du RR ou de l'OR dont la valeur peut être sous-estimée, surestimée ou, plus gravement, inversée. On distingue trois sortes de biais : les biais de sélection, les biais d'information et les biais de confusion.

Critères de causalité

1. **Force de l'association** : elle est donnée par la mesure de l'odds ratio ou du risque relatif. Plus cet indicateur est élevé, plus la liaison est forte.
2. **Stabilité de l'association.** Plus il existe d'études similaires, réalisées dans d'autres circonstances et sur d'autres populations, dont les résultats concordent avec les résultats présents, plus la probabilité d'une liaison réelle avec le facteur étudié est élevée. La stabilité est en rapport avec un degré de signification p élevé, lui-même équivalant à l'étroitesse de l'intervalle de confiance.
3. **Spécificité de l'association.** Si un seul facteur, parmi de nombreux facteurs testés, est relié à la survenue de la maladie cela renforce le lien de causalité.
4. **La relation temporelle.** C'est en fait une condition nécessaire à l'établissement d'un lien de causalité. L'action du facteur doit toujours précéder la survenue de la maladie.
5. **L'effet dose-réponse.** Si le risque relatif ou l'odds ratio est d'autant plus grand que le degré d'exposition est élevé, la probabilité d'une liaison entre facteur et maladie est très forte. Ce critère est le meilleur des critères de causalité.
6. **La plausibilité biologique.** La relation causale est d'autant plus probable qu'il existe des arguments biologiques allant dans le même sens que l'étude épidémiologique.
7. **Cohérence.** C'est la concordance des observations de l'étude en cours avec ce qui est déjà communément admis dans la littérature et le monde scientifique. Cet argument est le plus faible, car il ne tient pas compte de nouvelles découvertes.
8. **Expérimentation.** La relation causale est d'autant plus probable qu'il existe des arguments expérimentaux qui expliquent la nature de la chaîne causale.

Exemple 16.4. EFFET DOSE-RÉPONSE

On veut savoir si le tabac est un facteur de risque de cancer du poumon. On suit pendant plusieurs années une cohorte de sujets fumeurs et non-fumeurs. Les fumeurs sont divisés en trois strates selon leur degré de consommation (cigarettes par jour).

Exposition	Personnes-années à risque	Cancers du poumon	Incidence pour mille p.a.	RR
fumeurs	102 600	133	1,30	21,7
> 25	25 100	57	2,27	37,8
15-24	38 900	54	1,39	23,2
1-14	38 600	22	0,57	9,5
0	48 800	3	0,06	référence

Le risque relatif général est de 21,7. Il existe donc une forte association entre tabagisme et cancer du poumon. En outre le risque relatif, calculé en divisant l'incidence de chaque strate par l'incidence de la strate de référence, montre une franche augmentation du RR en fonction du niveau d'exposition. C'est ce qu'on appelle effet « dose-réponse ». Ce résultat milite très fortement en faveur d'une relation causale. Cet exemple est adapté d'une célèbre enquête britannique réalisée dans les années 50 et qui a pour la première fois mise en évidence cette relation.

1. Biais de sélection

Ils interviennent lors de l'inclusion des sujets de l'enquête :
- dans une enquête de cohorte, lorsque la sélection des exposés et des non-exposés dépend de la survenue de la maladie ;
- dans une enquête cas-témoins lorsque la sélection des sujets malades ou des témoins dépend du facteur d'exposition (exemple 16.5).

Exemple 16.5. BIAIS DE SÉLECTION

- *Non réponse* : omissions plus fréquentes chez les cas exposés ou les témoins non-exposés.
- *Perdus de vue* : plus fréquents chez les exposés malades ou les non-exposés sains.
- *Admission* : cas exposés plus à même d'être sélectionnés que les témoins.
- *Surveillance* : exposés plus facilement détectés (cas) que les non-exposés.
- *Survie sélective* : inclusion de cas survivants moins exposés que les cas décédés.

2. Biais d'information

Ils interviennent au moment du recueil des données :
- dans une enquête de cohorte lorsque les informations concernant la maladie sont recueillies ou mémorisées de façon différente entre les exposés et les non-exposés ;
- dans une enquête cas-témoins lorsque les informations concernant l'exposition sont recueillies de façon différentes chez les cas et les témoins (exemple 16.6).

Exemple 16.6. BIAIS D'INFORMATION

- *Mémorisation* : les cas exposés se souviennent mieux de leur exposition que les témoins.
- *Enquêteur* : il interroge plus longuement les cas non exposés.
- *Qualité des données* : meilleures chez les cas exposés que chez les témoins.
- *« Prévarication »* : réponses induites, ou suggérées chez les cas exposés.

Les biais de sélection et d'information sont définitifs et impossibles à corriger au moment de l'analyse. Ils doivent donc être discutés avant le début de l'enquête et leurs effets doivent être évalués au moment de l'analyse.

3. Biais de confusion

La découverte d'un facteur de risque peut être le train qui en cache un autre. Il est en effet possible qu'à l'exposition étudiée soit associée un autre facteur qui joue lui aussi un rôle dans la survenue de la maladie.
La mesure de la liaison exposition-maladie est alors perturbée par ce **tiers facteur** qu'on appelle facteur de confusion.

Il en existe en fait deux types de tiers facteur :
- les facteurs de confusion proprement dits qui entraînent des biais ;
- les modificateurs de l'effet.

L'étude de ces tiers facteurs nécessite un développement particulier.

VIII. PRISE EN COMPTE D'UN TIERS FACTEUR : ANALYSE STRATIFIÉE

Lorsqu'on étudie la liaison entre une maladie et une exposition, la méthode d'analyse stratifiée permet de prendre en compte un tiers facteur.
Son principe consiste à scinder l'analyse en strates selon les classes de la variable prise comme tiers facteur. On obtient ainsi dans chaque strate un tableau d'analyse à 4 cases et un RR ou un OR pour chacune des strates.

1. Modificateur de l'effet

On appelle modificateur de l'effet, un tiers facteur qui modifie la valeur du RR ou de l'OR d'une strate à l'autre. Il n'entraîne pas un biais mais modifie la mesure de la liaison exposition-maladie pour certaines de ses modalités. Un tel facteur est aussi appelé interaction.
Si les RR ou OR des strates sont très différents et divergent de part et d'autre du RR ou OR brut, cela signifie que la liaison entre le facteur étudié et la maladie n'est pas identique selon les classes du tiers facteur qui modifie l'effet étudié. Il faut examiner les résultats strate par strate et en tirer les conclusions.

Exemple 16.7. MODIFICATEUR DE L'EFFET

Lors d'une étude cherchant à comparer les effets indésirables de deux traitements A et B, on a observé une fréquence deux fois plus élevée d'incidents chez les sujets ayant pris le traitement A (RR = 2,2).

TRAITEMENT	EFFETS	TOTAL	*INCIDENCE*	RR	IC 95 %
A	69	101	*68,3*	2,2	1,6-3,0
B	31	99	*31,3*		

On veut vérifier si l'absorption simultanée de café joue un rôle dans la survenue de ces incidents. On divise donc les données en deux strates : absorption de café ou non et on obtient les résultats suivants.

	TRAITEMENT	EFFETS	TOTAL	*INCIDENCE*	RR	IC 95 %
Café	A	57	77	*74,0 %*	3,5	2,1-5,9
	B	12	57	*21,1 %*		
Pas de café	A	12	24	*50,0 %*	1,1	0,7-1,9
	B	19	42	*45,2 %*		

On constate que le RR de chaque strate est très différent du RR brut, mais surtout qu'ils sont très différents entre eux. Parmi les sujets consommant du café on observe une liaison très forte entre la prise du médicament A et la survenue d'incidents secondaires (RR = 3,5). À l'opposé chez les sujets n'en consommant pas, la liaison disparaît (RR = 1,1).
Un tel résultat suggère que le traitement A est générateur d'incidents secondaires mais seulement s'il est associé à la consommation de café. Le café est un modificateur de l'effet.

Le test de Woolf (non présenté dans cet ouvrage) est réalisé par certains logiciels statistiques comme Epiinfo. Il permet de vérifier l'hétérogénéité des OR ou RR. Si le test est significatif, les RR ou OR diffèrent. Une règle empirique qui peut être admise en première approximation consiste à considérer qu'il faut suspecter un modificateur de l'effet si les RR ou OR des strates diffèrent de plus de 20 %.

2. Facteur de confusion

On appelle facteur de confusion une variable accessoire (tiers facteur) qui fausse la mesure de la liaison entre exposition et maladie. Cette variable de confusion possède à la fois les deux propriétés suivantes :

- elle est liée à la survenue de la maladie indépendamment de l'exposition étudiée, (c'est-à-dire même chez les sujets non-exposés) ;
- elle est liée à l'exposition sans en être la conséquence.

Un facteur de confusion peut avoir des effets complètement opposés : soit surestimer le risque mesuré, soit le sous-estimer, soit le masquer complètement (exemple 16.8).
On contrôle un facteur de confusion :

- soit au début d'une enquête, (si on connaît par avance ce facteur), en appariant les sujets selon les modalités de ce facteur ;
- soit au moment de l'analyse, en stratifiant le calcul de RR ou d'OR sur les modalités du facteur de confusion.

Si les RR ou OR des strates sont proches entre eux (test de Woolf non significatif) et sont différents du RR ou OR brut, on a affaire à un facteur de confusion.
Afin d'obtenir une valeur globale du RR ou OR qui tienne compte de ce facteur, on calcule un RR ou un OR global « ajusté » (ou « pondéré ») selon la méthode de Mantel-Haenszel.

$$RR_{ajusté} = \frac{\Sigma[(a_i t_{0i})/T_i]}{\Sigma[(c_i t_{1i})/T_i]}$$

$$OR_{ajusté} = \frac{\Sigma[(a_i d_i)/T_i]}{\Sigma[(b_i c_i)/T_i]}$$

Strate i	Cas	Sains	Total
Exposé	a_i	b_i	t_{1i}
Non exposé	c_i	d_i	t_{0i}
Total	n_{1i}	n_{0i}	T_i

Le RR ou OR ajusté représente la moyenne pondérée des RR ou OR de chaque strate.
L'analyse statistique peut être effectuée à l'aide du test de χ^2 de Mantel-Haenszel (*cf.* Annexes, Formulaire 18).
Il existe des formules complexes pour calculer les intervalles de confiances de ces RR ou OR ajustés. En pratique, on utilise les logiciels statistiques (EpiInfo6/Epitable/Analyse/Etude de cohorte ou cas témoins/Stratifiée) pour effectuer ce type de calcul. Une méthode simple (mais approchée) est la méthode de Miettinen (*cf.* Annexes, Formulaire 18).

Exemple 16.8.

Lors d'une enquête sur les causes d'accidents post-chirurgicaux parmi les 1 600 patients opérés dans un service d'orthopédie, on constate que l'incidence des accidents est deux fois plus élevée chez les malades opérés par le chirurgien C.

EXPOSITION	ACCIDENTS	TOTAL OPÉRÉS	*INCIDENCE*	RR	IC 95 %
Dr C	200	800	*25,0 %*	2,0	1,6-2,5
Confrères	100	800	*12,5 %*		

Or, l'étude a également montré que le Dr C, spécialiste de la technique X, l'utilise plus souvent que ses confrères. Cette technique, indispensable dans certaines situations difficiles, est cependant dangereuse.

Le problème posé est donc de savoir si le Dr C est réellement dangereux, ou si cette constatation est un artefact car il utilise plus souvent une technique dangereuse.

	DR C	CONFRÈRES
Technique X	400	120
Autre	400	680
Total	800	800
%	50 %	15 %

test de $\chi^2 = 223$, $p < 10^{-5}$

En d'autres termes, la technique X est-elle un facteur de confusion dans le résultat incriminant le Dr C? On stratifie l'analyse en deux sous-groupes, l'un opéré par la technique X, le second opéré par d'autres techniques. À l'intérieur de chaque sous-groupe, on analyse indépendamment le facteur d'exposition « Dr C ».

	EXPOSITION	ACCIDENTS	TOTAL OPÉRÉS	*INCIDENCE*	RR	IC 95 %
Technique X	Dr C	170	400	*42,5 %*	1,02	0,8-1,3
	Confrères	50	120	*41,7 %*		
Autres techniques	Dr C	30	400	*7,5 %*	1,02	0,7-1,6
	Confrères	50	680	*7,4 %*		

On constate que dans chaque strate, l'exposition au Dr C n'est plus liée à la survenue des accidents. Leur incidence est identique chez le Dr. C ou chez ses confrères. Les RR de chaque strate sont proches de 1. Ainsi le résultat de l'analyse brute (RR brut = 2) était-il biaisé par le facteur de confusion que représentait l'emploi de la technique X.

On peut vérifier que la technique X est bien un facteur de confusion :

- elle est associée à l'exposition (le Dr C l'utilise chez 50 % des opérés alors que ses confrères ne l'utilisent que chez 15 % des opérés) ;
- elle est associée à la survenue des accidents indépendamment du chirurgien qui l'utilise :

DR C	ACCIDENTS	TOTAL OPÉRÉS	*INCIDENCE*	RR	IC 95 %
Technique X	170	400	*42,5 %*	5,6	3,9-8,1
Autres	30	400	*7,5 %*		
CONFRÈRES	**ACCIDENTS**	**TOTAL OPÉRÉS**	***INCIDENCE***	**RR**	**IC 95 %**
Technique X	50	120	*41,7 %*	5,6	4,0-7,9
Autres	50	680	*7,4 %*		

Exemple 16.9. CALCUL D'UN RR AJUSTÉ DE MANTEL-HAENSZEL

En reprenant les données de l'exemple 16.7, nous avions calculé un RR brut de 2,0 concernant le chirurgien C.
En reprenant les données des deux strates selon la technique chirurgicale utilisée, on peut calculer un RR ajusté.

TECHNIQUE X	ACCIDENTS	PAS D'ACCIDENTS	TOTAL OPÉRÉS	RR	IC 95 %
Dr C	170	230	400	1,02	0,8-1,3
Confrères	50	70	120		
Total			520		

AUTRES TECHNIQUES	ACCIDENTS	PAS D'ACCIDENTS	TOTAL OPÉRÉS	RR	IC 95 %
Dr C	30	370	400	1,02	0,7-1,6
Confrères	50	630	680		
Total			1 080		

Le RR ajusté est égal à $\frac{[(170 \times 120)/520] + [(30 \times 680)/1\,080]}{[(50 \times 400)/520] + [(50 \times 400)/1\,080]} = 1,02$

L'intervalle de confiance à 95 % est de 0,82-1,27.
On retrouve donc un RR ajusté équivalent à la moyenne des RR de chaque strate.
Le RR ajusté représente le véritable risque relatif lié au chirurgien C, indépendamment de facteur de confusion « technique X ». Ce RR ajusté est proche de 1. On peut donc conclure qu'il n'existe pas d'argument pour affirmer que le chirurgien C est plus dangereux que ses confrères !

3. Stratégie d'analyse stratifiée

On peut résumer la stratégie d'une analyse stratifiée sur un tiers facteur de la façon suivante :

1. Calculer le RR ou OR brut correspondant au facteur de risque étudié.
2. Stratifier l'analyse selon les modalités d'un tiers facteur suspecté.
3. Calculer les RR ou OR de chaque strate.
4. Évaluer la similarité des RR ou OR stratifiés (différences inférieures à 20 %).
5. Si les RR ou OR sont similaires entre eux et similaires au RR ou OR brut, il n'existe ni facteur de confusion, ni modificateur de l'effet. On analyse le facteur de risque avec le RR ou OR brut.
6. Si les RR ou OR de chaque strate diffèrent (test de Woolf significatif), on a affaire à un modificateur de l'effet. On arrête là l'analyse. Il ne faut pas tenir compte du RR ou OR brut. On interprète séparément les RR ou OR de chaque strate.
7. Si les RR ou OR de chaque strate sont similaires, mais différents du RR ou OR brut, le tiers facteur est un facteur de confusion. On calcule alors un RR ou OR ajusté de Mantel-Haenszel.

Résultats et interprétation d'une analyse avec un facteur de confusion

RR ou OR brut	Conclusion provisoire	RR ou OR ajusté*	Conséquence du biais de confusion
> 1	facteur de risque	= 1	pas de facteur de risque
		> 1 et < brut	surestimation
		> brut	sous estimation
		< 1	confusion +++ : facteur protecteur
= 1	absence d'association	> 1	facteur de risque masqué
		< 1	facteur protecteur masqué
< 1	facteur protecteur	= 1	pas de facteur protecteur
		< 1 et > brut	sur estimation
		< brut	sous estimation
		> 1	confusion +++ : facteur de risque

* sur un tiers facteur de confusion

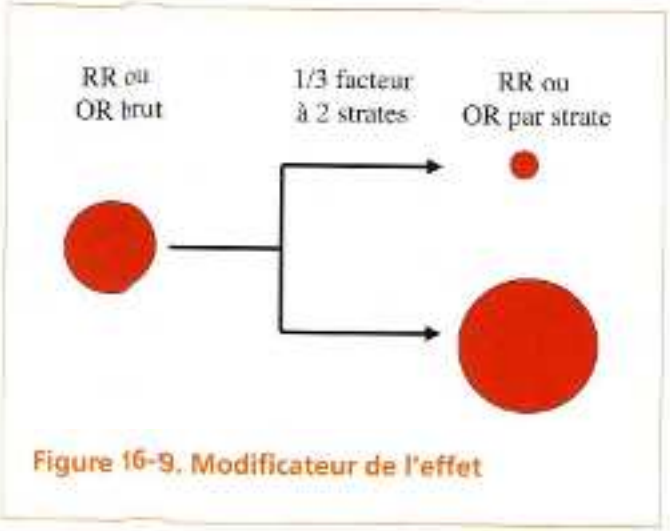


Figure 16-9. Modificateur de l'effet

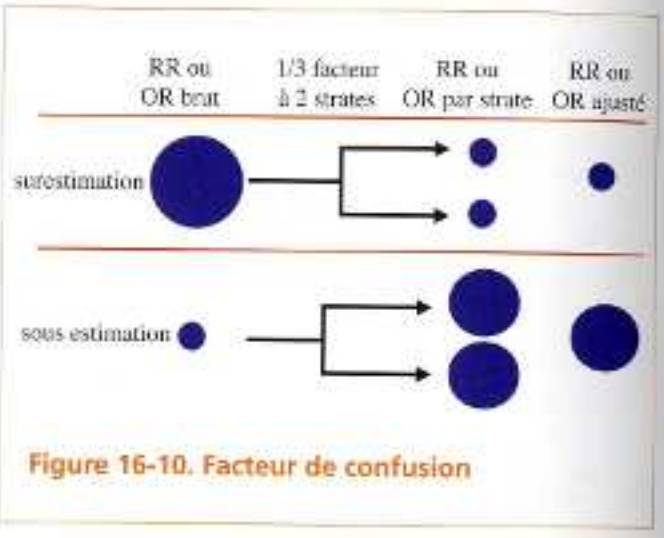


Figure 16-10. Facteur de confusion

4. Analyse multivariée

Nous avons jusqu'à maintenant utilisé des exemples d'analyse bivariée (un facteur de risque et un facteur de confusion).

Mais dans une enquête, il est fréquent de détecter plusieurs facteurs de confusion qui sont eux-mêmes des facteurs de risque. Il faudrait alors subdiviser l'analyse en strates à plusieurs dimensions, ce qui devient impossible au-delà de deux niveaux. Dans cette situation, il est nécessaire d'utiliser les méthodes d'analyse multivariée. La technique de **régression logistique** est la plus utilisée à cet effet.

Les résultats d'une analyse de régression logistique fournissent pour chaque facteur étudié un risque (sous forme d'OR) et un intervalle de confiance. Lorsque ce risque est significativement différent de 1, on peut alors affirmer (sous réserve d'avoir pris en compte tous les facteurs susceptibles d'être des facteurs de risque) que le facteur étudié est un facteur de risque, indépendamment des autres facteurs.

La technique de régression logistique est complexe et sort du cadre de cet ouvrage. Nous allons juste l'illustrer par un exemple des résultats qu'elle fournit (exemple 16.10).

Exemple 16.10. ANALYSE MULTIVARIÉE

Une étude cas-témoins sur les facteurs de risque de contracter la toxoplasmose par les femmes enceintes non immunisées a permis d'analyser une centaine de facteurs de risque possibles. L'analyse univariée a retenu une dizaine de facteurs présentant un odds ratio élevé. En raison de la possibilité d'une liaison entre certains de ces facteurs (confusion), une analyse multivariée par régression logistique a été effectuée en ajustant chacun de ces facteurs sur l'ensemble des autres. Le tableau ci-dessous présente une partie des résultats.

Étude des facteurs de risque de toxoplasmose chez les femmes enceintes : analyse multivariée par régression logistique. Étude cas-témoins sur 80 cas et 80 témoins :

FACTEURS DE RISQUE :	ANALYSE UNIVARIÉE		ANALYSE MULTIVARIÉE	
	OR BRUT	IC 95 %	OR AJUSTÉ	IC 95 %
consommation de viande mal cuite	4,5	3,8-6,9	5,4	3,2-7,8
consommation de crudités	1,6	0,8-2,9	2,8	1,1-4,9
présence d'un chat au domicile	2,5	1,3-4,7	1,6	0,8-2,5
hygiène des mains défectueuse	5,7	3,2-9,5	9,1	6,3-13,7
information non délivrée	2,4	1,7-3,6	2,7	1,9-3,8

Ce tableau, qui présente à la fois les OR bruts de l'analyse univariée et les OR ajustés de l'analyse multivariée, montre certaines différences dans leurs valeurs respectives : par exemple l'OR brut « consommation crudités » non significativement différent de 1, le devient lorsqu'il est ajusté en multivariée. À l'inverse, l'OR brut « chat » significativement différent de 1 ne l'est plus en multivariée. Enfin, d'autres OR bruts sont modifiés, mais conservent leur signification. Ces variations sont dues aux effets de confusion dans la simple analyse univariée. Ainsi, si les propriétaires de chat sont plutôt des femmes consommant de la viande mal cuite, et sachant que la consommation de viande mal cuite est associée à la maladie, on observe plus de propriétaires de chat associées à la toxoplasmose. Le processus d'ajustement de l'analyse multivariée a permis de gommer ce biais : indépendamment de la consommation de viande mal cuite, le fait de posséder un chat à domicile n'est plus lié à la toxoplasmose.

À l'inverse, si les consommatrices de crudités, sont plutôt des femmes ayant une hygiène correcte, on observe moins de consommatrices de crudités associées à la toxoplasmose. Après ajustement, indépendamment de l'hygiène des mains, on constate que la consommation de crudité est aussi associée à la toxoplasmose.

On peut ainsi conclure en affirmant que la survenue de toxoplasmose est associée à la consommation de viande mal cuite, à la consommation de crudités, à un manque d'information prophylactique au début de la grossesse et à une mauvaise hygiène des mains qui apparaît comme le principal facteur de risque.

Exercices

Exercice 16.1

Le tableau suivant présente les résultats de plusieurs enquêtes étiologiques de type cohorte ou cas-témoins.

	RR ou OR	IC 95 %
1)	18	0,5-48
2)	1,2	1,15-1,27
3)	0,01	0,005-0,017
4)	0,4	0,2-3,5
5)	0,9	0,81-0,95

Associez à chacun des résultats, l'interprétation qui lui convient.

a) Facteur de risque, liaison significative.

b) Facteur protecteur, liaison significative.

c) Facteur de risque possible, mais liaison non significative.

d) Facteur protecteur possible, mais liaison non significative.

e) Facteur protecteur, liaison très forte.

Exercice 16.2

Quel type d'enquête étiologique (cohorte ou cas-témoins) choisiriez-vous pour étudier :

1) la cause d'une épidémie d'une maladie infectieuse incubant en quelques jours ;

2) une maladie chronique fréquente due à une cause rare ;

3) une maladie rare due à une cause fréquente ;

4) une maladie liée à des facteurs de risque multiples ;

5) plusieurs maladies ayant un facteur de risque en commun.

Exercice 16.3

Dans une enquête cas-témoins lors d'une épidémie de gastro-entérite, on a trouvé une liaison entre la consommation de crustacés et la survenue de gastro-entérite : OR = 3,5[1,8 – 5,5]. L'analyse stratifiée selon la consommation de vin blanc pendant le repas (abstention, consommation modérée et consommation supérieure à 1 L) a montré respectivement les valeurs suivantes (OR_i et IC 95 %) : OR_1 = 4,8[2,6 – 7,7], OR_2 = 1,8[1,3 – 2,4], OR_3 = 0,1[0,02 – 0,24]. Quelle est votre interprétation de ces résultats ? Pouvez-vous calculer avec ces données un OR ajusté ?

Résumé

Les enquêtes épidémiologiques sont des études visant à fournir soit des observations sur la situation d'une maladie à un instant donné (enquête descriptive), soit à établir une relation entre la survenue d'une maladie et la présence de facteurs de risque (enquête étiologique).
Les enquêtes étiologiques se décomposent en deux grands types :

- les enquêtes de cohorte qui compare l'incidence de la maladie entre des groupes de sujets exposés et non-exposés à un facteur de risque ;
- les enquêtes cas-témoins qui comparent l'exposition à un facteur de risque entre des cas et des témoins non malades.

TYPES D'ENQUÊTES ÉPIDÉMIOLOGIQUES À VISÉE ÉTIOLOGIQUE

	CAS	SAINS
exposés	a	b
non exposés	c	d

ENQUÊTE	INDICATEURS MESURÉS		RÉSULTAT	
Cohorte	Incidence chez exposés	$I_e = a/(a + b)$	Risque relatif	$\frac{I_e}{I_{ne}}$
	Incidence chez non-exposés	$I_{ne} = c/(c + d)$		
Cas-témoins	% d'exposition chez les cas	$a/(a + c)$		
	% d'exposition chez témoins	$b/(b + d)$		
	Cote d'exposition chez les cas	a/c	Odds ratio ou	$\frac{a/c}{b/d}$
	Cote d'exposition chez les témoins	b/d	Rapport de cote	
Transversale	Prévalences chez les exposés	$P_e = a/(a + b)$	Rapport de	$\frac{P_e}{P_{ne}}$
	Prévalence chez les non-exposés	$P_{ne} = c/(c + d)$	prévalence	

17

INVESTIGATION D'UNE ÉPIDÉMIE

Le but de ce chapitre est de présenter la méthodologie qui sert de base à toute investigation d'une épidémie. Cette méthodologie se présente sous forme d'étapes nécessairement ordonnées. L'épidémiologiste de terrain, confronté à des situations très diverses sera souvent obligé de s'éloigner de cet ordre strict. Mais il devra garder en tête ces procédures de travail afin d'éviter des maladresses et des oublis.

I. DÉFINITIONS

Épidémie

On appelle épidémie, la survenue de cas d'une maladie quelconque, dont le nombre est supérieur au nombre de cas attendus pendant une période de temps donnée et en un lieu donné.

On remarque dans cette définition que la pathologie en cause peut être *quelconque*. Au sens moderne du terme, la cause d'une épidémie n'est pas limitée aux seules maladies transmissibles, bien que ce champ d'étude reste le principal. On peut aussi observer et analyser de nombreux phénomènes épidémiques non infectieux (exemple 17.1).

Exemple 17.1. EXEMPLES D'ÉPIDÉMIE

MALADIES INFECTIEUSES	MALADIES NON INFECTIEUSES
Listériose	Malnutrition sévère
Légionnellose	Leucémie
Salmonellose	Intoxications
Méningite	Accidents
Choléra	Suicides

La notion « supérieure au nombre de cas attendus » est volontairement vague et ne fait pas appel au nombre absolu de cas. Ainsi, la survenue de 2 cas de variole, en tout endroit du monde, serait à coup sûr considéré comme une épidémie à étudier d'urgence. Pour les maladies fréquentes faisant l'objet de surveillance systématique, on exige le dépassement de seuils prédéterminés.

Termes équivalents

Le mot « épidémie » est chargé d'une connotation suggérant un événement de grande ampleur et à forte létalité (peste, choléra…).

Pour des épidémies de faible importance et limitées dans le temps, on emploie les termes d'*épisode épidémique*, de *bouffée épidémique* ou de *flambée épidémique*. Lorsque l'épidémie est particulièrement bien circonscrite en un ou plusieurs lieux, on utilise le terme de *foyer(s) épidémique(s)* (*cluster* en anglais).

Investigation

C'est l'ensemble des opérations consistant à recueillir les données, décrire le phénomène et analyser les causes d'une épidémie. L'investigation est un processus relativement rapide, limité dans le temps. Il s'oppose au processus de surveillance des maladies qui nécessite la mise en place de structures permanentes.

Éclosion

Le terme d'éclosion caractérise le développement initial d'une épidémie.

Source

La source caractérise le point d'émergence de l'agent pathogène. La source est le plus souvent associée à un lieu fixe. Mais elle peut également être associée à un individu ou un objet circulant.
On distingue selon le temps, une source ponctuelle d'une source persistante. Une ***source ponctuelle*** est caractérisée par une émission de l'agent pathogène pendant une période très courte. C'est le cas, par exemple, des toxi-infections alimentaires collectives. Une ***source persistante*** continue à émettre l'agent pathogène, comme par exemple dans une pollution environnementale.
On distingue aussi, selon la topologie une **source commune**, qui contamine tous les individus directement en contact avec elle. Une épidémie à source commune s'oppose à une épidémie à transmission inter-humaine qui se propage dans la population.

Véhicule

Ce terme caractérise le support qui a contribué à la diffusion de l'agent pathogène entre la source et la population atteinte. Dans une toxi-infection collective, le véhicule est très souvent un aliment.

II. OBJECTIFS

Une investigation possède un coût en moyens matériels, en personnel qualifié et en temps de travail. Elle doit être justifiée par les objectifs suivants :

Justifications pour l'investigation d'une épidémie

- Enrayer la propagation du phénomène.
- Prévenir la survenue de nouveaux épisodes.
- Augmenter les connaissances sur la maladie.
- Évaluer la qualité du système d'alerte.
- Diffuser les principes et techniques d'investigation.

Si les 2 premiers objectifs de cette liste ne peuvent être remplis, il est inutile de réaliser une investigation.

Les **objectifs spécifiques** d'une investigation d'épidémie sont :

- Identifier l'***agent causal.***
- Localiser la ***source.***
- Déterminer le mode de transmission ou le ***véhicule.***
- Identifier la ***population à risque.***
- Déterminer les ***facteurs de risque*** de la maladie.

III. CHRONOLOGIE

L'accomplissement de ces objectifs se réalise en deux phases :
- une phase descriptive ;
- une phase analytique de recherche causale.

Chacune des phases comporte plusieurs étapes qui peuvent se chevaucher en raison des contraintes de temps. Schématiquement, l'investigation d'une épidémie en comporte 10.

Phase descriptive
1. Affirmer la réalité de l'épidémie
2. Confirmer le diagnostic
3. Définir un cas
4. Collecter les cas
5. Décrire l'épidémie dans ses composantes spatio-temporelles

Phase analytique
6. Formuler des hypothèses
7. Tester les hypothèses par une enquête étiologique
8. Rechercher la preuve biologique
9. Communiquer les conclusions de l'investigation
10. Prendre les mesures de prévention

Les cinq premières étapes doivent être réalisées avec un objectif purement descriptif, en évitant tout *a priori* sur l'éventuelle cause. C'est à partir des observations faites lors de cette phase descriptive que sont générées les hypothèses conduisant à l'étude analytique.

1. Affirmer la réalité de l'épidémie

- L'affirmation d'un nombre de cas observés supérieur au nombre de cas attendus dépend de la précision avec laquelle on est capable d'établir des prévisions :
- soit on dispose d'un *système de surveillance*. On affirme la réalité d'une épidémie lorsque le *seuil épidémique* a été franchi (figure 17.1). Ce seuil est le plus souvent un taux d'incidence supérieur d'environ 2 écarts type au seuil d'incidence prévu par des modèles. Ces modèles de prévision plus ou moins complexes tiennent compte de nombreux paramètres et notamment des variations saisonnières. Si l'incidence observée est supérieure à ce seuil, on conclut (avec un risque de première espèce) que le phénomène n'est pas une simple fluctuation. On affirme donc la survenue d'une épidémie greffée sur le fond endémique habituel ;

- soit la maladie est rare et il n'existe pas de système de surveillance. L'épidémie est affirmée sur la survenue d'un nombre de cas manifestement élevé, éventuellement après une enquête rapide pour confirmer des rumeurs évoquant l'épidémie. La loi de Poisson (chap. 6.II) est adaptée à ce type de question.

■ Fausses épidémies

Il existe plusieurs situations, où l'augmentation du nombre de cas n'est qu'apparente :

- mise en place récente d'un système de surveillance ;
- amélioration récente du système de surveillance ;
- amélioration des techniques de dépistage ou de diagnostic ;
- augmentation récente de la population.

Ainsi, avant de confirmer une épidémie, faut-il s'assurer qu'il n'existe aucun de ces artefacts.

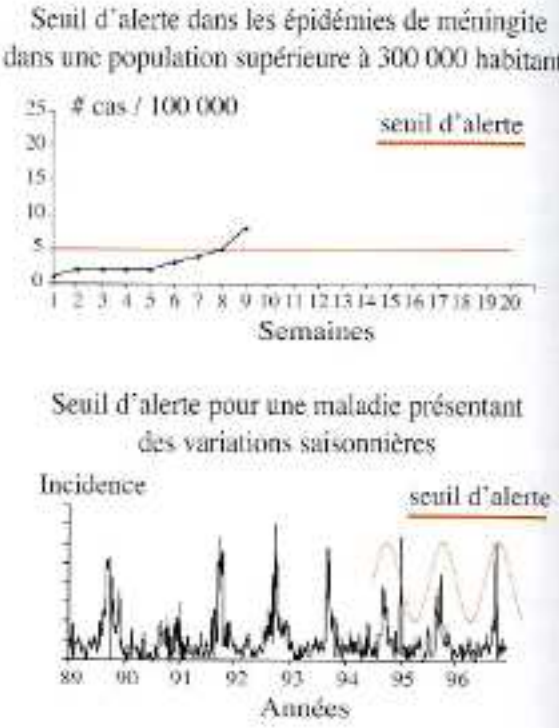


Figure 17-1.

2. Confirmer le diagnostic

Cette étape a pour but de vérifier la réalité du phénomène avec les spécialistes de la maladie. On vérifie notamment la cohérence des symptômes entre malades et on s'assure d'un diagnostic certain par des techniques de laboratoire. Pour l'épidémiologiste, cette vérification diagnostique peut être limitée à un nombre restreint de malades (10-20 %). Les autres cas peuvent être reliés à l'épidémie par analogie clinique.

Cette étape de vérification diagnostique est importante pour éliminer là encore de fausses épidémies (regroupement de malades présentant des symptomatologies diverses, rumeurs, phénomènes de suggestion collective favorisée par les forums sur Internet).

3. Définir un cas

Les principes généraux de définition d'un cas ont été abordés au chapitre 16.I.3.

■ ***Borne initiale temporelle :*** on sélectionne tous les cas ayant présenté des symptômes à partir d'une date précise. Il faut veiller à fixer cette date suffisamment en amont du début de l'événement pour ne pas éliminer des cas précoces passés inaperçus au début de l'enquête. Si on connaît la durée d'incubation de la maladie, on fixera la borne initiale à une date antérieure à la différence entre la date de survenue des premiers cas observés et la durée d'incubation.

■ ***Limites territoriales :*** la zone géographique de recherche des cas est fixée en fonction de l'origine des premiers cas. Elle doit être suffisamment vaste pour retrouver la majeure partie des cas constituant l'épidémie. Elle doit être raisonnablement limitée en raison des contraintes logistiques.

■ ***Critères clinico-biologiques***

Si la maladie est bien connue, responsable d'épidémies classiques, on utilise les définitions internationales. Si la maladie est mal connue, il faut privilégier la sensibilité de la définition afin de recueillir le maximum de cas et étudier les profils clinico-biologiques.

Il est toujours possible, au moment de l'analyse, de restreindre un nombre de cas trop élevé et douteux, en augmentant la spécificité. À l'inverse, si la définition initiale était trop spécifique, avec trop peu de cas collectés, il est trop tard pour agir, à moins de recommencer l'enquête sur le terrain.

Exemple 17.2. DÉFINITIONS DE CAS

Méningite : enfant de moins de 1 an présentant une fièvre supérieure à 38,5 °C et un bombement de la fontanelle depuis le 1er janvier 1997 dans le district de Bamako, Mali.

Rougeole : tout sujet présentant depuis le 1er février 2001 dans le district de Niamey, Niger, une fièvre supérieure à 38,5 °C ET une éruption cutanée depuis plus de 3 jours ET au moins un signe parmi les 3 suivants : toux OU coryza OU conjonctivite.

Trichinellose : tout sujet habitant la ville de Montauban et présentant une fièvre depuis le 15 juillet 1998.
- Cas certain : biopsie positive avec la présence du parasite OU sérodiagnostic positif (IFI > 1/100e).
- Cas probable : présence d'au moins 3 signes parmi les 4 suivants : fièvre supérieure à 39 °C, myalgies, œdème facial, éosinophilie supérieure à 1 000 c/mm³.
- Cas suspect : présence d'une éosinophilie supérieure à 1 000 c/mm³.

4. Collecter les cas et les données

Cette étape constitue le travail de terrain proprement dit. La collecte des cas s'effectue le plus souvent auprès des structures de santé (médecins, dispensaires, hôpitaux, laboratoires, services de médecine préventive…). La recherche des cas peut s'effectuer également par une recherche active en communauté générale.

Les données à recueillir sur les cas sont de plusieurs ordres :
- données d'identification : âge, sexe, adresse, *etc.* ;
- données cliniques et biologiques ;
- données temporelles : date de début des symptômes ++ ;
- données topographiques : lieux de vie, déplacements ;
- données sur les facteurs de risque si l'enquête étiologique est menée de front avec l'enquête descriptive.

Parallèlement aux données individuelles concernant les cas, on recueille des données équivalentes sur la population d'étude (structure âge/sexe, *etc.*).

5. Décrire l'épidémie

Au terme du recueil de données, l'épidémie est décrite de façon synthétique. Une épidémie se caractérise par :
- un nombre total de cas (certains, probables, suspects) ;
- un taux d'attaque ;
- un nombre total de décès liés à l'épidémie ;
- une létalité ;
- une cause éventuelle, si elle a déjà été découverte au moment de l'étape descriptive ;
- une distribution des cas en termes de temps, lieux et personnes.

a) Temps

La distribution des cas en fonction du temps aboutit à la **courbe épidémique**.

Cette courbe, représentée par un histogramme, est tracée en reportant en abscisses des périodes de temps, et en ordonnées le nombre de cas survenus pendant chaque période de temps. Le choix de l'unité de temps dépend de la durée de l'incubation de la maladie. Afin d'obtenir une courbe épidémique ni trop étalée, ni trop ramassée, on choisit une unité de temps à peu près égale au quart de la durée de l'incubation.

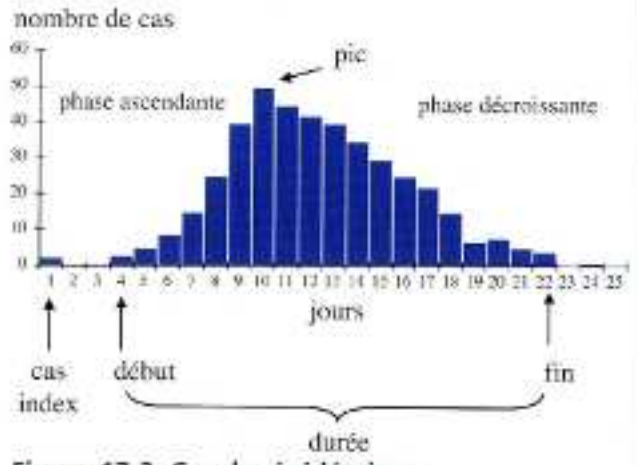


Figure 17-2. Courbe épidémique

Description de la courbe épidémique

On examine systématiquement (figure 17.2) :

- la date de début : d_{ini} ;
- la date de fin : d_{fin} ;
- la présence de cas aberrants ;
- la durée totale de l'épidémie : Δt ;
- la présence d'un ou de plusieurs pics ;
- la date du pic (ou des pics) : d_{pic} ;
- le profil général de la courbe.

Interprétation de la courbe épidémique

L'allure d'une courbe épidémique peut renseigner sur la nature de la source.

- ***Source ponctuelle :*** elle se traduit par une courbe épidémique unimodale (un seul pic) avec une ascension rapide et une décroissance légèrement étalée sur la droite. À l'aide d'une telle courbe, il est possible de situer la période d'exposition à la source (figure 17.3).
  Si on connaît les durées extrêmes d'incubation de la maladie, i_{min} et i_{max}, la période d'exposition se situe entre $[d_{ini} - i_{min}]$ et $[d_{fin} - i_{max}]$.
  Si on connaît la durée médiane de l'incubation i_{med}, la date d'exposition est proche de $[d_{pic} - i_{med}]$.
- ***Source persistante :*** elle se traduit par une courbe avec une ascension rapide, suivit d'un plateau (figure 17.4).
- ***Transmission inter-humaine :*** elle se traduit par une courbe qui évolue après un pic initial, en plusieurs vagues d'amplitude croissante qui traduisent les contaminations de proche en proche des groupes d'individus.

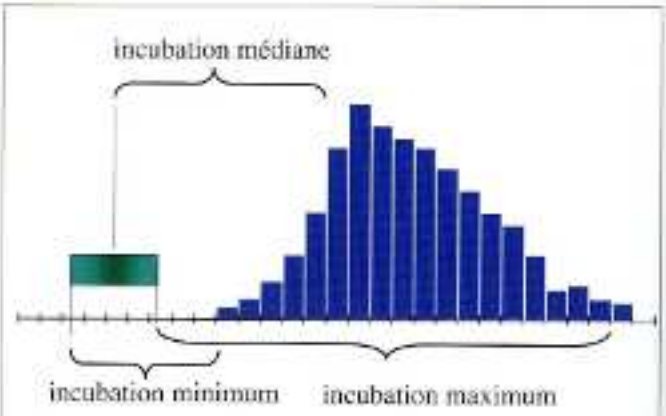


Figure 17-3. Estimation de la période d'exposition en fonction de la durée d'incubation. Épidémie à source ponctuelle

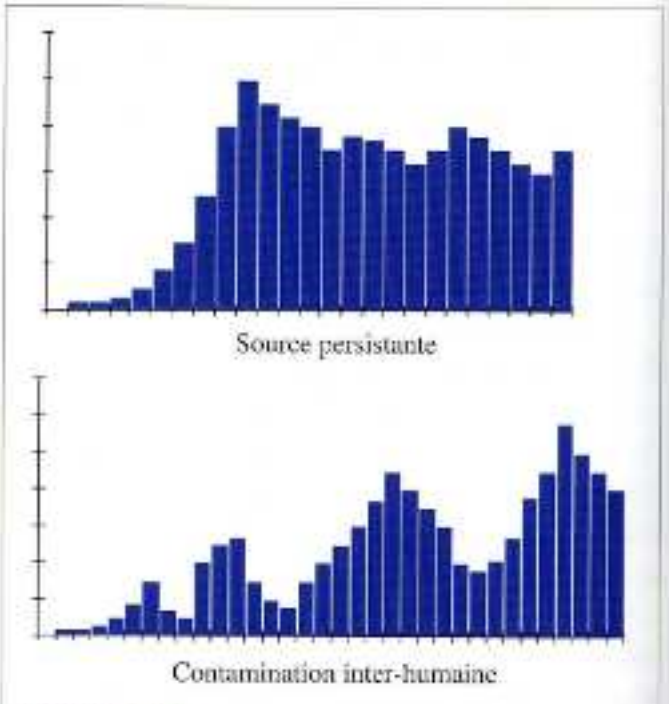


Figure 17-4.

b) Lieux

C'est l'étude de la distribution des cas en fonction de données topographiques. Plusieurs types de localisation des individus peuvent être analysés :

- lieu d'habitat ;
- lieu de travail ou d'activité journalière ;
- lieu de repas : cantines, restaurant ;
- lieu de loisir.

En cas d'épidémie très focalisée en collectivités fermées (école, hôpital, caserne, couvent, prison), l'enquête topographique garde la même importance à une échelle inférieure (classes, chambres, cellules, *etc.*).

La cartographie des cas s'effectue soit :

- par pointage de chaque cas sur une carte lorsqu'on ne dispose que de données sur les cas (figure 17.5) ;
- par des zones colorées proportionnellement à des taux d'attaque lorsqu'on dispose de données démographiques sur ces zones ;
- par des diagrammes en barres, où chaque barre représente un taux d'attaque pour un lieu donné (figure 17.7).

La lecture d'une carte permet de visualiser les foyers épidémiques par concentration des cas autour de points singuliers. L'identification d'un foyer nettement individualisé est en faveur d'une source commune à ce foyer. Le centre du foyer devient le lieu de recherche privilégié de la source commune.

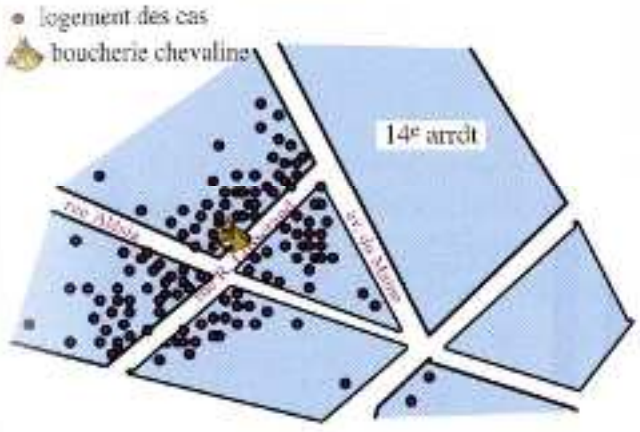


Figure 17-5. Épidémie de trichinellose, Paris 1985

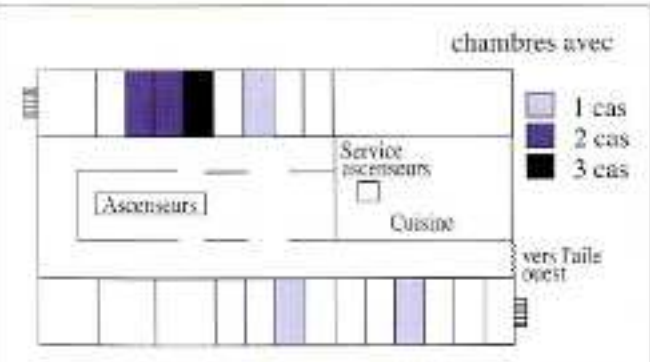


Figure 17-6. Épidémie d'aspergillose dans un hôpital

Figure 17-7. Épidémie d'intoxication à l'héliotrope, Tadjikistan, 1992-1993. Taux d'incidence par kolkhoze

c) Personnes

On étudie la distribution des cas selon les catégories humaines. Les catégories fondamentales sont l'âge et le sexe. Selon la nature de l'épidémie, d'autres critères peuvent être étudiés : catégorie socio-professionnelle, niveau d'étude, nationalité, ethnie, religion, *etc.*

Pour pouvoir être interprétés, les tableaux de distribution doivent présenter non seulement le nombre de cas, mais aussi les taux d'attaque pour chacune des catégories étudiées. Cela nécessite de disposer des dénominateurs, c'est-à-dire de données démographiques.

La lecture des distributions selon les personnes permet d'identifier les groupes à risque (exemple 17.3).

Exemple 17.3. DISTRIBUTION DES CAS DE MÉNINGITE PAR TRANCHE D'ÂGE, BAMAKO, MALI, 1997. SOURCE : ÉPICENTRE, 1997

ÂGE (en années)	POPULATION (N)	NOMBRE DE CAS (N)	*INCIDENCE POUR 10^5*
< 1	32 055	91	*283,9*
1-4	108 267	81	*74,8*
5-14	227 097	229	*100,8*
15-29	241 562	201	*83,2*
30-44	200 571	64	*31,9*
Total	809 552	666	*82,3*

On constate sur ce tableau que le groupe d'âge le plus à risque est le groupe des enfants de moins de un an, bien que ce groupe soit inférieur en nombre de cas.

6. Formuler des hypothèses

Au terme de la phase descriptive de l'épidémie, on connaît l'ampleur du phénomène et sa gravité. On dispose d'arguments sur la nature de la source, ponctuelle ou persistante, sur sa localisation. On a caractérisé les groupes à risque. L'étude de certains cas particuliers, comme les cas aberrants, précoces ou tardifs, les cas exposés de façon exclusive à un agent particulier, les groupes de cas familiaux, fournissent des pistes.

Tous ces éléments permettent de générer des hypothèses sur :

- la source de l'épidémie ;
- le véhicule ;
- les modes de transmission ;
- les facteurs favorisant la transmission de la maladie.

On liste de façon systématique ces hypothèses et on met en place un protocole d'enquête pour les tester.

7. Tester les hypothèses par une enquête étiologique

Les types d'enquête les plus adaptés à une investigation d'épidémie qui exige une réponse rapide, sont l'enquête cas-témoins ou l'étude de cohorte rétrospective en milieu fermé.

Les témoins d'une enquête cas-témoins doivent être recrutés avec les mêmes exigences que les cas. On établit une définition des témoins qui doivent être issus de la même population que les cas. La

définition des témoins (ou des sujets sains d'une cohorte) comporte des critères clinico-biologiques d'exclusion. Le recrutement des témoins doit être fait indépendamment de leur probabilité d'être ou non exposés aux risques testés, afin de ne pas biaiser les résultats.
On recueille auprès de chaque cas et de chaque témoin, à l'aide d'un questionnaire, les données sur les facteurs de risque à tester.
Chaque facteur de risque fait l'objet d'une analyse aboutissant par un calcul d'odds ratio ou de risque relatif, à retenir ou à éliminer une liaison entre le facteur et la survenue de la maladie.

Lorsque les causes d'une épidémie sont difficiles à établir avec certitude, il est parfois nécessaire, d'effectuer des enquêtes complémentaires :
- en élargissant le groupe témoin pour augmenter la puissance ;
- en travaillant sur des groupes plus restreints à risque élevé ;
- en quantifiant les facteurs de risque pour calculer des relations dose-effet (exemple 17.4).

Exemple 17.4. Épidémie de trichinellose, France, 1993

• Enquête cas-témoins sur la consommation de *viandes* dans le mois précédant la survenue des symptômes

CONSOMMATION DE VIANDE		CAS N = 239	% *EXP.*	TÉMOINS N = 177	% *EXP.*	OR	IC 95 %
bœuf	oui	227	*95,0*	168	*94,9*	1,01	0,4-2,5
	non	12		9			
porc	oui	203	*84,9*	150	*84,7*	1,01	0,6-1,7
	non	36		27			
cheval	oui	238	*99,6*	128	*72,3*	91,1	> 12,5
	non	1		49			

La viande de cheval est significativement et exclusivement associée à la survenue de la trichinellose.

• Enquête cas-témoins sur la fréquence de consommation de *viande chevaline* dans le mois précédant la survenue des symptômes

CONSOMMATION DE VIANDE CHEVALINE	CAS N = 239	% *EXP*	TÉMOINS N = 177	% *EXP*	OR	IC 95 %
< 1 fois par mois	40	*16,7*	41	*23,2*	47,8	> 6,3
1 fois par mois	89	*37,2*	47	*26,6*	92,8	> 12,4
1 fois par semaine	57	*23,8*	25	*14,1*	111,7	> 14,6
> 1 fois par semaine	52	*21,8*	15	*8,5*	169,9	> 21,6
jamais (référence)	1	*0,4*	49	*27,7*		

Dans ce tableau, le facteur de risque étudié est maintenant la *fréquence de consommation* de la viande chevaline isolée comme facteur de risque dans le tableau précédent. Chaque odds ratio est calculé en prenant comme groupe non exposé, le groupe ne consommant jamais de viande chevaline. On constate une augmentation des odds ratio en fonction de la fréquence de consommation. Ceci apporte un argument décisif dans la mise en cause de cette viande.

8. Rechercher la preuve biologique

L'enquête épidémiologique proprement dite a l'avantage d'une grande efficacité dans la détection du ou des facteurs étiologiques. Elle est notamment très rapide. Elle n'a pas pour autant valeur de preuve.
Parallèlement à l'investigation, on recherche l'agent causal à l'aide de prélèvement chez les cas, mais aussi sur les sources potentielles ou les véhicules suspectés lors des différentes étapes de l'investigation. Les prélèvements font l'objet de dosages biologiques divers selon la nature de la pathologie en cause : culture, dosage chimique, typages de souches, *etc.* Les techniques modernes d'analyse génomique dans les maladies infectieuses permettent de confronter les germes retrouvés chez les cas et dans l'environnement. La découverte d'un profil génomique identique entre des souches retrouvées chez les cas et dans la source ou le véhicule de l'épidémie, permet de certifier la relation causale. Le typage génomique permet aussi d'augmenter la spécificité de la définition dans des cas en distinguant les souches « épidémiques » des souches diverses constituant le bruit de fond.
Toutes ces actions représentent une grande part d'une investigation et exigent le concours de spécialistes concernés par l'événement : vétérinaires, ingénieurs sanitaires, biologistes, *etc.*

9. Communiquer les conclusions de l'investigation

L'objectif de toute investigation est l'intervention. Il est donc impératif de communiquer l'ensemble des résultats à tous les acteurs de santé publique concernés. Une investigation d'épidémie doit déboucher sur un rapport complet à la fois descriptif et analytique formulant la totalité des hypothèses soulevées et des résultats obtenus.

Un rapport d'investigation comprend :

- les circonstances de survenue de l'épidémie ;
- les faits anecdotiques initiaux ;
- la méthodologie de l'enquête descriptive et étiologique ;
- la quantification des cas, des formes compliquées, des décès ;
- les taux d'attaque, les pourcentages de complications et de létalité ;
- la description de l'épidémie en termes de temps, lieux et personnes ;
- les hypothèses testées ;
- les résultats des enquêtes étiologiques et biologiques ;
- une discussion comportant une critique du travail effectué, un exposé des biais éventuels, une comparaison avec des investigations antérieures ;
- des mesures de contrôle et de prévention préconisées.

10. Prendre les mesures de prévention

C'est l'ultime justification d'une investigation. Il s'agit là d'un acte de santé publique. Le rôle de l'épidémiologiste est d'argumenter les mesures de prévention et de contrôle jugées indispensables à la lumière des résultats de l'investigation.

IV. ASPECTS OPÉRATIONNELS

La mise en place d'une investigation nécessite une organisation planifiée et des moyens en personnel, en temps et en matériel. Tous ces aspects doivent être passés en revue au début d'une investigation.

1. Aspects administratifs

- Identifier dès le début :
  - origine et motivations de la demande ;
  - crédits alloués ;
  - responsables de l'enquête ;
  - autorités destinatrices du rapport.

- Rencontrer :
  - les responsables locaux ;
  - maire, directeurs des structures ;
  - médecins, chefs de service ;
  - infirmières.

2. Aspects logistiques

- Identifier :
  - les lieux de travail ;
  - les moyens de communication ;
  - le laboratoire de référence ;
  - les moyens informatiques.

- Se procurer :
  - la bibliographie ;
  - appareil de photo ;
  - micro-ordinateur portable ;
  - logiciel de saisie et d'analyse ;
  - papier graphique ;
  - ordre de mission.

3. Coopérations

- Scientifiques :
  - spécialistes médicaux ;
  - biologistes ;
  - toxicologues ;
  - vétérinaires ;
  - ingénieurs sanitaires.

- Administratives :
  - direction générale de la santé : DGS ;
  - institut de veille sanitaire : InVS ;
  - action sanitaire et sociale : DASS ;
  - cellule interrégionale d'épidémiologie : CIRE ;
  - agence de sécurité alimentaire : AFFSA ;
  - service de répression des fraudes ;
  - services de protection et environnement.

4. Communication

- Identifier les auteurs responsables.
- Lister les supports de communication :
  - rapports administratifs ;
  - publications scientifiques ;
  - communiqués de presse.

- Lister les cibles :
  - autorités sanitaires ;
  - professionnels impliqués ;
  - revues scientifiques ;
  - sociétés savantes ;
  - grand public.

Résumé

L'investigation d'une épidémie se déroule le plus souvent dans des situations d'urgence où se mêlent l'angoisse de populations et la pression des autorités responsables. Il est donc impératif d'appliquer une méthodologie systématique afin de ne pas oublier des étapes importantes.
L'investigation d'une épidémie comporte une partie purement descriptive du phénomène (les 5 premières étapes) et la recherche de la cause. Cette recherche utilise à la fois des moyens statistiques et des méthodes biologiques visant à remonter jusqu'à la cause initiale.
Le but ultime de toute investigation est de proposer des mesures de prévention immédiate et d'éviter qu'un nouvel épisode se reproduise.

LES 10 ÉTAPES DE L'INVESTIGATION D'UNE ÉPIDÉMIE

1) Affirmer l'existence de l'épidémie.
2) Confirmer le diagnostic.
3) Définir un cas.
4) Collecter les cas et les données.
5) Décrire l'épidémie (temps, lieux, personnes).
6) Formuler des hypothèses.
7) Tester les hypothèses (enquête étiologique).
8) Vérifier la cohérence biologique.
9) Rédiger un rapport.
10) Prescrire les mesures de prévention.

MESURES D'IMPACT

Les indicateurs de mesure d'impact servent à mesurer l'importance d'un facteur de risque ou d'un facteur protecteur en termes de santé publique.

Les mesures d'association (OR et RR), ne servaient en effet qu'à démontrer le rôle d'un facteur dans l'étiologie d'une maladie. Mais ils ne renseignent en rien, sur l'importance relative de ce facteur sur la fréquence de la maladie.

Par exemple, on peut démontrer que conduire une automobile les yeux fermés est un facteur de risque d'accident très élevé. Mais ce n'est pas le seul facteur et sa part dans les accidents de la circulation est sans doute assez faible.

Les indicateurs de mesure d'impact permettent d'établir des priorités dans les décisions de santé publique.

Remarque : tous les calculs de mesure d'impact n'ont de sens que s'il existe une relation causale spécifique démontrée entre le facteur de risque étudié et la maladie.

I. FRACTION ÉTIOLOGIQUE DU RISQUE

C'est la proportion de cas que l'on peut attribuer au facteur de risque qu'on étudie. On peut calculer cette proportion dans le groupe des exposés lors d'une enquête ou par extrapolation en population générale.

1. Fraction étiologique chez les exposés : FE_e

C'est, parmi un groupe exposé à un facteur de risque, la proportion de cas qu'on peut attribuer au facteur.

Si on appelle I_e : le risque chez les exposés (incidence cumulée chez les exposés) ;
I_{ne} le risque chez les non exposés (incidence cumulée chez les non exposés).

$$FE_e = \frac{I_e - I_{ne}}{I_e}$$

Formulations équivalentes

Enquête de cohorte

$$FE_e = \frac{RR - 1}{RR}$$

Enquête cas-témoins*

$$FE_e = \frac{OR - 1}{OR}$$

* sous réserve des conditions d'utilisation de l'OR (*cf.* § 16.4.4).

Résultat : la FE_e s'exprime par un chiffre compris entre 0 et 1 ou par un pourcentage.

Signification : la FE_e mesure la part de l'imputabilité du facteur dans la survenue de la maladie chez les exposés.

- Une FE_e proche de zéro est l'équivalent d'un RR ou un OR proche de 1 : le facteur étudié ne joue aucun rôle dans la survenue de la maladie.
- Une FE_e égale à 70 %, signifie que la maladie est imputable au facteur étudié chez 70 % des cas exposés (exemple 18.1).

Exemple 18.1. FRACTION ÉTIOLOGIQUE CHEZ LES EXPOSÉS

Une enquête de cohorte sur le rôle du tabac dans la survenue du cancer du poumon a montré un risque relatif de 21,7.

On a $FE_e = \frac{21,7 - 1}{21,7} = 0,954$.

Ce résultat signifie que 95,4 % des cancers du poumon chez les fumeurs sont attribuables au tabagisme.

2. Fraction étiologique dans la population FE_p

C'est, dans la population générale, la proportion de cas qu'on peut attribuer au facteur de risque étudié.

Si on connaît la proportion P_e de sujets exposés dans la population générale, on a :

Enquête de cohorte	Enquête cas-témoins*
$FE_p = \frac{P_e(RR - 1)}{P_e(RR - 1) + 1}$	$FE_p = \frac{P_e(OR - 1)}{P_e(OR - 1) + 1}$

* sous réserve des conditions d'utilisation de l'OR (*cf.* § 16.4.4)

Résultat : la FE_p s'exprime par un chiffre compris entre 0 et 1 ou un pourcentage.

Signification : la FE_p mesure le poids du facteur de risque sur la survenue de la maladie dans la population. En d'autres termes, elle mesure la proportion de cas évitables si le facteur de risque était supprimé (exemple 18.2).

Exemple 18.2. FRACTION ÉTIOLOGIQUE DANS LA POPULATION

Si l'on estime qu'il existe 30 % de fumeurs dans la population totale, en reprenant l'exemple 18.1 où le risque relatif de cancer du poumon lié au tabac était de 21,7.

On a : $FE_p = \frac{0,3 \times (21,7 - 1)}{0,3 \times (21,7 - 1) + 1} = 0,86$.

Ce résultat signifie que 86 % des cancers du poumon sont attribuables au tabagisme.
Ce résultat, comme celui de l'exemple précédent, implique que le RR de 21,7 soit bien établi, sans biais de confusion et que la relation soit causale.

Exemple 18.3. COMPARAISON DES FRACTIONS ÉTIOLOGIQUES CHEZ LES EXPOSÉS ET DANS LA POPULATION

Soit une population fictive dans laquelle la moitié des sujets sont exposés à un risque et la moitié sont non-exposés (*cf.* figure 18.1). Parmi les exposés, l'incidence des cas I_e est de 50 % et parmi les non-exposés, I_{ne} = 25 %.
On obtient par le calcul : RR = 2 , FE_e = 50 % et FE_p = 33,3 %.
Ce résultat est illustré par la figure 18.1 : si le facteur d'exposition n'avait aucune influence, la proportion de cas serait identique chez les exposés et les non-exposés (secteurs colorés en rouge). La part due au facteur d'exposition, colorée en noir, indique les cas en excès chez les exposés. Cette part représente dans cet exemple la moitié des cas exposés (FE_e), et un tiers du total des cas (FE_p).

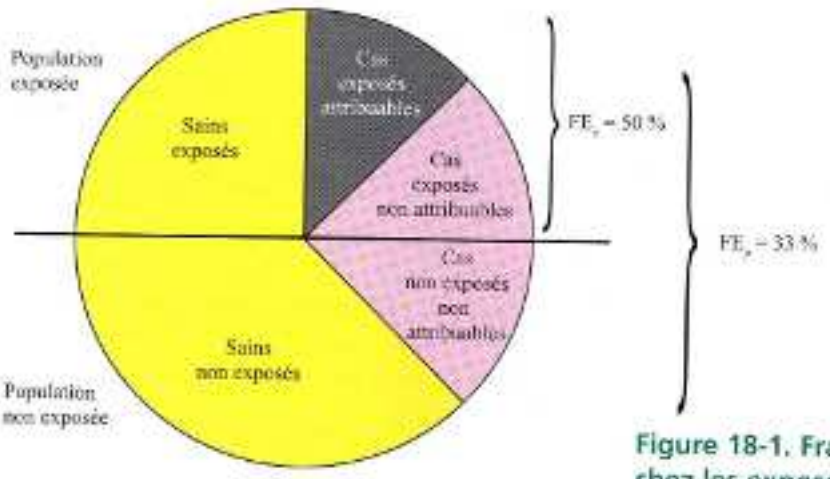


Figure 18-1. Fractions étiologiques d'un risque chez les exposés et dans la population

II. FRACTION PRÉVENTIVE

On calcule des fractions préventives de manière analogue aux fractions étiologiques, lorsque le facteur étudié est un facteur protecteur : vaccin, mesure de prévention, chimioprophylaxie, *etc.*).

1. Fraction préventive chez les exposés FP_e

On s'intéresse dans cette situation à la proportion de cas évités par le facteur protecteur chez les sujets exposés. Dans cette situation, le risque relatif (ou l'odds ratio) attendu est inférieur à 1.

Si on appelle :
I_e : le risque chez les exposés (incidence cumulée chez les exposés) ;
I_{ne} le risque chez les non exposés (incidence cumulée chez les non exposés).

$$FP_e = \frac{I_{ne} - I_e}{I_{ne}}$$

Formulations équivalentes

Enquête de cohorte	Enquête cas-témoins*
$FP_e = 1 - RR$	$FP_e = 1 - OR$

* sous réserve des conditions d'utilisation de l'OR (*cf.* § 16.4.4).

Résultat : la FP_e s'exprime par un chiffre compris entre 0 et 1 ou un pourcentage.

Signification : la FP_e donne chez les sujets protégés, la proportion de cas évités imputable au facteur de protection.

- Une FP_e proche de zéro est l'équivalent d'un RR ou un OR proche de 1 : le facteur étudié ne joue aucun rôle dans la prévention de la maladie.
- Une FP_e égale à 70 %, signifie que parmi les sujets soumis au facteur de protection, 70 % des cas potentiels ont été évités grâce à son action.

Lorsque le facteur protecteur est un vaccin, le terme FP_e est appelé **efficacité vaccinale** (exemple 18.4).

Exemple 18.4. FRACTION PRÉVENTIVE CHEZ LES EXPOSÉS

Une enquête sur l'efficacité d'un vaccin a montré les résultats suivants :

EXPOSITION	INCIDENCE POUR MILLE	RR
vaccinés	50	0,25
non vaccinés	200	

On a $FP_e = 1 - 0,25 = 0,75$. Ce résultat signifie que parmi les vaccinés, 75 % des cas potentiels ont été évités grâce au vaccin.

2. Fraction préventive dans la population FP_p

La fraction préventive dans la population donne la proportion de cas de la population générale qu'on peut éviter en la soumettant au facteur protecteur.

Si on connaît P_e, la proportion de la population qui est exposée au facteur préventif, on a :

Enquête de cohorte	Enquête cas-témoins*
$FP_p = P_e(1 - RR)$	$FP_p = P_e(1 - OR)$

* sous réserve des conditions d'utilisation de l'OR (*cf.* chap. 16.IV.4).

Résultat : la FP_p s'exprime par un chiffre compris entre 0 et 1 ou par un pourcentage.

Signification : la FP_p mesure la proportion de cas évités dans la population grâce au facteur protecteur (exemple 18.5).

Exemple 18.5. FRACTION PRÉVENTIVE DANS LA POPULATION

Si la proportion de sujets vaccinés est de 80 %, en reprenant l'exemple 18.3, on a $FP_p = 0{,}8 \times (1 - 0{,}25) = 0{,}6$. Ce résultat signifie que 60 % des cas potentiels de la maladie dans la population ont été évités grâce au vaccin.

Exemple 18.6. COMPARAISON DES FRACTIONS PRÉVENTIVES CHEZ LES SUJETS PROTÉGÉS ET DANS LA POPULATION

Soit une population fictive dont la moitié a été protégée par un vaccin et l'autre moitié non protégée. Parmi les vaccinés, l'incidence de la maladie est de 12,5 % et parmi les non-vaccinés l'incidence est de 50 % (*cf.* figure 18.2).
On obtient par le calcul : RR = 0,25, FP_v = 75 % et FP_p = 37,5 %.
Ce résultat est illustré par la figure 18.2 : si le vaccin était inefficace, la proportion de cas dans le groupe vacciné serait identique à celle du groupe non-vacciné (50 %). La part en grisé représente les cas évités grâce à l'action du vaccin. Parmi les vaccinés, cette part de cas évités représente les ¾ des cas « potentiels » (cas évités + cas réels). Si l'on considère l'ensemble de la population, cette part de cas évités représente 37,5 % des cas potentiels.

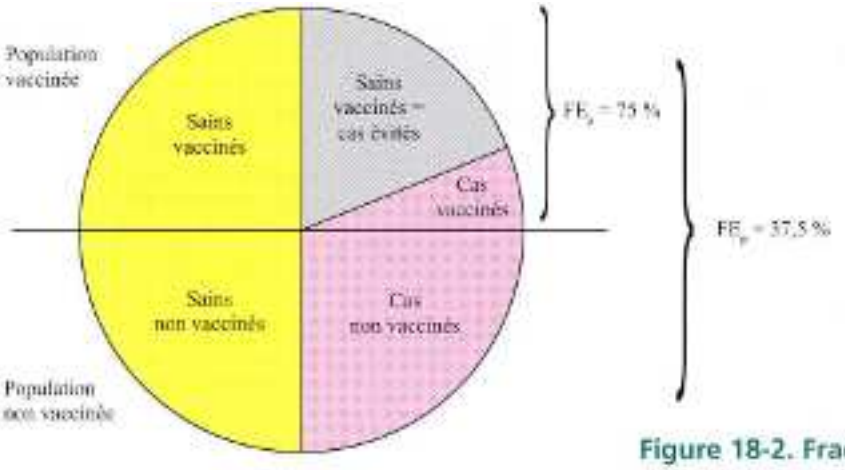


Figure 18-2. Fractions préventives chez les exposés (vaccin) et dans la population

III. INTERVALLE DE CONFIANCE DES FE ET FP

Tous les calculs de fraction étiologique et préventive chez les exposés et dans la population, ainsi que leur intervalle de confiance, peuvent être obtenus par analogie avec le cas particulier de l'efficacité vaccinale, quel que soit le facteur étudié, en utilisant EpiInfo/Epitable/Analyse/Efficacité

vaccinale/Cohorte ou Cas-Témoins. Il suffit de placer les sujets exposés dans les cases Vac+ et les non-exposés dans les cases Vac-. Dans le tableau des résultats les termes.
« Efficacité vaccinale » équivaut à « fraction préventive »
« Fraction attribuable » équivaut à « fraction étiologique »

Résumé

Les indicateurs de mesure d'impact servent à mesurer l'importance relative d'un facteur de risque ou d'un facteur protecteur.
Lorsque le facteur étudié est un facteur de *risque*, on mesure la *fraction étiologique* parmi les exposés ou dans la population. Cette proportion indique la part d'imputabilité du facteur de risques dans la survenue de la maladie.
Lorsque le facteur étudié est un facteur *protecteur*, on mesure la *fraction préventive* parmi les exposés ou dans la population. Cette proportion indique la proportion de cas évités grâce au facteur protecteur (exemple : vaccin).
Ces calculs de mesure d'impact ne sont valides que s'il existe une relation causale spécifique démontrée entre le facteur de risque étudié et la maladie.

FRACTION...	NOTATION		COHORTE	CAS-TÉMOINS
étiologique chez exposés	FE_e	$\frac{I_e - I_{ne}}{I_e}$	$\frac{RR - 1}{RR}$	$\frac{OR - 1}{OR}$
étiologique dans la population	FE_p		$\frac{P_e(RR-1)}{P_e(RR-1)+1}$	$\frac{P_e(OR-1)}{P_e(OR-1)+1}$
préventive chez les exposés	FP_e	$\frac{I_{ne} - I_e}{I_{ne}}$	$1 - RR$	$1 - OR$
préventive dans la population	FP_p		$P_e(1 - RR)$	$P_e(1 - OR)$

I_e et I_{ne} : risque (incidence cumulée) chez les exposés et les non-exposés.
P_e : proportion d'exposés dans la population.

STANDARDISATION DES TAUX

I. POSITION DU PROBLÈME

Supposons que l'on veuille comparer les taux bruts de mortalité de deux régions françaises, avec l'idée que certains facteurs de risque sont responsables d'une surmortalité dans une des deux régions. La simple comparaison des deux taux obtenus se heurte d'emblée à une difficulté : on sait en effet qu'indépendamment de ses causes, la mortalité est (hélas) liée à un facteur incontournable : l'âge. Si un des taux observés est plus élevé dans une des deux régions, cela peut être dû, non pas à des causes externes, mais tout simplement à une répartition différente des sujets selon l'âge. Si la région où la mortalité est la plus élevée possède la population la plus âgée, on n'aura tout juste démontré que les sujets âgés ont plus de probabilité de mourir que les jeunes !
Ainsi, pour pouvoir affirmer que la mortalité est plus élevée dans une des régions, il faudra soutenir qu'elle l'est, indépendamment de la structure par âge des régions comparées.
Dans ce type de problème, l'âge est un facteur de confusion. On résout cette difficulté en ajustant les taux sur le facteur de confusion.
La méthode employée est celle de la standardisation.

II. PRINCIPE

Nous continuerons à utiliser l'exemple de la mortalité et de l'âge comme facteur de confusion, car l'exemple est simple et évident. C'est d'ailleurs dans cette situation qu'on utilise le plus souvent cette méthode.
Pour pouvoir appliquer la méthode de standardisation des taux, il faut disposer des taux spécifiques observés dans chacune des classes du facteur de confusion. Il faut également disposer d'une population de référence qui va servir de point d'appui à la comparaison.
Prenons comme exemple deux régions A et B avec leurs taux de mortalité globale et par classes d'âge, pendant une année donnée. Pour simplifier, nous avons partagé les effectifs en larges classes d'âge, mais la méthode s'applique évidemment à des structures plus fines.

	Région A				Région B			
Classe d'âge	Effectifs N	%	Décès n	*Mortalité pour 1 000*	Effectifs N	%	Décès n	*Mortalité pour 1 000*
0-14	103 065	*14,5*	65	*0,63*	2 891 100	*26,4*	1 890	*0,65*
15-24	71 790	*10,1*	58	*0,81*	1 566 012	*14,3*	1 282	*0,82*
25-44	194 046	*27,3*	92	*0,47*	3 318 194	*30,3*	1 832	*0,55*
45-64	166 325	*23,4*	2332	*14,02*	2 168 325	*19,8*	33 753	*15,57*
> 64	175 566	*24,7*	6672	*38,00*	1 007 505	*9,2*	39 544	*39,25*
Total	710 792	*100,0*	9219	*12,97*	10 951 136	*100,0*	78 301	*7,15*

On constate que le taux de mortalité globale pour l'année considérée est de 12,97 pour mille dans la région A et de 7,15 pour mille dans la région B.
On constate parallèlement que la population A est plus âgée que la population B : 24,7 % des sujets de A ont plus de 64 ans contre seulement 9,2 % en B.
La question est de savoir si la mortalité est réellement plus élevée en A qu'en B, indépendamment de l'âge.
Il existe deux méthodes de standardisation : la méthode directe et la méthode indirecte.

III. MÉTHODE DIRECTE

Le principe consiste à calculer sur une population de référence, les taux de mortalité qu'on aurait observé *si les deux régions possédaient la même structure par âge que cette population*.
On calcule ainsi des ***taux de mortalité standardisés (TMS)***.
La population de référence naturelle pour deux régions françaises est de prendre la population de l'ensemble du pays.

1. Calcul

On calcule pour chaque région :
- le nombre de décès attendus (théoriques) pour chaque classe d'âge : effectif de la population de référence multiplié par le taux de mortalité spécifique dans la classe d'âge ;
- le nombre total de décès attendus : somme des décès attendus de chaque classe d'âge ;
- le TMS : nombre total de décès attendus divisé par l'effectif total de la population de référence.

Population de référence		Région A		Région B	
Classe d'âge	Effectifs N	Mortalité pour 1 000	Décès attendus[1]	Mortalité pour 1 000	Décès attendus[1]
0-14	11 178 318	*0,63*	7 042	*0,65*	7 266
15-24	7 743 422	*0,81*	6 272	*0,82*	6 350
25-44	17 286 620	*0,47*	8 125	*0,55*	9 508
45-64	13 462 806	*14,02*	188 749	*15,57*	209 616
> 64	9 295 668	*38,00*	353 235	*39,25*	364 855
Total	58 966 834	TMS_A	563 423	TMS_B	597 595

1. Décès attendus = (taux de mortalité spécifique de la classe d'âge dans la région/1 000) x N.

Dans la région A, TMS_A = 563 423/58 966 834 = 9,55 pour mille.
Dans la région B, TMS_B = 597 595/58 966 834 = 10,13 pour mille.

2. Interprétation

On constate que ces deux taux sont équivalents. La mortalité n'est pas plus élevée dans la région A lorsqu'on élimine le facteur âge. Elle paraît, au contraire, être légèrement plus basse.
L'interprétation de ce résultat doit être soigneusement réfléchie. La réalité mesurée par les taux bruts montre à l'évidence que la mortalité brute est plus élevée dans la région A. Mais, dans l'exemple qui

a été pris, cette différence ne s'explique que parce que la structure d'âge est en défaveur de cette région. On aurait pu d'ailleurs constater dans le premier tableau que les taux spécifiques par classe d'âge étaient très proches.
Si la différence entre les taux standardisés de mortalité était significative, on pourrait conclure que la mortalité ajustée sur l'âge est plus élevée dans la région B. Il faudrait alors en rechercher les causes.

IV. MÉTHODE INDIRECTE

Le principe est de calculer le nombre de décès attendus dans chaque groupe de comparaison *si les taux de mortalité spécifiques avaient été ceux de la population de référence*. On compare ensuite le nombre réel de décès observés au nombre attendu en calculant ***un ratio standardisé de mortalité (RSM)***.
Le produit de ces ratio par le taux brut de mortalité dans la population de référence donne les **taux de mortalité standardisés (TMS)**.
Cette méthode nécessite de connaître les taux spécifiques de mortalité dans la population de référence. Elle est intéressante à utiliser lorsque les effectifs des groupes à comparer sont trop petits pour pouvoir calculer des taux spécifiques précis.

On calcule pour chaque région :
- le nombre de décès attendus dans chaque classe d'âge : effectif de la classe multiplié par le taux spécifique de mortalité de cette classe d'âge dans la population de référence ;
- la somme des décès attendus ;
- le $\text{RSM} = \dfrac{\text{nombre de décès observés}}{\text{nombre de décès attendus}}$;
- le TMS : RSM multiplié par le taux brut de mortalité dans la population de référence.

Population référence		Région A			Région B		
Classe d'âge	Mortalité pour 1 000	Effectifs N	Décès observés	Décès[1] attendus	Effectifs N	Décès observés	Décès[1] attendus
0-14	*0,62*	103 065	65	63,9	2 891 100	1 890	1 792,5
15-24	*0,73*	71 790	58	52,4	1 566 012	1 282	1 143,2
25-44	*0,39*	194 046	92	75,7	3 318 194	1 832	1 294,1
45-64	*13,30*	166 325	2 332	2 212,1	2 168 325	33 753	28 838,7
> 64	*36,77*	175 566	6 672	6 455,6	1 007 505	39 544	37 046,0
Total	*9,16*	710 792	9 219	8 859,7	10 951 136	78 301	70 114,5

1. Décès attendus = (Taux de mortalité spécifique de référence/1 000) x effectif de la classe d'âge de la région.

$\text{RSM}_A = 9\,219 / 8\,859{,}7 = 1{,}041$
$\text{RSM}_B = 78\,301 / 70\,114{,}5 = 1{,}117$

$\text{TMS}_A = 1{,}041 \times 9{,}16 = 9{,}53$ pour mille
$\text{TMS}_B = 1{,}117 \times 9{,}16 = 10{,}23$ pour mille

On constate que les taux de mortalité standardisés calculés par cette méthode sont quasiment identiques à ceux de la méthode directe.
L'interprétation des résultats est identique.

V. CONDITIONS D'APPLICATION

- La standardisation ne peut être utilisée que si on dispose de la même structure de données démographiques dans les groupes et dans la population de référence.
- Dans la standardisation directe, si on ne dispose pas de population de référence, on peut prendre comme référence le total des effectifs des groupes de comparaison. On peut également prendre une distribution de référence fictive. Dans ce cas les taux ajustés, ne correspondent à aucune réalité. Ils permettent simplement d'effectuer la comparaison.
- La standardisation ne peut être utilisée que si les taux spécifiques sont tous inférieurs ou supérieurs dans chaque groupe de comparaison. En d'autres termes, il faut que les différences entre les taux spécifiques soient à peu près constantes. Notamment, il ne faut pas de croisement entre les taux, comme, par exemple, un taux spécifique chez les jeunes du groupe A supérieur à ceux de B et un taux chez les sujets âgés de A inférieur aux sujets âgés de B.

VI. EXTENSION DE LA MÉTHODE

On peut utiliser les deux méthodes de standardisation des taux :
- sur d'autres paramètres que la mortalité : prévalence, incidence, *etc.* ;
- en standardisant sur tout autre facteur de confusion que l'âge : sexe, catégorie socio-professionnelle, niveau d'étude, *etc.* ;
- en comparant plusieurs groupes.

Résumé

On utilise les méthodes de standardisation des taux lorsqu'on désire ajuster un taux brut calculé dans une population en tenant compte d'un tiers facteur qui joue un rôle sur le caractère étudié. La population étudiée est divisée selon les classes de ce tiers facteur. On calcule alors un taux standardisé qui donne une mesure de la fréquence du caractère étudié indépendamment du tiers facteur.
L'intérêt de calculer des taux standardisés est de pouvoir comparer deux ou plusieurs populations dont la structure varie selon le tiers facteur.
L'utilisation la plus fréquente de cette méthode est l'étude de la mortalité dans une population qu'on désire connaître indépendamment de sa structure par âge.

ANALYSE DE SURVIE

L'analyse de survie est une méthode permettant de synthétiser les probabilités de survenue d'un événement chez des sujets ayant en commun un événement d'origine, en tenant compte du délai écoulé entre ces deux événements.

L'exemple type est l'étude de la survenue de décès chez des individus ayant fait l'objet d'un diagnostic de maladie grave. On peut également utiliser ce type d'analyse pour étudier le délai de survenue de décès chez des malades bénéficiant d'un traitement particulier.

L'analyse de survie peut servir également à l'étude de tout autre événement qu'un décès : rechute d'une maladie après traitement initial, survenue d'une grossesse après traitement stimulant, et même guérison d'une maladie après traitement.

Malgré ses différents champs d'application, on garde le terme de « survie » pour définir la variable quantitative mesurant, chez un individu, le délai écoulé entre un événement initial et un événement final.

Comme il s'agit d'une variable quantitative, on pourrait limiter l'étude descriptive à un simple calcul de moyenne et de variance de la durée de survie dans le groupe d'individus étudiés. L'analyse de survie permet d'affiner l'analyse en tenant compte de la cinétique des événements.

Si un seul groupe est suivi, l'analyse de survie est simplement descriptive. Si plusieurs groupes sont suivis, l'analyse de survie permet une étude comparative, par exemple entre traitements ou entre facteurs de risque.

I. PRINCIPE

Une analyse de survie a pour principe d'estimer la probabilité de survie à différents intervalles de temps. On appelle fonction de survie S_i à un instant donné la probabilité d'être vivant à cet instant.

$$S_i = \frac{\text{Nombre de survivants à un instant } t_i}{\text{Nombre d'individus suivis}}$$

Cet indicateur serait aisé à calculer si on connaissait le statut vivant/décédé de tous les sujets au moment de chaque mesure. En pratique, les données sont souvent incomplètes car il existe des perdus de vue pour lesquels on ne connaît pas le statut. On appelle les données concernant les perdus de vue, ***données censurées***.

Période d'étude

On distingue trois dates :

- la date du début de l'étude lorsqu'on commence à inclure les patients ;
- la date de clôture de l'inclusion des patients ;
- la date de fin de suivi de l'étude. Cette date peut être fixée à l'avance. Dans ce cas, les individus inclus tardivement ne seront pas suivis pendant la même durée. On peut également ne pas la fixer et attendre l'événement final pour tous les sujets de l'étude.

Événement initial

Pour chaque individu, l'événement initial survient à l'instant t_0.

Intervalles d'observation

Ils sont définis par des instants t_i, pour lesquels on effectue les bilans intermédiaires de la survie.

Vivant

C'est un individu n'ayant pas encore subi l'événement final à un instant t_i.

Décédé

C'est un individu ayant subi l'événement final pendant l'intervalle précédent.

Exclu

C'est un individu dont on ne connaît plus le statut vivant/décédé à un instant t_i, soit parce qu'il est perdu de vue, soit parce que l'enquête s'est terminée avant que l'individu ait subi l'événement final.
Les instants t_i servant à calculer la courbe de survie sont définis selon deux méthodes possibles : la méthode de Kaplan-Meier et la méthode actuarielle.

II. MÉTHODE DE KAPLAN-MEIER

Dans cette méthode, on calcule la probabilité de survie à chaque fois qu'au moins un décès est enregistré.

Si on appelle :

- V_i : le nombre de vivants au début de l'intervalle de temps $t_i - t_{i-1}$;
- D_i : le nombre de décédés pendant l'intervalle $t_i - t_{i-1}$;
- E_i : le nombre de perdus de vue (exclus) au début de l'intervalle $t_i - t_{i-1}$;
- q_i : la probabilité de décès pendant l'intervalle $t_i - t_{i-1}$: $q_i = D_i/(V_i - E_i)$;
- p_i : la probabilité de survie pendant l'intervalle $t_i - t_{i-1}$: $p_i = 1 - q_i$;
- S_i : la fonction de survie à l'instant t_i : $S_i = p_0 p_1 \ldots p_i = p_i S_{i-1}$.

TEMPS	VIVANTS	DÉCÉDÉS	EXCLUS	PROB (DÉCÈS)	PROB (SURVIE)	FONCTION DE SURVIE
t_0	V_0	0	0	$q_0 = 0$	$p_0 = 1$	$S_0 = 1$
t_1	V_1	D_1	E_1	$q_1 = D_1/(V_1 - E_1)$	$p_1 = 1 - q_1$	$S_1 = S_0 p_1$
t_2	V_2	D_2	E_2	$q_2 = D_2/(V_2 - E_2)$	$p_2 = 1 - q_2$	$S_2 = S_1 p_2$
...	...	...	...	...	...	...
t_i	V_i	D_i	E_i	$q_i = D_i/(V_i - E_i)$	$p_i = 1 - q_i$	$S_i = S_{i-1} p_i$

L'analyse de la fonction de survie par la méthode de Kaplan-Meier s'exprime par un graphique portant en abscisses le temps et en ordonnées la probabilité de survie (figure 20.1). La courbe est en marche d'escalier descendant. Chaque plateau représente la probabilité de survie pendant l'intervalle. Le montant de la marche représente la chute de la probabilité de survie due au décès constaté à l'instant t_i (exemple 20.1).

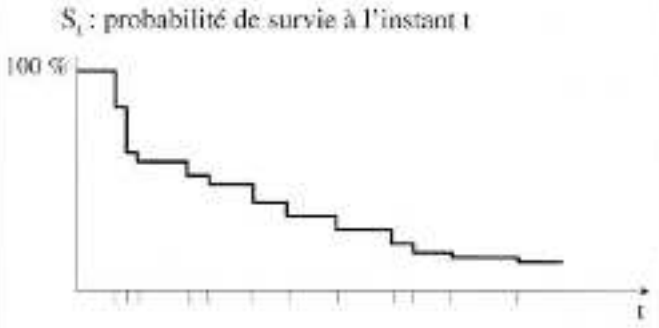


Figure 20-1. Analyse de survie. Méthode de Kaplan-Meier

Exemple 20.1. MÉTHODE DE KAPLAN-MEIER

On veut analyser la survie après une 2e cure d'un groupe de 23 patients ayant rechuté après une première cure d'un traitement A.
L'événement initial dans cet exemple, est l'administration d'une seconde cure du traitement A. L'événement final est le décès du patient.

Jour de décès	vivants	décédés	exclus	q	p	S
50	23	1	0	0,043	0,957	0,957
100	22	8	0	0,364	0,636	0,609
130	14	1	0	0,071	0,929	0,565
135	13	1	0	0,077	0,923	0,522
140	12	1	0	0,083	0,917	0,478
152	11	1	0	0,091	0,909	0,435
165	10	1	0	0,100	0,900	0,391
170	9	1	0	0,111	0,889	0,348
258	8	1	0	0,125	0,875	0,304
305	7	1	0	0,143	0,857	0,261
329	6	1	0	0,167	0,833	0,217
365	5	1	1	0,250	0,750	0,163
424	3	1	0	0,333	0,667	0,109
445	2	1	0	0,500	0,500	0,054

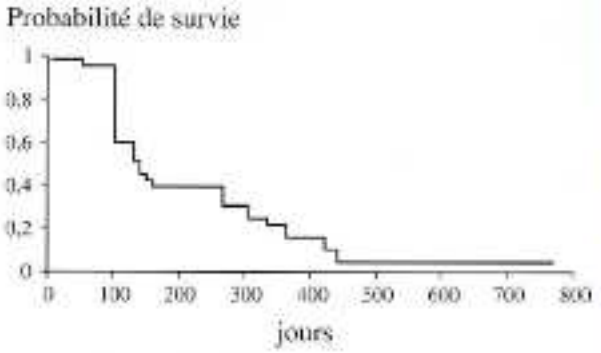


Figure 20-2.

III. LA MÉTHODE ACTUARIELLE

Elle se différencie de la méthode de Kaplan-Meier par deux points.

- Les intervalles de temps entre deux mesures sont des intervalles de temps systématiques.
- On considère que les exclus ont été suivis en moyenne pendant la moitié de l'intervalle. On divise donc leur nombre par deux. La probabilité de décès pendant l'intervalle est donc égale à D/(V-1/2 E).

L'analyse de la fonction de survie par la méthode actuarielle s'exprime par un graphique portant en abscisses le temps divisé en intervalle régulier, et en ordonnées la probabilité de survie (figure 20.3). Chaque point de la courbe représente la probabilité moyenne de survie pendant l'intervalle.

Figure 20-3. Analyse de survie. Méthode actuarielle

IV. COMPARAISON DE COURBES DE SURVIE : TEST DU LOG RANK

L'intérêt des courbes de survie réside surtout dans la comparaison de plusieurs courbes. La comparaison de deux courbes de survie est utilisée principalement pour comparer deux traitements ou pour comparer deux groupes exposés à des facteurs de risque (figure 20.4).

Le test statistique couramment utilisé à cet effet est le test du **log rank** qui s'applique lorsque les deux courbes de survie sont calculées par la méthode de Kaplan-Meier. La méthode figure en Annexe, Formulaire 22.

Sous H_0, les deux courbes de survie ont des profils identiques. Sous H_1 les deux courbes ont des profils différents.

Le rejet de H_0 signifie que la survenue de l'événement étudié est en moyenne plus tardive dans l'un des deux groupes.

Lorsqu'on désire prendre en compte un ou plusieurs tiers facteurs pouvant influer sur la survie (analyse multivariée), on utilise le modèle de Cox (proportionnal hazard model).

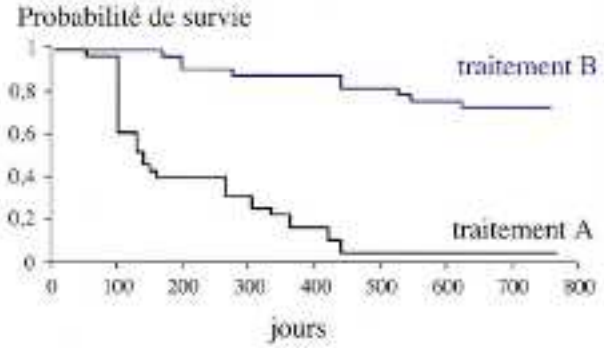


Figure 20-4.

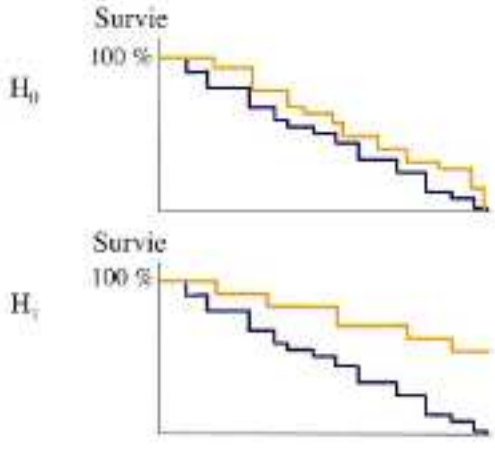


Figure 20-5. Test du log rank

Exemple 20.2. TEST DU LOG RANK

On veut comparer le taux de guérison d'une maladie entre deux groupes de patients traités par deux médicaments différents (traitement 1 et 2). L'évolution de la maladie est suivie quotidiennement par un test biologique. La guérison est affirmée par un examen biologique négatif.

Dans ce type de problème, l'événement initial est le début du traitement et l'événement final (qu'on aurait appelé « décès » dans un problème de survie) est la guérison.
Les sujets n'ayant pas encore subi l'événement final (qu'on aurait appelé « vivants » dans un problème de survie) sont les sujets encore malades (V).
Les sujets ayant subi l'événement final (qu'on aurait appelé « décédés » dans un problème de survie) sont les sujets guéris (D).

t	Traitement 1		Traitement 2		c_{1i}
	V_1	D_1	V_1	D_2	
1	20	2	20	0	1,00
2	18	2	20	1	1,42
3	16	1	19	0	0,46
4	15	3	19	1	1,76
5	12	0	18	2	0,80
6	12	2	16	0	0,86
7	10	2	16	0	0,77
8	8	1	16	0	0,33
9	7	0	16	3	0,91
10	7	1	13	0	0,35
11	6	0	13	2	0,63
12	6	0	11	2	0,71
13	6	1	9	0	0,40
14	5	0	9	2	0,71
15	5	2	7	0	0,83
16	3	2	7	0	0,60
17	1	0	7	1	0,13
18	1	0	6	2	0,29
19	1	0	4	2	0,40
20	1	1	2	0	0,33
Total		20		18	13,69

On obtient le tableau de résultats ci-contre.
Les effectifs théoriques c_{1i} s'obtiennent dans chaque ligne en calculant :
le terme $V_1 (D_1 + D_2)/(V_1 + V_2)$
(*cf.* Annexes, Formulaire § 22).
La somme des effectifs théoriques est $c_1 = 13,69$
La somme des effectifs théoriques c_2 s'obtient par différence

$c_2 = 20 + 18 - 13,69 = 24,31$

$$\chi^2 = \frac{(20 - 13,69)^2}{13,69} + \frac{(18 - 24,31)^2}{24,31} = 4,55$$

La différence entre les deux traitements est significative ($p < 0,04$), alors que la simple comparaison entre les deux taux de guérison (100 % versus 90 %) aurait montré une différence non significative (test de χ^2 à 4 cases = 2,1).
La comparaison des deux courbes de survie par cette méthode permet d'affirmer que traitement 1 guérit plus rapidement les patients.

Résumé

Les méthodes d'analyse de survie s'appliquent pour mesurer la probabilité d'un événement dans une population en tenant compte, pour chaque individu, du délai écoulé entre le début du suivi et la survenue de cet événement. L'événement étudié peut être le décès, mais aussi tout autre type de phénomène. On utilise la méthode de Kaplan-Meier ou la méthode actuarielle. Ces méthodes s'expriment par des graphes qui montrent les probabilités de « survie » au cours du temps.
Lorsqu'on désire comparer des courbes de survie entre deux groupes de sujets, on peut utiliser le test statistique du log rank.

PERFORMANCES D'UNE TECHNIQUE

Ce chapitre, traite de l'évaluation de techniques diagnostiques utilisées en pratique clinique ou biologique. Nous utiliserons plutôt le mot « test » pour technique. On entendra par « test », une technique diagnostique au sens large (on ne parle pas ici de test statistique). Un test peut être :

- soit la recherche d'un signe clinique ;
- soit une technique paraclinique (examen biologique, cliché d'imagerie, examen physiologique, *etc.*) ;
- soit une combinaison de plusieurs signes cliniques ou paracliniques ;
- soit toute évaluation d'une performance.

En épidémiologie, on peut aussi assimiler à un test, un système d'alerte dont on désire étudier les performances dans sa capacité à détecter les épidémies. Dans cette situation, un « cas » est une épidémie.

Schématiquement, le résultat d'un test s'exprime :

- soit par une variable qualitative binaire : test positif ou négatif ;
- soit par une variable quantitative : valeur d'une mesure biologique, note, indice, *etc.*

Les performances d'un test doivent être successivement mesurées de façon expérimentale, et évaluées en situation réelle sur le terrain. Les deux procédures sont différentes.

I. MESURE EXPÉRIMENTALE DES PERFORMANCES D'UN TEST

Un test doit posséder deux qualités majeures : la sensibilité et la spécificité. Il s'agit des qualités « intrinsèques » du test.

1. Sensibilité

La sensibilité d'un test est sa capacité à détecter les cas d'une maladie.
Pour mesurer la sensibilité, il faut donc disposer d'un groupe de malades. Ce groupe de malades doit avoir été préalablement sélectionné par des méthodes indiscutables qui permettent de certifier la présence de la maladie. Ces méthodes d'inclusion des cas doivent impérativement être ***indépendantes*** du test qu'on cherche à analyser.
Si nous prenons l'exemple simple d'un test qualitatif appliqué à une série de cas, deux situations vont se présenter en fonction des résultats :

- résultat positif : ce sont les vrais positifs (VP) ;
- résultat négatif : ce sont des faux négatifs (FN).

En termes mathématiques, la sensibilité est la proportion de vrais positifs sur le nombre total de cas :

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

La sensibilité est donc un nombre compris entre 0 et 1. On l'exprime en général en pourcentage.

Exemple 21.1. SENSIBILITÉ

On veut tester la sensibilité d'un test de dépistage de la toxoplasmose congénitale. On dispose d'un groupe de 58 prélèvements, correspondant à des enfants nés ultérieurement et atteints de façon certaine de toxoplasmose congénitale. Parmi eux, le test a été positif dans 54 cas.

La sensibilité est Se = 54/58 = 93,1 %

2. Spécificité

La spécificité d'un test est sa capacité à identifier correctement les individus qui ne sont pas atteints par la maladie.

Pour mesurer la spécificité, il faut donc disposer d'un groupe de sujets sains. Ce groupe de sujets doit avoir été préalablement sélectionné par des méthodes indiscutables qui permettent de certifier l'absence de la maladie. Ces méthodes d'inclusion des sujets sains doivent impérativement être ***indépendantes*** du test qu'on cherche à analyser.
Si nous prenons l'exemple simple d'un test qualitatif appliqué à une série de cas, deux situations vont se présenter en fonction des résultats :

- résultat négatif : ce sont les vrais négatifs (VN) ;
- résultat positif : ce sont des faux positifs (FP).

En termes mathématiques, la spécificité est la proportion de vrais négatifs sur le nombre total de sujets sains :

$$Sp = \frac{VN}{VN + FP}$$

La spécificité est donc un nombre compris entre 0 et 1. On l'exprime en général en pourcentage.

Exemple 21.2. SPÉCIFICITÉ

On veut tester la spécificité d'un test de dépistage de la toxoplasmose congénitale. On dispose d'un groupe de 125 prélèvements, correspondant à des enfants nés ultérieurement et indemnes de façon certaine de toxoplasmose congénitale. Parmi eux, le test a été négatif dans 114 cas.

La spécificité est Sp = 114/125 = 91,2 %

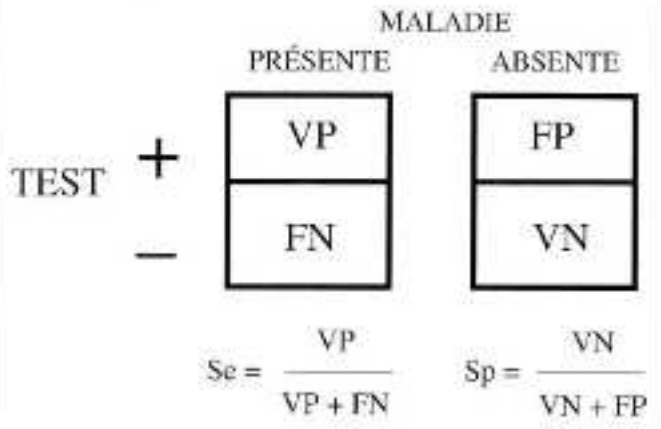


Figure 21-1.

3. Cas d'un test quantitatif

Le test peut se traduire par un résultat exprimé sous forme d'une valeur numérique. En raison de la variabilité biologique, ces valeurs sont différentes d'un sujet à l'autre. Si une série de sujets est examinée, les résultats vont s'afficher sous forme d'une distribution. Le problème est donc de déterminer une valeur seuil qui permettra de classer les malades et les sujets sains.

Si le test est parfaitement discriminant, la distribution des valeurs dans le groupe des cas, sera bien séparée de la distribution des valeurs dans un groupe de sujets sains (figure 21.2). Il sera alors aisé de choisir une valeur seuil qui permettra une sensibilité et une spécificité de 100 %.
Hélas, cette situation est rarement observée en biologie médicale.

Le plus souvent, il y a chevauchement des deux distributions (figure 21.3). Certains sujets sains présentent des valeurs qui peuvent être identiques à celle de sujets malades et à l'inverse certains malades présentent des valeurs qui peuvent être identiques à celles de sujets sains. Le choix d'un seuil devient une opération délicate, car il divise les deux groupes de sujets en 4 sous-groupes, VP, FN, VN, FN.

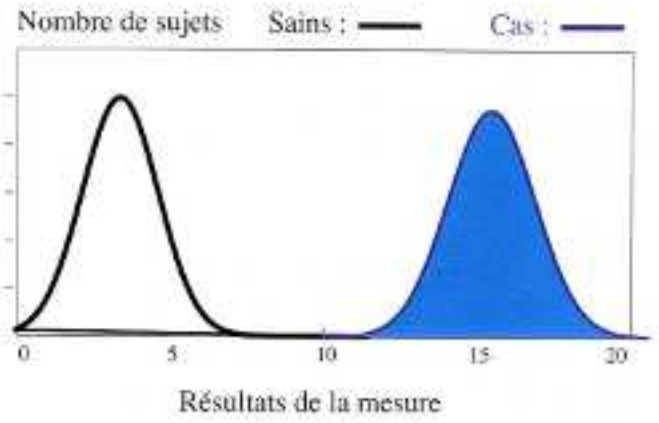


Figure 21-2. Test quantitatif : distribution des valeurs observées chez les cas et les sujets sains

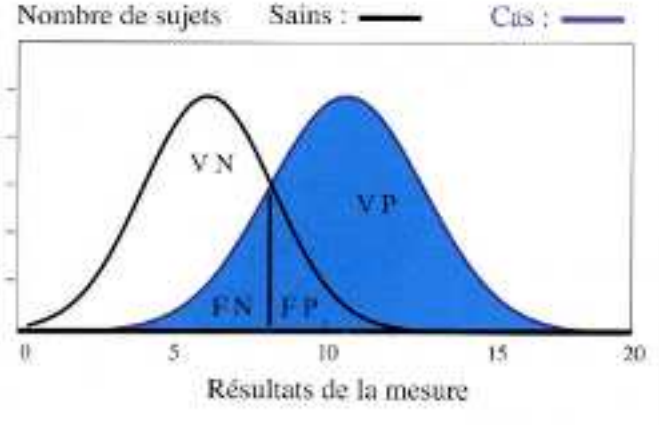


Figure 21-3. Test quantitatif : distribution des valeurs observées chez les cas et les sujets sains

On conçoit que les qualités diagnostiques du test vont varier selon le seuil choisi.
Si on décide de baisser le seuil, le nombre de faux négatif diminue, donc la sensibilité augmente. Mais, parallèlement, le nombre de faux positifs augmente et la spécificité diminue (figure 21.4).
Si, à l'inverse, on décide d'élever le seuil, le nombre de faux positifs diminue, donc la spécificité augmente, mais, parallèlement, le nombre de faux négatif augmente, donc la sensibilité diminue (figure 21.5).
On constate que sensibilité et spécificité d'un test quantitatif varient en sens inverse. Le choix d'un seuil est le résultat d'un compromis qui est fonction de l'objectif assigné au test.

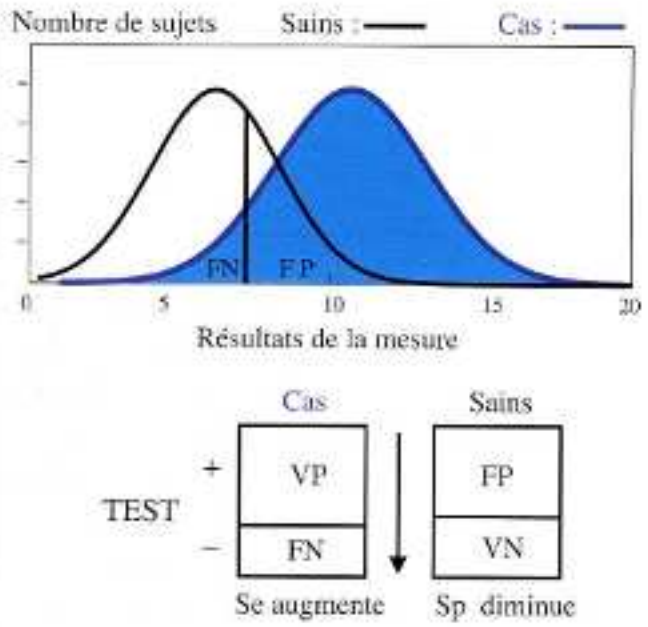


Figure 21-4. Baisse du seuil

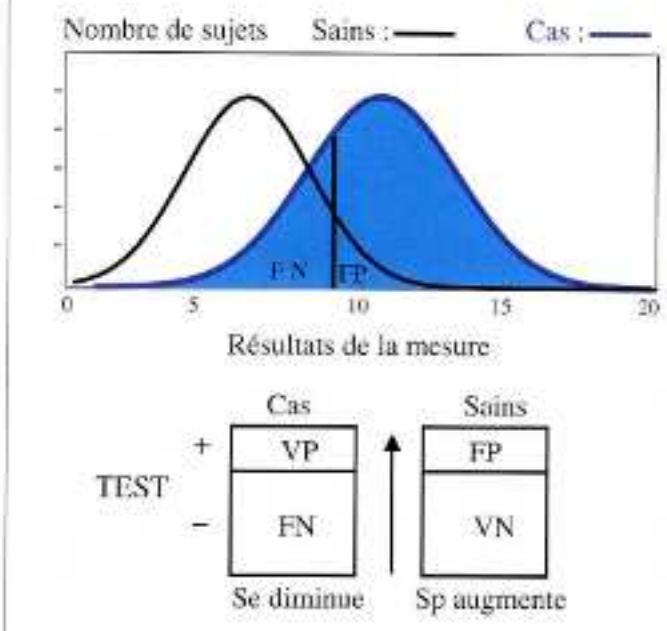


Figure 21-5. Hausse du seuil

4. Choix d'un seuil

- Lorsque les erreurs par excès sont plus graves que les erreurs par défaut, on cherche à minimiser le nombre de faux positifs, donc à privilégier la spécificité. Il faut donc hausser le seuil de positivité (exemple 21.3).

Exemple 21.3. EXIGENCE D'UNE BONNE SPÉCIFICITÉ

Dans le dépistage anténatal de la toxoplasmose congénitale ou d'anencéphalie, un dépistage faussement positif entraîne des conséquences très lourdes (mise sous traitement prolongé ou interruption de grossesse). À l'inverse, un faux résultat négatif, pourra être rattrapé ultérieurement par le suivi échographique.

- Lorsque les erreurs par défaut sont plus graves que les erreurs par excès, on cherche à minimiser le nombre de faux négatifs, donc à privilégier la sensibilité. Il faut donc baisser le seuil de positivité (exemple 21.4).

Exemples 21.4. EXIGENCE D'UNE BONNE SENSIBILITÉ

- Dans le dépistage de la phénylcétonurie à la naissance (test de Guthrie), un dépistage faussement négatif entraînera le développement de la maladie chez l'enfant. À l'inverse, un résultat faussement positif n'aura entraîné chez l'enfant qu'une prévention inutile qui pourra être corrigée par la suite.
- Dans le dépistage pré-transfusionnel du paludisme dans les flacons de la banque du sang, un dépistage faussement négatif entraînera l'administration de sang parasité chez un receveur, alors qu'un résultat faussement positif n'entraînera que le rejet à tort du flacon.

Courbe ROC

Lorsqu'on cherche à fixer le seuil d'une méthode quantitative, on applique le test à un groupe de malades et un groupe de sujets sains. Pour chaque seuil possible, on calcule la sensibilité et la spécificité. On obtient ainsi une liste de couples Se-Sp. Une méthode, permettant de visualiser de façon synthétique les différentes performances en fonction des seuils choisis, consiste à dessiner un graphe appelé courbe ROC (*receiver operating characteristics*). On porte en ordonnée la sensibilité de chaque seuil et en abscisses le pourcentage de faux positifs (1-Sp). On obtient ainsi un graphe convexe vers le haut et la gauche du graphe. Le seuil optimum (indépendamment de la stratégie choisie) est celui qui correspond au point le plus près du coin haut et gauche du graphe. Une courbe qui se rapproche de la diagonale illustre une technique très peu discriminante, donc de qualité médiocre (exemple 21.5).

Les courbes ROC prennent tout leur intérêt lorsqu'il faut comparer plusieurs techniques entre elles. Elles permettent de résumer clairement plusieurs tableaux de sensibilité-spécificité en un seul graphe. La technique la plus discriminante est celle dont le graphe est le plus convexe vers le coin haut et gauche du graphe. Des techniques statistiques permettent de comparer les courbes entre elles en comparant les surfaces sous les courbes (exemple 21.6).

Exemple 21.5. COURBE ROC

On dépiste les anticorps du paludisme par la technique d'immunofluorescence indirecte (IFI). Les résultats sont exprimés en dilutions. Une dilution faible (1/10) indique un faible taux d'anticorps. Une dilution élevée indique un taux élevé d'anticorps. Pour évaluer les performances du test et fixer un seuil, on a étudié la technique sur un groupe de 100 patients atteints de paludisme à *Plasmodium falciparum* confirmé par d'autres techniques et 100 sujets indemnes de paludisme n'ayant jamais quitté une zone tempérée. Pour chaque dilution, considérée comme un seuil, on calcule la sensibilité et la spécificité. Ainsi par exemple sur le tableau de données suivant, à la dilution de 1/80, 80 malades sont positifs (20 qui se négativent au-dessus et 60 qui sont encore positifs aux dilutions suivantes). La sensibilité pour ce seuil est donc de 80/100 = 80 %. Toujours à la dilution de 1/80, 10 sujets sains sont positifs (5 qui se négativent au-dessus, et 5 encore positifs aux dilutions suivantes). La spécificité pour ce seuil est donc de (100 – 10)/100 = 90 %.

On reporte sur le graphe, pour chaque seuil, sa sensibilité en ordonnée, et le pourcentage de FP (1-Sp) en abscisses. On obtient donc le graphe de la figure 21.6.

Exemple 21.5. COURBE ROC (*Suite*)

Performances du test d'IFI dans le dépistage sérologique du paludisme

Dilutions IFI	Cas de paludisme à *P. falciparum* n	Se (%)	Sujets sains n	Sp (%)
négatifs	0	-	40	-
1/10	0	*100*	10	*40*
1/20	10	*100*	30	*50*
1/40	10	*90*	10	*80*
1/80	20	*80*	5	*90*
1/160	20	*60*	4	*95*
1/320	30	*40*	1	*99*
1/640	10	*10*	0	*100*

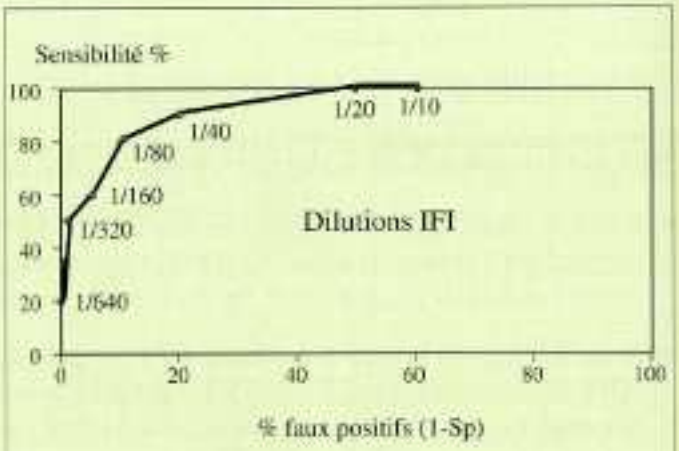


Figure 21-6. Performance de la technique d'IFI dans le dépistage sérologique du paludisme

La dilution représentant le meilleur compromis est la dilution au 1/80, plus proche du coin haut et gauche. Le choix du seuil dépend en fait de l'objectif assigné au test. Dans le cadre d'enquête épidémiologique, on conservera ce seuil optimum. En revanche, si le test doit être utilisé dans la prévention du paludisme transfusionnel chez des donneurs suspects de paludisme, on choisira le seuil donnant le moins de faux négatifs, 1/20. Mais dans ce cas, il faudra éliminer 50 % des donneurs sains.

Exemples 21.6.

On veut comparer 2 techniques sérologiques ELISA et CATT dans le diagnostic de la trypanosomiase africaine (maladie du sommeil ou THA). Les courbes obtenues figurent sur le graphe 21.7. On constate que l'ELISA (en noir) donne les meilleurs résultats pour tous les seuils de dilution étudiés. En revanche, le test statistique de comparaison des surfaces sous les courbes montre une différence non significative. Le test CATT, plus facile à utiliser sur le terrain peut donc être recommandé.

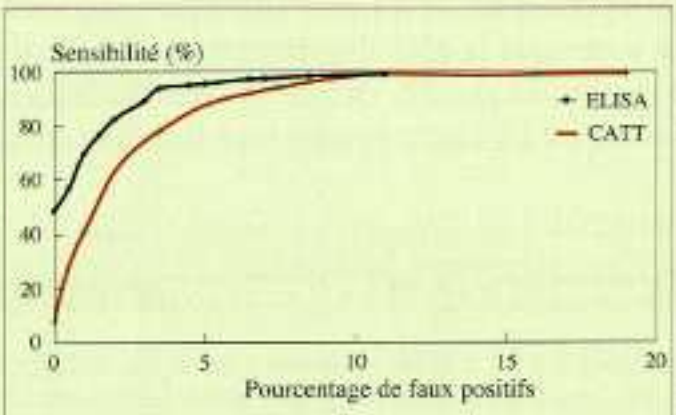


Figure 21-7. Comparaison de l'ELISA et du CATT dans le dépistage de la THA. Courbe ROC

II. PERFORMANCES D'UN TEST EN SITUATION RÉELLE

Un test dont on connaît la sensibilité et la spécificité est conçu pour être appliqué sur l'ensemble d'une population.

- Cette population peut être la population générale ou une population ciblée pour un dépistage systématique ; dans cette situation la prévalence de la maladie est en général rare.
- La population peut aussi être une population déjà sélectionnée : consultants chez un médecin ou un spécialiste, prélèvements pour un laboratoire, examens paracliniques prescrits à partir d'une indication clinique. On se trouve dans une situation diagnostique. La prévalence de la maladie recherchée est alors plus élevée que dans la population générale.

Le résultat d'un test partage la population d'étude en deux groupes : groupe avec résultats positifs et groupe avec résultats négatifs.

La question fondamentale qui se pose est de savoir quelle confiance accorder au résultat du test. En d'autres termes, on désire connaître quelle est la probabilité d'être malade chez un sujet présentant un test positif et quelle est la probabilité de ne pas être malade si le résultat est négatif.

Ces probabilités sont appelées **valeurs prédictives** d'un test.

1. Valeur prédictive positive

Lorsqu'un test est positif il existe deux possibilités : soit le sujet est malade, soit le sujet n'est pas malade malgré ce résultat contradictoire.

> On appelle valeur prédictive positive d'un test (VPP),
> la probabilité d'être malade lorsque le résultat est positif.

On démontre par le théorème de Bayes (*cf.* Annexes, Rappel et Formulaire 19) que :

- si Pr est la prévalence de la maladie ;
- Se la sensibilité ;
- Sp la spécificité :

$$VPP = \frac{SePr}{SePr + (1 - Sp)(1 - Pr)}$$

2. Valeur prédictive négative

Lorsqu'un test est négatif il existe deux possibilités : soit le sujet est sain, soit le sujet est malade malgré ce résultat contradictoire.

> On appelle valeur prédictive négative d'un test (VPN),
> la probabilité d'être sain lorsque le résultat est négatif.

On démontre par le théorème de Bayes (*cf.* Annexes Rappel et Formulaire 19) que :

- si Pr est la prévalence de la maladie ;
- Se la sensibilité ;
- Sp la spécificité :

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

Remarque : il apparaît dans de nombreux ouvrages ou articles des formules plus simples des valeurs prédictives : VPP = VP/(VP + FP) et VPN = VN/(FN + VN). Ces modes de calcul ne tiennent compte que des effectifs utilisés dans un échantillon d'étude pour lequel on connaît le nombre de vrais et faux positifs et négatifs. Elles ne sont valables que si l'échantillon est représentatif d'une population donnée et si la prévalence de la maladie étudiée est donnée.

3. Interprétation des VPP et VPN

a) Variation en fonction de la prévalence

Les formules montrent que les valeurs prédictives d'un test, pour une sensibilité et une spécificité donnée varient en fonction de la prévalence. Une VPP et une VPN non accompagnées de la prévalence estimée de la maladie n'ont aucun sens.

> Les valeurs prédictives d'un test dépendent de la prévalence de la maladie.

Ainsi un même test, n'aura pas les mêmes valeurs prédictives s'il est appliqué dans une population à forte ou à faible prévalence.

Le graphique 21.8 exprime les valeurs prédictives en ordonnées en fonction de la prévalence en abscisses.
On constate que pour de faibles prévalences, la VPP (courbe bleue) est très faible et varie rapidement. À l'inverse, la VPN (courbe rouge) est élevée et varie peu. Pour de fortes prévalences, la VPP est élevée et varie peu tandis que la VPN est faible et varie rapidement.

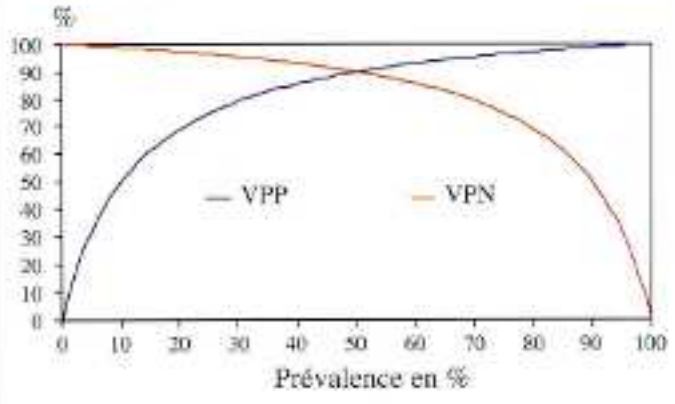


Figure 21-8. Valeurs prédictives en fonction de la prévalence de la maladie. Se = 90 %, Sp = 90 %

Ainsi, un test appliqué en situation de dépistage en population générale (Pr basse) aura une faible VPP et une forte VPN. De nombreux sujets seront alertés à tort, mais un résultat négatif sera rassurant. À l'inverse, le même test appliqué en situation de diagnostic, dans un service spécialisé (Pr élevé), aura une VPP élevée et une VPN moindre : un résultat positif sera hautement en faveur de la maladie tandis qu'un résultat négatif aura une signification moindre.
Le diagramme 21.9 illustre comment une variation de la prévalence peut affecter la VPP d'un test, alors que la sensibilité et la spécificité ne varient pas.

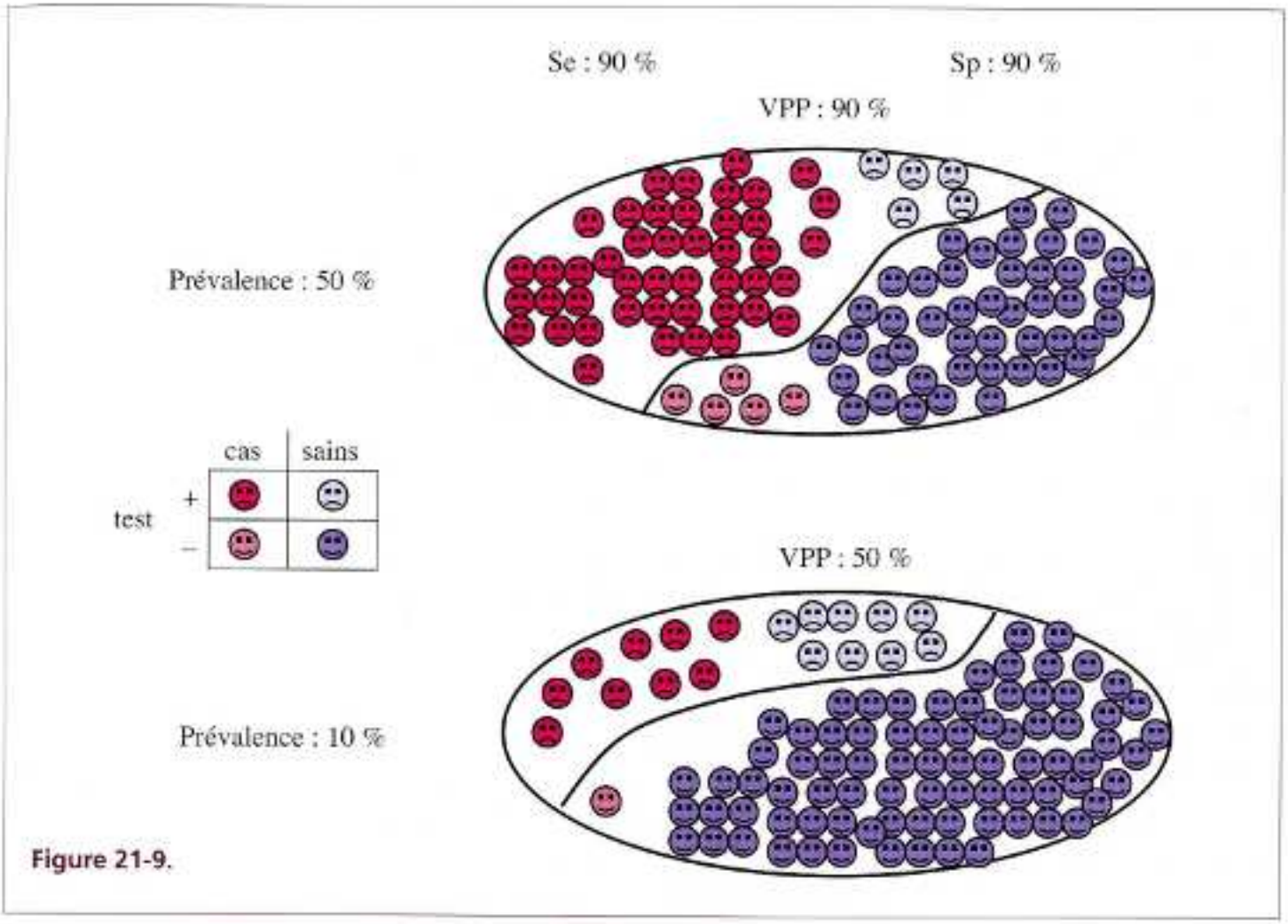


Figure 21-9.

b) Variation fonction de la sensibilité et de la spécificité

Sur les deux graphes de la figure 21.10, les courbes en couleur représentent des tests avec des performances Se et Sp différentes.

- VPP. D'après sa formule, on constate que la VPP dépend essentiellement du terme 1-Sp au dénominateur. Donc, plus Sp est élevée et plus la VPP est élevée.

> La valeur prédictive positive d'un test dépend de la spécificité.

Si l'on désire avoir une certitude diagnostique élevée, par exemple avant de prendre une décision chirurgicale lourde, il faudra donc exiger une haute spécificité.

- VPN. D'après sa formule, on constate que la VPN dépend essentiellement du terme 1-Se au dénominateur. Donc, plus Se est élevée, plus la VPN est élevée.

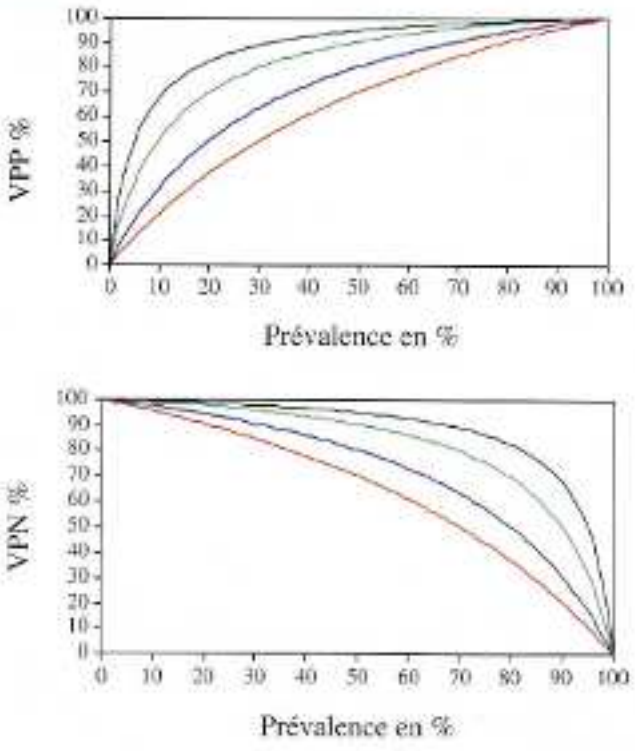


Figure 21-10.

La valeur prédictive négative d'un test dépend de la sensibilité.

Si l'on désire pratiquer un dépistage laissant échapper le moins possible de sujets malades, il faudra donc exiger une haute sensibilité (exemple 21.7).

Exemple 21.7. VALEURS PRÉDICTIVES D'UN TEST

Un radiologue spécialiste en mammographie dépiste dans sa clientèle de patientes adressées par des gynécologues, une proportion de 20 % de femmes atteintes de cancer du sein. L'examen mammographique qu'il utilise a une sensibilité de 98 % et une spécificité de 90 %. Quels sont les valeurs prédictives de ce test ?

$VPP = (0{,}98 \times 0{,}20)/[(0{,}98 \times 0{,}20) + ((1 - 0{,}90) \times (1 - 0{,}20))] = 0{,}71$ VPP = 71,0 %
$VPN = (0{,}90 \times (1\text{-}0{,}20)) / [(0{,}90 \times (1 - 0{,}20)) + (1 - 0{,}98) \times 0{,}20)] = 0{,}944$ VPN = 99,4 %

Ainsi lorsque l'examen est positif, il y a 71 % de probabilité que la patiente soit atteinte d'un cancer du sein. Une proportion d'environ 1/3 des patientes subiront des examens complémentaires plus invasifs, alors qu'elles ne sont pas malades. Lorsque l'examen est négatif, la quasi-totalité des consultantes sont rassurées. Moins d' 1 % d'entre elles sont porteuses d'un cancer du sein non dépisté.

Ce radiologue décide de changer de fonction, et de consacrer son installation exclusivement au programme de dépistage systématique du cancer du sein. La prévalence de cette affection est de 1 % dans la population cible de ce programme. Sachant que l'examen mammographique est strictement identique à la situation précédente, les valeurs prédictives du test ont-elles changé ?
La sensibilité du test est toujours de 98 % et la spécificité de 90 %. On a :

$VPP = (0{,}98 \times 0{,}01)/[(0{,}98 \times 0{,}01) + (1 - 0{,}90) \times (1 - 0{,}01)] = 0{,}09$ VPP = 9,0 %
$VPN = (0{,}90 \times (1 - 0{,}01)) / [(0{,}90 \times (1 - 0{,}01)) + (1 - 0{,}98) \times 0{,}01)] = 0{,}999$ VPN = 99,9 %

La VPP du test est maintenant effondrée. Un examen positif a une très faible probabilité de détecter un cancer du sein. À l'inverse, un résultat négatif sera tout à fait rassurant. Cette baisse de la VPP est due à la faible prévalence de la maladie dans la population cible.

Afin d'améliorer la VPP de son test, le radiologue décide d'acheter un nouvel appareillage, fournissant une spécificité de 95 % avec la même sensibilité. Quelle est l'amélioration de la VPP ?

$VPP = (0{,}98 \times 0{,}01)/[(0{,}98 \times 0{,}01) + ((1 - 0{,}95) \times (1 - 0{,}01))] = 0{,}165$ VPP = 16,5 %

Comme on le voit le gain reste médiocre. Lorsque la prévalence de la maladie est très basse, il faut que le test ait une spécificité proche de 100 % pour obtenir une VPP correcte.

III. REPRODUCTIBILITÉ ET CONCORDANCE

Un test biologique doit être reproductible d'une séance de travail à l'autre, ou concordant d'un expérimentateur à l'autre. Il existe une méthode permettant de comparer la reproductibilité entre deux séances de travail ou la concordance entre deux expérimentateurs (exemple 21.8).

Exemple 21.8.

Imaginons un test qualitatif dont le résultat peut s'exprimer par une variable à 3 classes : négatif, douteux, positif. La même série d'examen a été soumis à deux expérimentateurs A et B. On peut donc observer les résultats suivants :

A / B	négatif	douteux	positif
négatif	− −	− ±	− +
douteux	± −	± ±	± +
positif	+ −	+ ±	+ +

Les cases bleues représentent le nombre de résultats concordants.

De façon plus générale, on obtient le tableau de résultats suivants portant sur N tests réalisés par deux expérimentateurs ou lors de 2 séances de travail A et B :

A / B	A_1	A_2	...	A_i	Total
B_1	o_{11}	o_{12}	...	o_{1i}	t_1
B_2	o_{21}	o_{22}	...	o_{2i}	t_2
...	...	...	...	...	
B_i	o_{i1}	o_{i2}	...	o_{ii}	t_i
Total	n_1	n_2		n_i	N

Les résultats de A et B sont exprimés selon les modalités 1 à i. Le nombre de test selon le résultat de A et de B est exprimé dans chaque case par le nombre o_{ii}.

1. Coefficient de concordance

Le coefficient de concordance Cc est égal à la somme des résultats concordants sur le nombre total d'examens. Elle s'exprime par un nombre entre 0 et 1 (pourcentage) :

$$Cc = \frac{\text{nombre d'examens concordants}}{\text{nombre d'examens comparés}} = \frac{o_{11} + o_{22} + \ldots + o_{ii}}{N}$$

Le coefficient de concordance, aisé à interpréter, a cependant l'inconvénient, de comporter une part uniquement due au hasard. En effet, en imaginant que deux examinateurs lancent des pièces de monnaie pour décider si chaque candidat est reçu ou collé, il est fort probable qu'on obtiendrait des paires concordantes pile-pile ou face-face et on conclurait même à une concordance de 50 %.
Pour pallier cet inconvénient, on utilise un autre coefficient.

2. Coefficient kappa

On calcule d'abord la concordance attendue **Ca** de la façon suivante :

$$Ca = \frac{t_1 n_1 + t_2 n_2 + \ldots + t_i n_i}{N^2}$$

On appelle coefficient kappa le terme $\kappa = \frac{Cc - Ca}{1 - Ca}$

Le coefficient kappa s'exprime par un nombre compris entre – 1 et + 1 (exemple 21.9).

- Un κ proche de – 1 signifie une discordance complète.
- Un κ proche de 0 signifie une concordance moyenne due au hasard.
- κ proche de + 1 signifie une concordance absolue.

En médecine et en biologie, un coefficient kappa correct entre deux observateurs ou deux techniques doit être supérieur à 0,8 (80 %).

Mieux que le coefficient de concordance, le coefficient kappa exprime la concordance réelle, en éliminant la part due au hasard.

Sa valeur peut être testée (avec H_0 : kappa = 0, H_1 bilatéral : kappa ≠ 0) au moyen d'un simple test $Z = (\kappa - 0)/ s_\kappa$ où s_κ est l'écart type de kappa. Ces calculs sont effectués sur EpiInfo6 (Epitable/Compare/Proportion/Concordance).

L'écart type s_κ permet de calculer l'intervalle de confiance du coefficient kappa : $IC95\ \% = \kappa \pm 1{,}96\ s_\kappa$.

Exemple 21.9.

Les copies d'examen de statistique de 100 étudiants ont été corrigées en double par 2 correcteurs indépendants 1 et 2. Les résultats exprimés en trois classes : reçus, admis au rattrapage et collés sont exprimés sur le tableau ci-dessous :

		Correcteur 2			
		Reçus	Admis	Collés	Total
Correcteur 1	Reçus	14	7	1	22
	Admis	1	23	4	28
	Collés	10	15	25	50
	Total	25	45	30	100

$Cc = \frac{14 + 23 + 25}{100} = 0{,}62$ soit une concordance de 62 %

$Ca = \frac{22 \times 25 + 28 \times 45 + 50 \times 30}{100^2} = 0{,}33$

$\kappa = \frac{0{,}62 - 0{,}33}{1 - 0.33} = 0{,}43$ soit un coefficient kappa de 43 %.

$s_\kappa = 0{,}067$ (EpiInfo6) $Z = 0{,}43/0{,}067 = 6{,}4$ $\quad p < 10^{-5}$

Ces résultats signifient qu'il existe une concordance réelle de 43 % non attribuable au hasard. Bien que le coefficient soit significativement différent de zéro, est-il pour autant satisfaisant ? Ceci est une tout autre question. On observe en effet qu'il existe une forte différence entre correcteurs. Le correcteur 1 a tendance à coller les candidats plus souvent que le correcteur 2 (50 % *versus* 30 %) et le correcteur 2 a tendance à les admettre plus souvent au rattrapage (45 % *versus* 28 %).

3. Coefficient kappa pondéré

La formule de calcul du coefficient kappa vue dans le paragraphe précédent considère la concordance entre deux observateurs selon une alternative purement dichotomique : accord parfait ou désaccord total. Dans de nombreux cas, certaines modalités de jugement peuvent être considérées comme non identiques, mais pas franchement en désaccord absolu. Il existe une formule de coefficient kappa qui « pondère » les effectifs de chaque cellule en fonction de la gravité du désaccord. La formule du coefficient kappa pondéré est complexe (*cf.* Annexes, Formulaire statistique 26). On préférera les logiciels statistiques. La plupart de ces logiciels proposent plusieurs manières de pondérer les désaccords : soit de façon mathématique selon des formules qui tiennent compte de l'éloignement des désaccords, soit de façon raisonnée en laissant l'utilisateur fixer lui-même pour chaque case du tableau le poids qu'il juge pertinent de lui attribuer : poids = 1 pour un accord parfait, poids = 0 pour un désaccord total, et poids compris entre 0 et 1 pour les désaccords intermédiaires. Nous conseillons vivement cette méthode qui permet de comprendre exactement l'influence de son choix sur le résultat final.

Exemple 21.10. COEFFICIENT KAPPA PONDÉRÉ

Dans l'exemple 21.9, on pourrait considérer que le fait d'être reçu par un correcteur et admis au rattrapage par le second est un désaccord beaucoup moins grave que le fait d'être reçu par l'un et collé par l'autre.

On pourrait ainsi attribuer le poids 1 à chaque cellule pour lequel l'accord est total (reçu-reçu, admis-admis, collés-collés), le poids 0,8 à chaque cellule où le désaccord est modéré (reçu-admis), le poids 0,1 en cas de désaccord grave (admis-collé) et le poids 0 en cas de désaccord total (reçu-collé).

Matrice des pondérations attribuées en fonction de la gravité des discordances entre les deux correcteurs :

		Correcteur 2		
		Reçus	Admis	Collés
Correcteur 1	Reçus	1	0,8	0
	Admis	0,8	1	0,1
	Collés	0	0,1	1

En reprenant les données de l'exemple 21.9, et en appliquant les formules données en Annexe, Formulaire § 26, on a :

$$Cc_w = \frac{14 + 23 + 25 + 0{,}8\,(7 + 1) + 0{,}1\,(4 + 15)}{100} = 0{,}703 \text{ soit une concordance de } 70{,}3\ \%$$

$$Ca_w = \frac{22 \times 25 + 28 \times 45 + 50 \times 30 + 0{,}8\,(22 \times 45 + 28 \times 45) + 0{,}1\,(28 \times 30 + 50 \times 45)}{100^2} = 0{,}497$$

$$\kappa = \frac{0{,}703 - 0{,}497}{1 - 0.497} = 0{,}410 \text{ soit un coefficient kappa de } 41{,}0\ \%$$

En ayant ainsi pondéré les discordances, on aboutit à un coefficient inférieur au coefficient brut initial (43 %).

Exercices

Exercice 21.1

On veut tester une nouvelle technique de **diagnostic** du paludisme. Les cas étudiés pour la sensibilité sont des sujets présentant un paludisme clinique et biologique prouvé par une technique de référence. Quels témoins devra-t-on choisir pour tester la spécificité de cette nouvelle technique ?

Exercice 21.2

Dans un programme de dépistage-traitement de la maladie du sommeil en Afrique, on utilise un test rapide (CATT) applicable à toute la population. Ce test a une sensibilité de 95 % et une spécificité de 75 %. Au début du programme, la prévalence de la maladie était de 20 %. Au bout de 8 ans, grâce au programme la prévalence a chuté à 0,5 %.

Calculez les VPP et VPN du test appliqué au début et en fin de programme.

Résumé

Les performances d'un test diagnostique (clinique ou biologique) peuvent être appréciées soit de façon expérimentale, soit en situation de terrain.
En situation expérimentale, on compare les résultats du test entre un groupe de sujets malades et un groupe de sujets sains. Les performances « intrinsèques » du test se mesurent par la sensibilité et la spécificité.
En situation de terrain, on cherche à connaître la probabilité d'être ou de pas être malade en fonction du résultat positif ou négatif d'un test dont on connaît les performances intrinsèques. Ces probabilités se mesurent par les valeurs prédictives et négatives du test. Ces valeurs prédictives sont liées à la prévalence de la maladie.

PERFORMANCES D'UN TEST

Sensibilité	Se	$\frac{VP}{VP + FN}$
Spécificité	Sp	$\frac{VN}{FP + VN}$
Valeur prédictive positive	VPP	$\frac{SePr}{SePr + (1 - Sp)(1 - Pr)}$
Valeur prédictive négative	VPN	$\frac{Sp(1 - Pr)}{Sp(1 - Pr) + (1 - Se)(Pr)}$

VP, FP : vrais et faux positifs VN, FN : vrais et faux négatifs.
Pr : prévalence de la maladie.

ANNEXES

ANNEXES

Réponses aux questions des exercices

Exercice 1.1

- B : qualitative nominale
- C : qualitative binaire
- D : de type date
- E : quantitative continue
- F : qualitative nominale
- G : qualitative nominale
- H : qualitative ordinale

Exercice 2.1

Dilution : x	n	% = $x_i/121$	$x' = \log_2(1/x)$	% fréquence équivalente	% amplitude x = 2 dilutions	% amplitude de convenance logique
1/2	4	3,3	1			
1/4	5	4,1	2	32,2	7,4	14,0 % non significatifs
1/8	8	6,6	3			
1/16	22	18,2	4		24,8	
1/32	25	20,7	5			
1/64	16	13,2	6	33,9	33,9	74,4 % suspects
1/128	11	9,1	7			
1/256	7	5,8	8		14,9	
1/512	9	7,4	9	33,9		
1/1 024	6	5,0	10		12,4	11,6 %
1/2 048	5	4,1	11			paludismes évolutifs
1/4 096	3	2,5	12		6,6	
Total	121	100,0		100,0	100,0	100,0 %

Exercice 4.1

On classe les nouveau-nés par ordre croissant de poids :

poids (g)	2 985	3 043	3 122	3 250	3 359	3 482	3 498	3 507	3 634	3 743	3 854

- Médiane = 3 482 g : cette valeur partage la série en deux groupes de taille identique.
- Moyenne m :

$$m = \frac{3\,250 + 3\,482 + 3\,122 + 3\,498 + 3\,743 + 3\,854 + 3\,359 + 2\,985 + 3\,043 + 3\,634 + 3\,507}{11} = \frac{37\,477}{11} = 3\,407$$

Moyenne = 3 407 g.

Exercice 4.2

médiane = (3 359 + 3 482)/2 = 3 420 g : 50 % des valeurs sont situées de part et d'autre de cette valeur.

Exercice 4.3

On trie d'abord les valeurs par ordre croissant :

0,6	2,2	2,5	3,6	4	4,2	4,5	4,9	5,2	5,8	6,6	6,8
7	7,6	7,9	8	8,2	8,3	8,7	8,9	8,9	9,3	9,4	9,5
9,5	9,5	9,6	9,7	9,7	10	10,2	10,6	10,6	10,8	11,1	11,5
11,9	13,1	13,7	14	14,4	14,6	15,3	15,7	16,1	16,3	16,7	17,6
17,9	18,6	18,8	19,7	20,5							

• médiane	9,6	valeur centrale
• 1er quartile	7,6	
• 3e quartile	14,0	
• mode	9 à 9,9	données regroupées par classes d'amplitude 1,0
• intervalle interquartile	6,4	14-7,6
• minimum	0,6	
• maximum	20,5	
• étendue	19,9	20,5-0,6
• moyenne	10,38	550,3/53
• variance	23,25	$(6922{,}71 - (550{,}3)^2/53)/53$
• écart type	4,82	racine (23,25)
• coefficient de variation	46,4 %	4,82/10,38 × 100

Exercice 6.1

$x = 4 \quad n = 10 \quad p(x) = 0{,}15$

On utilise la loi binomiale.

La probabilité d'observer au moins 4 sujets présentant un effet indésirable est égale à la somme des probabilités d'en observer 4, 5, 6, 7, 8, 9 et 10.

$P(X \geq 4) = P(X = 4) + P(X = 5) + \ldots\ldots P(X = 10).$

Il est plus simple de calculer $1 - P(X < 4) = 1 - [P(0) + P(1) + P(2) + P(3)]$

$P(0) = 0{,}197$, $P(1) = 0{,}347$, $P(2) = 0{,}276$, $P(3) = 0{,}130$

$P(X \geq 4) = 1 - 0{,}950 = 0{,}05$

La probabilité d'observer au moins 4 sujets présentant un effet indésirable est de 5 %.

Exercice 6.2

On peut encore utiliser la loi binomiale

$x = 1 \quad n = 300 \quad p(x) = 1/11\,000 = 0{,}000090909$

$P(x = 1) = 0{,}026$: Il y a 2,6 % de chances de rencontrer un incident au cours d'un voyage de ce type.

Exercice 6.3

1) Dans un premier temps, il faut faire l'hypothèse que s'il n'y avait pas de transmission secondaire, les cas se répartiraient de façon aléatoire dans les chambres doubles. L'application de la loi binomiale permet de calculer les probabilités d'observer 0, 1 et 2 cas par chambre.
Dans ce type de problème, l'échantillon est une chambre et la taille de l'échantillon est de 2 (2 lits).
La caractéristique étudiée est d'observer éventuellement un cas de gale par lit.
La proportion de sujets atteints dans la population d'étude est $p = 25/80 \cong 0{,}31$
La probabilité d'observer 0 cas est $P(0) = (2!/0!\,(2-0)!\,) \times 0{,}31^{0} \times 0{,}69^{2-0} = 0{,}476$
La probabilité d'observer 1 cas est $P(1) = (2!/1!\,(2-1)!\,) \times 0{,}31^{1} \times 0{,}69^{2-1} = 0{,}428$
La probabilité d'observer 2 cas est $P(2) = (2!/2!\,(2-2)!\,) \times 0{,}31^{2} \times 0{,}69^{2-2} = 0{,}096$

2) Dans un deuxième temps, on calcule les fréquences réellement observées de chambres doubles comportant 0,1 et 2 cas.

- Fréquence des chambres avec 0 cas : $f(0) = 24/40 = 0{,}6$
- Fréquence des chambres avec 1 cas : $f(1) = 7/40 = 0{,}175$
- Fréquence des chambres avec 2 cas : $f(2) = 9/40 = 0{,}225$

3) Dans un troisième temps, il ne reste plus qu'à comparer les probabilités attendues et les fréquences réellement observées.
On constate que la fréquence des chambres comportant 2 cas est plus de 2 fois supérieure à la probabilité attendue (22,5 % *versus* 9,6 %). À l'opposé, la fréquence des chambres ne comportant qu'un seul cas est plus de deux fois inférieure à la probabilité attendue.
Conclusion : on peut supputer que le fait de côtoyer un cas de gale dans une chambre à deux lits est un facteur de risque supplémentaire se rajoutant à la problématique générale de cet établissement.

Exercice 6.4

On utilise la loi de Poisson de moyenne 1.
1) $P(0) = 36{,}8\ \%$, $P(1) = 36{,}8\ \%$, $P(2) = 8{,}4\ \%$, $P(3) = 6{,}1\ \%$, $P(4) = 1{,}5\ \%$
2) $1 - P(0) = 1 - 0{,}368 = 0{,}632$ soit 63,2 %
3) $P(0) + P(1) = 0{,}368 + 0{,}368 = 0{,}736$ soit 73,6 %

Exercice 6.5

On utilise la loi de Poisson de moyenne 1,4

Les probabilités d'observer 0, 1, 2, 3 cas sont de :
$P(0) = 0{,}247$, $P(1) = 0{,}345$, $P(2) = 0{,}242$, $P(3) = 0{,}113$
La probabilité d'observer au plus 3 cas est de : $P(0) + P(1) + P(2) + P(3) = 0{,}947$
La probabilité d'observer au moins 4 cas est donc $1 - 0{,}947 = 0{,}053 = 5{,}3\ \%$

Il y donc plus de 5 chances sur 100 d'observer au moins 4 cas au cours d'une année. On ne peut donc pas conclure à un risque majoré.

Exercice 6.6

1) 50 %.
2) 34 % : 68 % des valeurs sont comprises entre -1σ et $+1\sigma$ de part et d'autre de la moyenne. Il y a donc la moitié de 68 %, soit 34 % de valeurs comprises entre -1σ et m.
3) 81,5 % : la moitié de 95 % [-2σ et $+2\sigma$] plus 34 %.

Exercice 7.1

Les réponses sont données ici à titre indicatif. D'autres méthodes que celles proposées peuvent être discutées.

1) La population en France des personnes âgées vivant en communauté est très grande (plusieurs centaines de milliers de personnes). Il n'existe pas de base de sondage. En revanche, il existe une liste de tous les établissements d'hébergement.
Il faut donc pratiquer un sondage à deux degrés en tirant d'abord un échantillon des établissements (sondage élémentaire sur la liste des établissements), puis en tirant au sort les individus parmi les établissements sélectionnés.
D'autre part, les conditions de vie, et donc les facteurs de risque de gale, doivent être différents selon le type d'établissement, maison de retraite non médicalisée et centres de long et moyen séjour médicalisés.
Il serait donc judicieux d'équilibrer ces trois groupes et de pratiquer un sondage stratifié.
Au total, on aura réalisé un sondage stratifié en trois strates et à deux degrés.

2) La population d'étude est relativement faible (2000). Si les moyens de l'étude le permettent, on peut envisager une étude exhaustive. Sinon, on dispose de la liste des élèves inscrits dans l'établissement. On peut aisément réaliser un sondage élémentaire en numérotant chaque individu.

3) Grâce au recensement, on dispose en France de la liste de tous les logements d'un département et d'une ville. En raison du nombre élevé de logements dans la ville choisie, le moyen le plus simple est de pratiquer un sondage systématique après avoir déterminé un pas de sondage dépendant de la taille de l'échantillon choisie.

4) Il n'existe pas de base de sondage des enfants de moins de 5 ans dans un département. D'autre part, tous les enfants ne fréquentent pas obligatoirement une crèche ou une école. Il faut donc les rechercher dans leur famille.
D'autre part, comme dans l'exemple précédent, la liste de tous les logements d'un département est sûrement très élevée. On peut donc dans une première étape échantillonner des sous-unités administratives du département : commune ou cantons. On disposera ensuite de la liste des logements des communes ou cantons sélectionnés. On tirera ensuite au sort les logements. On interrogera tous les chefs de famille des logements sélectionnés. Dans une telle enquête, il faut évidemment prévoir une taille d'échantillon suffisamment grande, car une grande proportion de familles interrogées ne comprendra pas d'enfants de moins de 5 ans.

5) Pour un enjeu commercial de ce type, on peut se contenter d'un sondage empirique par méthode des quotas.

6) Il existe une liste des travailleurs inscrits à la caisse. Puisque la maladie est rare et à prédominance rurale, il est judicieux de focaliser l'échantillon sur les travailleurs ruraux. On pratiquera donc un sondage à deux strates, ruraux et urbains, et un sondage élémentaire ou systématique à l'intérieur de chaque strate.

7) Il n'existe pas de base de sondage des logements, et encore moins des individus, au niveau de la région dans les pays en voie de développement. Il faut donc dresser une liste des villages de la région. On pratiquera un sondage systématique sur la liste des villages. À l'intérieur de chaque

village sélectionné, on détermine aléatoirement par un procédé mécanique (lancement d'une baguette), une direction à partir du centre du village. On note tous les logements situés sur la ligne ainsi définie. Ensuite on tire au sort un logement parmi cette liste. On examine ensuite tous les enfants de chaque logement sélectionné qui constituent une grappe. Afin de tenir compte de la taille respective de chaque village, on peut tirer un nombre de grappes proportionnel à la taille du village.
Au total, on aura ainsi réalisé un sondage à plusieurs degrés. Premier degré à sondage systématique (les villages), deuxième degré à sondage aléatoire (direction), troisième degré à sondage en grappes (ensemble des sujets d'un logement tiré de façon élémentaire).
Ce type de sondage, simple à réaliser sur le terrain, est couramment pratiqué lors des enquêtes en PVD. L'interprétation des résultats doit être soigneusement évaluée, en raison de biais fréquent dus à l'hétérogénéité des grappes.

Exercice 9.1

L'écart type de la moyenne est $s_m = s/\sqrt{n} = 0,4/5 = 0,08$

1) La taille de l'échantillon étant inférieure à 30, il faut chercher la valeur $t_{5\,\%}$ dans la table de T en supposant que la distribution de la glycémie est « normale » dans la population d'étude. À la ligne ddl = 25 – 1 = 24, on trouve la valeur 2,064.
L'intervalle de confiance à 95 % de part et d'autre la moyenne est donc 0,08 × 2,064 = 0,165
soit m = 1,52 ± 0,165 g/L ou **m** compris dans l'intervalle[1,355-1,685]

2) La valeur $t_{1\,\%}$ pour ddl = 24 est de 2,8
L'intervalle de confiance à 99 % de la moyenne est donc 0,08 × 2,8 = 0,224
m = 1,52 ± 0,224 g/L ou **m** compris dans l'intervalle [1,296-1,744]
L'intervalle de confiance de la moyenne est plus large, puisqu'on désire un risque d'erreur plus faible.

Exercice 9.2

p = 30 %

$s_p = \sqrt{0,3(1-0,3)/3\,500} = 0,0077$

L'intervalle de confiance à 95 % est donc de 1,96 × 0,0077 = 0,015
p = 30 % ± 1,5 % ou p compris dans l'intervalle [28,5 %-31,5 %]

Exercice 9.3

Le pourcentage de complications d'infection fungique est de 8/12 soit 66,7 %
Il n'est pas nécessaire de calculer un intervalle de confiance sur ce pourcentage, si on ne considère que la population étudiée des 12 sujets greffés. On a bien observé 66,7 % de complications.
Mais si l'on considère que les 12 greffés sont représentatifs d'une population plus large qui seraient les greffés de moelle dans ce service pendant une période plus longue, il faudrait évidemment calculer un intervalle de confiance.
Mais on constate que n (1 – p) = (1 – 0,667) × 12 = 4. Ce chiffre est inférieur à 5. Pour calculer l'IC à 95 % on ne peut donc pas utiliser la formule contenant l'écart type d'un pourcentage. Il faut utiliser la loi binomiale ou regarder dans une table. On trouverait un intervalle compris entre 34,9 % et 90,1 %

Exercice 9.4

$p = 10/70 = 0{,}143$

$$s_p = \sqrt{\frac{0{,}143(1-0{,}143)}{70}} \sqrt{\frac{200-70}{200-1}} = 0{,}042 \times 0{,}81 = 0{,}034 :$$

il faut utiliser le facteur d'exhaustivité car la taille de l'échantillon est élevée par rapport à la taille de la population étudiée (70/200 = 35 %)

IC 95 % = 1,96 × 0,034 = 0,067
On a donc p = 14,3 % ± 6,7 % ou p compris dans l'intervalle [7,6 %-21,0 %]

Exercice 9.5

Il faut :

- connaître approximativement la fréquence attendue des résistances ;
- proposer une précision souhaitée ;
- proposer un risque α.

Exercice 10.1

1) H_0 : les deux traitements sont équivalents.
   H_1 bilatérale : les deux traitements ont une efficacité différente.
2) H_0 : les quatre traitements sont équivalents.
   H_1 bilatérale : les quatre traitements ont une efficacité différente.
3) H_0 : le traitement A et le placebo sont équivalents.
   H_1 unilatérale : le traitement A une activité supérieure au placebo.
4) H_0 : il n'existe aucune liaison entre la hauteur des arbres et leur altitude.
   H_1 : il existe une liaison négative entre la hauteur des arbres et leur altitude.

Exercice 10.2

3) Degré de signification *p*.

Exercice 10.3

3) De 1 % : il faut avoir le moindre risque de conclure à tort à une efficacité qui n'existerait pas.

Exercice 10.4

2) Le risque β : il ne faudrait pas passer à coté d'une efficacité réelle de ce vaccin. On choisira donc un risque β plutôt faible. Un tel choix entraîne automatiquement une augmentation de la puissance $(1 - \beta)$ et l'augmentation de la taille des échantillons.

Exercice 11.1

1) Test T de comparaison de 2 moyennes (un des effectifs est inférieur à 30).
2) Test F ou test de Kruskal-Wallis (comparaison de plusieurs moyennes). Le test de KW est sûrement plus approprié car il est peu probable que les temps réalisés dans une course contre la montre se distribuent de façon normale.
3) Test de χ^2 d'homogénéité entre pourcentages.
4) Test de χ^2 de conformité entre une distribution observée et la distribution dans la population. Comme on ne dispose que de la structure en classes d'âge, on ne peut pas comparer la moyenne d'âge observée à la moyenne d'âge théorique.

5) Test exact de Fisher. Étant donné le faible nombre des effectifs, il est fort probable que certains effectifs théoriques soient inférieurs à 3.
6) Test de χ^2 d'homogénéité.

Exercice 11.2

Soit H_0 a été rejetée à tort dans la première étude (il y a 4 chances sur 100 que cela se produise), soit H_0 n'a pas été rejetée à tort dans la seconde étude (risque de deuxième espèce β).

Exercice 11.3

Tableau de contingence des effectifs observés

		Échantillons			Total	
		E_1	E_2	E_3	N	%
Évolution de la maladie	guérison	5	6	16	27	*33,3*
	rechute	9	9	10	28	*34,6*
	décès	15	4	7	26	*32,1*
	Total	29	19	33	81	*100,0*

Hypothèses : nous sommes dans la situation d'un test de χ^2 d'homogénéité.
- H_0 : les 3 groupes de malades sont identiques et appartiennent à la même population.
- H_1 : les 3 groupes sont différents, et n'appartiennent pas à la même population.

Calculons les effectifs qui auraient été observés si les distributions de chaque échantillon étaient identiques à la distribution marginale totale.
Dans la première colonne, parmi les 29 sujets on aurait respectivement 33,3 % de guérison, 34,6 % de rechute et 32,1 % de décès. Les effectifs théoriques sont donc respectivement 9,7, 10,0 et 9,3.
On obtient un second tableau de contingence composé d'effectifs théoriques. On vérifie que les totaux sont toujours les mêmes (aux arrondis près) et que les 3 distributions sont maintenant identiques entre elles en fréquence relative et identiques à la distribution totale.

Tableau de contingence des effectifs théoriques

	E_1	E_2	E_3	N	%
guérison	9,7	6,3	11,0	27	*33,3*
rechute	10,0	6,6	11,4	28	*34,6*
décès	9,3	6,1	10,6	26	*32,1*
Total	29	19	33	81	*100,0*

On vérifiera que le raisonnement et les calculs auraient été strictement similaires si on avait pris les pourcentages des colonnes pour calculer les effectifs théoriques.
On vérifie que tous les effectifs théoriques sont supérieurs à 5. Le fait qu'un des effectifs *observés* soit inférieur à 5 n'est pas à prendre en compte.

$$\chi^2_o = \frac{(5-9{,}7)^2}{9{,}7} + \frac{(9-10{,}0)^2}{10{,}0} + \frac{(15-9{,}3)^2}{9{,}3} + \frac{(6-6{,}3)^2}{6{,}3} + \frac{(9-6{,}6)^2}{6{,}6} + \frac{(4-6{,}1)^2}{6{,}1} + \frac{(16-11{,}0)^2}{11{,}0} + \frac{(10-11{,}4)^2}{11{,}4} + \frac{(7-10{,}6)^2}{10{,}6}$$

$\chi^2_o = 2{,}28 + 0{,}10 + 3{,}49 + 0{,}01 + 0{,}87 + 0{,}72 + 2{,}27 + 0{,}17 + 1{,}22 = 11{,}13$

$\chi^2_o = 11{,}13$

ddl = (3 – 1) × (3 – 1) = 4

Pour ddl = 4 la table du χ^2 montre que la valeur seuil de $\chi^2_{5\%} = 9{,}49$.

La valeur trouvée est supérieure à la valeur seuil. Elle est encore supérieure à la valeur de $\chi^2_{3\%}$ qui est égale à 10,7. Le degré de signification est donc $p < 0{,}03$.

Conclusion

On rejette H_0. On conclut que ces trois groupes de malades proviennent de populations différentes ($p < 0{,}03$). Il faudrait examiner les proportions de chaque type d'évolution pour chaque échantillon. On constate que la guérison est moins fréquente dans l'échantillon E_1 et plus fréquente dans l'échantillon E_3. À l'inverse, le décès est plus fréquent dans l'échantillon E_1 et moins fréquent dans l'échantillon E_3.

Tableau des fréquences relatives des classes de la variable pour chaque échantillon

		Échantillons			Total	
		E_1 %	E_2 %	E_3 %	N	%
Évolution de la maladie	guérison	*17,2*	*31,6*	*48,5*	27	*33,3*
	rechute	*31,1*	*47,4*	*30,3*	28	*34,6*
	décès	*51,7*	*21,0*	*21,2*	26	*32,1*
	Total	*100,0*	*100,0*	*100,0*	100,0	*100,0*

Exercice 12.1

1) Régression : les deux variables sont quantitatives, mais ne jouent pas un rôle symétrique : la variable « hauteur des arbres » est présumée dépendante de la variable « altitude ».
2) Régression : le nombre de leucocytes est la variable dépendante Y, la dose du traitement est la variable explicative X.
3) Régression. Le poids est dépendant de la taille. Mais un test de corrélation peut être effectué de façon équivalente.
4) Test du χ^2 de tendance : la variable dépendante est de type binaire. Un test de régression serait équivalent en affectant la valeur 1 à la présence d'un effet secondaire, et la valeur 0 à son absence.

Exercice 15.1

	région 1	région 2	région 3	région 4
incidence du paludisme (‰)	36,48	109,40	18,88	16,73
mortalité brute (‰)	19,58	34,78	34,90	32,73
la mortalité spécifique (‰)	4,54	13,57	5,89	2,01
la mortalité proportionnelle (%)	23,20	39,03	16,89	6,15
la létalité du paludisme (%)	12,45	12,41	31,21	12,02

L'incidence est la plus élevée en région 2. Comme la létalité est moyenne, les mortalités spécifique et proportionnelle sont donc également plus élevées dans cette région.
La létalité est la plus élevée dans la région 3. Il y a sans doute un problème dans la prise en charge de la maladie dans cette région.
Les mortalités spécifique et proportionnelle sont plus faibles dans la région 4, alors que la létalité est moyenne. Il y a sans doute une autre maladie à mortalité élevée plus fréquente dans cette région.

Exercice 16.1

1) **c** : malgré la valeur élevée du RR ou OR, son intervalle de confiance englobe la valeur 1. On ne peut donc pas conclure.
2) **a** : la valeur du RR ou OR est faible certes, mais l'intervalle de confiance exclut la valeur 1. On peut donc conclure en affirmant que le facteur étudié multiplie le risque de contracter la maladie par un facteur 1,2.
3) **e** (ou b) : une valeur de 0,01 correspond à un facteur protecteur très marqué. Son opposé aurait un RR ou OR égal à 100!
4) **d** : la valeur ponctuelle du RR ou OR est inférieure à 1, mais l'intervalle de confiance englobe la valeur 1.
5) **b** : la valeur ponctuelle du RR ou OR est inférieure à 1 et la borne supérieure de l'intervalle de confiance est elle-même inférieure à 1.

Exercice 16.2

1) Cas-témoins : il faut travailler rapidement.
2) Cohorte : on dispose de nombreux cas. L'enquête cas témoins est peu adaptée pour étudier les expositions rares.
3) Cas témoins : on dispose de nombreux sujets dans le groupe exposé. L'enquête de cohorte est peu adaptée pour étudier les maladies rares.
4) Cas-témoins : on peut recueillir dans le même temps plusieurs variables d'exposition et calculer des odds ratio pour chacune d'elles. Pour monter une enquête de cohorte il faudrait suivre autant de cohortes que de facteurs à étudier.
5) Cohorte. On peut recueillir des informations sur plusieurs maladies dans les groupes exposés et non-exposés et calculer l'incidence de chacune des maladies.

Exercice 16.3

Les OR de chaque strate sont très différents entre eux. Ils diffèrent de plus de 20 %. Le tiers facteur sur lequel les données ont été stratifiées est à l'évidence un modificateur de l'effet (interaction). Dans la première strate (non consommateurs de vin blanc) la liaison entre consommation de crustacé et gastro-entérite est encore plus élevée que dans l'ensemble de l'étude. Dans la deuxième strate la liaison

persiste, mais est moins forte. Dans la troisième strate, la consommation élevée de vin blanc fait disparaître la liaison positive entre consommation de crustacé et survenue de gastro-entérite et au contraire produit un effet protecteur très marqué.
Il n'est pas pertinent de calculer un OR ajusté lorsqu'on identifie un modificateur d'effet. Seule l'analyse strate par strate est justifiée. Cette étude soulève l'hypothèse que la consommation d'alcool lors d'un repas contaminé par un germe analogue à celui de cette épidémie, peut prévenir la survenue de gastro-entérite.

Exercice 21.1

Témoins ne présentant aucun des signes cliniques définissant un cas de paludisme et ayant subi le test biologique de référence avec un résultat négatif. Contrairement à l'exemple 21.5 portant sur le dépistage des flacons de la banque du sang, il faudrait dans cette situation diagnostique choisir les témoins parmi des sujets ayant voyagé dans les mêmes pays que les cas afin que la population des témoins soit le plus proche possible de celle des cas.

Exercice 21.2

1) Début de programme Pr = 0,20

$$\text{VPP} = \frac{0{,}95 \times 0{,}20}{0{,}95 \times 0{,}20 + (1-0{,}75)(1-0{,}2)} = 0{,}487 = 48{,}7\ \%$$

$$\text{VPN} = \frac{0{,}75(1-0{,}20)}{0{,}75(1-0{,}20) + (1-0{,}95)(0{,}20)} = 0{,}984 = 98{,}4\ \%$$

2) En fin de programme Pr = 0,005

$$\text{VPP} = \frac{0{,}95 \times 0{,}005}{0{,}95 \times 0{,}005 + (1-0{,}75)(1-0{,}005)} = 0{,}0187 = 1{,}9\ \%$$

$$\text{VPN} = \frac{0{,}75(1-0{,}005)}{0{,}75(1-0{,}005) + (1-0{,}95)(0{,}005)} = 0{,}9997 = 99{,}97\ \%$$

La chute de la prévalence a entraîné une baisse considérable de la VPP. La certitude diagnostique en cas de résultat du CATT positif est quasiment nulle. Il faut changer la stratégie du programme.

Rappels mathématiques

Puissance

	Exemples
$a^n = a.a.a.....n$ fois	$2^{10} = 2.2.2.2.2.2.2.2.2.2 = 1\,024$
$a^1 = a$	$2^1 = 2$
$a^0 = 1$	$2^0 = 1$
$a^m . a^n = a^{m+n}$	$2^3 2^5 = 2^8 = 256$
$(a^m)^n = a^{mn}$	$(2^2)^3 = 2^6 = 64$
$a^m/a^n = a^{m-n}$	$2^6/2^2 = 2^4 = 16$
$a^{-n} = 1/a^n$	$10^{-3} = 1/10^3 = 0{,}001$

Racines

$b = a^n \Rightarrow a = \sqrt[n]{b}$

$\sqrt[n]{a} = a^{1/n}$

$\sqrt[n]{a^m} = a^{m/n}$

Logarithmes

logarithme de base a : $\log_a$	log népérien : ln	e = 2,7182...
$\log_a 1 = 0$	$\ln 1 = 0$	
$\log_a a = 1$	$\ln e = 1$	
$\log_a x = \ln x / \ln a$	$\ln 10 = 2{,}302\,59$	
$\log_{10} x = 0{,}43429 \ln x$	$\ln x = 2{,}302\,59 \log_{10} x$	
$\log_a xy = \log_a x + \log_a y$	$\ln xy = \ln x + \ln y$	
$\log_a 1/x = -\log_a x$	$\ln 1/x = -\ln x$	
$\log_a x/y = \log_a x - \log_a y$	$\ln x/y = \ln x - \ln y$	
$\log_a x^y = y \log_a x$	$\ln x^y = y \ln x$	
$\log_a \sqrt[y]{x} = \log_a x^{1/y} = 1/y \log_a x$	$\ln \sqrt[y]{x} = \ln x^{1/y} = 1/y \ln x$	
$y = \log_a x \Rightarrow x = a^y$	$y = \ln x \Rightarrow x = e^y$ ou $x = \exp(y)$	

Factorielles

Factorielle d'un nombre entier n : $n! = 1 \times 2 \times 3 \ldots\ldots (n-1) \times n$ $\quad 1! = 1$ $\quad 0! = 1$

Combinaisons

Combinaison de n éléments pris k à k : $C_n^k = \frac{n!}{k!(n-k)!}$

Exemple
Combien de combinaisons de rois peut-on constituer avec trois cartes ?
n = 4 couleurs k = 3 cartes
C = 4!/3!1 = 4 ♠♣♥ ♠♣♦ ♣♥♦ ♠♥♦

Probabilités

Probabilité d'un événement A : P(A) avec $0 \leq P(A) \leq 1$
$P(\overline{A}) = 1 - P(A)$ $\Rightarrow P(A) + P(\overline{A}) = 1$

Probabilité de deux événements :

- probabilité de A ou bien B : P (A ou B) = P (A ∪ B)
- probabilité de A et de B : P (A et B) = P (A ∩ B)

Probabilité d'un événement B si un événement A a eu lieu : P (B si A) = P (B|A)

1) A et B sont deux événements incompatibles (disjoints)
P(A ou B) = P(A) + P(B)
P(A et B) = 0
P(B|A) = 0

Exemple : on tire une carte dans un jeu de 52 cartes
Événements : A = roi de cœur, B = roi de trèfle :
P (roi de cœur ou roi de trèfle) = 1/52 + 1/52 = 0,019

2) A et B sont compatibles et indépendants
P(A ou B) = P(A) + P(B) – P(A et B)
P(A et B) = P(A) × P(B)
P(B|A) = P(B)

Exemple : on tire une carte sur 52 cartes
Événements : A = cœur, B = roi
P(cœur ou roi) = 1/4 + 4/52 – 1/52 = 0,31
P(cœur et roi) = 1/4 × 4/52 = 1/52
P(roi si cœur) = P(roi) = 4/52 = 1/13

3) A et B sont compatibles et dépendants : probabilités conditionnelles
P(A et B) = P(A|B) P(B)
= P(B|A) P(A)

$P(B|A) = \frac{P(A \text{ et } B)}{P(A)}$

Exemple : on tire 2 cartes dans un jeu de 52
Événements : A = un roi au 1er tirage,
B = un autre roi au 2e tirage
P(roi au 1er tirage) = 4/52
P(roi au 2e tirage si 1 roi est sorti au 1er) = 3/51
(au 2e tirage il reste 3 rois et 51 cartes).
P(2 rois) = (3/51) × (4/52) = 0,0045 = 4,5 chances sur mille

Formule de Bayes

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

Combinaisons

Combinaison de n éléments pris k à k : $C_n^k = \frac{n!}{k!(n-k)!}$

Exemple
Combien de combinaisons de rois peut-on constituer avec trois cartes ?
n = 4 couleurs k = 3 cartes
C = 4!/3!1 = 4 ♠♣♥ ♠♣♦ ♣♥♦ ♠♥♦

Probabilités

Probabilité d'un événement A : P(A) avec $0 \leq P(A) \leq 1$
$P(\overline{A}) = 1 - P(A)$ $\Rightarrow P(A) + P(\overline{A}) = 1$

Probabilité de deux événements :

- probabilité de A ou bien B : P (A ou B) = P (A ∪ B)
- probabilité de A et de B : P (A et B) = P (A ∩ B)

Probabilité d'un événement B si un événement A a eu lieu : P (B si A) = P (B|A)

1) A et B sont deux événements incompatibles (disjoints)

P(A ou B) = P(A) + P(B)
P(A et B) = 0
P(B|A) = 0

Exemple : on tire une carte dans un jeu de 52 cartes
Événements : A = roi de cœur, B = roi de trèfle :
P (roi de cœur ou roi de trèfle) = 1/52 + 1/52 = 0,019

2) A et B sont compatibles et indépendants

P(A ou B) = P(A) + P(B) – P(A et B)
P(A et B) = P(A) × P(B)
P(B|A) = P(B)

Exemple : on tire une carte sur 52 cartes
Événements : A = cœur, B = roi
P(cœur ou roi) = 1/4 + 4/52 – 1/52 = 0,31
P(cœur et roi) = 1/4 × 4/52 = 1/52
P(roi si cœur) = P(roi) = 4/52 = 1/13

3) A et B sont compatibles et dépendants : probabilités conditionnelles

P(A et B) = P(A|B) P(B)
= P(B|A) P(A)

$P(B|A) = \frac{P(A \text{ et } B)}{P(A)}$

Exemple : on tire 2 cartes dans un jeu de 52
Événements : A = un roi au 1er tirage,
B = un autre roi au 2e tirage
P(roi au 1er tirage) = 4/52
P(roi au 2e tirage si 1 roi est sorti au 1er) = 3/51
(au 2e tirage il reste 3 rois et 51 cartes).
P(2 rois) = (3/51) × (4/52) = 0,0045 = 4,5 chances sur mille

Formule de Bayes

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

Théorème de Bayes

$$P(B|A) = \frac{P(B)P(A|B)}{P(B)P(A|B) + P(\bar{B})P(A|\bar{B})}$$

Loi binomiale

n : nombre de tirage AVEC remise
k : nombre d'événements favorables parmi les n tirages
P : probabilité de l'événement favorable

$$P(k) = C_n^k P^k (1-P)^{n-k}$$

Exemple : on lance 5 fois une pièce. Quelle est la probabilité d'obtenir 3 fois face ?
P = 0,5, n = 5, k = 3
P(3 faces) = (5.4.3.2)/(3.2) 2. $0,5^3$. $0,5^2$ = 0,31

Loi hypergéométrique

n : nombre de tirage SANS remise: tirage simultané
k : nombre d'événements favorables parmi les n tirages
P : probabilité de l'événement favorable
N : limite maximum du nombre de tirage, taille de l'urne

$$P(k) = \frac{C_{NP}^k \, C_{N(1-P)}^{n-k}}{C_P^n}$$

Exemple
On tire 5 cartes dans un jeu de 32 cartes. Quelle est la probabilité d'obtenir un carré d'as ?
n = 5; k = 4; N = 32; P = 4/32; NP = 4; N(1 – P) = 28
P(4 as) = [4!/4! 1. 28!/1.27!] /28!/(5! 23!) = 0,00014 = 1,4 chances sur 10 000

Formulaire statistique

1. Loi binomiale

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

k : nombre de sujets porteurs d'une caractéristique dans un échantillon
n : taille de l'échantillon
P : fréquence de la caractéristique

2. Fonctions de répartition de la loi binomiale

$P(X < k) = P(0) + P(1) + \ldots\ldots\ldots\ldots + P(k-1) = 1 - P(X \geq k)$
$P(X \leq k) = P(0) + P(1) + \ldots\ldots + P(k-1) + P(k) = 1 - P(X > k)$
$P(X > k) = P(k+1) + \ldots\ldots\ldots\ldots + P(k_n) = 1 - P(X \leq k)$
$P(X \geq k) = P(k) + P(k+1) + \ldots\ldots\ldots\ldots + P(k_n) = 1 - P(X < k)$

3. Loi de Poisson

$$P(X = k) = \frac{e^{-\mu} \mu^k}{k!}$$

μ : nombre moyen d'événements observés dans la population pendant une période donnée
e = 2,718...
X : la variable du nombre d'individus ayant subi l'événement observé pendant la période donnée
k : une valeur de cette variable X
P (X = k) : la probabilité d'observer la valeur k

4. Fonctions de répartition de la loi de Poisson

$P(X < k) = P(0) + P(1) + \ldots\ldots\ldots\ldots + P(k-1)$
$P(X \leq k) = P(0) + P(1) + \ldots\ldots\ldots\ldots + P(k-1) + P(k)$
$P(X > k) = 1 - P(X \leq k)$
$P(X \geq k) = 1 - P(X < k)$

5. La loi normale ou loi de Gauss

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ : moyenne de toutes les valeurs de la variable normale
σ : écart type des valeurs
x : une valeur de la variable
f(x) : la densité de probabilité de chaque valeur de x

6. La variable centrée réduite Z

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

X : variable normale d'origine
μ : moyenne de la variable X
σ : écart type de la variable X

7. Loi normale centrée réduite de Z

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$

8. Paramètres de position et de dispersion d'une variable

VARIABLE	PARAMÈTRE	ÉCHANTILLON	POPULATION
quantitative	Moyenne	$m = \frac{\Sigma x}{n}$	$\mu = \frac{\Sigma x}{N}$
	Variance	$s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$	$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}$
	Écart type	$s = \sqrt{s^2}$	$\sigma = \sqrt{\sigma^2}$
	Coefficient de variation	$CV = \frac{s}{m} \times 100$	$CV = \frac{\sigma}{\mu} \times 100$
qualitative binaire	Pourcentage	$p = k/n$	$P = K/N$
	Variance	$s^2 = p(1-p)$	$\sigma^2 = P(1-P)$
	Écart type	$s = \sqrt{s^2} = \sqrt{p(1-p)}$	$\sigma = \sqrt{\sigma^2} = \sqrt{P(1-P)}$

Σx : sommes des valeurs de la variable n : taille de l'échantillon

9. Estimation d'un paramètre sur un échantillon

	Moyenne	Pourcentage
Variance	$s_m^2 = \frac{s^2}{n}$	$s_p^2 = \frac{p(1-p)}{n}$
Écart type	$s_m = \sqrt{s_m^2} = \frac{s}{\sqrt{n}}$	$s_p = \sqrt{s_p^2} = \sqrt{\frac{p(1-p)}{n}}$
Intervalle de confiance à 95 %	$m - 1{,}96\, s_m < \mu < m + 1{,}96\, s_m$ $\mu = m \pm 1{,}96\, s_m$	$p - 1{,}96\, s_p < P < p + 1{,}96\, s_p$ $P = p \pm 1{,}96\, s_p$
Intervalle de confiance (cas général)	$\mu = m \pm Z_\alpha \frac{s}{\sqrt{n}}$	$P = p \pm Z_\alpha \sqrt{\frac{p(1-p)}{n}}$

n : taille de l'échantillon
s^2 et s : variance et écart type de la variable quantitative calculé dans l'échantillon
μ et P : moyenne et pourcentage inconnu de la population
m et p : moyenne et pourcentage calculé dans l'échantillon

10. Écart type d'une moyenne et d'un pourcentage lorsque la taille n de l'échantillon est grande par rapport à la taille N de la population (n/N > 0,1)

Moyenne	Pourcentage
$s_m = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$	$s_p = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}$

Le terme $(N - n)/(N - 1)$ est appelé ***facteur d'exhaustivité***.
Il faudra utiliser ces formules pour calculer les intervalles de confiance.

11. Intervalle de confiance d'une moyenne dans le cas des petits échantillons

Lorsque la taille de l'échantillon est inférieure à 30, on utilise la loi de Student au lieu de la loi Z. Dans ce cas la formule générale est :

μ : moyenne inconnu de la population
m et s : moyenne et écart type calculé dans l'échantillon
n : taille de l'échantillon

$$\mu = m \pm t_\alpha \frac{s}{\sqrt{n}} \qquad \text{avec ddl} = n - 1$$

La valeur t_α se lit dans la table T de Student (*cf.* Annexes, Formulaire 27, Table 2) sur la ligne correspondant au nombre de ddl. Si on désire un intervalle de confiance à 95 % la valeur se lit dans la colonne $\alpha = 5\ \%$.
Condition d'application : il faut que la distribution dans la population suive une loi normale. Lorsqu'on utilise la distribution de ***t*** pour estimer une moyenne à partir d'un petit échantillon, il faudrait donc en principe soit vérifier cette condition, soit l'assumer en présentant le résultat.

12. Taille des échantillons pour réaliser un test Z

Le tableau donne les tailles minimales d'échantillon pour détecter une différence significative :

TYPE DE COMPARAISON	TAILLE ÉCHANTILLON	
moyenne observée/théorique	$n \geq \frac{\sigma^2}{\Delta^2}(Z_\alpha + Z_{2\beta})^2$	
2 moyennes		
taille échantillon différente	$n_1 \geq \frac{k+1}{k}\frac{\sigma^2}{\Delta^2}(Z_\alpha + Z_{2\beta})^2$	$n_2 = n_1/k$
taille échantillon identique	$n_1 \geq \frac{2\sigma^2}{\Delta^2}(Z_\alpha + Z_{2\beta})^2$	$n_1 = n_2$

Δ : la différence escomptée entre moyennes
$\boldsymbol{\sigma^2}$: la variance de la population d'où sont issus les échantillons sous H_0. Cette variance doit être estimée sur des connaissances préalables
k : le rapport n_1/n_2 (avec $n_1 \geq n_2$). Il faut se fixer préalablement un rapport k de taille raisonnable. En pratique, il est déconseillé de déséquilibrer les échantillons dans un rapport supérieur à 4
n, $\mathbf{n_1}$ et $\mathbf{n_2}$: taille des échantillons
$\mathbf{Z_\alpha}$: on choisit en général un risque α de 5 %, donc $Z_\alpha = 1{,}96$. Si le test est unilatéral il faut remplacer Z_α par $Z_{2\alpha}$ soit 1,645
$\mathbf{Z_{2\beta}}$: on choisit en général un risque β de 20 %, donc $Z_{2\beta} = 0{,}842$. Cette valeur est identique en cas d'hypothèse bilatérale ou unilatérale

13. Corrélation

Soit un échantillon de **n** couples de valeurs x_i et y_i et de moyennes m_x et m_y :

Covariance : $$\text{cov}(xy) = \frac{\Sigma(x_i - m_x)(y_i - m_y)}{n-1}$$

Coefficient de corrélation : $$r = \frac{\Sigma(x_i - m_x)(y_i - m_y)}{\sqrt{\Sigma(x_i - m_x)^2 \Sigma(y_i - m_y)^2}} = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\sqrt{\left[\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}\right]\left[\Sigma y_i^2 - \frac{(\Sigma y_i)^2}{n}\right]}}$$

14. Régression

Soit un échantillon de **n** couples de valeurs x_i et y_i, de moyennes m_x et m_y et de variance s_x^2 et s_y^2 :

Droite de régression	$y = a + bx$
Pente de la droite de régression	$b = \frac{\Sigma(x_i - m_x)(y_i - m_y)}{\Sigma(x_i - m_x)^2} = \frac{cov(xy)}{var(x)}$
Ordonnée à l'origine de la droite de régression	$a = m_y - bm_x$
Écart type de la pente de la droite de régression	$s_b = \sqrt{\frac{\frac{s_y^2}{s_x^2} - b^2}{n-2}}$
Test de la pente de la droite de régression	$t = \frac{\lvert b - 0 \rvert}{s_b}$ $\quad ddl = n - 2$
Coefficient de détermination : r^2	$r^2 = \frac{b^2\Sigma(x_i - m_x)^2}{\Sigma(y_i - m_y)^2} = b^2\frac{s_x^2}{s_y^2}$ (1)
Variance liée (variance de Y pour X fixé) : $s^2_{y/x}$	$s^2_{y/x} = \frac{n-1}{n-2}(s_y^2 - b^2 s_x^2)$
Variance de la moyenne de Y pour X fixé : $s^2_{\hat{y}}$	$s^2_{\hat{y}} = s^2_{y/x}\left(\frac{1}{n} + \frac{(x - m_x)^2}{(n-1)s_x^2}\right)$
Estimation de la valeur moyenne de Y pour X fixé : $\hat{Y}$	$\hat{Y} = \hat{y} \pm t_\alpha\sqrt{s^2_{\hat{y}}}$ (2)
Estimation de la valeur Y pour un individu ayant une valeur X fixée	$y = (a + bx) \pm Z_\alpha\sqrt{s^2_{y/x}\left(1 + \frac{1}{n} + \frac{(x - m_x)^2}{(n-1)s_x^2}\right)}$ (3)

(1) On peut vérifier que r^2 est le carré du coefficient de corrélation.

(2) $\hat{y}$: valeur estimée à partir de l'échantillon.

(3) Lorsque la taille de l'échantillon est supérieure à 30, $y = (a + bx) \pm Z_\alpha\sqrt{s^2_{y/x}}$

15. Intervalles de confiance d'un risque relatif et d'un odds ratio

Méthode de Miettinen pour le calcul d'un intervalle de confiance à 95 %.
χ^2 : valeur du test de χ^2 effectué sur le tableau à 4 cases

IC 95 % D'UN RISQUE RELATIF	IC 95 % D'UN ODDS RATIO
$IC\ 95\ \% = RR^{1 \pm \frac{1,96}{\sqrt{\chi^2}}}$	$IC\ 95\ \% = OR^{1 \pm \frac{1,96}{\sqrt{\chi^2}}}$

16. Nombre de sujets nécessaires à une enquête de cohorte

Cohorte simple entre deux groupes exposés et non-exposés de taille identique.

n : le nombre de sujets nécessaires dans chaque groupe

I_{ne} : incidence de la maladie chez les non-exposés

Z_α : la valeur de Z pour le risque de première espèce (pour $\alpha = 5\ \%$, $Z_\alpha = 1{,}96$)

$Z_{2\beta}$: la valeur de Z pour une puissance $1 - \beta$ (pour une puissance de 80 %, $\beta = 20\ \%$ et $Z_{2\beta} = 0{,}84$)

RR : RR minimum qu'on se fixe pour que l'étude présente un intérêt de santé publique

p : incidence moyenne de la maladie dans les 2 groupes

Calculs intermédiaires :

$$p = \frac{I_{ne}(1 + RR)}{2}$$

$$n \geq \frac{[Z_\alpha \sqrt{2p(1-p)} + Z_{2\beta}\sqrt{I_{ne} + I_{ne}RR - I_{ne}^2 - I_{ne}^2 RR^2}]^2}{[I_{ne}(1 - RR)]^2}$$

17. Nombre de sujets nécessaires à une enquête cas-témoins

Enquête avec un groupe de cas et un groupe témoin.

n : le nombre de cas

c : le nombre de témoins par cas

p_0 : la proportion de témoins exposés

Z_α : la valeur de Z pour le risque de première espèce (pour $\alpha = 5\ \%$, $Z_\alpha = 1{,}96$)

$Z_{2\beta}$: la valeur de Z pour une puissance $1 - \beta$ (pour une puissance de 80 %, $\beta = 20\ \%$ et $Z_{2\beta} = 0{,}84$)

OR : OR minimum qu'on se fixe pour que l'étude présente un intérêt de santé publique

p_1 : proportion de cas exposés

p : proportion de sujets exposés dans les deux groupes cas et témoins

Calculs intermédiaires :

$$p_1 = \frac{p_0 OR}{1 + p_0(OR - 1)} \qquad p = \frac{p_1 + cp_0}{1 + c}$$

$$n \geq \frac{p(1-p)\left(1 + \frac{1}{c}\right)(Z_\alpha + Z_{2\beta})^2}{(p_0 - p_1)^2}$$

Le nombre de témoins est supérieur ou égal à **nc.**

18. Analyse d'enquête étiologique stratifiée

STRATE i	CAS	SAINS	TOTAL
Exposé	a_i	b_i	t_{1i}
Non exposé	c_i	d_i	t_{0i}
Total	n_{1i}	n_{0i}	T_i

Test du χ^2 de Mantel-Haenszel stratifié :

$$\chi^2_{MH} = \frac{[\Sigma a_i - \Sigma(t_{1i}n_{1i}/T_i)]^2}{\Sigma[(t_{1i}t_{0i}n_{1i}n_{0i})/T_i^2(T_i-1)]}$$

Risque relatif et odds ratio ajustés :

	RISQUE RELATIF : RR	ODDS RATIO : OR
Valeur ajustée	$\frac{\Sigma[(a_i t_{0i})/T_i]}{\Sigma[(c_i t_{1i})/T_i]}$	$\frac{\Sigma[(a_i d_i)/T_i]}{\Sigma[(b_i c_i)/T_i]}$
Intervalle de confiance IC 95 %*	$RR_{MH}^{[1\pm(1,96/\sqrt{\chi^2_{MH}})]}$	$OR_{MH}^{[1\pm(1,96/\sqrt{\chi^2_{MH}})]}$

* Méthode approchée de Miettinen.

19. Valeur prédictives d'un test

On appelle
M le fait d'être malade ; $\overline{M}$ le fait d'être sain
S : existence d'un résultat positif au test (signe présent)
$\overline{S}$: résultat négatif au test (signe absent)
P (M) : probabilité d'être malade : c'est la prévalence **Pr** de la maladie dans la population d'étude
P ($\overline{M}$) : probabilité de ne pas être malade. C'est le complément de la prévalence **1 – Pr**
P (S si M) : probabilité d'un résultat positif si le sujet est malade : c'est la sensibilité **Se**
P (S si $\overline{M}$) : probabilité d'un résultat positif si le sujet n'est pas malade : **1 – Sp**
P ($\overline{S}$ si $\overline{M}$) : probabilité d'un résultat négatif si le sujet n'est malade : c'est la spécificité **Sp**
P ($\overline{S}$ si M) : probabilité d'un résultat négatif si le sujet est malade : **1 – Se**

VALEUR PRÉDICTIVE POSITIVE : VPP	VALEUR PRÉDICTIVE NÉGATIVE : VPN
$VPP = P(M \text{ si } S)$	$VPN = P(\overline{M} \text{ si } \overline{S})$
$P(M \text{ si } S)^* = \frac{P(S \text{ si } M)P(M)}{P(S \text{ si } M)P(M) + P(S \text{ si } \overline{M})P(\overline{M})}$	$P(\overline{M} \text{ si } \overline{S}) = \frac{P(\overline{S} \text{ si } \overline{M})P(\overline{M})}{P(\overline{S} \text{ si } \overline{M})P(\overline{M}) + P(\overline{S} \text{ si } M)P(M)}$
$VPP = \frac{SePr}{SePr + (1-Sp)(1-Pr)}$	$VPN = \frac{Sp(1-Pr)}{Sp(1-Pr) + (1-Se)(Pr)}$

* *cf.* théorème de Bayes, Annexes § 23

20. Test exact de Fisher

Les formules suivantes s'appliquent exclusivement à un tableau d'effectifs à 4 cases.

ÉCHANTILLONS / CLASSES DE LA VARIABLE	1	2	TOTAUX
caractère présent	a	b	t_1
caractère absent	c	d	t_2
totaux : effectifs des échantillons	n_1	n_2	N

a, **b**, **c**, **d** : les effectifs observés dans chaque case du tableau.
$\mathbf{n_1}$ et $\mathbf{n_2}$: les effectifs des deux échantillons $n_1 = a + c$ et $n_2 = b + d$.
$\mathbf{t_1}$ et $\mathbf{t_2}$: les totaux des effectifs observés pour les 2 classes de la variable.
N : le total général des effectifs observés dans toutes les cases.

En faisant varier les effectifs des cases sans changer les totaux marginaux pour chaque combinaison **i**, telle que $a_i \geq a - t_1 n_1/N$, on calcule $p_i = n_1!n_2!t_1!t_2!/a!b!c!d!N!$

Le calcul du test donne directement la valeur de p : $p = \Sigma\, p_i$

Le test exact de Fisher peut aussi s'appliquer à des tableaux de contingence comportant plus de 2 lignes et 2 colonnes. Le grand nombre d'itérations de calculs imposent d'utiliser un logiciel.

21. Correction de Yates : test du χ^2 de Yates

Ce test corrigé s'applique exclusivement lorsqu'on désire comparer deux pourcentages (ou bien un pourcentage observé à un pourcentage théorique), et, lorsque au moins un des effectifs théoriques du tableau est inférieur à 5 et supérieur ou égal à 3. Son utilisation est devenue caduque si on dispose de logiciels statistiques et il est préférable d'utiliser le test exact de Fisher.
o_{ij} : effectifs observés dans chacune des cases
c_{ij} : effectifs théoriques dans chacune des cases

$$\chi^2_o = \Sigma \frac{(|o_{ij} - c_{ij}| - 0{,}5)^2}{c_{ij}}$$

22. Test du log rank pour comparer deux courbes de survie

V_{1i} et V_{2i} : nombre de survivants à chaque période i dans chacun des groupes 1 et 2
D_{1i} et D_{2i} : nombre de décédés observés à chaque période i dans chacun des groupes 1 et 2
o_1 et o_2 : total des décédés observés dans les groupes 1 et 2 : $o_1 = \Sigma D_{1i}$ et $o_2 = \Sigma D_{2i}$
c_i : nombre théorique de décédés pendant chaque période i dans le groupe 1 (*cf.* tableau)
c_1 : somme des effectifs théoriques décédés du groupe 1 : $c_1 = \Sigma c_i$
c_2 : somme des effectifs théoriques décédés du groupe 2 : $c_2 = o_1 + o_2 - c_1$

TEMPS	GROUPE 1		GROUPE 2		EFFECTIFS THÉORIQUES
	VIVANTS	DÉCÉDÉS	VIVANTS	DÉCÉDÉS	DU GROUPE 1
t_i	V_{1i}	D_{1i}	V_{2i}	D_{2i}	$c_i = V_{1i}\frac{D_{1i}+D_{2i}}{V_{1i}+V_{2i}}$
Total		$o_1 = \Sigma D_{1i}$		$o_2 = \Sigma D_{2i}$	$c_1 = \Sigma c_i$

$$\chi^2 = \frac{(o_1 - c_1)^2}{c_1} + \frac{(o_2 - c_2)^2}{c_2} \qquad \text{avec ddl} = 1$$

Les résultats et l'interprétation du test de log rank sont identiques à ceux d'un χ^2 à 1 ddl (*cf.* chap. 11.IV).

23. Nombre de sujets nécessaires pour comparer deux pourcentages

$$p = \frac{p_1 + c\,p_2}{1+c}$$

$$n' = \frac{\left(Z_\alpha \sqrt{(c+1)p(1-p)} + Z_{2\beta}\sqrt{c\,p_1(1-p_1) + p_2(1-p_2)}\right)^2}{c(p_2 - p_1)^2}$$

$$n = \frac{n'}{4}\left(1 + \sqrt{1 + \frac{2(c+1)}{n'c\,|p_2 - p_1|}}\right)^2$$

p_1 et p_2 : les deux pourcentages à comparer et à défaut leur estimation probable
c : le ratio de sujets groupe 2/groupe 1
Z_α : la valeur de Z pour un risque α consenti : pour $\alpha = 5\ \%$, $Z_{0,05} = 1,96$
$Z_{2\beta}$: la valeur de Z pour un risque β consenti :
pour une puissance de 80 %, $\beta = 20\ \%$, $2\beta = 40\ \%$ et $Z_{0,40} = 0,842$
n : nombre de sujets nécessaires dans le groupe 1
nc : nombre de sujets nécessaires dans le groupe 2

24. Coefficient de dissymétrie (*skewness*)

$$\gamma_1 = \frac{\sum[(x_i - \mu) / \sigma]^3}{N}$$

X_i : valeurs de la variable
N : taille de la population
μ : moyenne
σ : écart type

Dans une loi normale, $\gamma_1 = 0$.
Calcul à partir des données d'un échantillon (estimateur non biaisé)

$$\gamma_1 = \frac{n}{(n-1)(n-2)} \sum[(x_i - m) / s]^3$$

X_i : valeurs de la variable
n : taille de l'échantillon
m : moyenne de l'échantillon
s : écart type de l'échantillon

Une distribution symétrique se caractérise par γ_1 proche de zéro.
Une distribution étalée vers la droite se caractérise par $\gamma_1 > 0$.
Une distribution étalée vers la gauche se caractérise par $\gamma_1 < 0$.

25. Coefficient d'aplatissement (*kurtosis*)

$$\gamma_2 = \frac{\sum[(x_i - \mu) / \sigma]^4}{N}$$

Dans une loi normale, $\gamma_2 = 3$.
Certains logiciels utilisent l'excès d'aplatissement (Kurtosis excess) (EA) en retranchant la valeur 3, soit $EA = \gamma_2 - 3$.

Calcul à partir des données d'un échantillon (estimateur non biaisé).

Les logiciels qui calculent l'excès d'aplatissement (EA) utilisent la formule :

$$EA = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum[(x_i - m) / s]^3 - 3\frac{n(n-1)^2}{(n-2)(n-3)}$$

Une distribution normale se caractérise par $\gamma_2 = 3$ ou EA proche de 0.
Une distribution pointue se caractérise par $\gamma_2 > 3$ ou $EA > 0$.
Une distribution aplatie se caractérise par $\gamma_2 < 3$ ou $EA < 0$.

26. Coefficient kappa pondéré entre 2 observateurs A et B

Soit le tableau de concordance :

A / B	A_1	A_2	...	A_j	TOTAL
B_1	O_{11}	O_{12}	...	O_{1j}	t_1
B_2	O_{21}	O_{22}	...	O_{2j}	t_2
...	...	...	...	...	
B_i	O_{i1}	O_{i2}	...	O_{ij}	t_i
Total	n_1	n_2	...	n_j	N

Et la matrice des poids attribués :

A / B	A_1	A_2	...	A_j
B_1	w_{11}	w_{12}	...	w_{1j}
B_2	w_{21}	w_{22}	...	w_{2j}
...	...	...	...	...
B_i	w_{i1}	w_{i2}	...	w_{ij}

$$Cc_w = \frac{\sum_{i=1}^{r}\sum_{j=1}^{r} w_{ij}\, o_{ij}}{N} \qquad Ca_w = \frac{\sum_{i=1}^{r}\sum_{j=1}^{r} w_{ij}\, t_i\, n_j}{N^2}$$

r : nombre de modalités de jugement

$$\kappa = \frac{Cc_w - Ca_w}{1 - Ca_w}$$

Bibliographie

STATISTIQUES

- Bouyer (J.). *Méthodes statistiques. Médecine-Biologie*[1]. ESTEM, Éditions INSERM Paris, 1996.
- Bouyer (J.) *et al. Épidémiologie. Principes et méthodes quantitatives*[2]. Éditions INSERM, Paris, 1993.
- Castle (W.M.). *Statistical in small doses*[1]. Churchill Livingstone, New York, 1977.
- Grais (B.). *Méthodes statistiques*[1]. Dunod, Paris, 1992.
- Huguier (M.) et Flahaut (A.). *Biostatistiques au quotidien*[1]. Elsevier, Paris, 2000.
- Jacquard (A.). *Les probabilités*[1]. Que sais-je n° 1571, Presses Universitaires de France, Paris, 1974.
- Lazar (Ph.) et Schwartz (D.). *Éléments de probabilités et statistique*[1]. Flammarion Médecine-Sciences, Paris, 1967.
- Salamon (R.). *Statistique médicale*[1]. Masson, Paris, 1988.
- Schwartz (D.). *Méthodes statistiques à l'usage des médecins et des biologistes*[1]. Flammarion Médecine Sciences, Paris, 1963.
- Schwartz (D.) *et al. L'essai thérapeutique chez l'homme*[1]. Flammarion Médecine Sciences, Paris, 1970.
- Valleron (A.J.). *Introduction à la biostatistique*[1]. Masson, Paris, 1998.
- Wonnacott (T.H.) et (R.J.). *Statistiques. Économie, Gestion, Sciences, Médecine*[1]. Economica, Paris, 1990, John Willey and Sons Inc, New York, 1972.

ÉPIDÉMIOLOGIE

- Ardilly (P.). *Les techniques de sondage*[2]. Éditions Technip, Paris, 1994.
- Bernard (P.M.) et Lapointe (C.). *Mesures statistiques en Épidémiologie*[1]. Presses de l'Université du Québec, Québec, 1995.
- Dabis (F.), Drucker (J.), Moren (A.). *Épidémiologie d'intervention*[1]. Arnette, Paris, 1992.
- Gregg (M.) *et al. Field Epidemiology*[1]. Oxford University Press, New York, Oxford, 1996.
- Hennekens (C.H.) et Buring (J.E.). *Epidemiology in Medicine*[2]. S.L. Mayrent Ed. Little, Brown & C°, Boston Toronto, 1987.
- Jenicek (M.) et Cleroux (R.). *Épidémiologie. Principes Techniques Applications*[1]. Edisem Inc, Montréal, 1982.
- Last (J.M.). *A Dictionary of Epidemiology*[1]. Oxford University Press, New York, Oxford, Toronto, 1995.
- Rothman (K.). *Modern Epidemiology*[2]. Little, Brown and Company, Boston Toronto, 1986.
- Rumeau-Rouquette (C.) *et al. Méthodes en épidémiologie*[1]. Flammarion Médecine Sciences, Paris, 1970.
- Schlesselman (J.J.). *Case-Control Studies. Design, Conduct, Analysis*[2]. Oxford University Press, New York, Oxford, 1982.

1. Initiation.
2. Perfectionnement.

Glossaire

Agrégées (données…) *agregated data* : données regroupées afin de les représenter de façon synthétique, facile à analyser. S'opposent aux données brutes, individuelles.

Ajustement *adjustment* : méthode d'analyse consistant à analyser des données en fonction d'une variable de confusion.

Aléatoire (variable…) *random variable* : qui est produit par le hasard. Une variable aléatoire est une variable qui peut prendre toute valeur produite par le hasard. Un tirage aléatoire est un processus de sélection dans lequel n'intervient que le hasard.

Alternative : *cf.* hypothèse alternative.

Aplatissement (coefficient d') *kurtosis* : paramètre mesurant le caractère plus ou moins aplati d'une distribution par rapport à une distribution de référence.

Appariées (séries…) *matched groups* : séries de variables, d'individus ou d'unités statistiques dont chaque élément de l'une est reliée à un élément de l'autre. Deux séries appariées sont de même taille. L'ensemble est composé de paires.

Appariement *matching* : action de sélectionner deux échantillons, de façon que chaque individu d'un échantillon soit relié par certains critères à un individu correspondant de l'autre échantillon. L'appariement peut s'effectuer également sur plusieurs échantillons, ou sur plusieurs individus du même échantillon.

Attaque (taux…) *attack rate* : incidence cumulée d'une maladie, mesurée lors d'une épidémie.

Base de sondage *sampling list* : liste d'individus ou d'unités statistiques sur laquelle sera réalisé un processus d'échantillonnage.

Bernouilli (variable de…) : type de variable nominale à deux classes (binaire) prenant les valeurs 0 et 1 (*cf.* booléenne, dichotomique, binaire)

Biais *bias* : erreur systématique engendrée par un mauvais échantillonnage. Un calcul effectué sur un échantillon non représentatif de la population, même s'il est juste mathématiquement, sera biaisé sur le plan statistique.

Bilatérale (Hypothèse H_1…) : *cf.* hypothèse bilatérale.

Bimodale (distribution) *bimodal distribution* : se dit d'une distribution possédant deux modes distincts, *cf.* mode.

Binaire (variable…) *binary variable* : type de variable nominale à deux classes (voir Bernouilli, dichotomique, booléenne).

Binomiale (loi, distribution) *binomial distribution*. Loi, modèle de distribution théorique permettant d'estimer un nombre d'événements sur un échantillon connaissant la probabilité générale de l'événement.

Booléenne (variable…) *booelan variable* : type de variable nominale à deux classes prenant les valeurs vrai ou faux (*cf.* v. de Bernouilli, v. dichotomique, v. binaire).

Cas *case* : sujet inclus dans une étude et présentant des critères de définition strictement établis par un protocole.

Cas-témoins (enquête…) *case control study* : type d'enquête épidémiologique consistant à comparer la fréquence d'exposition à un ou plusieurs facteurs de risque entre un groupe de malades (cas) et un groupe de non-malades

Chi-2 (tests du...) *chi-square test* : ensemble de tests servant à comparer des distributions de variables qualitatives ou à tester la liaison entre 2 variables.

Classe *class* : catégorie d'une variable. Pour une variable quantitative, catégorie regroupant plusieurs valeurs entre deux bornes. Ex. : classe d'âge des 5-15 ans.

Coefficient de corrélation *correlation coefficient* : paramètre mesurant la liaison entre deux variables quantitatives.

Coefficient de variation *coefficient of variation* : paramètre de dispersion d'une distribution. C'est le rapport de l'écart type à la moyenne (multiplié par 100). Il permet de mesurer l'indice de dispersion indépendamment des unités de la variable.

Cohorte (enquête de...) *cohort study* : type d'enquête épidémiologique consistant à comparer la survenue d'une maladie entre des groupes exposés et non-exposés à un facteur de risque.

Confusion *confounding* : biais introduit par une variable cachée qui modifie les résultats d'une analyse.

Contingence (tableau de...) *contingency table* : tableau de variables dans lequel les totaux des lignes et des colonnes doit rester fixe.

Corrélation *correlation* : étude de la liaison entre deux variables quantitatives.

Cote *odds* : ratio de la probabilité de survenue d'un événement sur la probabilité de non-survenue de cet événement.

Cote d'exposition *exposure odds* : rapport du nombre de cas exposés et non-exposés ou du nombre de témoins exposés et non-exposés dans une enquête cas-témoins.

Courbe épidémique *epidemic curve* : graphe représentant le nombre de cas survenant au cours d'une épidémie en fonction du temps.

Covariance *covariance* : mesure de la variance combinée de deux variables quantitatives appariées dans une étude de corrélation ou de régression.

Cumulée (incidence...) *cumulative incidence* : mode de calcul d'une incidence s'exprimant par le taux de nouveaux cas d'une maladie pendant une période de temps donnée.

Déciles *deciles* : valeurs qui partagent une distribution en dix groupes d'effectifs égaux.

Degrés de liberté (ddl) *degrees of freedom* : nombre de cases d'un tableau de contingence que l'on peut remplir librement sans modifier les totaux marginaux. De façon plus générale, dans un tableau à i lignes et j colonnes, ddl = (i – 1) (j – 1).

Degré de signification *level of statistical significance, p value* : limite maxima d'un risque pris lorsqu'on accepte une hypothèse alternative. S'exprime par une probabilité *p* inférieure à 0,05.

Densité d'incidence *incidence density* : mode de calcul d'une incidence s'exprimant par un nombre de nouveaux cas d'une maladie par personnes-temps d'exposition.

Détermination (coefficient de...) : carré du coefficient de corrélation. Exprime en pourcentage la part de la variance due à la corrélation.

Dichotomique (variable...) *dichotomous variable* : type de variable nominale à deux classes (*cf.* binaire, Bernouilli, booléenne).

Différence significative *significant difference* : se dit d'une différence observée entre paramètres ou entre des distributions après calcul d'un test statistique ayant abouti au rejet de H_0. Se mesure par la valeur de la probabilité *p*. *Cf.* degré de signification.

Discrète (variable...) *discrete variable* : type de variable quantitative ne pouvant prendre que des valeurs discontinues (par opposition à variable continue).

Dissymétrie (coefficient de...) *skewness* : paramètre mesurant le degré d'asymétrie d'une distribution autour de sa valeur centrale.

Distribution *distribution* : ensemble des effectifs d'une série de données classées selon les valeurs d'une variable.

Donnée *data* : c'est le résultat de la mesure particulière d'une variable faite sur une unité statistique. Une étude aboutit dans un premier temps à un tableau de données brutes.

Écart type *standard deviation* : paramètre de dispersion d'une distribution. C'est la racine carrée de la variance. Il s'exprime dans la même unité que les valeurs de la variable.

Écart type d'une moyenne ou d'un pourcentage *standard error* : appelé parfois erreur-type de la moyenne ou du pourcentage. Paramètre de dispersion d'une moyenne ou d'un pourcentage, eux-mêmes pris comme des variables aléatoires lorsqu'ils sont mesurés sur des échantillons.

Échantillon *sample* : sous-ensemble d'individus, ou d'unités statistiques, choisis dans une population pour réaliser une mesure. Le procédé du choix de l'échantillon conditionne sa représentativité. Seuls les procédés de tirage aléatoire, c'est-à-dire soumis au hasard, garantissent que l'échantillon est représentatif de la population d'où il est issu.

Échantillonnage *sampling* : processus de sélection d'un échantillon.

Échelle *scale* : type de graduation d'une série de valeurs. Échelle arithmétique, logarithmique, arbitraire…

Effectif *size of the class* : dénombrement des individus ou unités statistiques présentant une valeur donnée ou comprise dans une classe d'une variable.

Efficacité vaccinale *vaccine efficacy* : mesure d'impact d'un vaccin : fraction préventive de la maladie dans le groupe des sujets vaccinés. Rapport de la différence de l'incidence de la maladie entre non vaccinés et vaccinés sur l'incidence chez les non vaccinés.

Endémie *endemic disease* : situation propre à l'évolution d'une maladie sur un mode persistant.

Épidémie *epidemic, outbreak, cluster* : survenue d'un nombre de cas d'une maladie (infectieuse ou non) supérieur au nombre de cas habituellement attendu. Termes équivalent : épisode, flambée ou bouffée épidémique, éclosion de cas.

Épidémiologie *epidemiology* : discipline étudiant la distribution des maladies dans une population et analysant les facteurs qui conditionnent leurs fréquences.

Estimateur *estimator* : fonction (ou formule de calcul) appliquée dans un échantillon pour estimer la vraie valeur inconnue d'un paramètre dans la population.

Estimation *estimation* : méthode statistique visant à estimer à partir d'un échantillon un paramètre quantitatif dans une population inconnue. Une estimation aboutit à une mesure encadrée par un intervalle de confiance dans lequel on parie que se trouve la vraie valeur inconnue.

Étendue *range* : paramètre de dispersion mesurant la différence entre les valeurs extrêmes (maximum – minimum) d'une distribution.

Exposition *exposure* : caractérise, dans une enquête épidémiologique, le fait d'être soumis à un facteur pouvant influer sur le risque de survenue d'une maladie.

Facteur de confusion *confounding factor* : tiers facteur dans une enquête étiologique introduisant un biais dans l'analyse.

Facteur de risque *risk factor* : tout événement (alimentaire, comportemental, génétique, environnemental, iatrogène, *etc.*) pouvant favoriser la survenue d'une maladie.

Fluctuations d'échantillonnage *sampling distribution* : ensemble des valeurs possibles qu'un paramètre d'une population (moyenne, pourcentage, *etc.*) peut prendre lorsqu'il est mesuré sur des échantillons.

Foyer *focus, cluster* : localisation géographique délimitée de cas d'une maladie.

Fraction de sondage *sampling rate* : rapport entre la taille d'un échantillon et la taille de la population d'où il est issu.

Fraction étiologique *etiologic fraction* : mesure d'impact. Proportion de cas d'une maladie qu'on peut attribuer à un facteur de risque.

Fraction préventive *prevented fraction* : mesure d'impact. Proportion de cas évités de maladie qu'on peut attribuer à un facteur protecteur.

Fréquence *frequency* : rapport de l'effectif observé pour une classe d'une variable sur le total des individus étudiés pour cette variable. On l'appelle aussi fréquence relative. La fréquence cumulée est la somme des fréquences relatives de plusieurs valeurs ou classes d'une variable, inférieures ou égales à une valeur seuil.

Graphique *graph* : représentation imagée de données statistiques. Il comprend en général des axes, des échelles définissant les valeurs portées sur les axes, un titre, des légendes et un graphe.

Graphe *graph* : partie d'un graphique illustrant la variation d'une variable en fonction d'une autre ou de plusieurs autres variables. Ex. : courbe, histogramme, barres, *etc.*

Grappe (effet de...) *cluster effect* : distorsion de la variance calculée sur une variable issue d'un échantillon sélectionné par un sondage à plusieurs degrés.

Grappe *cluster* : groupe d'individus ou d'unités statistiques sélectionnés dans un échantillon par un processus de sondage en grappes.

Hasard *chance* : de az-zahr, le dé. Des traités entiers de philosophie ont tenté de définir ce qu'est le hasard et s'il existait. La notion de rencontre de deux événements totalement indépendant (si tant est qu'on puisse admettre l'indépendance) peut convenir pour son emploi en statistiques.

Histogramme *histogram* : graphe constitué de barres verticales jointives servant à illustrer la distribution d'une variable quantitative regroupée en classes. La surface de chaque barre est proportionnelle à l'effectif de la classe.

Hypothèse nulle H_0 *null hypothesis* : hypothèse préalable à tout test statistique supposant qu'il n'existe pas de différence entre les paramètres ou les distributions étudiées, ou bien, qu'il n'existe pas de liaison entre les variables étudiées.

Hypothèse alternative H_1 *alternative hypothesis* : hypothèse qu'on acceptera au cas où H_0 serait rejetée. H_1 suppose qu'il existe une différence entre les paramètres ou les distributions étudiées, ou bien qu'il existe une liaison entre les variables étudiées.

Hypothèse bilatérale *two-sided (two-tails) hypothesis* : cas général d'hypothèse alternative où on suppose qu'il existe une différence entre les paramètres étudiés, indépendamment du sens de cette différence : $A \neq B$.

Hypothèse unilatérale *one-sided (one- tail) hypothesis* : type particulier d'hypothèse alternative dans laquelle on s'intéresse à une différence ne s'exerçant que dans un seul sens : $A > B$.

Iatrogène *iatrogenic* : lié aux soins que l'on prodigue à un malade. Maladie iatrogène : maladie engendrée par une action thérapeutique.

Impact (mesure d'...) *impact factor* : indicateur mesurant la proportion d'une maladie qu'on peut attribuer à un facteur de risque ou à un facteur protecteur.

Incidence *incidence* : indicateur épidémiologique mesurant la fréquence de survenue d'une maladie pendant une période de temps donnée.

Interaction *interaction* : *cf.* modificateur de l'effet.

Intervalle de confiance *confidence interval* : bornes inférieure et supérieure entre lesquelles on estime la position d'un paramètre inconnu dans une population à partir des données d'un échantillon.

Intervalle de fluctuation *sampling distribution range* : intervalle contenant toutes les valeurs possibles d'un paramètre (moyenne, pourcentage, *etc.*) mesuré sur des échantillons issus d'une population. L'intervalle de fluctuation est symétrique de part et d'autre du paramètre de l'ensemble de la population.

Intervalle interquartile *interquartile range* : paramètre de dispersion exprimant la différence entre le troisième et le premier quartile. Intervalle semi-interquartile : paramètre de dispersion mesurant la moitié de l'intervalle interquartile.

Kurtosis : *cf.* aplatissement.

Létalité *lethality, case fatality rate* : indicateur particulier de mortalité mesurant la proportion de décès parmi les malades atteints d'une maladie donnée. Indicateur de gravité d'une maladie ou de la qualité des soins.

Liaison *association* : caractérise la notion de dépendance entre deux variables.

Médiane *median* : paramètre de position centrale indiquant la valeur qui partage une distribution en deux effectifs égaux.

Mode *mode* : paramètre de position indiquant la valeur la plus fréquemment observée dans une distribution.

Modèle *model* : fonction mathématique plus ou moins complexe à laquelle on assimile les valeurs d'une variable observée. En supposant que la variable suit le modèle proposé, on peut alors utiliser les lois mathématiques du modèle pour faire des estimations, des comparaisons ou des prédictions sur la variable.

Modificateur de l'effet *effect modificator* : dans une enquête étiologique, variable (ou tiers facteur) dont certaines modalités modifient la liaison observée entre un facteur de risque et la survenue d'une maladie. Appelé également interaction.

Morbidité (indicateur de...) *morbidity rate* : indicateur mesurant la fréquence des maladies.

Mortalité (indicateur de...) *mortality rate* : indicateur mesurant la fréquence des décès.

Moyenne *mean* : paramètre de position centrale d'une distribution mesuré par le rapport de la somme des valeurs à l'effectif.

Moyenne géométrique *geometric mean* : moyenne calculée sur une distribution dont les valeurs suivent une progression géométrique. La moyenne géométrique est égale à la racine énième du produit des n valeurs.

Multivariée (analyse...) *multivariete analysis* : méthode d'analyse d'une enquête étiologique prenant en compte plusieurs facteurs d'exposition et de confusion en même temps.

Nominale (variable...) *nominal variable* : variable qualitative dont les classes ne peuvent être ordonnées. Ex. : nationalité. S'oppose aux variables qualitatives ordinales.

Normale (distribution...) *Gaussian distribution, normal distribution* : se dit d'une distribution suivant le modèle d'une loi normale. La densité de probabilité d'une variable quantitative suivant une loi normale s'exprime sous forme d'une courbe en cloche, symétrique de part et d'autre de la moyenne.

Normale centrée réduite (loi...) *standardized normal distribution* : loi normale particulière de moyenne nulle et d'écart type égal à 1.

Normale centrée réduite (variable...) : *cf.* variable normale centrée réduite.

Nosocomial *nosocomial* : lié aux activités prodiguées dans un établissement de soins. Maladies nosocomiales : maladies engendrées par la fréquentation d'un établissement de soins.

Odds ratio *odds ratio* : estimateur du risque relatif de survenue d'une maladie calculé à l'issue d'une enquête cas-témoins. Signifie l'excès de risque de survenue d'une maladie chez des sujets exposés à un facteur de risque.

Ordinale (variable...) *ordinal variable* : variable qualitative dont les classes peuvent être ordonnées par ordre croissant. Ex. : niveau d'étude primaire, secondaire, supérieur... S'oppose aux variables qualitatives nominales.

Paramètre de position *measure of central tendency* : valeur servant à situer et à résumer l'ensemble d'une distribution. Ex. : moyenne, médiane, mode.

Paramètre de dispersion *measure of dispersion* : valeur servant à résumer la dispersion des valeurs d'une distribution. Ex. : variance, écart type, étendue.

Percentiles *percentiles* : valeurs qui partagent une distribution en 100 groupes d'effectifs égaux. Ex. : le 35^e percentile partage la distribution telle que 35 % des valeurs lui sont inférieures et 65 % lui sont supérieures.

Poisson (loi, distribution) *Poisson distribution* : loi ou modèle de distribution théorique permettant d'estimer la probabilité de survenue d'un événement rare, connaissant sa fréquence moyenne.

Polygone de fréquence *frequency polygon* : graphe servant à illustrer la distribution d'une variable quantitative continue regroupée en classes. Analogue à l'histogramme.

Population *population* : ensemble de tous les individus auxquels s'intéresse une étude.

Pourcentage *percentage* : valeur résultant d'une proportion multipliée par 100. Ex. : proportion 0,15, pourcentage $0{,}15 \times 100 = 15\ \%$.

Précision *precision* : dans une estimation par intervalle de confiance, la précision est le produit de l'écart type du paramètre par la valeur Z_α. La précision est d'autant plus élevée que ce produit est petit.

Prévalence *prevalence* : indicateur épidémiologique caractérisant la fréquence d'une maladie dans une population.

Proportion *proportion* : rapport de deux valeurs dont le numérateur est une part du dénominateur. Le résultat d'une proportion est compris entre 0 et 1.

Prospective (étude, enquête.) *prospective study* : étude recueillant des données sur des événements à venir.

Puissance d'un test *power of a test* : capacité d'un test statistique à faire accepter une hypothèse alternative qui est vraie. La puissance d'un test dépend de la taille des échantillons et du risque de deuxième espèce β.

Pyramide (des âges) *pyramid-shape diagram* : diagramme constitué de barres horizontales et jointives de part et d'autre d'un axe central, servant à illustrer la distribution par âge et par sexe d'une population.

Quartiles *quartiles* : valeurs qui partagent une distribution en quatre groupes d'effectifs égaux. Il existe trois quartiles. Le deuxième quartile est la médiane.

Randomisation *randomization* : action consistant à sélectionner au hasard les sujets d'un échantillon. Équivalant à un tirage au sort.

Rapport de cotes : *cf.* odds ratio.

Ratio *ratio* : rapport de deux valeurs dont le numérateur est une quantité non incluse dans son dénominateur.

Recensement *census* : opération consistant à dénombrer les sujets de l'ensemble d'une population.

Régression *regression* : étude de la liaison entre une variable quantitative X et une variable quantitative Y dépendante de X.

Remise : se dit d'un tirage « avec » ou « sans » remise. Un tirage avec remise consiste à remettre dans la base de sondage un individu ou une unité statistique qui a déjà été tiré. Un tirage sans remise à l'inverse consiste à poursuivre l'opération de sondage en éliminant au fur et à mesure les individus ou unités statistiques déjà tirés.

Représentativité, représentatif *representativeness* : se dit d'un échantillon choisi de telle manière que l'on peut considérer qu'il représente la population dont il est issu. Seul un procédé de sélection aléatoire garantit la représentativité d'un échantillon.

Rétrospective (étude, enquête) *retrospective study* : étude recueillant des données sur des événements passés.

Risque de première espèce (risque α) *type I error, alpha error* : risque d'erreur consentie pour établir un intervalle de confiance ou pour rejeter une hypothèse nulle (H_0) lors d'un test statistique. On choisit très généralement un risque de 5 %. Un test statistique avec un risque α de 5 %, a 5 chances sur 100 de faire rejeter à tort H_0.

Risque de deuxième espèce (risque β) *type II error, beta error* : risque d'erreur lors d'un test statistique aboutissant à ne pas rejeter H_0 alors que H_1 est vraie. S'oppose à la puissance (1-β).

Risque relatif *relative risk, risk ratio* : rapport des taux d'incidence mesurés dans une enquête de cohorte. Signifie l'excès de risque de survenue d'une maladie chez des sujets exposés à un facteur de risque.

Sensibilité *sensitivity* : capacité d'un test clinico-biologique à détecter les malades. Rapport du nombre de vrais positifs sur le nombre total de malades testés.

Série *serie* : ensemble de données élémentaires faisant l'objet d'une étude statistique.

Signification : *cf.* degré de…

Significative (différence…) : *cf.* différence significative.

Skewness : *cf.* dissymétrie.

Sondage *sampling* : méthode utilisée pour extraire un échantillon d'une population.

Source *source* : lieu ou objet à partir duquel se déclenche une épidémie.

Spécificité *specificity* : capacité d'un test clinico-biologique à détecter les sujets sains. Rapport du nombre de vrais négatifs sur le nombre total de sujets sains testés.

Sporadiques (cas) *sporadic cases* : se dit de cas rares de survenue d'une maladie (par opposition aux maladies endémiques ou aux épidémies).

Standardisation *standardization* : méthode d'analyse consistant à analyser et à comparer des taux entre deux ou plusieurs populations en les ajustant sur une variable de confusion. *Cf.* ajustement.

Strates *stratum* : catégories d'une population faisant l'objet d'échantillonnage et/ou d'analyses différentiées.

Stratification *stratification* : action de partager une population d'étude ou un échantillon en sous-groupes selon les classes d'une variable, afin de pratiquer un sondage ou une analyse spécifique dans chaque groupe.

Survie (courbe, analyse de…) *survival analysis* : méthode générale d'analyse de survenue d'un événement dans un groupe de sujets.

Taux *rate* : en épidémiologie, rapport dans lequel intervient la notion de temps. Ex. : taux d'incidence.

Taux de sondage : *cf.* fraction de sondage.

Témoin *control* : individu servant à constituer un groupe de comparaison, issu de la même population que les cas d'une étude.

Tendance *trend* : évolution d'une fréquence en fonction du temps ou de toute autre variable.

Test statistique *statistical test* : méthode statistique consistant à poser et à vérifier un système d'hypothèses préalablement établies.

Test biologique *biological assay* : examen médical à visée diagnostique.

Tirage au sort *random sampling* : processus de sélection d'un individu ou d'une unité statistique en vue de composer un échantillon. Les termes tirage au sort, tirage au hasard, tirage aléatoire, signifient que la sélection a été effectuée en utilisant le hasard.

Transversale (étude, enquête…) *cross sectionnal survey* : étude recueillant sur un intervalle de temps court l'ensemble des données d'une enquête, c'est-à-dire à la fois les observations sur les cas de maladie et sur les facteurs de risque.

Unilatérale (Hypothèse H_1…) : *cf.* hypothèse unilatérale.

Unimodale : se dit d'une distribution ne présentant qu'un seul mode.

Unité statistique *statistical unit* : unité élémentaire d'observation. Une unité statistique peut être soit un être humain, soit un animal, soit toute sorte d'objets sur lesquels une mesure peut-être faite. Une unité statistique peut aussi être un groupe d'êtres ou de choses sur lequel une mesure élémentaire est opérée. Par exemple une école, un immeuble, une boite de Pétri, *etc.* Une étude en statistique porte sur un ensemble d'unités statistiques.

Valeur *value* : expression chiffrée, codée ou nommée d'une donnée. On la formalise souvent par x_i.

Valeur prédictive positive *positive predictive value* : proportion de sujets réellement malades après résultat positif d'un test clinico-biologique appliqué sur une population définie. Probabilité d'être malade si le résultat du test est positif.

Valeur prédictive négative *negative predictive value* : proportion de sujets non malades après résultat négatif d'un test clinico-biologique appliqué sur une population définie. Probabilité de ne pas être malade si le résultat du test est négatif.

Variable *variable* : c'est une caractéristique de la population que l'on étudie. Dans les sciences de la vie, la plupart des caractéristiques que l'on mesure ou que l'on observe sont différentes d'un individu à l'autre.

Variable normale centrée réduite : *standard normal deviate* : variable quantitative X ayant subit une transformation mathématique du type $(X - \mu)/\sigma$, avec μ : moyenne des valeurs de la variable et σ : écart type.

Variance *variance* : paramètre de dispersion d'une distribution : moyenne de la somme des carrés des écarts à la moyenne.

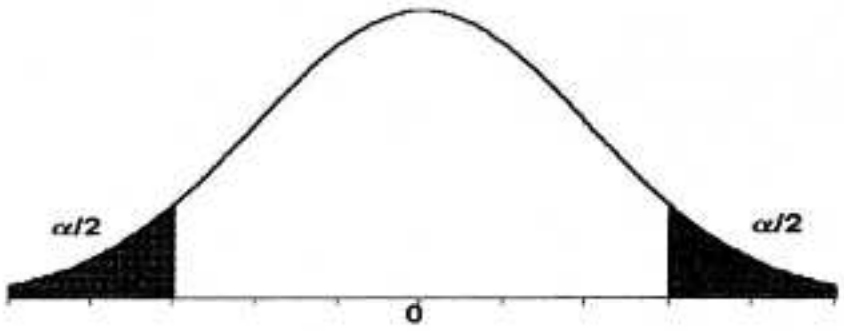
Véhicule *vehicle of infection transmission* : moyen de propagation d'une épidémie.

Tables statistiques

Table 1 : loi normale centrée réduite Z

Test bilatéral

α	Z
0,00001	4,414
0,0001	3,891
0,001	3,290
0,01	2,576
0,02	2,326
0,03	2,170
0,04	2,054
0,05	1,960
0,06	1,881
0,07	1,812
0,08	1,751
0,09	1,695
0,1	1,645
0,2	1,282
0,3	1,036
0,4	0,842
0,5	0,674
0,6	0,524
0,7	0,385
0,8	0,253
0,9	0,126
1,0	0,000



La table, conçue pour un test bilatéral, donne la probabilité α que la valeur absolue |Z| soit supérieure à une valeur donnée. Si le test est unilatéral, il faut **diviser** la valeur α par 2.

Exemples

1) Test bilatéral : il y a 5 chances sur 100 pour que |Z| soit supérieur à 1,960.
2) Test unilatéral :
   - Il y a 2,5 chances sur 100 pour que |Z| soit supérieur à + 1,960.
   - Il y a 5 chances sur 100 pour que |Z| soit supérieur à 1,645.

Une valeur z_0 observée de 2,18 permet de rejeter H_0 avec $p < 3$ % si le test est bilatéral, ou bien avec $p < 1,5$ % si le test est unilatéral.

Table 2 : loi T de Student

Test bilatéral

α / ddl	0,0001	0,001	0,01	0,02	0,03	0,04	0,05	0,1	0,2	0,3	0,5	0,9
1	6370,544	636,578	63,656	31,821	21,205	15,894	12,706	6,314	3,078	1,963	1,000	0,158
2	100,136	31,600	9,925	6,965	5,643	4,849	4,303	2,920	1,886	1,386	0,816	0,142
3	28,014	12,924	5,841	4,541	3,896	3,482	3,182	2,353	1,638	1,250	0,765	0,137
4	15,534	8,610	4,604	3,747	3,298	2,999	2,776	2,132	1,533	1,190	0,741	0,134
5	11,176	6,869	4,032	3,365	3,003	2,757	2,571	2,015	1,476	1,156	0,727	0,132
6	9,080	5,959	3,707	3,143	2,829	2,612	2,447	1,943	1,440	1,134	0,718	0,131
7	7,888	5,408	3,499	2,998	2,715	2,517	2,365	1,895	1,415	1,119	0,711	0,130
8	7,120	5,041	3,355	2,896	2,634	2,449	2,306	1,860	1,397	1,108	0,706	0,130
9	6,594	4,781	3,250	2,821	2,574	2,398	2,262	1,833	1,383	1,100	0,703	0,129
10	6,212	4,587	3,169	2,764	2,527	2,359	2,228	1,812	1,372	1,093	0,700	0,129
11	5,923	4,437	3,106	2,718	2,491	2,328	2,201	1,796	1,363	1,088	0,697	0,129
12	5,695	4,318	3,055	2,681	2,461	2,303	2,179	1,782	1,356	1,083	0,695	0,128
13	5,513	4,221	3,012	2,650	2,436	2,282	2,160	1,771	1,350	1,079	0,694	0,128
14	5,364	4,140	2,977	2,624	2,415	2,264	2,145	1,761	1,345	1,076	0,692	0,128
15	5,239	4,073	2,947	2,602	2,397	2,249	2,131	1,753	1,341	1,074	0,691	0,128
16	5,134	4,015	2,921	2,583	2,382	2,235	2,120	1,746	1,337	1,071	0,690	0,128
17	5,043	3,965	2,898	2,567	2,368	2,224	2,110	1,740	1,333	1,069	0,689	0,128
18	4,966	3,922	2,878	2,552	2,356	2,214	2,101	1,734	1,330	1,067	0,688	0,127
19	4,899	3,883	2,861	2,539	2,346	2,205	2,093	1,729	1,328	1,066	0,688	0,127
20	4,838	3,850	2,845	2,528	2,336	2,197	2,086	1,725	1,325	1,064	0,687	0,127
21	4,785	3,819	2,831	2,518	2,328	2,189	2,080	1,721	1,323	1,063	0,686	0,127
22	4,736	3,792	2,819	2,508	2,320	2,183	2,074	1,717	1,321	1,061	0,686	0,127
23	4,694	3,768	2,807	2,500	2,313	2,177	2,069	1,714	1,319	1,060	0,685	0,127
24	4,654	3,745	2,797	2,492	2,307	2,172	2,064	1,711	1,318	1,059	0,685	0,127
25	4,619	3,725	2,787	2,485	2,301	2,167	2,060	1,708	1,316	1,058	0,684	0,127
26	4,587	3,707	2,779	2,479	2,296	2,162	2,056	1,706	1,315	1,058	0,684	0,127
27	4,556	3,689	2,771	2,473	2,291	2,158	2,052	1,703	1,314	1,057	0,684	0,127
28	4,531	3,674	2,763	2,467	2,286	2,154	2,048	1,701	1,313	1,056	0,683	0,127
29	4,505	3,660	2,756	2,462	2,282	2,150	2,045	1,699	1,311	1,055	0,683	0,127
30	4,482	3,646	2,750	2,457	2,278	2,147	2,042	1,697	1,310	1,055	0,683	0,127
∞	3,892	3,291	2,576	2,327	2,170	2,054	1,960	1,645	1,282	1,036	0,675	0,126

La table construite pour un test bilatéral donne la probabilité a que la valeur absolue de T soit supérieure à une valeur donnée en tenant compte du nombre de degré de liberté (ddl). Si le test est unilatéral il faut **diviser** le risque obtenu par 2.

Exemples

Avec ddl = 10

1) Test bilatéral : Il y a 5 chances sur 100 pour que |T| soit supérieur à 2,228.
2) Test unilatéral
   - Il y a 2,5 chances sur 100 pour que |T| soit supérieur à 2,228.
   - Il y a 5 chances sur 100 pour que T soit supérieur à 1,812.

Une valeur t_o observée de 2,53 et ddl = 10 permet de rejeter H_0 avec $p < 3\ \%$ si le test est bilatéral, ou bien avec $p < 1{,}5\ \%$ si le test est unilatéral.

Remarque : on constate que si ddl ≥ 30, la distribution T de Student se rapproche de la distribution de Z.

Tables 3 : loi F de Fisher (test unilatéral)

$\alpha = 0{,}05$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	∞
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,62	4,56	4,52	4,50	4,46	4,44	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,94	3,87	3,83	3,81	3,77	3,75	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,51	3,44	3,40	3,38	3,34	3,32	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,22	3,15	3,11	3,08	3,04	3,02	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,01	2,94	2,89	2,86	2,83	2,80	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,85	2,77	2,73	2,70	2,66	2,64	2,54
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,40	2,33	2,28	2,25	2,20	2,18	2,07
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,20	2,12	2,07	2,04	1,99	1,97	1,84
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,09	2,01	1,96	1,92	1,87	1,84	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,01	1,93	1,88	1,84	1,79	1,76	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,92	1,84	1,78	1,74	1,69	1,66	1,51
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03	1,87	1,78	1,73	1,69	1,63	1,60	1,44
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93	1,77	1,68	1,62	1,57	1,52	1,48	1,28
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,67	1,57	1,51	1,46	1,39	1,35	1,00

$\alpha = 0{,}025$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	∞
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,43	6,33	6,27	6,23	6,18	6,14	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,27	5,17	5,11	5,07	5,01	4,98	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,57	4,47	4,40	4,36	4,31	4,28	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,10	4,00	3,94	3,89	3,84	3,81	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,77	3,67	3,60	3,56	3,51	3,47	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,52	3,42	3,35	3,31	3,26	3,22	3,08
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,86	2,76	2,69	2,64	2,59	2,55	2,40
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,57	2,46	2,40	2,35	2,29	2,25	2,09
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,41	2,30	2,23	2,18	2,12	2,08	1,91
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,31	2,20	2,12	2,07	2,01	1,97	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,18	2,07	1,99	1,94	1,88	1,83	1,64
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32	2,11	1,99	1,92	1,87	1,80	1,75	1,55
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18	1,97	1,85	1,77	1,71	1,64	1,59	1,35
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,83	1,71	1,63	1,57	1,48	1,43	1,00

Les tables conçues pour un test **unilatéral** donnent les probabilités α que F soit supérieur à une valeur donnée en tenant compte des degrés de libertés de chaque échantillon k_1 et k_2.
Si le test est bilatéral il faut **multiplier** le risque obtenu par deux.

Exemples

Avec $k_1 = 30$ et $k_2 = 40$

- Il y a 5 chances sur 100 pour que F soit supérieur à 1,74 et 2,5 chances sur 100 pour que F soit supérieur à 1,94.
- Une valeur observée de $Fo = 1{,}75$ permet de rejeter H_0 avec $p < 5\,\%$ si le test est unilatéral (cas de l'analyse de variance) et ne permet pas de rejeter H_0 si le test est bilatéral.
- Une valeur observée de $F = 1{,}95$ permet de rejeter H_0 avec $p < 2{,}5\,\%$ si le test est unilatéral et $p < 5\,\%$ si le test est bilatéral.

Tables 3 (suite) : loi F de Fisher (test unilatéral)

$\alpha = 0{,}01$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	∞
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,72	9,55	9,45	9,38	9,29	9,24	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,56	7,40	7,30	7,23	7,14	7,09	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,31	6,16	6,06	5,99	5,91	5,86	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,52	5,36	5,26	5,20	5,12	5,07	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,96	4,81	4,71	4,65	4,57	4,52	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,56	4,41	4,31	4,25	4,17	4,12	3,91
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,52	3,37	3,28	3,21	3,13	3,08	2,87
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,09	2,94	2,84	2,78	2,69	2,64	2,42
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,85	2,70	2,60	2,54	2,45	2,40	2,17
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,70	2,55	2,45	2,39	2,30	2,25	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,52	2,37	2,27	2,20	2,11	2,06	1,80
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70	2,42	2,27	2,17	2,10	2,01	1,95	1,68
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50	2,22	2,07	1,97	1,89	1,80	1,74	1,43
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,04	1,88	1,77	1,70	1,59	1,52	1,00

$\alpha = 0{,}001$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	∞
5	47,18	37,12	33,20	31,08	29,75	28,83	28,17	27,65	27,24	26,91	25,91	25,39	25,08	24,87	24,60	24,44	23,79
6	35,51	27,00	23,71	21,92	20,80	20,03	19,46	19,03	18,69	18,41	17,56	17,12	16,85	16,67	16,44	16,31	15,75
7	29,25	21,69	18,77	17,20	16,21	15,52	15,02	14,63	14,33	14,08	13,32	12,93	12,69	12,53	12,33	12,20	11,70
8	25,41	18,49	15,83	14,39	13,48	12,86	12,40	12,05	11,77	11,54	10,84	10,48	10,26	10,11	9,92	9,80	9,33
9	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,24	8,90	8,69	8,55	8,37	8,26	7,81
10	21,04	14,90	12,55	11,28	10,48	9,93	9,52	9,20	8,96	8,75	8,13	7,80	7,60	7,47	7,30	7,19	6,76
15	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,54	5,25	5,07	4,95	4,80	4,70	4,31
20	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,56	4,29	4,12	4,00	3,86	3,77	3,38
25	13,88	9,22	7,45	6,49	5,89	5,46	5,15	4,91	4,71	4,56	4,06	3,79	3,63	3,52	3,37	3,28	2,89
30	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	3,75	3,49	3,33	3,22	3,07	2,98	2,59
40	12,61	8,25	6,59	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,40	3,15	2,98	2,87	2,73	2,64	2,23
50	12,22	7,96	6,34	5,46	4,90	4,51	4,22	4,00	3,82	3,67	3,20	2,95	2,79	2,68	2,53	2,44	2,03
100	11,50	7,41	5,86	5,02	4,48	4,11	3,83	3,61	3,44	3,30	2,84	2,59	2,43	2,32	2,17	2,08	1,62
∞	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96	2,51	2,27	2,10	1,99	1,84	1,73	1,00

Les tables conçues pour un test **unilatéral** donnent les probabilités α que F soit supérieur à une valeur donnée en tenant compte des degrés de libertés de chaque échantillon k_1 et k_2.
Si le test est bilatéral il faut **multiplier** le risque par deux.

Exemples

Avec $k_1 = 30$ et $k_2 = 40$

- Il y a 1 chance sur 100 pour que F soit supérieur à 2,20 et 1 chance sur mille que F soit supérieur à 2,87.
- Une valeur observée de $F = 2{,}5$ permet de rejeter H_0 avec $p < 1$ % si le test est unilatéral (cas de l'analyse de variance) et $p < 2$ % si le test est bilatéral.

Table 4 : loi du χ^2

α ddl	0,0001	0,001	0,01	0,02	0,03	0,04	0,05	0,10	0,20	0,30	0,50	0,90
1	15,13	10,83	6,63	5,41	4,71	4,22	3,84	2,71	1,64	1,07	0,45	0,02
2	18,42	13,82	9,21	7,82	7,01	6,44	5,99	4,61	3,22	2,41	1,39	0,21
3	21,10	16,27	11,34	9,84	8,95	8,31	7,81	6,25	4,64	3,66	2,37	0,58
4	23,51	18,47	13,28	11,67	10,71	10,03	9,49	7,78	5,99	4,88	3,36	1,06
5	25,75	20,51	15,09	13,39	12,37	11,64	11,07	9,24	7,29	6,06	4,35	1,61
6	27,85	22,46	16,81	15,03	13,97	13,20	12,59	10,64	8,56	7,23	5,35	2,20
7	29,88	24,32	18,48	16,62	15,51	14,70	14,07	12,02	9,80	8,38	6,35	2,83
8	31,83	26,12	20,09	18,17	17,01	16,17	15,51	13,36	11,03	9,52	7,34	3,49
9	33,72	27,88	21,67	19,68	18,48	17,61	16,92	14,68	12,24	10,66	8,34	4,17
10	35,56	29,59	23,21	21,16	19,92	19,02	18,31	15,99	13,44	11,78	9,34	4,87
11	37,36	31,26	24,73	22,62	21,34	20,41	19,68	17,28	14,63	12,90	10,34	5,58
12	39,13	32,91	26,22	24,05	22,74	21,79	21,03	18,55	15,81	14,01	11,34	6,30
13	40,87	34,53	27,69	25,47	24,12	23,14	22,36	19,81	16,98	15,12	12,34	7,04
14	42,58	36,12	29,14	26,87	25,49	24,49	23,68	21,06	18,15	16,22	13,34	7,79
15	44,26	37,70	30,58	28,26	26,85	25,82	25,00	22,31	19,31	17,32	14,34	8,55
16	45,93	39,25	32,00	29,63	28,19	27,14	26,30	23,54	20,47	18,42	15,34	9,31
17	47,56	40,79	33,41	31,00	29,52	28,44	27,59	24,77	21,61	19,51	16,34	10,09
18	49,19	42,31	34,81	32,35	30,84	29,75	28,87	25,99	22,76	20,60	17,34	10,86
19	50,79	43,82	36,19	33,69	32,16	31,04	30,14	27,20	23,90	21,69	18,34	11,65
20	52,38	45,31	37,57	35,02	33,46	32,32	31,41	28,41	25,04	22,77	19,34	12,44
21	53,96	46,80	38,93	36,34	34,76	33,60	32,67	29,62	26,17	23,86	20,34	13,24
22	55,52	48,27	40,29	37,66	36,05	34,87	33,92	30,81	27,30	24,94	21,34	14,04
23	57,07	49,73	41,64	38,97	37,33	36,13	35,17	32,01	28,43	26,02	22,34	14,85
24	58,61	51,18	42,98	40,27	38,61	37,39	36,42	33,20	29,55	27,10	23,34	15,66
25	60,14	52,62	44,31	41,57	39,88	38,64	37,65	34,38	30,68	28,17	24,34	16,47
26	61,67	54,05	45,64	42,86	41,15	39,89	38,89	35,56	31,79	29,25	25,34	17,29
27	63,17	55,48	46,96	44,14	42,41	41,13	40,11	36,74	32,91	30,32	26,34	18,11
28	64,66	56,89	48,28	45,42	43,66	42,37	41,34	37,92	34,03	31,39	27,34	18,94
29	66,15	58,30	49,59	46,69	44,91	43,60	42,56	39,09	35,14	32,46	28,34	19,77
30	67,62	59,70	50,89	47,96	46,16	44,83	43,77	40,26	36,25	33,53	29,34	20,60
40	82,06	73,40	63,69	60,44	58,43	56,95	55,76	51,81	47,27	44,16	39,34	29,05
50	95,97	86,66	76,15	72,61	70,42	68,80	67,50	63,17	58,16	54,72	49,33	37,69
100	161,33	149,45	135,81	131,14	128,24	126,08	124,34	118,50	111,67	106,91	99,33	82,36

La table construite pour un test bilatéral donne la probabilité α que la valeur de χ^2 soit supérieure à une valeur donnée en tenant compte du nombre de degré de liberté (ddl). Si le test est unilatéral il faut diviser le risque obtenu par 2.

Exemples

Avec ddl = 10

1) Test bilatéral : il y a 5 chances sur 100 pour que χ^2 soit supérieur à 18,31.
2) Test unilatéral : il y a 2,5 chances sur 100 pour que χ^2 soit supérieur à 18,31 et 5 chances sur 100 pour que χ^2 soit supérieur à 15,99.

Une valeur χ^2_0 observée de 23,3 et ddl = 10 permet de rejeter H_0 avec $p < 1$ % si le test est bilatéral, ou bien avec $p < 0{,}5$ % si le test est unilatéral.

Remarque : on constate que, si ddl = 1, $\sqrt{\chi^2} = Z$: (par exemple: $\sqrt{3{,}84} = 1{,}96$)

Index

Cet ouvrage se propose de rendre attractives et compréhensibles les disciplines de la statistique et de l'épidémiologie pour les étudiants en sciences de la santé, mais aussi pour tous les professionnels de santé (médecins, pharmaciens, biologistes, infirmières, professions paramédicales, vétérinaires).

Il facilite la compréhension des principes fondamentaux grâce auxquels il devient possible, à partir de nombreux exemples et exercices, d'utiliser les tests statistiques les plus appropriés pour une recherche ou pour la conduite d'une enquête épidémiologique.

La première partie étudie les outils servant à écrire les données. La deuxième aborde les méthodes d'estimation d'un paramètre inconnu à partir d'un échantillon. La troisième concerne l'emploi des tests statistiques. Elle comporte de nombreux tableaux pratiques d'aide au choix d'un test en fonction de la nature des problèmes, des paramètres à comparer et des conditions d'application ; cette partie est complétée par une série de « fiches pratiques » des principaux tests usuels. La quatrième partie est orientée vers les concepts statistiques utilisés en épidémiologie de terrain.

Cet ouvrage a pour ambition de fournir au lecteur une aide pratique et de lui communiquer l'envie d'approfondir les notions de statistique et d'épidémiologie qu'il aura entrevues.

ISBN : 978-2-224-03042-1
9 782224 030421

www.maloine.fr