وزارة التعليـــم العـــالي و البحـــث العلـــمي

جامعـة وهـران للعلـوم والتكنـولـوجيا ۩ محـمد بـوضيـاف ۩

كـلـيـة الـفيزيـــاء

L'essentiel de la Biostatistique pour un Physicien Médical

Polycopié de cours avec exercices corrigés

Auteur : Samia Bahlouli

Selon le programme du **Master 2** option **physique médicale**

Avant propos

Ce polycopié est un petit fascicule destiné aux étudiants inscrits en Master 2 option Physique médicale. Il est rédigé selon le canevas proposé par le ministère de l'enseignement supérieur. Comme son nom l'indique, ce module est le lien entre la statistique et le monde du vivant, que ce soit en biologie ou en médecine. L'essentiel des outils statistiques utiles pour ce type d'étude a été présenté avec quelques applications.

Ce cours peut être aussi utilisé par les étudiants inscrits en PASV ou en Science Médicale.

Pour sa rédaction, les références suivantes ont été utilisées :

- *Exercices corrigés de statistique et probabilités avec rappels de cours, Maurice Lethielleux, DUNOD 2nd édition*
- *Éléments de probabilités et de statistique pour ingénieur, M. Henkouche, (tirage spécial USTO)*
- *Introduction à la biostatistique, Alain-Jacques Valleron, collection Masson*
- *Cours de biostatistique de la PACES - UE4 - de la Faculté de Médecine, Pierre et Marie Curie (Paris VI), 2013-2014.*
- *EI - Exercices de Probabilités Corrigés*, https://docplayer.fr/2601494-Ei-exercices-de-probabilites-corriges.html
- *Exercices corrigés* http://www.iamin.be/umdb/biostats/?q=book/export/html/249

Sommaire

Chapitre I : Généralités sur les probabilités

Chapitre2 : Variables aléatoires

Chapitre I : Généralités sur les probabilités

I/ Pourquoi de la statistique en médecine?

- En 1835 : à l'Académie des sciences de Paris de vifs débats sur l'applicabilité des méthodes numériques à la médecine ;
- en 1837, la querelle s'enflamma de nouveau à l'Académie de médecine de Paris, mais le monde médical ne fut pas convaincu
- C'est ainsi que selon **J Bouillaud** : « La somme de nos certitudes en matière d'étiologie, d'anatomie pathologique, de diagnostic et de thérapeutique est énorme : que dis-je ? La médecine ne serait pas une science, mais une sorte de jeu de hasard, sil elle ne roulait tout entière que sur des probabilités. » **Essai sur la philosophie médicale et sur les généralités de la clinique médicale (1856)**

La statistique constitue, en médecine, l'outil permettant de répondre à de nombreuses questions qui se posent en permanence au médecin :

1. Quelle est la valeur normale d'une grandeur biologique, taille, poids, glycémie ?

2. Quelle est la fiabilité d'un examen complémentaire ?

3. Quel est le risque de complication d'un état pathologique, et quel est le risque d'un traitement ?

4. Le traitement A est-il plus efficace que le traitement B ?

Toutes ces questions, proprement médicales, reflètent une propriété fondamentale des systèmes biologiques qui est leur variabilité.

- **variabilité totale = variabilité biologique + variabilité métrologique**
- **variabilité biologique = variabilité intra-individuelle + variabilité inter-individuelle**
- **variabilité métrologique = variabilité expérimentale + variabilité appareil de mesure**

Pour prendre une décision diagnostique ou thérapeutique le médecin doit avoir des éléments lui permettant de prendre en compte cette variabilité naturelle, pour distinguer ce qui est normal de ce qui est pathologique (décision à propos d'un patient) et pour évaluer la qualité d'un nouvel examen, ou d'une nouvelle thérapeutique (décision thérapeutique).

II/ Termes et concepts importants

- ***population P*** *: un ensemble généralement très grand, voire infini, d'individus ou d'objets* de même nature.
- ***Échantillon:*** une partie de la population (Il est le plus souvent impossible, ou trop coûteux, d'étudier l'ensemble des individus constituant une population).
- Chaque individu, ou unité statistique, appartenant à une population est décrit par un ensemble de caractéristiques appelées *variables ou caractères. Ces variables peuvent être quantitatives (numériques)* ou qualitatives (non numériques)
- **Quantitatives:** pouvant être classées en variables continues (taille, poids) ou discrètes (nombre d'enfants dans une famille)
- **Qualitatives:** pouvant être classées en variables catégorielles (couleurs des yeux) ou ordinales (intensité d'une douleur classée en nulle, faible, moyenne, importante).
- ***Épreuve****:* Une expérience aléatoire dont le résultat n'est pas prévisible*.*
- **Ensemble fondamental:** ensemble des résultats possibles que nous noterons *E dans la suite du cours*
- **Evénement:** Un événement *A est un sous ensemble de E, c'est-à-dire un ensemble de résultats.*
- L'événement {*a}, constitué par un seul point de E, donc par un seul résultat , est appelé* **événement élémentaire.**
- L'ensemble vide Æ ne contient aucun des résultats possibles : il est appelé **événement**

impossible.

- L'ensemble *E contient tous les résultats possibles : c'est l'**événement certain.***
- Si *E est fini, ou infini dénombrable, tout sous-ensemble de E est un événement ; ce n'est* pas vrai si *E est non dénombrable (ceci sort du cadre de ce cours).*
- On note parfois **Ω** l'ensemble de tous les événements

III/ Opérations sur les événements

- Si *A et B* sont deux événements, les opérations de combinaison sont :

1. A ∪B est l'événement qui se produit si *A ou B (ou les deux) est réalisé.* Il est parfois noté ou *A ou B.*

2. A ∩B est l'événement qui se produit si *A et B sont réalisés tous les deux.* Il est parfois noté ou *A et B.*

3. C_A est l'événement qui se produit quand *A n'est pas réalisé. On l'appelle aussi négation de A.* Il est parfois noté «non A », ou Ā.

- **Evénements incompatibles**: c'est quand deux événements *A et B sont tels que* $A \cap B = \varnothing$, *ils ne peuvent être réalisés simultanément.* On dit qu'ils s'excluent mutuellement
- **Système complet d'événements**: On dit que les événements *A1, A2, ..., An forment une famille complète si les Ai constituent* une partition de *E, c'est-à-dire si :*

1. les événements sont deux à deux disjoints : $\forall(i \neq j), Ai \cap Aj = \varnothing$,

2. ils couvrent tout l'espace : $\cup_i Ai = E$

IV/ Règles du calcul des probabilités

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

Pr est dite fonction de probabilité, et Pr(A) est appelée probabilité de l'événement A, si les

conditions ou règles suivantes sont satisfaites :

1. $Pr(A) \geq 0$ pour tout événement A : une probabilité est positive ou nulle

2. $Pr(E) = 1$: la probabilité de l'événement certain est 1

3. $A \cap B = \varnothing \Rightarrow Pr(A \cup B) = Pr(A) + Pr(B)$: permet le calcul de la probabilité de la réunion de deux événements **disjoints**

4. Soit un ensemble dénombrable (fini ou non) d'événements Ai deux à deux disjoints $Ai \cap Aj = \varnothing$, *alors* $Pr(A1 \cup A2 \cup ....) = Pr(A1) + Pr(A2) + ...$

V/ Ensembles probabilisés

- **Ensemble probabilisé fini:** Soit $E = \{a1, a2, ..., an\}$ *un ensemble fondamental fini. On probabilise cet ensemble en attribuant* à chaque point *ai un nombre pi, probabilité de l'événement élémentaire {ai}, tel que :*
  - $p_i \geq 0$
  - $P_1 + P_2 + ... + P_n = 1$
  - La probabilité d'un événement quelconque *A est la somme des probabilités des* a_i *qu'il contient* $Pr(A) = \sum_{ai \in A} p_i$

- **Ensemble fini équiprobable:** C'est un ensemble fini probabilisé tel que tous les événements élémentaires ont la même probabilité.
- On dit aussi qu'il s'agit d'un espace probabilisé uniforme.

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

avec $p_1 = p_2 = ... = p_n = 1/n$

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

VI/ Probabilité Conditionnelle ; Indépendance et Théorème de Bayes

Soient A et B deux événements quelconques d'un ensemble fondamental E muni d'une loi de

probabilité *Pr*. On s'intéresse à ce que devient la probabilité de *A* lorsqu'on apprend que *B* est déjà réalisé, c'est-à-dire lorsqu'on restreint l'ensemble des résultats possibles *E* à *B*.

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

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

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

On en tire immédiatement

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

VI-1 . Indépendance entre événements

On dit que deux événements *A* et *B* sont indépendants si la probabilité pour que *A* soit réalisé n'est pas modifiée par le fait que *B* se soit produit. On traduit cela par

Pr(*A* / *B*) = *Pr*(*A*).

D'après la définition d'une probabilité conditionnelle, on tire la définition, A et B sont indépendants si et seulement si

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

i- Indépendance, inclusion et exclusion de deux événements

Considérons deux événements *A* et *B*.

- Si A ⊂ B (*A* est inclus dans *B*) : si *A* est réalisé, alors *B* aussi

Alors $\Pr(A \cap B) = \Pr(A)$ d'où

$$\Pr(B/A) = \frac{\Pr(A \cap B)}{\Pr(A)} = 1 \; et \; \Pr(A/B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{\Pr(A)}{\Pr(B)}$$

A et *B* ne sont **pas indépendants**

- Si A ∩ B =∅ (*A* et *B* sont exclusifs) : si *A* est réalisé, *B* ne peut pas l'être

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

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

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

- Si A et B **sont indépendants** et $\boldsymbol{A \cap B \neq \emptyset} \Rightarrow Pr(A \cup B) = Pr(A) + Pr(B) - \Pr(A \cap B)$

VI-2. Théorème de Bayes

Considérons, pour illustrer ce théorème, le problème du diagnostic d'une douleur aiguë de l'abdomen. Il s'agit d'un patient arrivant aux urgences pour un « mal au ventre ».

Si l'on ne sait rien d'autre sur le patient (on n'a pas fait d'examen clinique ou complémentaire), on ne connaît que les probabilités d'avoir tel ou tel diagnostic si on observe une douleur.

Soient $D1$, $D2$ et $D3$ les 3 diagnostics principaux (il y en a en fait au moins une douzaine) et exclusifs ; par exemple $D1$ = appendicite, $D2$ = perforation d'ulcère, $D3$ = autres diagnostics.

Soit un signe $s1$ pour lequel on connaît $Pr(s1/D1)$, $Pr(s1/D2)$, et $Pr(s1/D3)$.

Par exemple, $s1$ serait « présence d'une fièvre 38,5°C » ; $Pr(s1/D1) = 0{,}90$; $Pr(s1/D2) = 0{,}30$; et $Pr(s1/D3) = 0{,}10$.

Ces probabilités peuvent être estimées sur une population de patients en dénombrant le nombre de sujets ayant le diagnostic $D1$ et présentant le signe $s1$. De même, on peut

connaître $Pr(D1)$, $Pr(D2)$ et $Pr(D3)$.

Le problème diagnostique se pose comme celui de choisir par exemple le diagnostic le plus probable connaissant le signe $s1$; pour ce faire, on calcule $Pr(D1/s1)$, $Pr(D2/s1)$, $Pr(D3/s1)$ et on retient le diagnostic qui a la plus grande probabilité : c'est l'application de l'approche bayesienne au problème de l'aide au diagnostic

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

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

Exercices d'applications

EX1 : Soient A, B et C des événements définis sur le même espace probabilisé.

1- Exprimer en utilisant les symboles des opérations sur les événements ci-dessous:

a) B seul se réalise
b) A et B se réalisent mais pas C
c) A, B, C se réalisent simultanément puis aucun des trois ne se réalise simultanément
d) Au moins un des événements se réalise
e) Un seul événement se réalise
f) Au moins deux événements se réalisent
g) Deux événements au plus se réalisent
h) Deux événements et deux seulement se réalisent

EX2 : simplifier les expressions:

$(A \cup B) \cap (A \cup \overline{B})$

$(A \cup B) \cap (A \cup \overline{B}) \cap (\overline{A} \cup B)$

$[(\overline{A \cap B}) \cap (\overline{A \cap C})] \cup A$

EX3: Les automobilistes se répartissent en trois catégories:

- Ceux qui n'ont pas un téléphone portable, événement X
- Ceux qui ont un téléphone portable et ne téléphonent pas en conduisant, événement Y
- Ceux qui ont un téléphone portable et téléphonent en conduisant, événement Z

Soit A l'événement un automobiliste a un accident et l'on donne les probabilités suivantes:

$P(X) = 0.2 \quad P(Y) = 0.5 \quad P(Z) = 0.3$

Px(A) = 0.01 Py(A) = 0.02 Pz(A) = 0.04

Décomposé A en un système complet d'événement à partir de X, Y, Z et en déduire P(A)

EX4 : Dans une usine de fabrication de prothèses, on dispose de deux machines de control M1 et M2 pour éliminer les prothèses non conformes. On procède d'un premier tri par M1, les pièces refusées par M1 sont éliminées, celles non refusées par M1 sont triées par M2 qui les accepte ou les refuse. Une erreur de tri par une machine consiste donc à accepter une pièce non conforme ou à rejeter une pièce conforme. On suppose que :

- La probabilité pour M1 d'accepter une prothèse non conforme est p.
- La probabilité pour M1 de refuser une prothèse conforme est p.
- La probabilité pour M2 d'accepter une prothèse non conforme est q.
- La probabilité pour M2 de refuser une prothèse conforme est q.
- Les erreurs de tri par les deux machines sont indépendantes.

1. Déterminer la probabilité d'éliminer une prothèse conforme
2. Déterminer la probabilité d'accepter une prothèse non conforme
3. Si p>q faut-il permuter les machines ?

EX5 : Le test de dépistage d'un certain virus n'est pas infaillible :

– 1 fois sur 100, il est positif, alors que l'individu n'est pas contaminé ;

– 2 fois sur 100, il est négatif, alors que l'individu est contaminé.

Il est donc important de répondre aux questions suivantes :

1. Étant donné que son test est positif, quelle est la probabilité qu'un individu ne soit pas porteur du virus ?

2. Étant donné que son test est négatif, quelle est la probabilité qu'un individu soit porteur du virus ?

Sachant que dans la population totale, la fraction de porteurs est approximativement de 1/1000

Corrigés des exercices

EX1 : On utilise les opérations sur les évènements

a) $B \cap \bar{A} \cap \bar{C}$
b) $A \cap B \cap \bar{C}$
c) $A \cap B \cap C$ puis $\bar{A} \cap \bar{B} \cap \bar{C}$
d) $A \cup B \cup C$
e) $(A \cap \bar{B} \cap \bar{C}) \cup (\bar{A} \cap B \cap \bar{C}) \cup (\bar{A} \cap \bar{B} \cap C)$
f) $(A \cap B) \cup (A \cap C) \cup (B \cap C)$
g) $\overline{A \cap B \cap C} = \bar{A} \cup \bar{B} \cup \bar{C}$
h) $(A \cap B \cap \bar{C}) \cup (\bar{A} \cap B \cap C) \cup (A \cap \bar{B} \cap C)$

EX2 :

- $(A \cup B) \cap (A \cup \bar{B}) = (A \cup B) \cap (A \cup \bar{B}) = A \cup \emptyset = \mathbf{A}$
- $(A \cup B) \cap (A \cup \bar{B}) \cap (\bar{A} \cup B) = A \cap (\bar{A} \cup B)$

$$= (A \cap \overline{A}) \cup (A \cap B) = \emptyset \cup (A \cap B) = \mathbf{A} \cap \mathbf{B}$$

- $[(\overline{A \cap B}) \cap (\overline{A \cap C})] \cup A = [(\bar{A} \cup \bar{B}) \cap (\bar{A} \cup \bar{C})] \cup A$ (loi de Morgan)

$$= [\bar{A} \cup (\bar{B} \cup \bar{C})] \cup A \text{ (distributivité des lois)}$$

$$= (\bar{A} \cup A) \cup (\bar{B} \cup \bar{C}) = \Omega \cup (\bar{B} \cup \bar{C}) = \mathbf{\Omega}$$

EX3 :

Des évènements A_1, A_2, …., A_n forment un système complet d'évènement si un évènement et un seul du système se produit à la fois, ils sont donc disjoints deux à deux. Ils constituent ce qu'on appelle on mathématique une partition.

Les évènements X, Y, et Z vérifient bien cette condition et on a :

$X \cap Y = X \cap Z = Y \cap Z = \emptyset$ et $A = (A \cap X) \cup (A \cap Y) \cup (A \cap Z)$

En appliquant l'axiome des probabilités totales aux évènements :

$$P(A) = P(A \cap X) + P(A \cap Y) + P(A \cap Z)$$

$$= P(X) \times P_X(A) + P(Y) \times P_Y(A) + P(Z) \times P_Z(A)$$

$$= 0{,}2 \times 0{,}01 + 0{,}5 \times 0{,}02 + 0{,}3 \times 0{,}04 = \mathbf{0,024}$$

EX4 :

On note PC : Prothèse conforme, et PNC : Prothèse non conforme

EC : Eliminer PC , APNC : Accepter une PNC

1 . Pour éliminer une prothèse conforme, les machines procèdent comme suit :

EC = (M1 élimine PC) ou (M1 accepte PC) et (M2 élimine PC)

Ce qui s'écrit avec les opérations sur les événements comme suit :

P(EC) = P(M1 élimine PC) $\cup$ P(M1 accepte PC) $\cap$ P(M2 élimine PC)

= P(M1 élimine PC) + P(M1 accepte PC) $\times$ P(M2 élimine PC)

P(EC)= p + (1-p)q = p+ q – pq

2. Pour accepter une prothèse non conforme, les machines procèdent comme suit :

APNC = (M1 accepte PNC) et (M2 accepte PNC)

Ce qui s'écrit avec les opérations sur les événements comme suit :

P(APNC) = P(M1 accepte PNC) $\cap$ P(M2 accepte PNC)

= P(M1 accepte PNC) × P(M2 accepte PNC)

P(APNC) = pq

3. Si p > q, ce qui signifie que M1 a plus de probabilité de commettre des erreurs que M2 et si on permute entre les machines cela revient a permuter p et q dans les formules, en examinant les relations, cela n'influe pas sur le résultat final.

EX5 :

On introduit les évènements suivants :

T^+ : test positif ; T^- : test négatif ; C : individus contaminé ; NC : individus non contaminé

On a donc les informations suivantes :

$P(T^+/NC) = 1/100$; $P(T^-/C) = 2/100$; $P(C) = 1/1000$: $P(NC) = 1 - P(C)$

Et on veut calculer $P(NC/T^+)$

En appliquant la formule de Bayes :

$$P(NC/T^+) = \frac{P(T^+/NC)\,P(NC)}{P(T^+/NC)P(NC) + P(T^+/C)P(C)}$$

sachant que : $P(T^+/C) = 1 - P(T^-/C) = 98/100$

En remplaçant les valeurs on trouve :

$P(NC/T^+) = 0{,}91$,

Même si son test est positif, un individu a plus de 90% de chance de ne pas être porteur du virus

2. Un calcul similaire donne le résultat suivant :

$P(C/T^-) = 0,00002$

La probabilité de déclarer non porteur un individu contaminé est de l'ordre de 2/10000, c'est là où réside l'importance de ce test.

Il faut sans doute mentionner que ces calculs ne s'appliquent pas à un individu appartenant à une population à risque, la probabilité d'être porteur devient proche de 1, cela change complètement les conclusions : dans ce cas la probabilité d'être non porteur alors que le test est positif est très petite, tandis la probabilité d'être porteur alors que le test est négatif est très importante.

Chapitre 2 : Variables aléatoires

I/ Définition d'une variable aléatoire

Considérons un ensemble fondamental E correspondant à une certaine expérience. Les éléments de E, résultats possibles de l'expérience, ne sont généralement pas des nombres. Il est cependant utile de faire correspondre un nombre à chaque élément de E, en vue de faire ensuite des calculs.

Une variable aléatoire X, sur un ensemble fondamental E, est une application de E dans R : à tout résultat possible de l'expérience (à tout élément de E), la variable aléatoire X fait correspondre un nombre.

Lorsque E est fini ou infini dénombrable, toute application de E dans R est une variable aléatoire.

Lorsque E est non dénombrable, il existe certaines applications de E dans R qui ne sont pas des variables aléatoires.

II/ Variables aléatoires finies

Soit X une variable aléatoire sur un ensemble fondamental E à valeurs finies :

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

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

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

$$\forall_i , p_i \geq 0 \; et \; n \sum_{i=1}^{n} p_i = 1$$

III/ Espérance mathématique d'une variable finie

L'espérance mathématique traduit la tendance centrale de la variable aléatoire. Il s'agit d'une moyenne où chacune des valeurs *xi* intervient d'autant plus que sa probabilité est importante, c'est-à-dire d'un barycentre ou d'un centre de gravité. On définit alors la **moyenne théorique**, ou **espérance mathématique** d'une variable *X* par :

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

Théorèmes

1. Soit *X* une variable aléatoire et *k* une constante réelle. On a :

- $E(kX) = kE(X)$
- $E(X + k) = E(X) + k$

2. Soient *X* et *Y* deux variables aléatoires définies sur le même espace fondamental *E*. On a :

- $E(X + Y) = E(X) + E(Y)$

On en déduit que pour *n* variables aléatoires X_i, définies sur le même espace fondamental :

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i)$$

IV/ Variance et écart-type d'une variable finie

Après avoir traduit la tendance centrale par l'espérance, il est intéressant de traduire la dispersion autour de l'espérance par une valeur (la variance ou l'écart-type).

La variance (vraie ou théorique) de *X*, notée *var*(*X*) ou σ^2_X, est définie par :

$$\boldsymbol{\sigma_X^2 = var(X) = E((X - \mu_X)^2)} \quad o\grave{u}\ \mu_X = E(X)$$

L'écart-type de *X*, noté σ (X) ou σ_X, est défini par

$$\boldsymbol{\sigma(X) = \sigma_X = \sqrt{var(X)}}$$

On définit la variable centrée réduite par

$$\boldsymbol{Y = \frac{X - \mu}{\sigma}}$$

Si *a* est une constante, on a : $\boldsymbol{var(X + a) = var(X)}$ **et** $\boldsymbol{var(aX) = a^2 var(X)}$

V/ Loi de probabilité produit

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

$X(E) = \{x1, x2, ..., xn\}$

$Y(E) = \{y1, y2, ..., ym\}$.

Considérons l'ensemble produit

$X(E)$ x$Y(E) = \{(x1, y1), (x1, y2), ..., (xn, ym)\}$ (ensemble des couples (xi, yj) pour $i = 1, ..., n$ et $j = 1, ..., m$)

Cet ensemble produit peut être transformé en ensemble probabilisé si on définit la probabilité du couple ordonné (*xi*, *yj*) par ***Pr([X=xi] ∩[Y=yi])*** que l'on note ***pxi,yj***. Cette loi de probabilité de *X*, *Y* est appelée **distribution jointe de *X* et *Y***

X / Y	x_1	x_2	x_3	....	x_n	$\sum_{i=1,n} x_i$
y_1	$P_{x1,y1}$	$P_{x2,y1}$				P_{y1}
y_1	$P_{x1,y2}$					P_{y2}
…						
y_1	$P_{x1,yn}$					
$\sum_{i=1,n} y_i$	P_{x1}	P_{x2}				1

Les probabilités $P_{xi} = \sum_{i=1}^{m} P_{xi,yi}$ et $P_{yj} = \sum_{i=1}^{n} P_{xi,yi}$ sont souvent appelées **lois de probabilité marginales de *X* et de *Y***. Il s'agit simplement de leurs distributions.

Soient μ*X* et μ*Y* les espérances de *X* et de *Y*, σ*X* et σ*Y* leurs écart-types. On montre facilement que

$$\boldsymbol{var(X+Y) = \sigma^2_X + \sigma^2_Y + 2cov(X,Y)},$$

où *cov*(*X*, *Y*) représente la **covariance de *X* et *Y*** et est définie par :

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

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

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

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

VI/ Variables aléatoires indépendantes

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

$$P_r([X=x_i]\cap[Y=y_j]) = P_r(X=x_i)\,.\,P_r(Y=y_j)$$

pour tous les couples (i, j).

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

- $E(XY) = E(X)E(Y)$
- $var(X + Y) = var(X) + var(Y)$
- $cov(X, Y) = 0$ et $\rho(X, Y) = 0$

VII/ Fonction de répartition

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

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

Si X est une variable aléatoire discrète on a

$$F(x) = \sum_{x_i \leq x} \Pr(X = x_i) = \sum_{x_i \leq x} p_i$$

Dans tous les cas, $F(x)$ est une fonction monotone croissante, c'est-à-dire F(a)>F(b) si a>b

$$\lim_{x \to -\infty} F(x) = 0 \quad et \quad \lim_{x \to \infty} F(x) = 1$$

VIII/ Variables aléatoires continues

On définit la loi de probabilité de X, ou distribution de X, à l'aide d'une fonction $f(x)$, appelée **densité de probabilité** de X, telle que

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

on admettra les définitions suivantes pour une variable aléatoire *X*, continue, de distribution *f*(*x*)

- $f(x) \geq 0$ analogue à $p_i \geq 0$
- $\int f(x)dx = 1$ analogue à $\sum_i p_i = 1$
- $\mu_X = E(X) = \int xf(x)dx$ analogue à $\sum_i x_i p_i$
- $\sigma_X^2 = var(X) = \int (X - \mu_X)^2 f(x)dx$ analogue à $\sum (x_i - \mu_X)^2 p_i$
- $\sigma(X) = \sigma_X = \sqrt{var(X)}$
- $F(x) = Pr(X \leq x) = \int_{-\tau}^{\tau} f(\tau)d\tau$ analogue à $\sum_{x_i \leq x} p_i$
- $Pr(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$

IX/ Lois de distributions

A/ Lois discrètes :

Les lois décrites ici ne concernent que des variables dont les valeurs sont des nombres entiers.

A.1 - Loi de Bernoulli

On considère une expérience n'ayant que deux résultats possibles, par exemple succès et échec (ou présence et absence d'une certaine caractéristique). On introduit la variable aléatoire *X* qui associe la valeur 0 à l'échec (ou à l'absence de la caractéristique) et la valeur 1 au succès (ou à la présence de la caractéristique). Cette variable aléatoire est appelée variable de Bernoulli.

- **Distribution de *X***

Appelons P la probabilité de l'élément succés :

Pr(succès) = Pr (X = 1) = P d'où Pr(échec) = Pr (X = 0) = 1 - P

- **Espérance de X**

$$\mu_X = E(X) = \sum x_i \Pr\ (X = x_i = 1 \times \Pr(X = 1) + 0 \times \Pr(X = 0) = \boldsymbol{P}$$

- **Variance de X**

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

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

$$\Rightarrow \sigma_X^2 = P - P^2$$

$$\boldsymbol{\Rightarrow \sigma_X^2 = P(1 - P)}$$

A.2- Loi binomiale

Soient les épreuves répétées et indépendantes d'une même expérience de Bernoulli.

Chaque expérience n'a que deux résultats possibles : **succès ou échec**, la probabilité d'avoir k succès lors de n épreuves répétées est

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

On dit que X est distribuée selon une loi binomiale **B(n, p)**

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

A.3- Loi de Poisson

La loi de Poisson (due à Siméon Denis Poisson en 1837) 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**. Elle peut s'appliquer au nombre d'accidents, à l'apparition d'anomalies diverses, à la gestion des files d'attentes, au nombre de colonies bactériennes dans une boîte de Pétri, etc

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 :

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

Cette loi dépend d'un paramètre λ, nombre réel strictement positif.

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

Remarques

- Si deux variables aléatoires indépendantes X_1 et X_2 sont distribuées selon des lois de Poisson de paramètres λ_1 et λ_2, alors la variable X_1+X_2 est distribuée selon une loi de Poisson de paramètre $\lambda_1+\lambda_2$.
- Si on connaît la probabilité de n'observer aucun événement $Pr(X=0) = p$:
- $\lambda = -\ln p$
- $\Pr(X = 1) = e^{-\lambda} \frac{\lambda^1}{1!} = p\lambda$

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

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

…….

$$\Rightarrow \mathbf{Pr}(\boldsymbol{X = k}) = \boldsymbol{e^{-\lambda}\,\frac{\lambda^k}{k!}} = \mathbf{Pr}(\boldsymbol{X = k-1})\,\boldsymbol{\frac{\lambda}{k}}$$

i- Lien avec la loi binomiale

Si une variable aléatoire *X* est distribuée selon une loi binomiale B(*n*, p), on montre que si p 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 $\boldsymbol{\lambda = np}$.

B/ Lois continues

B.1 Loi normale

La distribution normale, ou de Laplace-Gauss, appelée aussi gaussienne, est une distribution continue qui dépend de deux paramètres μ et σ. On la note **N(μ, σ²)**. Le paramètre μ peut être quelconque mais σ est positif. Cette distribution est définie par :

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

Loi normale N(μ, σ²)	
Espérance	μ
Variance	σ^2
Ecart-type	σ

a) Approximation de la distribution binomiale par la loi normale

Lorsque *n* est grand, et que p et 1-p ne sont pas trop proches de 0, alors on constate que la distribution binomiale tend vers la distribution normale de moyenne ***n*p** et de variance ***n*p(1-p)**

b) Approximation de la loi de Poisson par la loi normale

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

c) La distribution normale centrée réduite

On dit que la distribution est centrée si son espérance μ est nulle ; elle est dite réduite si sa

variance σ^2(et son écart-type σ) est égale à 1. La distribution normale centrée réduite N(0, 1) est donc définie par la formule

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

B.2- Loi du $\chi 2$

C'est une loi dérivée de la loi normale, très importante pour ses applications en statistiques.

Soient *X*1, ..., *Xn* des variables aléatoires indépendantes, chacune étant distribuée selon une loi normale centrée réduite :

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

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

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

B.3 Loi de Student

Il s'agit encore d'une loi dérivée de la loi normale, très utilisée dans les tests statistiques, on considère une première variable aléatoire *X*, distribuée selon une loi normale centrée réduite, puis une seconde variable *Y*, indépendante de *X*, distribuée selon un $\chi 2$ à *n* degrés de liberté.

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

loi de Student	
Espérance	0
Variance	$\dfrac{n}{n-2}$
Ecart-type	$\sqrt{\dfrac{n}{n-2}}$

B.4 Loi exponentielle

Cette loi décrit par exemple le processus de mortalité dans le cas où le « risque instantané » de décès est constant. La loi correspondante est :

$f(x) = \lambda e^{-\lambda x}$ avec $\lambda > 0$ et $x \geq 0$

où *x* est la durée de vie.

loi exponentielle	
Espérance	$1/\lambda$
Variance	$1/\lambda^2$
Ecart-type	$1/\lambda$

Exercices d'applications

EX1 :

On étudie la suite de naissance dans une famille

Fi est l'événement "le ième enfant est une fille"

Gj est l'événement "le jème enfant est un garçon"

On suppose que ces événements sont équiprobables est indépendants entre eux.

1/ Soit X la variable aléatoire égale au rang de la première fille (nombre de naissance nécessaire pour avoir la première fois une fille). Déterminer la loi de probabilité de X et en déduire son espérance mathématique.

2/ Retrouver la loi de probabilité de la variable Y égale au nombre de filles dans une famille de 6 enfants, puis donner son espérance mathématique et son écart type.

EX2 :

Les vendredi, et entre 8h et 10 h du matin, le nombre de malades X reçus par le service de diabétologie suit une loi de poisson de paramètre $\lambda = 3$.

1/ Déterminer la probabilité que le service reçoit 0 malade, 1 malade, 2 malades entre 8h et 10 h.

2/ Déterminer la probabilité que le service reçoit plus que 2 malades.

EX3 :

1/ Une variable aléatoire T suit une loi normale centrée réduite, déterminer les probabilités suivantes:
$P(T \leq 1.42)$; $P(T \leq -1.42)$; $P(-1 \leq T \leq 1)$; $P(-1{,}96 \leq T \leq 1{,}96)$

2/ Déterminer t sachant que $P(T < t) = 0.0718$

3/ Une variable aléatoire X suit une loi normale de moyenne 178 et de variance 25, déterminer les probabilités suivantes: $P(X \leq 173)$; $P(X > 188)$; $P(173 < X \leq 188)$

Corrigés des exercices

EX1 :

1/ La variable aléatoire X égale au rang de la première fille peut prendre les valeurs 1, 2, …,k, ..,n , en fonction des évenements, on écrit

$(X=k) = G_1 \cap G_2 \cap \ldots G_{k-1} \cap F_k$

Les événements sont indépendants $\Rightarrow P(X=k) = P(G_1)\, P(G_2) \ldots P(G_{k-1})\, P(F_k) = (0,5)^k$

X suit donc une loi géométrique de paramètre p =0,5 et son épérance E(X) = 1/p = 2, on obtient une fille pour la première fois après 2 naissances en moyenne

2/ à la naissance « i » on associe une variable aléatoire Xi de bernoulli de paramètre ½

$$\begin{cases} X_i = 1 \; Si \; le \; i^{ème} enfant \; est \; une \; fille \; avec \; la \; probabilité \; p = 1/2 \\ X_i = 0 \; Si \; le \; i^{ème} enfant \; est \; un \; garçon \; avec \; la \; probabilité \; q = 1 - p = 1/2 \end{cases}$$

$Y = \sum X_i = X_1 + X_2 + \ldots + X_6$ est une somme de 6 variables de Bernoulli indépendantes de paramètre p = ½

Donc Y suit une loi binomiale de paramètre p = ½ et n = 6

$$P(Y = k) = \binom{6}{k} \times (\frac{1}{2})^k \times (\frac{1}{2})^{6-k} = \frac{1}{64} \binom{6}{k}$$

Y=y	0	1	2	3	4	5	6	Total
P(Y=y)	1/64	6/64	15/64	20/64	15/64	6/64	1/64	1

$E(Y) = np = 3$

$\sigma = \sqrt{nP(1-P)} = 1{,}5$

EX2 :

X suit une loi de poisson de paramètre 3, sa loi de probabilité pour k entier s'écrit :

$$P(X = k) = p_k = e^{-3}\frac{3^k}{k!}$$

$p_0 = e^{-3} = 0{,}0498$ $\quad p_1 = 3e^{-3} = 0{,}1494$ $\quad p_2 = 4{,}5\, e^{-3} = 0{,}224$

2/ On sait que pour une valeur quelconque t, la probabilité totale de toutes les variables X est égale à 1:

$$P(X > t) + P(X \leq t) = 1$$

Donc : $P(X > 2) = 1 - P(X \leq 2) = 1 - p_0 - p_1 - p_2$

$$P(X > 2) = 1 - 8{,}5\, e^{-3} = 0{,}5768$$

La probabilité que le service reçoit plus que 2 malades est de 0,5768

EX3

On utilise la table de la loi normale centrée réduite donnée en annexe : P(T <t) pour t > 0

On trouve $P(T \leq 1{,}42) = 0{,}9222$

Lorsque t <0, on utilise la symétrie car la densité de T est une fonction pair :

$$P(T \geq -1{,}42) = P(T \leq 1{,}42) = 0{,}9222$$

Donc $P(T \leq -1{,}42) = 1 - P(T \geq -1{,}42) = 1 - 0{,}9222$

$$P(T \leq -1{,}42) = 0{,}0778$$

$$P(-1 \leq T \leq 1) = P(T \leq 1) - P(T \leq -1) = P(T \leq 1) - P(T \geq 1)$$

$$= P(T \leq 1) - \big(1 - P(T \leq 1)\big)$$

$P(-1 \leq T \leq 1) = 2P(T \leq 1) - 1$

$P(-1 \leq T \leq 1) = 0{,}6826$

De la même façon on trouve

$P(-1{,}96 \leq T \leq 1{,}96) = 0{,}95$

2/ sur la table, on a : $P(T \geq 0) = P(T \leq 0) = 0{,}5$

$Si\ P(T \leq t) > 0{,}5\ \ alors\ t > 0$

$Si\ P(T \leq t) < 0{,}5\ \ alors\ t < 0$

$P(T < t) = 0{,}0718\ donc\ t < 0$ et par symétrie $P(T > -t) = 0{,}0718$

$P(T < -t) = 1 - P(T > -t) = 1 - 0{,}0718 = 0{,}9282$

La valeur sur la table qui correspond à la probabilité de 0,9282 est environ 1,46

Donc –t = 1,46 $\Rightarrow$ t = -1,46

3/ On passe par la variable centrée réduite T

$X \rightarrow N(m,\sigma) \Rightarrow T = \frac{X-m}{\sigma} \rightarrow N(0,1)$

Pour cet exerciuce, m $=$ 178 et $\sigma = \sqrt{25} = 5$

Ainsi $P(X \leq 173) = P(\frac{X-178}{5} \leq \frac{173-178}{5})$

$\Rightarrow P(X \leq 173) = P(T \leq -1) = P(T \geq 1) = 1 - P(T \leq 1)$

$P(X \leq 173) = 0{,}159$

$P(X > 188) = P\left(\frac{X-178}{5} > \frac{188-178}{5}\right) = P(T > 2)$

On trouve $P(X > 188) = 0{,}0228$

$$P(173 < X \leq 188) = P(\frac{173 - 178}{5} < \frac{X - 178}{5} \leq \frac{188 - 178}{5})$$

$P(173 < X \leq 188) = P(-1 < T \leq 2)$

On suit la même procédure que la question 1 on trouve

$P(173 < X \leq 188) = 0{,}8185$

Chapitre 3: Méthodologie des études épidémiologiques

Une étude épidémiologique peut être réalisée afin de mettre en évidence un lien entre une maladie et un facteur de risque supposé. Les résultats d'une étude épidémiologique peuvent être représentés sous la forme d'une table de contingence. Il existe plusieurs types d'études épidémiologiques, les plus utilisées sont les études de cohorte et les enquêtes cas-témoins.

I/ Etudes de cohorte

Une cohorte était le dixième d'une légion romaine. C'est plus généralement un ensemble de sujets. Dans une étude dite de cohorte les sujets sont répartis en groupes en fonction de leur exposition (par exemple, fumeur/non fumeur) et l'événement n'est pas survenu au moment où cette répartition est faite. En d'autre termes, Deux groupes de personnes sont constitués: le premier est composé de personnes exposées à un facteur de risque déterminé tandis que le second comporte des personnes non-exposées à ce facteur de risque. Ces personnes sont suivies durant un certain temps afin de constater si elles développent ou non la maladie étudiée.

II/ Etudes cas-témoins

Dans une étude cas-témoins (ou cas-contrôle), les groupes de sujets sont constitués en fonction de leur réalisation ou non de l'événement de santé : les cas sont par exemple les malades atteints d'un cancer et les témoins, des sujets non atteints de ce cancer. On compare les niveaux d'exposition dans ces deux groupes pour étudier l'association entre exposition et événement de santé. En d'autre termes, Deux groupes de personnes sont constitués: un groupe composé de personnes atteintes de la maladie et un groupe composé de personnes non-atteintes de la maladie (témoins). Le passé de chaque personne est analysé afin de déterminer si elle a été exposée au facteur de risque étudié.

III/ Mesures d'association utilisées en épidémiologie

On traite le cas le plus simple où une exposition est répartie en deux niveaux (oui/non, présent/absent, exposé/ non exposé), et on notera E+ l'exposition, E- l'absence d'exposition au facteur étudié. L'événement d'intérêt est également catégorisé en deux niveaux, M+ pour malade, M- pour non malade. On notera que dans le cas d'un essai thérapeutique E+ est le traitement à l'étude, et M- peut être défini comme le succès thérapeutique. A partir de cette catégorisation, il est possible de dresser la table de contingence suivante :

	M+	**M-**
E+	n1	n2
E-	n3	n4

On définit

- **le risque absolu** chez les exposés, comme la proportion vraie de malades parmi les exposés P(M+ | E+), estimé par n1/(n1+n2), noté R_1

— le risque absolu chez les non exposés, comme la proportion de malades chez les non exposés, P(M+ | E-), estimé par n3/(n3+n4), noté R_0

- **le risque relatif (RR)** est une mesure d'association, défini comme le rapport des risques absolus chez les exposés et non exposés, P(M+ | E+) / P(M+ | E-). Ce risque est estimé par n1/(n1+n2) / n3/(n3+n4)
- **le rapport des cotes (OR)**(odds-ratio en anglais) est une autre mesure d'association très utilisée en biomédecine. Il 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-). Il est estimé par le rapport des produits croisés (n1n4) / (n2n3).

Le rapport des cotes est la seule quantité pertinente qui peut être estimée dans une étude cas-témoins puisque le nombre total de sujets non malades est déterminé par le nombre de témoins choisi par cas. Si la maladie est rare dans la population cible, aussi bien chez les exposés que chez les non exposés, P(M+) est proche de 0 et donc P(M-) voisin de 1, et P(M+ | E+)/P(M- | E+) est voisin de P(M+ | E+) ; P(M+ | E-)/P(M- | E-) proche de P(M+ | E-) et donc le rapport des cotes défini ci-dessus est proche du risque relatif.

Un risque relatif ou un rapport de cotes supérieur à 1 signifie que l'exposition est un facteur de risque de l'événement étudié.

Un risque relatif ou un rapport de cotes inférieur à 1 signifie que l'exposition est un facteur protecteur de l'événement. Un risque relatif de 50 (par exemple) pour l'exposition « fumeur » et l'événement « cancer du poumon » s'interprète littéralement comme « il y a 50 fois plus de cancer du poumon chez les fumeurs que chez les non fumeurs ».

IV/ Un test de χ^2

Un test de χ^2 est réalisé afin de vérifier si le risque relatif est significatif, autrement dit, si la probabilité d'être malade pour une personne exposée est significativement plus grande de celle d'être malade pour une personne non-exposée. Il s'agit d'un **test d'indépendance** comportant deux états pour chaque critère.

Hypothèse initiale (H0): RR=1. Cela signifie que la probabilité d'être malade pour une personne exposée n'est pas plus grande que la probabilité d'être malade pour une personne non-exposée; autrement dit, le fait d'être malade ne dépend pas de l'exposition ou non au facteur de risque. Hypothèse alternative (H1): RR>1. Cela signifie que la probabilité d'être malade pour une personne exposée est supérieure à la probabilité d'être malade pour une personne non-exposée; autrement dit, il y a dépendance entre l'apparition de la maladie et l'exposition au facteur de risque.

La détermination du χ^2 observé peut se faire en utilisant la formule suivante :

$$\chi^2 = \frac{(a\,.\,d - b\,.\,c)^2\,.\,N}{e_1\,.\,e_0\,.\,m_1\,.\,m_0}$$

Le χ^2 calculé peut alors être comparé à une valeur seuil (table de χ^2 en annexe) pour 1 dl et une confiance de 95%, 99% ou 99,9%.

V/ Intervalle de confiance du risque relatif

Le risque relatif étant une estimation, il est nécessaire de déterminer son intervalle de confiance.

a/Méthode de Miettinen

Cette méthode peut être appliquée aussi bien lors d'une étude de cohorte que lors d'une enquête cas-témoins.

Les limites inférieure (RRi) et supérieure (RRs) de l'intervalle de confiance sont déterminées au moyen de la formule suivante :

$$[RR_i\,;\,RR_s\,] = RR^{\,1\,\pm(\frac{Z}{\sqrt{\chi^2}})}$$

b/Méthode de Katz

Cette méthode ne peut être appliquée que dans le cadre d'une étude de cohorte.

Les limites inférieure (RRi) et supérieure (RRs) de l'intervalle de confiance sont déterminées au moyen de la formule suivante.

$$[RR_i\,;\,RR_s\,] = RR\,.\,e^{\pm Z\sqrt{\frac{1}{a}-\frac{1}{e_1}+\frac{1}{c}-\frac{1}{e_0}})}$$

c/ Méthode de Woolf

Cette méthode ne peut être appliquée que dans le cadre d'une enquête cas-témoins.

Les limites inférieure (RRi) et supérieure (RRs) de l'intervalle de confiance sont déterminées au moyen de la formule suivante :

$$[RR_i\,;\,RR_s\,] = RR\,.\,e^{\pm Z\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}})}$$

VI/ Le risque attribuable (RA)

Le risque attribuable (RA), aussi appelé fraction étiologique, correspond à la proportion des cas qui seraient évités si le facteur de risque était absent.

$$RA = \frac{E\ .\ (RR-1)}{1+E\ .\ (RR-1)}$$

On note E, la proportion de sujets exposés dans la population. Si la maladie est rare, celle-ci peut être estimée par la proportion de personnes exposées parmi les personnes non-malades.

$$E = \frac{b}{m_0}$$

Le risque attribuable peut également être calculé à partir de la formule suivante :

$$RA = 1 - \frac{c\ .m_0}{d\ .\ m_1}$$

Remarque

Le risque attribuable ne peut pas être déterminé dans le cadre d'une étude de cohorte. En effet, dans une telle étude, la proportion de sujets exposés dans la population n'est pas connue. L'exposition est un facteur arbitraire; c'est l'expérimentateur qui décide du nombre de personnes exposées dans son étude.

Le tableau suivant résume l'essentiel des deux études épidémiologiques.

<table>
<tr><th></th><th>Etude de cohorte</th><th>Enquête cas-témoins</th></tr>
<tr><td>Table de contingence</td><td>
<table>
<tr><th></th><th>Malade</th><th>Non-malade</th><th></th></tr>
<tr><td>Exposé</td><td>a</td><td>b</td><td>e_1</td></tr>
<tr><td>Non-exposé</td><td>c</td><td>d</td><td>e_0</td></tr>
</table>
</td><td>
<table>
<tr><th></th><th>Malade</th><th>Non-malade</th></tr>
<tr><td>Exposé</td><td>a</td><td>b</td></tr>
<tr><td>Non-exposé</td><td>c</td><td>d</td></tr>
<tr><td></td><td>m_1</td><td>m_0</td></tr>
</table>
</td></tr>
<tr><td>Risques absolus</td><td>$R_1 = \frac{a}{e_1} \quad R_0 = \frac{c}{e_0}$</td><td>Ne peuvent être déterminés</td></tr>
<tr><td>Risque relatif</td><td>$RR = \frac{R_1}{R_0}$</td><td>Si la maladie est rare:
$RR \approx \mathrm{OR} = \frac{a \cdot d}{b \cdot c}$</td></tr>
<tr><td>χ^2</td><td>$\chi^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{e_1 \cdot e_0 \cdot m_1 \cdot m_0}$</td><td>$\chi^2 = \frac{(a \cdot d - b \cdot c)^2 \cdot N}{e_1 \cdot e_0 \cdot m_1 \cdot m_0}$</td></tr>
<tr><td>Limites intervalle de confiance</td><td>$[RR_i ; RR_s] = RR^{1 \pm (\frac{z}{\sqrt{\chi^2}})}$
$[RR_i ; RR_s] = RR \cdot e^{\pm z \sqrt{\frac{1}{a} - \frac{1}{e_1} + \frac{1}{c} - \frac{1}{e_0}}}$</td><td>$[RR_i ; RR_s] = RR^{1 \pm (\frac{z}{\sqrt{\chi^2}})}$
$[RR_i ; RR_s] = RR \cdot e^{\pm z \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$</td></tr>
<tr><td>Risque attribuable</td><td>Ne peut être déterminé</td><td>$RA = 1 - \frac{c \cdot m_0}{d \cdot m_1}$</td></tr>
</table>

Problème :

Afin d'étudier les risques de l'accouchement liés à l'âge de la mère, une équipe de chercheurs a suivi 180 femmes de plus de quarante ans et 532 âgées entre vingt et trente ans. Parmi les femmes de plus de quarante ans, 29 ont dû accoucher par césarienne. Parmi les femmes plus jeunes, 53 ont eu recours à cette technique.

1. De quel type d'étude épidémiologique s'agit-il ? Justifiez votre réponse.
2. Présentez les résultats de l'étude sous forme d'un tableau.
3. Si c'est possible, déterminez le risque absolu de césarienne chez les femmes de plus de quarante ans.
4. Si c'est possible, déterminez le risque absolu de césarienne chez les femmes âgées entre vingt et trente ans.
5. Si c'est possible, déterminez le risque relatif. Est-il significatif ? Quel est son intervalle de confiance ?

Corrigé du problème :

1/ Deux groupes de femmes sont analysés: un groupe de femmes exposées (plus de quarante ans) et un groupe de femmes non-exposées (agées entre vingt et trente ans). Il s'agit donc d'une étude de cohorte.

2/

	Césarienne	**Césarienne**	
Exposée (>40 ans)	29	151	180
Non exposée (20-40 ans)	53	479	532

3/ le risque absolu pour f>40 ans :

$$\boldsymbol{R_1 = \frac{29}{180} = 0,161}$$

La probabilité de recours à une césarienne pour une femme de plus de quarante ans est donc de 16,1%.

4/ le risque absolu pour f>40 ans :

$$\boldsymbol{R_0 = \frac{53}{532} = 0,100}$$

La probabilité de recours à une césarienne pour une femme agée entre vingt et trente ans est donc de 10,0%.

5/ Le risque relatif (RR) est le rapport des incidences de la maladie étudiée (dans ce cas, le recours à une césarienne) pour les personnes exposées ou non au facteur de risque (dans ce cas, l'age de la mère). Il s'agit donc du rapport des risques absolus.

$$RR = \frac{R_1}{R_0} = \frac{0,161}{0,100} = 1,617$$

Le risque relatif étant plus grand que 1, on peut supposer une association entre l'age élevé de la mère et le recours à une césarienne.

Pour savoir si ce risque relatif est significatif ou non, on réalise un test de χ^2.

$$\chi^2 = \frac{(29 \times 479 - 151 \times 53)^2 \times 712}{180 \times 532 \times 82 \times 630} = 4,990$$

Comme la valeur du χ^2observée est plus grande que la valeur de χ^2seuil (χ^2 (1;0,95)=3,84), on peut conclure que le risque de recours à une césarienne est significativement plus élevé chez les femmes de plus de quarante ans que chez les femmes agées entre vingt et trente ans.

Détermination des limites de l'intervalle de confiance du risque relatif par la méthode de Miettinen.

$$RR_i = 1,617^{1-(1,96/\sqrt{4,990}} = 1,06$$

$$RR_r = 1,617^{1+(1,96/\sqrt{4,990}} = 2,47$$

L'intervalle de confiance du risque relatif est compris entre 1,06 et 2,47 (méthode de Miettinen).

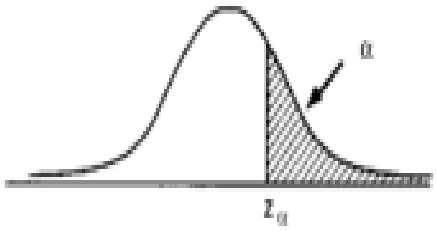
Détermination des limites de l'intervalle de confiance du risque relatif par la méthode de Katz.

$$RR_i = 1{,}617 \times e^{-1{,}96\times\sqrt{\frac{1}{29}-\frac{1}{180}+\frac{1}{53}-\frac{1}{532}}} = 1{,}06$$

$$RR_r = 1{,}617 \times e^{1{,}96\times\sqrt{\frac{1}{29}-\frac{1}{180}+\frac{1}{53}-\frac{1}{532}}} = 2{,}46$$

L'intervalle de confiance du risque relatif est compris entre 1,06 et 2,46 (méthode de Katz).

Table 1 : Table de la loi normale centre réduite



La table donne la valeur z_α telle que $\alpha = P(z > z_\alpha)$

α	0,000	0,005	0,010	0,015	0,020	0,025	0,030	0,035	0,040	0,045	0,050	0,055	0,060	0,065	0,070	0,075	0,080	0,085	0,090	0,095
0,0	∞	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695	1,645	1,598	1,555	1,514	1,476	1,440	1,405	1,372	1,341	1,311
0,1	1,282	1,254	1,227	1,200	1,175	1,150	1,126	1,103	1,080	1,058	1,036	1,015	0,994	0,974	0,954	0,935	0,915	0,896	0,878	0,860
0,2	0,842	0,824	0,806	0,789	0,772	0,755	0,739	0,722	0,706	0,690	0,674	0,659	0,643	0,628	0,613	0,598	0,583	0,568	0,553	0,539
0,3	0,524	0,510	0,496	0,482	0,468	0,454	0,440	0,426	0,412	0,399	0,385	0,372	0,358	0,345	0,332	0,319	0,305	0,292	0,279	0,266
0,4	0,253	0,240	0,228	0,215	0,202	0,189	0,176	0,164	0,151	0,138	0,126	0,113	0,100	0,088	0,075	0,063	0,050	0,038	0,025	0,013
0,5	0,000	-0,013	-0,025	-0,038	-0,050	-0,063	-0,075	-0,088	-0,100	-0,113	-0,126	-0,138	-0,151	-0,164	-0,176	-0,189	-0,202	-0,215	-0,228	-0,240
0,6	-0,253	-0,266	-0,279	-0,292	-0,305	-0,319	-0,332	-0,345	-0,358	-0,372	-0,385	-0,399	-0,412	-0,426	-0,440	-0,454	-0,468	-0,482	-0,496	-0,510
0,7	-0,524	-0,539	-0,553	-0,568	-0,583	-0,598	-0,613	-0,628	-0,643	-0,659	-0,674	-0,690	-0,706	-0,722	-0,739	-0,755	-0,772	-0,789	-0,806	-0,824
0,8	-0,842	-0,860	-0,878	-0,896	-0,915	-0,935	-0,954	-0,974	-0,994	-1,015	-1,036	-1,058	-1,080	-1,103	-1,126	-1,150	-1,175	-1,200	-1,227	-1,254
0,9	-1,282	-1,311	-1,341	-1,372	-1,405	-1,440	-1,476	-1,514	-1,555	-1,598	-1,645	-1,695	-1,751	-1,812	-1,881	-1,960	-2,054	-2,170	-2,326	-2,576

Exemple : pour $\alpha = 0,025$, la table donne $z_\alpha = 1,960$. C'est-à-dire que $P(z > 1,960) = 0,025$, ou, de façon équivalente, que $P(|z| > 1,96) = 0,05$.

Petites valeurs de α

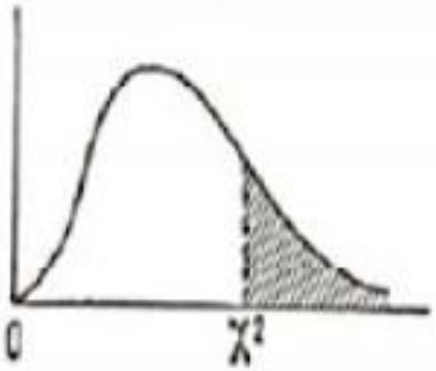
α	10^{-3}	$5\ 10^{-4}$	10^{-4}	$5\ 10^{-5}$	10^{-5}	$5\ 10^{-6}$	10^{-6}
z_α	3,09	3,29	3,72	3,89	4,26	4,42	4,75

Les grandes valeurs de α se déduisent par symétrie. Exemple : $P(z > -3,09) = 1 - 10^{-3} = 0,999$.

Table 2 : Table de χ^2

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.).



d.d.l. \ α	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 41	5,412	6,635	10,827
2	0,211	1,386	2,408	3,219	4,605	5 91	7,824	9,210	13,815
3	0,584	2,366	3,665	4,642	6,251	7 15	9,837	11,345	16,266
4	1,064	3,357	4,878	5,989	7,779	[illegible]	11,668	13,277	18,467
5	1,610	4,351	6,064	7,289	9,236	11 70	13,388	15,086	20,515
6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	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