PASS **2023-2024**

FICHES DE STATISTIQUES

Rédaction et mise en page par Ludovic Rousseau

Généralités et définitions

Définitions :

- Incidence : Nombre de **nouveaux cas** (ex : nombre de malades) **au cours d'une période t**
- Prévalence : **Proportion** de personnes malades **à un INSTANT t**

Axiomes des probabilités :

1) $0 \leq P(A) \leq 1$
2) $P(\Omega) = 1$
3) Toute famille dénombrable d'événements deux à deux disjoints (on dit aussi : deux à deux incompatibles), A 1 , A 2 , … satisfait :

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots) = \sum_{i=1}^{+\infty} \mathbb{P}(A_i)$$

Formule de Bayes avec Proba totale :

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

P conditionnelle à 3 evt° :

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

Toujours vrai : **$P(A \cup B) = P(A)+P(B) - P(A \cap B) \leq P(A)+P(B)$**

Indépendance et incompatibilité:

- Indépendance:
  - $P(A \cap B) = P(A) \times P(B)$
  - $P(A \mid B) = P(B \mid A) = P(A)$
- Incompatible :
  - $(A \cap B) = 0$
  - $(A \cup B) = 0$

Deux événements incompatibles ne sont pas indépendants.
Deux événements indépendants ne sont pas incompatibles.
Certains événements ne sont ni indépendants, ni incompatibles.

Tests et Intérêts diagnostiques :

La **sensibilité** d'un signe pour une maladie est la probabilité que le signe soit présent si le sujet est atteint de la maladie considérée. **Se = *P (S / M)***

La **spécificité** d'un signe pour une maladie est la probabilité que le signe soit absent si le sujet n'est pas atteint de la maladie. **Sp = *P ($\bar{S}$ / $\bar{M}$)***

VPP = P (M / S)

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

Se et **Sp : indépendantes** de la **prévalence**.
VPP et **VPN : dépendent** de la **prévalence**

Courbe **ROC** : **Se = *f* (1 − Sp)**

Si : **Seuil Augmente : Se ↓ VP↓ FP↓ Sp↑ FN↑ VN↑**
Si : **Seuil Diminue : Se↑ VP↑ FP↑ Sp↓ FN↓ VN↓**

Un test est dit à **valeur diagnostique** lorsque **Se + Sp > 1.**
Lorsque l'on inverse le seuil, la courbe ROC s'inverse aussi.

Quand **P(M) Augmente :**
VPP Augmente et **VPN Diminue** et inversement.

Formules complémentaires en vrac :

$$P(M) = P(M/S) \times P(S) + P(M/\bar{S}) \times P(\bar{S})$$

$$= VPP \times P(S) + (1 - VPN) \times P(\bar{S})$$

$$Se = P(S|M) = [P(M/S) \times P(S)] / P(M) = [VPP \times P(S)] / P(M) \approx \frac{VP}{VP + FN}$$

$$Sp = P(\bar{S}|\bar{M}) = [P(\bar{M}/\bar{S}) \times P(\bar{S})] / (1\text{-}P(M)) = [VPN \times P(\bar{S})] / (1\text{-}P(M)) \approx \frac{VN}{VN + FP}$$

$VP = Se \times M = P(M/S) \times S = VPP \times S$	$FP = (1 - Sp) \times \bar{M} = P(\bar{M}/S) \times S = (1 - VPP) \times S$
$FN = (1 - Se) \times P(M) = P(M/\bar{S}) \times P(\bar{S}) = (1 - VPN) \times P(\bar{S})$	$VN = Sp \times P(\bar{M}) = P(\bar{M}/\bar{S}) \times P(\bar{S}) = VPN \times P(\bar{S})$

Variables aléatoires :

L'entité sur laquelle peut s'observer la variable aléatoire s'appelle l'**unité statistique.**

Un échantillon de taille n *d'une variable aléatoire X est* ***un ensemble*** $X_1, X_2, ..., X_n$ *de n variables aléatoires, indépendantes entre elles, et ayant chacune la même distribution que X. On peut donc dire qu'un échantillon de valeurs de X est* ***une*** *réalisation de l'échantillon de la variable X tel qu'il vient d'être défini.*

1) Espérance propriétés :

- $E(k \times X) = k \times E(X)$
- $E(X + k) = E(X) + k$
- $E(k) = k$
- $E(X + Y) = E(X) + E(Y)$
- $E(X - E(X)) = E(X) - E(X) = 0$
- $E(g(X)) = \sum_{i=1}^{n} g(x_i)\, p_i$
- $E(X^2) = \sum_{i=1}^{n} (x_i)^2\, p_i$
- $Z = X - \mu_X = X - E(X)$, $E(Z) = 0$
- Unité de $E(X)$ = unité de X

2) Variance et Écart type :

$$\sigma^2 = var(X) = E[(X - E(X))^2] = \sum_{i=1}^{n} (x_i - \mu_X)^2 \times P(X = x_i)$$

$var(X + Y) = var(X - Y) = var(X) + var(Y)$ SI ET SEULEMENT SI X et Y sont INDEPENDANTES

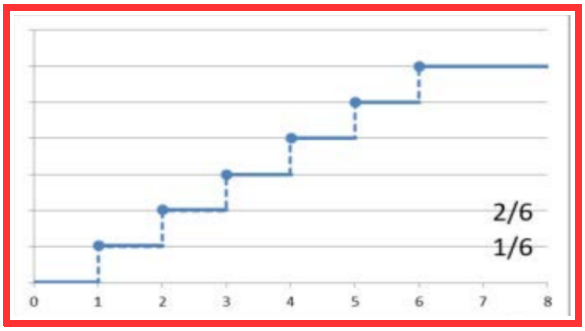
L'unité de la variance est l'unité de la moyenne au carré

- $var(X) = E(X^2) - E(X)^2$
- $var(X) \geq 0$
- $var(X + a) = var(X)$
- $var(aX) = a^2\, var(X)$
- $var(a) = 0$

F(x) : Fonction de répartition

$F(x) = P(X \leq x)$

$P(a < X \leq b) = F(b) - F(a)$



Densité de probabilité :

- $f(x)$ la dérivée de F(X). f(x) est la densité de probabilité de X au point x.
- $f(x)$ n'est **pas** une probabilité.

$f(x)dx$ est une probabilité (et c'est aussi l'aire sous la courbe f dans l'intervalle]x, x + dx]

$$\mu_X = E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx$$

$$\sigma^2 = var(X) = E(X^2) - E(X)^2 = \int_{-\infty}^{+\infty} x^2 f(x)\, dx - \left(\int_{-\infty}^{+\infty} x\, f(x)\, dx\right)^2$$

Distributions Usuelles :

Loi Normale

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

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

Distribution Centrée : $\mu = 0$

Distribution Réduite : $\sigma^2 = \sigma = 1$

Distribution normale centrée réduite N(0, 1) est donc définie par la formule

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

Soit une variable X distribuée selon une loi normale d'espérance μ et d'écart-type σ.

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

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

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

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

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

Approximation de la distribution Binomiale par la loi Normale :

Lorsque *n* est grand, **et** que π et $1 - \pi$ et ne sont **pas trop proches de 0**

(*en pratique si :* $n\pi > 5$ et $n(1-\pi) > 5$) *a*vec $K \sim B(n, \pi) \approx N(\mu = n\pi, \sigma^2 = n\pi(1-\pi))$

Approximation de la loi de Poisson par la loi normale

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

La loi d'une variable X suivant un $\chi^2(n)$ tend vers une loi normale lorsque $n \to +\infty$

On a donc, après avoir centré et réduit cette variable : $\frac{X-n}{\sqrt{2n}} \sim N(0, 1)$

1) Bernoulli : $P(X = k) = (k \text{ parmi } n)\ \pi^{k}\ (1 - \pi)^{n-k} = \frac{n!}{k!(n-k)!}\ \pi^{k}\ (1 - \pi)^{n-k}$

$Var(X) = \sigma^2 = \pi - \pi^2 = \pi(1 - \pi)$

2) Binomiale :

Y~ B(n , π)
→ μ = nπ
→ $Var(X) = \sigma^2 = n\pi(1-\pi)$

3) Chi 2 : χ^2

E(X) = *n*
Var (X)= *2n*
n = ddl

si *X* est une variable de Bernoulli,

— sa moyenne «vraie»= π

— sa varianc e«vraie» $Var(X) = \sigma^2 = \pi(1 - \pi)$

Loi de Poisson : *(normalement HP mais si jamais jl'ai mis..)*

C'est la loi du nombre d'événements observé pendant une période de temps donnée dans le cas où ces **événements** sont **indépendants et faiblement probables**.

Ex : Nombre d'accidents, apparition d'anomalies diverses, gestion des files d'attentes, nombre de colonies bactériennes dans une boîte de Pétri, etc.
Def : Soit X la variable aléatoire représentant le nombre d'apparitions indépendantes d'un événement faiblement probable dans une population infinie. ***La probabilité d'avoir k apparitions*** *de l'événement est :*

$$Pr(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \quad \text{avec } \lambda > 0 \text{ et } k \in \mathbb{N} :$$

Si : X ~ P (λ) :
- E(X) = λ
- σ^2 = λ

Valeurs remarquables :

On appelle $Pr(X=0) = p = e^{-\lambda}$
- $\lambda = -\ln p$
- $Pr(X=1) = \lambda e^{-\lambda}$
- $Pr(X=2) = Pr(X=1) \times \frac{1}{2}\lambda$
- $Pr(X=3) = Pr(X=2)\ \lambda/3$
- $Pr(X=k) = Pr(X=k-1)\ \lambda / k$

Lien avec la loi binomiale

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

Application : On approxime une loi de Bernoulli a une loi de poisson :

A) Risque individuel = proba : π CONNUE
n nombre de patients (quelque soit-il ex :naissance administrations...)
X → nb d'effets indésirables $Pr(X = 0) = (1-\pi)^n \approx e^{-n\pi}$

B) X=0 π INCONNUE

On observe l'événement ***X=0 pour n traitements*** ou qqch du genre .
On approche X par : $X \sim P(n\pi)$ (càd $\lambda = n\pi$)
méthode : on écarte les valeurs de π pour lesquelles X=0 serait « invraisemblable »

L'événement « X=0 parmis n répétitions » est **invraisemblable si :** $e^{(-n\pi)} < 5\ \%$
on en déduit le risque individuel :

$$\Rightarrow \pi \in [0\ ;\ 3/n[$$
$$\Rightarrow \pi < 3/n$$

Conclusion :

1) Quand on observe 0 effet indésirable parmi *n* répétitions, ceci est **compatible** avec un **risque individuel compris entre 0 et 3/*n*.** En revanche, les risques supérieurs à 3/*n* sont jugés invraisemblables**.**

2) Ne signifie pas qu'il y a 5 chances sur 100 pour que le risque sanitaire soit de 3/n

3) MAIS : « Si le risque de mort était de 3/n (il ne l'est peut-être pas), il y aurait 5 % de chance d'observer 0 décès parmi n prescriptions

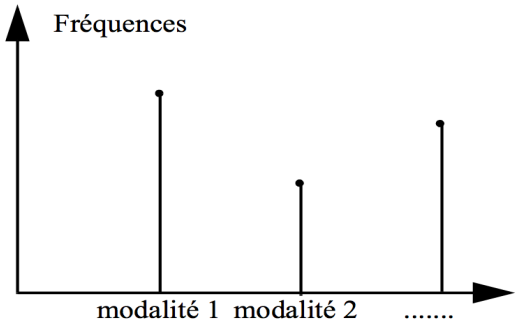
Statistiques déscriptives :

Variable Qualitative

La variable est décrite par la suite des probabilités des différentes modalités.

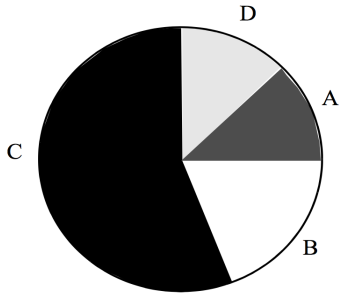
Diagramme en Baton :

Si l'on connaissait ces probabilités, on produirait le diagramme en bâtons (ou répartition « vraie ») de cette variable;
On va produire la **répartition observée** par substitution aux probabilités inconnues des fréquences observées :



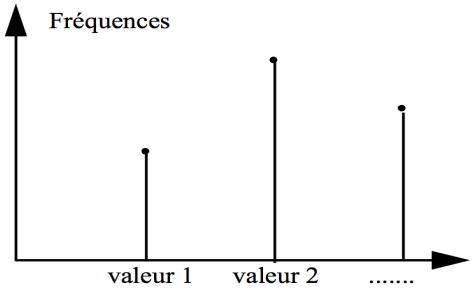
Représentation en Camembert:

Les différentes modalités sont représentées par secteurs angulaires d'angles au centre proportionnels aux **fréquences observées**.



Variable Quantitative Discrète :

La situation est similaire si ce n'est qu'il existe un ordre et une échelle naturels en abscisses, la répartition observée se nomme également **Histogramme en bâtons :**



Variable Quantitative Continue:

Histogramme

Variables continues :, on **représente** les données graphiquement **d'une façon qui soit proche de la représentation d'une densité de probabilité d'une variable aléatoire continue**.

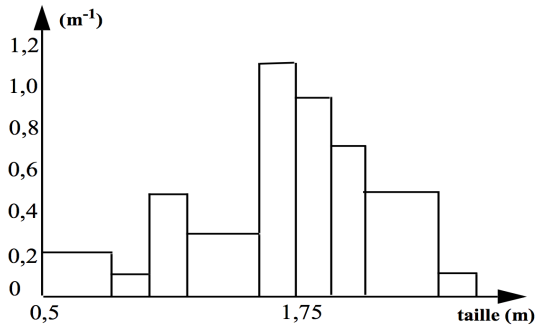
On découpe l'ensemble du domaine des valeurs possibles de la variable étudiée en intervalles contigus dont on choisit le nombre et les bornes.

Représentation indirecte de la fréquence des valeurs observées comprises entre deux bornes consécutives **:**

Surface d'un rectangle :

- **base → l'intervalle entre les bornes**.
- **Hauteur → rapport : fréquence observée de ces valeurs / largeur** (différence entre les bornes ou largeur de la classe).

Si la taille de l'échantillon croît, la **surface de chaque rectangle** tend vers la **probabilité** que la variable ait une valeur incluse dans l'intervalle correspondant



INDICATEUR DE DISPERTIONS DES VALEURS :

Variance OBSERVÉE :

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

Écart type observé : $s = \sqrt{s^2}$

***SYNONYMIE et TERMINOLOGIE* :**

Moyenne VRAIE = Espérance Mathématique = Moyenne théorique

sont synonymes et sont des grandeurs théoriques

s^2 = Variance OBSERVÉE

σ^2 = Variance VRAIE

m = moyenne OBSERVÉE

μ = moyenne THÉORIQUE

p= proportion OBSERVÉE

$\prod$ = π = Proportion VRAIE

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

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

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

	moyenne observée	**espérance, ou moyenne « vraie »**
variable discrète	$m = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\mu = \sum_{j=1}^{k} \text{val}_j Pr(\text{variable} = \text{val}_j)$
variable continue	$m = \frac{1}{n}\sum_{i=1}^{n} x_i$	$\mu = \int_{\Re} x f(x) dx$
variable de Bernoulli	m est notée p	$\mu = Pr(\text{variable} = 1)$ est notée Π

4. **Variances (variables quantitatives)**

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

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

Fluctuation de la moyenne observée :

Variable aléatoire moyenne arithmétique

***M* (ou *Mn* si nécessaire) est la VARIABLE ALEATOIRE MOYENNE ARITHMETIQUE ASSOCIEE A LA VARIABLE ALEATOIRE *X*, FONDEE SUR *n* REPETITIONS**

Dans le cas où ***X* est une variable de Bernoulli, *Mn* sera notée *Pn*** (et *M* simplement *P*). Il s'agit d'une variable aléatoire proportion dont on connaît déjà pratiquement la distribution puisque

$n*P_n \sim B(n, \pi)$

Formules :

ESPÉRANCE : $$E(M_n) = E(X)$$

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

ÉCART-TYPE : $$\sigma(M_n) = \frac{1}{\sqrt{n}}\sigma(X)$$

Dans le cas où ***X* est une variable de Bernoulli de paramètre π** **(on a : (*Pr*(*X* = 1) = π)**

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

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

Le Théorème Central Cimite

Seconde propriété de la variable aléatoire moyenne arithmétique

Soit *X* une variable aléatoire quantitative d'espérance mathématique μ, de variance « vraie » σ^2.

Soit *Mn* la variable aléatoire moyenne arithmétique associée à *X* construite sur *n* répétitions.

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

DISTRIBUTIONS :

La distribution de *Mn* est exactement une loi normale (la mention *à peu près* est inutile), **quel que soit *n***, **si *X*** elle-même **est gaussienne** (= est distribuée normalement)

si *X* n'est pas gaussienne :

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

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

Dans ce cas, on a : $\mu=\pi$, $\sigma^2 = \pi(1-\pi)$

On aura donc :

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

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

Intervalle de Pari (I.P)

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

Lorsqu'une variable aléatoire X remplit les conditions du TCL on peut construire :

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

$$Pr(a<Mn<b) = 1-\alpha$$

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

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

Cas d'une variable de Bernoulli, $\mu = \pi$ et $\sigma^2 = \pi(1-\pi)$

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

C'est la Probabilité 1 - α *(ici 0,95)* d'obtenir une valeur de la moyenne OBSERVÉE comprise dans cet intervalle

Rq : Probabilité **α** *de se tromper (ici 5% du temps)*

Rq : Dans le cas d'une variable de ***Bernoulli****, on remplace la moyenne observée par la* ***PROPORTION observée***

Longueur de l'I.P : $2u_\alpha \frac{\sigma}{\sqrt{n}}$

REMARQUES :

- Pour diviser par 2 la longueur de l'IP, il faut un échantillon 4 fois plus grand (2^2)

- La Longueur de $IP_{1-\alpha}$ décroît avec *n*

- Si $\alpha' < \alpha$ Alors : Longueur $IP_{1-\alpha'}$ > Longueur $IP_{1-\alpha}$

Intervalle de Confiance (I.C)

= Estimation par intervalle

De façon générale, **l'intervalle de confiance au risque α *d'une valeur que l'on cherche à estimer* est un intervalle qui contient avec une probabilité 1 – α la valeur VRAIE**

On parlera alors d'intervalle de confiance DE NIVEAU 1 – α ou d'intervalle de confiance AU RISQUE α

La valeur α sera alors le risque (ou la probabilité) pour qu'un intervalle de confiance ne contienne pas la proportion/moyenne « vraie » π/μ.

I.C approché d'une proportion « vraie » :

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

Notons $\prod_{min}$ et $\prod_{max}$ les bornes de cet intervalle. Cette approximation n'est jugée satisfaisante que sous les **CONDITIONS DE VALIDITÉ :** $n \prod_{min} > 5$ **ET** $n(1-\prod_{max}) > 5$

I.C approché d'une moyenne « vraie » : *(variable continue)*

Le calcul de cet intervalle suppose en outre le calcul de la variance ***observée*** s^2

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

CONDITION DE VALIDITÉ : $n \geq 30$

Si la variable étudiée est NORMALE, alors, et **SANS AUTRE CONDITION DE VALIDITÉ**, un intervalle de confiance de niveau 1 - α a pour expression

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

où t_α est associé à une nouvelle **distribution**, dite de **Student, à (n-1) degrés de liberté**

Remarque *(pour une variable normale encore) Si la variance « vraie » de la variable étudiée,* σ^2 *est connue, l'intervalle de confiance a la forme suivante :*

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

Précision Des Intervalles Et Nombre De Sujets

IC d'une proportion de longueur 2*i*, il faut *(au moins* **sujets** *(au risque α)*

$$u_\alpha^2 \frac{p(1-p)}{i^2}$$

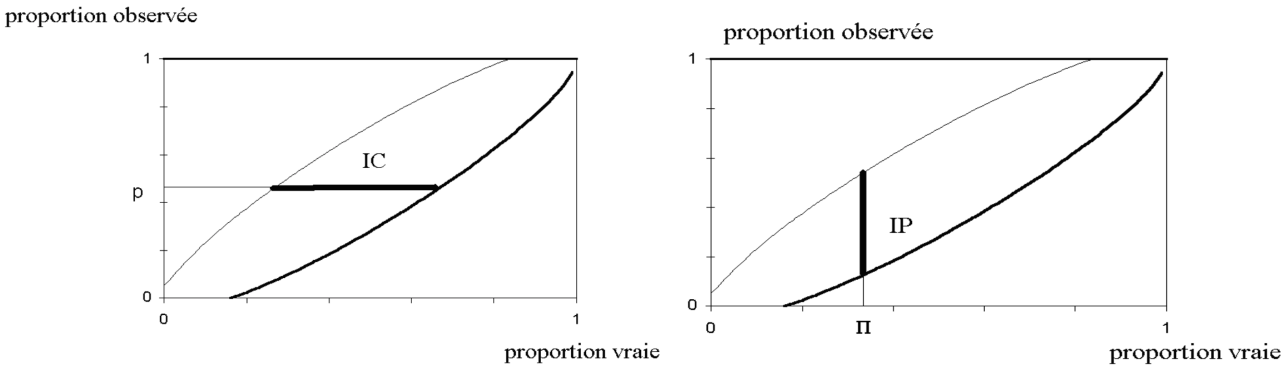
i* = *demi longueur d'intervalle

IC d'une Moyenne de longueur 2*i*, il faut *(au moins* **sujets** *(au risque α)*

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

i* = *demi longueur d'intervalle

Représentation Graphique des Intervalle de Confiance et de Pari :



CONFIANCE : on connais p/m on veut estimer ∏/μ **PARI : on Connais ∏/μ on cherche p/m**

Estimation ponctuelle

Définition : A partir d'un échantillon ($X1$, $X2$, ..., Xn) de la variable aléatoire X, on construit une nouvelle **Variable aléatoire *t(X1, X2, ..., Xn)*** dont les réalisations « se rapprochent » de la valeur **Θ** . Cette nouvelle variable est appelée **estimateur** de **Θ** . *Pour simplifier, cette variable* ***t(X1, X2, ..., Xn)*** *est* ***notée Tn ou T*** .

Par exemple ***t(X1, X2, ..., Xn) = Mn =*** $\frac{1}{n}\sum_{i=1}^{n} X_i$ *« se rapproche » de l'espérance de X*

Les estimateurs sont des fonctions des échantillons : ce **sont** donc des **variables aléatoires** qui possèdent une densité de probabilité, et le plus souvent, une moyenne (espérance mathématique) et une variance. *Ces deux grandeurs permettent de comparer, dans une certaine mesure, les estimateurs entre eux.*

Le Biais : Le **biais** d'un estimateur, noté $B(T)$, est la différence moyenne entre sa valeur et celle de la quantité qu'il estime.

$B(T) = E(T - \Theta) = E(T) - \Theta$ **Si *B*(T)=0** on parle d'**Estimateur Sans Biais**

Variance d'un Estimateur : $var(T) = E[T - E(T)]^2$

Si deux estimateurs sont sans biais, le ***meilleur*** *est celui qui a la* ***variance la plus petite*** *: en effet, ses valeurs sont « en moyenne » plus proches de la quantité estimée*

Erreur Quadratique Moyenne : *permettant de comparer des estimateurs entre eux, qu'ils soient biaisés ou sans biais*

$$EQM(T) = var(T) + [E(T) - \Theta]^2 = var(T) + B(T)^2$$

l'erreur quadratique moyenne des estimateurs sans biais est égale à leur variance. ***Lorsqu'on compare deux estimateurs, on considère que le meilleur est celui qui présente l'erreur quadratique moyenne la plus faible.***

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

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

***Remarque*: S_n^2 est un estimateur sans biais de σ^2**

Principe général des tests d'hypothèses

Etape 1

Avant le recueil des données.

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

H_0 : hypothèse nulle

Exemple : la fréquence « vraie » d'apparition du cancer chez les souris traitées est 0,2, ce qui se transcrit par $\varphi = 0{,}2$ (plus généralement $\varphi = \varphi_0$).

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

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

H_1 : hypothèse alternative

Exemple : la fréquence « vraie » d'apparition du cancer chez les souris traitées est différente de 0,2, qui se transcrit par $\varphi \neq 0{,}2$ (généralement $\varphi \neq \varphi_0$).

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

Etape 2

Avant le recueil des données.

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

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

Si possible, vérifier les conditions de validité.

Etape 3

Avant le recueil des données.

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

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

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

Etape 4

Avant le recueil des données.

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

Exemple : $z = \dfrac{p_{\text{réellement observé}} - \varphi_0}{\sqrt{\dfrac{\varphi_0(1-\varphi_0)}{n}}}$

Alors décider que :

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

Etape 5

Recueil des données

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

Etape 6

Interprétation des résultats

Cette étape concerne l'interprétation des résultats en des termes compatibles avec le problème médical initialement soulevé, et concerne en particulier le problème de la causalité. Ce point sera détaillé au chapitre 15.

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

Notions Importantes :

α = Proba de conclure H1 alors que H0 est vraie

(risque de 1ère espèce)

si on se fixe ***α = 0,*** *on ne conclut* ***jamais, H0 n'est jamais rejetée***

Puissance = 1-ß = Probabilité de Rejeter Ho (face à une hypothèse alternative) alors que H_1 est VRAI

ß = Probabilité de NE PAS Rejeter Ho alors que H_1 est VRAI

ß s'appelle le risque de deuxieme espèce

Si on fait PLUSIEURS TESTS

Probabilité de CONCLURE À TORT AU MOINS UNE FOIS :

$\longrightarrow Pr = 1 - (1-\alpha)^n$

Probabilité de NE PAS REJETER Ho au moins une fois Alors qu'il aurait fallu

$\longrightarrow Pr = 1 - (1-\beta)^n = 1 - (P)^n$

(où P est la puissance)

Résumé du chapitre

A. Etapes de mise en œuvre des tests :

1. Examiner le problème médical, aboutir à une formulation sous forme d'une question simple mettant en jeu deux hypothèses H_0 (précise, dite hypothèse nulle) et H_1 (contraire de H_0, dite hypothèse alternative). Enoncer ces hypothèses.
2. Construire un paramètre dépendant des données à venir dont on connaisse la distribution si H_0 est juste.
3. Choisir le seuil α ; $\alpha = 0,05$
4. Mettre en place la règle de décision sur la base d'un intervalle de pari au risque α.
5. Faire l'expérience, les calculs et conclure sur le plan statistique. En particulier indiquer le degré de signification du test en cas de rejet de l'hypothèse nulle.
6. Se livrer à une interprétation médicale des résultats du test (ce point sera revu au chapitre 15).

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

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

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

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

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

Tests Usuels : Les Hypothèses

Test d'égalité d'une proportion « vraie » à une valeur donnée *(ou test de comparaison d'une proportion observée à une valeur donnée)* ***valide que lorsque : $n\prod o \geq 5$ et $n(1-\prod o) \geq 5$***

H0 : la proportion « vraie » est égale à $\prod 0$ (proportion hypothétique ou supposée qu'on se donne pour le test). H0 : $\prod = \prod o$

H1 : la proportion « vraie » est différente de $\prod 0$. H1 : $\prod \neq \prod o$

Test d'égalité de deux proportions « vraies »*(ou test de comparaison de deux proportions observées)* ***valide que lorsque :*** $n_A\hat{\Pi} \geq 5, n_A(1-\hat{\Pi}) \geq 5$ ***et*** $n_B\hat{\Pi} \geq 5, n_B(1-\hat{\Pi}) \geq 5$

H0 hypothèse nulle : les fréquences « vraies » sont égales $\prod_A = \prod_B$

H1 hypothèse alternative : les fréquences « vraies » sont différentes $\prod_A \neq \prod_B$

Test d'égalité d'une moyenne « vraie » à une valeur donnée *(ou test de comparaison d'une moyenne observée à une valeur donnée)* ***valide que lorsque $n \geq 30$***

H0 :la moyenne «vraie» est égale à la valeur donnée μ_0 : $\mu = \mu_0$

H1 :$\mu \neq \mu_0$

Test de symétrie d'une variable (*X*) par rapport à une valeur donnée (μ_0) : *test de Wilcoxon*

H0 : les variables X - μ_0 et μ_0 - X ont même densité de probabilité

H1 : les variables X - μ_0 et μ_0 - X n'ont pas la même densité de probabilité

Test d'égalité de deux moyennes « vraies » *(ou test de comparaison de deux moyennes observées)* ***valide que lorsque nA et nB ≥ 30***

H0 hypothèse nulle : les moyennes « vraies » dans les deux populations sont égales $\mu_A = \mu_B$

H1 hypothèse alternative : $\mu_A \neq \mu_B$

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

H0 les densités de probabilité coïncident dans les deux populations : ***fA = fB***

H1 les densités de probabilité ne coïncident pas : ***fA ≠ fB***

- Test de comparaison de deux moyennes observées sur séries appariées

Valide si n ≥ 30 **H0** :la moyenne «vraie» de *d* est nulle, soit ***d=0*** **H1** :la moyenne «vraie» de *d* est non nulle, soit ***d ≠ 0***

- Test de symétrie de la distribution des différences

H0 : La densité de probabilité de la variable aléatoire *d* est **symétrique par rapport à zéro.**

H1 : La densité de probabilité de la variable *d* n'est **pas symétrique par rapport à zéro**

Résumé du chapitre

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

$$z = \frac{p - \varphi_0}{\sqrt{\frac{\varphi_0(1-\varphi_0)}{n}}}$$; v.a. ~ N(0, 1) ; validité $n\varphi_0 \geq 5$ et $n(1 - \varphi_0) \geq 5$

2. Comparaison de deux proportions observées

$$z = \frac{p_A - p_B}{\sqrt{\frac{\hat{\Pi}(1-\hat{\Pi})}{n_A} + \frac{\hat{\Pi}(1-\hat{\Pi})}{n_B}}}$$; v.a. ~ N(0, 1) ; $\hat{\Pi} = \frac{n_A p_A + n_B p_B}{n_A + n_B}$

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

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

$$z = \frac{m - \mu_0}{\sqrt{\frac{s^2}{n}}}$$; v.a. ~ N(0, 1) ; validité $n \geq 30$

4. Test de symétrie d'une variable par rapport à une valeur donnée
Ordonner les valeurs absolues des écarts à la valeur donnée et calculer T^+, somme des rangs des écarts positifs.

$$z = \frac{T^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$$; v.a. ~ N(0, 1) si $n > 15$; v.a. ~ Wilcoxon sinon.

5. Comparaison de deux moyennes observées

$$z = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$$; v.a. ~ N(0, 1) ; validité n_A et $n_B \geq 30$

6. Test d'égalité de deux distributions (on suppose $n_A \leq n_B$)
Ordonner les valeurs. T_A = somme des rangs des données A. $\delta = T_A - \frac{n_A(n_A + n_B + 1)}{2}$.
$T'_A = T_A$ - 0,5 si $\delta > 0$, $T'_A = T_A$ + 0,5 sinon

$$z = \frac{T'_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B(n_A + n_B + 1)/12}} \sim N(0, 1)$$ lorsque n_A **ou** $n_B > 10$

$$z = \frac{T_A - n_A(n_A + n_B + 1)/2}{\sqrt{n_A n_B(n_A + n_B + 1)/12}}$$ ~ Mann-Whitney-Wilcoxon si n_A **et** $n_B \leq 10$

7. Comparaison de deux moyennes observées sur séries appariées
On utilise le test 3 en comparant la moyenne de la variable différence d à 0
8. Test de symétrie des différences (séries appariées)
On utilise le test 4 de symétrie de la variable d par rapport à 0.

Tests concernant les Variables Qualitatives

1- Comparaison d'une répartition observée à une répartition donnée : Test du χ^2 d'ajustement

Deux hypothèses sont en présence :

1. la répartition « vraie » de la variable dans la population étudiée coïncide avec la réparti- tion donnée (hypothèse nulle H0)
2. les répartitions diffèrent (hypothèse alternative H1)

Avec les notations précédemment introduites, cela s'écrit :

H0 : hypothèse nulle : $\varphi_i = \varphi_{hi}$ pour tous les i de 1 à k.

H1 : hypothèse alternative : $\varphi_i \neq \varphi_{hi}$ pour au moins une modalité, c'est-à-dire pour au moins un i.

Règle de décision

Si $Q_c \leq K_{ddl,\alpha}$ on ne conclut pas

Si $Q_c > K_{ddl,\alpha}$ H0 est rejetée. Cela signifie que l'on conclut que la répartition du caractère étudié **ne coïncide pas (**ou **ne s'ajuste pas)** avec la répartition donnée

On admet, en formulant cette conclusion, un risque d'erreur égal à α.

2- Comparaison de plusieurs répartitions observées : Test du χ^2 d'homogénéité

H0 : $\varphi 1_i = \varphi 2_i$ pour toutes les modalités i.

H1 : $\varphi 1_i \neq \varphi 2_i$ pour au moins une modalité i

TABLEAU DE CONTINGENCE !!

Calcul de Q en faisant la somme de TOUTES les cases (calculette+++)

Règle de décision

Si $Q_c \leq K_{ddl,\alpha}$ On ne conclut pas, Il n'est pas démontré que les deux répartitions « vraies » diffèrent.

Si $Q_c > K_{ddl,\alpha}$ On conclut que les deux répartitions observées diffèrent significativement.

On admet, en formulant cette conclusion, un risque d'erreur égal à α.

3- Test d'Indépendance

TRES IMPORTANT (des erreurs sont souvent commises)

HYPOTHESE NULLE : LES DEUX VARIABLES SONT INDEPENDANTES

HYPOTHESE ALTERNATIVE : LES DEUX VARIABLES SONT LIEES

TABLEAU DE CONTINGENCE pour trouver Q *(Qc=Q calculé)*

la règle de décision s'établit comme suit :

- Si $Q_c \leq K_{ddl,\alpha}$ on **ne rejette pas l'hypothèse d'indépendance** des deux variables.

- Si $Q_c \leq K_{ddl,\alpha}$ on **rejette l'hypothèse d'indépendance** des deux variables. On dira alors que les deux variables sont liées, **au risque α**

Cas particulier d'homogénéïté : Deux répartitions appariées

Test du χ^2 de Mac Nemar

Dans le cas de données appariées, **les résultats doivent être présentés comme dans le tableau ci-dessous, en croisant les succès et échecs des deux crèmes chez les mêmes sujets**. *Par exemple, on voit que 4 sujets ont eu un succès avec le placebo et un échec avec la crème X.*

		Crème X	
		Succès	Echec
Crème Placebo	Succès	7 (a)	4 (c)
	Echec	19 (b)	10 (d)

H0 : les re partitions « vraies » de la variable succe s/e chec sont identiques

H1 : les re partitions « vraies » sont diffe rentes

H0 : $\varphi 1i = \varphi 2i$ *(dans notre cas H0 :* $\varphi 1 = \varphi 2$) (puisque c les memes sujets :$\varphi 1 = \varphi a + \varphi b$)

H1 : $\varphi 1i \neq \varphi 2i$ *(dans notre cas H1 :* $\varphi 1 \neq \varphi 2$) (idem $\varphi 1 = \varphi a + \varphi c$)

On est chez les memes sujets donc :

H0 s'e crit aussi H0 : $\varphi a + \varphi b = \varphi a + \varphi c$ *, et en simplifiant*

H0 : $\varphi b = \varphi c$

H1 :$\varphi b \neq \varphi c$

ON A DONC :

$$Q = \frac{(b-c)^2}{(b+c)}$$

Q suit une distribution du χ^2 à 1 degré de liberté

Conditions de validité* sont les mêmes dans les cas non appariés: *Effectifs attendus ≥ 5 :

→ ((b+c)/2) ≥ 5

Cf **2-** *pour la Conclusion*

Résumé du chapitre

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

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

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

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

Comparaison de plusieurs répartitions observées (homogénéité)

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

Test d'indépendance de deux variables qualitatives

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

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

Méthodologie des études épidémiologiques

1. **L'essai contrôlé randomisé permet de mesurer l'effet causal** d'une intervention de santé, un traitement par exemple.

2. La **randomisation** qui **consiste à tirer au sort l'attribution de l'intervention**, **permet d'assurer que les individus** constituant l'échantillon **sont comparables** en tout (homogènes) sauf pour ce qui concerne le caractère contrôlé.

3. **Dans un essai randomisé, le critère de jugement est la variable** qui sera **comparée** entre les groupes pour juger de l'efficacité de l'intervention. On distingue critères de jugements objectifs (ex : décès) et subjectifs (ex : douleurs), ces derniers pouvant être facilement influencés par d'autres effets que les effets propres de l'intervention

4. **L'effet thérapeutique** dans un essai est la **somme de l'effet pharmacologique propre et de l'effet placebo.**

5. **La mise en aveugle** qui signifie que **ni le patient, ni le médecin qui le suit, ni l'évaluateur du critère ne savent dans quel groupe est randomisé le patient**, **est utilisée pour limiter les biais**

6. **L'analyse en intention-de-traiter** signifie que l'on compare le critère de jugement entre les groupes tels qu'ils ont été constitués par la randomisation. Elle **implique que tous les patients randomisés sont conservés dans l'analyse**

7. **Dans une Étude d'Observation, il n'est PAS POSSIBLE DE CONCLURE CAUSALEMENT,** juste **de mettre en évidence des associations** entre expositions (par exemple fumer) et événements de santé (par exemple un cancer).

8. **Les études d'observations visent à identifier les facteurs associés à des événements de santé, il s'agit souvent de risques.**

Données longitudinales :

Lorsque qu'il existe Plusieurs mesures à travers le temps PAR sujet

(Les mesures longitudinales chez un même sujet ne peuvent pas être considérées comme réalisation de variables aléatoires indépendantes)

Les études d'observation

1. Etudes de cohorte :

Les sujets sont répartis en groupes en FONCTION de leur EXPOSITION

Ex : *fumeur / Non fumeur*

Comparaison du taux de SURVENUE *(de l'évènement)* **afin mesurer l'Association entre Exposition et Événement**

(D'un point de vue pratique l'étude de cohorte est la démarche d'observation la plus « proche » de l'essai randomisé, la principale différence étant que dans un essai, l'attribution de l'exposition (le traitement) est réalisée par tirage au sort.)

Remarque : Le plus souvent, une étude de cohorte sera prospective, et aura recueilli des données longitudinales.

2. Etudes Cas-témoins : *(= Cas-contrôle)*

Les sujets sont *répartis* en groupes en FONCTION de leur RÉALISATION ou non de l'évènement de santé

Ex : *Malade / Non Malade*

Comparaison du taux d'Exposition afin mesurer l'Association entre Exposition et Événement

(En général, on choisit de un à 4 témoins pour chaque cas et la proportion de malades dans l'étude est complètement déterminée (de 50 % pour 1 témoin pour 1 cas, à 20 % pour 4 témoins par cas), et ne correspond en rien à la proportion de ma- lades dans la population cible.)

Remarque : Le plus souvent une étude cas-témoins sera rétrospective.

3. Etudes Transversales :

Recueillir Simultanément des infos sur l'Exposition et Événement sur un échantillon représentatif de la population cible

Ex : *Etudes de prévalence* *(nb de malades à un instant t)*

Idendifier les facteurs associés aux variations de prévalence.

Ces études transversales sont limitées par l'absence de description temporelle des expositions (et des événements), mais peuvent permettre d'identifier des relations entre événement de santé et exposition lorsque celles-ci sont invariables dans le temps (par exemple, le sexe, le groupe sanguin, ...).

4. Etudes Prospective :

Lorsque l'Exposition est mesurée AVANT la Survenue de l'Evénement étudié

5. Etudes Rétrospective :

Lorsque l'Exposition est mesurée APRÈS la Survenue de l'Evénement étudié

Mesures d'association utilisées en épidémiologie

	M+	M-	RA
E+	**a**	**b**	$RA_{E+} = a/(a+b)$
E-	**c**	**d**	$RA_{E-} = c/(c+d)$
	P(M+)=(a+c)/(n tot)	**P(M-)=1-P(M+)**	

Ra= Risque absolu

RR=Risque Relatif

les formules sont des estimations

$RA_{E+} = P(M+ \mid E+)$

Proportion vraie de malades ***parmi les*** ***Exposés***, *estimé par* ***a/(a+b)***

$RA_{E-} = P(M+ \mid E-)$

Proportion vraie de malades ***parmi les*** ***Non Exposés*** *estimé par* ***c/(c+d)***

$RR = RA_{E+} / RA_{E-}$

Mesure d'association, *défini comme le rapport des risques absolus chez les exposés et non exposés,* ***→ P(M+ | E+) / P(M+ | E-). Estimé par*** ***(a/(a+b)) / (c/(c+d))***

OR : ***Il est estimé par le rapport des produits croisés*** ***(n1n4) / (n2n3)***

Le rapport des cotes est défini comme le rapport de la cote de la maladie chez les exposés P(M+|E+)/P(M-| E+) sur la cote de la maladie chez les non-exposés P(M+ | E-)/P(M- | E-), mais aussi, par application du théorème de Bayes, comme le rapport de la cote des expositions chez les malades P(E+ | M+)/P(E- | M+), par la cote des expositions chez les non malades P(E+ | M-)/P(E- | M-)

Le rapport des cotes est la seule quantité pertinente qui peut être estimée dans une étude cas-témoins !!!

Risque Attribuable et proportion de cas évitables

Le Risque attribuable à un facteur est la proportion des cas que l'on pourrait éviter en supprimant ce facteur, lorsqu'il est CAUSALE !!

Risque Atribuable = La proportion maximale de cas que l'on peut éviter est donc :

$$R_{attribuable} = \%_{évitable} = [\ P(M+) - P(M+ \mid E-)\ ] / P(M+)$$

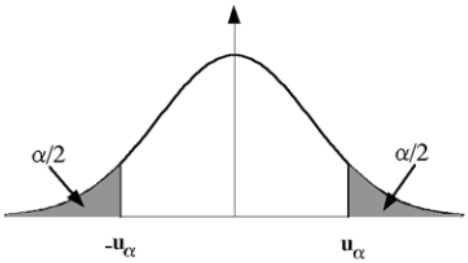
$$= [\ P(E+) * (RR-1)\ ] / [\ P(E+) * (RR-1) - 1\ ]$$

La proportion calculée grâce à cette formule est « maximale » : elle n'est atteinte que si le facteur E a un rôle causal dans le déclenchement de la maladie

Voila, un peeeeetiiit résumé du cours...

Par la suite, on trouve les Tables ainsi que des fiches sur quelques notions non traitées dans le cours vidéo mais pas marquées HP dans le poly, on sait jamais... J'espère que ces fiches vous seront utiles. Elles m'ont pris beaucoup de temps à faire... Bonne chance a vous <3 !

A.1 TABLE DE LA VARIABLE NORMALE REDUITE *Z*



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

La probabilité α s'obtient par addition des nombres inscrits en marge
exemple : pour $u_\alpha = 0{,}994$, la probabilité est $\alpha = 0{,}30 + 0{,}02 = 0{,}32$

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
u_α	3,29053	3,89059	4,41717	4,89164	5,32672	5,73073	6,10941

A.2 TABLE DU TEST DE WILCOXON

Table adaptée de Siegel

n	α		
	0,05	**0,02**	**0,01**
6	2,118		
7	1,961	2,299	
8	2,044	2,324	2,464
9	2,026	2,263	2,381
10	1,947	2,253	2,456
11	2,009	2,276	2,454
12	2,008	2,322	2,479
13	1,964	2,313	2,523
14	1,952	2,329	2,517
15	1,965	2,306	2,533

Indique, pour $n \leq 15$ les valeurs de W_α pour $\alpha = 0,05$, 0,02 et 0,01.

A.3 TABLE DU TEST DE MANN-WHITNEY-WILCOXON

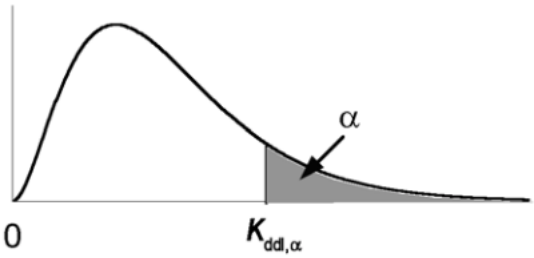
Table adaptée de Siegel

n_B	α	n_A = 3	4	5	6	7	8	9	10
4	**0,05**	2,333	1,905						
	0,01	*2,687*	*2,483*						
5	**0,05**	2,117	2,107	2,110					
	0,01	*2,415*	*2,596*	*2,528*					
6	**0,05**	1,962	2,047	2,118	2,018				
	0,01	*2,479*	*2,473*	*2,483*	*2,498*				
7	**0,05**	2,074	2,003	1,965	2,086	2,057			
	0,01	*2,530*	*2,570*	*2,615*	*2,514*	*2,568*			
8	**0,05**	1,960	1,970	1,991	2,014	2,037	1,953		
	0,01	*2,572*	*2,480*	*2,576*	*2,530*	2,500	*2,584*		
9	**0,05**	2,052	2,099	2,013	1,956	2,022	1,982	2,040	
	0,01	*2,422*	*2,561*	*2,680*	*2,546*	*2,551*	*2,560*	*2,570*	
10	**0,05**	1,961	2,065	2,033	2,017	2,010	2,008	2,009	2,011
	0,01	2,366	2,489	2,523	2,560	2,498	2,541	2,580	2,540

Indique, pour $n_A \leq 10$ et $n_B \leq 10$, $n_A \leq n_B$, les valeurs de M_α, pour α=0,05 et α=0,01.
Exemple : n_A=5, n_B=8 : $M_{0,05}$=1,991

A.4 TABLE DE χ^2

La table donne la probabilité α pour que χ^2 égale ou dépasse une valeur donnée, en fonction du nombre de degrés de liberté (d. d. l.)
Quand le nombre de degrés de liberté est élevé, $\sqrt{2\chi^2}$ est à peu près distribué normalement autour de $\sqrt{2(\text{d.d.l.}) - 1}$ avec une variance égale à 1



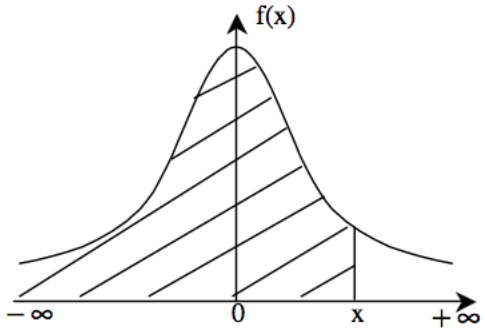
α / ddl	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	0,0158	0,455	1,074	1,642	2,706	3,841	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5,991	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7,815	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	9,488	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11,070	13,388	15,086	20,515
6	2,204	5,348	7,231	8,558	10,645	12,592	15,033	16,812	22,457
7	2,833	6,346	8,383	9,803	12,017	14,067	16,622	18,475	24,322
8	3,490	7,344	9,524	11,030	13,362	15,507	18,168	20,090	26,125
9	4,168	8,343	10,656	12,242	14,684	16,919	19,679	21,666	27,877
10	4,865	9,342	11,781	13,442	15,987	18,307	21,161	23,209	29,588
11	5,578	10,341	12,899	14,631	17,275	19,675	22,618	24,725	31,264
12	6,304	11,340	14,011	15,812	18,549	21,026	24,054	26,217	32,909
13	7,042	12,340	15,119	16,985	19,812	22,362	25,472	27,688	34,528
14	7,790	13,339	16,222	18,151	21,064	23,685	26,873	29,141	36,123
15	8,547	14,339	17,322	19,311	22,307	24,996	28,259	30,578	37,697
16	9,312	15,338	18,418	20,465	23,542	26,296	29,633	32,000	39,252
17	10,085	16,338	19,511	21,615	24,769	27,587	30,995	33,409	40,790
18	10,865	17,338	20,601	22,760	25,989	28,869	32,346	34,805	42,312
19	11,651	18,338	21,689	23,900	27,204	30,144	33,687	36,191	43,820
20	12,443	19,337	22,775	25,038	28,412	31,410	35,020	37,566	45,315
21	13,240	20,337	23,858	26,171	29,615	32,671	36,343	38,932	46,797
22	14,041	21,337	24,939	27,301	30,813	33,924	37,659	40,289	48,268
23	14,848	22,337	26,018	28,429	32,007	35,172	38,968	41,638	49,728
24	15,659	23,337	27,096	29,553	33,196	36,415	40,270	42,980	51,179
25	16,473	24,337	28,172	30,675	34,382	37,652	41,566	44,314	52,620
26	17,292	25,336	29,246	31,795	35,563	38,885	42,856	45,642	54,052
27	18,114	26,336	30,319	32,912	36,741	40,113	44,140	46,963	55,476
28	18,939	27,336	31,391	34,027	37,916	41,337	45,419	48,278	56,893
29	19,768	28,336	32,461	35,139	39,087	42,557	46,693	49,588	58,302
30	20,599	29,336	33,530	36,250	40,256	43,773	47,962	50,892	59,703

Exemple : avec d. d. l. = 3, pour $K_{3;\alpha} = 0{,}584$ la probabilité est $\alpha = 0{,}90$

Table a Utiliser pour le calcul de *P(Z<u2ß)* UNIQUEMENT

Loi Normale centrée réduite

Probabilité de trouver une valeur inférieure à x.



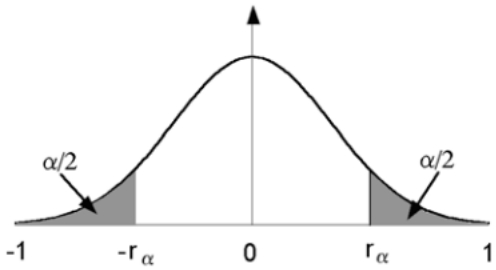
$$F(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}du$$

X	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

A.5 TABLE DU COEFFICIENT DE CORRELATION

(Pas marqué HP mais NON TRAITÉ en cours)

La table indique la probabilité α pour que le coefficient de corrélation égale ou dépasse, en valeur absolue, une valeur donnée r_α, c'est-à-dire la probabilité extérieure à l'intervalle $(-r_\alpha, +r_\alpha)$, en fonction du nombre de degrés de liberté (d. d. l.)

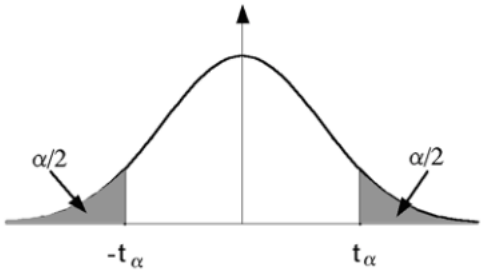


ddl \ α	0,10	0,05	0,02	0,01
1	0,9877	0,9969	0,9995	0,9999
2	0,9000	0,9500	0,9800	0,9900
3	0,8054	0,8783	0,9343	0,9587
4	0,7293	0,8114	0,8822	0,9172
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540

Exemple : avec d. d. l. = 30, pour r_α = 0.3494 la probabilité est α = 0.05

A.6 TABLE DU *t* DE STUDENT

HORS PROGRAMME, mais Utile pour la Corrélation



α / ddl	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	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Exemple : avec d. d. l. = 10, pour $t = 2{,}228$, la probabilité est $\alpha = 0{,}05$

Loi de Student

HORS PROGRAMME, mais ***Utile pour la Corrélation***

On Considère :

- une première V.A : *X*, distribuée selon une **loi normale centrée réduite.**

- une seconde V.A : *Y*, indépendante de *X*, distribuée selon un χ^2 **à *n* degrés de liberté.**

Alors la variable aléatoire $Z = \sqrt{n}\frac{X}{\sqrt{Y}}$ est distribuée selon une loi de Student à n degrés de liberté, notée t(n).

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

La courbe correspondante est symétrique autour de 0, et son allure est proche de celle de la loi normale.

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

Lorsque n croît, en pratique pour $n > 30$, la variance peut être prise égale à 1, et la distribution assimilée à celle d'une loi normale centrée réduite.

Liaison entre deux variables continues :

Notion de corrélation

(Pas marqué HP mais NON TRAITÉ en cours)

Coefficient de Corrélation : *(observé)*

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

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

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

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

Ainsi

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

Retenons, exprimé sur la base des valeurs observées :

$$r = \frac{\frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y}$$

Le numérateur de cette expression est appelé la covariance observée des deux variables X et Y, notée $cov_0(X, Y)$, dont on montre qu'elle s'exprime aussi sous la forme

$$cov_0(X, Y) = \frac{n}{n-1}\left(\frac{1}{n}\sum_i x_i y_i - \bar{x}\bar{y}\right)$$

Propriétés numériques fondamentales de r :

- r a toujours une valeur comprise entre -1 et 1 ;
- r prend la valeur -1 (respectivement 1) si et seulement si il existe des valeurs a et b telles qu'on ait pour tout i $y_i = ax_i + b$ avec a négatif (respectivement $a > 0$).

Remarques :

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

RÉSUMÉ DU CHAPITRE

le coefficient de corrélation vrai est noté ρ

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

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

   Propriétés :

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

2. Disposant d'un échantillon de n couples (x_i, y_i) on définit le coefficient de corrélation observé :

   $$r = \frac{\frac{1}{n-1}\sum_i (x_i - m_x)(y_i - m_y)}{s_X s_Y} = \frac{\frac{n}{n-1}\left(\frac{1}{n}\sum_i x_i y_i - m_x m_y\right)}{s_X s_Y}$$

   Propriété : $r \in [-1 ; 1]$

3. Il existe un test de nullité du coefficient de corrélation « vrai » dont le paramètre est r.
4. Indépendance et corrélation sont des notions différentes ; deux variables dont le coefficient de corrélation « vrai » est nul peuvent être liées.

Autre formulation du test

On peut montrer que $t = r\sqrt{\frac{n-2}{1-r^2}}$ est, sous H_0, distribué selon une loi de Student à n-2 ddl.

(c'est plus simple à utiliser)

Le Test en bref :

Hypothèses

H0 : ρ = 0 [les variables ne sont pas corrélées]

H1 : ρ ≠ 0 [les variables sont corrélées]

Paramètre du Test

Coefficient de corrélation observé : r *(cf au dessus)*
mais utiliser *t*

Méthodo :
- Calculer r avec les données, c un peu relou mais pas le choix,
- Calculer t à partir de r
- Établir l'intervalle de PARI à 95% (en général) à partir d'une loi de STUDENT de n-2 DDL

Conslusion:

Si t n'est PAS dans l' I.P on rejette Ho au risque α

Si t EST dans l'I.P l'est on ne conclut PAS