Ministère de l'enseignement supérieur
Université de Sfax
Faculté de médecine de Sfax

Certificat de Médecine préventive et d'épidémiologie

Module : STATISTIQUE

Année universitaire : 2024/2025

Service de médecine communautaire et d'épidémiologie
CHU H. CHAKER Sfax

Dr. H. MAAMRI
Dr. M. BEN JMAA
Pr. Ag. Y. MEJDOUB
Pr. J. JDIDI
Pr. S. YAICH

Service d'hygiène hospitalière
CHU H. CHAKER Sfax

Dr. N. KETATA
Dr. M. BEN HAMIDA
Dr. H. BEN AYED

Service de médecine préventive et d'hygiène hospitalière
CHU H. BOURGUIBA Sfax

Dr. M. TRIGUI
Pr. M. KASSIS

GENERALITES

1- STATISTIQUE ET VARIABILITE

Le domaine de la biologie, et plus encore celui des sciences humaines et la médecine, est placé sous le signe de la variabilité.

La variabilité est au maximum dans l'espèce humaine.

Chaque être humain est le résultat de facteurs nombreux :

- sa constitution génétique : hors le cas des vrais jumeaux, il n'y a pas deux personnes qui possèdent le même génome, le nombre des hommes ayant vécu, vivant actuellement et vraisemblablement à vivre étant très inférieur à l'ensemble des génomes possibles.
- son histoire psychologique, qui lui est propre.
- son environnement physique et social.

Ce qui constitue sa personnalité, c'est l'interaction permanente et complexe entre sa constitution génétique (l'inné) et les données psychologiques et environnementales (l'acquis).

Chaque être humain est unique et par suite différent, en raison de son extrême complexité.

Ainsi pour se limiter aux seuls paramètres biologiques, une population d'individus humains peut être classée selon le sexe (mâle ou femelle), le groupe sanguin ABO (A, B, AB, O), la taille (de 0,40m à 2,10m), l'âge (de 0 à 110 ans), la couleur des yeux, etc...

La variabilité que l'on observe en Biologie et en Médecine peut être considérée comme le résultat de l'intervention simultanée d'un grand nombre de chaînes causales (le génome, les événements de la vie, le milieu où s'est effectué le développement), le plus souvent interdépendantes, et qu'il est impossible d'analyser toutes en détail. A cette intervention simultanée, il est convenu de donner le nom de hasard (c'est une série de facteurs incontrôlés).

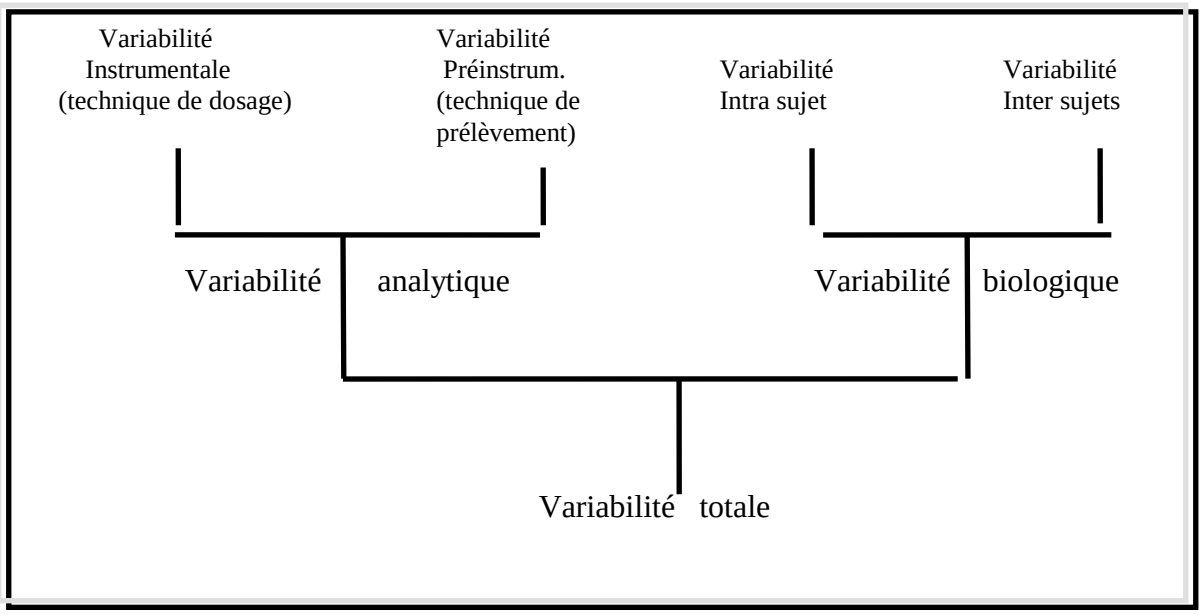
Un exemple de variabilité : celui d'un paramètre biologique. Considérons par exemple le dosage d'une des composantes du plasma (glucose ou créatinine). Examinons les facteurs qui peuvent influer sur le résultat final du dosage.

Chez un même sujet, le résultat peut varier d'un moment à un autre : il y a donc une variation intra sujet.

Il peut varier aussi -et en principe d'avantage- entre des sujets différents : variation inter sujets.

Mais les techniques de prélèvements (garrot plus ou moins serré, introduction de l'aiguille plus ou moins bien réalisée, etc...) et les techniques de dosage peuvent aussi influencer le chiffre obtenu. Il faut donc ajouter une nouvelle source de variation, la variation analytique, à la variation biologique indiquée ci-dessus.

Le tableau suivant résume ces considérations :



La variabilité et le hasard amènent à s'interroger sur la signification et sur l'interprétation des données, et sur la validité des décisions qu'elles entraînent.

2- LES DEUX DOMAINES DE LA STATISTIQUE

La statistique est une science et une méthode, qui ont un double but.

2-1- Décrire des ensembles de données complexes en opérant des réductions de ces données. C'est la statistique descriptive.

Exemples:

2-1-1- sur un ensemble de 250 personnes, on a déterminé les systèmes sanguins ABO. Plutôt que de garder les 250 résultats, on dénombre les individus A, B, O, AB. On a réduit les 250 résultats à 4 nombres dont la somme fait 250.

2-1-2- Sur le même ensemble de sujets, on note la taille. On obtient le plus souvent un récapitulatif de ces 250 valeurs, en calculant la moyenne et l'écart type de cette distribution.

2-2- Débusquer dans une variabilité constatée ce qui peut être expliqué par le hasard seulement ou ce qui relève d'une autre explication. C'est ce que l'on appelle la statistique inférentielle ou inductive.

Exemples:

2-2-1- chez un sujet, on a trouvé l'an passé une glycémie à jeun à 4,5 mmol/l. On trouve aujourd'hui un chiffre de 5,5 mmol/l. Cette différence s'inscrit-t-elle dans la variation intra sujet et la variation analytique normales ou est-elle le signe d'une évolution vers un état diabétique ? Faut-il ou non décider de traiter le sujet ?

2-2-2- deux traitements anticancéreux appliqués à des malades atteints de cancer de même type donnent respectivement des taux de survie à 5 ans de 43 % et 48 %. Cette différence peut-elle être expliquée par le hasard ou bien est-elle l'indication que le deuxième traitement est meilleur que le premier ? Faut-il ou non systématiquement prescrire le deuxième traitement ?

La statistique est justement la méthode qui permet de résoudre ce type de question et d'adopter une attitude très critique sur les données et les conséquences pratiques qu'on en tire.

Sous cet aspect, la statistique, science de la signification et de l'interprétation de la variabilité, apparaît comme une science de la décision.

Sous son double aspect descriptif et classificatoire d'une part, décisionnel d'autre part, la méthode statistique représente un des outils essentiels du raisonnement médical.

QUELQUES CONCEPTS DE BASE DE LA STATISTIQUE

1- SERIE STATISTIQUE

On appelle série statistique une collection d'objets de même nature, chez lesquels on s'intéresse à des caractéristiques communes ou variables, mais dont la valeur varie d'un sujet à l'autre, et qui sont susceptibles de mesure ou de classement.

Exemple: les étudiants d'une promotion constituent un ensemble « d'objets » de même nature -adultes humains jeunes- candidats à devenir médecins.

On peut étudier cette série statistique sous l'angle de plusieurs caractéristiques.

- leur sexe, variable qualitative susceptible de classement (deux classes ; puisque la variable sexe peut prendre 2 modalités).
- leur département d'origine, variable qualitative susceptible de classement (une centaine de classes) ;
- leur âge, variable quantitative, pouvant prendre un grand nombre de valeurs et susceptible de mesure ;
- leurs diverses notes aux épreuves d'examen, variables quantitatives, résultats d'une mesure, etc.

2- VARIABLES

On définie deux types de variables :

2-1- Variable quantitative : C'est le résultat d'une mesure effectuée sur chaque objet de la série statistique. Elle s'exprime par une valeur numérique. On distingue :

* Variable quantitative discrète

C'est le résultat d'un dénombrement : nombre $\in$ **N** (entre 2 valeurs il y a un nombre fini de valeurs possibles) : **Exemple:** Nombre d'enfants dans la famille.

* Variable quantitative continu

Résultat de la mesure d'une grandeur : nombre $\in$ **R** (entre 2 valeurs il y a une infinité de valeurs possibles)

Exemple: la taille pour un groupe de personnes.

Remarque: une variable quantitative continue peut être exprimée sous la forme d'une variable quantitative discrète. C'est remplacer une échelle élémentaire en une échelle par classes dont chacune sera représentée par sa valeur centrale. On perd en information et on gagne en simplicité.

2-2- Variable qualitative : Elle représente un attribut de l'objet, non susceptible de mesure. Cet attribut ne peut prendre qu'un nombre fini de modalités dites classes ou catégories. C'est le résultat d'un classement. Toutes les classes d'une variable qualitative doivent être conjointement exhaustives (chaque unité de la série statistique doit appartenir à une classe) et mutuellement exclusives (chaque unité de la série statistique ne peut appartenir qu'à une et une seule classe).

- Variable qualitative ordinale : les différentes catégories respectent un certain ordre croissant ou décroissant.

Exemple: âge : 3 classes enfants, adultes, âgés

Remarque: Une variable quantitative peut être exprimée sous la forme d'une variable qualitative ordinale. On perd en information, on gagne en simplicité.

- Variable qualitative nominale : les différentes catégories ne respectent aucun ordre.

Exemple: la couleur des yeux.

- Un cas particulier de la variable qualitative c'est la variable logique ou variable booléenne. C'est une variable qui ne peut prendre que 2 états.

Exemple: décès- non décès, santé- maladie, vrai- faux

VARIABLE QUANTITATIVE : résultat d'une mesure

VARIABLE QUALITATIVE : résultat d'un classement

Nous verrons plus loin que sur les variables quantitatives nous pourrons effectuer des calculs comme ceux de la moyenne, de la variance, de l'écart type... alors que pour les variables qualitatives nous ne pouvons procéder qu'à des dénombrements, c'est-à-dire à la détermination de l'effectif de chacune des classes.

3- POPULATION ET ECHANTILLON

On distingue deux grands types de série statistique :

- population (au sens statistique)
- échantillon.

3-1- La population**,** au sens statistique, est une série statistique exhaustive, c'est-à-dire qu'elle est l'ensemble de tous les objets de même nature que l'on veut étudier.

Une population peut être finie (ex : l'ensemble des tunisiens inscrits sur une liste électorale) ou infinie (ex : l'ensemble des mesures que l'on peut faire d'une grandeur).

3-2- On appelle échantillon un sous-ensemble fini, c'est à dire d'effectif limité, extrait de la population.

Sauf lorsque la population est d'effectif faible, le statisticien travaille sur des échantillons.

3-2-1- Pourquoi faut-il échantillonner ?

On ne peut faire autrement quand la population est potentiellement infinie.

Même quand la population est finie, on est amené à prélever un échantillon pour les raisons suivantes :

- les ressources affectées à l'étude sont limitées.

Ainsi, pour un sondage d'opinion ou pour un sondage électoral, on ne peut envisager d'interroger toutes les personnes dont on veut recueillir l'avis.

- on ne peut attendre de disposer de la population.

Par exemple, si l'on veut tester la qualité de la production d'une nouvelle machine, on prélèvera l'échantillon parmi les premiers objets produits. De même, dans un essai thérapeutique concernant la maladie de Hodgkin, on ne peut attendre de disposer de tous les malades atteints de cette maladie, passés, présents et futurs !

- la détermination de la variable à étudier est destructrice. Par exemple, si l'on étudiait la durée de vie d'un objet manufacturé sur l'ensemble de ces objets, il ne resterait plus rien à vendre.

3-2-2- Comment échantillonner ?

Bien entendu, même si on focalise l'étude sur l'échantillon, celui-ci n'a pas d'intérêt en soi. Ce que l'on veut connaître, c'est la population.

Il convient donc que l'échantillon ressemble le plus possible à la population : ou encore qu'il soit représentatif de la population.

Pour arriver à ce résultat, le moyen le plus simple est de donner à chaque élément de la population une chance égale de faire partie de l'échantillon. Et pour cela, la meilleure technique est celle du tirage au sort. Cette procédure de tirage au sort s'appelle randomisation, et l'échantillon obtenu est dit randomisé.

Le mot randomisation vient de l'Anglais « random » : au hasard (on retrouve cette racine dans le mot français «randonnée»).

La randomisation, procédure d'extraction d'échantillons, est une procédure objective, qui ne laisse aucune place à la subjectivité, ni à l'inspiration, ni au flair. « Au hasard » ne signifie pas « au petit bonheur ». Pour randomiser, on peut utiliser des procédures de type pile ou face, extraction d'une urne bien mélangée (comme la loterie nationale). En pratique, on utilise des tables de nombres au hasard. Ce sont des séquences de chiffres

telles que, pour chaque partie de la séquence, les dix chiffres O, 1, ...8, 9 ont une chance égale d'apparaître.

3-3- Relations entre population et échantillon randomisé

Deux situations bien distinctes peuvent être décrites :

3-3-1- Dans la première situation, peu fréquente en pratique, la population est connue, et on cherche à prévoir ce que sera l'échantillon randomisé que l'on va tirer. « L'événement -c'est-à-dire l'échantillon- est incertain, mais la cause -c'est-à-dire la population- est connue », disait LAPLACE en 1774.

Dire que la population est connue signifie de façon précise que l'on connaît comment sont distribuées dans cette population les variables qualitatives et quantitatives auxquelles on s'intéresse.

Cette situation correspond donc aux paris que l'on peut faire sur la distribution des variables dans l'échantillon. Elle correspond aussi à un processus mental qui va du général au particulier, et qui est donc déductif.

L'outil adapté pour traiter de cette situation est la théorie -ou calcul- des probabilités.

3-3-2- Dans la deuxième situation, à l'inverse, l'échantillon randomisé est connu (c'est-à-dire que l'on connaît comment y est distribuée la variable à laquelle on s'intéresse), et on cherche à deviner ce qu'est la population.

Ceci correspond à la situation la plus usuelle, où « l'événement est connu, mais la cause est inconnue » (LAPLACE).

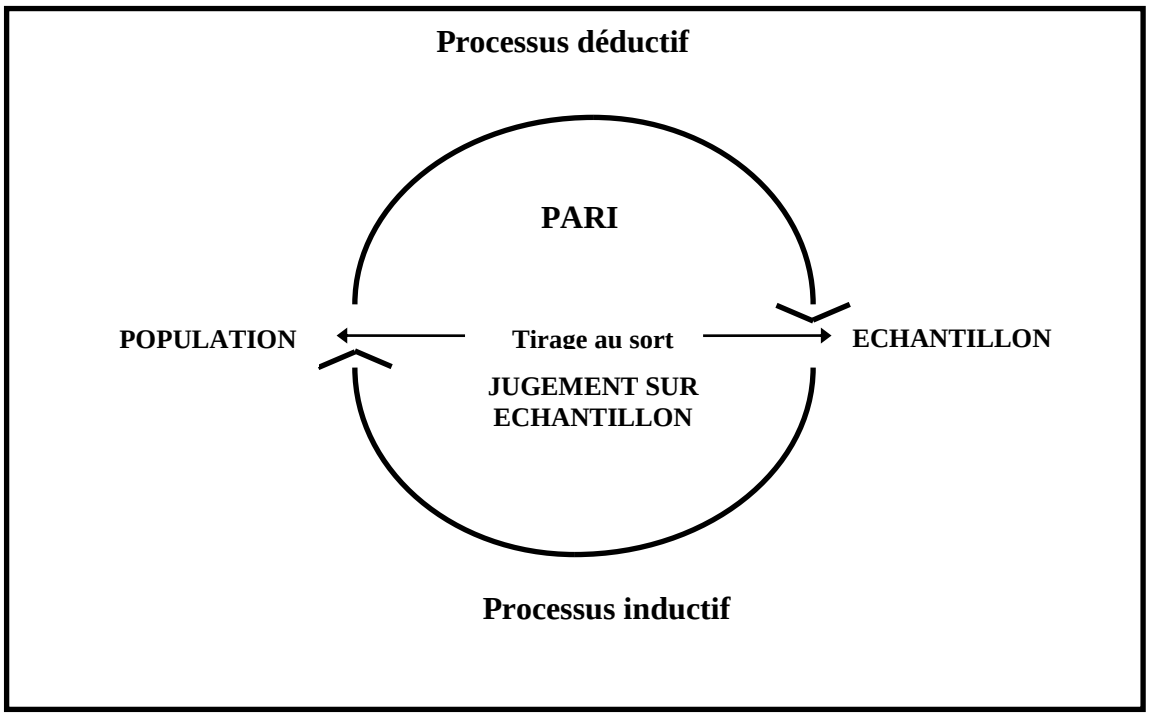
Le problème est donc celui du jugement sur échantillon ou encore d'inférence statistique. On juge (on infère) de ce que doit être la population d'après ce qu'on sait de l'échantillon supposé représentatif.

On procède du particulier au général : il s'agit donc d'un processus inductif.

Les outils de la statistique inductive sont la théorie des tests et la théorie de l'estimation. Nous en verrons des éléments à

propos de cas particuliers dans la suite du cours. Nous verrons qu'elles reposent sur la théorie des probabilités.

Le schéma suivant résume les relations entre <u>population</u> <u>et échantillon :</u>



ELEMENTS DE PROBABILITE

1- NOTION D'EVENEMENT

Soit une épreuve qui peut donner des résultats variables : un jet de dé (résultat de 1 à 6) ; tirage d'une famille de n enfants (résultat de n = 0 à n = 20) ; un malade se présente à un médecin : il peut avoir l'une (ou plusieurs) des maladies figurant dans la classification internationale des maladies (résultats : typhoïde, maladie de Basedow, hypertension artérielle...)

Soit S l'ensemble des résultats possibles, appelés éventualités. On appelle événement A un sous-ensemble de S.

Dans le cas du jet de dé, les éventualités sont 1, 2, 3, 4, 5, 6. Un événement A peut être le sous-ensemble élémentaire {2}, le sous-ensemble {1, 2, 3} ou encore {2, 4, 6} ou encore S = {1, 2, 3, 4, 5, 6}. Un autre événement pourrait être l'ensemble vide 0, à savoir ni 1, ni 2,... ni 6.

Dans le cas du malade qui se présente à un médecin, l'ensemble S est l'ensemble des diagnostics possibles (tels qu'ils figurent dans la classification internationale). A peut être la maladie de Basedow ou les maladies de la thyroïde ou les maladies endocriniennes, etc.

On dit que l'événement A s'est réalisé, si l'une des éventualités qui le constituent s'est réalisée.

AXIOMES ET THEORIE DES PROBABILITES

Les divers événements A peuvent être plus ou moins vraisemblables. Ainsi un malade dont on ne sait rien à priori et qui vient consulter en hivers, a beaucoup plus de chances d'avoir la « grippe » qu'une insolation.

Ce concept de vraisemblance est très usuel, mais est assez flou. Les axiomes et théorèmes de probabilité visent à lui donner de la rigueur et à le rendre opératoire.

On associe à chaque événement A, un nombre p(A) qui obéît aux règles suivantes :

1- Quelque soit A : $0 \leq p(A) \leq 1$

2- $p(S) = 1$

S est l'événement certain : à partir du moment où on fait l'épreuve, on est sûr de voir survenir une de ses éventualités : 1 ou 2... ou 6.

3- $p(\emptyset) = 0$

La probabilité de l'ensemble vide est nulle : il s'agit de l'événement impossible. Si on a réalisé le jet du dé, il est impossible de n'obtenir aucun des six chiffres 1, 2,..6

4- Quels que soient les deux événements A et B :

p(A ou B) = p(A) + p(B) - p (A et B)

Exemple: dans un jet de dé, je prends A = {1, 2, 3} et B = {2, 4, 6}

p(A) = p (B) = ½ p (A ou B) = p{1, 2, 3, 4, 6} = 5/ 6 p(A et B) = p {2} = 1/6

on a bien 5/6 = ½ + ½ - 1/6

Cas particulier : si A et B sont deux événements incompatibles, c'est-à-dire qu'ils ne peuvent se réaliser ensemble, on a p (A et B) = 0. Dans ce cas :

p (A ou B) = p(A) + p (B)

C'est l'axiome dit des probabilités totales, qui s'applique aux événements disjoints.

Exemple: dans un tirage de cartes d'un jeu de 32 cartes :

Proba (Roi ou Dame) = Proba (Roi) + Proba (Dame)

¼ = 1/8 + 1/8

5- Quelque soit A, si on appelle $\overline{A}$ événement complémentaire

(A ou $\overline{A}$ = S ; A et $\overline{A}$ = 0) :

P(A) = 1 - p ($\overline{A}$)

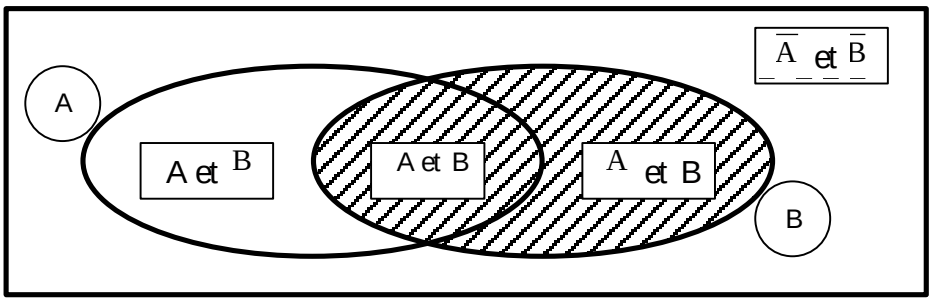
2- DEFINITION DE LA PROBABILITE CONDITIONNELLE

2-1- Soient deux événements A et B compatibles, c'est à dire qu'il peuvent se réaliser ensemble. On appelle probabilité conditionnelle et on note « p(A/B) » la probabilité que A se réalise si B est réalisée. P(A/B) se lit « Proba de A quand B » ou « Proba de A si B ».

Par définition, on pose : $p(A/B) = \frac{p(AetB)}{p(B)} \quad p(B) \neq 0$

Remarques:

2-1-1- Admettre que B est réalisé revient à diminuer le référentiel. Au lieu de se mouvoir dans le référentiel total où les quatre éventualités ($\overline{A}$ et, $\overline{A}$ et B, A et, A et B) sont possibles (c'est à dire tout le rectangle), on est réduit aux deux éventualités $\overline{A}$ et B, A et B (c'est à dire la zone hachurée).



2-1-2- On a :

p(A/B) + p($\overline{A}$/B) = 1 puisque :

$$p(A/B) + p(\overline{A}/B) = \frac{p(AetB)}{p(B)} + \frac{p(\overline{A}etB)}{p(B)}$$

$$= \frac{p(AetB) + p(\overline{A}etB)}{p(B)}$$

$$= \frac{p(B)}{p(B)} = 1 \quad p(B) \neq 0$$

2-2- Indépendance des deux événements A et B

Par définition, deux événements sont indépendants, si la probabilité de l'un n'est pas modifiée si on connaît l'existence de l'autre.

p(A/B) = p(A/$\overline{B}$) =p(A) et de même p(B /A) = p(B /$\overline{A}$) =p(B)

Dans ce cas p(A et B) = p(A) x p(B) (Théorème des probabilités composées qui s'applique aux événements indépendants).

Exemple : Dans un tirage de cartes, si A est « tirer un as », et B est « tirer un cœur » : p(as cœur) = p(as) x p(cœur) = 1/8 x 1/4 = 1/32

NB. : Si A et B sont indépendants, il en est de même de $\overline{A}$ et.

2-3- Théorème de BAYES

La théorie des probabilités peut être très utile dans la progression vers un diagnostic au cours de l'examen clinique. En effet, examiner un malade, c'est s'efforcer d'accumuler progressivement des preuves pour se rapprocher de la certitude d'un diagnostic. Autrement dit pour en faire augmenter la probabilité en la rapprochant le plus possible de 1.

Imaginons qu'à un moment donné d'un examen, la probabilité d'un certain diagnostic soit p (D). On recherche un nouveau signe S qui

est susceptible, s'il est présent, de modifier la probabilité du diagnostic.

Autrement dit : p(D/S) ≠ p(D)

probabilité à priori ≠ probabilité à posteriori

Il est facile de voir que :

$$p(D/S) = \frac{p(D) \times p(S/D)}{p(S)} \quad P(S) \neq 0$$

C'est le théorème de BAYES.

En effet, on peut écrire de deux façons la probabilité d'avoir à la fois D et S p(D et S) = p(D) x p(S/D) = p(S) x p(D/S).

Une formulation plus simplifiée consiste à dire que la probabilité à posteriori p(D/S) varie comme le produit de la probabilité à priori p(D) et de la probabilité du signe dans la maladie p(S/D). :

p(D/S) # p(D) x p(S/D) (# : proportionnel)

Une formulation plus complète et aussi plus compliquée explicite le terme p(S). On peut écrire :

S = (S et D) ou (S et $\overline{D}$)

d'où p(S) = p(S et D) + p(S et $\overline{D}$) = p(D) x p(S/D) + p($\overline{D}$) x p(S/$\overline{D}$)

La forme complète du théorème de BAYES s'écrit donc

$$p(D/S) = \frac{p(D) \times p(S/D)}{p(D) \times p(S/D) + p(\overline{D}) \times p(S/\overline{D})}$$

COMMENT ESTIMER LA PROBABILITE D'UN EVENEMENT

Deux approches sont possibles :

■ Une approche objective :

- par le raisonnement, au moins dans des cas simples (ex : des cartes à jouer) ;

- par l'expérience : on observe la fréquence de l'événement, lors d'épreuves répétitives, et on verra que cette fréquence observée permet une estimation de la probabilité.

L'expérience permet parfois de corriger un raisonnement trop simpliste. Ainsi, un modèle simple de la fécondation ferait conclure à un taux de masculinité (proportion de mâles parmi les naissances) de 0,5. L'expérience montre qu'il vaut en fait 0,513.

■ Une approche subjective :

Il s'agit là d'un degré de croyance. Il faut se contenter de cela lorsqu'on ne peut répéter l'épreuve. C'est le cas des paris sportifs ou des paris sur les courses de chevaux.

En médecine, on fait souvent appel à de tels degrés de certitude : « je suis sûr à 5 contre 1 que ce malade a une appendicite ».

3- LOI DE PROBABILITE D'UNE VARIABLE ALEATOIRE QUANTITATIVE CONTINUE

3-1- Variable aléatoire

Le résultat d'une épreuve telle que celle que nous avons décrite est appelé variable aléatoire

3-1-1- Nous avons vu des cas où cette variable aléatoire est qualitative : résultat d'un jet de pièces de monnaie (pile ou face) ; résultat d'un tirage de cartes (as, roi, dame,...) ; résultat de ce tirage qu'est une consultation (une des maladies possibles) etc.

3-1-2- Mais il est aussi des cas où la variable aléatoire est un nombre entier ; (variable quantitative discrète) tel est le cas du nombre d'enfants d'une famille tirée au hasard ou encore le résultat d'un jet de dé...

C'est par exemple aussi le cas où l'épreuve envisagée est une suite d'épreuves élémentaires. Par exemple, une suite de 100 jets de pièce de monnaie. Le nombre de fois où on tire pile est une variable aléatoire entière, qui peut prendre toutes les valeurs entre 0 et 100.

De même, si on tire un échantillon de 200 familles, le nombre de fois où on tire une famille de 5 enfants : cette variable peut prendre toutes les valeurs entières entre 0 et 200.

3-1-3- Enfin, la variable aléatoire X peut être quantitative et continue ; ainsi en est-il de la plupart des dosages et mesures en biologie.

Loi de probabilité d'une variable aléatoire.

Affecter une probabilité à chacune des valeurs possibles d'une variable aléatoire, c'est la doter d'une loi de probabilité.

Ceci ne pose aucun problème théorique lorsque la variable aléatoire est discrète. Ainsi, dans un jet de dé non pipé, la probabilité de sortir une des faces quelconques est de 1/6. Dans un tirage de cartes bien mélangées et non biseautées, la probabilité de sortir un cœur est de 1/4.

La chose est en principe plus difficile quand la variable aléatoire X est continue. En effet, la probabilité d'une valeur particulière est nulle, de la même façon que le choix d'un point particulier sur une droite a une probabilité nulle. Autrement dit, Proba (X = x) = 0, si x est un nombre réel quelconque.

Par contre, si l'événement auquel on s'intéresse est un intervalle, par exemple (a < x < b), alors la probabilité peut être définie.

Aussi les lois de probabilité pour les variables continues sont-elles définies comme affectées à des segments. On écrit :

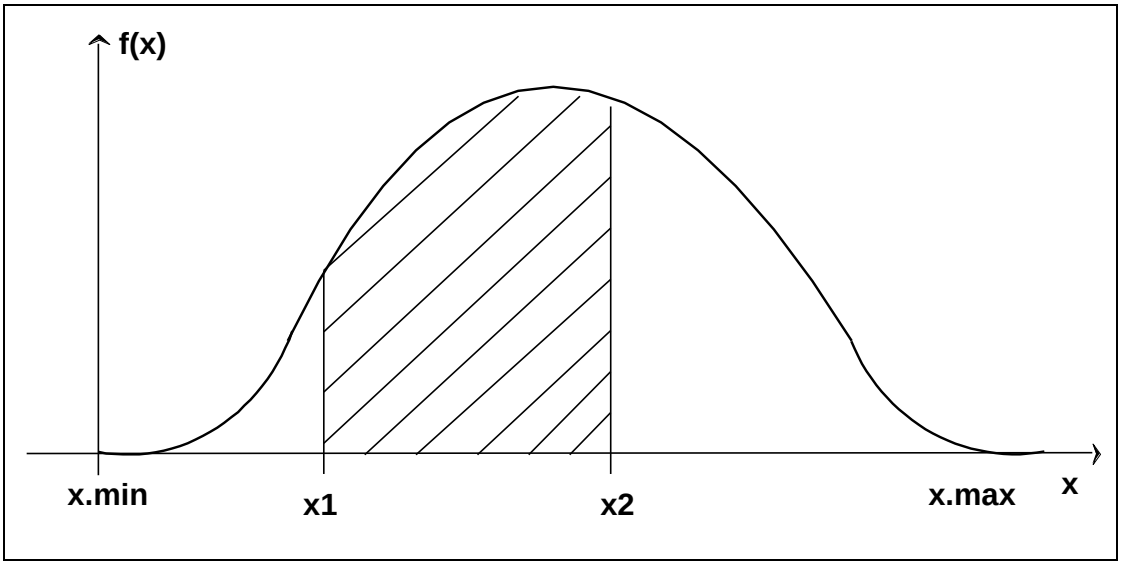
$$\text{Proba}\,(x \leq x < x + dx) = f(x)\,dx$$

f(x) s'appelle la densité de probabilité de X au point x. Autrement dit, l'événement élémentaire est ici l'intervalle infiniment petit, mais non nul :

(x, x + dx).

Bien entendu, si on additionne les probabilités de tous ces événements élémentaires, étendues à tout le domaine de variation possible de X (de x.min à x.max), on obtient l'unité.

On peut représenter graphiquement la densité de probabilité f(x) en fonction de x.



La surface hachurée représente Proba (x1 ≤ X < x2).(1).

La surface totale comprise entre la courbe et l'axe des x vaut par définition l'unité.

3-2- ***Un cas particulier très important*** est la **loi dite de LAPLACE-GAUSS** (dite encore « loi normale », expression impropre qu'il vaut mieux éviter), dans la mesure où cela voudrait dire que les autres lois de probabilité sont « anormales », voire pathologiques »). La densité de probabilité de la loi de LAPLACE-GAUSS a la forme suivante

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

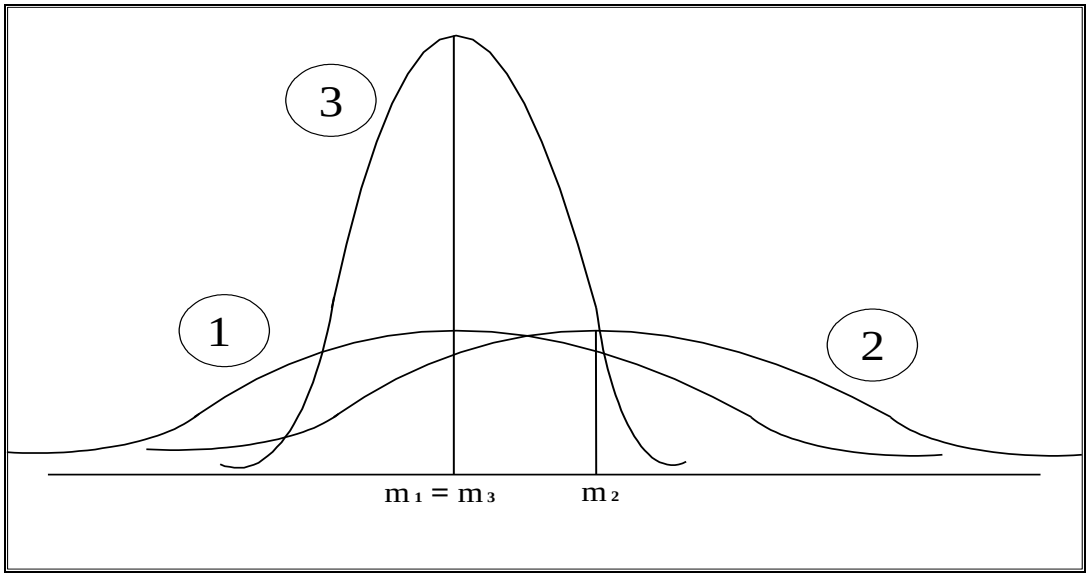
Cette forme nous montre qu'elle ne dépend que des deux paramètres μ et σ.

On parle souvent de loi L.G. (μ , σ) ou N (μ , σ).

μ est un nombre réel quelconque σ est un nombre réel strictement positif.

Graphiquement, la courbe qui représente la loi de L.G. a une forme de cloche (mais la réciproque n'est pas vraie, toute distribution en forme de cloche n'obéit pas forcément à la loi de L.G.). Le domaine de variation de la variable aléatoire X va de $-\infty$ à $+\infty$. Le paramètre μ est une valeur centrale, par rapport à laquelle la courbe est symétrique. Le paramètre σ est lié directement à la largeur de la cloche : plus σ est grand, plus la cloche est large.

Le graphique suivant représente trois distributions de L.G. :



° Les distributions 1 et 2 ont même largeur car $\sigma_1 = \sigma_2$

° Par contre, les valeurs centrales diffèrent : $\mu_1 < \mu_2$

° Les courbes 1 et 3 ont même valeur centrale $\mu_1 = \mu_3$

° mais des largeurs différentes : $\sigma_3 < \sigma_1$

Dans les trois cas, l'aire comprise entre l'axe des x et la courbe est égale à l'unité.

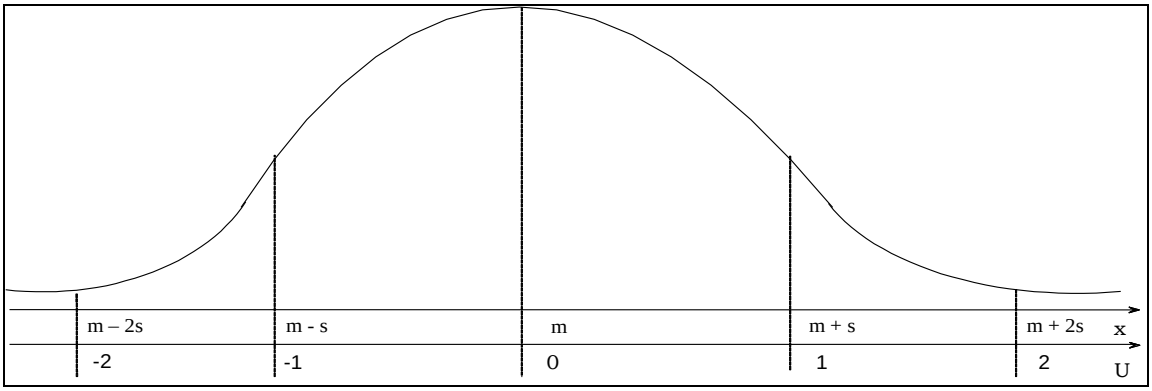
LOI DE LAPLACE-GAUSS CENTREE ET REDUITE : Il y a autant de lois de L.G. que de couples (μ,σ), c'est-à-dire une double infinité. Il est toutefois possible de ramener toutes ces lois à une seule, dite loi de L.G. centrée réduite.

Introduisons en effet une variable U reliée à la variable X par la relation :

$$U = \frac{x - \mu}{\sigma} \qquad x = \mu + U.\ \sigma$$

Ce changement de variable revient à prendre la valeur 0 pour valeur centrale (si $x = \mu$, U = 0), et la valeur l pour le paramètre σ (si $x = \mu + \sigma$, U = 1 et si $x = \mu - \sigma$, U = - 1)

Le graphique suivant illustre la correspondance entre une loi de L.G. quelconque et la loi centrée et réduite qu'on en déduit.



REMARQUE: les deux points $\mu - \sigma$ et $\mu + \sigma$ correspondent aux deux points d'inflexion de la courbe (points où celle-ci traverse sa tangente).

TABULATION DE LA LOI CENTREE REDUITE DE L.G.

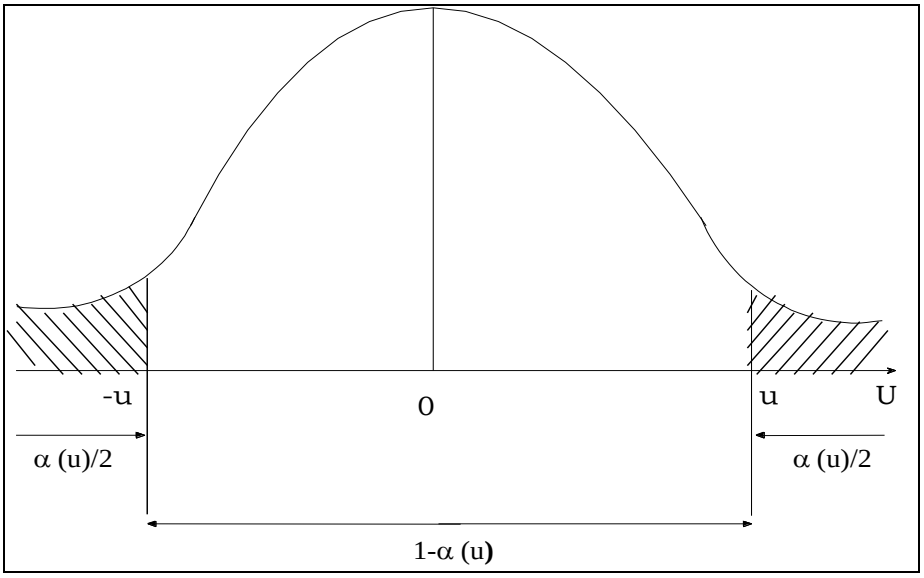
L'intérêt de cette transformation réside dans le fait que la loi de L.G. centrée réduite a été tabulée ; la table en est donnée, et il faut savoir la lire sans l'ombre d'une hésitation.

Cette table met en concordance, d'une part une série de valeurs positives u que peut prendre la variable aléatoire U, et d'autre part la probabilité α (u), telle que :

α (u) = Proba (U < -u ou U > u) = Proba ($|U|$ > u)

* **N.B. :** u est un nombre positif

Graphiquement, à chaque valeur de u, la table α (u) donne l'aire hachurée double représentée ci-dessous :



Par symétrie, chacune des deux aires vaut α (u)/2.

La lecture de la table se fait de la façon suivante : à chaque valeur de u lue en un endroit quelconque du tableau, la probabilité (u) s'obtient en sommant les valeurs marginales de la ligne et de la colonne correspondantes.

Ainsi, pour

U = 1 (en fait 0,994458) α (u) = 0,3 + 0,02 = 0,32

U = 2 (en fait 1,959964) α (u) = 0,0 + 0,05 = 0,05

U = 2,58 (en fait 2,575829) α (u) = 0,0 + 0,10 = 0,01

De même, dans le bas du tableau, on lit :

pour α (u) = 0,001 U = 3,29

Trois couples de valeurs doivent être très bien connus :

$\alpha = 0{,}05$	**u = 1,96 (ou à la rigueur 2)**
$\alpha = 0{,}01$	**u = 2,58 (ou à la rigueur 2,6)**
$\alpha = 0{,}001$	**u = 3,29 (ou à la rigueur 3,3)**

Car ils correspondent aux probabilités utilisées conventionnellement dans les tests statistiques.

Bien utilisée, cette table peut donner d'autres renseignements. Ainsi, on peut calculer :

3-2-1- probabilité ($-u \leq U \leq u$) = 1 - α (u)

Par exemple, Proba ($-2 \leq U \leq +2$) $\approx$ 1 - 0,05 = 0,95

Proba ($-1 \leq U \leq +1$) $\approx$ 1 - 0,32 = 0,68

3-2-2- Probabilité (U > u) = $\frac{\alpha(u)}{2}$

Par exemple, Proba (U > 2) $\approx$ 0,025

De même, en lisant la table convenablement (2ème ligne, 1ère colonne), on trouve : Proba (U > 1, 645) = $\frac{0{,}10}{2}$ = 0,05

3-2-3- Proba (U < u) = 1 - $\frac{\alpha(u)}{2}$

Ainsi, Proba (U < 1,645) = 0,95

3-2-4- On peut aussi calculer des probabilités telles que Proba ($u_1 < U < u_2$)

Proba ($u_1 < U < u_2$) = $\frac{\alpha(u_1) - \alpha(u_2)}{2}$ pour u_1 et u_2 positifs

Cherchons par exemple

Proba (1 < U < 2) = $\frac{0.32 - 0.05}{2}$ = 0,135

STATISTIQUE DESCRIPTIVE

Nous distinguerons dans ce chapitre :

- La description des populations, ensembles vastes, variés, infinis, de toutes façons souvent impossibles à étudier de façon exhaustive. Elles seront caractérisées par une distribution THEORIQUE de probabilité, et les paramètres qui apparaîtront dans ces lois de probabilité seront appelés paramètres THEORIQUES.

- La description des échantillons, ensembles de taille limitée, caractérisés par un effectif N Effectivement OBSERVE. La distribution des variables y sera dite distribution OBSERVEE, et les paramètres seront des paramètres OBSERVES (ou empiriques).

Nous ferons aussi la distinction entre les variables quantitatives et les variables qualitatives.

1- 1er CAS : UNE VARIABLE QUALITATIVE A K CATEGORIES

Exemple: l'état civil, qui est une variable à 4 catégories : célibataire, marié, veuf, divorcé. Les 4 catégories sont exclusives (on ne peut appartenir à deux catégories à la fois), et exhaustives (il n'y a pas d'autres catégories possibles).

1-1- Dans la POPULATION, chaque catégorie d'une variable est caractérisée par une probabilité pi. Celles-ci sont telles que :

$$(1) \qquad p_1 + p_2 + \ldots\ldots + p_k = \sum p_i = 1$$

Cette égalité traduit l'exclusivité (possibilité d'additionner les probabilités) et l'exhaustivité (la somme vaut 1).

Les probabilités p_i s'appellent aussi fréquences THEORIQUES.

1-2- Dans un ECHANTILLON observé dont l'effectif est N, on classe les objets dans chacune des classes et on dénombre les effectifs n_i de chacune. On a :

$$(2) \qquad n_1 + n_2 = ... + n_k = \sum n_i = N$$

On considère souvent, pour chaque classe, la <u>fréquence</u> relative <u>OBSERVEE</u> : c'est la quantité $\frac{n_i}{N} = f_i$

En divisant la relation (2) par N, on obtient :

$$(3) \qquad f_1 + f_2 + ... + f_k = \sum f_i = 1$$

Comme il est évident que $0 \leq f_i \leq 1$, on voit que cette relation (3) est formellement très proche de la relation (1).

2- 2ème CAS : ETUDE SIMULTANEE DE DEUX VARIABLES QUALITATIVES :

<u>Par exemple</u> : répartition simultanée selon le sexe (1ère variable à deux catégories) et l'état -civil (2ème variable à 4 catégories).

On peut, bien entendu, se ramener à une variable composée à 2 x 4 = 8 catégories. Il est commode de représenter celle-ci sous forme d'un tableau rectangulaire à 8 cases.

2-1- <u>Dans la POPULATION</u>, à chaque case du tableau, correspond une probabilité ou fréquence <u>THEORIQUE</u>. Dans l'exemple précédent, on obtient :

	C	**M**	**V**	**D**	**TOTAL**
H	p_{11}	p_{12}	p_{13}	p_{14}	$\mathbf{p_{1*}}$
F	p_{21}	p_{22}	p_{23}	p_{24}	$\mathbf{p_{2*}}$
TOTAL	$\mathbf{p_{*1}}$	$\mathbf{p_{*2}}$	$\mathbf{p_{*3}}$	$\mathbf{p_{*4}}$	**1**

Chaque probabilité est affectée d'un double indice qui indique la ligne puis la colonne à laquelle elle appartient.

En fin de ligne et de colonne, on a écrit la <u>probabilité marginale</u>.

Ainsi: $p_{1*} = p_{11} + p_{12} + p_{13} + p_{14}$ représente la probabilité d'être un homme, quel que soit l'état civil.

De même : $p_{*3} = p_{13} + p_{23}$ est la probabilité d'être veuf, quel que soit le sexe.

Bien évidemment, l'exclusivité et l'exhaustivité des catégories permettent d'écrire :

$$(4) \qquad p_{11} + p_{12} + \ldots + p_{24} = \sum_i \sum_j p_{ij} = 1$$

2-2- Dans un ECHANTILLON OBSERVE d'effectif N, on effectue de même la répartition des N individus entre les 4 x 2 cases (dans le cas général, k_1 x k_2 cases, où k_1 et k_2 sont le nombre de catégories des deux variables). On obtient ainsi un tableau que l'on appelle **tableau de contingence**, où, dans chaque case, est inscrit l'effectif OBSERVE correspondant.

	C	**M**	**V**	**D**	**TOTAL**
H	n_{11}	n_{12}	n_{13}	n_{14}	$\mathbf{n_{1*}}$
F	n_{21}	n_{22}	n_{23}	n_{24}	$\mathbf{n_{2*}}$
TOTAL	$\mathbf{n_{*1}}$	$\mathbf{n_{*2}}$	$\mathbf{n_{*3}}$	$\mathbf{n_{*4}}$	**N**

En fin de ligne ou de colonne, sont les effectifs marginaux OBSERVES. Bien évidemment : $$(5) \qquad n_{11} + n_{12} + \ldots + n_{24} = \sum_i \sum_j n_{ij} = N$$

On pourrait, en divisant tous les effectifs par N, faire apparaître un tableau des fréquences observées. On ne le fait pas de façon habituelle. En effet, le test que l'on effectue pour analyser ce type de tableau utilise les effectifs et non pas les fréquences.

3- 3ème CAS : VARIABLE QUANTITATIVE

3-1- DANS LA POPULATION Exemple: nombre d'enfants d'une famille ; nombre de globules rouges dans une case d'hématimètre au cours d'une numération globulaire.

Ce cas n'est pas fondamentalement distinct du suivant qui traite d'une variable quantitative continue, mais son expression mathématique est plus simple. Il s'agit donc simplement d'une approche du cas suivant. Chaque valeur possible x_i de la variable aléatoire x a une probabilité p_i avec :

$$(6) \qquad p_1 + p_2 + \ldots = \sum p_i = 1$$

On définit (ce qui n'était pas possible dans le cas d'une variable aléatoire qualitative) et on utilise habituellement 3 paramètres pour caractériser cette distribution.

La moyenne THEORIQUE μ appelée encore « espérance de X » : E(X)

$$(7) \qquad \mu = E(X) = x_1 p_1 + x_2 p_2 + \ldots = \sum x_i p_i$$

Chaque valeur possible x_i est pondérée par sa probabilité p_i. La moyenne indique une <u>valeur centrale</u> de la distribution.

La variance THEORIQUE DE X, que l'on écrit V(X) ou encore σ^2 (c'est toujours une valeur positive).

$$(8) \qquad V(X) = \sigma_x^2 = (x_1 - \mu)^2 p_1 + (x_2 - \mu)^2 p_2 + \ldots = \sum (x_i - \mu)^2 p_i$$

La variance est la somme des carrés des écarts à la moyenne $(x_i - \mu)^2$ pondérés par les probabilités respectives p_i c'est <u>un indice de dispersion autour de la moyenne</u>. Si les écarts les plus grands ont une forte probabilité, la variance est grande. La variance est toujours positive.

On voit que, sur le plan dimensionnel, la variance n'est pas homogène à la variable d'origine, mais à son <u>carré</u>. Aussi définit-on souvent.

L'écart type THEORIQUE σ, qui n'est autre que la racine carrée de la variance :

$$(9) \qquad \sigma_x = \sqrt{V(x)}$$

et qui est homogène à la variable d'origine.

Ainsi, une distribution de revenus imprimée en dinars aura-t-elle un écart type exprimé en dinars D, alors que la variance était exprimée en « Dinars-carré » D².

3-2- DANS L'ECHANTILLON observé d'effectif N, la distinction entre la variable de nature discrète ou de nature continue n'est plus pertinente, car on se trouve toujours en présence d'une suite discontinue de N valeurs.

On définit, de façon comparable au cas de la population :

Une moyenne OBSERVEE mx (ou encore $\bar{x}$)

$$m_x = \frac{x_1 + x_2 + \ldots\ldots x_n}{N} = \frac{\sum x_i}{N} \qquad (7bis)$$

C'est une valeur centrale de la distribution observée.

Une variance OBSERVEE Sx².

$$S_x^2 = \frac{\sum (x_i - m_x)^2}{N-1} \qquad (8bis)$$

Un écart type OBSERVE Sx

$$S_x = \sqrt{\frac{\sum (x_i - m_x)^2}{N-1}} \qquad (9bis)$$

En pratique

$$S^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N-1}$$

REPRESENTATION GRAPHIQUE DE L'INFORMATION

1- CAS D'UNE VARIABLE QUANTITATIVE : histogramme et polygone de fréquence.

1-1- HISTOGRAMME

Après avoir transformé l'échelle élémentaire en échelle par classes, on peut représenter graphiquement la distribution de fréquence de la variable en question. Pour cela, on utilise un système de coordonnées rectangulaires. Sur ce système, l'échelle de classification est disposée suivant l'axe horizontal (axe des X) alors que les fréquences (nombre, %) sont disposées sur l'axe vertical. Ce principe nous permet de construire des histogrammes : un histogramme est un graphique servant à représenter les distributions des fréquences. Il est constitué d'un ensemble de rectangles adjacents, dont chacune des bases coïncide avec un intervalle de classe et chacune des surfaces mesure la fréquence de la classe correspondance. Si les intervalles de classe (donc la largeur des rectangles) sont tout égaux, alors la hauteur du rectangle mesure ainsi la fréquence. Si les intervalles ne sont pas tous égaux, alors seule l'aire du rectangle mesure la fréquence.

1-2- POLYGONE DE FREQUENCE

Le polygone de fréquence est aussi une représentation graphique d'une variable continue. Ce polygone est obtenu à partir de l'histogramme en rejoignant le point milieu du sommet de chaque rectangle au milieu de sommet d'un rectangle adjacent.

Le polygone de fréquence doit être construit de façon à ce que l'aire comprise sous le polygone soit approximativement égale à l'aire de l'histogramme. Il faut par ailleurs que la courbe du polygone se ferme sur l'axe horizontal.

Cette représentation des données permet de mieux mettre en évidence le caractère de continuité ; en outre plusieurs distributions de fréquence peuvent être présentées sans trop surcharger le graphique.

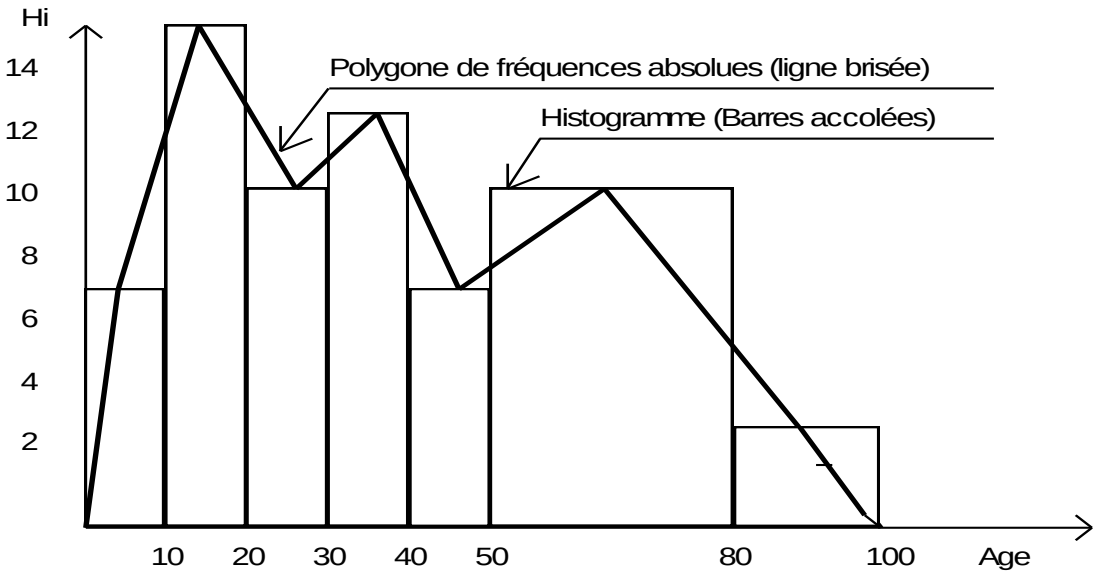
Exemple : dans une population donnée, on a pour l'âge la distribution suivante :

AGE	**FREQUENCE**	**INTERVALLE DE CLASSE**	**HI**
0-9	75	10	7,5
10-19	150	10	15
20-29	100	10	10
30-39	125	10	12,5
40-49	75	10	7,5
50-79	300	30	10
80-99	50	20	2,5
100 et plus	0	Indéterminé	

Pour construire l'histogramme, on doit déterminer les hauteurs des différents rectangles et les bases correspondant aux intervalles de classes. Comme on est tenu par le fait que ce sont les aires qui sont proportionnelles aux fréquences la hauteur de chaque rectangle va être calculée par la formule :

$$\boxed{hi = \frac{fi}{i}}$$

fi = fréquence de la classe
i = intervalle de la classe



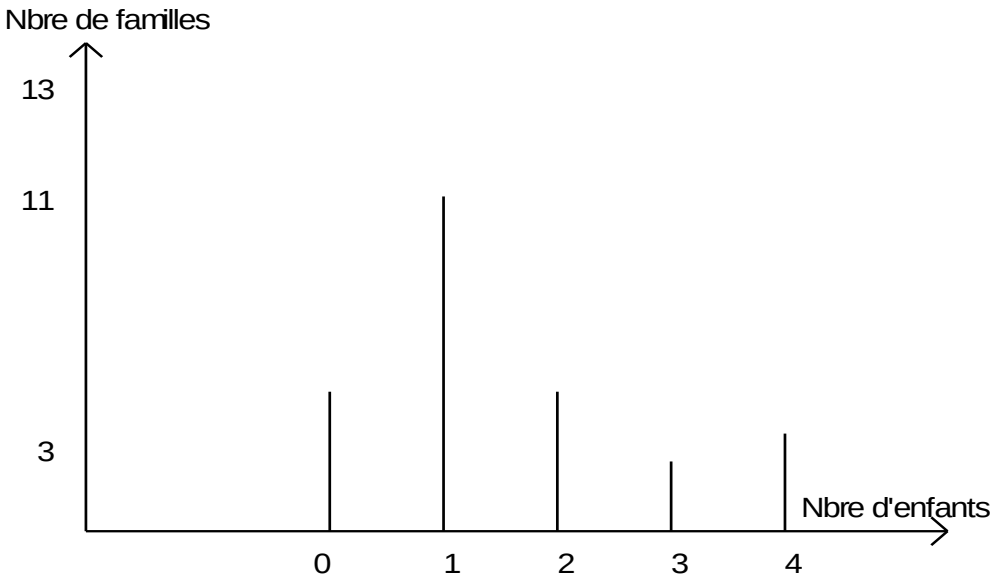
Pour respecter le caractère de continuité de la variable, les intervalles de classe sur l'échelle ne doivent pas être séparés.

2- CAS D'UNE VARIABLE QUANTITATIVE DISCONTINUE : diagramme en bâtonnets

Si les données sont non groupées, le mode de représentation de ce type de variable est le diagramme en bâtonnets. Pour chaque valeur de la variable, on construit un bâtonnet dont la longueur mesure la fréquence.

Exemple : distribution de fréquence du nombre d'enfants par famille dans une population de 30 familles.

Nombre de familles	**Nombre d'enfants/famille**
6	0
11	1
6	2
3	3
4	4

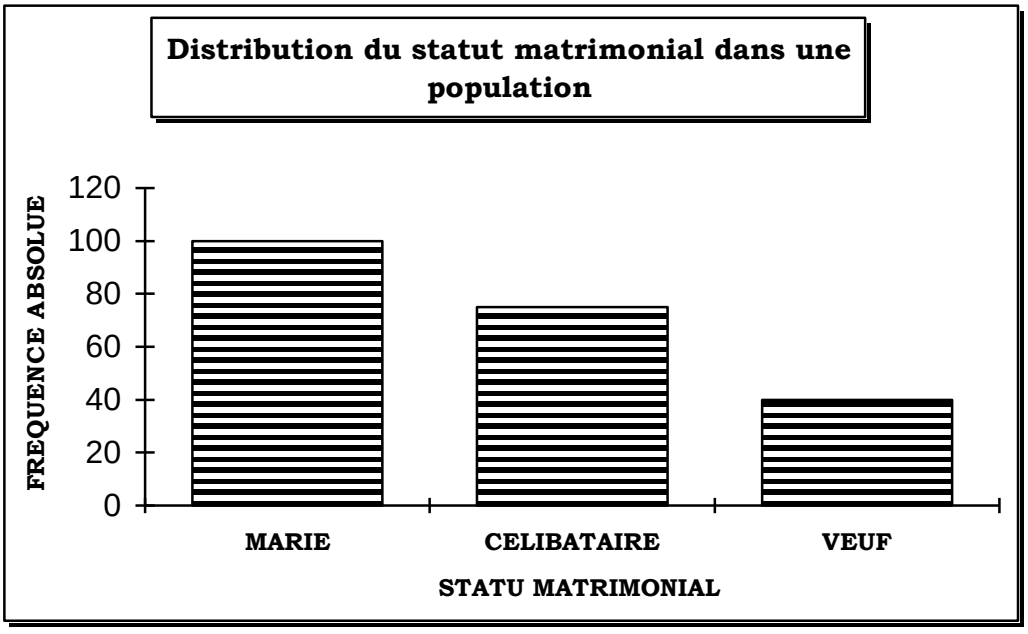


3- CAS D'UNE VARIABLE QUALITATIVE : diagramme en barres et diagramme en cercle.

3-1- DIAGRAMME EN BARRES

Chaque classe de la variable a une barre horizontale ou verticale correspondante dont la longueur mesure la fréquence (relative ou absolue). Contrairement à l'histogramme, les barres de ce diagramme ont toutes la même largeur et sont généralement non adjacentes.

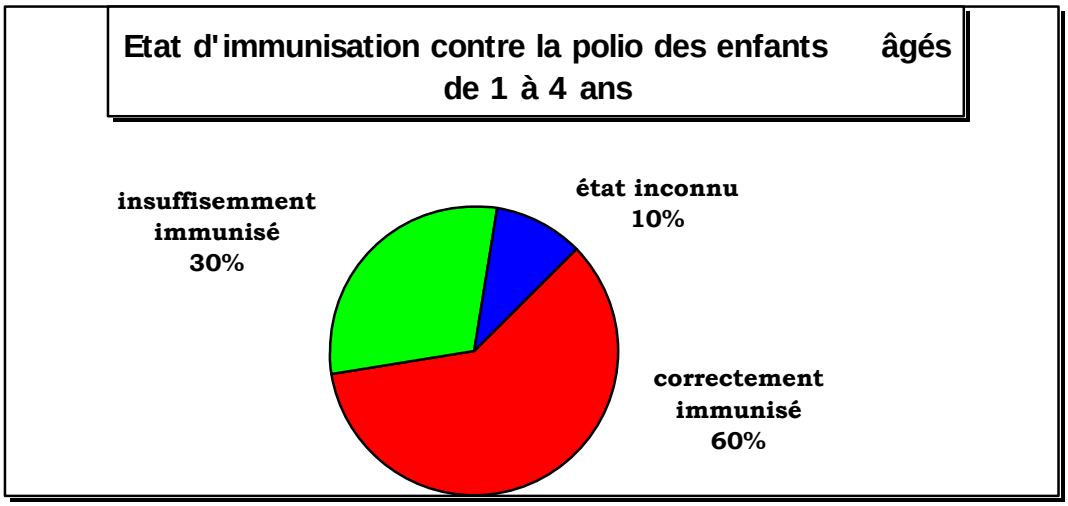
Exemple : distribution du statut matrimonial dans une population



3-2- DIAGRAMME EN CERCLE

Le principe de représentation consiste à diviser un cercle en secteurs proportionnels aux fréquences de classe de la variable. La division se fait habituellement en partant de la position "MIDI" pour disposer dans le sens horaire en ordre décroissant, les secteurs représentant les classes.

Exemple : état d'immunisation contre la polio des enfants âgés de 1 à 4 ans dans la ville de x. en 1969.



STATISTIQUE INFERENTIELLE
VARIABLES QUALITATIVES

1- LES TROIS PROBLEMES PRINCIPAUX A RESOUDRE

Nous nous plaçons dans la perspective de la **statistique inférentielle**. C'est-à-dire que nous disposons d'un échantillon observé et d'une population non ou mal connue: **que peut nous apprendre le premier sur la seconde ?**

De façon plus précise, dans ce chapitre, nous nous intéresserons à la fréquence d'un certain caractère. On observe une certaine valeur f de la fréquence dans l'échantillon. Que pouvons-nous savoir de la fréquence théorique P dans la population? Ce problème s'appelle **la comparaison d'une fréquence observée f à une fréquence théorique p**.

Deuxième problème: je dispose d'un même échantillon de 100 boules avec f = 21%. Mais je ne sais rien sur la population, c'est-à-dire que je ne fais aucune hypothèse a priori sur la valeur de p. Comment à partir de f, puis-je estimer au mieux ce que vaut p. Ce problème s'appelle **estimation de la fréquence théorique p**.

Troisième problème: je dispose de 2 échantillons, l'un de 100 boules avec f_1 = 21%; l'autre de 50 boules avec f_2 = 34%.

Peut-on accepter l'idée ou l'hypothèse que ces deux échantillons proviennent par tirage au sort d'une même urne (de composition d'ailleurs inconnue). Autrement dit, la différence entre f_1 et f_2 est-elle significative, c'est-à-dire non explicable par le seul hasard d'échantillonnage.

Ce problème s'appelle la **comparaison de deux fréquences observées**.

Quelques urnes médicales:

- une formule sanguine faite sur une lame après dénombrement de 200 leucocytes donne 75% de polynucléaires. C'est un échantillon tiré d'une population qui est l'ensemble de tous les leucocytes sanguins. Quelle est la proportion de polynucléaires dans le sang? **(Problème N°2).**

- une race de souris est atteinte de cancer spontané dans une proportion de 80% (population). Dans un échantillon de 100 souris traitées par une drogue antimitotique, la proportion de cancers est de 72%. Ce traitement a-t-il un effet? **(Problème N°1).**

- un échantillon de 1000 infarctus traités par anticoagulants font une récidive dans 15% des cas. Un autre échantillon de 2000 traités par placebo font une récidive dans 19% des cas. Les anticoagulants sont-ils des agents de prévention efficaces? **(Problème N°3).**

Nous allons maintenant reprendre les 3 problèmes de façon plus formelle et plus précise.

1-1- Comparaison d'une fréquence observée avec une fréquence théorique:

Le concept à la base de ce problème est le suivant: bien que l'on connaisse mal la population, on fait **l'hypothèse Ho** que, dans la population, le caractère a une certaine fréquence P (cette hypothèse peut être fondée sur des considérations théoriques ou sur l'opinion d'un expert). Dans ces conditions, l'échantillon, supposé randomisé, est-il compatible avec cette hypothèse? Autrement dit, la fréquence observée f est-elle suffisamment proche de P pour que cette hypothèse puisse être acceptée? La réponse à cette question est binaire:

- ou bien on répond OUI parce que la différence (f - P) semble

assez petite pour apparaître comme le fait du hasard; dans ce cas, on accepte l'hypothèse Ho;

- ou bien on répond NON parce que la différence (f - P)

apparaît trop grande pour n'être que le fait du hasard. On dit alors que cette différence est significative. Dans ce cas, on rejette l'hypothèse Ho, et on accepte une hypothèse alternative H_1. Celle-ci s'exprime ainsi: dans la population, la fréquence théorique est P différente de f. Dans cette procédure, on dit que l'on a testé l'hypothèse Ho.

Ce test d'hypothèse peut se résumer de la façon suivante:

Ho: cette hypothèse s'exprime ainsi: Dans la population, d'où l'échantillon a été tiré au hasard, la fréquence du caractère étudié a une valeur définie P, c'est-à-dire f = P. Donc, par définition, **l'hypothèse nulle stipule que l'échantillon provient de la population et il est randomisé.**

H_1: Hypothèse alternative. Elle est le complémentaire logique de la précédente: l'échantillon n'a pas été tiré au hasard d'une population P où la fréquence est P. Ceci peut se ramener à deux sous hypothèses non forcément exclusives: $f \neq P$ ou tirage biaisé, c'est-à-dire non randomisé. Donc, par définition, **l'hypothèse alternative stipule que l'échantillon ne provient pas de la population et/ou il n'est pas randomisé.**

On voit que l'hypothèse alternative est beaucoup moins précise que l'hypothèse nulle. Elle concerne la population et/ou le mode de tirage.

1-2- Estimation d'une fréquence théorique à partir d'une fréquence observée

Ici, la problématique est légèrement différente: on ne sait rien, a priori, sur la valeur de P, et on n'a aucun doute sur le caractère randomisé de l'échantillonnage.

Alors, que peut-on dire de la vraie valeur de P dans la population, c'est-à-dire comment **estimer** P?

1-3- Comparaison de deux ou plusieurs fréquences observées

Ici, on dispose de deux (ou plusieurs) échantillons. Dans chacun, on a observé le caractère avec une fréquence f_i. Ces fréquences sont le plus souvent différentes.

La question est de savoir si les différences sont suffisamment petites pour que l'on puisse admettre que ces différents échantillons sont extraits d'une même population.

Ici encore, on va confronter deux hypothèses.

Ho (hypothèse nulle): les différents échantillons ont été tirés de façon randomisée d'une même population.

H_1 (hypothèse alternative): les différents échantillons n'ont pas été tirés au hasard d'une même population. Par exemple, chacun a été tiré d'une population différente, ou bien, ils ont été tirés d'une même population, mais de façon non randomisée, c'est-à-dire avec des biais.

Ici, encore, c'est un test d'hypothèse. On teste l'hypothèse nulle Ho, très précise, contre une hypothèse alternative plus floue.

Deux réponses sont possibles:

- on accepte Ho et on rejette H_1 si les différences entre

fréquences sont suffisamment petites pour être attribuées au hasard;

- on rejette Ho et on accepte H_1 si les différences entre

fréquences sont significatives, c'est-à-dire assez grandes pour ne pas être expliquées par le simple hasard.

L'hypothèse nulle Ho est une hypothèse d'homogénéité entre les échantillons: ainsi dit-on que le test de comparaison de plusieurs fréquences observées est un **test d'homogénéité.**

Ainsi, dans le cas du sondage préélectoral, peut-on disposer d'un échantillon venu du nord, un autre du centre, un autre du sud,... L'hypothèse Ho exprime que ces différents départements sont homogènes quant à leurs intentions de vote vis à vis de Monsieur ALI.

S'il n'apparaît pas de différence entre les fréquences observées, ceci veut dire que la variable département n'a pas d'influence sur la variable intention de vote. Autrement dit, l'hypothèse Ho est une hypothèse **d'indépendance** entre deux variables: celle qui préside à la sélection des échantillons, et celle du caractère dans les échantillons.

Le test de comparaison des fréquences observées est donc **un test d'indépendance.**

2- LA FREQUENCE OBSERVEE f EST UNE VARIABLE ALEATOIRE

Soit une population P (une urne) où le caractère a une fréquence théorique p.

Tirons plusieurs échantillons de N individus (boules). Nous avons vu que la fréquence observée dans ces différents échantillons peut varier de l'un à l'autre. On a affaire à des fluctuations d'échantillonnage dues au hasard des tirages. **Ainsi, la fréquence observée est-elle une variable aléatoire.**

Le calcul des probabilités nous apprend beaucoup de choses sur la distribution théorique de cette variable.

Nous ne démontrerons pas les résultats essentiels que voici.

2-1- La moyenne théorique de cette distribution E (f) est égale à p.Ceci est assez intuitif, et conforme à l'expérience, les valeurs observées doivent se grouper autour de p, pour peu qu'à chaque tirage la probabilité de tirer une boule blanche soit toujours égale à p (tirage randomisé). E (f) = p

2-2- La variance théorique de cette distribution V(f) est égale à $\frac{p\,(1 - p)}{N}$ ou N est l'effectif de l'échantillon.

$$V(f) = \frac{p\,(1-p)}{N}$$

2-3-L'écart-type théorique δ(f) vaut donc: $\delta(f) = \sqrt{\frac{p(1-p)}{N}}$

Ce résultat est moins évident, mais il contient encore une idée assez intuitive, à savoir que plus l'effectif de l'échantillon est grand, plus la dispersion des fréquences observées va être faible, plus elles vont se tasser autour de la moyenne théorique P.

Il faut remarquer que la réduction de la dispersion (de l'écart type) n'est pas proportionnelle à l'effectif N, mais seulement à $\sqrt{N}$. Ainsi, dans le cas précédent, en multipliant par 16 l'effectif et par suite les dépenses du sondage, on n'a réduit la dispersion que d'un facteur 4.

2-4- Le calcul des probabilités nous apprend aussi la **loi exacte de probabilité** de f. Celle-ci est déduite de la loi binomiale et est assez compliquée. Mais le calcul des probabilités nous apprend que plus les quantités NP et N (1 - P) sont grandes, plus la loi de f se rapproche d'une loi de LAPLACE GAUSS.

Concrètement, on peut retenir:

<u>Si NP et N(1 - P) sont supérieurs ou égales à 5</u>, alors on peut assimiler la distribution de f à une loi de LAPLACE-GAUSS: de moyenne P et d'écart type

$$\sqrt{V}(f) = \sqrt{\frac{P(1-P)}{N}} \qquad \boxed{L.G.\ (P,\ \sqrt{\frac{P(1-P)}{N}})}$$

Ou encore, la distribution de la variable centrée réduite, appelée écart réduit: $\varepsilon = \dfrac{|f - P|}{\sqrt{\dfrac{P(1-P)}{N}}}$ **suit approximativement une loi L.G. (0,1)**

3- SOLUTION DU PREMIER PROBLEME

Cette solution utilise directement le résultat précédent.

<u>Si l'hypothèse Ho est vraie</u> (à savoir que π = P et que l'échantillon est tiré de façon correcte), alors on peut dire que <u>l'écart réduit</u> $\varepsilon = \dfrac{|f - P|}{\sqrt{\dfrac{P(1-P)}{N}}}$ doit se comporter comme une variable obéissant à une loi de L.G. (0,1). On peut donc parier qu'avec une probabilité égale à 0,95 la quantité ε va satisfaire l'inégalité.

$-1{,}96 < \varepsilon < +1{,}96$ (voir table de L.G. pour la probabilité 0,05)

Ceci veut dire qu'avec une probabilité 0,95 la différence f - P va se situer dans l'intervalle: $\left[\ -1{,}96\sqrt{\dfrac{P(1-P)}{N}}\ ;+1{,}96\sqrt{\dfrac{P(1-P)}{N}}\ \right]$

Autrement dit, on peut écrire, si Ho est vrai

$$\text{Proba}\left[\ -1{,}96\sqrt{\frac{P(1-P)}{N}}\ ;+1{,}96\sqrt{\frac{P(1-P)}{N}}\ \right] = 0{,}95$$

On a ainsi défini un <u>intervalle de pari</u> avec un degré de certitude de 0,95 ou un <u>risque d'erreur</u> de 0,05 (5%).

Dès lors, ou bien la différence (f - P) se retrouve effectivement dans l'intervalle de pari, ou bien elle ne s'y trouve pas.

➔ Si elle s'y trouve, on peut conclure que la différence (f - P) peut être expliquée par le hasard, puisqu'elle fait partir d'un événement dont la probabilité est 0,95, et on accepte Ho (c'est-à-dire que l'on rejette H_1). On dit qu'il n'y a pas de différence significative entre f et P. Remarquons qu'en acceptant Ho, on n'est pas à l'abri de l'erreur, car la valeur observée de f est compatible avec d'autres valeurs que P. Tout ce que l'on peut dire est que l'on n'a pas d'argument convaincant pour rejeter Ho.

➔ Si elle ne s'y trouve pas, on peut dire que la différence (f - P) fait partir d'un événement peu probable (moins de 5% de chances). On peut alors considérer que la fréquence observée serait peu vraisemblable si l'hypothèse Ho était vraie. On préfère donc rejeter l'hypothèse Ho (et, par suite, admettre H_1). Mais, ce faisant, on prend le risque de se tromper, car sous l'hypothèse Ho, la fréquence observée n'est pas très vraisemblable, mais elle n'est pas impossible. En rejetant l'hypothèse Ho, on prend un risque inférieur à 5%.

On dit que la différence (f - P) est significative au seuil 5%.

Que fait-on en pratique?

a- On calcule l'écart type réduit $\varepsilon = \dfrac{|f - P|}{\sqrt{\dfrac{P(1-P)}{N}}}$

b- On compare conventionnellement la valeur trouvée à trois valeurs dans la table de L.G. et correspondant aux risques 0,05; 0,01; 0,001, c'est-à-dire 1,96; 2,58 et 3,29 respectivement.

La communauté scientifique admet les règles suivantes:

➔ Si $|\varepsilon| < 1{,}96$ **f ne diffère pas significativement de P**.

➔ Si $1{,}96 < |\varepsilon| < 2{,}58$ Il existe une **différence significative** entre f et P. On rejette Ho. Le risque d'erreur en rejetant Ho, alors

que cette hypothèse est vraie, est compris entre 0,05 et 0,01. On peut lire dans la table la vraie valeur de ce risque.

➔ Si 2,58 < $|\varepsilon|$ < 3,29 **La différence (f - P) est dite très significative**. On rejette Ho avec un risque d'erreur compris entre 0,01 et 0,001.

➔ Si $|\varepsilon|$ > 3,29 **La différence est dite hautement significative**. On rejette Ho avec un risque de se tromper inférieur à 0,001 (un pour mille).

ATTENTION: le propre de la décision statistique est d'être toujours entachée du risque de se tromper. Ces risques sont de deux espèces:

- quand on rejette Ho est qu'on accepte H_1, il se peut qu'en fait Ho soit vrai. C'est **l'erreur dite de 1ère espèce**, erreur que l'on commet avec un risque α dont on vient de voir qu'on peut le chiffrer, ou du moins le comparer à des valeurs prédéterminées: 0,05; 0,01; 0,001.
- quand on accepte Ho et qu'on rejette H_1, il se peut qu'en fait H_1 soit vrai. On commet alors **une erreur dite de 2ème espèce**.

EXEMPLES:

➔ La proportion de leucémies spontanées chez la souris est P = 0,80. On essaie un traitement anti-leucémique sur un lot de 96 souriceaux. Dans cet échantillon, on constate l'apparition de 68 leucémies. Le traitement est-il efficace? Autrement dit, la fréquence $f = \frac{68}{96}$ =0,708 est-elle significativement différente de P = 0,80?

On calcule $\varepsilon = \frac{|f - P|}{\sqrt{\frac{P(1-P)}{N}}} = 2{,}245$

Les effectifs théoriques Np et Nq étant suffisants, on peut comparer aux valeurs lues dans la table de L.G. (0,1).

Comme 1,96 < ε < 2,58, on peut dire qu'avec un risque d'erreur compris entre 5 % et 1 % le traitement est efficace. On peut même préciser que ce risque est un peu inférieur à 3% (voir la table).

Si on avait observé la même fréquence f = 0,708 sur un effectif N = 192, on aurait trouvé ε = 1,59, et on aurait conclu à une différence non significative.

➔ Sur un échantillon de 10000 naissances, on a enregistré une proportion de garçons de 0,513. Ce résultat est-il compatible avec l'hypothèse que dans la population, le sex-ratio (proportion de mâles) est de 0,50?

$$\varepsilon = \frac{0{,}513 - 0{,}500}{\sqrt{\frac{1}{2}.\frac{1}{2}.\frac{1}{10000}}} = 2{,}6$$

L'échantillon étant grand (NP et N(1-P) $\geq$ 5), on consulte la table de L.G (0,1). La différence est très significative avec un risque inférieur à 1%. Si la même proportion avait été trouvée avec un échantillon de 40000 naissances, le calcul eût montré ε = 5,2. La différence est hautement significative avec un risque d'erreur α inférieur à 10^{-5}.

A RETENIR

La comparaison entre un pourcentage f observé sur n cas et un pourcentage théorique P est basée sur l'écart réduit $\varepsilon = \frac{|f - P|}{\sqrt{\frac{P(1-P)}{N}}}$

Si $|\varepsilon| < 1{,}96$ (pratiquement 2), la différence n'est pas significative (à 5%).

Si $|\varepsilon| \geq 1{,}96$ (pratiquement 2), la différence est significative, et le risque correspondant à ε, lu dans la table de l'écart réduit fixe le degré de signification.

N.B. Le test n'est valable que pour de grands échantillons

(Voir conditions d'application).

4- SOLUTION DU DEUXIEME PROBLEME: L'ESTIMATION DE LA FREQUENCE THEORIQUE

Ce cas se pose dans deux cas schématiques: ou bien on ne sait rien à priori sur la valeur de p; ou bien, à la suite d'un test comme celui du paragraphe précédent, on a rejeté une valeur théorique P, et il faut estimer la vraie valeur.

Cette estimation peut se faire de deux façons:

➔ **Estimation dite ponctuelle**: on donne pour P une valeur. Faute de mieux, on attribue à P la valeur observée f. Ce n'est pas déraisonnable puisqu'on a vu que l'espérance de f, E(f), est égale à P. Il est donc raisonnable de considérer la valeur observée f comme un bon estimateur de P.

Toutefois, il est tout à fait, possible et même très probable que la vraie valeur de P soit différente de f, tout en lui étant voisine. Aussi pratique-t-on souvent l'estimation par **intervalle de confiance**.

➔ **Estimation par intervalle de confiance**: si nous supposons que le tirage est correct, nous avons vu que: $\varepsilon = \dfrac{|f - P|}{\sqrt{\dfrac{P(1-P)}{N}}}$ suit une loi de L.G. (0,1).

Ceci signifie qu'avec une probabilité 1 -α (par ex. 0,95) $|\varepsilon| = \dfrac{|f - P|}{\sqrt{\dfrac{P(1-P)}{N}}}$ est inférieure à u α (dans l'exemple 1,96)

où $u\alpha$ est la valeur lue dans la table de la loi réduite et correspondant à la probabilité α .

De la relation précédente, on déduit par un calcul simple:

$$f - u\alpha\sqrt{\frac{P(1-P)}{N}} \leq P \leq f + u\alpha\sqrt{\frac{P(1-P)}{N}}$$

On a ainsi défini pour P un intervalle de confiance au risque α (5%). Il y a une probabilité 1-α (95%) que l'intervalle défini ainsi contienne la vraie valeur P (et une probabilité α qu'il ne la recouvre pas).

Une difficulté tient au fait que la valeur à estimer P figure dans la définition de l'intervalle, pour le produit P (1 - P).

En fait, on commet une erreur faible en remplaçant P par son estimation ponctuelle f (erreur d'autant plus faible que le produit x (1 - x) varie très lentement avec x).

Finalement, l'intervalle de confiance au risque 5% s'écrit:

$$\left[f - u\alpha.\sqrt{\frac{f(1-f)}{N}}, f + u\alpha.\sqrt{\frac{f(1-f)}{N}} \right]$$

EXEMPLES:

➔ On a vu plus haut que la proportion 0,708 de leucémies chez les souris traitées est différente de 0,80, si l'effectif est de 96 souris. Quelle est donc la vraie valeur de cette proportion? L'intervalle de confiance.

$$f - 1{,}96\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + 1{,}96\sqrt{\frac{f(1-f)}{n}}$$

au seuil de 0,05 est donc égal à: [0,617, 0,799], intervalle qui ne contient pas 0,80.

Si l'effectif du lot traité avait été de 192 souriceaux, l'intervalle de confiance eût été: [0,644, 0,772]

En doublant l'échantillon, on divise l'intervalle de confiance par $\sqrt{2}$.

REMARQUE: L'estimation de la fréquence théorique est la démarche logique lorsqu'un test nous a amené à rejeter une valeur théorique P. Il ne suffit pas en effet de dire que P_1 n'est pas une valeur acceptable, il faut dire où se trouve la vraie valeur P. Il est clair que, pour le même risque d'erreur, la valeur rejetée P_1 n'est pas incluse dans l'intervalle de confiance.

Cette remarque vaut pour l'exemple suivant.

➔ D'un sex-ratio de 0,513 dans un échantillon de 10000 naissances, on peut déduire que, dans la population, il est compris dans l'intervalle [0,503, 0,523] avec une probabilité de 0,95.

➔ Soit un hémogramme. On a, sur 100 globules blancs, compté 50 polynucléaires. Quel est le taux de polynucléaires dans la circulation générale.

On calcule $\sqrt{\frac{f(1-f)}{100}} = \sqrt{\frac{0,5x0,5}{100}} = 0,05$

On déduit: - avec un risque de 5% $0,40 < P < 0,60$

- avec un risque de 1% $0,37 < P < 0,63$.

Si on avait compté 200 cellules, on aurait:

$0,43 \leq p \leq 0,57$ au risque de 5%.

A RETENIR

L'observation d'un pourcentage f sur un échantillon de n cas permet d'assigner au pourcentage inconnu P l'intervalle de confiance à 5%:

$$f - 1,96\sqrt{\frac{f(1-f)}{n}} \leq p \leq f + 1,96\sqrt{\frac{f(1-f)}{n}}$$

5- SOLUTION DU TROISIEME PROBLEME:

5-1- COMPARAISON DE DEUX FREQUENCES OBSERVEES

Nous disposerons de deux échantillons d'effectifs respectifs N_1 et N_2, où le caractère étudié a les fréquences f_1 et f_2.

Que pensez de la différence entre f_1 et f_2 ? Peut-elle être attribuée au hasard? Ou au contraire est-elle trop grande pour être explicable par de simples fluctuations d'échantillonnage ?

On va répondre à cette question par un test. Deux hypothèses sont en présence

Ho : les deux échantillons sont tirés de la même population, où la fréquence théorique du caractère est P (valeur inconnue).

H_1 : les deux échantillons sont tirés de deux populations différentes P_1 et P_2 et/ou l'un ou les 2 échantillons n'ont pas été tirés au hasard.

On raisonne dans le cadre de l'hypothèse Ho: Si Ho est vraie, alors

f_1 suit une loi de L.G. $(P, \sqrt{\frac{P(1-P)}{N_1}})$

et f_2 suit une loi de L.G. $(P, \sqrt{\frac{P(1-P)}{N_1}})$

Que dire de la différence f_1 - f_2 ?

- l'espérance de la différence est la différence des espérances

$$E(f_1 - f_2) = E(f_1) - E(f_2) = P - P = 0$$

Les deux échantillons ayant été tirés indépendamment l'un de l'autre, la variance de la différence est la somme des variances:

$$-V(f_1 - f_2) = V(f_1) + V(f_2) = \frac{P(1-P)}{N_1} + \frac{P(1-P)}{N_2} = P(1-P)(\frac{1}{N_1} + \frac{1}{N_2})$$

De plus, le calcul des probabilités nous apprend que la différence de deux variables de L.G. obéit à une loi de L.G.

Par conséquent, si Ho est vrai, alors f_1-f_2 suit approximativement une loi

$$L.G.(0, \sqrt{P(1-P)(\frac{1}{N_1}+\frac{1}{N_2})})$$

Reste à estimer P.

Si Ho est vrai, les deux échantillons viennent de la même population. On peut donc considérer que la meilleure estimation de P est celle obtenue en tenant compte des deux échantillons à la fois. L'effectif total est e $N_1 + N_2$ individus. Parmi eux, $N_1 f_1 + N_2 f_2$ présentent le caractère. On va donc estimer P par la quantité (estimation ponctuelle) $P = \frac{N_1 f_1 + N_2 f_2}{N_1 + N_2}$

En pratique, on calcule l'écart réduit entre f_1 et f_2 ,

$$\varepsilon = \frac{|f_1 - f_2|}{\sqrt{P(1-P)(\frac{1}{N_1}+\frac{1}{N_2})}}$$ **qui doit suivre une loi de L.G. (0;1)**

A partir de là, l'interprétation est la même que ci-dessus:

➔ $\varepsilon < 1{,}96$ On accepte Ho, c'est-à-dire l'homogénéité des deux échantillons. **La différence observée n'est pas significativement différente de zéro**. On rejette H_1.

➔ $1{,}96 \leq \varepsilon < 2{,}58$ On rejette Ho au seuil 5%, on accepte H_1. **La différence observée est significative**. Le risque d'erreur de 1ère espèce est compris entre 5% et 1%.

➔ $2{,}58 \leq \varepsilon < 3{,}29$ On rejette Ho au seuil 1%. **La différence est très significative**. Le risque d'erreur de 1ère espèce est compris entre 1% et 1%°.

➔ 3,29 $\leq$ ε On rejette Ho au seuil 1%°. **La différence est hautement significative**. Le risque d'erreur de 1ère espèce est inférieur à 1%°.

Rappelons que le fait d'accepter Ho ne signifie pas qu'elle est forcément vraie. Cela signifie seulement que l'information dont on dispose ne permet pas de la rejeter. Peut-être, avec des effectifs plus grands, la même différence deviendrait-elle significative.

EXEMPLE:

➔ Dans une enquête sur 357 asphyxies par le gaz, celle-ci, se subdivise en deux échantillons, l'un E_1 de 136 hommes, l'autre E_2 de 221 femmes.

Dans chacun de ces échantillons, on dénombre ceux qui sont morts de leur asphyxie et ceux qui ont survécu.

	E_1 Hommes	**E_2 Femmes**
Décédés	53	69
Non décédés	83	152
Total	**136**	**221**

Les fréquences de décès sont respectivement:

$$f1 = \frac{53}{136} = 0{,}390 \quad f2 = \frac{69}{221} = 0{,}312$$

Ces deux fréquences observées sont-elles significativement différentes?

L'hypothèse nulle Ho affirme: ces deux échantillons sont tirés d'une même population. Autrement dit, la fréquence théorique est la même pour les hommes et pour les femmes. Ou encore, le sexe n'a pas d'influence sur le décès. Ce sont deux variables indépendantes.

L'hypothèse alternative H_1 affirme au contraire que le sexe a une influence sur le taux de décès.

Si Ho est vraie, ALORS, d'une part, on peut estimer la fréquence de décès dans la population par $\frac{53+69}{357}=\frac{122}{357}=0,342$ ET, d'autre part, la quantité $\varepsilon=\frac{|f_1-f_2|}{\sqrt{P(1-P)(\frac{1}{N_1}+\frac{1}{N_2})}}$ qui suit une loi de L.G. (0 ; 1). Calculons:

$$\varepsilon=\frac{0,390-0,312}{\sqrt{0,342x0,658(\frac{1}{136}+\frac{1}{221})}}=1,5$$

$$\varepsilon=1,5<1,96$$

La différence n'est pas significative. Au vu de ces échantillons, on ne peut affirmer qu'il y a une différence de la fréquence de décès selon le sexe des asphyxiés. On ne peut affirmer que la variable décès dépendra de la variable sexe. J'accepte donc l'hypothèse Ho d'homogénéité entre les échantillons, ou d'indépendance entre les deux variables.

REMARQUES

En complétant le tableau par la colonne des sommes marginales des lignes, on obtient le tableau de contingence qui croise les deux variables sexe et décès.

	E_1 Hommes	**E_2 Femmes**	**Total**
E_1 Décédés	53	69	122
E_2 Non décédés	83	152	235
Total	**136**	**221**	**357**

On aurait pu alors considérer ces données comme deux échantillons, E_1 de décédés et E_2 de survivants. Le caractère étudié est alors le sexe. On aurait étudié par exemple la fréquence du sexe masculin. Parmi les décédés, on a $f'_1 = \frac{53}{122} = 0{,}434$. Parmi les non décédés, on a $f'_2 = \frac{83}{235} = 0{,}353$. Si Ho est vraie, on peut mélanger les deux échantillons et estimer la fréquence des hommes dans la population des asphyxiés par: $\frac{136}{357} = 0{,}381.$

Et on peut calculer: $\varepsilon = \frac{0{,}434 - 0{,}353}{\sqrt{0{,}381 xx\, 0{,}619(\frac{1}{122} + \frac{1}{235})}} = 1{,}5$

On trouve la même valeur de ε, ce qui amène à la même conclusion. On n'a pas de raison de rejeter l'indépendance des deux variables sexe et décès. On voit donc que les deux variables jouent un rôle parfaitement symétrique.

A RETENIR

La comparaison entre deux pourcentages f1, f2 observés sur N1 et N2 cas respectivement, est basée sur l'écart réduit: $\varepsilon = \frac{|f_1 - f_2|}{\sqrt{P(1-P)(\frac{1}{N_1} + \frac{1}{N_2})}}$ **où p et (1-p) désignent les proportions évaluées sur l'ensemble des deux échantillons.**

Si $|\varepsilon| < 1{,}96$ (pratiquement 2), la différence n'est pas significative (à 5%).

Si $|\varepsilon| \geq 1{,}96$ (pratiquement 2), la différence est significative, et le risque correspondant à ε, lu dans la table de l'écart réduit fixe le degré de signification. Le test n'est valable que pour de grands échantillons (voir conditions d'application).

5-2- CAS DE SERIES APPARIEES

Un essai destiné à comparer deux traitements A et B a porté sur 100 malades, qui ont été leurs propres témoins, tous recevant successivement les deux traitements dans un ordre tiré au sort pour chacun d'eux. Le résultat de chaque traitement a tété noté par « + » ou « - », de sorte qu'on a finalement 100 couples de réponses ; on peut les présenter simplement, en donnant l'effectif des 4 couples de réponses possibles.

Résultat avec		***Nombre de malades***
A	B	
-	-	35
-	+	5
+	-	15
+	+	45
Total……		100

Pour comparer les deux traitements, on pourrait penser à :

1- comparer les pourcentages de succès, soit :

$$\text{pour A, } P_A = \frac{15+45}{100} = 60\%$$

$$\text{pour B, } P_B = \frac{5+45}{100} = 50\%$$

On formerait alors l'écart réduit :

$$\varepsilon = \frac{0{,}60 - 0.50}{\sqrt{\frac{0{,}55x0{,}45}{100} + \frac{0{,}55x0{,}45}{100}}} = 1{,}42$$

Et on conclurait à une différence non significative.

Cette méthode serait correcte si l'essai avait porté, comme dans les problèmes étudiés précédemment, sur deux séries indépendantes, de 100 malades chacune, traitées, la première avec A, la deuxième avec B, et ayant donné 60 et 50 succès respectivement.

Mais il n'en est pas de même ici ; les traitements sont été comparés chaque fois sur le même malade, et il faut en tenir compte dans le test statistique.

2- On montre qu'il faut laisser de côté les paires de réponses concordantes, c'est-à-dire les 35 réponses – –et les 45 réponses ++, qui n'apportent rien à la question de savoir quel est le meilleur traitement ; considérant alors les paires de réponses divergentes, au nombre de 20, on examine si les effectifs observés de – + et de + –, soit 5 et 15, sont compatibles avec l'hypothèse d' équivalence des deux traitements ; on a donc à comparer le pourcentage observé $\frac{5}{20}\left(\text{ou } \frac{15}{20}\right)$; ou encore, en raisonnant sur les nombres, l'effectif observé 5 (ou 15) à l'effectif théorique 10 ; dans cette dernière formulation, l'écart réduit est :

$$|\varepsilon| = \frac{|5-10|}{\sqrt{20x\frac{1}{2}x\frac{1}{2}}} = 2{,}24.$$

La supériorité de A sur B est donc significative au risque α= 3 %, contrairement au résultat obtenu à tort par la première méthode.

La notion de séries appariées, examinée ici dans le cas particulier où les deux jugements portent sur le même sujet, s'étend plus généralement au cas où les sujets des deux séries se correspondent par un élément commun : par exemple souris de même portée, de même âge, de même poids, etc.

La méthode à utiliser pour comparer les pourcentages de réponses positives dans les deux séries reste la même : si a et b désignent les nombres de paires à réponses différentes, soit - + et + -, elle conduit à comparer : $\frac{a}{a+b}$ à $\frac{1}{2}$, ou a à $\frac{a+b}{2}$ par l'écart réduit

$$|\varepsilon| = \frac{\left|a - \frac{a+b}{2}\right|}{\sqrt{(a+b)\frac{1}{2}.\frac{1}{2}}} = \frac{|a-b|}{\sqrt{a+b}}.$$

En résumé :

Pour comparer les pourcentages de réponses positives de deux séries appariées, on compte les nombres a et b des paires – + et + –, et on examine si a diffère significativement de b en formant l'écart réduit :

$$\varepsilon = \frac{|a - b|}{\sqrt{a + b}} = 2{,}24.$$

Si $|\varepsilon| < 1{,}96$ (pratiquement 2), les pourcentages ne diffèrent pas significativement (à 5 %).

Si $|\varepsilon| \geq 1{,}96$ (pratiquement 2), les pourcentages diffèrent significativement et le risque correspondant à ε , lu dans la table de l'écart type, fixe le degré de signification.

N.B. : La méthode n'est applicable que si $\frac{a+b}{2} \geq 5$, c'est-à-dire si le nombre des paires considérées égale ou dépasse 10.

La suppression des paires – – et ++, qui a apporté au test un supplément de puissance, peut paraître étonnante. Selon en effet que ces paires sont très nombreuses, ou au contraire peu nombreuses, on doit penser que les séries sont peu différentes, ou au contraire très différentes.

LE TEST DU CHI DEUX

But = Comparaison de fréquences quand il y a un nombre quelconque d'échantillons, et un nombre quelconque de modalités par variable.

1- COMPARAISON DE DEUX OU PLUSIEURS FREQUENCES OBSERVEES

1-1- Le cas où les deux variables ont chacune deux modalités, par exemple le pronostic vital de l'intoxication oxycarbonée selon le sexe. Le tableau de contingence des effectifs observés était le suivant:

	Hommes	**Femmes**	**Total**
Décédés	53	69	122
Non décédés	83	152	235
Total	**136**	**221**	**357**

Si l'hypothèse Ho d'indépendance des deux variables est vraie, ces quatre effectifs observés ne doivent pas être très éloignés des effectifs théoriques. On peut calculer ceux-ci de la façon suivante: Si l'hypothèse Ho -homogénéité des taux de décès pour les hommes et pour les femmes- est vraie, alors la meilleure estimation de la fréquence des décès est $\frac{122}{357} = f$

Par suite, l'espérance du nombre de décès parmi les 136 hommes est $136 \times f = \frac{136 \times 122}{357} = 46{,}48$

C'est la meilleure estimation possible du nombre de décès masculins sous l'hypothèse Ho (c'est-à-dire si Ho est vraie).

De même, l'espérance du nombre de décès féminins est: $\frac{221 \times 122}{357} = 75{,}52$

On peut ainsi réaliser un tableau de contingence théorique:

	Hommes	**Femmes**	**Total**
Décédés	$\frac{122 \times 136}{357} = 46{,}48$	$\frac{122 \times 221}{357} = 75{,}52$	122
Non décédés	$\frac{235 \times 136}{357} = 89{,}52$	$\frac{235 \times 221}{357} = 145{,}48$	235
Total	**136**	**221**	**357**

On voit que dans un tableau de contingence, chacun des effectifs théoriques Ti s'obtient en faisant le calcul:

$$\text{Ti} = \frac{\text{total ligne i x total colonne i}}{\text{total général}}$$

On remarque que, par construction, les effectifs théoriques respectent les sommes marginales. Exemple: 46,48 + 89,52 = 136

Si l'hypothèse Ho est vraie, le tableau observé doit être "voisin" du tableau théorique.

On calcule donc : $\chi 2 = \Sigma \frac{(Oi - Ti)2}{Ti} = \Sigma \frac{Oi2}{Ti} - N$

Pour l'ensemble des cases, et on cherche le risque α correspondant donné par la table du chi-deux pour le nombre de degrés de liberté : d.d.l. = (nombre de lignes - 1) x (nombre de colonnes -1)
Si $\alpha \geq 5\%$, il n'y a pas de liaison significative.
Si $\alpha < 5\%$, la liaison est significative et α mesure son degré de signification. N.B. La méthode n'est valable que si tous les effectifs théoriques Ti $\geq$ 5.

Ainsi $\chi^2 = \frac{(53-46{,}48)^2}{46{,}48} + \frac{(69-75{,}52)^2}{75{,}52} + ... + \frac{(152-145{,}48)^2}{145{,}48} = 2{,}25$

Puisque chacun des effectifs Ti est $\geq$ 5, on peut considérer que χ^2 suit une loi du CHI-DEUX. A combien de d.d.l. ? Il suffit de remarquer que si l'on calcule un seul des effectifs théoriques, on déduit les trois autres par soustraction. Ainsi, si on a calculé l'effectif 46,48 par la formule (1), pour la case en haut et à gauche, on déduit successivement:

75,52 = 122 - 46,48

89,52 = 136 - 46,48

145,48 = 235 - 89,52 (ou bien 221-75,52).

La quantité χ^2 obéit donc à une loi du CHI-DEUX à UN d.d.l.

Comme χ^2 = 2,25 < 3,84, on ne se sent pas le droit de rejeter Ho. On voit qu'on arrive à la même conclusion qu'au chapitre précédent. Ce n'est pas étonnant, puisque au fond, on a réalisé le même test (remarquons que 2,25 = (1,5)², et que 1,5 était la valeur de l'écart réduit).

1-2- Prenons le cas où les deux variables ont plus de deux modalités

L'intérêt du test du CHI-DEUX est qu'il permet d'analyser des tableaux de contingence d'une taille quelconque.

1-2-1- considérons d'abord un tableau (2 x k) à 2 lignes et k colonnes. Une variable à 2 modalités est croisée avec une variable à k modalités.

Variable 2

		mod 1	**mod 2**	**mod k**	**Total**
Variable 1	**mod 1**	n_{11}	n_{12}	n_{1k}	n_1
	mod 2	n_{21}	n_{22}	$n_{2\,k}$	n_2
	Total	$\mathbf{n._1}$	$\mathbf{n._2}$	$\mathbf{n._k}$	**N**

Selon le cas, on pourra considérer ce tableau de plusieurs façons:

➔ comme un seul échantillon de N éléments où l'on s'intéresse simultanément à k variables qualitatives, dont on veut tester l'indépendance.

➔ comme k échantillons à 2 classes et d'effectifs $n._1$, $n._2$, ..., n.k, dont on veut tester l'homogénéité.

➔ comme deux échantillons à k classes d'effectifs $n_1.$ et $n_2.$, dont on veut tester l'homogénéité.

Quel que soit le point de vue, la pratique du test est la même:

- on calcule les 2 xk effectifs théoriques Ti,

- on calcule les 2 xk quantités $\frac{(Oi - Ti)^2}{Ti}$

- on en fait la somme,

- on compare la somme obtenue aux chiffres lus sur la table de loi du CHI-DEUX à k-1 d.d.l.

Pourquoi k - 1 d.d.l. ? Parce que les sommes marginales étant données, il suffit de calculer k - 1 des Ti, les k + 1 autres se déduisant par différence.

EXEMPLE:

➔ On a fait un prélèvement de gorge chez 1 348 enfants pour rechercher la présence de streptocoques. Et on a examiné l'état de leurs amygdales.

	Taille normale	**Hypertrophiées**	**Très hypertrophiées**	**Total**
Streptocoques Oui	19	29	24	72
Streptocoques Non	447	560	269	1276
Total	**466**	**589**	**293**	**1348**

A un examen rapide, il semble que la fréquence des streptocoques augmente quand la taille des amygdales augmente.

L'hypothèse nulle Ho consiste à affirmer qu'il n'y a pas de dépendance entre la présence de streptocoques et la taille des amygdales, et que les différences de fréquences observées ne sont que le fait du hasard.

Le test du CHI-DEUX va nous permettre d'en juger.

On trouve χ^2 = 7,88, valeur qu'on compare aux nombres lus dans la table du CHI-DEUX à 2 d.d.l. On constate que 5,99 < χ^2 < 9,21.

Si l'hypothèse Ho est vraie, on a donc moins de 5% de chances et plus de 1% de chances que la valeur trouvée soit atteinte ou dépassée. On rejette donc Ho avec un risque d'erreur compris entre 5% et 1%. On peut préciser que ce risque est inférieur à 2%.

1-2-2- <u>On peut généraliser au tableau quelconque r x k à r lignes et k colonnes</u>.

Il peut être considéré comme décrivant:

- un seul échantillon où l'on croise deux variables qualitatives à r et k modalités respectivement;
- r échantillons classés selon une variable à k modalités;
- k échantillons classés selon une variable à r échantillons.

On calcule les Ti (en fait il suffit d'en calculer (k - 1) (r - 1)). On somme les $\frac{(\boldsymbol{Oi}-\boldsymbol{Ti})^2}{\boldsymbol{Ti}}$, et on compare cette somme aux valeurs lues dans la table du CHI-DEUX à (k - 1) (r - 1) d.d.l.

2- METHODE PRATIQUE DE CALCUL DU X²

Pour terminer, un problème pratique: comment calculer simplement $\chi^2 = \Sigma \frac{(Oi-Ti)^2}{Ti}$?

L'erreur serait d'appliquer cette formule: r x k différences, r x k élévations au carré, r x k divisions + 1 somme de r x k termes!

On peut montrer que: $\chi^2 = \Sigma \frac{(Oi-Ti)^2}{Ti} \Leftrightarrow \chi^2 = \Sigma \frac{Oi^2}{Ti} - N$

formule plus facile à calculer car on évite les r x k différences, source à la fois de travail et d'erreurs d'arrondi.

3- CONDITIONS DE VALIDITE DU TEST DU CHI-DEUX

De la même façon -et pour les mêmes raisons- que f obéit d'autant mieux à la loi de L.G. que les effectifs théoriques Np et N (1-p) sont plus grands, la quantité χ^2 obéit à la loi du CHI-DEUX d'autant mieux que les effectifs théoriques sont plus grands.

En pratique, on peut mettre le test en pratique si, pour chaque classe, on a: Ti $\geq$ 5

Dans le cas contraire, on peut essayer de regrouper les classes de petit effectif. Sinon, il faut faire appel à la loi exacte de probabilité (loi multinomiale). Ceci dépasse l'objet de ce cours.

UNE REMARQUE FONDAMENTALE

Le test du CHI-DEUX permet de suspecter ou d'affirmer (au risque d'erreur de 1ère espèce près) l'existence d'une liaison. Mais il ne constitue pas une MESURE de celle-ci. Ce n'est pas parce que le test sera hautement significatif, que la liaison sera forte. Tout ce qu'on peut dire, c'est qu'il existe presque certainement une liaison.

Il faut toutefois tempérer cette distinction en remarquant qu'une liaison forte sera plus facilement détectée: un petit échantillon pourra suffire pour arriver à un CHI-DEUX significatif.

EXEMPLE :

➔ trois formes cliniques A, B et C d'une même maladie ont l'évolution décrite par le tableau ci-dessous:

	Survie à 5 ans sans complications	**Survie à 5 ans avec complications**	**Décès avant 5 ans**	**Total**
A	6	7	16	29
B	22	19	48	89
C	43	58	147	247
Total	**71**	**84**	**211**	**366**

Peut-on considérer que ces trois formes cliniques ont la même évolution?

Si oui (hypothèse Ho), la distribution des 9 valeurs observées ne doit pas différer de la distribution des 9 valeurs théoriques suivantes:

	Survie à 5 ans sans complications	Survie à 5 ans avec complications	Décès avant 5 ans	Total
A	5,63	6,66	16,71	29
B	17,27	20,43	51,30	89
C	48,10	56,91	142,99	247
Total	**71**	**84**	**211**	**366**

Calculons $X^2 = \Sigma \frac{Oi^2}{Ti} - N \qquad = \frac{6^2}{5{,}63} + \frac{7^2}{6{,}66} + \ldots + \frac{147^2}{142{,}99} - 366 =$ 2,35

Cette valeur observée doit être comparée aux valeurs lues sur la table du CHI-DEUX à (3-1) (3-1) = 4 d.d.l.

Elle est très inférieure à la valeur pour p = 0,05, soit 9,49.

On peut donc garder l'hypothèse nulle.

Au vu de cet échantillon, on n'a pas le droit de conclure à une évolution différente entre les trois formes cliniques.

Remarque: Mais on ne peut pas non plus affirmer "mordicus" qu'elles ont la même évolution. Peut être avec un échantillon plus important, observerait-on une différence. Si par exemple tous les effectifs observés étaient multipliés par 10, on aurait trouvé:

$\chi^2 =$ 23,5 valeur très hautement significative car supérieure à 18,47.

A RETENIR

Pour prouver l'indépendance de deux variables qualitatives, à partir du tableau de contingence à L lignes et C colonnes, on détermine d'abord pour chaque case l'effectif théorique dans l'hypothèse d'indépendance, qui est le produit du total de sa ligne (L) par le total de sa colonne (C), divisé par le total général. On forme ensuite

$$\chi^2 = \Sigma \frac{(Oi - Ti)^2}{Ti} = \Sigma \frac{Oi^2}{Ti} - N$$

Pour l'ensemble des cases, et on cherche le risque α correspondant donné par la table du chi-deux pour le nombre de degrés de liberté.

d.d.l. = (Nb L - 1) x (Nb C -1)

Si $\alpha \geq 5\%$, il n'y a pas de liaison significative (indépendance des deux variables).

Si $\alpha < 5\%$, la liaison est significative et α mesure son degré de signification.

N.B. : La méthode n'est valable que si tous les effectifs théoriques Ti $\geq$ 5.

VARIABLES QUANTITATIVES
PROBLEMES DE MOYENNES

1- LES PROBLEMES A RESOUDRE

Note préliminaire : dans ce chapitre, nous voulons étudier des échantillons sous l'angle d'une variable quantitative. Le faire de façon générale nous amènerait à nous intéresser non seulement à la moyenne de ces échantillons, mais aussi à son écart type, à sa médiane, etc. et même à la forme de sa distribution.

Ici, en fait, nous n'examinerons en détail que le problème de la moyenne observée des échantillons, en négligeant les autres aspects (nous ne ferons qu'une brève allusion à l'estimation de l'écart type car on en a besoin pour résoudre le problème des moyennes).

Les problèmes qui peuvent poser les moyennes observées sont de même type que pour les fréquences observées et au nombre de trois.

La logique du raisonnement est parfaitement superposable

1-1- comparaison d'une moyenne observée m et d'une moyenne théorique μ_0

Bien que la population ne soit pas connue, on est parfois amené à faire l'hypothèse Ho que, dans la population, la moyenne μ des caractères a une valeur donnée μ_0.

On dispose d'un échantillon ou la moyenne observée de même caractère est m.

L'observation de m est-elle compatible avec l'hypothèse que dans la population $\mu=\mu_0$?

Il s'agit là d'un test d'hypothèse, auquel on répond par OUI ou par NON

- Si la différence m - μ_0 est petite, on peut considérer qu'elle est le fruit du pur hasard et des fluctuations d'échantillonnage. On accepte alors l'hypothèse Ho que $\mu = \mu_0$.
- Si la différence m -μ_0 est grande, elle ne peut guère être expliquée par le hasard. Elle apparaît significative. Dans ce cas, on rejette Ho et on accepte une hypothèse alternative, à savoir que $\mu \neq \mu_0$. On a donc le schéma suivant :

Ho l'échantillon est tiré au hasard d'une population P où la moyenne est μ_0. C'est l'hypothèse "nulle" :

H_1 Hypothèse alternative. C'est l'hypothèse complémentaire de Ho. L'échantillon n'a pas été tiré au hasard d'une population où la moyenne est μ_0. Ce qui veut dire que l'on a : $\mu \neq \mu_0$ ou tirage biaisé

Remarquons que les deux aspects de l'hypothèse H_1 ne sont en fait pas distincts. Car si on a tiré un échantillon biaisé, ceci ne veut pas dire qu'on l'a tiré d'une sous- population particulière de P dont la moyenne est différente de la moyenne générale dans celle-ci. Cette remarque vaut aussi pour les fréquences au chapitre précédent

1-2- Estimation d'une moyenne théorique à partir d'une moyenne observée

Ici, la moyenne μ est supposée totalement inconnue, et l'échantillon est supposé parfaitement randomisé. que dire alors de μ , connaissant m ? c'est un problème d'estimation.

1-3- Comparaison de deux moyennes observées

(Nous ne traiterons pas ici de la comparaison de plusieurs moyennes observées qui fait appel à une méthode plus complexe, dite analyse de variance).

On dispose de deux échantillons où les moyennes (pour le même caractère bien sûr) sont respectivement m_1 et m_2.

La différence m_1 - m_2 est-elle suffisamment petite pour qu'on puisse admettre que les deux échantillons soient tirés d'une même

population de moyenne théorique $\mu = \mu_1 = \mu_2$? C'est un test d'hypothèse. On va confronter en effet deux hypothèses.

Ho (hypothèse nulle) : les deux échantillons proviennent de façon randomisée d'une même population P.

H_1 (hypothèse alternative) : les deux échantillons n'ont pas été tirés au hasard d'une même population. Par exemple, ils ont été tirés de deux populations différentes.

Si $m_1 - m_2$ est "petite" , on accepte Ho et on rejette H_1.

Si $m_1 - m_2$ est suffisamment "grande" pour que le hasard seul ne puisse l'expliquer, on rejette Ho et on accepte H_1.

2- LA MOYENNE m D'UN ECHANTILLON RANDOMISE EST UNE VARIABLE ALEATOIRE

Soit une population P où une variable quantitative X est distribuée avec une moyenne μ et un écart type σ ; on ne fait aucune hypothèse sur la façon de la distribution de X, qui peut être quelconque.

Extrayons de façon randomisée de cette population un certain nombre d'échantillons de même effectif n, et calculons les moyennes $m_1, m_2, ... m_n$ de ces échantillons. Il est évident, et l'expérience le confirme, que ces moyennes ne sont pas identiques. Elles varient, et ces variations expriment les fluctuations d'échantillonnage. Autrement dit, la moyenne même d'un échantillon tiré au hasard, est une variable aléatoire.

Que nous apprend le calcul des probabilités sur la distribution théorique de la moyenne observée d'un échantillon d'effectif n ?

Les résultats principaux sont les suivants :

2-1- La moyenne théorique de cette distribution E(m) est égale à μ.

Ce résultat semble assez naturel. Les valeurs de m sont regroupées autour de μ.

2-2- La variance de m est telle que V(m) = $\frac{s^2}{n}$

Ici encore, on retrouve l'idée que plus l'échantillon est grand, plus les moyennes observées sont groupées autour de μ et, comme pour les fréquences, la réduction de la dispersion n'est proportionnelle qu'à $\sqrt{n}$ (et non pas à n).

2-3- La forme de la distribution : La Loi de probabilité de m ne peut être connue de façon exacte, que si on connaisse celle de X dans la population.

En particulier, si le caractère obéit à une loi de LAPLACE-GAUSS dans la population, alors la distribution de m suit une loi de LAPLACE-GAUSS. Donc : m suit L.G. $(\mu, \frac{\sigma}{\sqrt{n}})$ ou encore :

U = $\frac{|m-\mu|\sqrt{n}}{\sigma}$ suit L.G.(0,1).

Toutefois, on peut montrer que lorsque l'effectif n tend vers l'infini, la loi de probabilité de m tend vers une loi de LAPLACE-GAUSS et ceci quelle que soit la forme de distribution de X dans P. Elle y tend d'autant plus vite que la distribution de X est plus proche d'une distribution de L.G. En pratique, pourvu que la distribution de X ne soit pas trop compliquée ou tordue, on peut assimiler la distribution de m à une distribution L.G. à partir d'une valeur de n $\geq$ 30.

On peut résumer ces résultats par le tableau suivant qui distingue quatre cas selon le caractère de la distribution de X dans P et selon la taille de l'échantillon, et qui indique ce que l'on sait de la distribution de m.

DISTRIBUTION DE X DANS P

Taille de l'échantillon	L.G	Quelconque
grand n $\geq$ 30	L.G. $(\mu, \frac{\sigma}{\sqrt{n}})$	L.G. $(\mu, \frac{\sigma}{\sqrt{n}})$
petit n < 30	$\simeq$ L.G. $(\mu, \frac{\sigma}{\sqrt{n}})$	?

Dans les trois cas non hachurés, en toute rigueur, ou de façon approximative :

$$\varepsilon = \frac{|m-\mu|\sqrt{n}}{\sigma} \text{ suit une loi L.G. (0,1)}.$$

Intervention de la Loi de Student

Le problème semble résolu, mais il persiste en fait une inconnue, car le plus souvent σ, écart type de X, est inconnu. Que faire?

On ne peut que remplacer σ par une estimation. Le calcul des probabilités montre que la meilleure estimation (ponctuelle) de σ^2, variance de X dans P, est:

$$S^2 = \frac{\sum (xi - mx)^2}{n-1}$$

Le remplacement de σ^2 par une estimation S^2 (entachée d'incertitude) a pour effet de modifier quelque peu les résultats précédents. On montre que lorsque X suit une loi d L.G. dans P, m ne suit pas une loi de L.G. mais une loi dérivée de celle-ci: la loi dite de Student. C'est cette loi en effet et non pas la loi de L.G (0,1) que suit la quantité: $\varepsilon = \frac{|m-\mu|\sqrt{n}}{S}$

Mais la forme de cette loi montre que dès que n augmente, la loi de Student tend vers la loi de L.G (0,1). En pratique, on peut faire l'assimilation de la loi de Student à la loi de L.G., dès que n $\geq$ 30.

Le tableau précédent est donc finalement modifié de la façon suivante quand σ est inconnue, qui donne la distribution de

$$\varepsilon = \frac{|m-\mu|\sqrt{n}}{S}$$

DISTRIBUTION DE X

DANS P

Taille de l'échantillon	L.G.	Non L.G.
Grand n $\geq$ 30	Student à n-1 ddl # L.G. (0,1)	# L.G. (0,1)
Petit n < 30	Student à n - 1 ddl	??

Comment lire la table de Student?

Il en est du Student comme du CHI-DEUX. Il n'y a pas une loi de Student, mais une infinité, qui diffèrent par le nombre de degrés de liberté (d.d.l.). Que choisir ?

Pour un échantillon de taille n, la moyenne m est telle que: $\varepsilon = \frac{|m-\mu|\sqrt{n}}{S}$ suit une loi de Student à n - 1 d.d.l.

En examinant la table des lois de Student, on constate que plus le nombre de d.d.l. est grand, plus on tend vers une loi de L.G. (0,1). En effet, pour n = 30, on retrouve les valeurs bien connues de la loi de L.G. centrée réduite.

On voit aussi que pour n - 1 = 30 d.d.l., les valeurs lues sont très peu supérieures aux valeurs de la loi de L.G. (0,1). Et ceci d'autant plus que α est plus petit. Ceci a pour effet d'élargir l'intervalle de pari (1er problème), et l'intervalle de confiance (2ème problème). Cet élargissement n'est pas surprenant, puisque la non-connaissance de σ ajoute une incertitude supplémentaire.

3- SOLUTION DU PREMIER PROBLEME: COMPARAISON D'UNE MOYENNE OBSERVEE m ET D'UNE MOYENNE THEORIQUE

On applique directement les résultats précédents. C'est un test d'hypothèse avec: Ho (hypothèse nulle): $\mu = \mu_0$ ET échantillon randomisé.

H_1 (hypothèse alternative): $\mu \neq \mu_0$ OU échantillon non randomisé.

Si Ho est vraie, alors on peut faire un pari sur m, ou sur la différence m - μ.

On calcule $\varepsilon = \frac{|m-\mu|\sqrt{n}}{S}$

Avec une probabilité 0,95, la valeur calculée doit se trouver dans un intervalle de pari:

- u 0,05 = -1,96 < ε < 1,96 = u 0,05 Si n ≥ 30 ou

- t 0,05, n-1 < ε < t0,05, n-1 Si n < 30

où t0,05, n-1 est la valeur lue dans la colonne 0,05, et la ligne n-1 de la table des lois de Student. Dans ce deuxième cas l'intervalle de pari est plus grand que dans le premier:

De façon plus générale, on a:

Probabilité $(|\varepsilon| < u\alpha) = 1-\alpha$ Si n ≥ 30

Probabilité $(|\varepsilon| < t\alpha, n-1) = 1-\alpha$ Si n < 30

uα = valeur lue dans la table L.G. (0,1) réduite

tα, n-1 = valeur lue dans la table de Student à n - 1 d.d.l.

L'interprétation du test est toujours la même:

➔ $|\varepsilon| < u0{,}05 \; ou \; t0{,}05, n-1$

m ne diffère pas systématiquement de μ_0. On accepte H_0, parce qu'on manque d'arguments pour faire autrement. On rejette H_1.

➔ u 0,05 ou t0,05, n-1 < $|\varepsilon| < u0{,}01 \; ou \; t0{,}01, n-1$

La différence entre m et μ_0 est significative et on accepte H_1. On rejette Ho, avec un risque d'erreur de 1ère espèce inférieur à 5% et supérieur à 1%.

➔ u 0,01 ou t0,01, n-1 < $|\varepsilon| \leq u0{,}001 \; ou \; t0{,}001, n-1$

La différence entre m et μ_0 est très significative. On rejette Ho avec un risque d'erreur beaucoup plus petit puisqu'il est compris entre 1% et 1‰. On accepte H_1.

➔ $|\varepsilon| \geq u0{,}001 ou\ t0{,}001, n-1$

La différence entre m et μ_0 est hautement significative. Le rejet de H_0 et l'acceptation de H_1 se font avec une quasi-certitude (risque d'erreur inférieure à 1‰).

EXEMPLES:

➔ On dispose d'un échantillon de 400 adultes jeunes de sexe masculin, la moyenne de leur taille est de m = 172,23 cm avec un écart type de s = 2,50 cm.

On se demande si cet échantillon est représentatif de la population générale, où la taille moyenne est = 171,33 cm. "Est représentatif" signifie avoir été tiré dans des conditions honnêtes, sans biais.

On calcule $\varepsilon = \dfrac{(172{,}23 - 171{,}33)\sqrt{400}}{2{,}50} = 7{,}2$

Cette valeur est très supérieure à 3,29 (u = 0,001). Il y aurait donc beaucoup moins de 1 chance sur 1000, si l'échantillon avait été tiré honnêtement de cette population, de trouver une valeur égale ou supérieure à 7,2 (en regardant les dernières lignes de la table de L.G., on peut même voir qu'il y a moins de 1 chance la table de L.G., on peut même voir qu'il y a moins de 1 chance sur un milliard). m est différent de μ_0 de façon hautement significative. On peut affirmer en toute sécurité que cet échantillon n'est pas représentatif de la population au moins en ce qui concerne la taille des sujets. Il a été tiré dans une sous-population de taille plus grande: c'est un échantillon biaisé.

Même problème, avec les mêmes données, sauf que l'effectif de l'échantillon n'est que de 16.

Pour avoir droit de résoudre ce problème, il faut examiner si le caractère est raisonnablement gaussien dans la population. Si oui, on peut calculer

$$\varepsilon = \frac{|m-\mu|\sqrt{n}}{S}$$ avec n = 16 , t = à 1,44 .

Cette valeur doit être comparée aux valeurs lues dans la table de student à n - 1 d.d.l. Or, 1,44 < t 0,05, 15 = 2,131.

On ne rejette pas Ho. Rien ne permet de dire que ce petit échantillon n'est pas représentatif de la population. m n'est pas significativement différent de μ.

On voit que m tombe à l'intérieur de l'intervalle de pari au risque de 5%:

$$\left[171{,}33 - 2{,}13\frac{2{,}5}{\sqrt{16}}, 171{,}33 + 2{,}13\frac{2{,}5}{\sqrt{16}}\right] : [170,\ 00 \quad - \quad 172{,}66]$$

➔ Même problème, mêmes données, mais n = 36

Ici, on a $\varepsilon = \frac{|m-\mu|}{S}\sqrt{n} = 2{,}16$

Cette valeur est telle que: t0,05,35 = 2,03 < ε < t0,01,35 = 2,75.

On est en droit de rejeter Ho avec un risque légèrement inférieur à 5%. La différence m - μ_0 est significativement différente de zéro, mais de façon limitée.

A RETENIR

Pour les grands échantillons n $\geq$ 30

La comparaison d'une moyenne m, observée sur n cas, à une valeur théorique μ est basée sur l'estimation de l'écart réduit

$$\varepsilon = \frac{|m-\mu|\sqrt{n}}{S}$$

où s désigne l'écart type estimé sur l'échantillon.
Si $|\varepsilon| < 1{,}96$ (pratiquement 2), la différence n'est pas significative (à 5%).
Si $|\varepsilon| \geq 1{,}96$ (pratiquement 2), la différence est significative, et le risque correspondant à ε, lu dans la table de l'écart réduit fixe le degré de signification.

Pour les petits échantillons n < 30

La comparaison d'une moyenne m, observée sur n cas, à une valeur théorique μ est basée sur le rapport

$$t = \frac{|m-\mu|}{\frac{s}{\sqrt{n}}}$$

où s désigne l'écart type estimé sur l'échantillon.
Si $|t|$ est inférieur à la valeur lue dans la table de t pour d.d.l. = n-1 et le risque 5%, la différence n'est pas significative; dans le cas contraire elle est significative, et le risque indiqué par la table pour la valeur $|t|$ trouvée fixe le degré de signification.
N.B. le test n'est utilisable que si le caractère étudié est distribué selon la loi normale.

4- ESTIMATION D'UNE MOYENNE THEORIQUE SUPPOSEE INCONNUE A PARTIR DE LA MOYENNE m D'UN ECHANTILLON RANDOMISE

On peut faire deux types d'estimation:

- Estimation ponctuelle: on donne pour μ une valeur. La moins mauvaise est évidemment la valeur m, puisque E(m) = μ. Toutefois, il est plus prudent d'avoir recours à une estimation par intervalle de confiance.
- Estimation par intervalle de confiance: on part du même concept que si l'échantillon est tiré au hasard, on a selon la taille n de l'échantillon:

Probabilité $\left[\frac{|m-\mu|}{S}\sqrt{n} < ua \text{ ou } t\alpha, n-1 \right]$ = 1-α

Ceci permet de calculer l'intervalle de confiance au risque α:

Si n est grand $m - u\alpha.\frac{S}{\sqrt{n}} < \mu < m + u\alpha.\frac{S}{\sqrt{n}}$

Si n est petit et si L.G. dans P $m - t\alpha, n-1.\frac{S}{\sqrt{n}} < \mu < m + t\alpha, n-1.\frac{S}{\sqrt{n}}$

EXEMPLES:

➔ On a un échantillon de n = 400 sujets dont la taille moyenne vaut 172,23 cm et l'écart type s = 2,50 cm.

En admettant qu'il ait été tiré de façon loyale de la population P, que peut-on dire de la moyenne de la taille de celle-ci?

Puisque n = 400, on considère la table de L.G. au risque 5% et à la sécurité 95%, on peut dire que μ se trouve dans l'intervalle de confiance: $\left[m - 1{,}96\frac{S}{\sqrt{n}}, m + 1{,}96\frac{S}{\sqrt{n}} \right]$ soit [171,98-172,48]

➔ Même problème, mêmes données, mais l'échantillon est réduit: n = 16.

Ici, il faut considérer Student à 15 d.d.l. L'intervalle de confiance au risque 5% est $\left[m - 2{,}13\frac{S}{\sqrt{n}}, m + 2{,}13\frac{S}{\sqrt{n}} \right]$ soit [170, 90, 173,56]

On constate que l'intervalle de confiance varie beaucoup avec la taille de l'échantillon: plus celle-ci est petite, plus grande est l'incertitude.

A RETENIR

Pour les grands échantillons n $\geq$ 30

L'observation d'une moyenne m sur un échantillon de n observations permet d'assigner à la moyenne inconnue μ l'intervalle de confiance à 5% de risque : $m - \frac{1{,}96s}{\sqrt{n}} \leq \mu \leq m + \frac{1{,}96s}{\sqrt{n}}$ **;**

s étant l'écart type estimé sur l'échantillon.

Pour les petits échantillons n < 30

L'observation d'une moyenne m sur un petit échantillon de n cas permet d'assigner à la moyenne inconnue μ l'intervalle de confiance à 5% de risque : $m - \frac{t.s}{\sqrt{n}} \leq \mu \leq m + \frac{t.s}{\sqrt{n}}$ **;**

s étant l'écart type estimé sur l'échantillon, et t la valeur donnée par la table de t pour le nombre de degrés de liberté (n-1) et le risque 5%.

N.B. Cette formule n'est valable que si le caractère étudié est distribué selon la loi normale.

5- TROISIEME PROBLEME: COMPARAISON DE LA MOYENNE DE DEUX ECHANTILLONS INDEPENDANTS

Nous disposons de deux échantillons qui ont été tirés indépendamment l'un et l'autre. Leurs effectifs respectifs sont n_1 et n_2. Le caractère étudié X a pour moyenne m_1 et m_2 et pour écart type S_1 et S_2 respectivement.

Que penser de la différence m_1 -m_2? Peut-elle raisonnablement être attribuée au seul hasard d'échantillonnage?

Ici encore, on va résoudre le problème par un test. Deux hypothèses en présence:

Ho (Hypothèse nulle) les deux échantillons sont tirés de la même population où la moyenne théorique est μ. Dans ces conditions, la différence m_1-m_2 ne peut être que l'effet du hasard.

H_1 (hypothèse alternative): les deux échantillons sont tirés de deux populations distinctes où les moyennes sont respectivement μ_1 et μ_2. Dans ces conditions, la différence m_1 -m_2 démontre la différence systématique μ_1 -μ_2.

Il faut distinguer deux cas:

5-1- 1er cas: n_1 et n_2 sont grands (chacun d'eux $\geq$ 30).

Comme d'habitude, on raisonne dans le cadre de Ho qu'on suppose vraie. Si Ho est vraie, Alors: m_1 suit une loi L.G. $(\mu_1, \frac{S_1}{\sqrt{n_1}})$

et m_2 suit une loi L.G. $(\mu_2, \frac{S_2}{\sqrt{n_2}})$

m_1 - m_2 suit une loi L.G. dont la moyenne est μ_1 -μ_2 = 0, et la variance est la somme des variances $\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$ C'est-à-dire une loi L.G. $(0, \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}})$

$$\varepsilon = \frac{|m_1 - m_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$ suit une loi L.G (0,1)

Le test en résulte immédiatement. On calcule ε, et on le compare aux valeurs normales 1,96; 2,58; 3,29 et l'interprétation se fait selon l'habitude.

EXEMPLE:

➔ on dispose de deux échantillons de nouveau-nés. 82 sont des prématurés (nés avant terme mais de développement normal, compte tenu de la durée de la grossesse). 87 sont des dysmatures (le développement intra-utérin a été perturbé et retardé par rapport à la durée de grossesse). On a mesuré chez tous les périmètres crâniens.

Pour le premier échantillon, on a m_1= 30,06 cm ; S_1= 2,04 cm

et pour le deuxième m_2= 30,72 cm ; S_2= 1,96 cm

On calcule: $$\varepsilon = \frac{30,72 - 30,06}{\sqrt{\frac{2,04^2}{82} + \frac{1,96^2}{87}}} = 2,14$$

On constate: u 0,05 = 1,96 < ε < u0,01 = 2,58

La différence m_1-m_2 est significative au risque 5%, mais pas au risque de 1%. En rejetant Ho, on compte un risque inférieur à 5% mais supérieur à 1% (en examinant la table, on constate qu'il est de l'ordre de 3%).

5-2- 2ème cas: L'un au moins des deux effectifs n_1 et n_2 est petit (< 30).

Dans ce cas, le problème n'est soluble que si deux conditions supplémentaires sont remplies:

5-2-1- Le caractère est gaussien (au moins de façon approchée) dans la population.

5-2-2- Les deux variances distinctes S_1^2 et S_2^2 ne sont pas significativement différentes.

Tester ces deux hypothèses va au-delà des objectifs de ce cours. Pour la première condition, on peut se faire une idée approximative de la distribution en regardant l'histogramme des échantillons.

En admettant que ces deux conditions sont vérifiées, on procède ainsi:

Puisque les deux variances S_1^2 et S_2^2 ne sont pas significativement différentes, on peut faire la meilleure estimation de leur valeur commune en mêlant les échantillons. Cette estimation montre:

$$S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1-1+n_2-1} = \frac{\sum(xi-m_1)^2 + \sum(xi-m_2)^2}{n_1+n_2-2}$$

Le test s'effectue alors de la façon suivante:

Si H_0 est vraie, ALORS on peut montrer que:

$t = \frac{|m_1-m_2|}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ suit une loi de Student à $n_1 + n_2 - 2$ d.d.l.

EXEMPLES:

➔ Chez n_1 =18 prématurés, on a mesuré la natrémie (taux de na dans le plasma). On a trouvé m_1 = 133,0 mmol/l ; S_1 = 7,52 mmol/l

Chez n_2 = 24 dysmatures, on a trouvé pour le même paramètre: m_2 = 136,6 mmol/l ; S_2 = 9,01 mmol/l

Le test de comparaison des variances S_1 et S_2 montre qu'elles ne diffèrent pas significativement. On peut donc estimer leur valeur commune par:

$$S^2 = \frac{17(7,52)^2 + 23(9,01)^2}{17+23} = 70,66\,(\text{mmol/l})^2 \qquad S = 8,41 \text{ mmol/l}$$

On calcule $t = \frac{|133,0-136,6|}{8,41\sqrt{\frac{1}{18}+\frac{1}{24}}} = 1,38$

On constate que: $|t| = 1,38 < t_{0,05,40} = 2,02$.

On ne se sent donc pas en droit de rejeter Ho car la différence m_1 -m_2 n'apparaît pas significative, pouvant être expliquée par le hasard.

A RETENIR

Pour les grands échantillons n1 > 30 et n2 > 30

La comparaison entre deux moyennes m1 et m2 observées sur n1 et n2 cas, est basée sur l'écart réduit:

$$\varepsilon = \frac{|m_1 - m_2|}{\sqrt{\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2}}}$$

où $S1^2$ et $S2^2$ désignent les variances estimées.

Si $|\varepsilon| < 1{,}96$ (pratiquement 2), la différence n'est pas significative (à 5%).

Si $|\varepsilon| \geq 1{,}96$ (pratiquement 2), la différence est significative, et le risque correspondant à ε, lu dans la table de l'écart réduit fixe le degré de signification.

Pour les petits échantillons n1 < 30 ou n2 < 30

La comparaison entre deux moyennes m1 et m2 observées sur deux échantillons de n1 et n2 cas, dont l'un au moins est petit, est basée sur la valeur de:

$$t = \frac{|m_1 - m_2|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

où S^2 désigne l'estimation de la variance, supposée commune, par la formule :

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}$$

Si $|t|$ est inférieur à la valeur lue dans la table de t pour

d.d.l. = n1 + n2 - 2

et le risque 5%, la différence n'est pas significative; dans le cas contraire elle est significative et le risque indiqué par la table pour la valeur de $|t|$ trouvée fixe le degré de signification.

N.B. le test n'est utilisable que si le caractère étudié est distribué, dans les 2 populations d'où proviennent les échantillons, selon des lois normales de même variance.

6- TROISIEME PROBLEME BIS: COMPARAISON DE DEUX ECHANTILLONS APPARIES

Dans le cas précédent, les deux échantillons étaient tirés indépendamment l'un de l'autre. Tel n'es pas le cas, si les deux échantillons ont même effectif n, et s'il y a une même correspondance qui associe chaque valeur de l'un à une valeur de l'autre. On dit alors que les échantillons sont appariés.

On parle de série appariée quand on mesure la même chose par deux personnes différentes; par deux techniques différentes; dans deux conditions différentes; en deux lieux différents soit en deux moments différents.

Un des meilleurs exemples d'échantillons appariés est celui où n sujets d'expérience sont l'objet de deux mesures d'un même paramètre, et qu'il s'agit de comparer.

	1ère mesure	**2ème mesure**
Sujet 1	X_1	Y_1
Sujet 2	X_2	Y_2
-		
-		
Sujet n	Xn	Yn

Il s'agit de comparer les xi et les yi.

EXEMPLES:

- Sur chacun des n échantillons de sang, on fait deux dosages d'une même substance par deux méthodes de mesure qu'on veut comparer.
- On mesure sur chacun des n malades, un même paramètre physiologique dans deux circonstances différentes.
- Sur un paquet de copies d'examen, on pratique une double correction.

CE QU'IL NE FAUT PAS FAIRE:

Il ne faut pas procéder comme si ces deux échantillons étaient indépendants. Car ceci reviendrait à comparer l'ensemble des xi à l'ensemble des Yi, alors qu'il convient de comparer chaque xi à l'yi correspondant. Ce serait méconnaître la planification qui a présidé au recueil des données.

PROCEDURE A SUIVRE :

Il convient pour tenir compte de cette planification d'effectuer les n différences : di = Xi - Yi.

On va agir désormais sur cet échantillon unique de n valeurs di.

On en calcule la moyenne d et l'écart -type Sd.

Si les xi ne sont pas différents des Yi alors la moyenne d des différences ne diffère pas de zéro. Les tests se présentent ainsi :

H_0 les di sont tirés d'une "population de différences" dont la moyenne théorique est $\mu = 0$. H_1 $\mu \neq 0$.

Selon une formule qui fait la joie des polytechniciens, "on est ramené à un problème précédent": celui de la comparaison d'une moyenne observée d à une moyenne théorique $\mu = 0$.

Si donc H_0 est vraie, ALORS : $\varepsilon = \dfrac{|d-0|\sqrt{n}}{Sd}$ **suit une loi L.G (0,1) si n est grand ou une loi de student à n - 1 d.d.1. si n< 30 . Dans ce dernier cas, il faut supposer en outre que la "population des différences" est gaussienne. EXEMPLES :**

➔ Soit à comparer chez 10 malades la pression artérielle systolique moyenne après l'administration d'un médicament hypotenseur, et après un placebo.

	Après placebo X_1	après hypotenseur X_2	$d = X_1 - X_2$
1er malade	17	16	+1
2ème malade	15	11	+4
3ème malade	15	12	+3
4ème malade	13	13	0
5ème malade	12	14	-2
6ème malade	17	11	+6
7ème malade	15	13	+2
8ème malade	16	13	+3
9ème malade	19	17	+2
10ème malade	11	10	+1

L'hypothèse nulle H_0 s'exprime ainsi : dans la population, l'effet de l'hypotenseur est en moyenne identique à l'effet du placebo. Autrement dit, la différence moyenne μ est égale à zéro. L'hypothèse alternative H_1 est : la différence moyenne μ est ≠ 0.

On calcule facilement $\sum di = T_1 = 20$; $\sum di\dagger = T_2 = 84$

$$d = \frac{\sum di}{10} = \frac{20}{10} = 2$$

$$Sd^2 = \frac{T_2 - \frac{T_1^2}{10}}{9} = \mathbf{4{,}89\ et\ par\ suite}\ t = \frac{|d-0|}{\sqrt{\frac{Sd^2}{10}}} = \frac{2}{\sqrt{0{,}489}} = 2{,}86$$

En se reportant à la table de student sur la ligne 9 d.d.l. on voit que : T0,05,9 = 2,26 < 2,86 < T0,01,9 = 3,25

On voit qu'on peut rejeter l'hypothèse nulle H_0 au risque d'erreur compris entre 5 et 1 % (en fait 2 %). Dans ce cas, H_1 signifie d ≠ 0, ce qui implique que l'hypotenseur a une influence sur la pression artérielle systolique.

A titre de curiosité (?) examinons ce qu'aurait donné la procédure incorrecte, celle de considérer le X_1 et le X_2 comme deux échantillons indépendants.

Pour chacun des échantillons, on calcule :

$m_1 = 15$ $\quad S_1^2 = 54$ à ddl=9

$m_2 = 13$ $\quad S_2^2 = 44$ à ddl=9

Les S_1 et S_2 sont peu différentes, on calcule un S commun :

$$S^2 = \frac{54 + 44}{18} = 5,44$$ On calcule:

$$t = \frac{|m_1 - m_2|}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{2}{1,04} = 1,92$$

On constate: t = 1,92 < t0,05, 18 = 2,10

La différence apparaît donc comme non significative. Cette perte de signification quand on passe de la procédure correcte (tenant compte de l'appariement) à une procédure incorrecte (qui néglige celui-ci) ne doit pas surprendre.

La procédure correcte ne tient compte que de la variation intra malade, alors que l'autre procédure prend en compte simultanément la variation intra-, mais aussi la variation inter malade.

A RETENIR

Pour les grands échantillons n ≥ 30

Pour comparer les moyennes de deux séries appariées, on forme pour chaque paire la différence des deux mesures et on compare la moyenne des n différences à 0 par l'écart réduit. $\varepsilon = \dfrac{|md|}{\dfrac{Sd}{\sqrt{nd}}}$

où md et sd désignent la moyenne et l'écart type estimés sur l'échantillon des n différences.

Si $|\varepsilon| < 1{,}96$ (pratiquement 2), la différence n'est pas significative (à 5%).

Si $|\varepsilon| \geq 1{,}96$ (pratiquement 2), la différence est significative, et le risque correspondant à ε, lu dans la table de l'écart réduit fixe le degré de signification.

Pour les petits échantillons n < 30

Pour comparer les moyennes de deux séries appariées de faible effectif, on forme pour chaque paire la différence des deux mesures et on compare la moyenne des différences à 0 par le rapport

$$t = \frac{|md|}{\dfrac{Sd}{\sqrt{nd}}}$$

où md et sd désignent la moyenne et l'écart type estimés sur l'échantillon des n différences.

Si $|t|$ est inférieur à la valeur lue dans la table de t pour le nombre de degré de liberté (n-1) et le risque 5%, les moyennes ne diffèrent pas significativement; dans le cas contraire les moyennes diffèrent significativement et le risque indiqué par la table pour la valeur de $|t|$ trouvée fixe le degré de signification.

N.B. Cette formule n'est applicable que si la différence est distribuée selon une loi normale.

CAS DE DEUX VARIABLES QUANTITATIVES

STATISTIQUE DESCRIPTIVE ET INFERENTIELLE

1- PRESENTATION DU PROBLEME

Très nombreux sont les cas où sur chaque sujet, on enregistre simultanément la valeur de plusieurs variables qualitatives et/ou quantitatives.

Ainsi en est-il des malades atteints d'une affection donnée lorsqu'on pratique un travail de recherche sur cette maladie.

Pour chacun, on prévoit à l'avance le type de la variable: qualitative ; quantitative et on prévoira les réponses possibles.

Nous nous limiterons à l'étude de DEUX variables simultanées. Déjà, avons-nous examiné le cas de deux variables qualitatives simultanées, et de l'analyse du tableau de contingence qui en résulte.

La liaison entre une variable qualitative et une variable quantitative se ramène à une comparaison de moyennes observées. Nous ne savons le faire que lorsque la variable qualitative a deux classes, car nous sommes amenés à comparer deux moyennes. Par contre, nous ne savons pas le faire si la variable qualitative a plus de deux modalités: la technique utilisée dans ces cas, l'analyse de variance, étant un peu délicate et au-delà du cadre de ce cours.

Dans ce chapitre, nous examinerons le cas de deux variables quantitatives simultanées.

2- STATISTIQUE DESCRIPTIVE

2-1- Dans la population: Covariance et corrélations théoriques

Quand on a affaire à deux variables simultanées, le problème principal est de savoir si elles sont dépendantes en probabilité ou non; si oui, de mesurer si possible leur degré de dépendance.

Considérons deux variables aléatoires X et Y continues prélevées sur le même sujet. On peut définir une loi de probabilité à deux dimensions. Si l'événement élémentaire est:
[x < x < x + dx] et [Y < Y < Y + dY], on peut définir sa probabilité par une densité de probabilité (f(x,y) telle que:
Probabilité (x < x < x+ dx, y < y < y + dy) = f(x,y) dxdy

Un cas particulier, mais très important, de densité de probabilité à deux dimensions, est la loi de LAPLACE GAUSS à deux dimensions. Elle a la forme d'un chapeau de coupe elliptique. En effet, les courbes de niveau sont des ellipses de même centre, même direction, même excentricité. De plus, la distribution de chacune des variables x et y prise isolément est une courbe de L.G. Ce sont les distributions dites marginales.
Elle est complètement déterminée par 5 paramètres que nous verrons dans quelques lignes: μx, μy, δx, δy, et p.

On peut alors déterminer:

μx = E(X) Espérance de x quel que soit Y

μy = E(Y) Espérance de Y quel que soit X

V(X) = δx² = E [(X-μx)²] Variance de x quel que soit y

V(Y) = δy²= E [(Y - μy)²] Variance de y quel que soit X.

On peut aussi déterminer la COV (X,Y) = E [(x-μx) (y-μy)]

Cette expression du double produit se nomme la covariance théorique de X et de Y.

On appelle <u>coefficient de corrélation linéaire</u> théorique la quantité:

$$r = \frac{\sum[(X-\mu x)(Y-\mu y)]}{\sqrt{\sum[(X-\mu x)^2]\sum[(Y-\mu y)^2]}} = \frac{\sum[(X-\mu x)(Y-\mu y)]}{\sigma x \sigma y}$$

C'est le quotient de la covariance par le produit des écarts types.

Il a les propriétés suivantes:

2-1-1- C'est un nombre pur (numérateur et dénominateur sont homogènes au produit XY).

2-1-2- Sa valeur absolue est comprise entre 0 et 1:

$0 < |r| < 1$ ou $-1 < r < +1$

2-1-3- La valeur absolue du coefficient de corrélation linéaire nous renseigne sur l'indépendance de la façon suivante:

1ère règle: SI $|r| = 1$, ALORS il y a une relation linéaire stricte entre X et Y de type $Y = a + bX$.

C'est-à-dire que la dépendance entre X et Y est totale: si X est connu, Y n'est plus aléatoire, il est connu et certain, et réciproquement. Que cette dépendance a une forme très particulière et très simple: une relation du premier degré. C'est ce qui justifie l'épithète de linéaire qui est accrochée au mot corrélation.

2ème règle: SI X et Y sont indépendants en probabilité, ALORS la covariance et le coefficient sont nuls. Mais la réciproque n'est pas vraie.

3ème règle: en fait le coefficient de corrélation mesure le degré de liaison linéaire. Quand il vaut 1 en valeur absolue, il y a une liaison linéaire parfaite. Quand il vaut 0, il n'y a aucune liaison linéaire, ce qui veut dire "aucune liaison ou liaison non-linéaire".

2-1-4- Quant au signe de r (et de la covariance), il a deux significations qui sont les suivantes: Si $r > 0$, ALORS X et Y varient dans le même sens. Quand X augmente, Y a tendance à augmenter. Ceci est d'autant plus sûr que r est plus proche de 1. Dans ce dernier cas, $Y = a + b X$ avec $b > 0$.

SI $r < 0$, ALORS X et Y varient en sens inverse: quand X augmente, Y a tendance à diminuer.

➢ STATISTIQUE INFERENTIELLE

3-1- Le coefficient r est bien entendu une grandeur aléatoire qui dépend des fluctuations d'échantillonnage.

Il se pose donc le même type de problème pour r que pour d'autres paramètres observés:

- Comparaison de r à une valeur r_0 théorique,
- Estimation de r_0 à partir de r,
- Comparaison de deux coefficients empiriques r_1 et r_2 de deux échantillons indépendants.

Des solutions existent pour résoudre ces trois problèmes. Elles reposent toutes sur l'hypothèse que la distribution est gaussienne à deux dimensions dans la population. En pratique, on prend beaucoup de liberté avec une telle hypothèse. Il suffit que la forme du nuage ne s'éloigne pas trop d'une ellipse!

Dans le cadre de ce cours, nous ne traiterons qu'un seul problème, et encore dans un cas particulier: **Comparaison de r à la valeur théorique $r_0 = 0$**

En toute rigueur, on a vu que la nullité de p n'implique pas de façon formelle l'indépendance des variables (l'implication est en sens inverse). Mais si le nuage de points a une forme "régulière", on peut admettre que tester la valeur théorique r = 0 signifie tester l'indépendance. Ce serait rigoureusement vrai si on était sûr que la population est gaussienne pour le couple (X,Y).

En pratique, et en dépit de toutes ces restrictions mentales, on teste la valeur théorique r = 0 pour tester l'indépendance de X et de Y.

On constate une valeur $r \neq 0$ qui semblerait suggérer qu'il y a un certain degré de dépendance linéaire. Cette valeur est-elle compatible avec l'hypothèse que dans la population $r_0 = 0$?

Le test est donc: H0: dans P $r_0 = 0$ et H_1: dans P $r_0 \neq 0$

La pratique du test est très simple. Elle ne suppose aucun calcul car des mathématiciens pleins de mansuétude l'ont calculé.

Cette table indique les lois de probabilité au quelles obéit r dans l'hypothèse où r = 0. Les lois au pluriel, car il y en a autant que d'effectifs possibles pour l'échantillon.

On trouve donc en colonne une série de lois correspondant au nombre de d.d.l.Celui-ci vaut N-2 (dans le calcul de r, interviennent deux paramètres calculés à partir des données $\bar{x}$ et $\bar{y}$).

Les valeurs lues sont r_{α}, N-2 telles que:

Proba (l r l > r_{α}, N-2) = α

Exemple:

Sur un échantillon d'effectif N, on a trouvé un coefficient de corrélation linéaire empirique de -0,40. Si N = 20, on considère la ligne N-2 = 18. On constate que: l r l < r0,5; 18

On n'est donc pas en droit de rejeter Ho et de conclure à une dépendance linéaire de X et de Y.

- Examinons le cas où N = 32 (r = -0,40).

Considérant la ligne N = 30, on constate:

r 0,05; 30 = 0,349 < l r l < r0,01;30 = 0,4487

Je constate que r est significativement différent de 0, et je conclus en rejetant H0. Ce faisant, je prends un risque d'erreur inférieur à 5% mais supérieur à 1%. De fait, la table nous dit qu'il est de l'ordre de 2%.

- Si maintenant, on avait N = 50, toujours avec r = -0,40:

Ici, il faut lire entre les lignes N= 50, et N = 45. Mais il est évident que: $|r|$ > $r_{0,01}$; 48

Donc, r est très significativement différent de 0, et je rejette H0 avec un risque inférieur à 1%.

Je conclus donc avec un risque de me tromper inférieur à 1% qu'il existe une liaison linéaire entre X et Y. Celle-ci est d'intensité

moyenne ($|r|$ = 0,40). De plus, elle exprime que X et Y varient en sens inverse (r négatif).

Cet exemple montre qu'une corrélation peut être très significative sans être forcément très forte.

A RETENIR

Le risque α correspondant à r peut être obtenu par la table du coefficient de corrélation pour d.d.l. = n- 2.

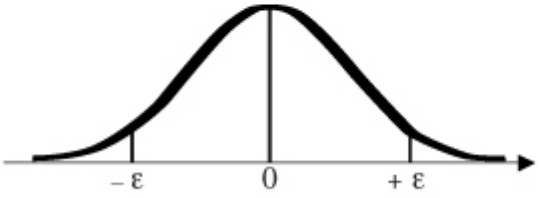
Si $\alpha \geq$ à 5%, la liaison n'est pas significative (à 5%).

Si α < 5%, la liaison est significative, et α mesure son degré de signification.

Table 1 :

Table de l'écart–réduit ϵ donnant la probabilité α pour que l'écartréduit soit supérieur ou égal à la probabilité extérieure de l'intervalle ϵ, + ϵ (d'après Fisher et Yates, *Statistical tables for biological, agricultural, and medical research*, Oliver & Boyd, Edimbourg).

La probabilité α s'obtient par addition des nombres inscrits en marge : pour ϵ = 1,960, la probabilité α = 0,00 + 0,05 = 0,05.



α	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695
0,10	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,20	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058
0,30	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,40	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690
0,50	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,60	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399
0,70	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,80	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138
0,90	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013

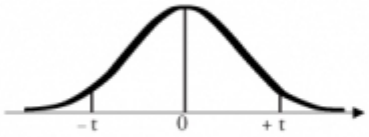
Table pour les petites valeurs de la probabilité

α	0,001	0,000 1	0,000 01	0,000 001	0,000 000 1	0,000 000 01	0,000 000 001
ε	3,29053	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

Table 2 :

– Table de Student donnant la probabilité α pour que t soit supérieur ou égal en valeur absolue à une valeur donnée en fonction du degré de liberté d.d.l. (d'après Fisher et Yates, *Statistical tables for biological, agricultural, and medical research,* Oliver & Boyd, Edimbourg).

Pour d.d.l. = 5 et t = 2,572, la probabilité α est égale à 0,05.

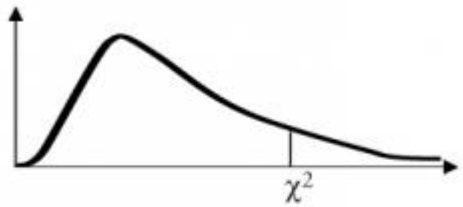


d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,703	1,100	1383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Table 3 :

Table de χ^2 donnant la probabilité α pour que χ^2 soit supérieur ou égal à une valeur donnée en fonction du degré de liberté d.d.l. (d'après Fisher et Yates, *Statistical tables for biological, agricultural, and medical research,* Oliver & Boyd, Edimbourg).

Pour d.d.l. = 5 et χ^2 = 11,070, la probabilité α est égale à 0,05.



d.d.l. \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703