Université Victor Segalen Bordeaux 2
Institut de Santé Publique, d'Épidémiologie et de Développement (ISPED)

Campus Numérique SEME

MODULE

Principaux outils en statistique

Version du 28 août 2008

Écrit par : **Geneviève Chêne, Marianne Savès**
Relu par : **Ahmadou Alioum, Marthe-Aline Jutand, Valériane Leroy, Louis-Rachid Salmi**

SOMMAIRE

L'épidémiologie s'appuie très largement sur les outils statistiques, qu'il s'agisse de la moyenne, de l'écart-type, de la proportion, de l'intervalle de confiance, mais aussi des tests statistiques. Les notions de base en statistique utiles à la pratique de l'épidémiologie sont donc abordées dès ce module, car elles seront essentielles tout au long de cette formation en épidémiologie.

Introduction

La **statistique** est une « méthode de raisonnement permettant d'interpréter le genre de données très particulières, qu'on rencontre notamment dans les sciences de la vie, dont le caractère essentiel est la variabilité » (D. Schwartz). En cela, c'est un outil indispensable à l'interprétation des résultats des enquêtes épidémiologiques.
La **variabilité** est un caractère essentiel des êtres vivants et donc de l'être humain, en particulier.

> Par exemple, certaines caractéristiques de l'être humain, comme le poids ou la quantité de sucre dans le sang, varient d'un sujet à l'autre ou de l'enfance à l'âge adulte, parfois même d'un moment à l'autre de la journée.

La présence d'une maladie peut également expliquer la variabilité d'une caractéristique.

> Par exemple, au cours de certaines maladies, le nombre de globules rouges circulant par unité de volume de sang peut être plus bas, témoignant d'une anémie.

L'épidémiologiste est confronté en permanence à ce phénomène de variabilité car les questions qu'il essaie de résoudre sont à l'échelon d'un groupe, d'une **population** et non pas d'un seul individu. Une des solutions est de décrire les propriétés moyennes des groupes d'individus. Pour cela, le traitement des données et la communication des résultats nécessitent l'utilisation de la statistique. Par ailleurs, le plus souvent, l'étude de la population entière est rarement possible car elle est trop vaste : des milliers, voire des centaines de milliers de sujets. Même si c'était possible, il faudrait des moyens trop importants. Il faut donc se résoudre à sélectionner un **échantillon**, le décrire et en tirer des conclusions sur la population (**inférence**). Là encore, la méthode statistique est indispensable.
Au total, la méthode statistique intervient à tous les échelons d'une enquête dont l'objectif est de recueillir des informations sur un groupe d'individus : choix du meilleur schéma d'étude, recueil des données, analyse des données. Il convient donc de bien en connaître les grands principes et les outils de base. Mais, avant tout, il vous faut bien comprendre la nature des informations qui sont recueillies afin de choisir ensuite les méthodes les plus appropriées pour les interpréter.

Remarque : tout au long de ce chapitre, les formules mathématiques ont été simplifiées aussi souvent que possible.

Différents types de variables

I Introduction

Une **variable** est une caractéristique dont on peut observer des valeurs différentes au sein d'un groupe de sujets. Une variable peut être de nature **catégorielle** ou de nature **quantitative**.

Dans la suite de ce chapitre, vous verrez que les modes de représentation et les méthodes d'analyse diffèrent selon que l'on a affaire à une variable catégorielle ou à une variable quantitative. Il est donc important de bien comprendre leurs différences pour les distinguer avec assurance.

II Variable catégorielle ou qualitative

Une variable dite **catégorielle** ou **qualitative** est une caractéristique ayant un certain nombre de catégories ou modalités, **exhaustives** et **mutuellement exclusives** : exhaustives car toutes les modalités possibles sont citées, mutuellement exclusives car chaque individu peut être classé dans une catégorie et une seule.

Quand il s'agit de classer les sujets selon deux catégories, la variable catégorielle est dite **dichotomique** (ou **binaire**).

> Par exemple, si l'on dénombre les hommes et les femmes dans un groupe, la variable « sexe » est une variable catégorielle à deux catégories : « hommes » et « femmes ». On peut également classer les sujets selon qu'ils sont fumeurs ou non fumeurs, selon qu'ils sont atteints ou non d'allergie, selon qu'ils ressentent ou non une douleur.

Certaines caractéristiques se décrivent naturellement en plus de deux catégories.
Certaines de ces variables catégorielles sont dites **nominales** : chaque classe désigne une catégorie de sujets (elle les nomme). Il n'existe pas d'ordre naturel entre les catégories.

> C'est, par exemple, le cas du groupe sanguin : A / B / AB / O ou encore de la situation familiale : marié / vivant en couple / célibataire / divorcé / séparé / veuf.

Pour d'autres variables, il existe un ordre naturel entre les différentes catégories. Ces variables sont dites **ordinales**.

> Par exemple, lorsque l'on interroge des sujets sur la sévérité d'une douleur : au lieu de deux catégories (douleur / pas de douleur), on peut classer les individus selon les catégories suivantes : aucune / minime / modérée / sévère / insupportable.

La transformation d'une variable catégorielle ordinale en variable catégorielle dichotomique est toujours possible. L'analyse des données est simplifiée, mais la transformation aboutit à une perte d'information.

> Par exemple, si la sévérité de la douleur a été recueillie selon les catégories : « aucune / minime / modérée / sévère / insupportable », on peut être moins précis en classant les sujets selon les 2 catégories : « douleur / pas de douleur » d'une variable catégorielle dichotomique.

III Variable quantitative

Les valeurs d'une variable **quantitative** sont obtenues par un instrument de mesure ou le résultat d'un dénombrement. Elles sont souvent accompagnées d'une unité de mesure. Avec une telle variable, on peut toujours répondre à une question commençant par : « combien … ?».
Une variable est **continue** si elle peut prendre, en théorie, un nombre infini de valeurs dans un intervalle donné, et si la précision avec laquelle on la mesure ne dépend que de l'exactitude de l'instrument de mesure.

> L'âge, la pression artérielle systolique et la quantité de sucre dans le sang en sont des exemples.

Lorsque l'on arrondit la valeur obtenue, on dit que l'on **discrétise** cette variable continue, car on lui impose de prendre certaines valeurs.

> Par exemple, au lieu d'exprimer l'âge calculé de la date de naissance au jour de visite d'un sujet en jours, on exprime couramment l'âge en années (22, 23, 24 ans, etc.) ou encore de dix ans en dix ans (20 à 29, 30 à 39 ans, etc.).

Dans le premier exemple, l'**intervalle** entre chaque valeur a une amplitude d'une année, dans le deuxième exemple, l**'amplitude d'intervalle** est de 10 ans.

Si les intervalles ne sont pas de même amplitude, on parle plutôt de **regroupement**.

> Par exemple, pour décrire l'âge d'enfants vus dans une consultation pédiatrique, on peut utiliser les regroupements : « 0-3 mois », « 4 –11 mois », « 1-3 ans », « 4-10 ans ».

On parle également de variable **discrète** lorsque la variable est, à l'origine, une variable qui ne peut prendre que certaines valeurs numériques.

> Par exemple, le nombre d'enfants d'une famille est une variable quantitative discrète qui peut prendre les valeurs : 0, 1, 2, 3, 4, 5, … Une famille ne peut avoir 1,4 enfants, ni 2,5 enfants.

Au premier abord, la distinction n'est donc pas simple entre certaines variables catégorielles ordinales, comme le stade d'un cancer, qui pourrait être codé par exemple 1, 2, 3 ou 4 et les variables quantitatives discrètes, comme le nombre d'enfants. Un petit test est néanmoins facile à réaliser pour distinguer les deux types de variables. Pour une variable catégorielle ordinale, chaque différence entre les catégories ne signifie pas la même chose. En revanche, pour une variable quantitative discrète, chaque différence entre les catégories a toujours la même signification sur toute l'étendue des valeurs.

> Par exemple, pour la variable « stade de cancer », on ne peut pas dire que le stade 2 est deux fois plus grave que le stade 1 ; c'est donc une variable catégorielle ordinale. Pour la variable « nombre d'enfants », on peut dire que deux enfants, c'est deux fois plus que un, et que trois enfants c'est trois fois plus que un ; c'est donc une variable quantitative discrète.

Représentation des données

I Introduction

Décrire les données que l'on a rassemblées pour répondre à une question est une première étape très importante en épidémiologie. Pour chaque type de variable, catégorielle ou quantitative, il existe des formes de représentations différentes qui permettent d'avoir une première impression visuelle. On peut utiliser un **tableau** ou un **graphique**.

Si un tableau est plus utile pour présenter de façon complète et précise les données, un graphique est, en revanche, plus utile pour donner une impression visuelle immédiate. Le principal **critère de choix** réside donc dans la façon dont on souhaite communiquer les résultats : si l'on souhaite disposer de l'ensemble des résultats chiffrés, on choisira plutôt un tableau ; si l'on souhaite visualiser une tendance évolutive, on choisira plutôt un graphique. On veillera à ne pas représenter les mêmes données par un tableau plus un graphique, mais à choisir l'un ou l'autre. Par ailleurs, il doit exister une cohérence (format du titre, contenu) entre des données similaires dans un même tableau ou un même graphique, ou entre des tableaux (ou des graphiques) similaires.

Quelle que soit la forme de représentation des données, quelques **principes simples** doivent être appliqués afin que l'interprétation soit évidente pour les lecteurs :

- chacune de ces représentations doit être **lisible indépendamment** de son éventuel texte d'accompagnement,
- elles doivent donc toujours être dotées d'un **titre informatif**, c'est-à-dire donnant suffisamment d'informations sur la population, le lieu et la période d'étude
- les acronymes doivent être définis (par exemple, en note sous le tableau ou le graphique, ou dans le titre),
- si seuls des pourcentages sont présentés (sans les effectifs correspondants), il faut préciser l'effectif total à partir duquel ils ont été calculés dans le titre
- les unités de mesure doivent systématiquement être indiquées pour les variables quantitatives (exemple : années pour la variable âge). Elles doivent figurer une seule fois à côté du nom de la variable. C'est également le cas pour le caractère % en ce qui concerne les variables catégorielles.

Enfin, dans le document rapportant les résultats d'une étude, il faudra savoir présenter les principaux résultats dans le corps du document, et les résultats secondaires en annexe.

II Tableau

Un **tableau** est une représentation des données, utilisable quelle que soit la nature de la variable à représenter, quantitative ou catégorielle. La construction d'un tableau permet de disposer de l'ensemble des données. La présentation des données sous forme de tableau est particulièrement indiquée pour des données répétées et précises. Comme il s'agit souvent de nombreux chiffres, il est important de simplifier le plus possible la présentation. Le tableau le plus simple a deux colonnes. Dans la première colonne, figure la liste des catégories d'une variable catégorielle ou des regroupements d'une variable quantitative. Dans la seconde colonne, figurent les effectifs correspondant à chacune de ces catégories.

> Par exemple, dans une enquête étudiant la relation entre tabac et cancer du poumon, on décrit les différentes catégories de consommateurs de tabac : jamais fumeur / ex-fumeur / tabac blond / tabac brun / tabac mixte. Le tableau OUTILS-STAT-1 permet de donner le nombre de sujets pour chaque catégorie de la variable « consommation de tabac » (variable catégorielle).

Tableau OUTILS-STAT-1. Répartition des individus selon le type de consommation de tabac. Étude de Fictif et al., 1988.

Consommation de tabac	Nombre de sujets
Jamais	500
Ex-fumeur	100
Tabac blond	200
Tabac brun	100
Tabac mixte	100
Total	1000

La construction d'un tableau obéit à quelques règles générales :

1) il existe toujours un bandeau de titre pour indiquer la nature des informations figurant dans les colonnes, ce bandeau a un trait horizontal au-dessus et au-dessous, la tête de colonne permet d'indiquer la nature de la variable figurant dans cette colonne,

2) un trait horizontal figure au-dessous de la dernière ligne,
3) en dehors de ces traits permettant de souligner les bandeaux, aucun autre trait n'est utile, en particulier aucun trait vertical,
4) les chiffres sont alignés par colonne : sur le dernier chiffre de droite (s'il s'agit d'entiers) ou sur la virgule (s'ils sont exprimés avec une décimale),
5) pour une même variable, le même nombre de chiffres après la décimale est employé ; en français, le séparateur décimal est la virgule, dans le système anglo-saxon, c'est le point qui est employé,

   Par exemple, on ne présenterait pas une proportion de 25,2% pour une catégorie et 34% pour une autre. On choisirait 25% et 34% ou 25,2% et 34,0%,
6) les totaux, s'il y a lieu, doivent être donnés,
7) le séparateur des milliers est un espace (et non un point comme dans le système anglo-saxon) ; on peut également ne pas marquer la séparation.

Enfin, par convention, le titre d'un tableau figure au-dessus du tableau.

Le tableau OUTILS-STAT-2 permet de représenter les données d'âge (variable quantitative) de 120 femmes venues en consultation dans un centre de dépistage du cancer du sein. Ce tableau illustre également l'application des règles simples de construction d'un tableau. La proportion correspond au rapport de l'effectif sur le total.

Tableau OUTILS-STAT-2. Distribution selon l'âge de 120 femmes ayant consulté dans le centre de Maville entre octobre et décembre 2000. Représentation par un tableau.

Age (en années)	Effectif	Proportion en %
43	1	0,8
44	1	0,8
45	4	3,3
47	2	1,7
52	4	3,3
53	3	2,5
54	5	4,2
56	6	5,0
57	4	3,3
58	4	3,3
59	8	6,7
60	8	6,7
61	12	10,0
62	8	6,7
63	8	6,7
64	8	6,7
65	4	3,3
66	6	5,0
67	6	5,0
68	4	3,3
70	4	3,3
72	3	2,5
73	3	2,5
76	2	1,7
78	2	1,7
Total	120	100,0

On pourra également effectuer des regroupements permettant de présenter les résultats de manière plus synthétique (Tableau OUTILS-STAT-3).

Tableau OUTILS-STAT-3. Répartition par catégories d'âge de 120 femmes ayant consulté dans le centre de Maville, octobre-décembre 2000.

Age (en années)	Effectif	Proportion en %	Fréquence cumulée en %
40-44	2	1,7	1,7
45-49	6	5,0	6,7
50-54	12	10,0	16,7
55-59	22	18,3	35,0
60-64	44	36,7	71,7
65-69	20	16,7	88,4
70-74	10	8,3	96,7
75-79	4	3,3	100,0
Total	120	100,0	

Ces données permettent de calculer très facilement la fréquence cumulée à la borne supérieure d'une catégorie, résultat qui vous sera souvent utile. La **fréquence cumulée** est la proportion d'observations dans toutes les catégories précédentes ajoutée à celle de la catégorie présente. Ces catégories représentent toutes les données (l'exhaustivité des données) et sont mutuellement exclusives.

Par exemple, pour trouver la fréquence cumulée à 49 ans, on va additionner la proportion dans la catégorie « 40-44 » et celle dans la catégorie « 45-49 ». L'écriture « 40-44 » sous-entend que toutes les valeurs entre 40 et 44 ans sont incluses, ou encore de 40 ans exactement à 44,99 ans (la valeur 45 ans exactement étant exclue). D'après le tableau OUTILS-STAT-3, la fréquence cumulée à 49 ans est de 6,7% (1,7% + 5,0%) : 6,7% des femmes avaient moins de 50 ans.

L'examen du tableau OUTILS-STAT-3 permet d'apporter une réponse descriptive à la question initiale : les femmes de 60 ans et plus sont plus nombreuses que celles de moins de 60 ans. La catégorie d'âge la plus représentée est celle des femmes ayant entre 60 et 64 ans.

III Graphique

Le **graphique** permet essentiellement de visualiser un phénomène remarquable : contrastes ou tendances. L'œil doit pouvoir observer les changements des valeurs en ordonnée (échelle verticale) pour un changement d'unité, de classe ou de modalité en abscisse (échelle horizontale). Le choix entre les graphiques possibles repose essentiellement sur le type et le nombre de variables à représenter. Par convention, le titre d'un graphique figure au-dessous du graphique.

Dans ce paragraphe, vous verrez les principes de construction de graphiques simples.

1 Variable quantitative continue

Pour une **variable quantitative continue**, si les bornes (valeur minimale et maximale) sont connues, le choix se porte sur l'histogramme ou le polygone de fréquence.

L'**histogramme** est un graphique où l'axe des abscisses représente les valeurs de la variable, regroupées en classes, et l'ordonnée représente l'effectif ou la fréquence de chacune des classes.

Par exemple, un histogramme permet de représenter la distribution de l'âge des femmes venues en consultation au centre de dépistage (Figure OUTILS-STAT-1).

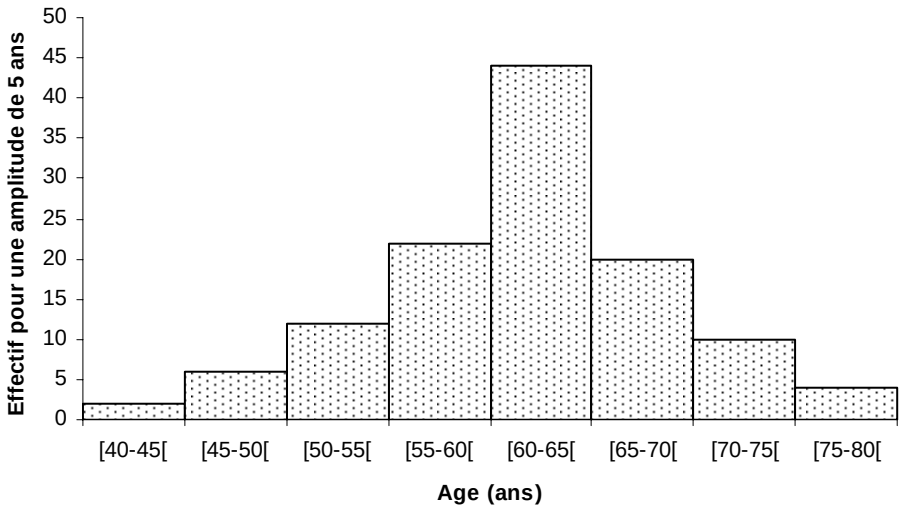

Figure OUTILS-STAT-1. Distribution selon l'âge des 120 femmes ayant consulté dans le centre de Maville entre octobre et décembre 2000. Représentation par un histogramme.

L'examen de cet histogramme permet d'apporter une réponse descriptive plus visuelle qu'un tableau à la question initiale : les femmes de 60 ans et plus sont plus nombreuses que celles ayant moins de 60 ans.

Par ailleurs, un graphique donne une bonne idée de la dispersion des valeurs autour de la catégorie d'âge la plus représentée. En revanche, il est plus difficile de connaître le nombre exact d'individus dans chaque intervalle.

L'**intégrité statistique** des représentations graphiques des variables quantitatives repose sur un certain nombre de règles :

1) le choix de l'échelle doit être correct pour ne pas exagérer ou sous-estimer un changement et, par conséquent, donner une impression fausse des données. La distance entre les marques de graduation doit donc être proportionnelle à l'amplitude réelle des catégories.

Dans l'exemple, un intervalle d'amplitude 5 ans a été choisi et la distance sur l'axe des abscisses est exactement la même entre 40 et 44 ans, entre 45 et 49 ans et ainsi de suite.

2) l'axe des abscisses doit couvrir toute l'étendue des données possibles.

Dans l'exemple, la plus petite valeur est 43 ans et la plus grande est 76 ans. Les extrémités du graphique vont donc de 40 à 80 ans.

3) la variable étant continue, il ne peut y avoir d'espace entre la base des différents rectangles en abscisse.

Dans l'exemple, chaque rectangle est contigu au rectangle précédent et au suivant.

Il y a une exception à cette règle si une catégorie a un effectif nul, car, dans ce cas, aucun rectangle n'occupe l'intervalle.

Une astuce si vous travaillez sur MS-Excel™ : par défaut, ce logiciel laisse un espace entre les différents rectangles. Il faut alors double-cliquer sur l'un des rectangles. La fenêtre « Format de série de données » apparaît. Il faut cliquer sur l'onglet « Options » et choisir une largeur d'intervalle égale à 0.

Les graphiques **tridimensionnels** ne respectent pas les règles d'intégrité statistique à cause de la distorsion liée à la perspective : ils sont donc déconseillés.

Par exemple, dans la figure OUTILS-STAT-2, la perspective choisie distord la tendance et donne l'impression que les femmes âgées de moins de 60 ans sont plus nombreuses que les autres.

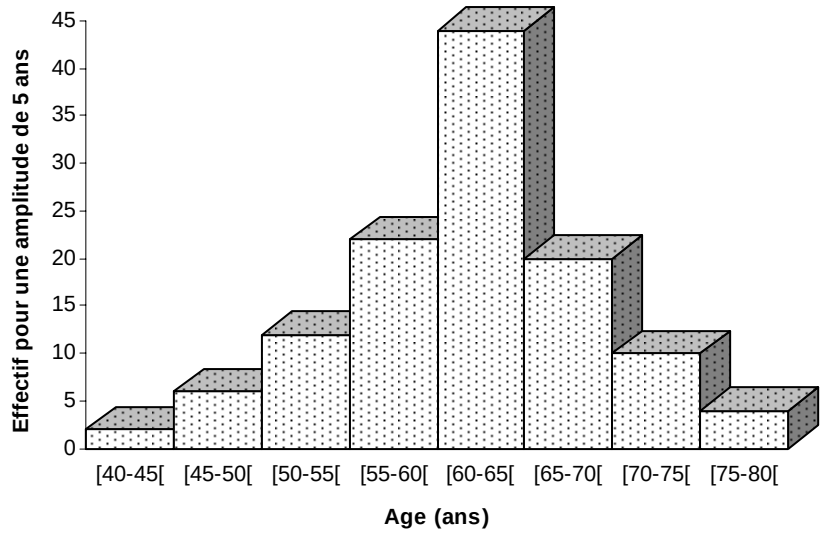

Figure OUTILS-STAT-2. Distribution selon l'âge des 120 femmes ayant consulté dans le centre de Maville entre octobre et décembre 2000. Représentation par un histogramme en 3 dimensions d'une variable à une dimension. Exemple de distorsion liée à la perspective.

Le **polygone de fréquence** est la courbe qui joint les milieux des sommets des rectangles de l'histogramme. Le terme est général, car on peut faire la représentation de chaque catégorie d'âge soit en fonction de son effectif, soit en fonction de sa fréquence.

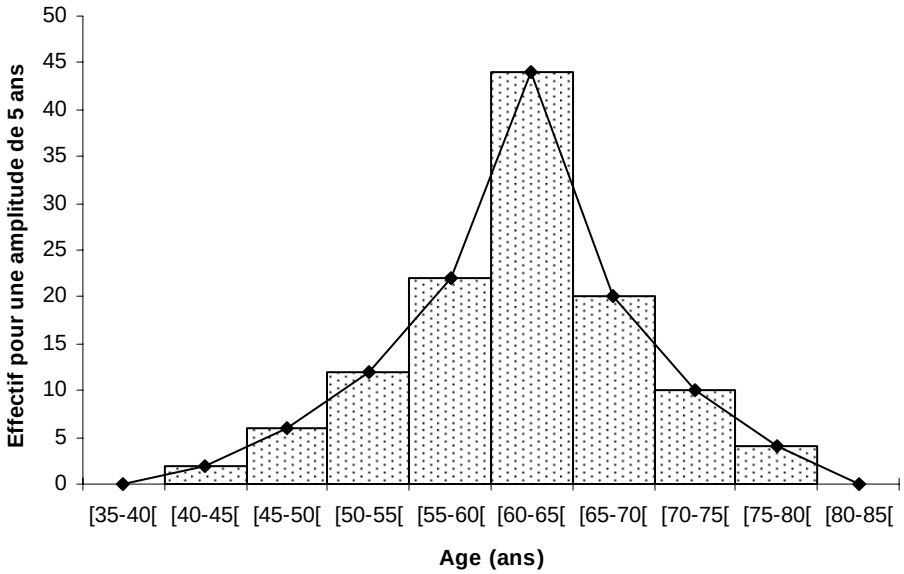

Figure OUTILS-STAT-3. Distribution de l'âge des 120 femmes ayant consulté dans le centre de Maville entre octobre et décembre 2000. Représentation par un polygone de fréquence.

Le **polygone de fréquence cumulée** est la courbe qui joint les valeurs des fréquences cumulées de chaque classe.

Un exemple de calcul des fréquences cumulées figure dans le tableau OUTILS-STAT-3 Le tracé correspondant à ces données est celui de la figure OUTILS-STAT-4.

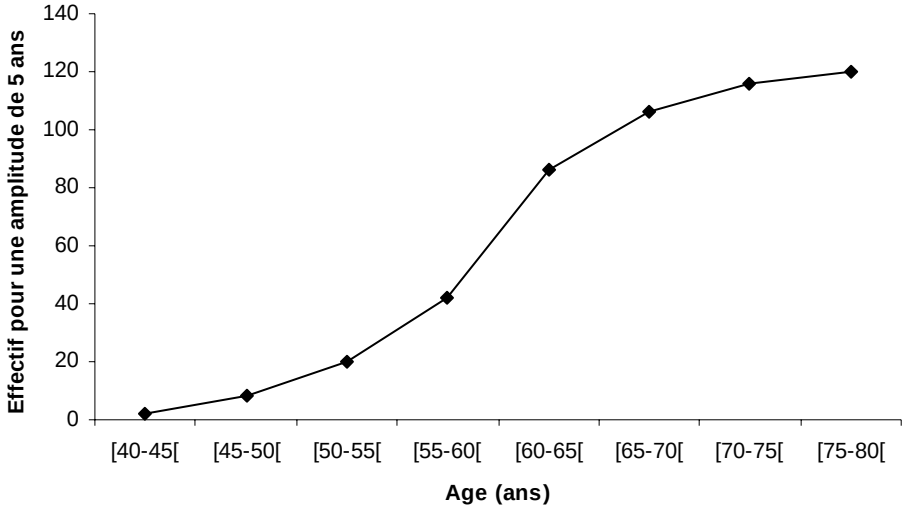

Figure OUTILS-STAT-4. Distribution de l'âge des 120 femmes ayant consulté dans le centre de Maville, octobre-décembre 2000. Représentation par un polygone de fréquence cumulée.

En pratique, si l'on ne dispose pas de logiciel adapté, ces graphiques peuvent être tracés à la main sur du papier à carreaux ou du papier millimétré. Si l'on dispose d'un tableur, MS-Excel™ par exemple, il est facile de trouver ce type de représentation ; il faut veiller néanmoins au respect des règles d'intégrité statistique, en particulier la règle n°3. Enfin, tous les logiciels statistiques couramment utilisés ont un module graphique permettant de réaliser ce type de graphique.

2 Variable catégorielle

On présente le nombre ou la proportion de sujets dans chaque catégorie. Le graphique permettant cette représentation est le **diagramme à barres**. Chaque barre a la même largeur, et contrairement à l'histogramme, un espace est laissé entre chaque barre.

Si l'on reprend l'exemple du tabac (Tableau OUTILS-STAT-1), on obtient la figure OUTILS-STAT-5.

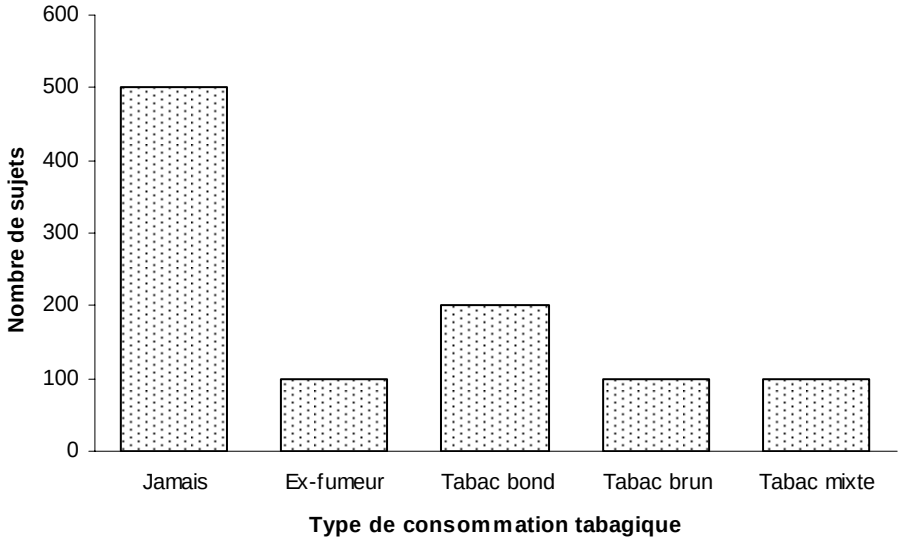

Figure OUTILS-STAT-5. Description de la consommation de tabac à l'inclusion dans une étude du risque de cancer du poumon chez 1000 sujets. Représentation à l'aide d'un diagramme à barres.

Le **graphique en secteurs** ou « **camembert** » illustre également la répartition d'une variable catégorielle. Il nous semble néanmoins inutile pour deux raisons : 1) il ne respecte pas les règles d'intégrité statistique car une surface n'est pas adaptée pour représenter une seule dimension et le risque de distorsion optique est d'autant plus important que l'on emploie plusieurs couleurs ou hachures ; 2) il est le plus souvent peu informatif en comparaison à un tableau.

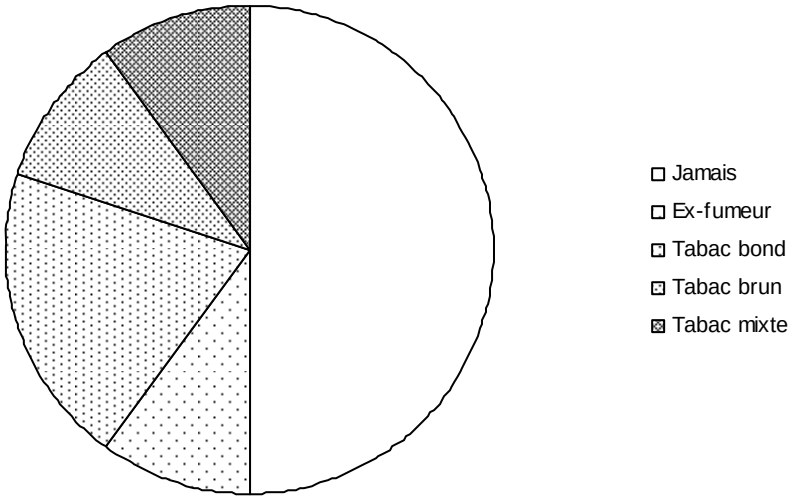

Figure OUTILS-STAT-6. Description de la consommation de tabac à l'inclusion dans une étude du risque de cancer du poumon chez 1000 sujets. Représentation à l'aide d'un diagramme en secteurs en 2 dimensions d'une information à une dimension.

Paramètres décrivant une distribution

I Introduction

Vous avez vu que la plupart des caractéristiques permettant de décrire les êtres vivants n'ont pas une valeur unique. L'ensemble des valeurs observées sur un échantillon pour une caractéristique est sa **distribution observée**. Pour donner un sens aux données dont il dispose, l'épidémiologiste va devoir les résumer. Il va utiliser pour cela des **paramètres**, qui sont des fonctions des observations : les « **paramètres de tendance centrale** », comme la moyenne ou la médiane ; ou les « **paramètres de dispersion** », comme la variance, l'étendue et les percentiles. Nous allons aborder la définition de ces termes et l'usage de ces paramètres dans ce nouveau chapitre.

II Paramètres de tendance centrale

1 Variable quantitative

Pour une variable quantitative, une façon simple de résumer les valeurs obtenues sur un échantillon est d'utiliser la **moyenne arithmétique**, appelée plus couramment **moyenne**. La moyenne est obtenue en faisant la somme des valeurs, puis en divisant cette somme par le nombre de valeurs, noté ici *n*.

> Par exemple, si vous recueillez l'âge en années d'une population de cinq femmes qui viennent d'accoucher de leur premier enfant : 24, 17, 35, 37, 32. La somme est : 145 ans, et comme il y a 5 valeurs, la moyenne est : 145/5=29 ans. L'âge moyen des femmes à l'accouchement de leur premier enfant est donc, pour la série de valeurs mesurées au sein de cet échantillon, de 29 ans.

Chaque valeur est notée x_i.

> Dans l'exemple, on a donc : $x_1 = 24$, $x_2 = 17$, $x_3 = 35$, $x_4 = 37$, $x_5 = 32$.

La somme est notée $\sum_{i=1}^{n} x_i$ (c'est-à-dire : somme de toutes les valeurs de la première à la dernière) et la moyenne, μ, est donc : $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$.

Une autre façon de résumer les valeurs est d'utiliser la **médiane**. La médiane est la valeur centrale de la distribution, qui divise l'échantillon en deux moitiés de taille égale. Pour trouver la médiane, il faut d'abord classer toutes les observations par ordre croissant.

- Si le nombre d'observations est impair, la médiane est la valeur correspondant à l'observation située au milieu, celle située au $\frac{(n+1)}{2}$ ème rang.

  > Pour notre série de 5 observations d'âge, après avoir ordonné les observations de façon croissante, la série s'écrit : 17, 24, 32, 35, 37 et l'on voit facilement que la médiane est égale à 32 ans. La médiane correspond bien à la valeur de la $3^{ème}$ observation, car : (5+1)/2 = 3.

- Si *n* est un nombre pair, on considère que la médiane est à mi-chemin entre les deux valeurs du milieu de la distribution (puisque l'on cherche la médiane, synonyme de milieu).

Par exemple, pour une série de 8 observations d'âge : 17, 24, 27, 27, 29, 32, 35, 37, la médiane se situe entre la valeur de la 4ème observation (27 ans) et celle de la 5ème observation (29 ans), car (8+1)/2 = 4,5. La médiane vaut donc : (27 + 29)/2 = 28 ans.

Enfin, le **mode** est, par définition, la valeur la plus représentée de la série. Une série peut ne pas avoir de mode ou au contraire avoir plusieurs modes.

Par exemple, pour la série des 8 observations d'âge, le mode est 27 ans, car cette valeur apparaît deux fois, alors que les autres valeurs n'apparaissent qu'une seule fois.

La 1ère série des 5 valeurs d'âge (17, 24, 32, 35, 37) n'a pas de mode.

Dans la série suivante : 17, 24, 27, 27, 29, 29, 32, 35, 37, il existe deux modes : 27 ans et 29 ans, valeurs qui apparaissent deux fois, alors que toutes les autres n'apparaissent qu'une seule fois. Dans ce cas, on parle de **distribution bimodale**.

2 Variable catégorielle ou quantitative discrétisée

Pour une variable catégorielle, qu'elle soit dichotomique, nominale ou ordinale, on présente la **proportion** des sujets dans les différentes catégories.

Par exemple, parmi 100 malades atteints d'arthrite rhumatoïde, on observe 76 femmes et 24 hommes. La proportion de femmes est le nombre de femmes rapporté au nombre total de sujets, soit 76%. La proportion d'hommes est de 24%.

Cette description selon les proportions peut également être utilisée pour une **variable quantitative** dont les valeurs ont été discrétisées ou bien regroupées.

Par exemple, si l'on reprend la série de femmes ayant consulté dans un centre de dépistage du cancer du sein, la proportion de femmes ayant plus de 60 ans est 70/120, soit 58,3%.

La **fréquence cumulée** à la borne supérieure de classe, que vous avez déjà vue lors du chapitre sur la représentation des données, est la proportion des observations dans toutes les classes précédentes ajoutée à celle de la classe présente. Elle est utile pour une variable quantitative ayant fait l'objet d'une discrétisation (Tableau OUTILS-STAT-3) ; elle est également utile pour une variable catégorielle ordinale.

III Paramètres de dispersion

Les paramètres qui viennent d'être cités sont tous des paramètres qui résument la tendance centrale des observations, mais ne donnent pas une idée de leur **dispersion**. Or, la dispersion des valeurs est importante à prendre en compte dans l'interprétation des résultats et les décisions qui en découlent.

Par exemple, si l'on s'intéresse à la durée d'incubation d'une infection (délai entre la date d'exposition à l'agent infectieux et la date du diagnostic), calculer que la moyenne vaut 13 jours n'apporte pas une information suffisante pour envisager les mesures pertinentes d'observation ou d'isolement à prendre pour les sujets qui sont exposés à l'agent infectieux. En effet, on ne recommanderait pas de garder les patients seulement 13 jours en observation ou en isolement parce que la moyenne des observations est de 13 jours. En fait, on a également besoin de savoir combien de personnes développent la maladie au 14ème jour, au 15ème jour, etc.

Afin d'apprécier la **distribution** observée d'une variable quantitative autour de la moyenne ou de la médiane, on peut simplement repérer la plus petite (**minimum**) et la plus grande valeur (**maximum**) de la distribution : il s'agit de l'**étendue** des observations. On peut également entendre par étendue la différence entre valeurs minimum et maximum. Si l'on présente la différence, il est recommandé de donner également la valeur minimum ou maximum.

Par exemple, dans la série suivante de valeurs d'âge (en années) : 17, 24, 27, 27, 29, 29, 32, 35, 37, le minimum est 17 ans et le maximum est 37 ans. L'étendue est 17-37 ans. On peut aussi considérer que l'étendue est de 20 ans.

Néanmoins, l'étendue est souvent insuffisante pour résumer la dispersion d'une distribution, car les valeurs extrêmes sont assez particulières. Les **quantiles** sont les valeurs d'une distribution définies par la proportion de sujets qui se trouvent au-dessous et au-dessus de cette valeur. On parle de quartiles, déciles, percentiles.

Les **quartiles** sont les trois valeurs qui partagent la distribution en quatre parties égales. Le **premier quartile** correspond à la valeur de l'observation qui a 25% de la distribution au-dessous et 75% au-dessus, le second quartile est donc … la médiane, et le **troisième quartile** correspond à la valeur de l'observation qui a 75% de la distribution au-dessous et 25% au-dessus.

Par exemple, la durée de survie de 42 patients atteints de cancer digestif a été recueillie lors d'une consultation de suivi de gastro-entérologie. La série de valeurs est ordonnée de manière croissante (Tableau OUTILS-STAT-4).

Tableau OUTILS-STAT-4. Durée de la survie (en mois) de 42 patients atteints de cancer digestif. Service de gastro-entérologie de l'hôpital de Maville, 2000.

Survie (mois)	Survie (mois)	Survie (mois)	Survie (mois)	Survie (mois)	Survie (mois)
1	15	30	40	52	62
3	16	32	41	54	64
3	17	33	41	54	72
5	23	34	42	58	74
8	24	36	44	58	74
10	28	36	47	59	85
12	28	38	49	60	90

Il y a au total 42 observations et la médiane correspond à la valeur située entre le rang 21 et le rang 22, car : (42+1)/2 = 21,5. Comme la durée de survie est respectivement de 38 et 40 mois à ces deux rangs, la médiane vaut (38 + 40)/2 = 39 mois.

Pour trouver la valeur de l'observation correspondant aux 1er et 3ème quartiles, on peut procéder avec la même méthode.

Le rang du 1er quartile est : (n+1)/4. Dans l'exemple, on trouve (42+1)/4 = 10,75 et il s'agit donc d'une valeur située entre la valeur classée au 10ème rang (17 mois) et celle classée au 11ème rang (23 mois). Si l'on utilise le même principe de calcul que pour la médiane, le 1er quartile vaut (17+23)/2 = 20 mois (ou de façon plus précise : 17 + 0,75x(23-17) = 21,5 mois).

Le rang du 3ème quartile est (n+1) x (3/4). Dans l'exemple, 43 x (3/4) vaut 32,25. Or, la valeur au 32ème rang et la valeur au 33ème rang valent toutes deux 58 mois. Le 3ème quartile de cette distribution est donc 58 mois.

Plutôt que les quartiles, on présente souvent l'**étendue inter-quartiles** (25% à 75%) qui est donc la partie centrale qui couvre 50% de la distribution observée.

Dans l'exemple, l'étendue inter-quartiles est 20-58 mois.

On peut raisonner de la même manière avec les quintiles, les **déciles** ou les **centiles** (**percentiles**), partageant la distribution en 10 ou 100 parties égales, respectivement. On peut ainsi calculer la valeur correspondant au $5^{ème}$ percentile et au $95^{ème}$ percentile et obtenir l'étendue centrale couvrant 90% de la distribution observée.

Une autre façon de mesurer la variabilité consiste à calculer la **variance**, σ^2, qui est une mesure des distances de chaque individu à la moyenne :

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$$

La variance a l'unité de la variable au carré.

Par exemple, pour la variance de l'âge, il peut s'agir d'années au carré (années²) ou de jours au carré (jours²).

Pour exprimer la variabilité dans la même unité que les valeurs observées, on en prend la racine carrée, qui s'appelle l'**écart-type** (ou **écart-type inter-individuel**) :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$$

On utilise plus souvent l'écriture suivante, plus commode à utiliser pour les calculs manuels :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i^2\right)-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}}{n}}$$

Par exemple, pour les cinq valeurs d'âge de l'exemple initial de ce chapitre, on trouve :

Tableau OUTILS-STAT-5. Exemple de décomposition du calcul pour la variance et l'écart-type. Série de 5 valeurs d'âge.

Age (en années), x_i	Age au carré (en années au carré), x_i^2
17	289
24	576
32	1024
35	1225
37	1369
Total : $\sum_{i=1}^{n} x_i = 145$	$\sum_{i=1}^{n}(x_i)^2 = 4483$

$$\sigma = \sqrt{\frac{4483 - 145^2/5}{5}} = 7{,}5 ans$$

Variance et écart-type sont très intéressants à titre descriptif car ils permettent d'apprécier à quel point la distribution est dispersée. **Plus la variance et l'écart-type sont grands, plus la dispersion est grande** (pour une même variable).

Lois en statistique

I Introduction

Comme il est habituellement impossible d'étudier la **population** entière, on dispose le plus souvent de données sur un **échantillon** d'individus. On utilise alors les informations obtenues sur cet échantillon pour en tirer des conclusions sur l'ensemble de la **population** que cet échantillon est supposé représenter. L'échantillon est considéré **représentatif** de cette population s'il n'a pas fait l'objet d'une sélection particulière. La méthode idéale pour constituer un échantillon représentatif d'une population est le **tirage au sort**.
Dans ce chapitre, vous verrez comment les **probabilités** et les **lois de probabilité** contribuent à utiliser les informations obtenues à partir d'un échantillon pour appréhender la population.
L'exemple de la loi Normale permettra d'illustrer ces notions.

II Rappels sur les probabilités

La **probabilité d'un événement** est la proportion de fois où cet événement se produit si on répète à l'infini les conditions où il peut se produire.
Par définition, la valeur d'une probabilité est comprise entre 0 et 1 (ou 0% et 100%). Un événement impossible, qui ne peut se produire, a une probabilité de 0. Un événement certain, qui se produit toujours, a une probabilité de 1.

III Loi Normale et son usage

Dans l'échantillon dont nous disposons, nous savons décrire la **distribution observée** d'une variable quantitative continue. Si l'on souhaite utiliser ces informations pour en déduire ce qui se passe dans la population dont cet échantillon est issu et représentatif, il faut faire l'hypothèse que la variable suit, dans la population, une **distribution théorique** ou **loi de probabilité**. Cette loi de probabilité est spécifiée mathématiquement. Dans cette écriture mathématique, la loi dépend de paramètres. La moyenne, notée μ, et l'écart-type, noté σ, sont, par exemple, les paramètres de la distribution théorique la plus utilisée, la **loi Normale** (ou loi de Gauss ; les majuscules N et G sont intentionnelles). On note $N(\mu,\sigma)$. On utilise des lettres grecques pour désigner la moyenne et l'écart-type de la population ou « théoriques ». On appelle ces paramètres théoriques car ils ne sont pas connus le plus souvent (on ne connaît que les valeurs observées dans l'échantillon). L'importance de la loi Normale est considérable dans le domaine du vivant car de nombreuses variables aléatoires suivent cette loi, en théorie.
La loi Normale a la forme d'une **courbe en cloche** comme on peut le voir sur la figure OUTILS-STAT-7. Elle tend à avoir un pic : on la dit **unimodale** (le pic correspond au mode). Le pic est obtenu autour de la valeur moyenne de la variable qui est aussi la valeur médiane. De plus, la distribution est **symétrique** autour de ce pic.

Toute distribution Normale peut être transformée en une seule distribution ayant pour moyenne 0, et pour écart-type 1 : la **distribution Normale, centrée et réduite**, notée : **N(0,1)**.
Comment faire ? On soustrait à chaque valeur d'une distribution Normale quelconque, la moyenne, μ, et l'on divise par l'écart-type, σ. Cela revient à écrire :

$$\frac{x_i - \mu}{\sigma}$$

On appelle cette quantité : l'**écart-réduit**. La distribution de l'écart-réduit suit une loi N(0,1).

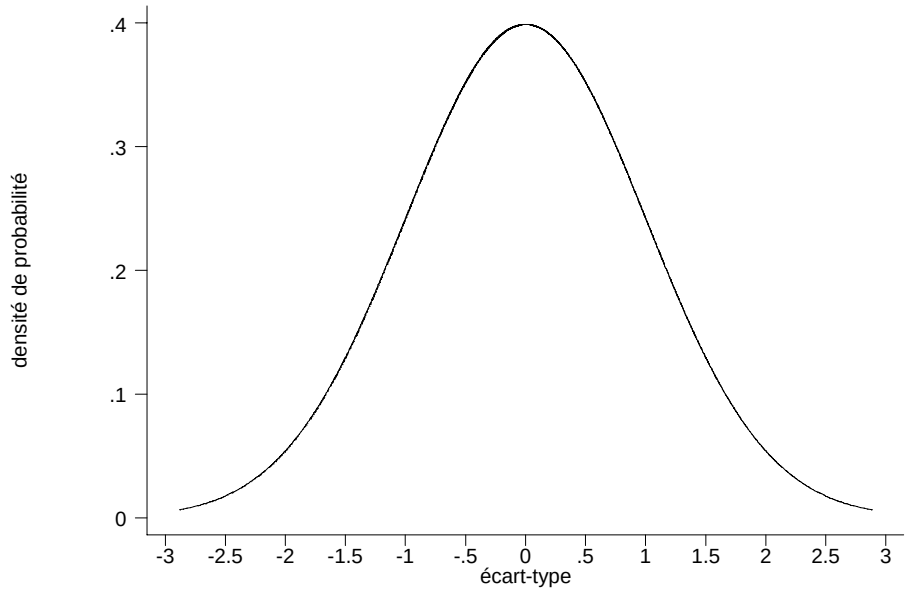

Figure OUTILS-STAT-7. La distribution Normale de moyenne 0 et d'écart-type 1.

L'intérêt de cette transformation réside dans les propriétés très intéressantes de cette distribution.
Avant de pouvoir explorer ces propriétés, nous avons besoin d'un tableau appelé « **Table de la loi normale centrée réduite** » ou « **Table de l'écart-réduit** ». Vous pouvez le consulter sur le site ou l'imprimer. Dans ce tableau, l'écart-réduit est appelé U_α. Comment l'interpréter ?

- à l'intérieur du tableau, figurent les valeurs d'une distribution Normale centrée réduite (les valeurs dites de U_α, en valeur absolue)
- et dans les bandeaux à gauche et au-dessus, figurent les valeurs des probabilités correspondantes α.

On trouve la probabilité α d'observer des valeurs comprises entre $]-\infty$ à $-U_\alpha[$ et $]+U_\alpha$ à $+\infty[$ en additionnant la valeur correspondante de la ligne sur le bandeau à gauche et la valeur correspondante de la colonne sur le bandeau de titre.

Voici une illustration à partir de données concrètes. Dans la population des hommes de 35-40 ans, la concentration moyenne de cholestérol total dans le sang est 1,84 g/l et l'écart-type 0,40 g/l. On fait l'hypothèse que la concentration de cholestérol total dans le sang a une distribution Normale.

Question : quelle est la probabilité d'observer une valeur de la concentration de cholestérol total > 2,50 g/l dans le sang si μ = 1,84 g/l et σ = 0,4 g/l ?

Réponse : selon l'hypothèse d'une distribution Normale, on peut calculer de combien d'écarts-types, la valeur 2,50 g/l est éloignée de la moyenne :

$$U_{\alpha} = \frac{2{,}50 - 1{,}84}{0{,}4} = 1{,}65$$

La valeur 2,50 g/l est éloignée de la moyenne de 1,65 écarts-types. Dans la table de la loi normale centrée réduite, on trouve que la probabilité d'être en dehors de 1,65 écarts-types (au-dessous ou au-dessus) est $\alpha = 0,10$. On en déduit que la probabilité de se situer au-dessus de 1,65 écarts-types est donc $\alpha/2 = 0,10/2 = 0,05$. Autrement dit, 5% de la population a, en théorie, une concentration de cholestérol total dans le sang au-dessus de cette valeur de 2,50 g/l (Figure OUTILS-STAT-8).

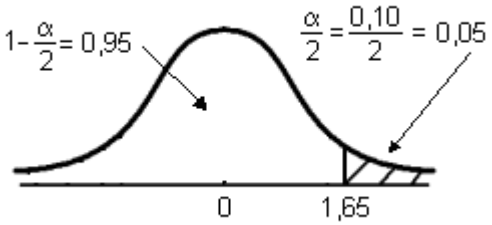

Figure OUTILS-STAT-8. Recherche de la probabilité d'observer une concentration de cholestérol total dans le sang > 2,50 g/l, si $\mu = 1,84$ g/l et $\sigma = 0,4$ g/l, dans le cas d'une distribution normale.

Principes de l'analyse des données en statistique

I Introduction

Le plus souvent, on dispose de données sur un échantillon et on utilise ces informations pour décrire la population dont cet échantillon est issu. **L'analyse se déroule en deux étapes** :

1) **la 1ère étape consiste à donner une description** résumée de la distribution de la variable.

Par exemple, à la suite d'une grande enquête sur un échantillon représentatif, on donne une estimation de la proportion de diabétiques ou une estimation de la consommation moyenne d'alcool en grammes par semaine dans la population ;

2) **la 2nde étape consiste à comparer** formellement la distribution observée d'une variable dans un échantillon par rapport à sa distribution attendue dans la population ou à comparer la distribution observée dans plusieurs groupes.

Par exemple, dans un essai clinique, on compare la fréquence des récidives de mélanome après la mise en route de deux traitements différents ou, dans une enquête visant à étudier l'association entre tabac et cancer du poumon, on compare la proportion de sujets exposés au tabac selon que les sujets sont atteints de cancer du poumon (cas) ou non (témoins).

Dans ce chapitre, vous verrez les grands principes de ces deux étapes.

II Étape de description : estimation

1 Introduction

On souhaite utiliser les résultats obtenus dans un échantillon pour estimer la vraie valeur dans la population.

Par exemple, dans un échantillon, on trouve que la pression artérielle systolique moyenne est 140 mmHg. Quelle information sur la vraie valeur dans la population ce résultat apporte-t-il ?

Faisons l'hypothèse que l'échantillon est **représentatif** de la population, c'est-à-dire constitué sans biais, au mieux par **tirage au sort**. Si l'on tirait au sort des échantillons successifs à

partir d'une même population, chaque échantillon fournirait une estimation ponctuelle du paramètre d'intérêt (par exemple, ici, la moyenne). On peut comprendre intuitivement que cette estimation va **varier d'un échantillon à l'autre**, en suivant une loi de probabilité. Cela nous aide à comprendre pourquoi au sein d'un échantillon représentatif, la moyenne, m, d'une variable, peut différer de la moyenne de la population, μ, uniquement du fait du hasard. Cette fluctuation est importante à prendre en compte, même si l'on s'attend à ce que m soit très proche de μ, puisque l'échantillon est représentatif de la population. Il est donc nécessaire d'apprécier l'**incertitude** associée à notre **estimation**, grâce à un **intervalle de confiance**. Dans ce chapitre, vous aborderez le cas des petits et des grands échantillons pour une variable quantitative, et seulement le cas des grands échantillons pour les variables catégorielles.

2 Fluctuations d'échantillonnage : exemple d'une moyenne, variable quantitative

Une façon de procéder est de supposer que l'échantillon dont nous disposons n'est en fait qu'un des multiples échantillons d'une taille donnée, *n*, représentatifs de la population, qu'il est possible de constituer. Comment varient alors les moyennes observées dans les échantillons, m_1, m_2, m_3, etc. par rapport à la moyenne de la population, μ ? Intuitivement, on peut penser que la moyenne de chaque échantillon varie autour de μ, comme les valeurs individuelles d'un échantillon varient autour de la moyenne, selon un certain écart, et selon une loi de probabilité connue.

De plus, on admettra que la **moyenne** obtenue dans chaque échantillon varie autour de μ selon les propriétés suivantes :

1) la variabilité autour de μ est moins importante avec les échantillons de grande taille qu'avec ceux de petite taille,
2) la variabilité autour de μ est moins importante que la variabilité des observations individuelles dans la population,
3) la variabilité autour de μ est proportionnelle à la variabilité des observations individuelles (écart-type).

Ces propriétés permettent de montrer mathématiquement (dans cet enseignement, vous admettrez les résultats) que :

1) en moyenne, la valeur observée dans un échantillon, m_i, est une bonne estimation de la moyenne de la population, μ. Cette estimation sera désormais notée $\hat{\mu}$, le chapeau signifiant qu'il s'agit d'une estimation de la vraie valeur.
2) la distribution des moyennes est Normale si la distribution des données individuelles est Normale. Quelle que soit la distribution des données individuelles, la distribution des moyennes suit une loi approximativement Normale, si les échantillons sont de taille suffisante, c'est-à-dire $n \geq 30$.
3) l'écart-type des moyennes de plusieurs échantillons est égal à $\sqrt{\frac{\sigma^2}{n}}$. On appelle ce rapport l'**erreur-type** pour le distinguer de l'écart-type inter-individuel (ou « écart-type »).

4) la variance de l'échantillon est une bonne estimation de la variance au sein de la population (σ^2) si l'on utilise n-1 au dénominateur, et non n. Dans ce cas, on utilise la notation $\hat{\sigma}^2$: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - m)^2}{n-1}$ ou $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$.

3 Intervalle de confiance d'une moyenne, grands échantillons

L'erreur-type est utile pour construire l'**intervalle de confiance de la moyenne**.

On peut remplacer σ par l'estimation de l'écart-type de la population que l'on peut obtenir à partir de l'échantillon. Si $n \geq 30$, on considérera, avec un risque d'erreur de 5%, que la moyenne inconnue, μ, doit être comprise entre : $\hat{\mu} - 1{,}96 \times \sqrt{\frac{\hat{\sigma}^2}{n}}$ et $\hat{\mu} + 1{,}96 \times \sqrt{\frac{\hat{\sigma}^2}{n}}$: c'est l'intervalle de confiance à 95% de la moyenne, ou encore : $IC_{95\%}(\mu) = \left[\hat{\mu} \pm 1{,}96 \times \sqrt{\frac{\hat{\sigma}^2}{n}}\right]$.

> Par exemple, dans un échantillon de 100 hommes jeunes, la moyenne de la concentration sanguine de cholestérol total sanguin est 1,84 g/l et l'écart-type $\hat{\sigma}$ est 0,4 g/l. L'intervalle de confiance à 95% de la moyenne est donc égal à : 1,84 ± 1,96 × 0,04 = [1,76-1,92] g/l.

L'intervalle de confiance à 95% est le plus couramment employé, mais il arrive que l'on construise un intervalle à d'autres niveaux de confiance. Pour calculer un intervalle de confiance à 99%, la seule différence sera U_α, qui, dans ce cas, vaut : 2,5758 (ou 2,58).

> Pour l'exemple, on trouve un intervalle de confiance à 99% de la moyenne égal à : 1,84 ± 2,58 × 0,04 = [1,74-1,94] g/l. Cet intervalle est plus large que le précédent car on a moins de chance de se tromper en disant que l'intervalle contient μ. En contrepartie, on a plus d'incertitude sur μ.

De façon générale, l'intervalle de confiance à (1-α)% s'écrit :

$$IC_{1-\alpha\%}(\mu) = \left[\hat{\mu} \pm U_\alpha \times \sqrt{\frac{\hat{\sigma}^2}{n}}\right].$$

Mais **comment énoncer ce résultat** ? Pour un intervalle de confiance à 95%, il n'est pas correct de dire qu'il y a 95% de chances que la moyenne de la population soit dans l'intervalle : la moyenne de la population étant une valeur unique, qui ne fluctue pas, elle n'est pas attachée à des probabilités. « 95% », c'est la probabilité que les limites calculées à partir d'un échantillon représentatif incluent la vraie valeur, μ. La construction de cet intervalle rend donc son interprétation délicate.

> La figure OUTILS-STAT-9 donne un exemple du calcul d'intervalles de confiance à partir de 20 échantillons tirés au sort d'une population d'hommes dont on sait que la taille moyenne, μ, est 172 cm. Le trait horizontal représente la moyenne dans la population, 172 cm. Certains des échantillons ont une moyenne proche de cette valeur, d'autres ont une moyenne qui en est bien plus éloignée. Pour certains, la moyenne est au-dessus, pour d'autres, elle est au-dessous. La moyenne de la population, 172 cm, est

contenue dans l'intervalle de confiance de 19 échantillons sur 20, soit 95% des intervalles de confiance.

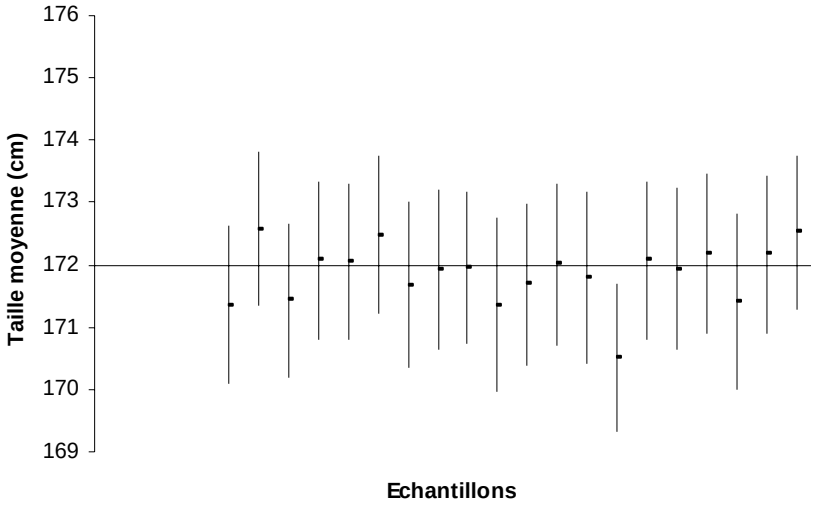

Figure OUTILS-STAT-9. Intervalles de confiance à 95% calculés
à partir de 20 échantillons de 100 sujets issus d'une population d'hommes
dont la moyenne est 172 cm. Pour chaque échantillon, le trait vertical représente
l'étendue entre la borne inférieure et la borne supérieure de l'intervalle de confiance
et le petit trait horizontal représente la moyenne observée.

En général, 95% des intervalles ainsi calculés contiennent la valeur de la moyenne de la population, μ. Comme nous ne pouvons pas savoir, néanmoins, quels sont les 95 échantillons sur 100 qui contiennent μ, on dit que l'on est « confiant » que la moyenne se trouve à l'intérieur des limites calculées à partir d'un seul échantillon, en particulier.

Enfin, la partie $\left[U_\alpha \times \sqrt{\frac{\hat{\sigma}^2}{n}} \right]$ est appelée **précision** de l'estimation.

Par exemple, dans l'échantillon des 100 hommes jeunes, la précision de l'estimation est de 0,08 g/l pour l'intervalle de confiance à 95%.

4 Intervalle de confiance d'une moyenne, petits échantillons

La distribution théorique des moyennes est Normale à condition que la taille des échantillons soit suffisamment grande ($n \geq 30$). Pour des échantillons de plus petite taille, et à condition que la distribution théorique de la variable soit Normale, la distribution des moyennes est un peu différente : la forme de la courbe est plus aplatie que la loi Normale correspondante. Cette **loi**, dite **de Student**, dépend du nombre *n* de sujets. La **distribution du *t* de Student** a un seul paramètre, le nombre de **degrés de liberté**. L'entrée dans **la table de Student** est le nombre de degrés de liberté. Vous pouvez consulter la table de Student sur le site et l'imprimer (les valeurs présentées sont des valeurs absolues).

Pour le calcul de l'intervalle de confiance de la moyenne, le nombre de degrés de liberté est égal à *n-1*. A condition de pouvoir faire l'hypothèse que la variable suive une loi Normale dans la population, on a donc pour l'intervalle de confiance de la moyenne :

$$IC_{1-\alpha\%}(\mu) = \left[\hat{\mu} \pm t_{n-1,\alpha} \times \sqrt{\frac{\hat{\sigma}^2}{n}} \right]$$

Sur un échantillon de 26 sujets, la moyenne de la concentration de glucose dans le sang est 5,50 mmol/l et la variance, $\hat{\sigma}^2$, 1,25 (mmol/l)^2. On considère que la concentration de glucose dans le sang a une distribution Normale dans la population dont est issu cet échantillon. Comme le nombre de sujets est inférieur à 30, on calcule l'intervalle de confiance à 95% de la moyenne en faisant : 5,50 ± 2,06 × 0,219 = [5,05-5,95] mmol/l, car pour α=5% et 25 degrés de liberté, on lit dans la table de Student 2,0595, soit 2,06. On est confiant que la vraie valeur de la moyenne de la concentration de glucose dans le sang est comprise entre 5,05 et 5,95 mmol/l.

Dans le cas de petits échantillons, ou si l'on a de bonnes raisons de penser que la variable quantitative ne suit pas une loi Normale, l'usage de la médiane et de l'étendue inter-quartiles ou de l'étendue (minimum-maximum) est souvent préférée pour décrire la distribution d'une variable quantitative plutôt que celui de la moyenne et de son intervalle de confiance.

5 Intervalle de confiance d'une proportion, grands échantillons

Pour une variable catégorielle dichotomique, le nombre de fois où une catégorie, par exemple « femmes » pour la variable « sexe », apparaît dans un échantillon de taille *n* suit une loi binomiale. La forme d'une telle distribution est asymétrique, mais devient de plus en plus symétrique lorsque *n* augmente et peut ressembler à une distribution Normale. On peut montrer que cette approximation par la loi Normale est correcte lorsque la proportion d'une des catégories dans la population, notée *P*, et son complémentaire, 1- *P*, sont tous deux plus grands que 5/*n* ou encore que : (*P* x *n*) et [(1- *P*) x *n*] sont tous deux plus grands que 5.

De façon similaire à la notation pour la moyenne de variables quantitatives, l'intervalle de confiance à (1-α)% de la proportion d'une des catégories d'une variable catégorielle s'écrit :

$$IC_{1-\alpha\%}(P) = \left[f \pm U_\alpha \times \sqrt{\frac{f(1-f)}{n}} \right].$$

Par exemple, si dans un échantillon de 100 enfants de 5 à 10 ans, représentatif de la population des enfants du même âge, on trouve que 20 d'entre eux présentent des signes d'allergie, la proportion des enfants ayant des signes d'allergie est 0,20 ou 20%. L'intervalle de confiance à 95% de cette proportion vaut : $\left[0{,}20 \pm 1{,}96 \sqrt{\frac{0{,}20 \times 0{,}80}{100}} \right]$ = [0,12-0,28].

On doit vérifier *a posteriori* que le calcul était possible de cette façon-là, en faisant : (*P* x *n*) et [(1- *P*) x *n*]. Comme on ne connaît pas *P*, on utilise son estimation à chacune des bornes de l'intervalle de confiance.

Dans l'exemple, on trouve respectivement : (0,12 × 100) et (0,88 × 100) et (0,28 × 100) et (0,72 × 100), qui sont tous supérieurs à 5.

En pratique, il suffit de faire la vérification avec la plus petite des quatre proportions (0,12 × 100).

De façon similaire à la notation pour la moyenne de variables quantitatives, la précision de l'estimation de la fréquence est : $\left[U_\alpha \times \sqrt{\frac{f \times (1-f)}{n}} \right]$.

III Étape de comparaison : test d'hypothèse

1 Introduction

Le plus souvent en épidémiologie et en santé publique, la question d'intérêt entraîne une **comparaison**.

On peut citer trois exemples :

1) la ration calorique des femmes travaillant de nuit est-elle différente de la ration théorique recommandée ?

2) la concentration moyenne de cholestérol total dans le sang est-elle différente chez les sujets présentant un infarctus du myocarde et chez les sujets n'en présentant pas ?

3) la glycémie moyenne de sujets diabétiques est-elle différente avant et après la prise d'un nouveau traitement anti-diabétique ?

Dans chacun des exemples, on veut mesurer une différence, un effet, mais selon la formulation de la question, on peut distinguer **trois grands types de situations** :

- 1ère situation : on souhaite comparer une information observée sur un échantillon à une valeur théorique,
- 2ème situation : on souhaite comparer deux échantillons indépendants,
- 3ème situation : on souhaite comparer la distribution d'une variable mesurée à deux reprises au sein du même échantillon.

Dans la 2ème situation, on a affaire à des **séries indépendantes** et dans la 3ème situation à des **séries appariées**.

Deux séries sont indépendantes si les sujets de ces deux séries sont totalement indépendants (pas d'élément commun entre les deux séries).

Par exemple, si l'on inclut 100 patients dans un essai thérapeutique évaluant des médicaments contre l'hypertension artérielle, le groupe des 50 sujets recevant le médicament A et le groupe des 50 sujets recevant le médicament B sont deux groupes indépendants.

Deux séries sont appariées si elles se correspondent par un élément commun.

Par exemple, si les 100 sujets reçoivent le même médicament et que l'on compare la pression artérielle avant et après traitement, les groupes sont appariés.

Autre exemple, dans une enquête comparant des cas de cancers de poumon à des sujets témoins (qui en sont indemnes) quant à leur exposition au tabac, si on sélectionne pour chaque cas un témoin de mêmes âge et sexe, les témoins sont « appariés » aux cas : on a constitué des paires (un cas et un témoin) ressemblantes pour l'âge et le sexe.

Dans ce chapitre, nous verrons la démarche générale d'un test statistique, puis le choix du test approprié pour la comparaison de deux groupes en fonction de ces trois situations selon le type de variable (quantitative/catégorielle) et la taille de l'échantillon (grande/petite). Dans ces rappels utiles à la pratique de l'épidémiologie, nous n'aborderons néanmoins pas tous les cas. Vous pourrez vous référer aux ouvrages spécialisés en biostatistique (cf. références bibliographiques).

2 Test d'hypothèse : raisonnement général et exemple

Afin d'expliquer le surpoids des femmes travaillant de nuit dans une certaine entreprise, le médecin du travail fait l'hypothèse que les femmes exerçant un métier de nuit ont une ration calorique quotidienne plus importante que celle qui est habituellement recommandée pour la même tranche d'âge, soit 2000 kcal/jour. Une enquête permet de décrire la moyenne observée, $\hat{\mu}$, des apports caloriques quotidiens pour 100 femmes.

Comment décider si cette moyenne observée sur un échantillon de la population est ou non différente de la valeur théorique ?

A ce stade, vous n'avez pas de méthode pour répondre. Néanmoins, votre raisonnement peut s'appuyer sur plusieurs acquis : d'une part, la notion de fluctuation d'échantillonnage, d'autre part, la notion d'intervalle de confiance.

Supposons que la population des femmes travaillant de nuit se conforme à la ration théorique. D'après ce que vous savez des fluctuations d'échantillonnage, la moyenne observée sur de nombreux échantillons de 100 femmes sera souvent proche de 2000 kcal/jour, mais il pourra arriver que, sur quelques échantillons, on observe des moyennes très différentes de 2000 kcal/jour. Inversement, supposons que la ration calorique des femmes travaillant de nuit soit, en réalité, beaucoup plus importante (ou beaucoup moins importante) que ce qui est recommandé. Il pourra arriver qu'un échantillon permette d'observer, malgré tout, une ration calorique proche des recommandations. Par conséquent, compte tenu que l'on ne dispose le plus souvent que d'un seul échantillon, on ne pourra jamais répondre avec certitude à ce genre de problèmes. On ne pourra donc apporter une réponse qu'en acceptant de prendre certains risques : nous y reviendrons plus tard.

Le phénomène de fluctuation d'échantillonnage nous amène à formuler la question de la façon suivante. Si, sur un échantillon, on observe la distribution d'une variable en la caractérisant par sa moyenne et son écart-type, comment savoir si cette distribution est différente de la valeur attendue ou si cette différence peut être attribuée au seul hasard (c'est-à-dire aux fluctuations d'échantillonnage) ?

Par exemple, sur un échantillon de 100 femmes exerçant un métier de nuit, on observe une ingestion moyenne de 2200 kcal/jour ($\hat{\sigma}$: 845 kcal/jour). Compte tenu que cette observation provient d'un échantillon, cette ingestion moyenne par jour est-elle plus importante que ce qui est recommandé ou bien cette différence peut-elle être attribuée au seul hasard ?

Pour commencer, faisons l'hypothèse qu'il n'y a pas de différence (« différence nulle ») entre la distribution observée et la distribution théorique. Cette hypothèse s'appelle l'**hypothèse nulle** : c'est l'hypothèse à tester.

Dans l'exemple, l'hypothèse nulle est que la ration moyenne observée ne diffère pas de la ration théorique.

On sait que la moyenne observée varie autour de la moyenne, du fait des fluctuations d'échantillonnage. On peut calculer l'intervalle en dehors duquel la moyenne observée n'a pas plus de α chances de sortir, sous cette hypothèse. D'après ce que vous avez déjà vu, l'écart (précision) est égal à :

$$U_\alpha \times \sqrt{\frac{\hat{\sigma}^2}{n}}$$

Dans l'exemple, la moyenne observée varie autour de 2000 kcal/j, du fait des fluctuations d'échantillonnage. L'écart de l'intervalle en dehors duquel la moyenne observée n'a pas plus de 5% de chance de sortir sous cette hypothèse s'écrit :

$$U_{5\%} \times \sqrt{\frac{\hat{\sigma}^2}{n}} = 1{,}96 \times \frac{845}{\sqrt{100}} = 165{,}62 \approx 166 kcal / j$$

Ainsi, sous l'hypothèse nulle, la moyenne observée n'a que 5% de chance de sortir de l'intervalle [1834-2166] kcal/j.

Le calcul de cet intervalle va nous donner un guide pour la prise de **décision** :

- si la moyenne observée tombe dans l'intervalle, nous admettrons que l'écart avec la moyenne théorique résulte des seules fluctuations d'échantillonnage.

  Dans l'exemple, on conclut alors que la ration calorique des femmes qui travaillent la nuit est conforme aux recommandations.
- si la moyenne observée tombe en dehors de l'intervalle, l'écart à la valeur théorique est trop grand pour pouvoir être attribué aux seules fluctuations d'échantillonnage, on dira alors qu'il est **significatif**.

  Dans l'exemple, on conclut alors que la ration calorique des femmes travaillant la nuit est différente de la ration théorique.

En prenant cette décision, on s'expose à deux **risques** :

- le **risque α**, ou **risque d'erreur**, est le risque de rejeter l'hypothèse nulle alors qu'en fait elle est exacte.

  Dans l'exemple, si, en réalité, les femmes travaillant de nuit se conforment à la ration théorique, il pourra arriver que la ration calorique moyenne observée soit très différente de 2000 kcal/j et se situe en dehors de l'intervalle : ce risque est connu, c'est le risque α, dit de **1ère espèce**, et le plus souvent, il est fixé à 5%,
- le **risque β**, ou **manque de puissance**, est le risque de ne pas rejeter l'hypothèse nulle alors qu'en fait elle est fausse.

  Dans l'exemple, si, en réalité, les femmes travaillant de nuit ne se conforment pas à la ration théorique, il pourra arriver que, sur quelques échantillons, la ration calorique moyenne observée soit malgré tout proche de 2000 kcal/j et se situe dans l'intervalle. Dans ce cas, on ne verra donc pas une différence qui existe en réalité, c'est donc un manque de puissance. Ce risque, β, dit de **2ème espèce**, n'est en revanche pas connu.

Ces deux risques, α et β, sont **antagonistes** et c'est bien cela qui permet de comprendre pourquoi on ne souhaite pas trop diminuer le risque d'erreur, 5% étant choisi comme seuil, le plus souvent. En effet, si l'on choisissait un risque α infiniment petit, on ne rejetterait jamais l'hypothèse nulle, mais en revanche, on ne mettrait jamais en évidence des écarts qui pourtant semblent importants.

Dans notre exemple, on a vu que pour un risque α de 5%, on a l'intervalle : [1834-2166] kcal/j. Que se passe-t-il si l'on veut réduire le risque de déclarer à tort qu'il existe une différence et que l'on choisit un risque α, par exemple, de 1 pour mille ? L'intervalle correspondant est : [1722-2278] kcal/j. On conclura à une différence moins souvent puisque l'intervalle est plus large. En conséquence, il faudra des écarts plus importants à la ration théorique pour conclure à une différence.

On s'en tiendra donc le plus souvent à un risque α de 5%, qui est certes un choix arbitraire, mais admissible. C'est la valeur de α qui permet de calculer l'intervalle à partir duquel on décidera si l'écart est ou non significatif. Le risque α ainsi fixé est appelé **seuil de signification**, le terme seuil étant employé car ce choix va permettre de prendre une décision.

Si on résume, le **test d'hypothèse** se décompose **selon les étapes suivantes** dans l'exemple :

1) **Hypothèse nulle** : « il n'y a pas de différence entre l'apport calorique de ce groupe de femmes et ce qui est recommandé »
   L'**hypothèse alternative**, Ha, est : « il y a une différence entre l'apport calorique de ce groupe de femmes et ce qui est recommandé ».

2) Écriture de la **statistique de test**, après vérification de ses conditions d'application :
   Dans l'exemple, la statistique de test ou **test de l'écart-réduit** :
   $U_\alpha = \frac{\hat{\mu} - \mu}{\sqrt{\hat{\sigma}^2 / n}}$ suit une loi Normale centrée, réduite, car $n \geq 30$.
   Cette écriture vient de l'intervalle précédemment cité. C'est une forme d'écriture très commune pour les tests d'hypothèse.

3) Choix du **seuil de décision** et définition de la **région de rejet** (« région critique ») de l'hypothèse nulle :
   Pour un seuil α = 0,05, on rejettera l'hypothèse nulle chaque fois que $|U_\alpha| > 1{,}96$ (Figure OUTILS-STAT-10),

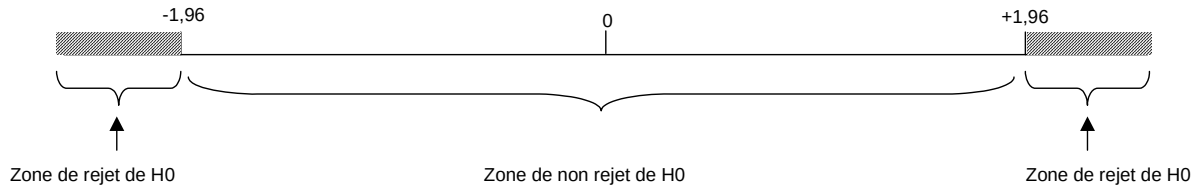

Figure OUTILS-STAT-10. Régions de rejet et de non rejet de l'hypothèse nulle (H_0) dans le cas où la statistique de test suit une loi Normale, centrée et réduite.

4) **Calcul** de la statistique de test : application numérique : $U_\alpha = \frac{2200 - 2000}{845 / \sqrt{100}} = 2{,}37$

5) **Comparaison** du résultat obtenu au seuil de la région de rejet de l'hypothèse nulle, **décision et conclusion** par rapport à la question posée initialement :
   Dans l'exemple, le résultat du calcul de la statistique de test : 2,37 est plus grand que 1,96 et l'on rejette donc l'hypothèse nulle. Dans la table de la loi Normale centrée réduite, on peut voir que 2,37 correspond à une probabilité *P* = 0,02. « *P* » est en fait le **degré de signification** : l'écart entre la moyenne observée et la moyenne théorique est significatif à 2 pour 100. Ceci signifie que, si en réalité les femmes travaillant de nuit n'ont pas une ration calorique différente de ce qui est recommandé, les seules fluctuations d'échantillonnage n'auraient que 2 chances sur 100 de conduire à une différence égale ou supérieure à celle que l'on observe.
   On conclut que l'apport calorique quotidien des femmes exerçant ce métier de nuit est donc significativement différent de celui recommandé (P<0,02).

Cette méthode est un **test** car on prend une décision à un moment donné sur le rejet ou non de l'hypothèse nulle, et on en tire des conséquences vis-à-vis de la question initiale.

3 Comparaison de moyennes selon deux modalités d'une variable catégorielle

Rappelez-vous les trois grands types de situations qui conduisent au choix du test approprié : comparaison à une valeur théorique, comparaison de groupes indépendants, comparaison de séries appariées. La première situation a déjà été abordée dans le chapitre précédent.
Pour chaque situation, vous devez examiner la taille de l'échantillon. Nous pourrions utiliser un test différent selon que le nombre de sujets est supérieur à 30 dans les deux groupes ou non. Néanmoins, nous avons vu que, pour des échantillons de taille < 30 sujets, on utilise la **distribution du t de Student** qui est proche de la distribution Normale et s'en rapproche de plus en plus quand *n* augmente. On proposera uniquement ce type de test dans ce chapitre.

A Comparaison de deux moyennes observées lorsque les groupes sont indépendants

Le principe du test est toujours le même, seule la statistique de test va changer. En présence de deux groupes, on forme le rapport entre la différence des moyennes observées et l'écart-type de la différence des moyennes : $t = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_c \times \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$

Or, afin de calculer cet écart-type, on peut montrer qu'il est nécessaire de prendre la moyenne pondérée des écarts-types de chacun des groupes, c'est-à-dire donnant un poids proportionnel à la taille de chaque groupe. On va l'appeler l'**écart-type « commun »** :

$$\hat{\sigma}_c = \sqrt{\frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}}$$

La statistique de test : $t = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_c \times \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ suit une loi du t de Student à $n_1 + n_2 - 2$ degrés de liberté.
Le déroulement du test et le principe de la décision sont toujours les mêmes.

Remarque sur les **conditions d'application du test du *t* de Student** : la distribution théorique de la variable doit être Normale (si $n < 30$) et de même variance dans les deux populations dont proviennent les échantillons. Ces deux conditions semblent drastiques et sont pourtant très souvent vérifiées pour les caractéristiques étudiées en épidémiologie. Il existe des tests permettant de vérifier la normalité ou l'égalité des variances. Ils ne sont pas au programme de cet enseignement, mais vous pouvez les trouver dans les livres de biostatistique ou dans les logiciels statistiques couramment utilisés. Enfin, il faut signaler que le test du *t* de Student est robuste, même en cas de violation mineure de ces hypothèses. En revanche, lorsque, au moins à l'examen des données, vous suspectez des violations majeures de ces hypothèses, il est recommandé d'employer d'autres tests : test du *t* de Student pour variances inégales ou tests non paramétriques. Ils ne sont pas non plus au programme de cet enseignement, mais, si vous avez compris le principe du test le plus simple, vous pourrez comprendre les explications figurant dans un livre.

> Voici un exemple commenté. Chez 950 hommes âgés de plus de 40 ans, la concentration sanguine moyenne de cholestérol total des 50 patients qui ont présenté un infarctus du myocarde : $\hat{\mu}_1$ = 5,14 mmol/l ($\hat{\sigma}_1$ = 1,06 mmol/l) par rapport aux 900 n'ayant pas présenté d'infarctus $\hat{\mu}_2$ = 4,49 mmol/l ($\hat{\sigma}_2$ = 1,00 mmol/l). Y a-t-il une différence entre ces deux groupes ?

Le test se déroule selon les étapes suivantes :

1. Hypothèse nulle : la concentration sanguine moyenne du cholestérol total ne diffère pas selon que les patients ont présenté ou non un infarctus du myocarde.
2. Statistique de test :

   Comme les effectifs sont supérieurs à 30, $t = \dfrac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_c \times \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ suit une loi du t de Student à 50 + 900 – 2 = 948 degrés de liberté.
3. Pour un seuil $\alpha = 0{,}05$, on rejettera l'hypothèse nulle lorsque $|t| > 1{,}96$.
4. Calcul de la statistique de test :

   L'écart-type commun est égal à : $\hat{\sigma}_c = \sqrt{\dfrac{(50-1)1{,}06^2 + (900-1)1{,}00^2}{50+900-2}} = 1{,}0032$ mmol/l.

   La statistique de test vaut : $t = \dfrac{5{,}14 - 4{,}49}{1{,}0032 \times \sqrt{\dfrac{1}{50} + \dfrac{1}{900}}} = 4{,}46$.
5. Décision :

   Le résultat de la statistique de test est supérieur à 1,96. On rejette donc l'hypothèse nulle au seuil $\alpha = 0{,}05$. Le taux moyen de cholestérol total est significativement différent chez les hommes de plus de 40 ans ayant fait un infarctus que chez ceux n'ayant pas fait d'infarctus. Dans la table de Student, on peut voir que 4,46 correspond à une probabilité $P < 0{,}001$. On conclut que la concentration sanguine moyenne de cholestérol total est significativement plus élevée chez les patients ayant présenté un infarctus du myocarde que chez les patients n'en ayant pas présenté (P<0,001).

Remarquez que :

- pour les **calculs intermédiaires**, celui de l'écart-type commun par exemple, on garde au moins 4 chiffres après la virgule, alors que pour le résultat final, on ne garde pas plus de deux décimales. Dans le premier cas, trop arrondir pourrait conduire à fausser les résultats, dans le deuxième cas, garder trop de décimales rendrait le résultat peu lisible, car incompatible avec la précision du dosage.
- on aurait pu choisir de mettre au numérateur $\hat{\mu}_2$ en premier dans la formule, soit $t = \dfrac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_c \times \sqrt{\dfrac{1}{n_2} + \dfrac{1}{n_1}}} = -4{,}46$. La table de Student présentée correspond en fait à des valeurs absolues, et, étant symétrique autour de la valeur 0, aurait conduit à la même conclusion et à la même valeur P que précédemment.

B Comparaison de deux moyennes observées lorsque les groupes sont appariés

Pour commencer avec ce type de données très particulières, voici un exemple.

Dix personnes diabétiques ont reçu un nouveau médicament qui aurait des propriétés antidiabétiques. La mesure de la glycémie à jeun a été réalisée avant traitement et 10 jours après. Les résultats figurent dans le tableau OUTILS-STAT-6.

Tableau OUTILS-STAT-6. Glycémie à jeun chez 10 personnes diabétiques avant et après la prise du nouveau médicament.

Sujet n°	Glycémie à jeun (mmol/l)		Différence (avant-après)
	Avant traitement	Après traitement	
1	7,2	6,5	+0,7
2	8,3	8,1	+0,2
3	9,2	9,3	-0,1
4	8,5	8,4	+0,1
5	7,9	5,8	+2,1
6	10,2	9,8	+0,4
7	6,8	6,7	+0,1
8	7,2	7,3	-0,1
9	8,3	8,4	-0,1
10	9,4	8,6	+0,8
Moyenne	8,30	7,89	+0,41
Ecart-type : $\hat{\sigma}$	1,08	1,28	0,68

Pour traiter cette question, on pourrait utiliser le test de comparaison de moyennes pour deux séries indépendantes. Or, cette manière de répondre n'est pas correcte, car on omet une information importante : chaque paire de mesures concernant le même sujet, les séries ne sont donc pas indépendantes. La solution correcte consiste donc à calculer la **différence pour chaque paire de résultats**.

Comme indiqué dans le tableau OUTILS-STAT-6, on dispose ainsi de 10 différences, positives ou négatives. Si le traitement n'a aucun effet sur la glycémie, la moyenne de ces différences doit être égale à 0, en théorie.

Le test à utiliser est donc celui qui permet de comparer la moyenne observée des différences à la valeur théorique 0. Ce test est basé sur l'écart-réduit dans le cas d'échantillons de taille supérieure à 30 (ce test ne sera pas présenté ici). Dans le cas d'échantillons de plus petite taille, sous réserve que la différence soit distribuée selon une loi Normale, la statistique de test : $t = \frac{\hat{\mu} - 0}{\hat{\sigma}/\sqrt{n}}$ suit une loi du t de Student à *n*-1 degrés de liberté.

Lorsqu'on analyse des groupes appariés, on diminue la variabilité des données puisqu'on tient compte de la différence des valeurs pour un même sujet, éliminant la variabilité inter-sujets.

D'après le tableau OUTILS-STAT-6, on peut observer que l'écart-type des différences est plus faible que l'écart-type des valeurs individuelles.

Pour le test, on a : $t = \frac{+0{,}41 - 0}{0{,}68/\sqrt{10}} = 1{,}90$.

Ce résultat est à comparer à la table de *t* pour 9 degrés de liberté et le risque 5%, soit 2,2622. Le résultat de la statistique de test est inférieur à cette valeur. On ne rejette donc pas l'hypothèse nulle. La différence observée de glycémie avant et après la prise de ce médicament est compatible avec une différence nulle. En concluant que le traitement n'a pas d'effet, on prend un autre risque : celui de ne pas mettre en évidence une différence alors qu'en réalité elle existe. C'est précisément le risque β.

4 Comparaison de proportions entre deux groupes d'une variable catégorielle

Pour cette comparaison, on utilise un test que l'on notera **Chi-2** (on prononce « qui deux ») ou **Chi-carré** ou χ^2 (notations employées). Vous verrez d'abord la situation de deux groupes indépendants, puis, la comparaison à une valeur théorique et l'analyse des séries appariées.

A Chi-2 d'indépendance pour un tableau 2 x 2

Par exemple, on voudrait savoir s'il existe un lien entre la bronchite chronique et l'exposition (oui/non) à un certain toxique employé dans un groupe industriel. Pour cela, une étude a permis de recueillir, dans ce groupe industriel, les informations sur 100 sujets exposés au toxique et 200 sujets non exposés (Tableau OUTILS-STAT-7).

Tableau OUTILS-STAT-7. Comparaison des employés d'un groupe industriel pour l'exposition à un toxique dans l'usine de Maville : effectifs observés.

	Bronchite chronique		Total
	Oui	Non	
Exposition au toxique			
Oui	25	75	100
Non	5	195	200
Total	30	270	300

Un tableau contenant les effectifs des différentes catégories de plusieurs variables est appelé **tableau de contingence**. Le tableau OUTILS-STAT-7 est le tableau de contingence des effectifs observés, e_o.

L'hypothèse nulle est qu'il n'y a pas de lien entre les deux caractéristiques étudiées, ou, autrement dit, qu'elles sont indépendantes.

Dans l'exemple, l'hypothèse nulle est : « le fait d'être atteint d'une bronchite chronique est indépendant de l'exposition au toxique ». Sous cette hypothèse, la probabilité d'être malade est la même chez les exposés et les non exposés. La proportion globale de malades est 30/300 = 0,10. Remarquez que c'est le rapport entre le total de la colonne « bronchite chronique, oui » et le total général. Chez les exposés, le nombre attendu (effectif théorique) de cas de bronchite chronique sous H_0 est donc 30/300 × 100 = 10. Remarquez qu'il s'agit de la proportion de malades appliquée au total des sujets exposés.

On peut ainsi en déduire un **tableau de contingence théorique** (Tableau OUTILS-STAT-8). Les effectifs théoriques sont notés e_t.

Tableau OUTILS-STAT-8. Comparaison des cas et des sujets témoins pour l'exposition à un toxique dans le groupe industriel X : effectifs théoriques.

	Bronchite chronique		Total
	Oui	Non	
Exposition au toxique			
Oui	10	90	100
Non	20	180	200
Total	30	270	300

Le test est basé sur la différence entre les effectifs observés et les effectifs théoriques calculés sous l'hypothèse nulle.

On peut montrer que : $\chi^2 = \sum \frac{(e_O - e_t)^2}{e_t}$ suit une loi connue appelée loi du Chi-2 dont la distribution dépend de **degrés de liberté** calculés en fonction du nombre de catégories de chaque variable. Ce test est appelé **test du Chi-2 de Pearson**. Si l'on appelle « *c* » le nombre de catégories de la variables en colonne du tableau et « *l* » le nombre de catégories de la variable en ligne, le nombre de degrés de liberté est égal à $(c-1)\times(l-1)$. La condition d'application de ce test est que **tous les effectifs théoriques sont supérieurs ou égaux à 5**. Si un effectif au moins est strictement inférieur à 5, on utilisera un autre test : le test exact de Fisher. Il n'est pas au programme, mais on peut en trouver l'explication dans les livres de statistique médicale.
L'entrée dans la **table du Chi-2** se fait par le nombre de degrés de liberté. Remarquez que, s'agissant d'un carré, le résultat de la statistique de test sera toujours supérieur ou égal à 0. Vous pouvez consulter la table du Chi-2 sur le site et l'imprimer.

Dans l'exemple, si l'on reprend les différentes étapes du test :
1. L'hypothèse nulle est qu'il y a indépendance entre l'exposition au toxique et la maladie. Il n'y a pas de lien entre l'exposition au toxique et la maladie.
2. Sous cette hypothèse nulle, on peut calculer le tableau de contingence théorique. Comme tous les effectifs théoriques sont supérieurs à 5, la statistique de test est : $\chi^2 = \sum \frac{(e_o - e_t)^2}{e_t}$ qui suit une loi du Chi-2 à 1 degré de liberté.
3. Si nous choisissons le seuil α = 0,05, nous pouvons lire dans la table du Chi-2 à 1 degré de liberté que la valeur correspondant à ce seuil est 3,84.

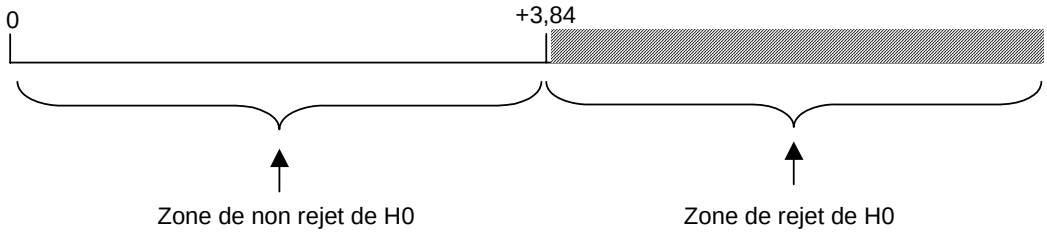

Figure OUTILS-STAT-11. Régions de rejet et de non rejet de l'hypothèse nulle (H_0) dans le cas où la statistique de test suit une loi du Chi-2 à un degré de liberté.

4. Le calcul de la statistique de test permet de trouver :
$$\chi^2 = \frac{(25-10)^2}{10} + \frac{(75-90)^2}{90} + \frac{(5-20)^2}{20} + \frac{(195-180)^2}{180} = 37{,}5\,.$$
5. Décision :
Ce résultat est supérieur à 3,84 : on rejette l'hypothèse nulle. On conclut à une association entre l'exposition au toxique et la bronchite chronique. La proportion de sujets présentant une bronchite chronique est significativement plus élevée chez les sujets exposés au toxique (25/100 = 25%) que chez les sujets non exposés au toxique (5/200 = 2,5%) (P<0,001).

B Chi-2 d'ajustement pour un tableau 2 x 2

On peut utiliser la même démarche pour comparer la distribution d'une seule caractéristique catégorielle par rapport à une distribution théorique.

Par exemple, dans un dispensaire, les médecins ont l'impression de voir de plus en plus souvent des femmes atteintes de migraine. Dans la population des femmes de la même tranche d'âge, on sait que la proportion des femmes ayant une migraine est de 12 %. En revoyant les dossiers des 200 femmes qui ont consulté au cours du mois précédent, on trouve 40 cas de migraine, soit 20 %.

Pour cela, on peut utiliser un test du Chi-2 dont le nombre de **degrés de liberté** est égal au nombre de catégories de la variable, k, moins 1. La condition d'application de ce test est que **tous les effectifs théoriques sont supérieurs ou égaux à 5**.

On peut donc utiliser également ce test pour comparer une proportion observée à une proportion théorique ou de façon plus générale une distribution observée à une distribution théorique.

Dans l'exemple, l'hypothèse nulle du test est qu'il n'y a pas de différence entre la proportion observée, *f*, de 20 % et la proportion théorique, *P*, de 12 %. On fait par ailleurs l'hypothèse que les femmes de la consultation sont représentatives de la population pour la prévalence de la migraine. On a le tableau OUTILS-STAT-9 des effectifs observés, e_o.

Tableau OUTILS-STAT-9 Effectifs observés, e_o, de migraine dans un échantillon de 200 sujets.

	Migraine		Total
	Oui	Non	
e_o	40	160	200

Le test se déroule suivant les étapes habituelles :

1. Hypothèse nulle : la proportion des femmes ayant une migraine ne diffère pas dans cet échantillon par rapport à la population générale.
2. Sous cette hypothèse, on peut dresser un tableau de contingence des effectifs théoriques. Dans ce tableau, les effectifs théoriques sont obtenus en appliquant la proportion théorique, *P*, à l'effectif de l'échantillon, n=200.

Tableau OUTILS-STAT-10. Proportion, P, de sujets souffrant de migraine dans une population et effectifs théoriques, e_t, dans un échantillon de 200 sujets.

	Migraine		Total
	Oui	Non	
P	0,12	0,88	1,00
e_t	24	176	200

Comme tous les effectifs théoriques sont supérieurs à 5, la statistique de test est : $\chi^2 = \sum \frac{(e_o - e_t)^2}{e_t}$ suit une loi du Chi-2 à 1 degré de liberté.

3. Au risque $\alpha = 0,05$, la valeur correspondante du Chi-2 à 1 degré de liberté est 3,84.
4. Le calcul permet de trouver : $\chi^2 = \frac{(40-24)^2}{24} + \frac{(160-176)^2}{176} = 12,1.$

5. Décision. Ce résultat est supérieur à 3,84. On rejette l'hypothèse nulle. Au cours de la période d'étude, la proportion de femmes ayant une migraine au sein de la consultation est significativement supérieure à celle de la population générale (P<0,001).

Attention à ne pas confondre les deux utilisations de la notation « P », soit P : proportion théorique, et P : degré de signification. Le contexte d'utilisation est différent.

C Chi-2 pour séries appariées

Une étude vise à étudier l'impact de mesures d'éducation pour la santé. Au total, 100 jeunes adultes sont interrogés avant et après la diffusion d'une émission de radio expliquant l'intérêt des préservatifs pour la prévention des maladies sexuellement transmises. On voudrait savoir si l'émission a eu un impact sur l'utilisation systématique de préservatifs.

Pour répondre à la question, on doit prendre en compte le fait qu'il s'agit d'observations appariées.

La solution consiste à classer les sujets selon 4 groupes comme dans le tableau OUTILS-STAT-11.

Tableau OUTILS-STAT-11. Effectif de sujets selon les différentes combinaisons des caractéristiques : notation générale et résultats d'une étude en guise d'exemple*.

		Utilise systématiquement un préservatif après l'émission		
		Oui	Non	Total
Utilise systématiquement un préservatif avant l'émission	Oui	e=10	f=18	a=28
	Non	g=29	h=43	c=72
	Total	b=39	d=61	T=100

* utilisation d'un préservatif avant et après une émission de radio dans Marégion, 1998.

On souhaite comparer une proportion « avant » et une proportion « après », soit respectivement : $p_1 = \frac{(e+f)}{n}$ et $p_2 = \frac{(e+g)}{n}$. Or, ces deux proportions ne sont pas indépendantes puisqu'elles contiennent « e » toutes les deux. Si l'on forme la différence, on trouve : $p_1 - p_2 = \frac{(e+f)}{n} - \frac{(e+g)}{n} = \frac{(f-g)}{n}$. Autrement dit, seules les paires discordantes comptent pour résoudre le problème.

La statistique de test utilisée s'écrit de la façon suivante :

$\chi^2 = \frac{(|f-g|-1)^2}{f+g}$ et suit une loi du Chi-2 à un degré de liberté.

Ce test est aussi connu sous le nom de **test de McNemar** corrigé (la correction est représentée par une soustraction de « 1 » au numérateur). Ce test est applicable, quelles que soient les valeurs de f ou g.

Si l'on résume les différentes **étapes du test** pour les données de l'exemple :
1. L'hypothèse nulle est qu'il n'y a pas de lien entre le fait d'avoir écouté l'émission de radio et l'utilisation systématique de préservatifs.

2. Sous cette hypothèse nulle, la statistique de test est : $\chi^2 = \frac{(|f-g|-1)^2}{f+g}$ qui suit une loi du Chi-2 à 1 degré de liberté.

3. Si nous choisissons le seuil $\alpha = 0,05$, nous pouvons lire dans la table du Chi-2 à 1 degré de liberté que la valeur correspondant à ce seuil est 3,84.

4. Le calcul de la statistique de test permet de trouver : $\chi^2 = \frac{(|18-29|-1)^2}{(18+29)} = 2,13$.

5. Décision :
Ce résultat est inférieur à 3,84, il ne tombe pas dans la région critique : on ne rejette pas l'hypothèse nulle. Il n'y a pas d'association entre le fait de regarder l'émission et l'utilisation systématique de préservatifs (P compris entre 0,10 et 0,20).

5 Choix du test statistique

Le choix du test statistique selon le type de comparaison et les conditions d'application est synthétisé sur la figure OUTILS-STAT-12.

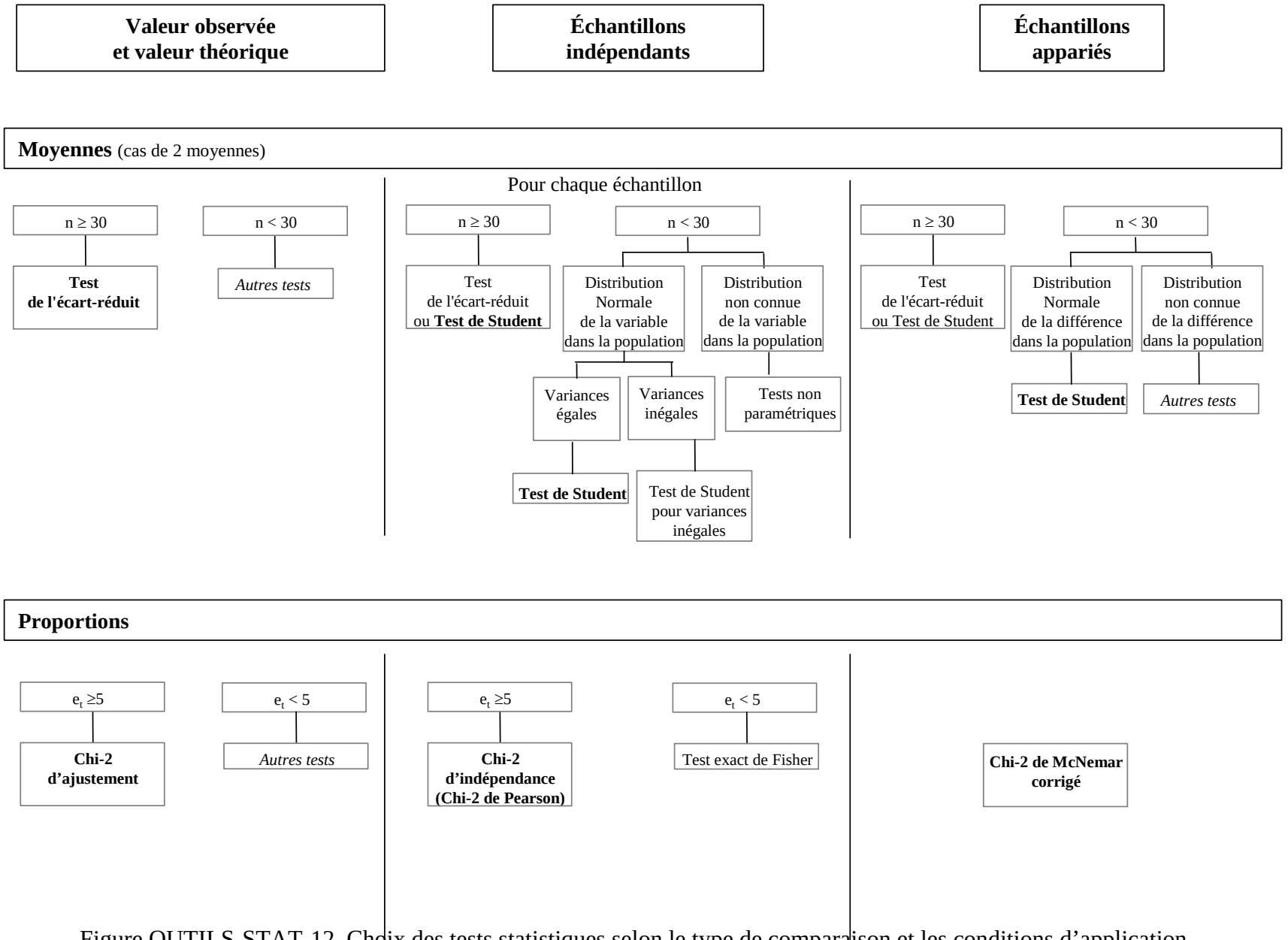

Figure OUTILS-STAT-12. Choix des tests statistiques selon le type de comparaison et les conditions d'application.

Le nom des tests détaillés dans ce module apparaît en caractères gras.
Deux tests différents peuvent porter le même nom (par exemple, test de l'écart-réduit pour comparer une moyenne observée et une moyenne théorique ou deux moyennes théoriques)

Références bibliographiques

Ancelle T.
Statistique. Épidémiologie.
Paris : Maloine; 2002.
ISBN 2 224 027060

Bouyer J.
Méthodes statistiques. Médecine - Biologie.
Paris : INSERM; 1996. 353 pages.
ISBN 2-909455-74-2

Falissard B.
Comprendre et utiliser les statistiques dans les sciences de la vie. 2ème édition.
Paris : Masson; 1998. 332 pages (Abrégés).
ISBN 2225850305

Valleron AJ.
Introduction à la biostatistique.
Paris : Masson; 1998. 448 pages (Évaluation et Statistique).
ISBN 2-225-83285-4

Altman, DG.
Practical statistics for medical research. 3rd edition.
London : Chapman & Hall; 1991. xii-611pages (Statistics texts).
ISBN 0412276305

Bland M.
An Introduction to Medical Statistics. 3rd edition.
Oxford : Oxford University Press; 2000. 422 pages.
ISBN 0-19-263269-8

Campbell MJ, Machin D.
Medical statistics: a commonsense approach. 3rd edition.
New York : Wiley; 1999. xiv-203 pages.
ISBN 0471987212