République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

Concepts de base en Biostatistiques, population, sante, causalité, raisonnement probabiliste (théorème de Bayes), inférence, normalité, risque

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Concepts de base en Biostatistiques

1-HISTOIRE DE LA STATISTIQUE

Dans l'évolution de la statistique au cours de l'histoire on peut distinguer trois étapes :

- Dans cette étape qui va de la plus haute antiquité jusqu'au 18 ième siècle, la statistique se réduit à des recensements et inventaires de caractère démographique ou comptable et reste une statistique de constatation.
- Ce n'est qu'au 18 ième siècle que l'on voit apparaitre le rôle prévisionnel des statistiques et c'est à Adolphe Quételet que l'on doit l'idée que la statistique est une science s'appuyant sur les probabilités.
- Sous l'impulsion de grands mathématiciens, la troisième étape est révolutionnaire car elle va permettre désormais d'ouvrir à la statistique des horizons illimités, surtout avec l'essor de l'informatique qui va permettre de traiter et d'analyser un plus grand nombre de données.

2-Définition de la STATISTIQUE

Ensemble des méthodes qui servent à organiser les épreuves fournissant des observations, à analyser celles-ci et à interpréter les résultats.

Le but est de présenter les données pour que l'on puisse en prendre connaissance facilement.

Etymologiquement : science de l'´état.

La statistique s'applique à la plupart des disciplines : agronomie, biologie, démographie, ´économie, sociologie, linguistique, psychologie.

- Connue sous le nom biométrie ,la Biostatistique est l'application des statistiques en biologie .
- la Biostatistique est la science dont l'objet est de recueillir, de traiter et d'analyser des données issues de l'observation des phénomènes biologiques.
- Elle utilise des concepts et principes statistiques à des données médicales, biologiques et de santé publique.

Exemples:

- ❑ Les effets d'un médicament.
- ❑ L'effet du niveau du cholestérol sur la pression artérielle.
- ❑ Le nombre de patients admis durant les fins de semaine aux urgences.
- ❑ Distribution des pandémies.

Elle englobe :

- ❖ La conception d'expériences biologiques ;
- ❖ La collecte d'informations ;
- ❖ L'analyse des données chiffrées ;
- ❖ L'interprétation des résultats et conclusion

Domaines d'applications de la statistique

- Démographie (recensement).
- Economie.
- Sociologie.
- Politique.

- Médecine.
- Physique.
- Climatologie.
- Ecologie

3-Biostatistique: champs et applications

La Biostatistique nous permet de

- ❑ Description des moyens et l'état de santé d'une population
  - Causes de décès, morbidité, surveillance sanitaire….
- ❑ Mesurer la précision d'une estimation
- ❑ Définir le dégrée d'association entre une série de caractères et d'´evenements.
- Évaluer un test ou d'un signe
  - Sémiologie quantitative : spécificité, sensibilité, valeurs prédictives
- Évaluer un traitement
  - Essai thérapeutique
- Recherche de facteurs étiologiques
- Économie de la santé
- Évaluation de la qualité et contrôle de production
- Prévision (Nbre de malades attendus,...)

4- Les méthodes de Biostatistique

- L'analyse statistique se subdivise en deux parties

1-Analyse "déductive" ou descriptive

- a pour but de résumer et de présenter les données observées pour que l'on puisse en prendre connaissance facilement : tableaux, graphiques, .paramètres de position, de dispersion, étude de la relation.

2-Analyse "inductive" ou Inférentielle

- l'ensemble des méthodes permettant de formuler un jugement.
- Elle nécessite des outils mathématiques plus pointus (théorie des probabilités).
- permet d'étendre ou de généraliser, dans certaines conditions, les conclusions obtenues. Cette phase comporte certains risques d'erreur qui peuvent être mesurés en faisant appel à la théorie des probabilités.
- Ces étapes ne sont pas indépendantes.

3- Statistique exploratoire

1-Statistique Descriptive

- C'est un ensemble de méthodes permettant de décrire et d'analyser des phénomènes susceptibles d'être dénombrés et classées.
- Synthèse de l'information: résumés statistiques.
- Elle a pour but de décrire et non d'expliquer
- Ces méthodes comportent :
  - **Les tableaux** : distributions de fréquences.
  - **Les diagrammes** : graphiques.
  - **Les paramètres** statistiques : Réduction des données à quelques valeurs numériques caractéristiques.

1-1-Concepts de Base en statistique descriptive:

Population : ensemble total d'objets ou d'individus à étudier, à partir duquel sont extraits des échantillons(ensemble des unités statistiques).

- Exemples :Population des lymphocytes,hommes âgés de 50 ans et plus et ayant déjà eu une attaque cardiaque,femmes atteintes du cancer du sein.

Echantillon :

- Un échantillon représentatif est un sous-ensemble de taille finie d'une population.
- Il est choisi au hasard dans la population.
- La moyenne et l'écart-type s de l'échantillon sont des variables aléatoires, variant d'un échantillon à l'autre, et sont appelées statistiques d'échantillon, statistiques aléatoires ou estimateurs ici respectivement (**μ** et **σ**).
- Échantillonnage : mécanisme de génération de l'échantillon.
- **Exemple :** échantillonnage aléatoire simple, échantillonnage stratifié.

Unité statistique:

Unité sur laquelle sont effectuées des observations: Individu, animal, organe, cellule, champ de microscope,...

Variables:

- Caractéristique mesurable pour toutes les unités statistiques.
- **Exemple** : sexe, taille, poids, âge, Groupe ABO: A, B, AB, O.

Types de Variable :

- **Qualitatives**
  - **Ordinales** (Niveau d'études; primaire, moyen,secondaire et universitaire)
  - **Nominales** (ex :Couleurs des cheveux : blond, brun, blond, noir….
- **Quantitatives**
  - **Discontinues** (ex: nombre d'enfants dans les familles : 1, 2, 1, 4, 0 ….)
  - **Continues** (ex: le poids d'un nouveau né)

1-2 Variabilité :

- Variabilité totale
  - Variabilité de la mesure
    - Essayer de mesurer plusieurs (100) fois la taille en mm d'un individu : vous trouverez des valeurs différentes cependant dans l'absolu un individu a une taille et une seule.

- Variabilité inter individus
  - Si vous observez des personnes dans la rue, vous constatez qu'elles n'ont pas toutes la même couleur de cheveux.

- Variabilité intra individu
- Si vous mesurez la tension artérielle d'un individu à différents moments de la journée ou au même moment mais plusieurs jours de suite, vous obtiendrez des valeurs différentes.
- Du fait de la variabilité, on est dans le domaine de l'incertain.
- Cette science de l'incertain, c'est le défi qu'a relevé la statistique en s'appuyant sur le concept de probabilité.
- Plutôt qu'une seule valeur, la prise en compte de l'incertain permet de déterminer un intervalle à l'intérieur duquel on a une certaine probabilité de se situer et donc un risque de ne pas y être.
- Exemple courbe de croissance dans le carnet de santé

2-Statistique Inférentielle

- La Statistique Inférentielle (Inductive) consiste à prédire les caractéristiques d'une Population inconnue à partir des statistiques déterminées dans un échantillon représentatif de cette population.
- À partir d'un nombre réduit d'observations, peut-on répondre à des hypothèses faites sur une population plus large et avec quel niveau de fiabilité.
- L'opération de "remontée" de l'échantillon à la population est appelée inférence statistique.
- Elle nécessite d'un modèle mathématique:(ex la régression)

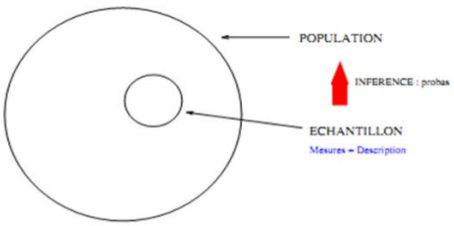
Statistique Inférentielle versus descriptive

- La statistique descriptive est la branche des statistiques qui regroupent les nombreuses techniques utilisées pour décrire un ensemble relativement important de données.
- La statistique Inférentielle est un ensemble de méthodes permettant de tirer des conclusions fiables à partir de données d'échantillons statistique

- Les données de l'échantillon ne nous intéressent pas en tant que telles Les résumer, les représenter est le domaine de la statistique descriptive.

- Elles nous intéressent car elles donnent une information sur une ensemble plus vaste dont elles proviennent : la population
- Si l'on prélève un nouveau jeu de données, les nouvelles observations seront différentes des précédentes
- L'inférence statistique suppose de prendre en compte l'aspect aléatoire des données.
- Elle s'appuie donc sur des outils probabilistes , chaque individu constitutif de la population doit avoir la même chance de figurer dans l'échantillon.

N.B: L'échantillon doit ainsi être prélevé au hasard.



3-LA STATISTIQUE EXPLORATOIRE

- Les techniques d'analyse des données ou, plus précisément, de statistique exploratoire multidimensionnelle, sont utilisées pour l'étude descriptive de tableaux présentant un nombre de variables en lignes, individus, colonnes, variant de quelques dizaines à quelques milliers.
- La production de graphiques et indicateurs synthétiques permettent de résumer les structures et principales caractéristiques des ces tableaux de grand format.
- Cette exploration présente un intérêt dans la recherche clinique et aussi ,développement,industriel, tertiaire.

5- Raisonnement probabiliste (Théorème de Bayes)

- C'est un résultat de base en théorie des probabilités, issu des travaux du révérend Thomas Bayes et retrouvé ensuite indépendamment par Laplace.
- Le théorème de Bayes est utilisée dans l'inférence statistique pour mettre à jour ou *actualiser* les estimations d'une probabilité ou d'un paramètre quelconque, à partir des *observations* et des lois de probabilité de ces observations.
- le théorème de Bayes énonce des probabilités conditionnelles :
  soit A et B deux *évènements*, le théorème de Bayes permet de déterminer la probabilité de A *sachant* B, si l'on connaît les probabilités de A,de B et de B *sachant* A.

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

6- La normalité :

- En statistiques, les **tests de normalité** permettent de vérifier si des données réelles suivent une loi normale ou non. Les tests de normalité sont des cas particuliers des tests d'adéquation (ou tests d'ajustement, tests permettant de comparer des distributions), appliqués à une loi normale.
- Ces tests prennent une place importante en statistique. En effet, de nombreux tests supposent la normalité des distributions pour être applicables. En toute rigueur, il est indispensable de vérifier la normalité avant d'utiliser les tests. Cependant, de nombreux tests sont suffisamment robustes pour être utilisables même si les distributions s'écartent de la loi normale.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

LES INDICATEURS DE MORBIDITE ET DE MORTALITE

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

LES INDICATEURS DE MORBIDITE

I/ INTRODUCTION :

Les concepts d'indicateurs ont été utilisés au départ en économie, avec les notions d'indicateurs sociaux qui étaient destinés à « décrire et à mesurer les résultats pour l'action à entreprendre ».
L'appropriation du terme d'indicateur de santé par les professionnels de santé (Francophones) s'est faite à partir des années 1970 dans la Revue d'épidémiologie et de santé publique.
les indicateurs (1981) comme « **des variables qui aident à mesurer, directement ou indirectement, les changements dans la situation sanitaire et à apprécier dans quelle mesure les objectifs et cibles d'un programme sont atteints** » OMS.
Les indicateurs de santé sont donc des paramètres de mesure qui reflètent diverses composantes de l'état de santé. Ils ont pour intérêt de décrire l'état de santé d'une population, planifier et évaluer les actions de santé et sont perçus comme des instruments de la surveillance épidémiologique.

Selon le phénomène observé, maladie ou décès, on distingue :
-Les indicateurs de morbidité qui décrivent la fréquence des maladies
-Les indicateurs de mortalité qui décrivent la fréquence des décès

II/ LES MESURES DE BASE :

Toute mesure en épidémiologie doit être précédée par une définition des deux termes du rapport (N/D) :
-Définition d'un cas au numérateur (N)
-Définition de la population d'étude au dénominateur (D)

1-Proportion :

Dans une proportion, le numérateur est une part du dénominateur. Le (N) et le (D) sont donc de même nature ; P= a / (a+b)
Une proportion s'exprime sous forme d'un nombre compris entre 0 et 1, ou bien sous forme d'un pourcentage
Exemple : Dans une population de 7500 enfants de moins de cinq ans, on constate que 5300 sont correctement vaccinés contre la rougeole. La proportion d'enfants vaccinés est de P= a / (a+b) = 5300/7500 = 0,707 = 70,7%. Cette proportion est communément appelée « couverture vaccinale ».

2-Ratio :

Un ratio représente le rapport entre les effectifs de deux classes d'une même variable. Le (N) et le (D) sont de même nature, mais exclusifs l'un de l'autre.
Un ratio s'exprime par un nombre sans unités
Exemple : Dans une population de 100 individus, on observe 49 hommes et 51 femmes. Le ratio h/f (sex ratio en anglais) = 49/51 = 0,96 (0,96 homme pour une femme).

3-Indice :

Un indice est le rapport de deux effectifs qui sont de nature différente. On l'utilise surtout comme indicateurs de fonctionnement, notamment en économie de la santé.
Exemple : Dans un hôpital on dispose de 850 lits et 10 médecins. On calcule l'indice lits d'hôpital / médecin ou le nombre de lits par médecins = 850/10 = 85 (85 lits par médecins).

4-Taux :
Un taux mesure la probabilité de survenue d'un événement au cours du temps.
Au (N) figurent des individus ayant subi un événement pendant une période de temps déterminé et au (D) figure l'ensemble des individus susceptibles de connaître l'événement pendant cette période (la population à risque).

III/ LES INDICATEURS DEMOGRAPHIQUES :

Les taux bruts sont calculés sur l'ensemble de la population et les taux spécifiques sont calculés pour une sous population définie par tranche d'age, sexe ou par catégorie socio professionnelle.

1-Taux brut de natalité (TBN) :

TBN= Nombre de naissances vivantes d'une région donnée pendant une période donnée / Population moyenne de la même région et durant la même période × 1000

2-Taux d'accroissement naturel :
Ce taux représente la différence entre le taux de natalité et le taux de mortalité

3-Taux de fécondité :
Nombre de naissances (filles et garçons) durant l'année × 1000 / Nombre de femmes en age de procréer (15-49) durant la même année

4- Espérance de vie à la naissance : vie moyenne
C'est le nombre moyen d'années qu'un nouveau né peut espérer vivre

IV/ LES INDICATEURS DE MORBIDITE :

1- La prévalence :
La prévalence est un indice important largement utilisé pour déterminer les besoins médicaux et sociaux surtout dans le cas des maladies chroniques. Elle mesure la présence d'une maladie.
Il peut s'agir d'une prévalence instantanée (à un moment donné) ou de période (pendant une période de temps).

Taux de prévalence = Nombre total de cas d'une maladie (anciens et nouveaux cas) dans une région donnée et pendant une période donnée / Population moyenne de la même région et de la même période × 100

2- L'incidence :
L'incidence est un indice important des besoins en soins préventifs, utile pour les maladies aigues et chroniques. L'incidence permet d'évaluer l'efficacité des mesures de contrôle d'une maladie à caractère de masse et exprime donc la vitesse d'apparition d'une maladie dans une population.

Taux d'incidence = Nombre de nouveaux cas d'une maladie dans une région donnée et pendant une période donnée / Population moyenne de la même région et de la même période × 100

Exemple :
Dans une population de 500 personnes relevés au cours de l'année 2002, 74 cas ont présenté un infarctus aigu du myocarde. Le taux d'incidence = 74/500= 0,148 =148 cas pour 1000 personnes

2-1/ Taux d'attaque :
Ce taux désigne la proportion de sujets atteints au cours d'une période définie et déterminée. On l'utilise en règle à la suite d'une exposition de courte durée, par exemple lors d'une épidémie.

Exemple :
La cantine d'une école recevant 250 enfants a été le siège d'une toxi-infection alimentaire collective. 52 enfants ont présenté l'affection. Le taux d'attaque est de 52/250 = 208 cas pour 1000.

2-2/ Densité de l'incidence (Id) :

Densité de l'incidence = Nombre de nouveaux cas d'une maladie pendant une période de temps
Nombre total de personnes années d'exposition au facteur étudié pendant la même période

Exemple : **une** personne exposée pendant **3 ans** à un facteur étudié =
3 personnes années d'exposition

3- Relation entre le taux d'incidence et le taux de prévalence:
Le taux de prévalence (P) dépend à la fois du taux d'incidence et de la durée de la maladie

P = taux d'incidence × durée moyenne de la maladie
(**I**) (**D**)

4- Létalité :
La létalité représente la proportion de malades atteints d'une affection qui décèdent de cette affection durant une période donnée. Cet indicateur témoigne de la gravité de la maladie et de la qualité des soins.
Un taux élevé de létalité implique un problème dans la prise en charge de la maladie.

Taux de létalité = Nombre de décès attribuable à une maladie donnée au cours d'une période donnée / Nombre total de cas de cette maladie au cours de la même période

5- Relation entre mortalité, incidence et létalité :

Létalité = Mortalité / Incidence

Mortalité = Incidence × Létalité

V-LES INDICATEURS DE MORTALITE

1-Taux brut de mortalité (TBM) :

TBM = **Nombre** de décès au cours de l'année × 1000
Population totale au milieu de la même année

2-Taux de mortalité spécifique par âge :

En raison de l'influence déterminante de l'âge sur la mortalité, il faut calculer le taux de mortalité pour chaque tranche d'âge.

Nombre de personnes d'âge particulier décédées au cours de l'année × 1000
Population de cet âge au milieu de la même année

3-Taux de mortalité spécifique selon le sexe :

Nombre de femmes ou d'hommes décédés au cours de l'année × 1000
Population moyenne du même sexe au milieu de la même année

4-Taux de mortalité proportionnelle (TMP) :

Ce taux représente la proportion de mortalité qui peut être attribuée à une cause spécifique

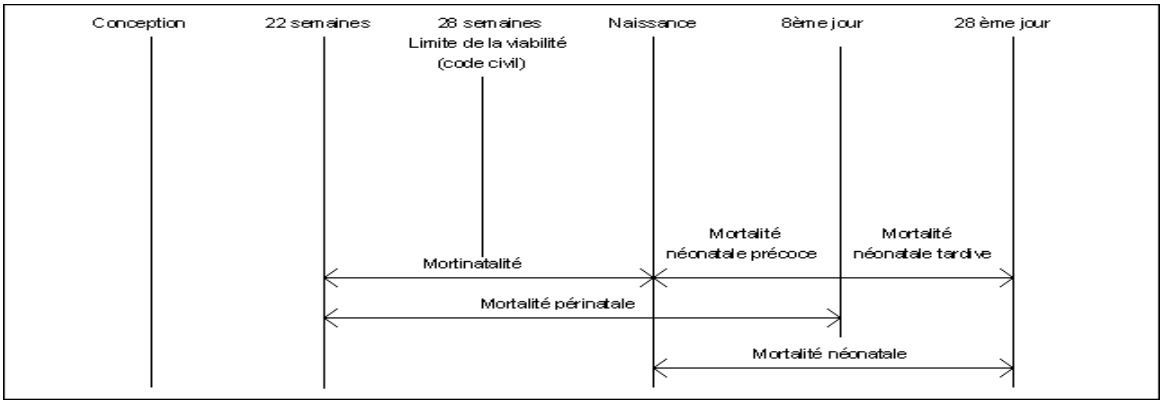
TMP = Nombre de décès attribuable à une maladie donnée au cours de l'année × 1000
Nombre de décès dans la population durant la même année

5-Taux de mortalité infantile (TMI) :

Ce taux mesure surtout le niveau médical et social de la collectivité

TMI = Nombre de décès d'enfants de moins de 1 ans (0 à 11 mois et 29 jours) durant l'année
Nombre de naissances vivantes de la même année × 1000

Ce taux de mortalité infantile peut se décomposer comme suit :



5-1/ Taux de mortalité néonatale précoce :

Nombre de décès des nouveaux nés de 0 à 07 jours durant l'année × 1000
Nombre de naissances vivantes de la même année

5-2/ Taux de mortalité néonatale tardive :

Nombre de décès des nouveaux nés de 8 à 28 jours durant l'année × 1000
Nombre de naissances vivantes de la même année

5-3/ Taux de mortalité post natale :

Nombre de décès de nourrissons de 29 jours à 11 mois et 29 jours durant l'année × 1000
Nombre de naissances vivantes de la même année

6-Taux de mortalité maternelle :

Nombre de décès des mères dus à l'accouchement,
aux complications de la grossesse et suites de couches durant l'année × 1000
Nombre de naissances vivantes de la même année

VI/ BIBLIOGRAPHIE :

Mesli MF, Bouziani M. Epidémiologie Objet et Méthodes. Laboratoire de biostatistiques. Faculté de médecine d'Oran. Mai 2007

Abdelouahab A. Cours à l'usage des étudiants en médecine .Les indicateurs de santé. Année universitaire 2005-2006
Dr.R.TALHI – Pr. MF. MESLI, Cours à l'usage des étudiants en sciences médicales LES INDICATEURS DE SANTE

Belateche F. Cours à l'usage des étudiants en médecine. Les indicateurs de santé. Année universitaire 2004-2005

Mesli MF, Mokhtari A. Biostatistique description et analyse des données en médecine et en biologie. Les éditions universitaires le fennec 2002

Ancelle T. Statistique Epidémiologie. Edition 2002

Bezzaoucha A. Epidémiologie et biostatistiques. Office des publications universitaires decembre 1996

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

TABLEAU DE PROPAGATION DES MALADIES (CARACTERISTIQUES DE PERSONNES, TEMPS ET LIEU) DEFINITION DES POPULATIONS (CIBLE, A RISQUE)

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

TABLEAU DE PROPAGATION DES MALADIES (CARACTERISTIQUES DE PERSONNES, TEMPS ET LIEU) DEFINITION DES POPULATIONS (CIBLE, A RISQUE)

A-Définition de l'épidémiologie :

1- Etymologie :

L'origine grecque du mot est simple :

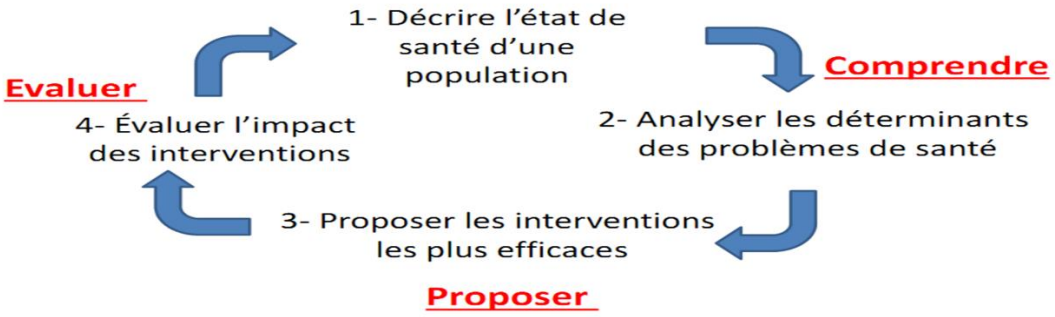
- Epi : au dessus de ,veut dire « sur» ;
- Demos :veut dire «peuple – population » ;
- Logos : discours
- EPI –DEMOS –LOGOS – veut dire « Etude ou connaissance » ;
- Par conséquent : l'Epidémiologie est l'étude de ce qui arrive aux individus »

2- de l'épidémiologie Concept:

« Étude de la distribution et des déterminants d'une maladie dans des populations humaines, et application des résultats de cette étude dans la lutte contre cette maladie. »

B-Cinq points importants :

- Distribution
- Problèmes de santé
- Déterminants :
- Populations humaines
- Prévention



Cycle de l'épidémiologie

Les cinq fonctions de l'épidémiologie

- 1. La surveillance épidémiologique
- 2. La mesure de l'importance des problèmes de santé
- 3. La recherche étiologique
- 4. L'identification des groupes à risque élevé
- 5. L'évaluation de la santé

C-Populations en épidémiologie :

Population cible :

La population cible est la population que nous voulons observer, tandis que la population observée est la population que nous pouvons observer. Ce processus a pour but de faire en sorte que la population observée se rapproche autant que possible de la population cible.

Population à risque :

Qualités et caractéristiques de divers types de populations dans un groupe social ou géographique en insistant sur la démographie, l'état de santé et les facteurs socio-économiques.

E-Définition du risque :

« Conséquences » négatives sur la santé de la population et probabilité d'observer ces conséquences à la suite d'une exposition à un agent dangereux ».

- Ainsi, le risque peut être défini par la formule suivante :
- Risque = conséquences x probabilité
- Les conséquences du risque correspondent aux effets négatifs sur la santé humaine d'une population résultants de l'exposition à un agent dangereux. Elles peuvent toucher la santé directement (morbidité, incapacité et mortalité) ou indirectement, par exemple à travers des impacts sociaux ou économiques.

- La probabilité du risque à la santé est associée à différents types de probabilité :
- la probabilité d'apparition de l'agent dangereux;
- la probabilité d'exposition à l'agent dangereux;
- la probabilité d'observer des effets négatifs sur la santé d'une population après l'exposition à l'agent dangereux.
- Cette probabilité peut être estimée quantitativement (pourcentage, nombre possible de cas, etc.) ou qualitativement (rare, probable, presque certain, etc.).

F-Branches d'épidémiologie

1-L'épidémiologie descriptive

- C'est l'étude de la distribution de la maladie dans les populations selon les caractéristiques de :

– Personne :

- • âge
- • sexe
- • état civil
- • profession

– Lieu : résidence, région, pays, lieu de travail,…

– Temps : saison, années, mois, etc.. (temps d'observation)

L'épidémiologie a été conçue pour répondre à la question:

Qui ? Quoi ? Quand ? où ? pourquoi ?

Objectifs de L'épidémiologie descriptive :
Elle a pour objet de décrire la fréquence et la répartition de phénomènes de santé ou de déterminants de santé dans les populations, en fonction de caractéristiques humaines, spatiales, temporelles. Il s'agit donc d'apporter des réponses pour les questions suivantes :

Chez qui ?	Personnes
Où ?	Lieu
Quand ?	temps
Quoi ?	
pourquoi ?	

Principes de l'épidémiologie descriptive :
Elle est basée sur l'utilisation d'indicateurs simples :

- des taux de mortalité lorsque l'on s'intéresse aux décès,
- des taux de prévalence et d'incidence lorsque l'on s'intéresse aux maladies.
- Les données utilisées par l'épidémiologie descriptive peuvent être issues des
  - statistiques sanitaires (statistiques de mortalité, registres des cancers, ou registres pour d'autres maladies,
  - données issues des déclarations obligatoires pour les maladies transmissibles)
  - ou d'études qui ont été réalisées dans le but de fournir une information spécifique adaptée à un objectif particulier,

2- L'épidémiologie analytique

Afin d'estimer le lien entre l'exposition à certains facteurs et la survenue ultérieure de maladie (ou événement de santé) au moyen d'enquêtes réalisées chez des individus.

- La question à laquelle on veut répondre ici est « pourquoi ?»
- Les études d'observations mises en place (cas-témoins, ou exposés-non exposés) permettent de comparer les groupes d'individus définis en fonction de la maladie (malade ou non malade) et d'un facteur d'exposition (exposé ou non exposé à ce facteur), en estimant un risque (= probabilité) associé.

- Lorsque cette relation est établie, on peut ainsi déterminer par combien est multiplié1 (ou divisé2) la probabilité de survenue de la maladie chez les sujets exposés au facteur par rapport aux sujets non exposés à ce facteur.
- Selon le champ d'application considéré, on parle d'épidémiologie étiologique (on s'intéresse à l'étude des causes des maladies), d'épidémiologie génétique, d'épidémiologie sociale, d'épidémiologie environnementale.

3- L'épidémiologie évaluative

Démontrer l'efficacité de l'intervention qui est exprimée sous forme d'un état de santé =>

Évaluation de recherche (ce sont les expériences)

Vérifier l'efficacité de l'intervention telle qu'elle a été mise en en place dans la pratique habituelle =>

Évaluation professionnelle (ce sont les études d'observation évaluatives qui utilisent les méthodes de l'épidémiologie descriptive)

G- Quelle est la place de l'épidémiologie en santé publique

- L'épidémiologie vise à quantifier l'état de santé des populations, permettant ainsi
- L'identification des problèmes de santé
- La priorisation des problèmes de santé
- La planification d'actions pour résoudre ces problèmes
- L'évaluation des actions mises en place

H- Conclusion:

La description des données d'une maladie ou d'un événement de santé en fonction du temps, du lieu et des caractéristiques individuelles est la première étape de toute analyse épidémiologique.

- En pratique, elle est souvent négligée, bien que faisant appel à des techniques statistiques et de représentation relativement conventionnelles et faciles à utiliser.
- L'expérience en épidémiologie de terrain montre cependant que cette phase d'épidémiologie descriptive est souvent décisive pour la suite des investigations.
- Elle doit donc être aussi complète, systématique et précise que possible, afin d'appréhender au mieux le phénomène de santé étudié.
- Le plus souvent, quelques tableaux et graphiques permettront assez facilement de représenter et de comparer les données en termes de caractéristiques de temps, de lieu et de personnes, et de dégager des hypothèses concernant la nature, la source et le mode de transmission d'une maladie, hypothèses qui seront ensuite testées au cours de la phase analytique de l'investigation et plus généralement de toute enquête épidémiologique.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

Etudes épidémiologiques : Structure des études descriptives et analytiques, Etude descriptive, transversale, cas-témoins, cohorte, Structure incomplète.

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Structure des Etudes épidémiologiques :

Structure des études descriptives et analytiques, Etude descriptive, transversale, cas-témoins, cohorte, Structure incomplète, échantillonnage et biais

1-Introduction :

L'épidémiologie :

Raisonnement et méthode appliqués à la description des phénomènes de santé, à l'explication de leur étiologie et à la recherche des méthodes d'intervention les plus efficaces.

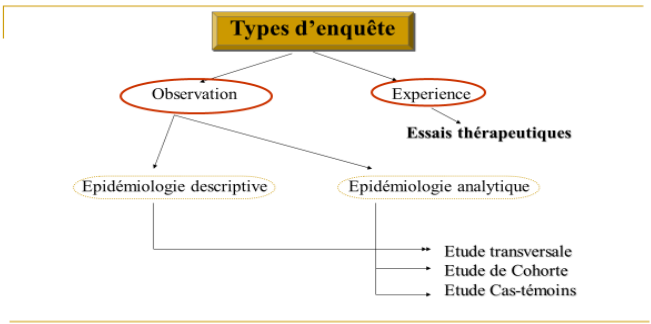
On distingue l'épidémiologie :

-**Descriptive :** Il y a plus de... chez les…

-**Explicative :** Etiologique (analytique) : Les gens exposés à… sont plus atteints par… que les non exposés.

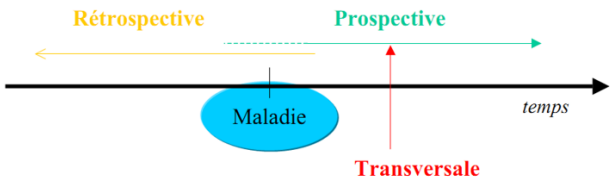
-**Evaluative : (**d'intervention) : Quand je donne… il y a moins de…

2-Structures des études épidémiologie :

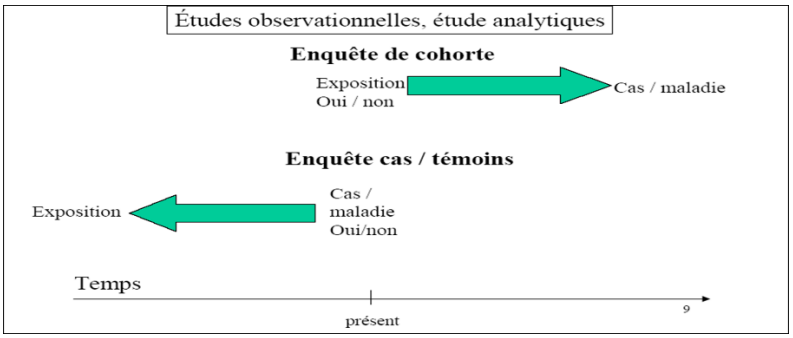


Types d'études selon la chronologie :

Classification chronologique :



Types d'études selon le recueil des données:



1- Enquêtes transversales (Observation)

• A un instant donné un enquêteur cherche à connaître le nombre de personnes qui présentent un caractère donné dans une population.

• Les Enquêtes transversales permettent de mesurer le nombre de cas (prévalence) du problème ciblé mais aussi de recueillir les facteurs associés : caractéristiques socio-économiques, niveau d'éducation, conditions environnementales, style de vie…

• Possibilité de déduction d'une hypothèse étiologique.

• **Exemple :** Enquête de prévalence des infections Nosocomiales dans un hôpital

– au Maroc cette prévalence était de 7,7 % dans les hôpitaux régionaux en 1994

Avantages :

- Facilité de mise en œuvre

Inconvénients :

- Biais de sélection (patients sortis de la cohorte car malades par ex)
- Relation temporelle exposition – maladie ?

2- Enquêtes analytiques :

- • Elle s'intéresse donc à la recherche des déterminants c.à.d. origines ou étiologies des phénomènes de santé.
- • Elle vise à comparer les fréquences d'une maladie dans différents groupes afin de mettre en évidence les « facteurs de risque »

Principe

- Etudier la relation entre un phénomène de santé (M) et des facteurs susceptibles de l'influencer (E)

– M = pathologie ou comportements de santé

– E = facteur biologique, comportemental, environnemental, pathologie… (exposition)

• Exemple :

– association entre contraception orale (E) et cancer du sein (M)

Buts :

• mettre en évidence l'association entre une exposition (E) et une maladie (M)

• Les gens exposés à (E) sont plus atteints par (M) que les non exposés.

• Comparaison +++

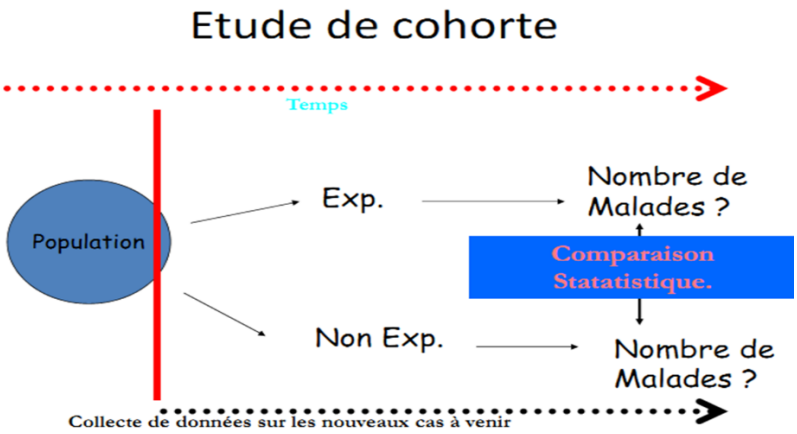
• Notion de risque

a-Enquêtes de cohorte :

Cohorte = groupe de sujets suivi dans le temps

Les sujets sont sélectionnés sur la base de leur exposition au facteur étudié et sont suivi de manière prospective pour voir s'ils développent la maladie.

Schémas d'étude :



Avantages :

- ✓ Étude du risque de plusieurs maladies (Ex : exposition à l'efavirenz et risque de tératogénicité et de troubles psychiatriques).
- ✓ Expositions rares (Ex: exposition à l'atazanavir et apparition de lithiases)
- ✓ Séquence chronologique exposition / maladie plus facile à établir
- ✓ Données d'incidence et donc meilleure estimation du risque (RR)

Inconvénients

- Souvent coûteuses et longues
- Perdus de vue
- Peu performantes pour les maladies rares, ou à temps de latence long

B- Enquêtes cas-témoins

Définition

• Étude rétrospective

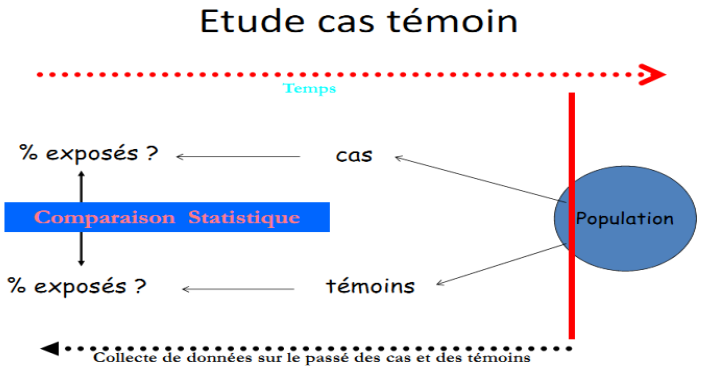
• Sélection de la population en fonction de la présence ou non du critère de la maladie

• Recherche rétrospective de l'exposition aux facteurs de risque

• **Observation :**

• L'enquêteur choisit les groupes étudiés sur la base de leur statut malade (cas) et non malade (témoins). Ce sont des études rétrospectives, c'est-à-dire que l'enquêteur va chercher dans le passé des individus des deux groupes l'exposition au facteur étudié.

Schémas d'étude :



Avantages :

• **Intérêt**

– Résultats rapides

– Coût faible

• **Indications**

– Maladies dont les périodes de latence sont longues

ex: cancer du poumon et tabac

– Maladies rares

– Étude de plusieurs facteurs de risque pour une seule maladie

ex: cancer de l'œsophage et exposition au tabac et à l'alcool

Inconveneints :

Non adapté pour l'étude

– De plusieurs maladies

– Des expositions rares

– De la relation chronologique entre exposition et maladie

• Ne permettent pas un calcul direct des taux d'incidence de la maladie chez les patients exposés et non exposés

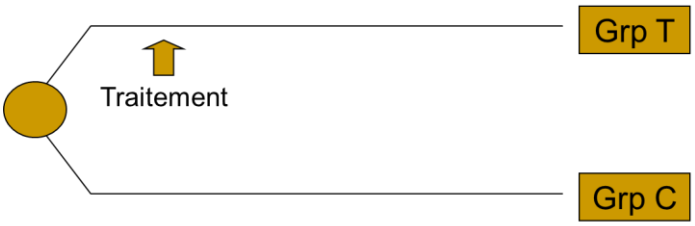
– Estimation du risque par OR (Odds Ratio)

• Les risques de biais sont importants particulièrement ceux de sélection et de mémoire (recueil rétrospectif)

c- Essai clinique :

- Groupe avec et sans intervention
- Mesure de l'effet de l'intervention
- 2 types de conclusions:
  - jugement de signification (statistique)
  - jugement de causalité (qualité de l'essai)

Schémas d'étude :



- Groupes identiques
  - même type de patients
  - même stade de la maladie, etc.
- qui ne diffèrent que par le traitement appliqué
- Si, à la fin, il existe une différence, celle-ci n'est due qu'au traitement

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

ESSAI RANDOMISE

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

L'essai randomisé

I-Structure d'un essai therapeutique controlé:

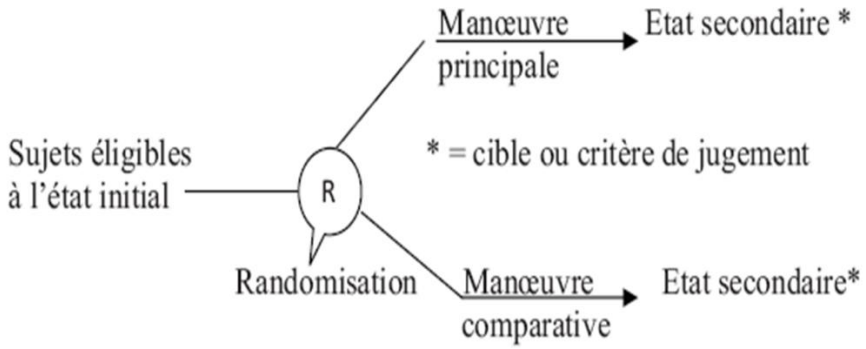
- Essai thérapeutique randomisé est un cas particulier de l'étude cohorte, il s'agit d'une expérience réalisée pour évaluer l'utilité de traitements, de vaccins ou de toutes autres manœuvres préventives ou thérapeutique.
- La répartition des sujets éligibles entre les différentes cohortes est réalisée par un tirage au sort, ce qui assure dès le départ la comparabilité de celles-ci.
- L'essai randomisé se distingue de l'étude cohorte par deux caractéristiques essentielles :

• Les sujets éligibles de l'étude sont sélectionnés par tirage au sort pour former la cohorte principale et la cohorte de comparaison, on dit aussi que les sujets sont **randomisés** ;

• L'application des manœuvres (traitements, vaccins…) est **contrôlée directement** par l'investigateur ou le chercheur en charge de l'étude.

- Il pourrait même être mieux indiqué de comparer les examens en les réalisant chez un même sujet selon le dessin d'une étude randomisée en cross-over : chaque sujet étant son propre témoin.

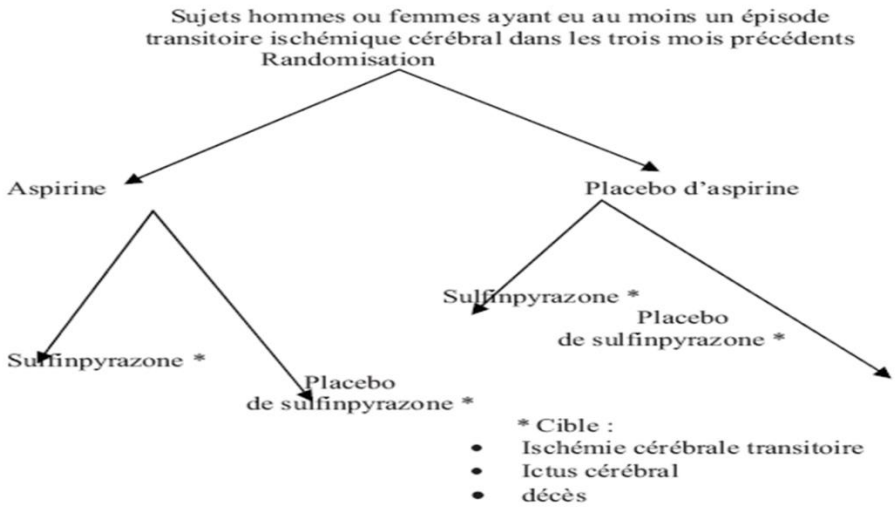
Le dessin (ou l'architecture) d'un essai randomisé peut être schématisé comme suit :



- Un sujet devrait avoir la même probabilité de subir l'une ou l'autre des manœuvres. Cela constitue le respect du principe d'ambivalence qui justifie le tirage au sort.
- La cible peut être mesurée par une densité d'incidence, un pourcentage ou une moyenne. La comparaison de ces paramètres s'effectue par un test statistique.

- Les résultats de l'essai, s'il a été correctement réalisé, peuvent être attribués à la manœuvre car le tirage au sort, qui répartit les différents facteurs pronostiques de manière identique dans les deux groupes, a constitué deux groupes comparables à l'état initial.

Le dessin de l'essai peut être schématisé comme suit :



Cet essai a envisagé d'étudier l'utilité de l'aspirine et du sulfinpyrazone et de leur association dans la prévention des accidents cérébraux, :
L'essai a été réalisé chez des sujets ayant présenté des accidents ischémiques cérébraux transitoires.
Ainsi, les sujets éligibles de l'étude furent soumis à l'un des quatre régimes thérapeutiques suivants possibles : aspirine seul, sulfinpyrazone seul, aspirine + sulfinpyrazone, ou deux placebos.

L'organisation d'un essai selon un plan factoriel permet d'étudier l'interaction entre les deux traitements. Les résultats de l'essai ont été les suivants :

• l'aspirine a réduit le risque d'accidents ischémiques cérébraux récidivants, d'ictus et de décès ;

• le sulfinpyrazone n'a pas d'utilité.

II-Substituts au tirage au sort :

Des méthodes proposées pour remplacer le tirage au sort, sont incorrectes car elles introduisent un biais susceptible d'entraver la validité des résultats de l'étude. Ces méthodes sont :

1-**Méthode des témoins historiques** ou approche historique : cette méthode repose sur l'idée que les malades sont comparables d'une époque à une autre. En fait, l'amélioration

des conditions diagnostiques et les modifications des contextes hospitaliers aboutissent à la constitution de groupes non comparables.

2- Comparaison des malades de deux hôpitaux ou approche géographique : les malades peuvent différer de façon considérable d'un hôpital à un autre.

3- Observation simple ou approche systématique : les groupes comparés risquent fort de n'être pas comparables, la diversité thérapeutique traduit une diversité initiale même en présence d'un dossier médical informatisé. Cette approche doit faire l'objet de plus de méfiance.

4- Cohorte de comparaison constituée par les sujets refusant la manœuvre : le refus peut être dicté par des raisons sociales, psychologiques ou médicales liées au traitement.

5- Répartition systématique : la distribution des malades dans chaque cohorte est réalisée par un moyen alphabétique, par la date de naissance… Cette technique peut être plus compliquée que le tirage au sort, n'offre aucune garantie éthique et cache des biais.

6-Techniques adaptatives : la distribution des sujets dans l'une ou l'autre des deux cohortes est fonction des résultats par l'accumulation des cas. Le plus grand nombre possible de malades est affecté au traitement le plus efficace en

administrant toujours le même traitement tant qu'un échec ne survient pas (play

the winner ». La qualité éthique est en réalité plus apparente que réelle.

III. Echantillonnage dans une expérimentation

- Le choix de la population dépend de la nature des interventions qui peuvent être des méthodes de prévention primaire (empêcher la survenue de la maladie) ou des méthodes de prévention secondaire (traitement des maladies) et tertiaire (réduction des séquelles).
- Les populations peuvent selon le cas être définies sur des bases administratives,

professionnelles ou recrutées dans des organismes de soins.

- Il est en effet parfois plus commode que la cohorte de la manœuvre principale et celle de la manœuvre comparative ne soient pas composées d'individus mais de groupes entiers. Mais il faut évidemment s'assurer que les deux cohortes ne diffèrent pas par des caractéristiques inégalement réparties d'une cohorte à l'autre.
- Par ailleurs, les essais peuvent porter sur des individus ou des groupes d'individus appelés unités collectives.

- On peut, par exemple, attribuer la population d'une école (ou d'un quartier) à une cohorte et celle d'une autre école (ou d'un autre quartier) à une autre cohorte en vue de la réalisation d'un essai sur un vaccin.
- Il est plus sûr cependant de disposer de plusieurs paires de groupes et d'attribuer les groupes de chaque paire au hasard à la cohorte principale et à la cohorte comparative. Le choix d'unités collectives est obligatoire lorsque :

• Les interventions sont appliquées de façon collective (effet sur la fréquence des caries dentaires par la fluoration de l'eau, campagnes d'éducation sanitaire réalisées par la voie des media…) ;

IV. Modalités pratiques de tirage au sort

- Le tirage au sort doit être effectué le plus tard possible après l'inclusion du malade dans l'essai juste avant de commencer le traitement. Le système de randomisation doit être disponible au moment où se présente un malade.
- Il s'agit de constituer à partir d'un groupe donné deux cohortes soumises à deux manœuvres différentes si on veut, par exemple, comparer un traitement A à un traitement B.
- Le système le plus simple est celui des enveloppes cachetées classées dans un ordre chronologique (malade n°1, malade n°2, etc.) dans lesquelles sont contenues les exigences du traitement A ou du traitement B.
- Ces exigences sont préalablement randomisées grâce à l'utilisation d'une

table de **nombres au hasard**.

- Pour effectuer la randomisation, deux techniques sont possibles : tirage au sort simple et raffinement.

Le tirage au sort simple:

- il consiste à utiliser une seule colonne de la liste des nombres au hasard (table statistique 4). Les chiffres 0, 1, 2, 3, 4 peuvent être attribués aux sujets de la cohorte A et les chiffres 5, 6, 7,8, 9 sont attribués aux sujets de la cohorte B.
- Le premier chiffre est désigné au hasard dans la colonne. S'il s'agit d'un 8 par exemple, la première enveloppe doit contenir les exigences relatives au traitement B.
- Si le chiffre suivant est 2, l'enveloppe correspondante doit contenir les instructions du traitement A et ainsi de suite.
- On peut généraliser aisément à plus de deux traitements.

- Pour trois traitements A, B, C, on peut établir la correspondance suivante : 0, 3, 6 sont attribués à A ; 1, 4, 7 peuvent désigner B et 2, 5, 8 sont attribués à C. Le chiffre 9 est tout simplement ignoré.
- Si le rapport de randomisation est de 2/1 en faveur d'un traitement, les chiffres 0, 1, 2, 3, 4, 5 peuvent désigner A tandis que les chiffres 6, 7, 8 sont attribués à B (le chiffre 9 est ignoré).
- Le tirage au sort simple aboutit à des effectifs à peu près égaux, mais

des inégalités restent possibles surtout si les effectifs sont faibles.

Le raffinement (tirage au sort équilibré):

- Il permet d'assurer l'égalité numérique des cohortes subissant chacune une des manœuvres de l'essai.
- Le raffinement sur le groupe total et le raffinement par séries sont les techniques utilisées pour réaliser l'équilibrage des cohortes.
- Le raffinement sur le groupe total est effectué si l'essai est conduit dans un seul centre et si l'effectif total des sujets éligibles est déterminé d'emblée. Pour répartir de façon équilibrée un groupe de 50 patients (numérotés de 00 à 49) en deux cohortes comprenant chacune 25 sujets, on utilise deux colonnes de la table des nombres randomisés.
- Les 25 premiers chiffres seront associés, par exemple, au traitement A tandis que les 25 suivants désigneront le traitement B (les chiffres 50 à 99 sont ignorés).
- **Le raffinement par séries:**
- est la technique usuelle, il est toujours effectué dans les essais randomisés multicentriques impliquant plusieurs centres.
- Pour deux traitements, l'équilibrage est réalisé par séries de 4, 6, 8 ou 10 patients. Le nombre usuel de malades dans une série est de 6.
- A cette fin, on utilise une table de permutations au hasard à 6 éléments

(en annexe de ce chapitre).

- On lit au hasard (en fermant les yeux et en pointant un crayon) une première permutation, suite de six chiffres de 1 à 6.
- Cette suite peut être, par exemple, 246315. Si les chiffres 1, 2, 3 sont associés au traitement A et les chiffres 4, 5, 6 au traitement B, on obtient la séquence ABBAAB qui détermine le contenu des six premières enveloppes de randomisation.

- Les permutations suivantes de la table, lues dans un sens quelconque, déterminent la suite du tirage au sort.
- Dans un essai comparant trois traitements, on peut désigner le

traitement A par les chiffres 1 et 2, le traitement B par les chiffres 3 et 4 et le traitement C par les chiffres 5 et 6.

- Si le rapport de randomisation est 2/1, le traitement A peut être associé aux chiffres 1, 2, 3, 4 et le traitement B aux chiffres 5 et 6.
- Si l'essai est multicentrique, le plus simple est d'établir une liste pour un centre et de reprendre la même liste pour les autres centres en commençant par une autre suite, quitte à revenir au début si la liste est insuffisante.

Elaboration d'un protocole d'essai randomisé

- Un essai randomisé obéit aux mêmes règles qui régissent la réalisation de toute étude épidémiologique.
- C'est ainsi que le protocole guide l'exécution de l'essai et assure à tous les participants une exécution identique de ses différentes phases.
- Le protocole d'un essai comporte obligatoirement :

• la définition de la maladie,

• la définition des malades,

• la définition des manœuvres,

• le choix et la définition des cibles.

Toutes les précautions relatives à la définition de la maladie, des malades et des manœuvres ont pour but d'assurer la comparabilité et l'homogénéité des cohortes tout au long de l'essai.

- Il convient, en matière de cibles, de retenir le moins possible de critères et seulement ceux utilisables avec profit au moment de l'analyse.
- Un critère très utilisé est le délai d'apparition d'un événement.
- Dans tous les cas, les critères retenus doivent tendre vers l'objectivité (examen bactériologique…) mais certains critères restent essentiellement subjectifs (appréciation de l'intensité d'une douleur...).
- Les techniques à l'aveugle sont susceptibles de réduire cette subjectivité.
- Le résultat du travail d'élaboration du protocole se concrétise par trois documents :

• le protocole proprement dit ;

• le questionnaire d'enregistrement avec cinq grandes rubriques :

o renseignements généraux et identification ;

o description du malade et de la maladie ;

o facteurs éventuellement pronostiques ;

o description des traitements reçus ;

o cibles.

6. Analyse des résultats

- L'analyse d'un essai randomisé est la comparaison des résultats, obtenus dans les cohortes, des critères de jugement retenus dans le protocole (comparaison de pourcentages, de densités d'incidence, de moyennes).
- Une étape obligatoire dans cette analyse est la description des sujets inclus :

• description des caractéristiques des sujets ;

• description des écarts au protocole.

- Les caractéristiques des sujets inclus, les traitements administrés, les modalités de détection des cibles doivent être en principe l'exact reflet de ce qui est exigé dans le protocole.
- L'existence d'écarts au protocole est cependant inévitable en pratique sujets inclus à tort, écarts au protocole de traitement, perdus de vue (biais).
- Les écarts au protocole concernent les sujets qui n'ont pas respecté les conditions d'administration de la manœuvre telles qu'elles ont été définies dans le protocole. Si les écarts au protocole dépendent du traitement, un biais d'information survient.
- Dans la plupart des cas, les écarts au protocole de traitement ne sont pas dus à des raisons identiques dans les deux cohortes.
- Si l'un des deux traitements entraîne des effets indésirables sévères, les deux groupes seront ainsi déséquilibrés étant donné que les malades qui ont reçu un tel traitement seront plus nombreux à l'abandonner par rapport à ceux qui ont reçu l'autre traitement.
- Dans de telles conditions, l'exclusion de la comparaison finale de malades dont la manœuvre s'écarte trop de celle qui leur a été attribuée est une attitude erronée et illicite. D'une cohorte à une autre, les malades qui s'écartent de leur traitement

peuvent avoir une attitude si différente vis-à-vis de l'atteinte des cibles qu'une comparaison effectuée sur les seuls malades restants sera lourdement biaisée.

- L'analyse des résultats doit toujours se faire en intention de traiter, avec tous les malades déclarés bons pour l'essai et inclus dans l'essai tels qu'ils ont été affectés par tirage au sort. L'analyse par traitement réellement reçu (analyse per-protocole) doit toujours être prudente sinon rejetée si ses résultats ne coïncident pas avec ceux de l'analyse en intention de traiter.
- L'analyse en intention de traiter n'augmente pas le risque. Par contre, le risque augmente et engendre une perte de puissance de l'étude. La probabilité de ne pas mettre en évidence une différence qui existe réellement augmente évidemment lorsque les sujets changent de groupe.
- Dans certains cas, le nombre de ces écarts a une importance telle qu'il conduira à mettre en doute l'applicabilité du protocole et la possibilité d'une analyse à moins de reformuler les objectifs de manière pertinente.
- Les perdus de vue sont les sujets pour lesquels le résultat de la cible n'a pu être obtenu. Les perdus de vue ne pouvant être pris en compte dans l'analyse, le test statistique est moins puissant que celui initialement prévu et un biais de répartition peut survenir.
- Dans les essais où la cible est le délai d'apparition d'un événement, les perdus de vue sont pris en compte jusqu'au moment de leur disparition grâce aux méthodes de survie.
- L'analyse d'un essai est effectuée le plus souvent lorsqu'on dispose du résultat de la cible pour le nombre de sujets nécessaire déterminé dans le protocole (analyse en fin d'essai).
- La comparaison doit :

• porter sur les critères de jugement définis dans le protocole ;

• comparer les « bons groupes », c'est-à-dire sans exclusion illicite ;

• déterminer les mesures d'association statistique et épidémiologiques appropriées.

- Une analyse intermédiaire est toute analyse effectuée avant le recueil du résultat de la cible pour le nombre de sujets nécessaire déterminé dans le protocole.
- Il est d'usage d'apprécier la comparabilité des groupes de l'essai et le maintien de cette comparabilité tout au long de l'essai (depuis l'inclusion des patients jusqu'à

l'analyse en passant par la randomisation, l'attribution des manœuvres et le suivi) par un organigramme appelé diagramme de flux (flowchart).

- La figure suivante est le diagramme de la grille Consort fréquemment utilisé.

Ethique

- Les problèmes éthiques, en matière d'essais randomisés, dominent
- tous les autres. Les difficultés éthiques dépendent :

• de la gravité de la maladie ;

• du type de traitement proposé ;

• de la spécificité du traitement vis-à-vis de la maladie.

- Le problème éthique est minimum si la maladie est bénigne, s'il existe un traitement classique de cette maladie et si l'on veut tester les effets symptomatiques de ce traitement.
- Le problème éthique est maximum si la maladie est grave, s'il n'existe jusqu'alors aucun traitement efficace et si l'on cherche à mettre en évidence un traitement curatif.
- Des conventions internationales fixent des règles strictes en ce qui concerne les recherches biomédicales portant sur l'être humain qui sont plus ou moins respectées.
- En santé communautaire, des populations considérables peuvent être impliquées. Ce qui, du point de vue éthique, amplifie chaque erreur.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

Mesures d'association épidémiologique : Concept de risque et de cote dans les études cas-témoins (OR) cohorte (RR)et transversale (OR, RR), Mesures d'impact ,intervalle de confiance

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Mesures d'Association épidémiologique

Interprétation de l'intervalle de confiance de RR et OR à 95%

I-Introduction

L'étude de l'association entre un facteur d'exposition et une maladie est l'une des étapes majeures de la recherche des facteurs étiologiques des maladies.

* Pour identifier un facteur d'exposition, il faut :

- mettre en évidence une liaison entre ce facteur et la maladie avec un test statistique approprié (le test du X2,…) ;
- mesurer la force de cette liaison.

II- Notion de risque et mesures d'association :

Notion de risque : Elle est définie par la probabilité de survenue de la maladie ou tout autres événements de santé (complication, décès…) dans une population à un moment donné ou pendant un intervalle de temps donné suite à l'**exposition** à des facteurs de risque.

Facteur de risque : désigne la source de risque qui augmente la probabilité pour un individu de développer une maladie, il peut être soit :

Individuel : biologique(hérédité), comportement (alimentation, habitudes toxique).
Communautaire : Hygiène de milieu.

III- Mesures d'association dans les études épidémiologiques :

- Afin d'établir clairement les mesures d'association entre la maladie et les facteurs de risque, on doit réaliser les études analytiques sont ont pour objectif de comparer des malades et des non malades selon leur niveau d'exposition à un ou plusieurs facteurs de risque.
- On distingue selon la chronologie du recueil des données :

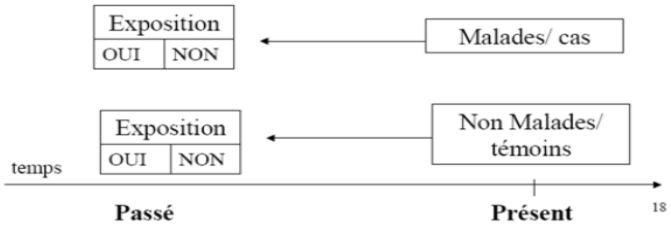
✓ Etude cas témoins
✓ Etude Cohorte

A-Etude Cas Témoins :

1-Principe :

Elle consiste à comparer la fréquence d'exposition au facteur parmi un groupe de malade et parmi un groupe de sujets non malades choisis comme témoins.
Si la survenue de la maladie est liée à l'exposition, on doit observer un pourcentage d'exposition plus élevé chez les cas que chez les témoins.

Schéma d'étude :



3-Presentation des données et mesures d'association :

A- Fréquence d'exposition

	cas	témoins	Total
Exposé	**a**	**b**	**a+b**
Non exposé	**c**	**d**	**c+d**
total	**a+c**	**b+d**	**a+b+c+d**

-Fréquence d'exposition chez les cas (EXP_c): $$EXPc = \frac{a}{a+c}$$

-Fréquence d'exposition chez les témoins (Exp_t): $$EXPt = \frac{b}{b+d}$$

Ces cotes en tant que telles ne servent à rien, mais leur rapport permet d'évaluer la liaison entre l'exposition et la maladie, ce rapport est appelé Rapport des Cotes (RC) ou Odds Ratio (OR).

B - Les cotes d'exposition :

*La cote d'exposition chez les cas : a /c

*La cote d'exposition chez les témoins : b/d

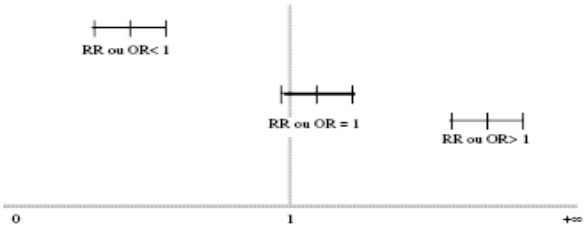
c-L' Odds ratios (OR) :

L'OR est la mesure la plus étulisée dans une étude cas témoins, l'OR est le rapport de la cote d'exposition chez les cas sur la cote d'exposition chez les témoins :

$$OR = \frac{a/c}{b/d}$$

4-Interprétation de l'OR :

L'OR est un nombre sans unité compris entre 0 et l'infini



OR = 1 : absence de risque, Intervalle de confiance à 95% inclut la valeur 1 : risque non mis en évidence (non significatif)

OR > 1, le facteur étudié est un facteur de risque

OR < 1 et le facteur étudié est un facteur protecteur

Exemple :

OR = 4, signifie que les sujets exposés ont un risque multiplié par 4 de contracter la maladie par rapport aux sujets non explosés.

Intervalle de Confiance De l'OR :

-Par définition les Cas et les témoins sont issus d'une population.

-L'OR est donc une variable aléatoire qui subit des fluctuations d'échantillonnage.
-on calcule donc un intervalle d confiance (IC)à 95% de RR

$$IC\ à\ 95\% = RR^{1 \pm [(1,96 / \sqrt{\chi^2})]}$$

Si l'IC inclut la valeur 1ca signifie l'absence d'association entre le facteur et la maladie.

Comparaison OR et X^2 :

L'intervalle de confiance à la même signification que le résultat de p(degré de signification)

du test X^2 :

*si p est supérieur à **0,05** (test non significatif), alors, l'IC à95% du RR contient la valeur 1, le

test n'est pas significatif, le facteur étudié n'est pas un facteur de risque.

*si p est inférieur à **0.05**, alors l'IC à 95% exclut la valeur1, et le facteur étudié est un facteur

de risque.
Donc l'intérêt du calcul de l'OR est de donner la force le sens et le dégrée de signification de l'association lors le test de chi deux ne donne que le dégrée de signification de l'association.

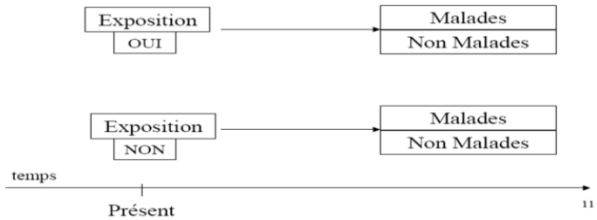
B-Etude de Cohorte :

1-Principe :

Une cohorte se définit comme un groupe de sujets suivis dans le temps, si l'événement observé est la survenue d'une maladie, on mesure, à la fin de l'étude, le nombre de sujets atteints par la maladie pendant la période d'étude.

Une étude de cohorte est parfois appelée étude exposés/non exposés.

2-Schémas de l'étude:



Une étude de Cohorte est une étude comparant une cohorte de sujets exposés à une cohorte de sujets non exposés. A l'issue d'une étude de cohorte, on compare donc le taux d'incidence entre exposés **(Ie)** et non exposés**(Ine).**

3-Présentation des données et Mesures d'association

a-Des taux d'incidences dans chaque groupe de comparaison. Ces taux d'incidence sont assimilés aux probabilités ou risque de survenue de la maladie

	malade	Non malade	Total
Exposé	**a**	**b**	**a+b**
Non exposé	**c**	**d**	**C+d**
total	**a+c**	**b+d**	**a+b+c+d**

Ie : incidence de la maladie chez les sujets exposés : $Ie = \frac{a}{a+b}$

Ine : incidence de la maladie chez les sujets non exposés $Ine = \frac{c}{c+d}$

b-La différence de risque: entre exposés et non exposés :

***C'est l'excès de maladie dû à la présence du facteur de risque.**

*c'est le risque de la maladie attribuable au facteur de risque chez les exposés.

DR=Ie-Ine

*Il représente la partie du risque exclusivement liée au facteur étudié**.**

C- Le risque relatif (RR) :

Le RR est le rapport entre l'incidence chez les exposés surs

L'incidence chez les non exposés, cet indicateur est appelé **encore le ratio de risque** ou

rapport de risque (RR). **RR= Ie/Ine**

-Il est utilisé très fréquemment en recherche étiologique.

-Il exprime la force de l'association entre un facteur et la maladie.

Interprétation du risque relatif(RR) :
***un risque nul n'a pour valeur 1 :** signifie l'absence d'association entre le facteur et la maladie.
***un risque inférieur à 1 :** signifie que le facteur étudie est un facteur protecteur
***un risque plus de 1 :** signifie que le facteur étudié est un facteur de risque, et plus le RR est éloigné de 1 plus l'association entre la survenue de la maladie et le facteur est forte.

Intervalle de Confiance du risque relatif :

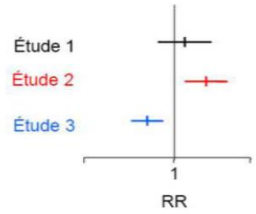
-Une étude de cohorte est rarement réalisée sur l'ensemble d'une population à risque d'une maladie donnée.

-Elle est effectuée sur un échantillon représentatif de cette population.

- Le RR est donc une variable aléatoire qui subit des fluctuations d'échantillonnage.

-on calcule donc un intervalle d confiance à 95% de RR. $IC\ à\ 95\% = RR^{1\pm(1,96/\sqrt{\chi^2})}$

Exemples :



	Risque Relatif RR (95% CI)	
Étude 1	1.80 (0.90,3.60)	p > 0.05
Étude 2	3.30 (1.80,5.90)	p < 0.05
Étude 3	0.30 (0.20,0.60)	p < 0.05

L'intérêt de calculer le **RR** et de son **intervalle de confiance** est de donner **la force, le sens et le degré de signification de l'association.**

Comparaison RR et X^2 :

On pourrait comparer les taux d'incidence dans les deux groupes : exposés et non exposés, pa un simple test de X^2, ce test permettra d'affirmer qu'il existe une différence significative entre les incidences observées :

*si p est supérieur à **0,05** (test non significatif), alors, l'IC à95% du RR contient a valeur 1, le test n'est pas significatif, le facteur étudié n'est pas un facteur de risque.

*si p est inférieur à **0.05**, alors l'IC à 95% exclut la valeur1, et le facteur étudié est un facteur de risque.

Conclusion :

La notion de risque est très importante en Epidémiologie.

*Elle se base sur les enquêtes épidémiologiques analytiques.

* L'enquête de cohorte reste la plus précise

*En pratique courante les enquêtes cas témoins et les enquêtes transversales sont les plus utilisées.

*La réduction des problèmes de santé repose sur la prévention et nécessite donc une connaissance rigoureuse des risques encourus d'où l'intérêt des études épidémiologiques.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

Biais dans les études épidémiologiques : Définition, détection, contrôle des biais de sélection, biais d'information, biais de confusion, effet de modification, principe de raisonnement causal

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Biais dans les études épidémiologiques : Définition, détection, contrôle des biais de sélection, biais d'information, biais de confusion, effet de modification, principe de raisonnement causal

1-DEFINITION DES BIAIS :

-Distorsion de l'estimation de la mesure d'une association entre l'exposition (E) à un facteur de Risque (R) et la survenue de la maladie (M).

-C'est une erreur systématique qui contribue à produire des estimations plus élevées ou plus basses que la valeur réelle des paramètres à estimer dans une enquête épidémiologique

Conclusions erronées.

L'association entre une facteur de risque et une maladie peut être soit due à :

- Hasard – Fluctuation d'échantillonnage
- Biais

2-FAMILLES DES BIAIS :

En épidémiologie, il existe de nombreuses sources de biais que l'on peut regrouper en 3 :

❶ Biais de sélection

❷ Biais de mesure, d'information, de classification

❸ Biais de confusion

2 grandes catégories de biais :

Peuvent être évités au moment de la conception d'un protocole, et de prendre en compte en cours de l'analyse	***Biais de confusion***
Peuvent pas être éviter lors de la conception du protocole, mais contrôlé au moment de l'analyse	***Biais de sélection et biais d'information***

A- BIAIS DE SELECTION OU D'ECHANTILLONAGE

Biais dans la constitution de l'échantillon, qui va se retrouver non représentatif de la population cible pour des facteurs liés au problème étudié.

- Ce type de biais réfère à une distorsion dans l'estimation d'un effet résultant de la façon par laquelle les sujets de la population étudiée ont été sélectionnés.
- Ce sont des erreurs induites par une sélection préférentielle des sujets à comparer qui n'est pas indépendante de leur statut Malade/Témoin dans une étude Cas/Témoins et de leur statut Exposé/Non Exposé dans une étude de Cohorte.

PROCESSUS :

- ❑ **Erreur sur le choix des groupes à comparer dans tous les types d'enquêtes épidémiologiques**

- Lors de la constitution de l'échantillon : échantillon non représentatif de la population source ou de la population à laquelle on veut extrapoler les résultats.

Exemple : sujets âgés de 70 ans ou plus sans comorbidité ne sont pas représentatifs de la population des sujets âgés.

- Au niveau des groupes analysés : les groupes analysés ne diffèrent pas seulement par le facteur étudié mais aussi par un autre facteur (facteur pronostique) qui peut modifier les résultats.
- Exemple : essai contrôlé randomisé avec 10% de sorties d'étude liées à des effets indésirables dans un groupe de traitement => résultats d'efficacité biaisés.

1-Biais de perdus de vue (de migration) :

- Les sujets perdus de vue dans les études de cohortes et les sujets non répondants, répétées au fil du temps
- Biais survenant quand les motifs des pertes de vue sont liés au résultat.

Exemple :

-Les enquêtes portant un nombre élevé de nos réponses : sujets impossible à recenser ou à contacter ou qui refusent de répondre.
-Les perdus de vue : sujets ayant déménagé sujets jeunes, et donc moins stables...

2- Biais de recrutement :

- Si probabilité d'inclusion dans l'étude est liée au facteur étudié.

Surtout si les sujets sont recrutés dans des institutions de soin.

- **Exemple :**

La sélection de diabétiques dans un service de diabétologie en CHU.

3- Biais du Healthy Worker Effect (effet du travailleur en bonne santé)

- En milieu de travail, les individus sont capables d'exercer un travail régulier et sont a priori en meilleure santé que la population générale dans laquelle on trouve des malades dans l'incapacité de tenir un emploi.
- La santé des travailleurs n'est pas comparable avec celle de la population générale.
- La population active a un meilleur état de santé (examen d'embauche), que celui de la population générale
- Cet effet est d'autant plus marqué que les conditions de travail sont difficiles et requièrent des capacités particulières (travaux physiquement pénibles, horaires de nuit/ou alternants, etc.).

4- Biais de BERKSON ou biais d'Admission

Les enquêtes analytiques sur les malades hospitalisés :

L'admission des malades à l'hôpital dépend de plusieurs paramètres liés à la renommée du service hospitalier, au malade lui-même (lieu de résidence)...
Les malades hospitalisés ne sont donc pas représentatifs de la population générale.

Pour les études qui se déroulent en milieu hospitalier : Probabilité donc différente d'être admis à l'hôpital chez les cas et les témoins.

Les cas exposés sont plus facilement admis à l'hôpital où se déroule l'étude, que les non exposés. (lié au fait que les témoins soient aussi des patients hospitalisés).

Les témoins peuvent ne pas provenir de la même population générale.

Contrôle des biais :

Avant l'étude :

1-Recrutement par tirage au sort : concerne surtout les enquêtes d'observation.

Exemple : enquête descriptive avec sondage aléatoire; enquête cas-témoin avec sélection par tirage au sort des témoins dans population d'où sont issus les cas.

2-Recrutement exhaustif :

Exemple : enquête cas-témoin : recrutement systématique de tous les cas définis par des critères diagnostiques valides et appliqués à tous les sujets. Enquête de cohorte, essai clinique, évaluation procédure diagnostique , recrutement systématique de tous les patients éligibles sur une période donnée.

3-Randomisation du facteur étudié : si l'objectif de l'étude le permet. S'assurer que la randomisation est bien faite.

4-Suivi complet et/ou relances : concerne toutes les études. Pour éviter les perdus de vue dans études prospectives et les données manquantes dans toutes les études.

5-Définition stricte des critères d'inclusion et d'exclusion.

Après l'étude : (limite le biais, mais ne le neutralise pas complètement)

Comparaison initiale des groupes :les facteurs pronostiques doivent être similaires (comparabilité clinique et non statistique). Sinon, ajustement sur les facteurs pronostiques déséquilibrés entre les groupes.

Analyse en ITT: Elle n'élimine pas le biais si le suivi des patients est incomplet et si les motifs de sortie d'étude ou de perdus de vue sont liés aux résultats.

Analyse en intention de dépister: Idem

Ajustement sur facteurs: Il s'agit des facteurs pronostiques déséquilibrés, des facteurs sur lesquels la randomisation a été stratifiée (centre …) et les facteurs d'appariement (enquêtes cas-témoins, enquêtes de cohorte).

Remarque : si les groupes analysés sont très différents, un ajustement ne suffit pas à neutraliser le biais car on ne peut pas ajuster sur tous les facteurs pronostiques (tous ne sont pas recueillis).

Analyse de sensibilité : Elle évalue l'influence sur les résultats, d'un changement de méthode, variable, critère de jugement. Elle permet de déterminer le sens du biais et de le quantifier dans la mesure du possible.

Exemple : analyse de la robustesse des résultats si les perdus de vue sont liés au facteur étudié ou à la réponse.

B- Biais d'information :

Définition :

Biais dans la mesure du facteur de risque ou dans la certitude de la maladie.

Cette erreur est quasi inévitable puisque aucun outil de mesure (interrogatoire, examen, test n'est parfait. ↺ Instrument d'observation mesure défectueux.:

Ils sont provoqués au moment où l'on a recueilli l'information concernant l'**exposition** chez les cas et les témoins ou la **maladie** chez les exposés et les non exposés
Ce biais peut toucher aussi bien le facteur étudié (par exemple l'exposition dans une enquête **de cohorte** que **les critères de jugement dans un essai** (exemple : score d'un questionnaire, décès pour cause spécifique).
Etude Cas/témoins information sur l'exposition des cas et/ou des témoins est inexacte car recueillie différemment dans les 2 groupes : surestimer la relation si les cas se souviennent mieux de leurs antécédents ou le contraire.

Deux types d'erreurs de classement :

Différentiel :(erreur dépendante de la maladie)

Si les erreurs sur les informations recueillies affectent différemment les groupes soumis à comparaison (Erreur différente chez les malades/non malades ou exposés/non-exposés, ce qui
• sur- estiment l'exposition ⇒ OR ou RR sur -estimé
• sous-estiment l'exposition ⇒ OR ou RR sous-estimé
Biaise l'étude et le sens du biais imprévisible

- ✓ Pour éviter ce biais : mesure de l'exposition à l'aveugle de la maladie (et vice et versa)

Exemple : Enquête entre « trouble du sommeil de l'enfant entre 0-6 mois » et « mort subite de l'enfant » (MSE) , les parents traumatisés par la MSE vont plus souvent surestimer l'exposition de l'enfant décédé
Non-différentiel : erreur indépendante de la maladie

Si l'erreur systématique de classification est identique chez les sujets malades et non malades (ou les exposés et les non exposés),

Sens du biais prévisible : toujours diminution de l'association entre le facteur étudié et l a maladie Diminution de la puissance.

– Erreur aléatoire (« bruit »)
– La force de l'association est sous-estimée
⇒ Les OR ou RR se « rapprochent » de 1

Exemple : Enquête entre tabagisme passif (exposition) et troubles lipidiques (maladie) du sujet non-fumeur
Auto -évaluation du temps passé avec des fumeurs
⇒ Erreur probable sur mesure
Processus :
Il peut résulter de

La subjectivité de l'enquêteur : variabilité.

La subjectivité des enquêtés : mémorisation, refus de répondre, déni.

La méthode de mesure : rythme de suivi différent, temps de mesure différent, outil non valide.

1- Biais d'information/mesure :

Définition :

Biais survenant quand le protocole de mesure n'est pas standardisé (le même pour tous) et/ou est sujet à interprétation.

Exemples :

- Une enquête basée sur la mesure d'un paramètre par un équipement non standardisé (cas des enquêtes biométriques)
- questionnaire de qualité de vie non validé.

- **biais de classification des cas :**

Exemple : Les biais liés à la validité des tests biologiques.

- **Les biais de mémorisation (« recall bias ») :**

Définition : biais survenant lors du rappel de l'exposition ou du critère de jugement dans la vie passée. Exemple : enquête cas-témoin : les cas peuvent faire plus d'effort pour se souvenir d'une exposition passée que les témoins.

Cas et témoins se souviennent avec une acuité différente de leur exposition au facteur de risque.

Exemple : Les enquêtes de morbidité portant sur d'anciens épisodes de maladies.

Dans les enquêtes cas-témoins, un malade se souvient plus d'une exposition à un facteur de risque qu'un témoin

Les biais de subjectivité :

Biais de subjectivité de l'enquêteur : réponses suggérée ou interprétées.

Exemples :

- Connaissance du statut du malade par l'enquêteur.
- Un enquêteur peut suggérer des réponses (enquêtes par questionnaire, ou dans les essais thérapeutiques)

Biais de prévarication :

omission volontaire ou mensonge. Ex : toxicomanie

Contrôle des biais :

Pour les critères subjectifs, il faut une référence dans le texte à une étude de validation ; à défaut, des résultats de reproductibilité. Exemple : questionnaire de qualité de vie (subjectif) doit avoir été validé. Vérifier qu'il y a une référence dans le texte à propos d'une étude de validation.

Standardisation des procédures. Les mêmes procédures doivent être appliquées à tous les sujets de l'étude : même procédé de recueil de mesure, même rythme de mesure (si prospectif), même définition du facteur étudié et des critères de jugement pour tous les sujets. Exemples : mêmes critères diagnostiques appliqués à tous les sujets dans une enquête cas-témoin, à chaque fois que c'est possible même contenu et rythme de suivi dans les enquêtes exposés/non exposés

Evaluation/recueil des données en aveugle/en insu.

Essais cliniques : évaluation des critères de jugement en insu du traitement.

Résultat d'un examen diagnostique en insu du résultat du gold standard et vice versa.

Enquête cas-témoin : évaluation de l'exposition en insu du statut malade/non malade du sujet.

Enquête de cohorte exposée/non exposé : évaluation de la maladie en insu de l'exposition.

Formation des enquêteurs pour avoir une standardisation des procédures.

C-BIAIS DE CONFUSION :

Ce sont des biais systématiques qui s'introduisent dans l'analyse et l'interprétation des résultats d'une enquête.

Position du problème :

Distorsion de la mesure de l'association entre une maladie **M** et un facteur d'Exposition **E**, introduite par un 3ème variable appelée **FACTEUR de CONFUSION**

● Un facteur de confusion est un facteur associé aussi bien à la maladie étudiée qu'au facteur de risque recherché.

● Un biais de confusion perturbe la relation entre un facteur d'exposition E et une maladie M .

La relation entre le facteur de confusion et un facteur de risque :

Illustration du type d'association qui lie un facteur de confusion F à la maladie M, et au facteur de risque E (déterminant de confusion)

Exemples :

⏩Dans les enquêtes portant sur la relation entre un risque professionnel et le cancer du poumon, le tabagisme est un facteur de confusion.

L'âge des sujets enquêtés peut être aussi un facteur de confusion

⏩ Dans les enquêtes portant sur l'association entre l'usage des contraceptifs et l'infarctus du myocarde, l'âge est un facteur de confusion, les utilisatrices de contraceptifs sont généralement plus jeunes.

Correction des biais de confusion :

Plusieurs méthodes permettent de les prendre en compte soit lors de l'élaboration du protocole, soit au moment de l'analyse :

A.-STRATIFICATION :

Dans un échantillon de malade et de témoins, elle consiste à former des classes de sujets par rapport au facteur de confusion.

Exemple : *Relation « tabac » et Cancer Bronchique, on classera les sujets par tranche d'âge et par sexe*.

B.-APPARIEMENT :

Consiste à neutraliser le facteur de confusion en groupant les sujets de telle sorte que ceux d'un même groupe partage le même facteur de confusion.

Chaque cas sera apparié avec un ou plusieurs témoins de même âge et de même sexe.

Une modalité particulière de la stratification

C.- AJUSTEMENT ET STANDARDISATION :

Consiste à éliminer d'une comparaison de série d'observations le lien entre un effet et une ou plusieurs causes autres que celles qui sont le sujet propre de l'étude

C.- AUTRES :

La sélection au hasard des groupes à étudier se restreindre à étudier les catégories d'âge ou de sexe dans les enquêtes cas-témoins susceptibles de porter des variables de confusion.

CONCLUSION :

La prévention des biais ou leur recherche dans l'analyse constituent des étapes importantes dans une enquête afin de juger de la causalité entre un ou plusieurs facteurs.

Pour identifier des biais potentiels, on doit se poser des questions :

- Si la la population de l'étude est bien identifiée.
- Si la population étudiée représente de manière adaptée la population cible.
- Si les définitions des maladies et des expositions sont bien claires.
- La définition des cas
- Les critères d'inclusion et d'exclusion
- Si l'identification ou la sélection des cas ou des témoins a pu influencer le statut d'exposition
- Si les mesures sont objectives
- Si le suivi est identique pour les cas et les témoins
- Si l'analyse est appropriée

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

Variables de confusion et d'interactions

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Variables de confusion et d'interactions

Objectifs du cours :

- Savoir définir et reconnaître un facteur confondant
- Savoir définir et reconnaître un biais de confusion
- Savoir définir une modification d'effet
- Connaître les stratégies visant à minimiser les biais et facteurs confondants
- Connaître les différentes méthodes de correction des facteurs confondants

Plan du cours :

I-Introduction
II- Définition des erreurs systématiques et biais
1II- Le biais de confusion
IV- Management des facteurs confondants
V- Modification d'effet (effet modification)

I-Introduction :

Toutes les études épidémiologiques sont susceptibles d'inclure des erreurs car elles sont basées sur des mesures qui ne sont jamais parfaites. Il existe 2 types d'erreurs pouvant affecter les conclusions d'une étude épidémiologique : Les erreurs aléatoires et les erreurs systématiques. Parmi les erreurs systématiques, le biais de confusion prédomine cet ensemble des biais

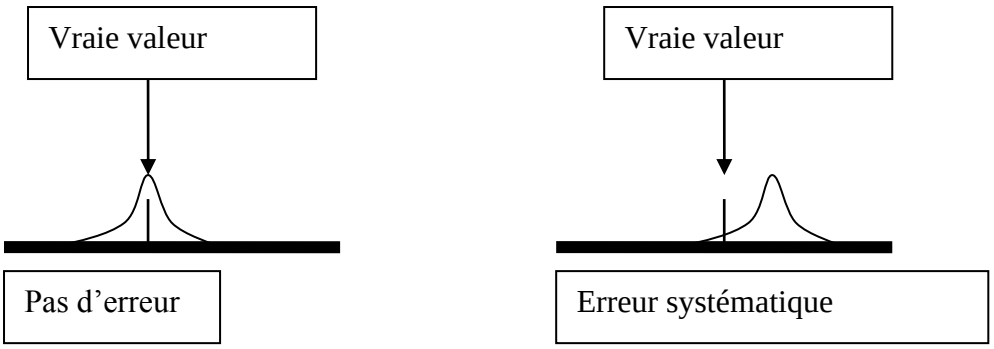
II- Définition des erreurs systématiques et biais :

Les **erreurs systématiques** sont des erreurs qui se produisent lorsque la probabilité de commettre une erreur est différente pour les 2 groupes à évaluer. Il s'agit donc d'erreurs qui se font toujours dans un sens (systématique). Ceci créé un déséquilibre menant à une conclusion biaisée.

On définit le **biais** par la présence d'erreurs systématiques dans une étude menant à une évaluation erronée de la relation entre l'exposition et la maladie. La présence de biais nuit à la validité des résultats. Contrairement au hasard, il est très difficile de quantifier un biais. Il faut donc tout faire pour minimiser les risques de biais dans l'élaboration de la méthodologie des études.

Il existe peu de liens entre la précision (reliée aux erreurs aléatoires) et les biais (relié aux erreurs systématiques) ; Une étude peut avoir des résultats très précis (reproductibles) mais qui ne reflètent pas la réalité (biaisé) et vice-versa.

Voici une représentation graphique de deux études ayant peu d'erreurs aléatoires. La première ne contient pas d'erreur systématique tandis que la deuxième en contient.



Les types de biais

On différencie trois grands types de biais :

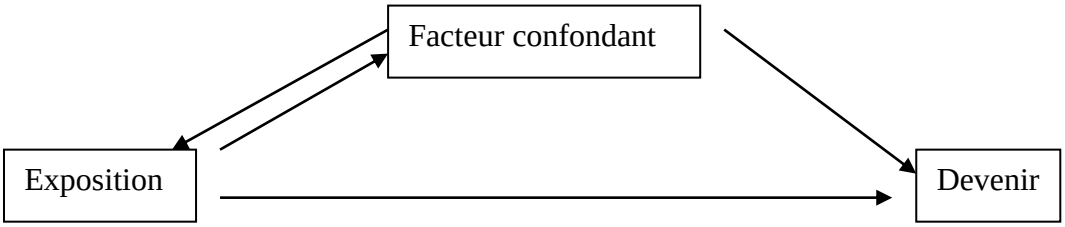
- Le biais de sélection qui est relié au mode de recrutement des patients dans l'étude
- Le biais d'observation est relié à l'évaluation du patient une fois qu'il est recruté
- Le biais d'analyse qui se caractérise par la présence d'un autre **facteur confondant** reliée à l'exposition et au devenir du patient. Ce type de biais sera discuté dans ce cours.

1II- Le biais de confusion :

Un facteur confondant est un facteur qui biaise les résultats de l'étude parce que son association avec l'exposition et la maladie fait faussement croire que l'exposition est associée à la maladie. Contrairement aux types de biais précédemment décrits, les facteurs confondants sont des facteurs conduisant à des résultats erronés que l'on peut quantifier. On peut aussi en tenir compte dans l'analyse afin d'en éliminer l'effet.

On définit un **facteur confondant** par la présence obligatoire de 3 critères :

- Le facteur confondant est associé à l'exposition
- Le facteur confondant est associé au devenir (outcome)
- Cette association se fait de façon indépendante de l'exposition : Le facteur confondant n'intervient pas directement dans la relation exposition-outcome.



Voici un exemple :
Dans une étude sur le taux d'accident de voiture et l'utilisation de téléphone cellulaire on constate une forte association.
Voici le tableau des résultats :

	Cellulaire +	Cellulaire -
Cas (accident)	35	90
Témoin (pas d'accident)	15	70

OR : (35/90) / (15/70) = 1.8

Un chercheur voit ces résultats et suggère qu'il pourrait y avoir un important facteur confondant : l'âge. Les jeunes de moins de 25 ans ont beaucoup plus d'accident de voiture et ils sont beaucoup plus fréquents à posséder un téléphone cellulaire. Voici 2 tableaux de la même étude selon que les participants aient < 25 ans ou > 25 ans.

Pour les moins de 25 ans

	Cellulaire +	Cellulaire -
Cas	30	60
Témoin	5	10

OR : (30/60) / (5/10) = 1.0 (pas d'association)

Pour les plus de 25 ans

	Cellulaire +	Cellulaire -
Cas	5	30
Témoin	10	60

OR : (5/30) / (10/60)= 1.0 (pas d'association)

On voit donc que l'utilisation n'est pas un facteur de risque d'accident pour les gens de moins de 25 ans ni pour ceux de plus de 25 ans !!!! Par contre, lorsqu'on les regroupe sans considérer l'âge, on obtient une association. Ceci est une belle démonstration d'un facteur confondant. On voit que l'âge est associé avec l'exposition (les plus jeunes ont plus de téléphones) et avec le outcome (les jeunes ont plus d'accident). De plus, l'âge n'intervient pas directement dans la relation cellulaire-accident.

D'autres exemples seraient :

- Relation calvitie- consommation de bière. Un facteur confondant pourrait être le sexe
- Relation # de téléviseur à la maison-cancer de la prostate. L'âge ou le niveau socio économique pourraient être des facteurs confondants.

Il existe 2 façons de démontrer la présence d'un facteur confondant:
1. On évalue l'association entre l'exposition et le devenir de façon séparée pour les diverses strates du présumé facteur confondant. Si les OR sont tous semblables à la valeur du OR total, il n'y a pas de facteur confondant. Si les OR sont semblables pour les diverses strates mais différentes du résultat total, il faut suspecter un facteur confondant. Ceci fut démontré dans les tableaux précédents
2. On peut aussi montrer qu'il existe une association entre le facteur confondant et l'exposition et une association entre le facteur confondant et le devenir (outcome). Voici les tableaux pour l'exemple précédent.

Tableau de l'association age et cellulaire :

	Cellulaire +	Cellulaire -
Âge < 25 ans	35	70
Âge > 25 ans	15	90

OR : (35/70) / (15/90) = 3.0

Tableau de l'association age et accident :

	Accident	Témoin
Âge < 25 ans	90	15
Âge > 25 ans	35	70

OR : (90/15) / (35/70) = 12.0

On voit donc une association entre l'âge et l'exposition (OR 3) et entre l'âge et le devenir (OR12). L'âge est donc un facteur confondant puisqu'il n'est pas impliqué directement dans la relation cellulaire-accident.

VI- Management des facteurs confondants

On peut prévenir les facteurs confondant à la phase de l'élaboration d'une étude ou en tenir compte lors de son analyse. Voici 3 méthodes de contrôle lors de l'élaboration:

1- La **restriction** : Pour confondre, un facteur doit être présent en fréquence variable dans les divers groupes (exposé vs non). Une façon de contrecarrer cette variabilité est de restreindre la participation à l'étude qu'à certains individus selon leur statut face au facteur confondant. On peut, par exemple, restreindre l'étude qu'aux hommes, qu'aux enfants d'un certain âge ou membre d'une seule communauté ethnique. La restriction est simple à appliquer mais elle diminue souvent de beaucoup le nombre de participants potentiels et elle ne permet pas de conclure pour les patients qui n'étaient pas dans les strates évaluées. On limite donc la généralabilité des résultats

- L'**appariement** : Dans ce type de procédé, les participants sont recrutés afin d'avoir une distribution équitable des facteurs confondants parmi les divers groupes. Cette technique permet d'évaluer toutes les strates du facteur confondant. Pour chaque participant recruté, il y aura un autre participant ayant une valeur semblable du facteur confondant recruté dans l'autre groupe. Les désavantages principaux de l'appariement sont que c'est difficile à faire, que l'on perd beaucoup de participants et que l'analyse est plus complexe.
- La **randomisation** : Cette technique est la procédure de choix pour éliminer les facteurs confondants. Un avantage indéniable par rapport aux autres méthodes, est que cette technique permet de minimiser les facteurs confondants inconnus. Malheureusement, elle ne s'applique que pour les études où il y a une intervention. Elle est inapplicable pour les études de cohorte ou cas-témoin.

Voici deux méthodes de contrôle appliquées à la phase d'analyse pour le contrôle des facteurs confondants :

2- **L'analyse stratifiée** : Il s'agit d'une technique où l'on évalue la relation exposition-outcome pour diverses strates du facteur confondant (voir l'exemple plus haut). On peut ainsi rapporter le ratio de cote pour chaque strate. Ceci se rapporte bien lorsqu'il n'y a que deux strates (homme-femme, jeune-vieux) mais il est beaucoup plus laborieux lorsqu'il y a plusieurs strates (multiple groupes d'âge, ethnie, etc.). Il serait, par exemple, beaucoup plus fastidieux de lire un article qui rapporterait les ratios de cotes pour toutes les strates d'une étude où l'on a divisé l'âge des participants en 6 groupes. Une valeur sommaire peut être calculée en accordant un poids relatif à chaque strate. Plusieurs techniques ont été décrites mais la plus populaire est celle qui fut décrite par **Mantel et Haenszel**. Pour cette technique, on accorde un poids relatif à chaque strate. Ce poids est inversement proportionnel à la variance de la strate. Ainsi les strates contenant plusieurs participants ou ayant une faible variabilité auront une faible variance et auront un poids plus important dans le calcul. D'autre part, les strates avec peu de participants et, par conséquent, des variances larges auront peu d'influence sur le résultat global. Une dérivation mathématique complexe (au-dessus du niveau de ce cours) a permis d'en arriver aux formules suivantes :

   a. Pour une étude cas-témoin : $OR_{MH} = \Sigma(ad/T) \ / \ \Sigma(bc/T)$
   b. Pour une étude de cohorte : $RR_{MH} = \Sigma(a*(c+d) / T) \ / \Sigma(c*(a+b)/T)$

c. Pour le tableau suivant :

d.		e.	Cas	f.	Control	g.	
h.	Exposé	i.	.a	j.	b	k.	a+b
l.	Non exposé	m.	.c	n.	d	o.	c+d
p.		q.	.a+c	r.	b+d	s.	T

Ces analyses sont fréquentes et facilement accomplies par un ordinateur. Elles permettent de rapporter une valeur unique, non-biaisée par les facteurs confondants. On peut calculer la magnitude d'effet confondant en comparant les valeurs obtenues en utilisant tout les résultats (Ratio de cote cru) et les valeurs ajustées (Mantel haenszel). Plus la différence est grande, plus il y a un effet confondant.

3- **L'analyse multivariée** est la technique de choix de nos jours. Le perfectionnement des ordinateurs a permis de généraliser l'utilisation de l'analyse multivariée. L'avantage principal de cette technique est qu'elle permet d'évaluer une multitude de facteurs en même temps. On peut donc faire l'analyse de plusieurs facteurs ainsi que voir les interactions qu'il y entre eux. Cette analyse se fait généralement par l'utilisation du modèle mathématique de la régression linéaire multiple. Un désavantage est qu'il faut connaître le facteur confondant pour recueillir les informations à son sujet.

Comment faire une régression linéaire ?
La régression linéaire multiple est une extension de la régression linéaire simple, que l'on a tous apprit au secondaire lorsque l'on voulait décrire une ligne droite simple à partir de points (y=mx +b où m est la pente, b est l'ordonnée, y est la variable dépendante et x est la variable indépendante). Lorsque l'on met les données dans l'ordinateur (une variable dépendante et une indépendante), celui-ci calcule la valeur de la pente et la valeur de l'ordonnée qui vont minimiser la distance entre les points et la ligne droite.
On dit que Y= a +bX. Ou a est l'ordonnée et b est la pente.
Les ordinateurs ont la possibilité de refaire ces analyses en utilisant plusieurs dimensions permettant d'évaluer plusieurs facteurs de risques en même temps. On obtient alors la formule :
$Y = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$ ou a est l'ordonnée $b_{1\text{-}n}$ est le coefficient pour les divers facteur de risque 1à n.
Tout ceci sera discuté plus loin au cours 9 et 10.

V- La modification d'effet (effect modification)

Nous avons déjà vu que
*si les ratios de cotes pour les diverses strates sont tous semblables au résultat global il n'y a pas de facteur confondant
*si les ratios de cotes pour les diverses strates sont tous semblables entre eux mais différents du résultat global il y a un facteur confondant
Que faire si les ratios de cotes pour les diverses strates sont différents ? Ceci signifie que la relation entre l'exposition et le devenir est différente pour les diverses strates. Ainsi, on pourrait dire que l'effet de la fluoration de l'eau diminue le taux de carie dentaire chez les enfants à faible niveau socio-économique mais que cet effet n'est pas remarqué chez les plus riches. Voici 3 tableaux qui résument cet exemple :

Tableau carie dentaire vs fluoration dans une étude cas-témoin (population totale) :

	Fluoration	Pas de fluoration
Carie	40	80
Témoin	60	60

OR : (40/80) / (60/60) = 0.5 la fluoration protège

VI- Conclusion :

Analyse brute : OR(ou RR) pour l'exposition et la maladie.
Analyse stratifiée : OR, RR dans chaque strate
Recherche d'une interaction (modification d'effet)
Différence de RR/OR entre les strates
D'un point de vue de santé publique
Différences dans les IDC (chevauchement ??, test d'homogénéité)
Si interaction : STOP ; pas de mesure pondérée, donner les OR par strate (analyse séparée).
Si pas d'interaction, recherche de CONFUSION
Calcul de RR/OR pondéré : si différence avec OR/RR de plus de 15-20% → confusion ➔ prendre RR/OR pondéré (Ajusté) comme mesure d'association entre l'exposition et la maladie.

VII- Références bibliographiques :

- BEZZAOUCHA Abdeljalil, épidémiologie et Biostatistique, à l'usage des étudiants en sciences médicales, 3éme édition OPU.
- Ancelle T. Statistique Épidémiologie. Édition 2002.
- BEZZAOUCHA Abdeljalil, Tests statistiques en sciences médicales, OPU 2004.

Tableau carie dentaire vs fluoration dans une étude cas-témoin (population riche) :

	Fluoration	Pas de fluoration
Carie	20	40
Témoin	20	40

OR : (20/40) / (20/40) = 1.0 pas d'effet chez les riches

Tableau carie dentaire vs fluoration dans une étude cas-témoin (population pauvre) :

	Fluoration	Pas de fluoration
Carie	20	40
Témoin	40	20

OR : (20/40) / (40/20) = 0.25 donc fluoration protège chez les pauvres

Lorsqu'il y a modification d'effet, la valeur globale se retrouve toujours entre les valeurs les plus extrêmes des diverses strates. Si la valeur globale se retrouve en dehors des valeurs extrêmes des strates, il y a aussi un facteur confondant. Une modification d'effet se définit donc comme un facteur qui a un effet différent sur le outcome selon les diverses strates d'une population.

Références bibliographiques :

1. Cours d'épidémiologie pour fellow de l'urgence version Janier 2008

2.
3.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

Analyse stratifiée des données et tests d'ajustement

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

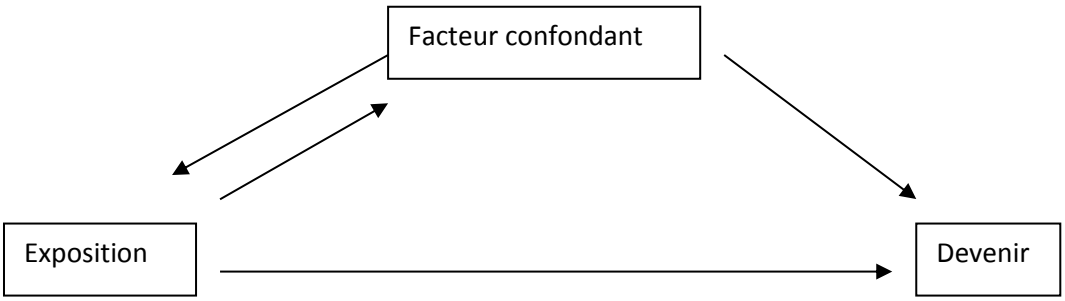
Analyse stratifiée des données et tests d'ajustement

Introduction

Un facteur confondant est un facteur qui biaise les résultats de l'étude parce que son association avec l'exposition et la maladie fait faussement croire que l'exposition est associée à la maladie. Contrairement aux types de biais précédemment décrits, les facteurs confondants sont des facteurs conduisant à des résultats erronés que l'on peut quantifier. On peut aussi en tenir compte dans l'analyse afin d'en éliminer l'effet.

On définit un **facteur confondant** par la présence obligatoire de 3 critères :

- Le facteur confondant est associé à l'exposition
- Le facteur confondant est associé au devenir (outcome)
- Cette association se fait de façon indépendante de l'exposition : Le facteur confondant n'intervient pas directement dans la relation exposition-outcome.



Voici un exemple :

Dans une étude sur le taux d'accident de voiture et l'utilisation de téléphone cellulaire on constate une forte association.

Voici le tableau des résultats :

	Cellulaire +	Cellulaire -
Cas (accident)	35	90
Témoin (pas d'accident)	15	70

OR : (35/90) / (15/70) = 1.8

Un chercheur voit ces résultats et suggère qu'il pourrait y avoir un important facteur confondant : l'âge. Les jeunes de moins de 25 ans ont beaucoup plus d'accident de voiture et ils sont beaucoup plus fréquents à posséder un téléphone cellulaire. Voici 2 tableaux de la même étude selon que les participants aient < 25 ans ou > 25 ans.

Pour les moins de 25 ans

	Cellulaire +	Cellulaire -
Cas	30	60
Témoin	5	10

OR : (30/60) / (5/10) = 1.0 (pas d'association)

Pour les plus de 25 ans

	Cellulaire +	Cellulaire -
Cas	5	30
Témoin	10	60

OR : (5/30) / (10/60)= 1.0 (pas d'association)

On voit donc que l'utilisation n'est pas un facteur de risque d'accident pour les gens de moins de 25 ans ni pour ceux de plus de 25 ans !!!! Par contre, lorsqu'on les regroupe sans considérer l'âge, on obtient une association. Ceci est une belle démonstration d'un facteur confondant. On voit que l'âge est associé avec l'exposition (les plus jeunes ont plus de téléphones) et avec le outcome (les jeunes ont plus d'accident). De plus, l'âge n'intervient pas directement dans la relation cellulaire-accident.

D'autres exemples seraient :

- Relation calvitie- consommation de bière. Un facteur confondant pourrait être le sexe
- Relation # de téléviseur à la maison-cancer de la prostate. L'âge ou le niveau socio économique pourraient être des facteurs confondants.

Il existe 2 façons de démontrer la présence d'un facteur confondant:

1. On évalue l'association entre l'exposition et le devenir de façon séparée pour les diverses strates du présumé facteur confondant. Si les OR sont tous semblables à la valeur du OR total, il n'y a pas de facteur confondant. Si les OR sont semblables pour les diverses strates mais différentes du résultat total, il faut suspecter un facteur confondant. Ceci fut démontré dans les tableaux précédents

2. On peut aussi montrer qu'il existe une association entre le facteur confondant et l'exposition et une association entre le facteur confondant et le devenir (outcome). Voici les tableaux pour l'exemple précédent.

Tableau de l'association age et cellulaire :

	Cellulaire +	Cellulaire -
Âge < 25 ans	35	70
Âge > 25 ans	15	90

OR : (35/70) / (15/90) = 3.0

Tableau de l'association age et accident :

	Accident	Témoin
Âge < 25 ans	90	15
Âge > 25 ans	35	70

OR : (90/15) / (35/70) = 12.0

On voit donc une association entre l'âge et l'exposition (OR 3) et entre l'âge et le devenir (OR12). L'âge est donc un facteur confondant puisqu'il n'est pas impliqué directement dans la relation cellulaire-accident.

Management des facteurs confondants

On peut prévenir les facteurs confondant à la phase de l'élaboration d'une étude ou en tenir compte lors de son analyse.

Voici 3 méthodes de contrôle lors de l'élaboration:

- La **restriction** : Pour confondre, un facteur doit être présent en fréquence variable dans les divers groupes (exposé vs non). Une façon de contrecarrer cette variabilité est de restreindre la participation à l'étude qu'à certains individus selon leur statut face au facteur confondant. On peut, par exemple, restreindre l'étude qu'aux hommes, qu'aux enfants d'un certain âge ou membre d'une seule communauté ethnique. La restriction est simple à appliquer mais elle diminue souvent de beaucoup le nombre de participants potentiels et elle ne permet pas de conclure pour les patients qui n'étaient pas dans les strates évaluées. On limite donc la généralabilité des résultats
- L'**appariement** : Dans ce type de procédé, les participants sont recrutés afin d'avoir une distribution équitable des facteurs confondants parmi les divers groupes. Cette technique permet d'évaluer toutes les strates du facteur confondant. Pour chaque participant recruté, il y aura un autre participant ayant une valeur semblable du facteur confondant recruté dans l'autre groupe. Les désavantages principaux de l'appariement sont que c'est difficile à faire, que l'on perd beaucoup de participants et que l'analyse est plus complexe.
- La **randomisation** : Cette technique est la procédure de choix pour éliminer les facteurs confondants. Un avantage indéniable par rapport aux autres méthodes, est que cette technique permet de minimiser les facteurs confondants inconnus. Malheureusement, elle ne s'applique que pour les études où il y a une intervention. Elle est inapplicable pour les études de cohorte ou cas-témoin.

Voici deux méthodes de contrôle appliquées à la phase d'analyse pour le contrôle des facteurs confondants :

- **L'analyse stratifiée** : Il s'agit d'une technique où l'on évalue la relation exposition-outcome pour diverses strates du facteur confondant (voir l'exemple plus haut). On peut ainsi rapporter le ratio de cote pour chaque strate. Ceci se rapporte bien lorsqu'il n'y a que deux strates (homme-femme, jeune-vieux) mais il est beaucoup plus laborieux lorsqu'il y a plusieurs strates (multiple groupes d'âge, ethnie, etc.). Il serait,

par exemple, beaucoup plus fastidieux de lire un article qui rapporterait les ratios de cotes pour toutes les strates d'une étude où l'on a divisé l'âge des participants en 6 groupes. Une valeur sommaire peut être calculée en accordant un poids relatif à chaque strate. Plusieurs techniques ont été décrites mais la plus populaire est celle qui fut décrite par **Mantel et Haenszel**. Pour cette technique, on accorde un poids relatif à chaque strate. Ce poids est inversement proportionnel à la variance de la strate. Ainsi les strates contenant plusieurs participants ou ayant une faible variabilité auront une faible variance et auront un poids plus important dans le calcul. D'autre part, les strates avec peu de participants et, par conséquent, des variances larges auront peu d'influence sur le résultat global. Une dérivation mathématique complexe (au-dessus du niveau de ce cours) a permis d'en arriver aux formules suivantes :

- Pour une étude cas-témoin : OR_{MH} = □□ad/T) / □□bc/T)
- Pour une étude de cohorte : RR_{MH} = □□a*(c+d) / T) / □□c*(a+b)/T)
- Pour le tableau suivant :

○	○ Cas	○ Control	○
○ Exposé	○ .a	○ b	○ a+b
○ Non exposé	○ .c	○ d	○ c+d
○	○ .a+c	○ b+d	○ T

Ces analyses sont fréquentes et facilement accomplies par un ordinateur. Elles permettent de rapporter une valeur unique, non-biaisée par les facteurs confondants. On peut calculer la magnitude d'effet confondant en comparant les valeurs obtenues en utilisant tout les résultats (Ratio de cote cru) et les valeurs ajustées (Mantel haenszel). Plus la différence est grande, plus il y a un effet confondant.

- **L'analyse multivariée** est la technique de choix de nos jours. Le perfectionnement des ordinateurs a permis de généraliser l'utilisation de l'analyse multivariée. L'avantage principal de cette technique est qu'elle permet d'évaluer une multitude de facteurs en même temps. On peut donc faire l'analyse de plusieurs facteurs ainsi que voir les interactions qu'il y entre eux. Cette analyse se fait généralement par l'utilisation du modèle mathématique de la régression linéaire multiple. Un désavantage est qu'il faut connaître le facteur confondant pour recueillir les informations à son sujet.

Comment faire une régression linéaire ?

La régression linéaire multiple est une extension de la régression linéaire simple, que l'on a tous apprit au secondaire lorsque l'on voulait décrire une ligne droite simple à partir de points (y=mx +b où m est la pente, b est l'ordonnée, y est la variable dépendante et x est la variable indépendante). Lorsque l'on met les données dans l'ordinateur (une variable dépendante et une indépendante), celui-ci calcule la valeur de la pente et la valeur de l'ordonnée qui vont minimiser la distance entre les points et la ligne droite.

On dit que Y= a +bX. Ou a est l'ordonnée et b est la pente.

Les ordinateurs ont la possibilité de refaire ces analyses en utilisant plusieurs dimensions permettant d'évaluer plusieurs facteurs de risques en même temps. On obtient alors la formule :

$Y = a + b_1X_1 + b_2X_2 + \ldots b_nX_n$ ou a est l'ordonnée b_{1-n} est le coefficient pour les divers facteur de risque 1à n.

Tout ceci sera discuté plus loin au cours 9 et 10.

4.4 La modification d'effet (effect modification)

Nous avons déjà vu que

- si les ratios de cotes pour les diverses strates sont tous semblables au résultat global il n'y a pas de facteur confondant
- si les ratios de cotes pour les diverses strates sont tous semblables entre eux mais différents du résultat global il y a un facteur confondant

Que faire si les ratios de cotes pour les diverses strates sont différents ? Ceci signifie que la relation entre l'exposition et le devenir est différente pour les diverses strates. Ainsi, on pourrait dire que l'effet de la fluoration de l'eau diminue le taux de carie dentaire chez les enfants à faible niveau socio-économique mais que cet effet n'est pas remarqué chez les plus riches. Voici 3 tableaux qui résument cet exemple :

Tableau carie dentaire vs fluoration dans une étude cas-témoin (population totale) :

	Fluoration	Pas de fluoration
Carie	40	80
Témoin	60	60

OR : (40/80) / (60/60) = 0.5 la fluoration protège

Tableau carie dentaire vs fluoration dans une étude cas-témoin (population riche) :

	Fluoration	Pas de fluoration
Carie	20	40
Témoin	20	40

OR : (20/40) / (20/40) = 1.0 pas d'effet chez les riches

Tableau carie dentaire vs fluoration dans une étude cas-témoin (population pauvre) :

	Fluoration	Pas de fluoration
Carie	20	40
Témoin	40	20

OR : (20/40) / (40/20) = 0.25 donc fluoration protège chez les pauvres

Lorsqu'il y a modification d'effet, la valeur globale se retrouve toujours entre les valeurs les plus extrêmes des diverses strates. Si la valeur globale se retrouve en dehors des valeurs extrêmes des strates, il y a aussi un facteur confondant. Une modification d'effet se définit donc comme un facteur qui a un effet différent sur le outcome selon les diverses strates d'une population.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

Validité d'un test diagnostique

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Validité d'un test diagnostique

PLAN DU COURS

1. Rappel sur les principes de mise en œuvre d'un programme de dépistage
    1.1. Définitions (étymologie du dépistage et du diagnostic)
    1.2. Les méthodes de détection de cas de maladies dans la population ;
    1.3. Les critères d'un bon dépistage
2. Performance d'un test en situation expérimentale :
    2.1. Définition d'un test
    2.2. Sensibilité et Spécificité d'un test
    2.3. Cas d'un test qualitatif
    2.4. Cas d'un test quantitatif
    2.5. Courbes Roc
    2.6. Combinaisons de plusieurs tests
3. Performance d'un test en situation de terrain :
    3.1. Valeur prédictive positive
    3.2. Valeur prédictive négative
    3.3. Interprétation des VPP VPN
    3.4. Variation en fonction de la prévalence
    3.5. Variation en fonction de la sensibilité et de la spécificité
4. Reproductibilité et concordance
5. Indice de Youden
6. Rapports de vraisemblance
7. Mode d'échantillonnage
8. Les biais

9. Bibliographie

Validité d'un test diagnostique

1. Rappel sur les principes de mise en œuvre d'un programme de dépistage

En médecine de soins traditionnelle, la démarche du médecin est de répondre à une demande de la part du patient : cette démarche aboutit le plus souvent à un diagnostic et la mise en place d'un traitement. Lorsqu'il s'agit de faire un **diagnostic de masse**, la démarche est différente : en effet, il est rarement possible de soumettre chaque sujet à un examen clinique détaillé et éventuellement à plusieurs examens complémentaires, ce qui représente une procédure longue et coûteuse. On doit donc se contenter le plus souvent de procédures plus simples et plus rapides, même si elles sont approximatives et associées à un certain taux d'erreur.

L'objectif du dépistage est d'**améliorer la santé des individus par le diagnostic précoce des maladies** à un stade où elles sont curables ou quand leurs conséquences peuvent être limitées.

1.1 - Définition (étymologie du dépistage et du diagnostic) :

Le dépistage est l'action de rechercher systématiquement et découvrir ce qui peu apparent (définition 1896).

Selon le dictionnaire d'épidémiologie (Last), le dépistage est l'identification dans une population à priori en bonne santé de sujets présentant :

- Soit une maladie inapparente ;
- Soit un risque d'une maladie donnée.

En vue d'examen complémentaires ou de mesures de prévention.

Donc un test de dépistage permet de faire la différence entre les personnes apparemment en bonne santé et ceux qui le sont réellement, il ne pose pas le diagnostic.

Par contre, un test diagnostic représente la méthode de référence qui permet de poser le diagnostic d'une maladie.
Le tableau ci-dessous indique schématiquement les principales différences entre test de dépistage et test de diagnostic.

Tableau n°1 : Différence entre test de dépistage et test diagnostic

Test de dépistage	**Test de diagnostic**
Appliqué aux personnes apparemment en bonne santé	Doit donner une certitude diagnostique (examens spécifiques)
Pratiqué sur des groupes d'individus	Appliqué aux personnes présentant des troubles définis
Ne constitue pas une base de traitement	Essentiellement individuel
Moins précis que le test de diagnostic	Constitue une base du traitement
Coûte moins cher que le test de diagnostic	Plus précis que le test de dépistage
	Coûte plus cher que le test de dépistage

1.2. Les méthodes de détection de cas de maladies dans la population :

Ils existent deux méthodes épidémiologiques pour détecter les cas de maladies dans la population, c'est soit la mise en place d'une surveillance épidémiologique soit la pratique d'un dépistage.

La surveillance est la mise en place d'un système qui permet la collecte systématique de données sur la survenue de maladies dans la population. Elle fournit des données sur les nouveaux cas donc des taux d'incidence.

Le dépistage, au contraire va rechercher des cas de maladie dans une population à un moment donné, donc, il va porter sur la prévalence et s'applique surtout aux maladies chroniques.

1.3. Critères de recours au dépistage

Le dépistage doit répondre à un certain nombre de critères définis par l'O.M.S. comme les 10 principes servant au choix d'un programme de dépistage.
1. La maladie dépistée doit constituer une menace grave pour la santé publique (fréquence de la pathologie, gravité des cas, ...).
2. Il doit exister un traitement d'efficacité démontrée.
3. Il faut disposer de moyens appropriés de diagnostic et de traitement.
4. L'intervention au stade précoce de la maladie doit influencer favorablement l'évolution et le pronostic.
5. Il existe un examen de dépistage efficace au stade précoce de la maladie.
6. les techniques de dépistage et de traitement doivent être acceptables par la population.
7. Il faut bien connaître l'histoire naturelle de la maladie.
8. Il faut que le programme cible la population à risque.
9. Il faut que le coût du programme soit acceptable.
10. Il faut que le programme de dépistage soit réalisé de façon continue.

2. Performance d'un test en situation expérimentale :

2.1. Définition d'un test :

On appel un test la méthode de référence qui permet de poser le diagnostic d'une maladie
Un résultat d'un examen clinique à la recherche d'un symptôme avec une simple donnée qualitative: présence/absence ;

Un résultat d'un examen biologique avec des données quantitatives : valeur d'une mesure (glycémie….);

Un résultat d'un examen paraclinique: mammographie dans le cancer du sein;

Un résultat d'un groupe d'examen (clinique, biologique, paraclinique);

Un résultat de l'évaluation d'un système de surveillance (déclaration obligatoire, recueil de données, capacité à détecter les épidémies Dans cette situation, un « cas » est une épidémie.

Le résultat d'un test s'exprime :

- soit par une variable qualitative binaire : test positif ou négatif;
- soit par une variable quantitative : valeur d'une mesure biologique, note, indice, etc.

2.2. Sensibilité et Spécificité d'un test :

Un test doit posséder deux qualités majeures : la sensibilité et la spécificité.
Il s'agit des qualités « intrinsèques » d'un test.
La sensibilité d'un test est sa capacité à détecter les cas d'une maladie.
C'est la probabilité quand on a la maladie d'être diagnostiqué comme malade.

La spécificité d'un test est sa capacité à identifier les non malades (sains).
C'est la probabilité quand on n'a pas la maladie d'être diagnostiqué comme non malade (sain).

Dans la situation idéale, le test permet de classer correctement tous les sujets (pas de faux négatifs ni faux positifs).
Cependant, dans la plupart des cas, le classement des sujets dans le groupe des malades et des non malades s'accompagne d'un certain taux d'erreur.

2.3. Cas d'un test qualitatif

Dans le cas d'un test de nature qualitative pour lequel la réponse est binaire (positif/négatif, présent/absent), la sensibilité et la spécificité sont fixes. La confrontation des résultats entre le test à évaluer et le test de référence est en général présentée dans un tableau de contingence de la manière suivante :

Maladie Test	Présente	Absente	Total
+	VP	FP	VP+FP
-	FN	VN	FN+VN
Total	VP+FN	FP+VN	

-VP (vrais positifs) : sujets effectivement malades pour lesquels le test est positif.
− VN (vrais négatifs) : sujets effectivement non malades pour lesquels le test est négatif.
− FP (faux positifs) : sujets en réalité non malades pour lesquels le test est positif
− FN (faux négatifs) : sujets en réalité malades pour lesquels le test est négatif.

Sensibilité (Se) = VP/VP+FN

$$\mathrm{Se} \pm \varepsilon \sqrt{\frac{\mathrm{Se}\,(1-\mathrm{Se})}{\mathrm{a}+\mathrm{c}}}$$

Spécificité (Sp) = VN/VN+FP

$$\mathrm{Sp} \pm \varepsilon \sqrt{\frac{\mathrm{Sp}\,(1-\mathrm{Sp})}{\mathrm{b}+\mathrm{d}}}$$

La Sensibilité et la Spécificité sont des valeurs comprises entre 0 et 1, elles s'expriment en pourcentage. Les résultats de la Se et de la Sp sont toujours exprimées selon leurs intervalles de confiance.

Exemple : diagnostic des hépatites virales (HV) par les transaminases (TA)

Maladie Test	Présente	Absente	Total
TA+	10	500	510
TA-	0	9 490	9 490
Total	10	9 990	10 000

Se = VP/VP + FN = 10 /10 =100 % **Sp = VN/VN + FP = 9490/9990 = 95 %**

La sensibilité et la spécificité varient en sens inverse, donc, il est difficile de choisir entre deux techniques (l'une sensible et peu spécifique, l'autre peu sensible et spécifique) d'où la nécessité de faire appel à d'autres indices.

2.4. Cas d'un test quantitatif

Le test peut se traduire par un résultat exprimé sous forme d'une valeur numérique.
En raison de la variabilité biologique, ces valeurs sont différentes d'un sujet à l'autre. Si une série de sujets est examinée, les résultats vont s'afficher sous forme d'une distribution.
Le problème est donc de déterminer une **valeur seuil** qui permettra de classer les malades et les sujets sains.

Si le test est parfaitement discriminant, la distribution des valeurs dans le groupe des cas, sera bien séparée de la distribution des valeurs dans un groupe de sujets sains. Il sera alors aisé de choisir une valeur seuil qui permettra une sensibilité et une spécificité de 100 %. Hélas, cette situation est rarement observée en biologie médicale (figure 1).

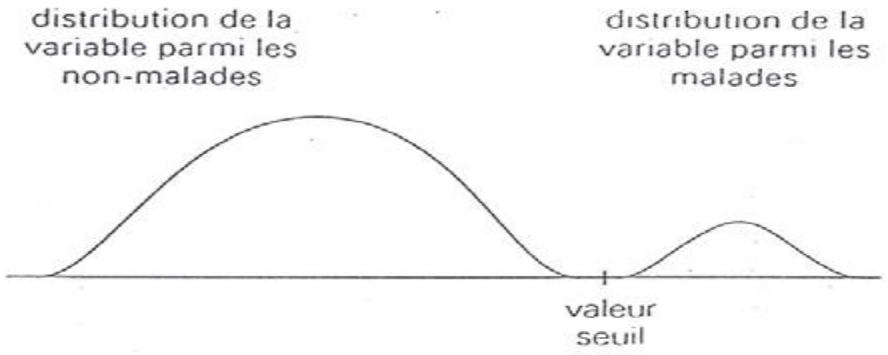
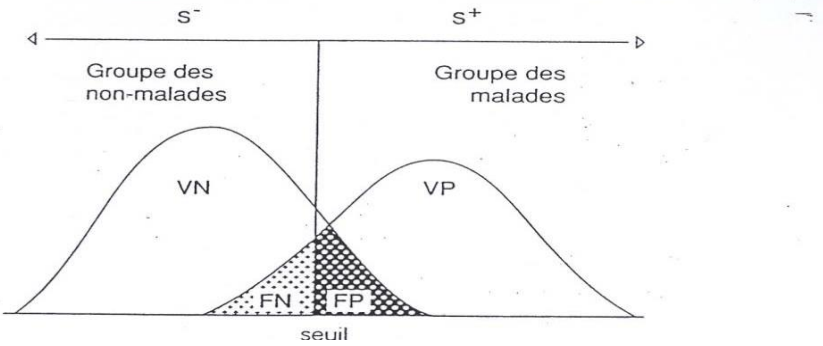


Figure 1 : Test quantitatif, distribution des valeurs observées chez les cas et les sujet sains

Le plus souvent, il y a chevauchement des deux distributions, certains sujets sains présentent des valeurs qui peuvent être identiques à celle de sujets malades et à l'inverse certains malades présentent des valeurs qui peuvent être identiques à celles de sujets sains. Le choix d'un seuil devient une opération délicate, car il divise les deux groupes de sujets en 4 sous-groupes, VP, FP, VN, FN (figure 2).

Fig. 2 : Distribution des valeurs d'un test quantitatif pratiqué sur des malades (M+) et des non malades (M-). Choix d'une valeur seuil.



On conçoit que les qualités diagnostiques du test vont varier selon le seuil choisi. Si on décide de baisser le seuil, le nombre de faux négatif diminue, donc la sensibilité augmente. Mais parallèlement, le nombre de faux positifs augmente et la spécificité diminue (figure 3).

Si à l'inverse on décide d'élever le seuil, le nombre de faux positifs diminue, donc la spécificité augmente, mais parallèlement le nombre de faux négatif augmente, donc la sensibilité diminue (figure 4).

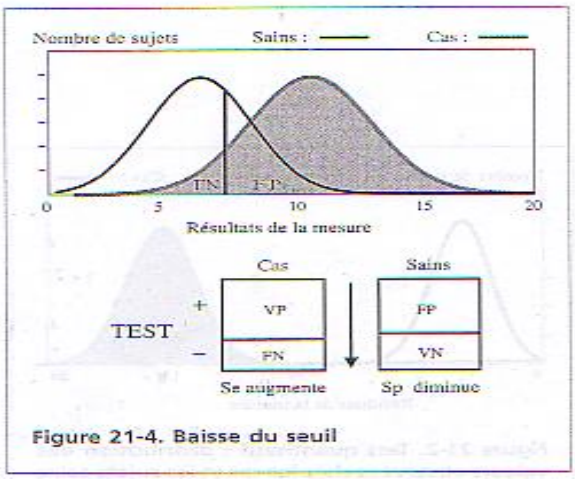


Figure 21-4. Baisse du seuil

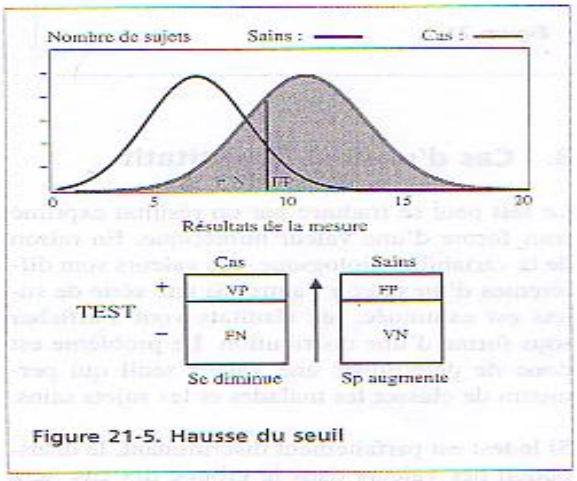


Figure 21-5. Hausse du seuil

On constate que sensibilité et spécificité d'un test quantitatif varient en sens inverse. Le choix d'un seuil est le résultat d'un compromis qui est fonction de l'objectif assigné au test.

Choix d'un seuil :

1. Lorsque les erreurs par excès (FP) sont plus graves que les erreurs par défaut:
Dans le dépistage anténatal de l'anencéphalie par un test biologique, un dépistage faussement positif (erreur par excès) donc faire un faux diagnostic d'anencéphalie entraîne des conséquences très lourdes (mise sous traitement prolongé ou interruption de grossesse).
A l'inverse, ne pas dépister grâce à ce test biologique, une anencéphalie, est une erreur par défaut (faux résultat négatif), qui pourra être rattrapé ultérieurement par le suivi échographique.

On veut dans ce cas: minimiser les FP donc améliorer la Spécificité donc élever le seuil de positivité.

2. Lorsque les erreurs par défaut (FN) sont plus graves que les erreurs par excès:
Dans le dépistage pré-transfusionnel du paludisme dans les flacons de la banque du sang, un dépistage faussement négatif entraînera l'administration de sang parasité chez un receveur, alors qu'un résultat faussement positif n'entraînera que le rejet à tort du flacon.

Dans ce cas: on minimise les FN donc on améliore la Se et on abaisse le seuil de positivité.

Le choix du seuil de positivité est un compromis et dépend de l'objectif assigné au test :
On augmente la sensibilité quand les conséquences d'un FN sont graves et on augmente la spécificité quand les conséquences d'un FP sont graves.

2.5. Courbe ROC :

La courbe ROC (Receiver Operating Characteristics) est une représentation graphique de la relation existante entre la sensibilité et la spécificité d'un test pour toutes les valeurs seuils possibles. C'est une courbe qui sert à fixer le seuil d'une méthode quantitative.
On appliquer le test pour deux groupes de sujets (malades et non malades). Pour chaque seuil possible, on calcule la Sensibilité et la Spécificité du test.

On obtient une liste de couples Sensibilité-Spécificité.
On porte en ordonné la Sensibilité et en abscisse le pourcentage des faux positifs (1- Spécificité).
La courbe de ROC est définie par:
-Un point d'inflexion: le point le plus prés du coin haut gauche, ça représente le seuil optimum « le seuil qui a la Se et la Sp les plus élevées » (figure 3).
- L' aire sous la courbe (AUC): la mesure de la surface doit être statistiquement supérieure à 0.5, plus elle s' éloigne de 0.5 plus le rendement du test est élevé (figure 3).

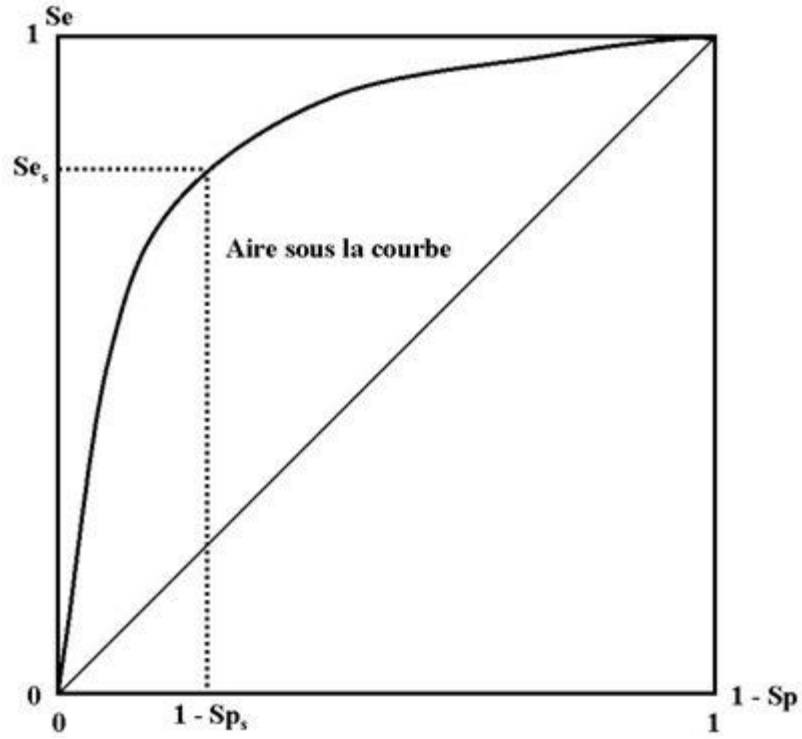


Figure 3 : courbe ROC

L'aire sous la courbe (AUC) varie de 0.5 à 1, le test est discriminant si AUC > 0.9
Si AUC est égale à 1, on parle d'une discrimination parfaite, si elle est égale à 0.5, on parle d'absence de discrimination (figue 4).

La courbe ROC est utilisée non seulement pour déterminer la valeur seuil optimale d'un test tout en prenant en compte les données épidémiologiques et médicoéconomiques de la maladie, mais, elle trouve surtout son intérêt lorsqu'on souhaite comparer plusieurs tests, par la comparaison des courbes ROC.

2.6. Combinaisons de plusieurs tests :

On peut calculer grâce au développement de nombreux logiciels informatiques, les aires sous les courbes ensuite on compare les aires sous les courbes à l'aide d'un test statistique (figure 5).
Donc la comparaison des aires sous la courbe permet d'apprécier et de classer les performances diagnostiques de plusieurs tests, mieux que la simple étude des couples sensibilité – spécificité.

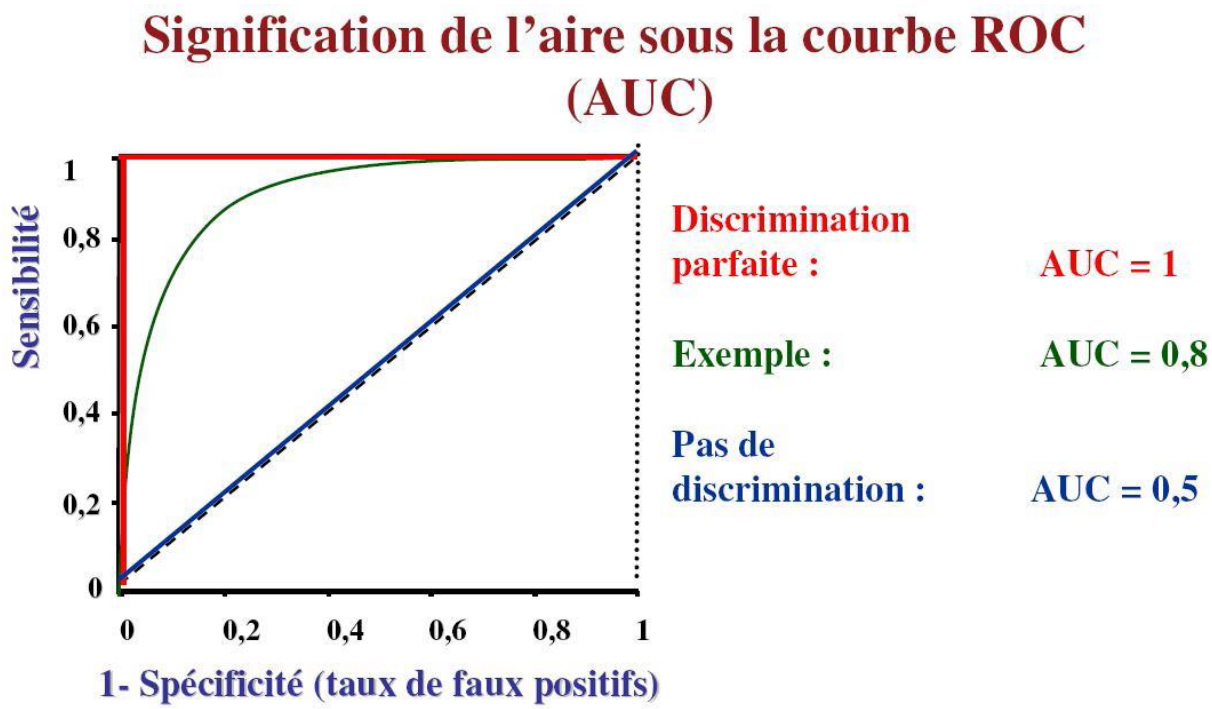


Figure 4 : l'aire sous la courbe

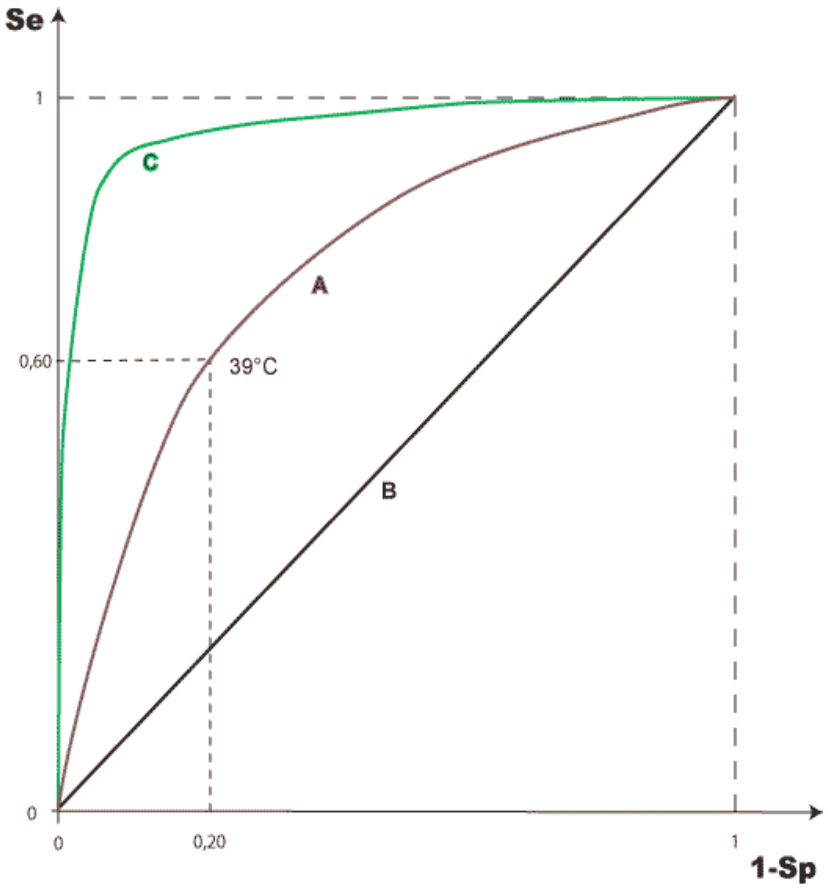


Figure 5 : comparaison de plusieurs courbes ROC

3. Performance d'un test en situation réelle (Valeurs extrinsèques d'un test) :

Un test dont on connaît la sensibilité et la spécificité est conçu pour être appliqué sur le terrain (population: malade non malade).

Cette population peut être la population générale ou une population ciblée pour un dépistage systématique. Dans ce cas la prévalence de la maladie recherchée est alors plus faible (Dépistage).

La population peut aussi être une population déjà sélectionnée : consultants chez un médecin ou un spécialiste, prélèvements pour un laboratoire, examens para cliniques prescrits à partir d'une indication clinique. On se trouve dans une situation diagnostique où la prévalence de la maladie recherchée est alors plus élevée que dans la population générale (Diagnostic).

3.1. Valeur prédictive positive:
Lorsqu'un test est positif, il existe deux possibilités : soit le sujet est malade, soit le sujet n'est pas malade malgré ce résultat contradictoire.

On appelle valeur prédictive positive d'un test (VPP) la probabilité d'être malade lorsque le résultat est positif.

3.2. Valeur prédictive négative:

Lorsqu'un test est négatif, il existe deux possibilités : soit le sujet est sain, soit le sujet est malade malgré ce résultat contradictoire.
On appelle valeur prédictive négative d'un test (VPN) la probabilité d'être sain lorsque le résultat est négatif.

3.3. Interprétation :

Exemple : reprenons notre exemple des hépatites virales et l'augmentation des transaminases.

Maladie Test	Présente	Absente	Total
TA+	10 (VP)	500 (FP)	510
TA-	0 (FN)	9 490 (VN)	9 490
Total	10	9 990	10 000

Sensibilité = 100 %
Spécificité = 95 %
VPP = VPP = VP/VP + FP = 10/510 = 2 %
Interprétation : si un sujet à une augmentation des TA, il a 2 % de « chance » d'avoir une hépatite virale
VPN = VPN = VN/VN+FN = 9490/9490 = 100 %
Interprétation : si un sujet n'à pas d'augmentation des TA, il a 100 % de « chance » d'être indemne d'hépatite virale.

3.4. Relation entre valeurs prédictives et prévalence de la maladie :

Ex : test ELISA de dépistage du VIH dans un centre de transfusion **(prévalence faible du VIH).**

Maladie Test	VIH +	VIH -	Total
ELISA+	999 (VP)	9999 (FP)	10998
ELISA -	1 (FN)	9989001 (VN)	9 989002
Total	1000	9999000	10^7

Se = 999/1000 = 0.99 %
Sp = 9989001/9999000 = 0.99 %
VPP = 999/10998 = 9.08 %
VPN = 9989001/9 989002 = 100 %

SI le même test appliqué dans un service de maladies infectieuses **(prévalence élevée)**

Se = Sp =99.9 %
VPP = 7992/8000 = 7994 99.97 %
VPN = 1998/2006 = 99.59 %

Les deux situations montrent que les valeurs prédictives d'un test, pour une sensibilité et une spécificité donnée varient en fonction de la prévalence.
Une VPP et une VPN non accompagnées de la prévalence estimée de la maladie n'ont aucun sens.
Ainsi un même test, n'aura pas les mêmes valeurs prédictives s'il est appliqué dans une population à **forte prévalence (diagnostique) ou à faible prévalence (dépistage)**.

Les valeurs prédictives d'un test dépendent donc de la prévalence de la maladie.

Maladie Test	VIH +	VIH -
+	(VP) **Se NP**	(FP) **(1-Sp) N (1-p)**
-	(FN) **(1-Se)NP**	(VN) SpN(1-P)
Total	**NP**	**N (1-P)**

N : Total de la population étudiée
P : prévalence
Se : sensibilité
Sp : spécificité

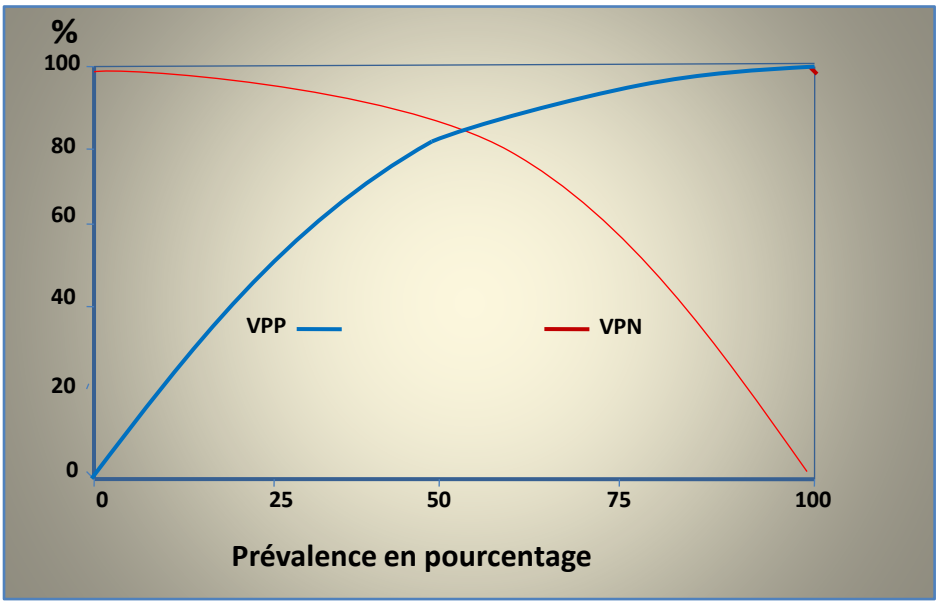
On démontre par le théorème de Bayes que :

VPP = SeP/ SeP + (1 – Sp)(1 – P)

VPN = Sp (1 - P)/ Sp (1 – P) + (1 – Se)(P)

Le graphique suivant ; exprime les valeurs prédictives en ordonnées en fonction de la prévalence en abscisses.

Valeurs prédictives en fonction de la prévalence de la maladie.
Se = 90% , Sp = 90 %.



On constate que pour de faibles prévalences, la VPP est très faible et varie rapidement. A l'inverse la VPN est élevée et varie peu. Pour de fortes prévalences, la VPP est élevée et varie peu tandis que la VPN est faible et varie rapidement.

Ainsi, un test appliqué en situation de dépistage en population générale (Prévalence basse) aura une faible VPP et une forte VPN.
De nombreux sujets seront alertés à tort, mais un résultat négatif sera rassurant.
A l'inverse, le même test appliqué en situation de diagnostic, dans un service spécialisé (Prévalence élevé), aura une VPP élevée et une VPN moindre : un résultat positif sera hautement en faveur de la maladie tandis qu'un résultat négatif aura une signification moindre.

3.5. Variation des valeurs prédictives en fonction de la sensibilité et de la spécificité:
1) D'après la formule de la VPP, on constate qu'elle dépend essentiellement du terme 1 -Sp au dénominateur. Donc, plus Sp est élevée et plus la VPP est élevée.

La valeur prédictive positive d'un test dépend de la spécificité

2) D'après la formule de la VPN, on constate qu'elle dépend essentiellement du terme 1-Se au dénominateur. Donc, plus la Se est élevée, plus la VPN est élevée.

La valeur prédictive négative d'un test dépend de la sensibilité

4. Reproductibilité et concordance :

La reproductibilité est la capacité du test à donner le même résultat lors d'essais répétés chez le même sujet. La concordance est la conformité ou la similitude de 2 ou plusieurs informations (tests) se rapportant au même sujet (ou concordance d'un expérimentateur à un autre).

Imaginons un test qualitatif dont le résultat peut s'exprimer par une variable à 3 classes:

Négatif, douteux et positif. La même série d'examen a été soumis à deux expérimentateurs A et B. on peut donc observer les résultats suivants :

A	**Négatif**	**Douteux**	**Positif**
B			
Négatif	--	±	- +
Douteux	± -	± ±	± +
Positif	± -	+±	+ +

De façon plus générale on obtient le tableau de résultats suivants portant su N tests réalisés par deux expérimentateurs ou lors de 2 séances de travail A et B

A **B**	**A1**	**A2**	**…**	**Ai**	**Total**
B1	O11	O12	…	O1i	t1
B2	O21	O22	…	O2i	t2
…	…		…	…	
Bi	Oi1	Oi2	…	Oii	ti
Total	n1	n2	…	ni	N

Les résultats de A et B sont exprimés selon les modalités 1 à i. Le nombre de test selon le résultat de A et de B est exprimé dans chaque case par le nombre Oii
Le coefficient de concordance Cc est égale à la somme des résultats concordants sur le nombre total d'examens. Elle s'exprime par un nombre entre 0 et 1 (pourcentage)

$$Cc = \frac{\text{nombre d'examens concordants}}{\text{nombre d'examenscomparés}} = \frac{O11 + O22 + \cdots + Oii}{N}$$

Le coefficient Kappa (K) :

- La mesure Kappa mesure l'accord entre deux observateurs lorsque le critère de jugement est un caractère qualitatif
- Le coefficient Kappa quantifie l'accord observé par rapport à l'accord lié au hasard. En d'autres termes, il calcule la concordance réelle après prise en compte de la concordance due au hasard

On calcule d'abord la concordance attendue Ca de la façon suivante

$$Ca = \frac{t1n1+t2n2+\ldots\ldots +tini}{N2}$$

- On appelle coefficient Kappa le terme

$$k = \frac{Cc - Ca}{1 - Ca}$$

Le coefficient Kappa s'exprime par un nombre compris entre -1 et +1

- Un K proche de -1 signifie une discordance complète
- Un k proche de 0 signifie une concordance moyenne du au hasard
- Un k proche de + 1 signifie une concordance absolue

5. Indice de Youden :

L'addition des deux qualités d'un test (sa sensibilité et sa spécificité) conduit à un indice synthétique tel que :

J= Se+Sp-1

L'indice de Youden varie entre –1 et 1. Un indice égal à 0 traduit un test qui n'a aucune efficacité d'orientation diagnostique. Sa valeur diagnostique est maximale lorsque l'indice est proche de 1.

6. Rapports de vraisemblance (RV)

Définition :
Les RV ou Likelihood ratio (LR), sont un autre mode d'expression des caractéristiques intrinsèques d'un test. Ils estiment le rapport entre la probabilité d'avoir un test positif (ou négatif) chez les sujets malades et celle d'avoir un test positif (ou négatif) chez les sujets sains. Calculés à partir de la Sensibilité et de la Spécificité du test, ils sont donc **indépendants de la prévalence de la maladie dans la population**.

Le rapport de vraisemblance positif (LR+) est égal au taux de tests positifs chez les malades (soit la Se) sur le taux de tests positifs chez les non malades (soit [1 — Sp]):
LR+ = Se /(1 − Sp)
Il quantifie le gain diagnostique d'un test positif, un individu malade ayant LR fois plus de chance d'avoir un test positif qu'un individu sain.

Le rapport de vraisemblance négatif (LR-) est égal au taux de tests négatifs chez les malades soit (1 — Se) sur le taux de tests négatifs chez les sujets sains (soit la Sp):

LR- = (1 − Se)/Sp

Il quantifie le gain diagnostique d'un test négatif, un individu malade ayant LR- fois plus de chance d'avoir un test négatif qu'un individu sain.

Plus le rapport de vraisemblance positif est élevé et plus le rapport de vraisemblance négatif est faible, plus le gain diagnostique du test est important (tableau n°2).

Exemple : test de diagnostic de la thrombose veineuse profonde par échodoppleur couleur.

TVP		Présence	Absence	
Test	**Echodoppleur couleur anormale**	**8**	**23**	**31**
	Echodoppleur couleur normale	**13**	**275**	**288**
		21	**298**	**319**

TVP : thrombose veineuse profonde

Sensibilité=38,1%
Spécificité=92.3%
RV+ = 4,94
RV- = 0,40

Le rapport de vraisemblance positif de 4,94 exprime qu'un écho-doppler couleur anormal a 4,94 fois plus de chances de provenir d'une population de sujets ayant une thrombose veineuse profonde que d'une population de sujets n'en ayant pas.

Un individu malade à 0.40 fois plus de chance d'avoir un test négatif qu'un individu sain.

Le rapport de vraisemblance a donc la même signification que celle du risque relatif en épidémiologie.

Tableau 2 : apport diagnostic d'un test en fonction de la valeur des rapports de vraisemblance positif et négatif.

RV+	RV-	Apport diagnostic
Sup 10	Inf 10	Très fort
5-10	0.1-0.2	Fort
2-5	0.2-0.5	Modéré
1-2	0.5-1	Faible
1	1	Nul

Calcul de la probabilité post-test de la maladie

L'un des principaux intérêts des RV est qu'ils permettent de calculer la probabilité post-test de la maladie par application du théorème de Bayes, donc de répondre à la question : quelle est la probabilité que le patient soit malade si le test est positif ou sain si le test est négatif? **Les RV étant indépendants de la prévalence de la maladie dans la population**, ce calcul est transposable d'une population à une autre. Le théorème de Bayes introduit la notion de cote (ou *odd*).

L'*odd* est le rapport de la probabilité qu'un évènement se produise sur celle qu'il ne se produise pas, soit: ***Odd* = P/(1 − p)**

Un odd d'être malade égal à dix signifie que le risque d'être malade est dix fois supérieur à celui de ne pas l'être.
Le théorème de Bayes stipule que **l'*odd* post-test est égal au produit de l'*odd* pré-test par le rapport de vraisemblance du test, soit**:

***Odd* post-test = *odd* pré-test × RV.**
Le passage de l'*odd* post-test à la probabilité post-test s'effectue à l'aide de la formule:
Probabilité post-test = *Odd* post-test/ (*Odd* post-test + 1)

Les variations entre probabilité pré-test et probabilité post-test observées sont fonction de la valeur du rapport de vraisemblance. Des RV positifs supérieurs à dix et des RV négatifs inférieurs à 0,1 donnent généralement des variations importantes de la probabilité permettant le plus souvent de confirmer ou d'exclure la maladie. À l'inverse, des RV positifs de deux à cinq et négatifs de 0,2 à 0,5 donnent des variations minimes de la probabilité pré-test (Tableau 2).

L'application du théorème de Bayes permet également d'intégrer l'ensemble des résultats des tests biologiques réalisés pour un patient. Ainsi, l'*odd* post-test final est
obtenu en multipliant simplement l'*odd* pré-test par les RV des différents tests effectués, soit:

Odd post-test = Odd pré-test × RV test 1 × RV test 2 × RV test 3...

Les résultats des différents tests biologiques ne sont donc pas exploités séparément mais associés afin d'agréger leur performance diagnostique (Figure 6).
Si cette approche peut paraître « très mathématique », elle est réalisée de manière intuitive chaque jour par le médecin. L'interprétation des résultats d'un test repose sur des notions de concordance ou de discordance entre la clinique et le résultat. En cas de concordance et si les performances diagnostiques du test sont suffisantes, la décision d'intervention ou d'abstention peut être prise. En cas de discordance et/ou si les performances diagnostiques du test sont insuffisantes, d'autre(s) test(s) est (sont) nécessaire(s) pour aboutir à une décision clinique.

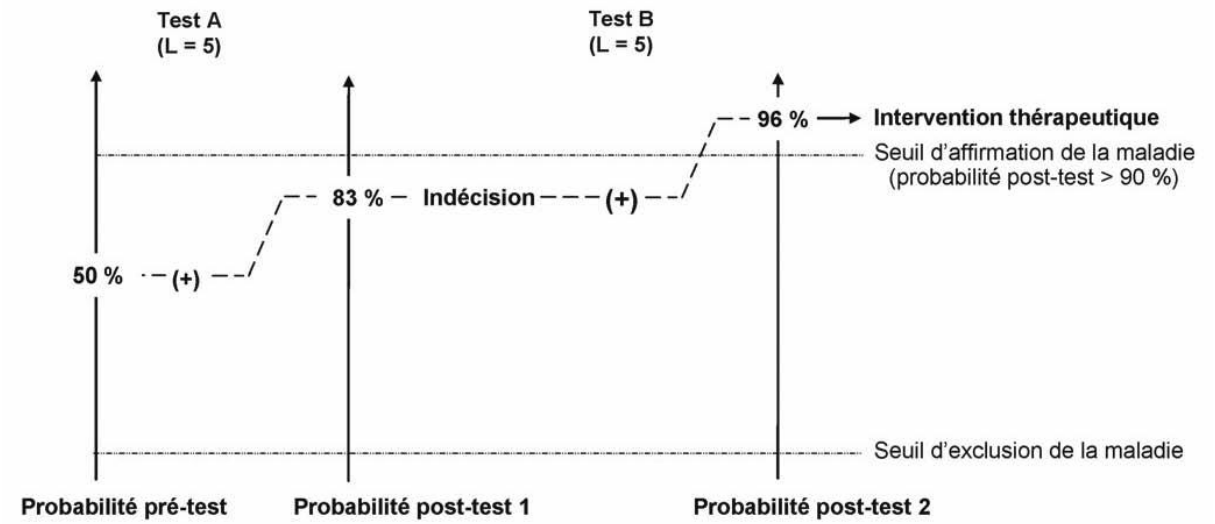


Figure 6 :

L'application du théorème de Bayes formalise cette approche et permet un calcul « objectif » de la probabilité post test.
Sa mise en œuvre, relativement fastidieuse avec les conversions de probabilité en *odds* et inversement, est facilitée par l'utilisation du **nomogramme de Fagan** comme l'illustre la Figure 7. Le nomogramme de Fagan permet un calcul rapide de la probabilité post-test en fonction de la probabilité pré-test et du rapport de vraisemblance d'un test. En traçant une droite qui part de la probabilité pré-test et qui passe par le rapport de vraisemblance du test, on peut lire la probabilité post-test de la maladie.

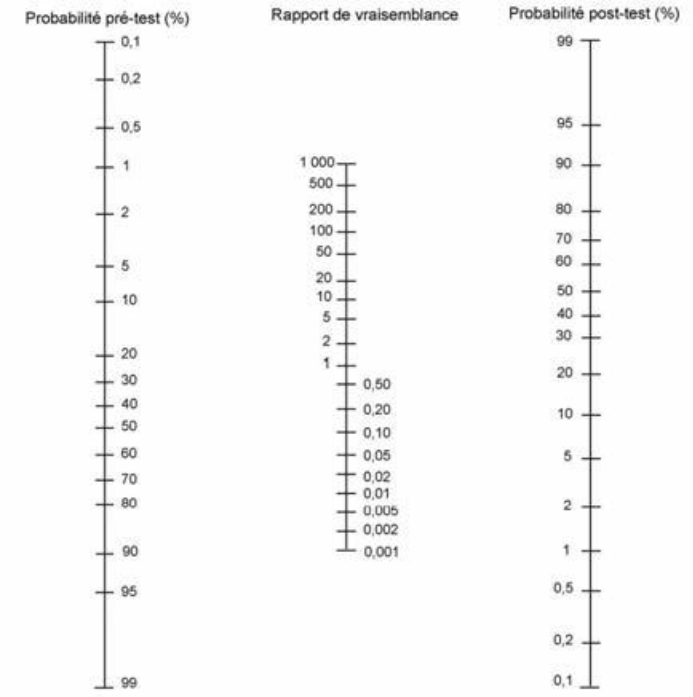
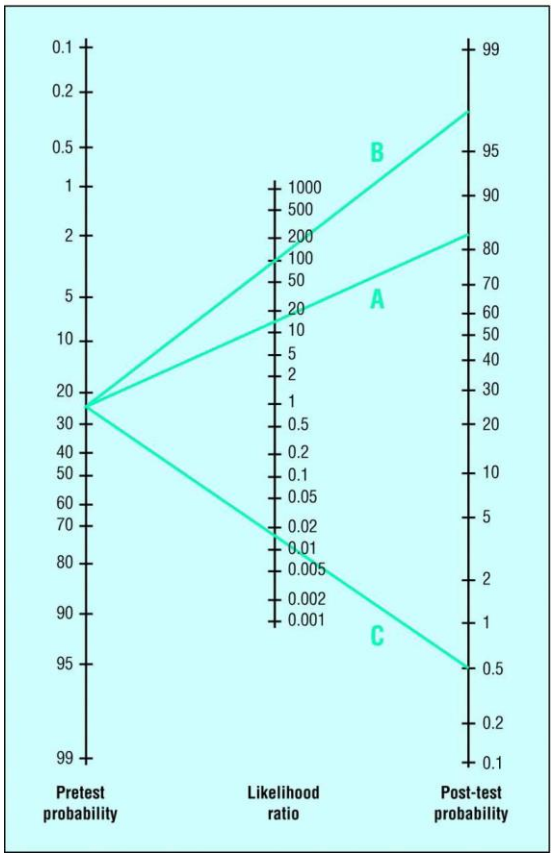


Figure : Nomogramme de Fagan.

Exemple :
Probabilité d'avoir un fumeur en face de soi
- Au départ 25% (prévalence anglaise)
- Si test B+ (LR+=100) P > 95%
- Si test A+ (LR+=15) P > 80% (moins bon test)
- Si test C+ (LR+=0.015) P = 0.5% (détecte les non-fumeurs !!!, mais très bon test)



7. Mode d'échantillonnage

1. Échantillon représentatif ++ :
• Réalisation des 2 tests (test de référence et nouveau test) auprès de tous les sujets sans connaître a priori le résultat de l'un ou de l'autre test. Tous les indices sont calculables.

2. Echantillonnage selon le résultat du nouveau test
• on fixe le nombre de sujets présentant un résultat positif au test et le nombre de sujets présentant un résultat négatif. On effectue ensuite le test de référence.
• les résultats du nouveau test sont fixés a priori ; pas de sens de calculer la sensibilité et la spécificité du nouveau test ; valeurs prédictives peuvent être calculées. Il s'agit d'une approche de type prédictif.

3. Echantillonnage selon le résultat du test de référence
• On peut convenir de fixer a priori le nombre de sujets malades et le nombre de sujets non-malades que l'on veut introduire dans l'étude.
• la sensibilité et la spécificité du nouveau test sont calculables. pas de sens de calculer les valeurs prédictives. Il s'agit d'une approche de type nosologique.
• Remarque : si l'on connaît la prévalence de la maladie dans la population, les valeurs prédictives pourront être calculées si dans le choix des effectifs de sujets malades et de non-malades, on respecte la proportion de malades de la population cible.

8. Les biais à recherché

-Biais d'observation : variabilité inter-observateurs et intra observateur selon l'examen réalisé et la connaissance du contexte ;
- Biais d'interprétation : connaissance du résultat de l'autre test (pas d'aveugle) ;
-Biais de spectre : profil clinique des sujets de l'échantillon d'étude différent de celui de la population cible ;
-Biais de sélection : critères d'inclusion dans l'étude non indépendants des résultats de certains tests ;
-Biais de vérification : résultat du nouveau test influençant la décision de réaliser celui de référence.

9. Bibliographie :

1. Thierry Ancelle : Statistiques épidémiologie. P 245-258. 3ème édition, Maloine.
2. H. Delacour, A. Servonnet, A. Perrot, J.F. Vigezzi, J.M. Ramirez : La courbe ROC (*receiver operating characteristic*) : principes et principales applications en biologie clinique. Ann Biol Clin 2005 ; 63 (2) : 145-54.

3. H. Delacoura, N. Francois, A. Servonneta, A. Gentile2, B. Roche : Les rapports de vraisemblance: un outil de choix pour l'interprétation des tests biologiques. da Immuno-analyse & Biologie spécialisée LigandAssay 14 (2) 2009.

4. A.BEZZAOUCHA : Précis d'épidémiologie pour lire et écrire des articles médicaux : chapitre 6. p 245 à 259.

1-Introduction :

Références bibliographiques :

- Précis d'épidémiologie **Pr BEZZAOUCHA Abdeljalil.**
- Ancelle T. Statistique Épidémiologie. Édition 2002.

Université Saad Dahlab Blida1

Faculté de Médecine

Département de Medecine

Responsable de l'enseignement : Dr BENILHA.S

Module d'Epidémiologie et Médecine préventive

1ème année résidanat d'épidémiologie et médecine préventive

Définitions préliminaires en Biostatistiques

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Définitions préliminaires en Biostatistiques

I-Définition de la statistique descriptive :

Ensemble des méthodes qui permettent de décrire les unités statistiques qui composent une population.

Buts de la statistique descriptive :

Toute série d'observations comporte un certain nombre de données relatives à un ou plusieurs caractères ou encore **variables.**

Le but des statistiques descriptives est de décrire un ensemble d'observations à l'aide de quelques éléments caractéristiques (tableaux, graphiques, numériques).

II-Concepts de base en statistique :

Population :

Désigne l'ensemble des entités ou des individus qui font l'objet de l'étude**.**

Echantillon :

Une partie (sous -ensemble) de la population dont l'effectif (la taille) est très réduit et qui permet d'effectuer l'étude expérimentale.

Individu ou unité statistique :

Unité de base sur laquelle l'observation réalise un certain nombre de mesures (une personne, un ménage, une ville,…).

III- Le caractère en statistique:

Contrairement à une constante -caractéristique ayant la même valeur pour tous les individus- **une variable** comporte nécessairement plus d'une modalité, les modalités sont les différentes catégories que peut présenter une variable.
Exemple : le « **groupe sanguin dans le système ABO** » est **une variable** car il comporte les modalités : groupe O, groupe A, groupe B, groupe AB.
Donc, chaque individu d'un échantillon ou d'une population sera décrit relativement à un ou plusieurs **caractères** (on dit aussi **variables**).

Types des caractères :

1-Variables qualitatives :

Une variable qualitative est un caractère dont les modalités s'expriment par des qualités et non pas par des valeurs numériques.

On dit d'un caractère qu'il est qualitatif, si ses diverses modalités ne sont pas mesurables.

Le « **groupe sanguin** dans le système ABO », le « **sexe** » ou « genre » avec ses deux modalités : masculin et féminin sont deux variables qualitatives,
Les variables qualitatives à **deux modalités** sont souvent désignées sous le vocable de « variables **dichotomiques** » parce que de type « l'un ou l'autre ».
Tandis que les variables comportant plus de deux modalités, comme le « groupe sanguin » sont appelées **multichotomiques.**

Classification des caractères qualitatifs :

***Nominale :** Ordre n'existe pas (ex: couleur des yeux,…).

Les variables multichotomiques dont les modalités ne sont pas soumises **à un tel ordre** sont appelées variables **nominales.**
Exemples :
-**groupe sanguin** dans le système ABO (O, A, B, AB).
-**statut matrimonial** (célibataire, marié, divorcé, veuf).

***Ordinale :** Peuvent- être **ordonnées**

Exemples :
- Stade de la maladie au moment du diagnostic : stades I à IV pour la maladie de Hodgkin.
- Gravité des lésions engendrées par les accidents de la route : mineure, modérée, sérieuse, sévère, critique.
- Intensité de la douleur : faible, modérée, importante.
- Niveau d'instruction : analphabète, primaire, secondaire, supérieur.

2-Caractère quantitatif:

Un caractère est dit quantitatif, si ses diverses modalités s'expriment par des valeurs numériques (**mesurables)**.

A chacune de ses modalités, on peut associer un nombre, ce nombre est appelé **variable statistique.**

Exemples : l'âge, le poids, la taille, la glycémie, le nombre de personnes par ménage, le temps, etc.

Classification des variables quantitatives :

***Variables discrètes** : quand la variable ne peut prendre que des valeurs isolées (nombre entier).

Exemple: nombre d'enfants par famille, nombre d'accidents du travail par ateliers, nombre de globules par unité de volume sanguin »,

***Variables continues** : Quand la variable est susceptible de prendre toute valeur appartenant à son intervalle de variation.

Une variable quantitative continue a des modalités en nombre infini qui se situent à un point quelconque d'une échelle numérique,

Exemples :

La glycémie, l'heure du jour, l'âge, la taille », la tension artérielle, le taux d'urée sanguine, le taux de cholestérol, le temps depuis l'instauration d'un traitement » ...

IV-Effectif :

L'effectif ou **fréquence absolue** est le nombre d'individus par classe.

Ce dénombrement donne lieu à une représentation des données sous forme de tableau.

Exemples :

L'effectif ou fréquence absolue est le nombre d'individus correspondant à une modalité donnée d'une variable.

Exemple 1:

La distribution de 50 malades selon le sexe, parmi ces 50 malades, 15 sont de sexe masculin et 35 de sexe féminin, les effectifs correspondant à chacune des deux modalités sont 15 et 35.

Exemple 2 :

On a dénombré sur un ensemble de 180 sujets, les individus qui appartenaient aux différents groupes sanguins (tableau).

Tableau : Description de l'échantillon des groupes sanguins.

A+	*A-*	*B+*	*B-*	*AB+*	*AB-*	*O+*	*O-*
80	*10*	*20*	*5*	*5*	*2*	*50*	*8*

V-Fréquence relative:

On peut définir les fréquences relatives (pourcentage ou proportion) qui sont, pour chaque classe, le rapport de son effectif au nombre total d'individus de la série des mesures.

Le rapport entre l'effectif d'une modalité de variable et l'effectif total de la série sur laquelle cette variable est mesurée. Le numérateur fait obligatoirement partie du dénominateur.

La somme des fréquences relatives est égale à 1.

Parfois, les résultats sont exprimés en pourcentage, chacune des fréquences relatives étant multipliée par 100. La fréquence relative s'exprime généralement en pourcentage parce que celui-ci est plus expressif.

Tableau : fréquences relatives d'un échantillon selon le groupe sanguin

A	B	AB	O
50	14	4	32

VI-Fréquences cumulées (relatives et absolues)

Les fréquences cumulées sont utilisées pour les données ordinales qui présentent des classes ordonnées.

Exemple :

Sur un échantillon de 500 malades cancéreux, on a noté le stade de la maladie.

On peut résumer ou présenter ces données par des fréquences relatives.

Les résultats obtenus sont présentés par le tableau.

Tableau : Répartition selon le stade de la maladie.

Stade	*Nombre de malades*	*Fréquence relative (%)*	*Fréquence relative cumulée (%)*
1	*350*	*70*	*70*
2	*110*	*22*	*92*
3	*30*	*6*	*98*
4	*10*	*2*	*100*

Interprétation : cette présentation permet de dire, par exemple, que **92%** des sujets examinés ont un stade inférieur ou égal à 2.

VII- Ratio :

Le Ratio : est le rapport entre les effectifs de deux modalités (classes) d'une même variable
Donc le numérateur et le dénominateur sont de même nature, mais exclusifs l'un de l'autre
C'est un nombre sans unité

Le sex-ratio : est le rapport numérique des sexes à la naissance, le sex-ratio = (effectif des garçons) / (effectif des filles).
Si le sex-ratio est égal à 1.06 (ou 106%), cela signifie que pour 100 naissances féminines, il y a 106 naissances masculines.
Lorsqu'on n'a plus affaire à des naissances, on peut parler de ratio hommes/femmes.

Exemple :
Dans une population de 100 individus, on observe 49 hommes et 51 femmes.
Quel est le ratio H/F (Sexe ratio) ? 49/51 = 0,96

VIII -Taux :

Le Taux : est un rapport qui prend en compte la notion de temps.

C'est La probabilité de survenue d'un évènement au cours du temps.

Le numérateur(N) : les individus ayant subis un évènement pendant une période de temps déterminé

Dénominateur (D): ensemble des individus susceptibles de connaître cet évènement pendant cette période.

Au (N) figurent des individus ayant subi un événement pendant une période de temps déterminé et au (D) figure l'ensemble des individus susceptibles de connaître l'événement pendant cette période (la population à risque).

Exemple :

Dans une population de 500 personnes,on a releve au cours de l'annee 2002,74 cas d'infarctus aigue du myocarde, le taux d'incidence est donc de 148/1000.

-la cantine d'une ecole recevant 250 enfants a été le siége d'une toxi-infection alimentaire collective.52 enfants ont présenté l'affection.

Le taux d'attaque est de 52/250 = 208 pour 1000.

Université Saad Dahlab Blida1
Faculté de Médecine
Département de Médecine

Responsable de l'enseignement : Dr BENILHA.S

Module d'Epidémiologie et Médecine préventive

Première année résidanat

PARAMETRES DE REDUCTION

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Paramètres de réduction

Introduction :

Afin de résumer une série statistique d'une façon simple tout en en conservant au mieux le contenu informationnel en limitant au maximum la perte d'information, on utilise:

A- Les paramètres de tendance centrale (de position):

Elles permettent d'obtenir une idée de l'ordre de grandeur des valeurs de la série et indiquent la position où semble se rassembler les valeurs de la série.

B- Les paramètres de dispersion:

Elles quantifient les fluctuations des valeurs observées et leur étalement.

Les paramètres de tendance centrale :

a-mode:

Le mode est la variable qui a l'effectif le plus grand:

1-Données non groupées :

Ex: série : 3,5,7,15,**16,16,16**,17,17,30 Mode = 16

2-Données groupées: (Variable quantitative discrète)

Le mode correspond à l'effectif maximal. Mode =4

Nombre d'enfant	Nombre de famille
0	4
1	5
2	10
3	16
4	18
5	14
6	7

3-Données groupées: (variable quantitative continue):

La classe modale correspond à l'effectif le plus élevé. Le mode correspond au centre de la classe modale. La classe modale : 3.5- 4

Le mode = (3.5+4)/2 = 3.75 Mode = 3.75 kg

POIDS en KG	effectif
2 - 2.5	2
2.5 - 3	4
3 - 3.5	6
3.5 - 4	30
4 - 4.5	8

B-médiane:

La médiane est une valeur de variable qui divise l'ensemble des observations en 2 parties égales ,50% de l'effectif se situe en dessous de la médiane et 50% de l'effectif se situe au dessus.

1-Données non groupées:

Nombre d'observation est impair :

Ex: 7, 9,13 , 45, 70, 101,115 Médiane = 45

Nombre d'observation est pair :

Ex: 2, 5, 9 , 10, 12, 14, 20, 22 Médiane = (10+12)/2=11

2-Données groupées :(variable quantitative discontinue),

La médiane est la valeur de la variable qui occupe le (n/2) éme rang

N=80 le (n/2) éme rang = 40

Nombre d'enfant	Nombre de famille	Effectif cumulé
0	4	4
1	5	9
2	10	19
3	16	35
4	18	53
5	14	67
6	7	74
7	6	80

Université Saad Dahlab Blida1

Faculté de Médecine

Département de medecine

Responsable de l'enseignement : Dr BENILHA.S

Module d'Epidémiologie et Médecine préventive

Première résidanat

Présentation tabulaire de l'information

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Représentations graphiques de l'information

I-Introduction :

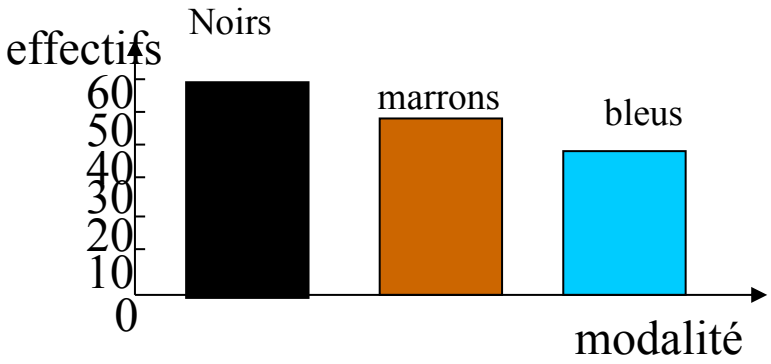
La représentation graphique est la première synthèse de l'information contenue dans un tableau statistique.

Le choix d'un type de graphe dépend de la nature des données du problème posé.

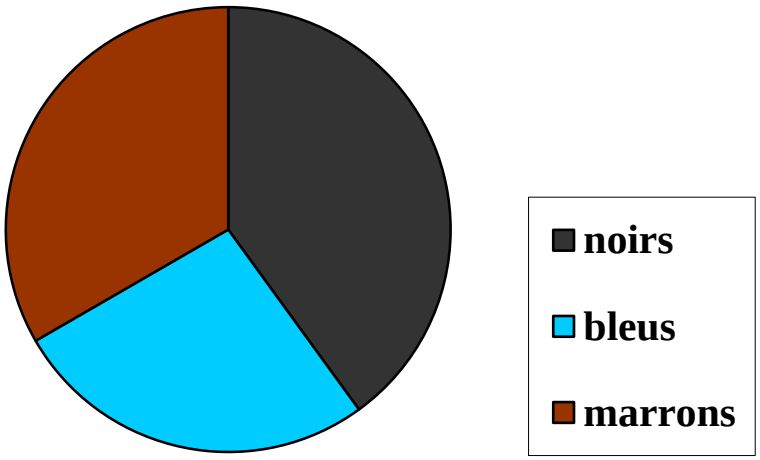
II-Représentation graphique :

1- Caractère qualitatif :

La représentation graphique par tuyaux d'orgue



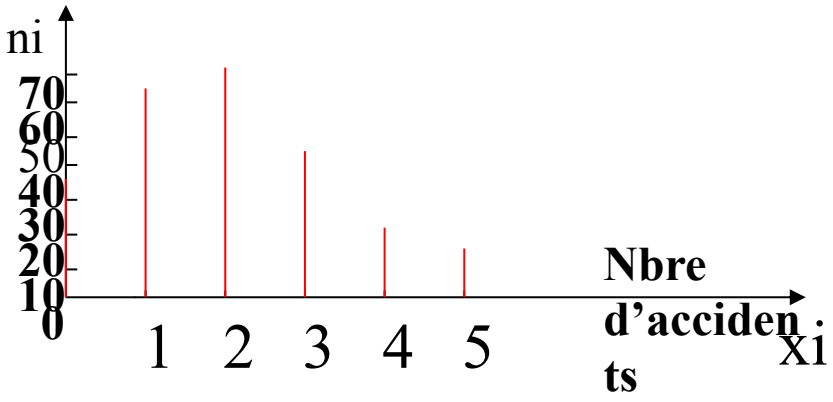
La représentation par secteur circulaire (nombre de modalités < 5)



2- Caractère quantitatif

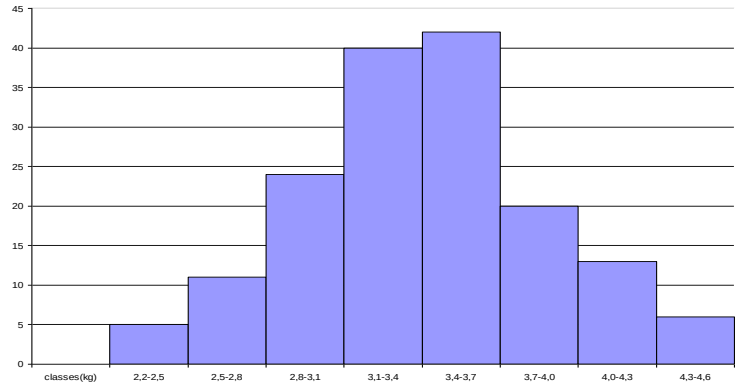
a-Variable statistique discrète : Diagramme en bâtons

Distribution des consultants en fonction du nombre d'accidents du travail



b-Variable statistique continue : Histogramme

La répartition du poids de 161 nouveau-nés



République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

Polycopié distribué aux résidents de première année résidanat d'épidémiologie et médecine préventive

Relation statistique entre deux variables
(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

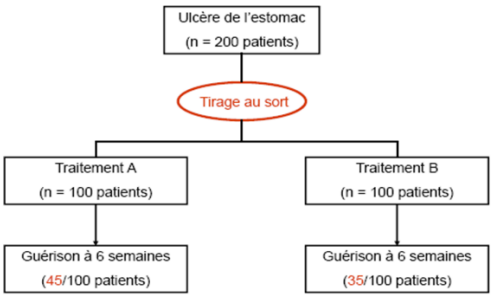
Relation statistique entre deux variables

I-Principe du test d'hypothèse

Introduction :

- Lorsqu'on effectue une comparaison entre deux ou plusieurs séries de données on observe toujours une différence entre les paramètres mesurés.
- Le but du test est de déterminer si la différence observée est simplement due au hasard, (fluctuations d'échantillonnage), ou au contraire la différence observée est bien réelle.

Exemple : Lequel de deux traitements est efficace ?



Que conclure ?

Les tests statistiques d'hypothèse permettent de se fixer une règle de décision objective

- Un test n'a de sens que s'il teste une hypothèse préalablement posée afin de répondre à une question.
- Tout test statistique doit donc avoir pour objectif de vérifier une hypothèse justifiée.

Observation → hypothèse → test

Principe des tests :

- Servent à comparer des séries de données entre elles ; deux situations :

-comparer un échantillon observé à une population de référence.

-comparer deux ou plusieurs échantillons entre eux.

- Le principe général d'un test est de regarder si la différence qu'on observe est due au hasard ou non.

Quelle que soit la nature d'un test son principe et sa chronologie sont toujours les mêmes

1- Etablir l'hypothèse nulle(H0):

Proposer H0 c'est de supposer que la différence observée provient seulement des fluctuation d'échantillonnage donc la différence n'est pas significative.

2- Proposer une hypothèse alternative :

On appelle hypothèse alternative H1 l'hypothèse qui sera retenue au cas où les résultats du test aboutirait à rejeter l'hypothèse nulle H0, donc lorsque la différence est significative.

3-calcul statistique :

Calculer une quantité mathématique X exprimant l'écart entre les paramètres ou les distributions (chaque test est sa formule).

Confronter cette quantité à un modèle de distribution théorique X` (table statistique de chaque test).

4- résultat d'un test

deux situations :

- la valeur de X < à X` (table) .On en conclut que la différence observée entre les paramètres étudiées n'est pas significative (fluctuations d'échantillonnage).

H0 est pas retenue

-la valeur de X ≥ à X`, donc la différence observée entre les paramètres étudiées est significative.

on rejette l'hypothèse nulle H0 et on accepte l'hypothèse H1.

5-Choix du risque d'erreur :

a) le risque α :

c'est le risque de se tromper en rejetant H0.

On l'appelle risque de première espèce ou risque α

=probabilité de rejeter H0 si H0 est vraie

on lui assigne communément la valeur 5%

b) le risque β :

risque β de deuxième espèce

= probabilité de ne pas rejeter H0,si H1 est vraie.

II_Test de chi deux

A-Définition :

Formulations équivalentes : Test de chi- deux = Test de chi- carré = Test de Pearson

Le principe général d'un test est de regarder si la différence qu'on observe est due au hasard ou si au contraire cette différence est telle qu'il est fort peu probable de l'observer par hasard

Quelle que soit la nature d'un test, son principe et sa chronologie sont toujours les

Mêmes.

B-Principe du test :

Le test de χ^2 permet de tester la liaison entre deux ou plusieurs distributions de caractères qualitatifs

C-Categories du test :

1-Comparaison d'une distribution observée et théorique

Il sert à comparer une distribution observée sur un échantillon à une distribution connue dans la population ou à une distribution théorique

Il s'agit d'un test statistique qui étudie l'écart entre la distribution théorique et la distribution observée

Exemple :

Dans une maternité, sur 100 naissances, on a observé 44 garçons et 56 filles ; cette observation est- elle compatible avec la statistique nationale donnant les proportions de naissances de garçons et de filles de respectivement 53% et 47% ?

Tableau de contingence:

	Garçon	Fille	Total
Statistique nationale	0.53	0.47	1
Effectif théorique (Ci/Ti)	53	47	100
Effectif observé	44	56	100

Démarche à suivre :

1- **Ho** :la distribution par sexe observée à la maternité est conforme à la distribution nationale

2-choix du test ➔ un test de χ^2

3-vérification des conditions d'application :

Tous les Ti ≥ 5 (53 et 47 sont > 5)

4-Calcul de la statistique (somme des écarts) :

$$\chi^2 = \sum \frac{(Ci - Oi)^2}{Ci}$$

$\chi^2 = (53-44)^2 / 53 + (47-56)^2 / 47 = 3.25$

5-on fixe le seuil de signification au risque **$\alpha = 5\%$**

6-on compare le χ^2 calculé au χ^2 de la table au risque $\alpha = 5\%$ et avec un nombre de degré de liberté **ddl = (l – 1)(c – 1)**

(k = nombre de modalités de la variable étudiée) ➔ k= 2-1=1

7-conclusion :

χ^2 calculé < **χ^2** de la table (3.25 < 3.84)

➔ **Ho n'est pas rejetée au risque** $\alpha = 5\%$

➔ La distribution observée est **conforme** à la répartition nationale par sexe des naissances

2-Comparaison entre plusieurs distributions observées :

Il sert à comparer deux ou plusieurs distributions observées sur des échantillons

Exemple :

Deux médicaments (A et B) ont été testés sur deux groupes de malades.

A l'issue de l'essai, on a observé les résultats suivants :

	Disparition des symptomes	Persistance des symptomes	aggravation	Réaction secondaire	Total
A	100	40	20	30	190
B	220	80	70	40	410

Total	320	120	90	70	600

Peut-on dire que ces deux traitements ont les mêmes effets ?
Correction :
1-H0 :
les deux traitements ont les mêmes effets ➔
il n'y a pas de différence entre les deux distributions ➔ les différences observées résulteraient des seules fluctuations d'échantillonnage.
2-tableau de contingence :

	Disparition des symptômes	Persistance des symptômes	aggravation	Réaction secondaire	Total
A	100 /**101.33**	40/**38**	20/**28.5**	30/**22.16**	190
B	220/**218.66**	80/**82**	70/**61.5**	40/**47.83**	410
Total	320	120	90	70	600

3-choix du test ➔ un χ^2
4-vérification des conditions d'application :

Tous les Ti ≥ 5

Calcul des Ti :

Ti = (total de la ligne × total de la colonne) / total general

(101.33- 218.66- 38- 82 - 28.5- 61.5- 22.16 et 47.83) sont tous > 5

5-Calcul de la statistique (somme des écarts) :

$$\chi^2 = \sum \frac{(Ci - Oi)^2}{Ci}$$

$$\chi^2 = (100 - 101{,}33)^2 / 101{,}33 + (220 - 218{,}66)^2 / 218{,}33 + (40 - 38)^2 / 38 +$$
$$(80 - 82)^2 / 82 + (20 - 28{,}5)^2 / 28{,}5 + (70 - 61{,}5)^2 / 61{,}5 +$$
$$(30 - 22{,}16)^2 / 22{,}16 + (40 - 47{,}83)^2 / 47{,}83$$

$$\chi^2 = 0{,}0175 + 0{,}0082 + 0{,}105 + 0{,}048 + 2{,}535 + 1{,}174 + 2{,}773 + 1{,}28$$

$$\chi^2 = 7{,}94$$

6-on fixe le seuil de signification au risque $\alpha = 5\%$

7-on compare le χ^2 calculé au χ^2 de la table au risque $\alpha = 5\%$ et avec un nombre de degré de liberté

$$ddl = (C-1) \times (L - 1)$$

C= nombre de colonnes

L= nombre de lignes

$ddl = (4-1) \times (2-1) = 3$

8-conclusion:

χ^2 calculé $> \chi^2$ de la table (7.94 > 7.81)

- ➔ Ho est rejetée
- ➔ La différence est significative au risque $\alpha = 5\%$
- ➔ les deux distributions **ne sont pas homogènes**. Ou

Les deux traitements ont des effets différents au risque

de 5%.

Avec α = **5 %** et **d.d.l = 3** χ^2 = **7,815**

3-Etude de la liaison entre les distributions de deux variables :

Il sert à étudier sur un même échantillon la liaison entre les distributions de deux variables

Exemple :

Un épisode d'intoxication alimentaire collective (TIAC) est survenu parmi les stagiaires d'un atelier, Le docteur chargé de l'enquête a dressé le tableau suivant croisant l'information sur la consommation de glace au chocolat, l'un des desserts proposés au cours du dernier repas pris en commun par les stagiaires, et le statut malade / non malade.

1-

Tableau de contingence :

Ho : il n'y a pas de liaison entre la consommation de glace au chocolat et la survenue de la gastro- entérite.

	Malade	Sain	Total
Glace au chocolat	270 (206.5)	7 (70.5)	277
Pas de glace	26 (89.5)	94 (30.5)	120
Total	296	101	397

Le principe et le calcul du test sont identiques

χ^2 calculé > χ^2 de la table (253.7 > 3.84)

- ➔ Ho est rejetée
- ➔ La différence est significative au risque α = 5%
- ➔ Il existe une liaison statistique entre la consommation de glace au chocolat et la survenue de la gastro- entérite au risque de 5%.

Université Saad Dahlab Blida1

Faculté de Médecine

Département de pharmacie

5ème année de Pharmacie

Responsable de l'enseignement : Dr BENILHA.S

Module d'Epidémiologie et Méthodologie de la recherche

Comparaison de pourcentages (Test de l'écart réduit (Z)

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Comparaison de pourcentages (Test de l'écart réduit (Z)

Plan du Cours

1-Introduction

2-Comparaison d'un pourcentage observé à un pourcentage théorique.

3-Comparaison de deux pourcentages observés : (Échantillons indépendants)

4-Comparaison de deux pourcentages observés : (Échantillons appariés)

Comparaison de pourcentages : (Test de l'écart réduit (Z)

1-Introduction :

Pourcentages : Variable qualitative dichotomique (Présence/Absence, Malades/Non malades,).

P: le pourcentage (inconnu) d'individus présentant la caractéristique dans la **population.**

p:le pourcentage observé sur un **échantillon** de taille n dont k individus présentent la caractéristique.

Types de Comparaison :

Il existe trois types de comparaison :

- **Comparaison d'un pourcentage observé à un pourcentage théorique**.
- **Comparaison de deux pourcentages observés : (Échantillons indépendants)**
- **Comparaison de deux pourcentages observés : (Échantillons appariés)**

2.Comparaison d'un pourcentage observé à un pourcentage théorique :

Intérêt : déterminer si un pourcentage **p** observé sur un échantillon de taille n'est pas différent d'une valeur théorique **P**.

Démarche :

➲ Comparer **p** à **P.**

➲ **1-Formuler les hypothèses :**

➲ **Hypothèse nulle H_0:**

$$p = P$$

P est le pourcentage de la population dont l'échantillon est issu.

➲ **Hypothèses alternatives H_1 : $p \neq P$**

2-Fixer le risque α à 5 %

3-Choisir la statistique:

Test z « Test de **l'écart réduit** » (loi normale centrée réduite)

4-Conditions d'application :

n. P ≥ 5 **et n. q** ≥ 5 **(N.B: q= 1-P)**

N.B:

P : pourcentage de la population

5-Calculer la statistique Z:

$$Z = \frac{|p - P|}{\sqrt{\frac{P * (1-P)}{n}}}$$

6-Comparer la valeur calculé de z avec la valeur critique de z_α (1,96):

_ si $|z| < z_\alpha$ **:**la différence n'est pas **significative** au risque α et donc **H_0 est retenue.**

_ si $|z| \geq z_\alpha$ **:** la différence est **significative** au risque α et donc H_0 est rejetée , **H_1 est retenue**.

Exemple :

En France, 7% des personnes hospitalisées contractent une infection nosocomiale dans l'établissement où elles sont soignées.

Sur un échantillon de 250 personnes soignées à l'hôpital H, 28 ont contracté une infection nosocomiale.

Le pourcentage observé sur l'échantillon diffère-t-il de la référence nationale au risque α = 5% ?

Démarches:

1.Poser les hypothèses :

H_0 : Le pourcentage observé sur l'échantillon **(p)** ne diffère pas de la référence nationale **(P)**, c.à.d : **p =P**

H_1 : Le pourcentage observé sur l'échantillon diffère de la référence nationale, c.à.d. : **$p \neq P$**

2.Détermination du risque α à 5 %:

3.Choix du test :Test de l'écart réduit(test Z)

4.Vérification des conditions d'application:

Sachant que : n= **250**, **P=0.07** **q**= 1-P = 1-0.07=**0.93**

np = 250 x 0,07 =**17,5** ≥ 5 , **nq** = 250 x 0,93=**232,5** ≥ 5

5. Calcul de la statistique Z:

$$Z = \frac{|p - P|}{\sqrt{\frac{P * (1-P)}{n}}} \qquad z = \frac{0,112-0,07}{\sqrt{\frac{0,07.(1-0.07)}{250}}} = 2,625$$

6.Comparaison et conclusion:

$z = 2,625 \geq z_{0,05} = 1,96$: rejet de H_0.

On montre, au risque 5%, une **différence significative** entre le pourcentage de personnes hospitalisées contractant une infection nosocomiale à l'hôpital H et dans l'ensemble du pays (**p < 0,01**).

2.Comparaison de deux pourcentages observés: Échantillons indépendants

Intérêt : comparer 2 proportions (p_1 et p_2) dans 2 groupes indépendants de tailles n_1 et n_2 :

Démarches:

1.Formuler une hypothèse:

Hypothèse nulle H_0 :

$p_1 = p_2$ (p_1 et p_2,pourcentages de la population dont sont issus les échantillons 1 et 2)

Hypothèses alternatives H_1:

$p_1 \neq p_2$

2.Fixer le risque α à 5 %

3.Choisir la statistique : Test de l'écart réduit (test z)

4.Conditions d'application :

- $n_1 . p_0 \geq 5$ et $n_1.(1- p_0) \geq 5$
- $n_2 . p_0 \geq 5$ et $n_2.(1- p_0) \geq 5$ Avec p_0

$$p_o = \frac{n_1.p_1 + n_2.p_2}{n_1 + n_2}$$

p_0 :pourcentage commune

5.Calculer la statistique Z :

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_0.(1-p_0)}{n_1} + \frac{p_0.(1-p_0)}{n_2}}} \qquad p_0 = \frac{n_1.p_1 + n_2.p_2}{n_1 + n_2}$$

6. Conclusion et prise de décision:

Confronter la valeur de z calculée à la valeur critique $z_{\alpha:}$

_ **Z<Zα** :**H_0 est retenue** la différence **n'est pas significative** au risque α .

_ **Z≥ Z α**: H_0 n'est pas retenue, donc **H_1 est retenue** la différence est **significative** au risque α.

Exemple

On désire comparer l'efficacité de deux traitements T1 et T2 sur 100 patients atteints d'une maladie M.

On tire au sort 2 deux groupes de 50 patients, un groupe est soumis à T1, le second à T2.

Le pourcentage de guérison chez les patients soumis à T1 est de 30%, chez ceux soumis à T2 de 40 %.

Le taux de guérison est-il différent entre les 2 traitements ?

Démarches :

1.Poser l'hypothèse nulle :

- H0: Le taux de guérison n'est pas différent entre les 2 traitements donc $\mathbf{p_1 = p_2}$.
- H1: Le taux de guérison est différent entre les 2 traitements donc $\mathbf{p_1 \neq p_2}$.

2.Détermination du risque α=5 %.

3.Choix du test : Test de l'écart réduit (Test Z).

4. Vérification des conditions d'application :

$n_1 = 50$; $p_1 = 0,3$ et $n_2 = 50$; $p_2 = 0,4$ $p_o = \frac{n_1.p_1 + n_2.p_2}{n_1 + n_2}$ $p_0 = \frac{50x0,3 + 50x0,4}{50 + 50} = 0,35$

$q_0 = 1 - p_0 = 1 - 0.35 = 0.65$

$n_1.p_0 = 50 \times 0,35 = 17.5 \geq 5$ et ……

$n_2.q_0 = 50 \times 0,65 = 32.5 \geq 5$

5. Calcul de la variable testée:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_0.(1-p_0)}{n_1} + \frac{p_0.(1-p_0)}{n_2}}} \qquad z = \frac{0,3 - 0,4}{\sqrt{\frac{0,35x0,65}{50} + \frac{0,35x0,65}{50}}} = 1,05$$

$z = 1,05 < z_{0,05} = 1,96$: H_0 est retenue

6.Conclusion : il y'a pas de différence significative entre les taux de guérison avec les 2 traitements, au risque 5%**.**

3.Comparer deux pourcentages observés - Séries appariées -

Intérêt : on s'intéresse aux taux de guérison chez des sujets ayant reçus un traitement T1 et des sujets **appariés** ayant reçus un traitement T2.

on cherche à comparer p_1 et p_2 (les taux de guérison avec T1 et T2).

		Éch. 2		
		+	-	Total
Éch. 1	+	a	b	a+b
	-	c	d	c+d
	Total	a+c	b+d	**n**

Les paires concordantes n'apportent pas d'information sur la liaison entre le traitement et la guérison. On doit donc se fonder sur la répartition des paires discordantes.

Conditions d'application : b+c ≥ 10

Calcul de la statistique Z:

$$z = \frac{b - c}{\sqrt{b + c}}$$

Comparaison et conclusion :

_ **Z<Zα** : H0 est retenue la différence n'est pas significative au risque α .

_ **Z≥ Z α**: H0 n'est pas retenue, donc la différence est significative au risque α.

Exemple

On désire comparer l'efficacité de deux traitements T_1 et T_2 chez 100 patients atteint d'une maladie M.

Les deux traitements sont administrés aux patients. L'ordre d'administration des 2 traitements est tiré au sort en ménageant une période dite de *Wash-out* entre les 2 administrations.

Les résultats sont les suivants :

		T_1	
		Succès	Échec
T_2	Succès	24	16
	Échec	6	54

Le taux de guérison est-il différent entre les deux traitements ?

On cherche à comparer les pourcentages observés :

$$p_1 = \frac{24 + 6}{100} = 0{,}3 \qquad p_2 = \frac{24 + 16}{100} = 0{,}4$$

1.Hypothèses :

H_0 : Le taux de guérison n'est pas différent entre les deux traitements ($p_1 = p_2$)

H_1 : Le taux de guérison est différent entre les deux traitements ($p_1 \neq p_2$)

2.fixer le risque à 5 %

3.choix du test :test de l'écart réduit (séries appariées)

4.conditions d'application vérifiées :

nombre de paires discordantes = 16 + 6 = 22 ≥ 10

$$Z = \frac{b - c}{\sqrt{b + c}} = \frac{16 - 6}{\sqrt{16 + 6}} = 2.13$$

6.Conclusion:

Z=2.13 > 1,96 donc

On montre, au risque 5%, une différence significative entre les taux de guérison avec les 2 traitements ($p<0,05$).

Travaux dirigés :

Exercice 01 :

L'étude expérimentale d'un médicament a été pratiquée sur **100**malades divisés par tirage au sort en deux groupes A et B.
Le groupe A composé de **60** malades a absorbé le médicament étudié.
Le groupe B composé de **40** malades n'a absorbé qu'un produit inactif << placebo>> extérieurement identique au médicament.
Les résultats sont les suivants :

Groupe A : **40** malades guéris.

Groupe B : **20** malades guéris.

Peut- on conclure à l'efficacité de médicament ?

Exercice 02 :

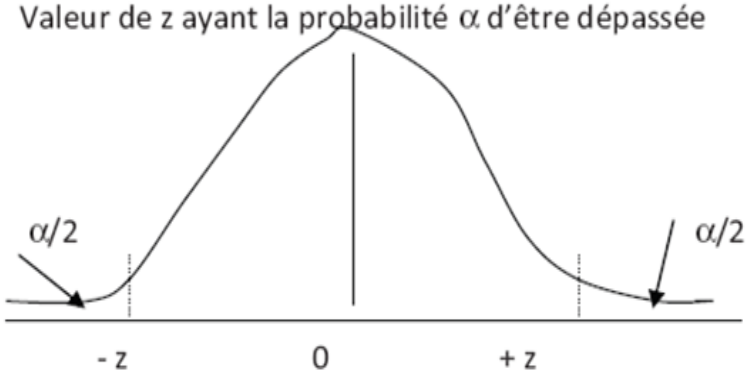
On a observé dans une maternité sur un échantillon de **400** naissances dont les mamans font un travail pénible que le nombre des naissances prématurées est de **52**.
On admet que la proportion des naissances prématurées est de **8%**.

Le travail pénible est-il cause des naissances prématurées ?

Table de l'écart réduit z

Table des aires limitées par la courbe N (0, 1)

Valeur de z ayant la probabilité α d'être dépassée



α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	∞	2.576	2.326	2.170	2.054	1.960	1.881	1.812	1.751	1.695
0.1	1.645	1.598	1.555	1.514	1.476	1.440	1.405	1.372	1.341	1.311
0.2	1.282	1.254	1.227	1.200	1.175	1.150	1.126	1.103	1.080	1.058
0.3	1.036	1.015	0.994	0.974	0.954	0.935	0.915	0.896	0.878	0.860
0.4	0.842	0.824	0.806	0.789	0.772	0.755	0.739	0.722	0.706	0.690
0.5	0.674	0.659	0.643	0.628	0.613	0.598	0.583	0.568	0.553	0.539
0.6	0.524	0.510	0.496	0.482	0.468	0.454	0.440	0.426	0.412	0.399
0.7	0.385	0.372	0.358	0.345	0.332	0.319	0.306	0.292	0.279	0.266
0.8	0.253	0.240	0.228	0.215	0.202	0.189	0.176	0.164	0.151	0.138
0.9	0.126	0.113	0.100	0.088	0.075	0.063	0.050	0.038	0.025	0.013

Exemple : si $\alpha(z) = 0.23$, $z = -1.200$

Table pour les petites valeurs de $\alpha(z)$

α	10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
z	3.291	3.891	4.417	4.892	5.327	5.731	6.109

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Premiere année de residanat

ANALYSE DE LA VARIANCE
(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

Analyse de la variance

Plan du cours :

1. **Position du problème**
2. **Principe du test d'ANOVA :**
3. **Procédure du test**
4. **Comparaison de deux variances**

ANALYSE DE LA VARIANCE :

1-Position du problème :

Comparaison de deux moyennes

→ test de l'écart réduit ou test T

Comparaison de plusieurs moyennes ?

Solution :

a-Comparaison de moyennes deux à deux :

Inconvénients : - méthode longue

-Vue d'ensemble flou

b-Analyse de la variance (terme souvent abrégé par le terme **ANOVA** : ***ANalysis Of VAriance***) :bon reflet de dispersion autour de la moyenne.

- **Test approprié** : test de **Fisher-Snedecor**

2-Principe de test d'ANOVA :

Il repose sur la **décomposition de la variation de la variable** X

a) La dispersion entre les classes (variation inter-groupe)

(D inter) Chaque classe est caractérisée par sa moyenne (X1 , ..., Xk) qui s'écarte plus ou moins de la moyenne générale (X).

b) La dispersion à l'intérieur de chaque classe (intra-groupe ou celle due au hasard)(D intra)

On test le rapport de deux variances :

1- variance **inter-groupes(** variance factorielle **Vf)**

2- variance **intra-groupes(** variance résiduelle **Vr)**

-**Ho:** teste donc l'hypothèse de l'homogénéité des k moyennes.(la différence entre les moyennes n'est pas significative)

-On dit aussi que le facteur (à partir duquel on a construit les k groupes) n'a pas d'influence sur la variable X

Formule pratique $D\ inter = n1(\overline{X}\ 1)^2 + n2(\overline{X2})^2 + \ldots\ldots - N\ (\overline{X})^2$

Xi = les moyennes respectives des K échantillons

X = moyenne générale = ΣXi/ n1+n2+n3+…..+nk

N=n1 + n2+ n3+……..+nk

1-variance factorielle Vf

Vf = Dinter / k-1

K: nombre d'échantillons

Dinter: Dispersion inter-groupes: analyse la variabilité **entre** les groupes

$$Dinter = n1 (X1 - X)^2 + n2 (X2-X)^2 + n3(X3 - X)^2 + ... + nk (Xk - X)^2$$

2- variance résiduelle Vr :

Vr = D intra /N-k

N= somme des tailles de tous les échantillons

Dintra: Dispersion intra-groupes :analyse la variabilité **à l'intérieur** des groupes

$$Dintra = \Sigma(Xi1 - \overline{X1})^2 + \Sigma(Xi2-\overline{X2})^2 + \Sigma(Xi3-\overline{X3})^2 + \ldots + \Sigma(Xik-\overline{Xk})^2$$

Formule pratique :

$$Dintra = \Sigma X^2i - (n1.\overline{X^2}1 + n_2.\overline{X^2}2 + \ldots + nk.\overline{X^2}k)$$

$$\Sigma x^2i = \Sigma xi^2 1 + \Sigma xi^2 2 + \Sigma xi^2 3 + ... + \Sigma xi^2 k$$

4-Procédure du test

1)Détermination du risque α

2)Formulation des hypothèses H0et H1

3) vérification des conditions d'application

4)Calcul de la dispersion inter –groupe

5)Calcul de la dispersion intra –groupe

6)Calcul de la variance factorielle Vf

7)Calcul de la variance résiduelle Vr

Conditions d'utilisation

1-L'ensemble des N individus est réparti au hasard (randomisation).

2-Normalité de la distribution des mesures.

3- L'égalité des variances.

Exercice :

5 milieux de culture de BCG (K)

10 tubes par milieu de culture (n1=n2=..=10; N=50)

on observe le nombre de colonies par tubes (xi)

Ces milieux sont-ils équivalents?

Milieu de culture	**1**	**2**	**3**	**4**	**5**
	10	11	7	12	7
	12	18	14	9	6
	8	12	10	11	10
	10	15	11	10	7
	6	13	9	7	7
	13	8	10	8	5
	9	15	9	13	6
	10	16	11	14	7
	8	9	7	10	9
	9	3	9	11	6
Moyenne	**9.5**	**13.0**	**9.7**	**10.5**	**7.0**

Corrigé :

1) $\alpha = 5\%$

2)H0 : les milieux sont équivalents

3) Dinter = n1 $(x1)^2$+n2$(x2)^2$ +……- N $(x)^2$=10$(9.5)^2$+10$(13)^2$+....-50$(9.94)^2$=**185.7**

4)Dintra = Σx^2i - (n1. x^21+n2. x^22+….nk. x^2k) =5083-(10 . $9,5^2$+10.13^2+.....+10.7^2)=**225.1**

5)Vf=185.7/5-1=**46.4**

6)Vr=225.1/50-5=**5**

7) conclusion :Vf /Vr=46.4/5=**9.3**

8) F de la table=**2.56** (ddl1=4 et ddl2=45)

le rapport est supérieur à F critique ;H0 est rejetée.

Lois de Fisher–Snédécor (0, 05)

4-Comparaison de deux variances

Intérêt :

1) comparer deux moyennes dans le cas des petits échantillon (la variable étudiée doit être normale et les variance des deux populations ne doivent pas être significativement différentes)

2) comparer la précision de deux méthodes

Principe:

Comparaison du rapport $F=S^2A/S^2B$ (S^2A est supérieure à S^2B**)** et F de la table de Fisher.

F critique: intersection de ddl=nA-1 et ddl=nB-1

Procédure:

Les étapes à suivre sont:

1) Détermination du risque α
2) construction de l'hypothèse nulle
3) Calcul de la variance testée **$F=S^2A/S^2B$**
4) Détermination de la valeur critique F de la table de Fischer, en fonction du risque α et les degrés de liberté V1=n1-1 et V2=n2-1

5) Conclusion :

5) Fc≥ Ft la différence est significative au risque α et H0 est rejetée
6) Fc < Ft la différence n'est pas significative au risque α et H0 est retenue

Exemple:

	A	B
	37	38
	36	39
	37	38
	38	37
	37	36
		39
		39
		38
		37
Moyenne	**37**	**38**
Variance	**0.5**	**1.11**

F calculé=**1,11 /0.5 =2.22**
F table=**6.04** (intersection de 8 et 4)
Fc < Ft pas de différence significative entre les deux variances

Lois de Fisher–Snedecor (0, 05)

ν_2 \ ν_1	1	2	3	4	5	6	8	10
1	161	200	216	225	230	234	239	242
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4
3	10,1	9,55	9,28	9,12	9,01	8,94	8,85	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,35

Lois de Fisher–Snedecor (0, 05)

ν_2 \ ν_1	1	2	3	4
1	161	200	216	225
2	18,5	19,0	19,2	19,2
3	10,1	9,55	9,28	9,12
4	7,71	6,94	6,59	6,39
5	6,61	5,79	5,41	5,19
6	5,99	5,14	4,76	4,53
7	5,59	4,74	4,35	4,12
8	5,32	4,46	4,07	3,84
9	5,12	4,26	3,86	3,63
10	4,96	4,10	3,71	3,48
11	4,84	3,98	3,59	3,36
12	4,75	3,89	3,49	3,26
13	4,67	3,81	3,41	3,18
14	4,60	3,74	3,34	3,11
15	4,54	3,68	3,29	3,06
16	4,49	3,63	3,24	3,01
17	4,45	3,59	3,20	2,96
18	4,41	3,55	3,16	2,93
19	4,38	3,52	3,13	2,90
20	4,35	3,49	3,10	2,87
22	4,30	3,44	3,05	2,82
24	4,26	3,40	3,01	2,78
26	4,23	3,37	2,98	2,74
28	4,20	3,34	2,95	2,71
30	4,17	3,32	2,92	2,69
40	4,08	3,23	2,84	2,61
50	4,03	3,18	2,79	2,56
60	4,00	3,15	2,76	2,53

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

Polycopié distribué aux résidents de première année résidanat d'épidémiologie et médecine préventive

Corrélation et régression linéaire simple

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Corrélation et régression linéaire simple

Objectifs du cours

Connaitre la définition de la corrélation et de la régression linéaire

Apprendre la différence entre la corrélation et de la régression linéaire

Connaitre les conditions d'application de la corrélation et de la régression linéaire simple

Appliquer les formules de calcul de différents coefficients de la corrélation et de la régression.

Plan du cours

1-Definition de la corrélation

2- Corrélation versus régression

3-Conditions d'application de la corrélation et de la régression linéaire simple

4-Corrélation linéaire simple :

a-Estimation du coefficient de corrélation

b-Test du coefficient de corrélation :

5-Regression linéaire simple :

a -Estimation par la méthode des moindres carrés

b- Test de la pente de la droite de régression

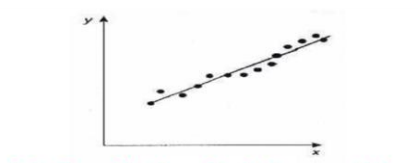
Corrélation et régression linéaire simple :

1-Définition de la corrélation :

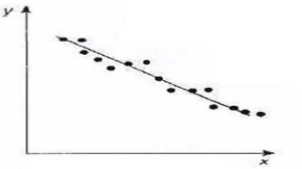
En statistique, le terme de corrélation désigne la liaison entre 2 variables **quantitatives** (le plus souvent continues).

Corrélation / régression : liaison entre 2 variables quantitatives

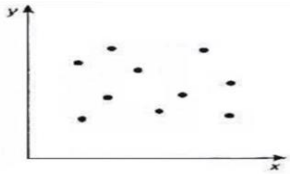
CORRÉLATION LINÉAIRE :
La corrélation est une mesure de la quantité d'association existant entre deux variables.
La corrélation est dite linéaire, si tous les points, expérimentaux tracés (nuage de points) sur un graphique se trouvent sur une ligne droite.



(a) Corrélation linéaire positive



(b) Corrélation linéaire négative



(c) Pas de corrélation

1. Exemple : corrélation (positive)

• X = ventes de paires de lunettes de soleil en été

• Y = ventes de crèmes glacées en été

• Il existe une liaison entre X et Y :

– Quand X augmente, Y augmente

– Quand X diminue, Y diminue

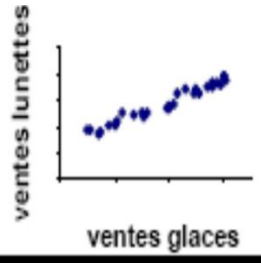
• La liaison est symétrique :

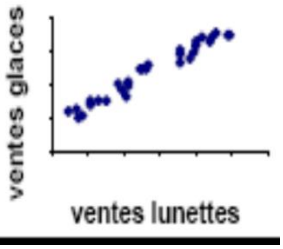
– X est liée à Y, et Y est liée à X

– mais X ne dépend pas de Y et Y ne dépend pas de X

– on peut permuter X et Y en abscisses et en ordonnées

• Y ne peut pas être prédite par X





2. Exemple : corrélation (négative)

• X = ventes de paires de lunettes de soleil en été

• Y = ventes de parapluies en été

• Il existe une liaison entre X et Y :

– Quand X augmente, Y diminue

– Quand X diminue, Y augmente

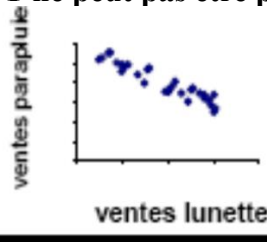
• La liaison est symétrique :

– X est liée à Y, et Y est liée à X

– mais X ne dépend pas de Y et Y ne dépend pas de X

– on peut permuter X et Y en abscisses et en ordonnées

• Y ne peut pas être prédite par X



2-Corrélation versus régression

Corrélation :

• Liaison entre 2 variables quantitatives X et Y

• Rôle **symétrique** (on peut permuter X et Y)

• Rôle asymétrique

Régression :

• Liaison entre 2 variables quantitatives X et Y

• Rôle **asymétrique** uniquement :

– X = variable explicative / Y = variable expliquée

– X = variable indépendante / Y = variable dépendante

• (on ne peut pas permuter X et Y)

3. Exemple : régression

• X = âge (de 0 à 15 ans)

• Y = taille (cm)

• Il existe une liaison entre X et Y :

– Quand l'âge augmente, la taille augmente

– Quand l'âge diminue, la taille diminue

• **La liaison est asymétrique :**

– la taille dépend de l'âge mais l'âge ne dépend pas de la taille

– on ne peut pas permuter X et Y en abscisses et en ordonnées

• On peut **prédire** la taille par l'âge à l'aide d'une équation de droite ou de courbe de régression (cf carnet de santé)

	Corrélation	Régression
Variables	X = quantitative Y = quantitative	X = quantitative Y = quantitative
Symétrie de la liaison	Oui / Non Y liée à X X liée à Y	Non Y dépend de X -
Exemples	Y = conso. cannabis X = température moyenne annuelle	Y= taille X = âge
Prédiction	Non	Oui (équation)

3-Conditions d'application de la corrélation et de la régression linéaire simple

Indépendance des observations

Liaison linéaire entre X et Y

Distribution conditionnelle normale et de variance constante

Liaison linéaire entre X et Y

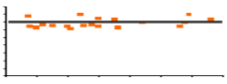
Avant d'appliquer le test du coefficient de corrélation ou d'estimer la droite de régression, il faut vérifier -empiriquement (graphiquement)que la liaison entre les 2 variables est de nature linéaire.

Coefficient de corrélation nul

Pente de la droite de régression nulle

Cas 1

La nature de la liaison est linéaire (le nuage de points est résumé au mieux par une droite horizontale d'équation y = a)

La condition d'application est vérifiée 

Il est possible d'utiliser le coefficient de corrélation et la régression linéaire simple pour quantifier la liaison entre les 2 variables

4-Corrélation linéaire simple :

a- Estimation du coefficient de corrélation

Le coefficient de corrélation estimé sur échantillon issu d'une population est noté r.
Il s'interprète comme le coefficient de corrélation ρ mesuré sur la population.
Il est calculé à partir des estimations de la covariance et des variances de X et de Y sur L'échantillon.

L'expression de l'estimateur du coefficient de corrélation r à partir d'un échantillon.

$$r = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right]}}$$

b-Coefficient de corrélation dans la population :

Le coefficient de corrélation entre 2 variables quantitatives X et Y est égal au rapport de la covariance de X et Y divisé par le produit des écart- types de X et Y.
Le coefficient de corrélation est noté ρ dans la population.

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}}$$

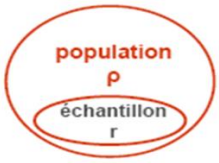
$$-1 \leq \rho \leq +1$$

Interprétation du coefficient de corrélation

- $-1 \leq r(x; y) \leq 1$
- $r(x; y) > 0 \rightarrow$ relation linéaire croissante
- $r(x; y) < 0 \rightarrow$ relation linéaire décroissante
- $r(x; y) = 0 \rightarrow$ pas de relation linéaire
- $r(x; y) = \pm 1 \rightarrow y = \beta 0 + \beta 1 x$

c-Test du coefficient de corrélation :

Après le calcul du coefficient de corrélation r estimé sur un échantillon, il faut déterminer si le coefficient de corrélation ρ est Significativement différent de 0.



$r \approx \rho$

H0 : ρ = 0 (absence de liaison [linéaire] entre X et Y)

H1 bilatérale : ρ ≠ 0 (existence d'une liaison entre X et Y)

Sous l'hypothèse nulle (H0) :

Le rapport de l'estimateur du coefficient de corrélation r sur son Écart-type suit une loi de Student à (n-2) degrés de liberté.

n :est l'effectif de l'échantillon.

$$\frac{r}{S_r} \rightarrow t_{(n-2)ddl}$$

L'estimateur de l'écart-type du coefficient de corrélation est égal à

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

Le test du coefficient de corrélation consiste à calculer la grandeur To et à la comparer à la valeur seuil tα sur la table de la loi de Student à (n-2) degrés de libertés.

$$t_o = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Conditions d'application

• indépendance des observations

• liaison linéaire entre X et Y

• distribution conditionnelle normale et de variance constante

5- Régression linéaire simple

La régression s'adresse à un type de problème où les 2 Variables quantitatives continues X et Y ont un rôle asymétrique : la variable Y dépend de la variable X. La liaison entre la variable Y dépendante et la variable X Indépendante peut être modélisée par une fonction de type $Y = \alpha + \beta X$, représentée graphiquement par une droite.

$$\mathbf{Y = \alpha + \beta X}$$

Y : variable dépendante (expliquée)

X : variable indépendante (explicative)
α : ordonnée à l'origine (valeur de Y pour x = 0)
β : pente (variation moyenne de la valeur
de Y pour une augmentation d'une unité de X)

a -Estimation par la méthode des moindres carrés

Chaque individu i est caractérisé par un couple de coordonnées (xi,yi) et est représenté par un point sur le graphique.
L'ensemble des individus forme un nuage de points.
a et b sont les estimations de l'ordonnée à l'origine α et de la pente β de la droite de régression**.**
L'estimation de la pente de la droite de régression b est égale au rapport de la covariance de X et Y sur la variance de X.

$$b = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)} \qquad b = \frac{\sum_{i=1}^{n}(x_i - m_x)(y_i - m_y)}{\sum_{i=1}^{n}(x_i - m_x)^2}$$

L'estimateur de l'ordonnée à l'origine a est déduit de la pente
b et des coordonnées du point moyen (mx, my) :

a = my – b mx

b- Test de la pente de la droite de régression

La droite de régression d'équation Y = α + βX comporte 2 paramètres (α et β).
L'hypothèse nulle est que la pente β de la droite de Régression de Y en X est égale à 0 (soit Y est égal à α, c'est-à-dire que la droite de régression est horizontale et qu'il n'y a pas de liaison entre X et Y).
H0 : β = 0 (droite de régression horizontale : Y = α)
H1: β ≠ 0
Sous l'hypothèse nulle (H0) :
Le rapport de l'estimateur de la pente b sur son écart-type suit une loi de Student à (n-2) degrés de liberté.
n: est l'effectif de l'échantillon.

$$\frac{b}{s_b} \rightarrow t_{(n-2)ddl}$$

L'estimateur de l'écart-type de la pente est égal à :

$$s_b = \sqrt{\frac{\frac{s_y^2}{s_x^2} - b^2}{n-2}}$$

Le test de la pente consiste à calculer la grandeur to et à la comparer à la valeur seuil tα sur la table de la loi de Student à (n-2) degrés de libertés

$$t_o = \frac{b}{\sqrt{\frac{\frac{s_y^2}{s_x^2} - b^2}{n-2}}}$$

Conditions d'application

- Indépendance des observations
- Liaison linéaire entre X et Y
- Distribution conditionnelle normale et de variance constante

Conclusion

La corrélation et la régression linéaire simple sont deux tests fréquemment utilisés en médecine afin de tester la relation linéaire entre deux variables quantitatives.

Les conditions d'applications doivent être vérifiées avant d'appliquer les deux tests.

Bibliographie :

- BEZZAOUCHA Abdeljalil, épidémiologie et Biostatistique, à l'usage des étudiants en sciences médicales, 3éme édition OPU.
- Ancelle T. Statistique Épidémiologie. Édition 2002.
- BEZZAOUCHA Abdeljalil, Tests statistiques en sciences médicales, OPU 2004.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

Polycopié distribué aux résidents de première année résidanat d'épidémiologie et médecine préventive

SONDAGE

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

Plan du cours :

I-Introduction

II-Définition de l'échantillon

III-Critères de choix d'un échantillon

IV- Choix d'un plan d'échantillonnage

V-Les méthodes d'échantillonnage

VI – Méthodes d'échantillonnage aléatoire simple :

VII- Les biais dans la réalisation de l'échantillon

VIII-Conclusion

XI-Références bibliographiques

SONDAGE

I-Introduction :

L'échantillonnage, lorsque bien fait, permet de mesurer des caractéristiques sur un nombre restreint d'unités du groupe d'intérêt et d'arriver à une estimation des paramètres à l'étude qui sera non seulement précise et exempte de biais, mais aussi représentative de l'ensemble des unités du groupe. On entend par paramètre une caractéristique quantifiable de la population dont la valeur est fixe au sein d'une région et d'une période de temps donnée, mais qui demeure inconnue.

Dans une étude empirique, qu'elle soit quantitative ou qualitative, l'échantillonnage permet de récolter des informations pertinentes à partir d'un public cible, afin de répondre à la problématique et aux hypothèses de départ.

II-Définition de l'échantillon

L'échantillonnage est un procédé qui permet de définir un échantillon dans un travail d'enquête. Il s'agit d'étudier une partie sélectionnée pour établir des conclusions applicables à un tout. En d'autres termes, l'échantillonnage est une sélection précise de personnes ciblées pour réaliser un entretien, un sondage ou un questionnaire.

III-Critères de choix d'un échantillon :

Toute action à réaliser par un chercheur pour mener à bien une étude se place dans quatre grandes catégories :

- **Définition de l'objectif :** La première étape de la conception d'un échantillonnage consiste à définir les objectifs de l'étude à mener, les buts de la collecte des données.

- **Variables (descripteurs) :** Les variables, appelées aussi descripteurs, désignent toute caractéristiques mesurable ou observable sur chacun des éléments de l'échantillon ou sur son environnement.

- **Échelles d'observation :** une échelle d'observation réfère à l'étendue (surface, durée) et à la résolution (taille de l'unité élémentaire) des observations dans l'espace et dans le temps.

- **Choisir le plan d'échantillonnage :** Choisir le plan d'échantillonnage consiste à choisir de quelle manière les données seront recueillies sur le terrain. Il conditionne aussi le mode de traitement des données et donc les résultats.

- **Interprétation des résultats :** à partir de comparaisons, de recherche des causes, et réponse à l'objectif défini au départ.

IV- Choix d'un plan d'échantillonnage

Le plan d'échantillonnage définit la manière dont les échantillons élémentaires sont répartis sur le terrain étudié (et éventuellement au long de la saison ou des années).

Il est conçu de manière à ce que l'échantillon sélectionné représente aussi fidèlement que possible l'ensemble du milieu étudié. Ce plan est inutile quand aucune extrapolation des données recueillies n'est nécessaire et notamment quand le site concerné est suffisamment petit pour être étudié en entier.

Le plus souvent, et toujours lorsqu'on emploie des méthodes indiciaires, l'échantillon est fractionné en un certain nombre d'unités d'échantillonnage. Pour la meilleure exploitation statistique des données, ces unités d'échantillonnage doivent être standardisées, restant identiques aussi bien dans l'espace la même année qu'au cours du temps entre années.

Définir le nombre d'échantillons

Le nombre d'échantillons peut être défini dans le temps et dans l'espace : des relevés peu fréquents (annuels par exemple) mais sur un nombre important de placettes, un certain nombre de relevés réguliers (un par semaine par exemple) sur peu de stations. Dans tous les cas, le nombre et la répartition des stations à observer doivent être fixés dans le cadre d'un plan d'échantillonnage.

Tenir compte de la représentativité

La représentativité constitue la première qualité que doit posséder un échantillon. Pour que les résultats soient généralisables à la population statistique, l'échantillon doit être représentatif de cette dernière, c'est-à-dire qu'il doit refléter fidèlement sa composition et sa complexité et fournir une estimation précise et non biaisée des paramètres mesurés sur les objets dans une aire donnée, à un moment donné.

Prendre en compte la taille des unités d'échantillonnage et du site

Plus les unités d'échantillonnage sont petites, plus elles doivent être nombreuses, pour les habitats notamment. Le nombre d'échantillons dépendra de la taille du site, de leurs nature, hétérogénéité et diversité ou de la population statistique.

Tenir compte des besoins pour l'analyse et l'interprétation des données

Le nombre d'échantillons doit être suffisamment élevé pour une analyse statistique pertinente des résultats. Classiquement, le nombre d'échantillons minimum proposé est de 30, par exemple pour des analyses factorielles. Cependant, les statistiques non paramétriques permettent de travailler avec un nombre d'échantillons plus faible. Il n'est pas évident de démontrer (statistiquement) des changements significatifs dans le temps pour des espèces qui ont une fréquence faible dans les relevés. Pourtant si on veut montrer des variations il est important que ces variations apparaissent entre les échantillons. Le gestionnaire définira un nombre d'échantillons suffisant pour mettre en évidence les changements dans le temps ou dans l'espace.

V-Les méthodes d'échantillonnage

Pour effectuer un échantillonnage, vous avez deux types : l'échantillonnage représentatif et l'échantillonnage aléatoire (probabiliste).

a) **L'échantillonnage représentatif :** ce type d'échantillon est

souvent utilisé dans une étude quantitative (questionnaire ou sondage), est défini comme représentatif lorsqu'il a les mêmes caractéristiques que la population étudiée (population mère). Un échantillon représentatif peut se faire à travers l'utilisation de la technique des quotas. Cette notion définit un groupe d'individus représentant une certaine partie de la population, en fonction de certains critères de segmentation ou d'étude.

Ce type d'échantillonnage a pour principal objectif de développer une certaine connaissance accrue d'une ou plusieurs populations en particulier, notamment par l'étude d'un grand nombre d'échantillons que l'on peut considérer comme réellement représentatif. A noter que le fait de mettre en place une action d'échantillonnage est généralement la réponse directe à une contrainte qui supprime la possibilité d'étudier exhaustivement la population, telle que le manque de temps pour étudier en totalité le groupe de population concerné, le coût financier, ou encore le manque de place.

Un échantillon sélectionné pour une étude quantitative est jugé comme représentatif quand il développe les mêmes caractéristiques et particularités que la population que l'on veut examiner et étudier. On peut notamment appeler cette population cible la "population mère". Il faut notamment retenir que cette représentativité doit absolument être basée sur les données et spécificités qui peuvent avoir une influence sur les réponses au sondage ou à l'enquête. Naturellement, si la représentativité de l'échantillon est jugée insuffisante, les résultats qui seront obtenus ne pourront pas être le support de généralisation sur la population mère.

Concernant la méthode de constitution de cet échantillon, on peut obtenir ce dernier par le biais de deux techniques : la méthode des quotas ou alors la méthode de l'échantillonnage aléatoire. A noter que ce principe de représentativité est d'autant plus important, et surtout nécessaire, pour ce qui concerne les panels, qui eux correspondent à une forme d'échantillon représentatif permanent.

b) **L'échantillonnage aléatoire :** L'échantillonnage aléatoire est

déterminé à partir d'une procédure de tirage aléatoire statistique. Malgré le hasard, la représentativité de l'échantillon aléatoire est assurée par les lois statistiques de la probabilité.

Parmi ses avantages, l'échantillonnage aléatoire simple ne requiert aucune connaissance a priori sur la population. De plus, son étude théorique est simple et les estimateurs courants pour la moyenne et la variance sont non-biaisés. L'absence de lien entre le choix des différents éléments est un avantage pour mettre en place un échantillonnage adaptatif. De plus, l'échantillonnage peut être interrompu ou poursuivi sans que cela ne cause de biais (à condition d'examiner les éléments dans l'ordre de tirage).

Parmi ses inconvénients, L'échantillonnage aléatoire simple offre une précision minimale se traduisant par une variance élevée des estimateurs. De plus, il peut être difficile à réaliser car il nécessite une procédure de tirage aléatoire qui puisse inclure n'importe quels éléments de la population, il faut donc que ces derniers soient tous repérés individuellement. Enfin, pour un échantillonnage aléatoire simple adaptatif, le coût en temps du déplacement et de repérage des unités statiques peut être élevé puisque elles peuvent être distantes spatialement les unes des autres. (***Tableau récapitulatif comparatif).***

Type d'échantillonnage	**Caractéristiques**	**Avantages**	**Limites**
L'échantillonnage représentatif	Possède les mêmes caractéristiques que	• Permet d'obtenir des résultats statistiques pertinents.	• Peut s'avérer difficile à mettre en place.

Type d'échantillonnage	Caractéristiques	Avantages	Limites
	la population étudiée.	• Le chercheur connaît précisément le taux de représentativité de son échantillon.	• Il faut du temps pour mener à bien un échantillonnage représentatif.
L'échantillonnage aléatoire	Se caractérise par le hasard à travers un tirage au sort aléatoire.	• Plus facile à mettre en place que l'échantillonnage représentatif. • Permet de calculer des seuils et intervalles de confiances face aux résultats obtenus.	• Malgré les lois statistiques de la probabilité, la représentativité d'une population peut être moins exacte. • Le chercheur contrôle moins la parfaite représentativité de son échantillon.

VI – Méthodes d'échantillonnage aléatoire simple :

L'échantillonnage aléatoire simple, ou au hasard, est une méthode qui consiste à prélever au hasard et de façon indépendante « n » unités d'échantillonnage d'une population de « N » éléments. Chaque point dans l'espace étudié a donc une chance égale d'être échantillonné.

Remarque :

Chaque élément sélectionné peut être remis dans la population après son tirage pour éventuellement être choisi une deuxième fois : on parle alors d'échantillonnage avec remise, appelé aussi échantillonnage non exhaustif. Si l'élément sélectionné n'est pas remis dans la population après son tirage, on parle d'échantillonnage sans remise ou échantillonnage exhaustif.

Exemples :

* Une méthode garantissant sécurité et représentativité consiste à dresser la liste complète et sans répétition des éléments de la population, à les numéroter, puis à tirer au sort « n » d'entre eux à l'aide d'un système générant des chiffres aléatoires.

* Tirage au sort d'un certain nombre d'heures de mesure dans l'année.

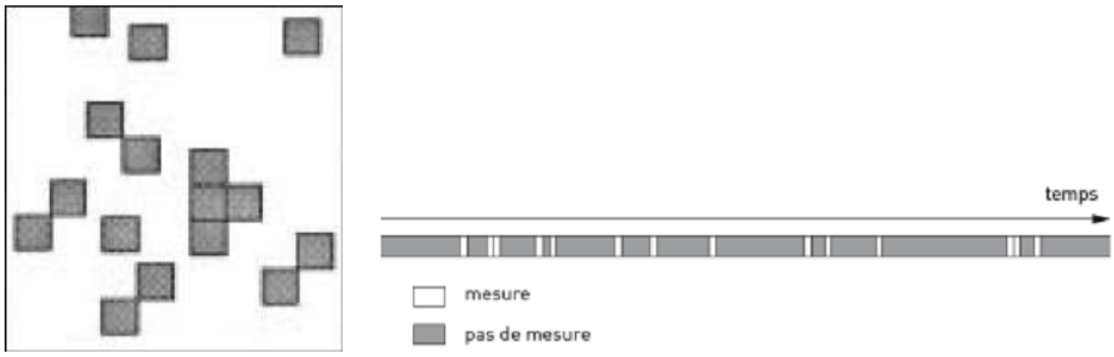


Figure 4. Illustration d'un échantillonnage aléatoire simple dans l'espace (à gauche) et dans le temps (à droite)

Pour utiliser la méthode d'échantillonnage aléatoire et simple dans un inventaire il faut d'abord disposer de la carte de végétation du site à inventorier.

Avantages et inconvénients :

L'échantillonnage aléatoire et simple présente des avantages importants:

- Estimation non biaisée de la moyenne de la population, calcul aisé de l'erreur d'échantillonnage.

-Avec l'échantillonnage aléatoire, les placettes sont sélectionnées indépendamment les uns des autres et respectent ainsi le caractère aléatoire des observations nécessaires pour les analyses statistiques.

-Il a pour inconvénient majeur les pertes de temps consécutives à la dispersion des échantillons.

-Aussi, il est assez rare que la végétation présente une homogénéité structurale justifiant l'utilisation de ce type d'échantillonnage. En cas de structure non homogène de la végétation, par exemple la présence de différents groupements végétaux au sein de la même végétation, l'échantillonnage aléatoire occasionne une perte de précision dans l'estimation des paramètres.

-Cette erreur étant surtout liée au fait que les formations végétales sont supposées dans ce type d'échantillonnage avoir le même poids en termes de superficie ou de densité d'arbres ou encore d'autres critères.

a. L'échantillonnage systématique

Un échantillonnage est systématique si les individus sont sélectionnés à intervalles réguliers (exemple une mesure journalière tous les six jours). **Il consiste aussi à répartir les échantillons de manière régulière** (p.ex. Tous les « x » mètres). Il est moins demandeur en temps qu'un échantillonnage aléatoire. On utilise habituellement un quadrillage (souvent positionné sur la photographie aérienne du territoire étudié). Les points d'échantillonnage sont ainsi faciles à localiser à chaque relevé.

Exemples :

 Si les espèces nichent au même endroit tous les ans, le comptage devient plus facile avec le temps.

 On peut réaliser un échantillonnage systématique lorsqu'on privilégie les inventaires dans les secteurs les plus susceptibles d'abriter les espèces (habitats potentiels). On porte alors une plus grande attention aux milieux répondant à leurs exigences écologiques. Par exemple, pour les chauves-souris, on cherchera en priorité dans les grottes mais aussi les mines, bâtiments, ponts, tunnels, arbres creux.

 Le positionnement des pièges pour les espèces difficilement observables (invertébrés ou encore mammifères) est souvent fait de manière systématique sur un secteur donné ou le long de gradients. La figure suivante nous montre comment sont localisés et répartis les pièges de micromammifères le long d'un transect dans une réserve naturelle.

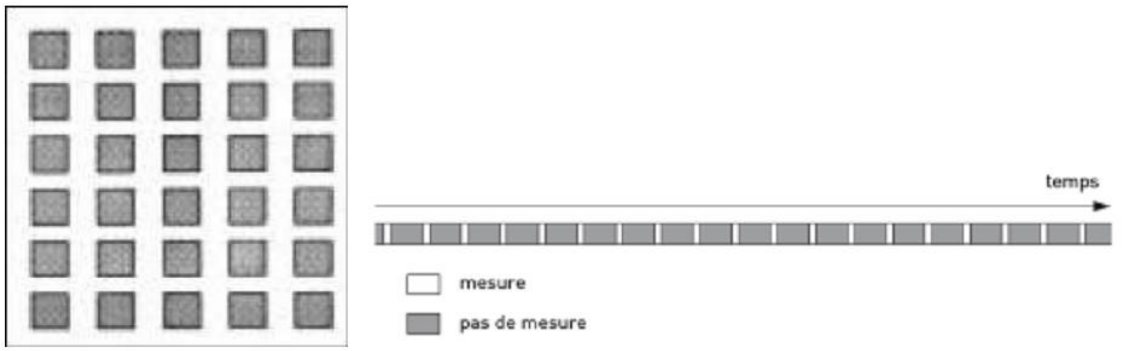


Figure 5. Illustration d'un échantillonnage systématique dans l'espace (à gauche) et dans le temps (à droite)

Avantage et inconvénients

L'avantage principal de ce type d'échantillonnage est qu'il est :

-plus facile à réaliser sur le terrain, du fait que l'échantillon est réparti de façon égale sur toute la superficie.

Comme inconvénients :

- le calcul de l'erreur d'échantillonnage peut être biaisé si l'on n'y prête pas attention.

- De même, la moyenne peut être aussi biaisée, notamment dans les cas où il existe une autocorrélation entre points de sondage (ici des placettes) géographiquement/spatialement très proches.

C'est un échantillonnage souvent recommandé dans les inventaires forestiers à grande échelle comme les inventaires forestiers nationaux.

b. L'échantillonnage stratifié

Il est particulièrement utilisé quand l'aire étudiée est divisée en zones différenciées (strates). Les strates peuvent correspondre à des divisions administratives, des zones à topographie différente,...etc.

Il consiste à subdiviser une population hétérogène en sous-populations ou strates plus homogènes. La stratification s'impose lorsque les résultats sont recherchés au niveau de chacune des sous-populations. On répartit alors les échantillons au sein des strates (en

procédant éventuellement par un échantillonnage au hasard par exemple) avec un nombre proportionnel à l'aire de chacune.

Exemple :

* On pourra utiliser toutes les connaissances acquises sur la végétation et le milieu pour découper la zone à étudier en sous-zones plus homogènes qui seront échantillonnées séparément.

*Un pré-échantillonnage est possible, notamment à l'aide de la cartographie (photographies aériennes, cartes géologique, pédologique, topographique,…).

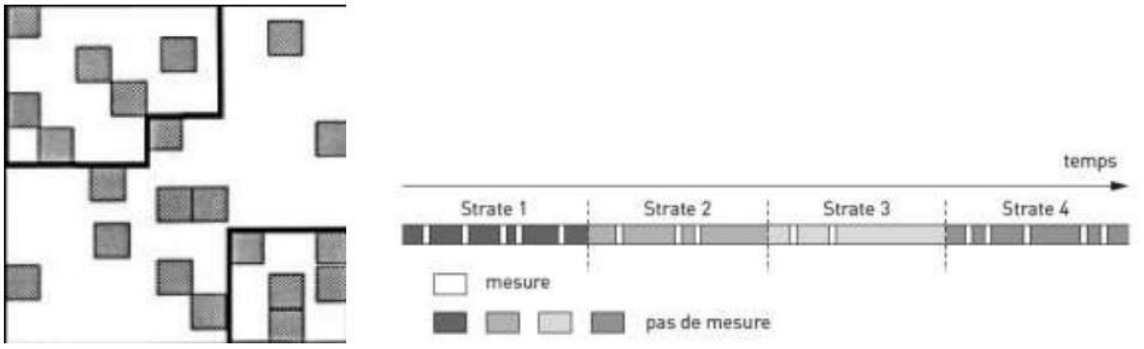


Figure 6. Illustration d'un échantillonnage stratifié dans l'espace (à gauche) et dans le temps (à droite)

Avantages et inconvénients

Les principaux avantages de l'échantillonnage aléatoire stratifié sont liés à la possibilité d'estimer pour chaque strate, les moyennes et les variances, et ceci de façon séparée ; les dispositifs d'échantillonnage différents peuvent être utilisés dans les différentes strates. Avec l'échantillonnage aléatoire stratifié, les placettes sont sélectionnées indépendamment les uns des autres et donne ainsi le caractère aléatoire de l'échantillonnage nécessaire pour les analyses statistiques.

En outre, la méthode suppose la connaissance préalable de la répartition de certaines strates dans la population et un échantillon doit être prélevé dans chaque strate si l'on souhaite effectuer une estimation relative à celle-ci. C'est la méthode d'échantillonnage la plus utilisée et recommandée pour l'étude de vastes formations végétales.

VII- Les biais dans la réalisation de l'échantillon

L'échantillonnage est efficace "à condition de respecter des règles rigoureuses". Afin de réaliser un échantillonnage correctement, il faut :

a) ***Bien choisir les personnes à interroger :*** les caractéristiques de

l'échantillonnage sont strictes : pour être efficace, il ne vous faut interrogez que les individus susceptibles de vous apporter des informations pertinentes pour votre recherche. Pour respecter la règle de la représentativité de l'échantillonnage, vous devez cibler les personnes interrogées.

b) *La taille de l'échantillon en fonction de l'étude :* Une étude empirique peut être menée par le biais d'une étude qualitative ou une étude quantitative. À travers le type d'étude et la technique d'enquête utilisée, la taille de l'échantillon varie.

- ***L'étude qualitative :*** *Lors d'une étude qualitative, la taille de l'échantillon peut être extrêmement restreinte : 1 à 2 personnes. En fonction des informations que vous voulez obtenir, il est important de sélectionner la bonne personne à interroger.*
- ***L'étude quantitative :*** *Dans le cadre d'une étude quantitative, l'échantillon à étudier doit être le plus représentatif possible de la population ciblée.*

Cet échantillon doit également être assez important, afin d'obtenir des données statistiques pertinentes et une conclusion efficace.

VIII-Conclusion :

- Définir une problématique et des hypothèses.
- Choisir le type d'étude (qualitative ou quantitative).
- Choisir l'outil le plus adéquat (entretien, questionnaire, sondage) car la taille de l'échantillon n'est pas la même selon l'outil employé.
- Définir un échantillon efficace : un questionnaire ou un sondage demanderons un échantillon large de personnes (+500) alors qu'un entretien beaucoup moins (une dizaine d'individus, voir juste 1 ou 2).
- Constituer un échantillon : si votre échantillon est volumineux (sondage ou d'un questionnaire), vous devez effectuer un échantillonnage représentatif de la population concernée par votre sujet.
- Cadrer votre échantillonnage : définissez un début et une fin à votre étude ("saturation théorique").

IX- Références bibliographiques :

1-Précis d'épidémiologie, Bezzaoucha.A, édition OPU, 2016

2-Methodes en épidémiologie, C. Rumeau Rouquette, 3 éme édition Flammarion.

3-Epidemiologie,Methodes et Pratique, Claude Rumeau Rouquette, édition Flammarion.

4-Epidemiologie en Médecine,Charles N,Edition Frison Roche.

République Algérienne Démocratique et Populaire
Ministère de l'enseignement Supérieur et de la Recherche Scientifique
Université Saad DAHLAB de Blida
Faculté de Médecine
Département de Médecine
Responsable de l'enseignement : Dr BENILHA.S

Polycopié distribué aux résidents de première année résidanat d'épidémiologie et médecine préventive

Etapes d'Analyse Statistique et choix du test statistique

(Cours à l'usage des étudiants en sciences médicales)

Dr BENILHA.S

MODULE DE D'EPIDEMIOLOGIE ET MEDECINE PREVENTIVE

ETAPES D'ANALYSE STATISTIQUE ET CHOIX DU TEST STATISTIQUE

1-Etapes d'Analyse Statistique

Les TESTS STATISTIQUES

Finalité d'une étude statistique

tirer des conclusions sur la population à partir de l'étude d'un échantillon issu de cette population

Obtention du résultat

par les tests statistiques qui sont des **tests d'hypothèse**

l'hypothèse étant imposée par construction du test

Definition :

Test statistique :

procédure qui permet , **avec un risque d'erreur connu**, d'effectuer un choix entre deux hypothèses complémentaires (H_o et H_1) au vu des observations réalisées sur un échantillon

Les résultats fournis par une étude statistique présentent de l'intérêt dans la mesure où ils sont accompagnés d'indications quantitatives fixant le degré de confiance qui peut leur être accordé

Choix du test statistique depend de la nature de variable

Les variables représentent Les caractéristiques étudiées sur les individus d'une population sont appelées des variables ou des caractères.

Sur les unités stratégiques c'est-à-dire les individus on mesure un caractère ou une variable. Les valeurs possibles d'une variable sont appelées des modalités, par exemple pour la variable couleur on a les modalités {vert, rouge, blanc}.

ITypes de variables Représente le domaine de variation de la variable,

On distingue deux grands types de variables (quantitative et qualitative) :

Les variables quantitatives, prennent des valeurs numériques ; leurs modalités sont mesurables, elles sont à leur tour composées de deux autres sous catégories : - Quantitatives discrètes, représente un ensemble de modalités fini ou dénombrable (nombre enfants par famille, nombre de places dans un cinéma, etc.). - Quantitatives continues, l'ensemble des modalités n'est pas dénombrable, elles sont représentées par des nombres en écriture décimale elles correspondent au résultat d'une mesure (poids, distance, moyenne, etc.).

Les variables qualitatives, sont des variables non quantitatives, les valeurs qu'elles prennent sont appelées catégories ou modalités, à leur tour elles peuvent être : - Nominales, elles ne présentent pas d'ordre particulier, par exemple la couleur des yeux est une variable quantitative nominale. - Ordinale, les modalités sont organisées selon un ordre hiérarchique, par exemple la mention du bac est une variable qualitative ordinale car les mentions sont ordonnées selon la moyenne obtenue:

a) Etapes de l'analyse statistique

Choix du test statistique en fonction des données du problème et de la (ou des) variable(s) étudiée(s) **.** nature de la (ou des) variable(s) (qualitative ou quantitative)

. nombre d'échantillons d'observations

. si plus d'un échantillon, sont-ils indépendants ?

taille des échantillons : grands ($n \geq 30$) ou petits ($n < 30$)

b) **Formulation** littérale puis statistique **des hypothèses**

H_o (hypothèse nulle)

H_1 (hypothèse alternative)

Définir les hypothèses à tester ***à partir des observations***

c) **Vérification** des **conditions de validité** (ou d'application) du **test**

d) **Calcul** de la **valeur expérimentale** de la **statistique de test** à partir des observations

e) **Choix du risque a** ou risque de première espèce

f) **Conclusion** de signification du test *

Criteres de choix du test statistique

Afin de déterminer le test adapté, on a besoin d'identifier certains critères :

- Nature des variables
- Nombre de groupes
- Appariement ou indépendance des groupes
- Taille
- Normalité de la distribution

A_Nature du test :

Un test bilatéral permet d'étudier une corrélation, qu'elle soit positive ou négative.

Un test unilatéral permet d'étudier une corrélation soit positive soit négative.

La plupart des tests permettent les tests bilatéraux et unilatéraux mais ce n'est pas le cas de tous.

B-Il faut distinguer les tests comparant

- 2 groupes
- > 2 groupes

C_Appariement :

- 2 groupes sont dits appariés lorsque chaque individu inclus dans un groupe correspondra à un sujet semblable (sur l'âge, le sexe, le poids…) dans l'autre groupe
- Cette méthode permet d'améliorer la puissance de l'étude en diminuant la variance d'un paramètre étudié (on diminue la variance parasite des autres paramètres).
- Ex : on fait une mesure à 2 reprises sur un même sujet (avant et après un traitement). En choisissant le même sujet, on diminue la variabilité inter-individuelle et on augmente la puissance.

- Cette méthode est très utilisée dans les études cas-témoins : pour chaque cas (malade) on va lui choisir un témoin (non malade) semblable sur certains critères (âge, sexe…). Si les sujets sont appariés, on utilisera une catégorie particulière de tests (très proches des tests pour les groupes non appariés).
- Ex : t Student pour données appariées

D_Taille

- Le nombre d'observations seuil est différent en fonction de la nature des variables :

Qualitative

- < 5 : petit échantillon, on choisira souvent un test exact de Fisher.
- ≥ 5 : on choisira souvent un test du Chi-2.

Quantitative :

- < 30 : petit échantillon, on utilisera souvent le test t de Student
- ≥ 30 : grands échantillons, on choisit le test Z, dont la distribution suivra une loi normale.

E_Normaliré ;

On distingue 2 situations :

- On regarde si la distribution de la variable suit une loi normale et peut être approximée par les paramètres caractérisant une loi normale, à savoir la moyenne et la variance.

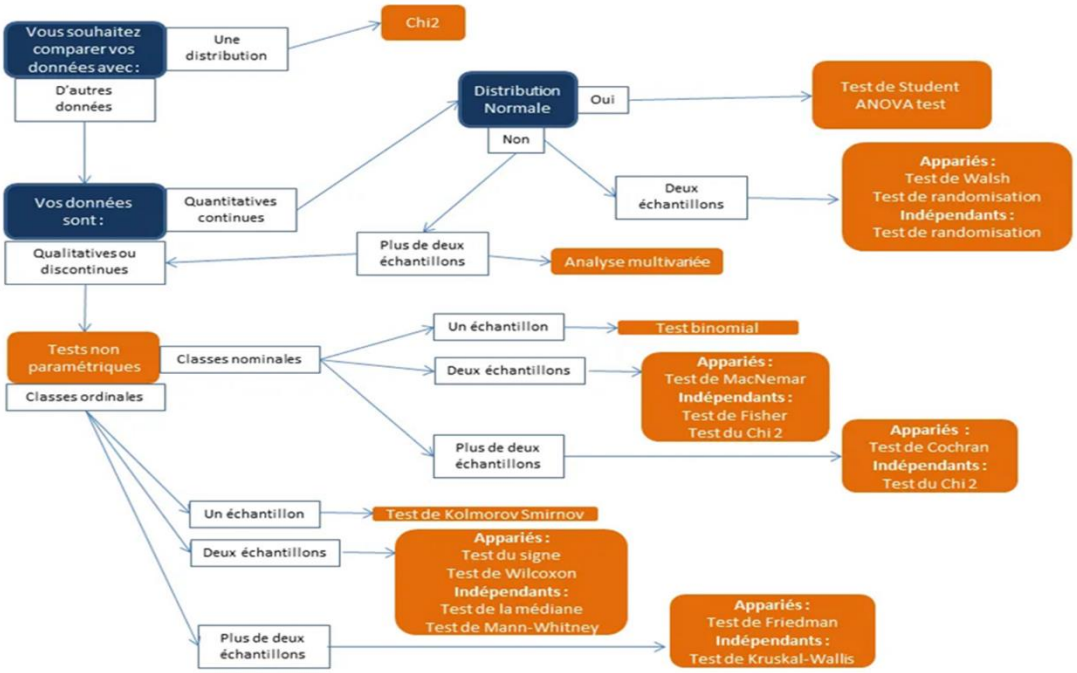
  On effectuera un test paramétrique.

- La distribution de la variable ne ressemble pas à une distribution normale, on ne pourra pas caractériser cette distribution par des paramètres.

  On effectuera un test non paramétrique.

- Les tests paramétriques sont un peu plus puissants que les tests non paramétriques.
- En revanche, ils ne peuvent être utilisés que dans des conditions de normalité alors que les tests non paramétriques sont plus robustes et peuvent s'appliquer indépendamment de la distribution et de la taille de l'échantillon.

Analyses Univariées



	Analyses multivariées		Variable de réponse			
			Qualitatif nominal (2 groupes)	Qualitatif nominal (plus de 2 groupes)	Qualitative Ordinale	Quantitative
Facteur d'étude	Qualitatifs (2 groupes)	Appariés	• Test de McNemar • Test exact de Fisher	Test Q de Cochran	• Tests des signes. • Tests des rangs signés de Wilcoxon	• Test t de Student pour données appariées • Tests des rangs signés de Wilcoxon
		Indépendants	• Z de comparaison de proportions • Chi² • Test exact de Fisher	Chi²	Test de Cochran-Armitage	• Test de Mann-Whitney • Test t de Student • Test de Welch
	Qualitatifs (plus de 2 groupes)	Appariés	Q de Cochran	Q de Cochran	Test de Friedman	Test de Friedman
		Indépendants	Chi²	Chi²	Test de Kruskal-Wallis (ordinal)	• Analyse de la variance • Test de Kruskal-Wallis (échelle quanti)
	Quantitatifs		Régression logistique	Régression logistique multinomiale	• Corrélation de Spearman • Tau de Kendall	• Corrélation de Pearson • Régression linéaire

Tests Paramétriques Tests non paramétriques