Cours de Biostatistique

3ème année cycle ingénieur

Rym Jaroudi

ESSAI

Table of Contents

Loi Binomiale

Épreuve de Bernoulli : expérience avec deux issues mutuellement exclusives A (succès) et B (échec), où :

$$P(X = 1) = p, \quad P(X = 0) = q = 1 - p$$

$$\mathbb{E}(X) = p, \quad \text{Var}(X) = p \cdot (1 - p) = pq$$

Loi Binomiale : répétition de n épreuves de Bernoulli indépendantes

$$P(K = k) = \binom{n}{k} p^k q^{n-k}$$

- K : nombre de succès (réalisation de A) parmi n épreuves
- k est une réalisation de la variable aléatoire K

- **Formule de récurrence :**

$$P(K = k + 1) = \frac{p \cdot (n - k)}{q \cdot (k + 1)} \cdot P(K = k)$$

- **Espérance et Variance :**

$$\mathbb{E}(K) = n \cdot p, \quad \text{Var}(K) = n \cdot p \cdot q$$

- **Fréquence relative** : pour $f = \frac{k}{n}$ avec F une réalisation de f,

$$\mathbb{E}(F) = p, \quad \text{Var}(F) = \frac{p \cdot q}{n}$$

Conditions d'applications :

- résultats binaires (deux évènements disjoints possibles uniquement).
- essais indépendants (les probabilités ne changent pas d'un essai à l'autre).
- n le nombre d'essais totaux est fixé à l'avance
- probabilité du 'succès' p constante (probabilité de 'l'échec' $= 1 - p$)

Loi de Poisson

Loi de Poisson : variable discontinue dans $\mathbb{N}$ et cas limite de la distribution binomiale : lorsque p est petit et n est grand, avec $m = n \cdot p$.

$$P(K = k) = \frac{e^{-m} \cdot m^k}{k!}$$

- m : paramètre qui correspond à la moyenne et à la variance de la distribution.

Conditions d'utilisation :

- résultat binaire
- essais indépendants
- taille d'échantillon ou laps de temps que le phénomène est observé fixe
- probabilité d'observation de l'évènement m faible.
  - On utilise la loi de Poisson si np et nq sont inférieurs à 5.
  - Si np et nq sont supérieurs à 5, on utilise la loi normale.

Loi Normale

Loi de Laplace-Gauss : variable continue sur $\mathbb{R}$

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

- μ : moyenne (position de la courbe).
- σ : écart type (dispersion autour de la moyenne).

Loi Normale Centrée Réduite : Toutes les lois normales peuvent être transformées en une loi normale centrée réduite, avec moyenne 0 et écart type 1, via la variable $u = \frac{x-\mu}{\sigma}$:

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

Loi symétrique autour de μ, où la moyenne = médiane = mode.

Importance de la Loi Normale :

- limite des lois binomiales et de Poisson pour des valeurs élevées de np et nq (souvent > 5 ou > 10 selon les auteurs).
- très fréquente dans les phénomènes biologiques et médicaux.

Théorème central limite : la moyenne de n variables indépendantes, même non normales, tend vers une distribution normale si n est grand ($n > 10$ souvent suffisant).

Loi log-normale : certaines distributions peuvent être transformées en distribution normale par des opérations simples (logarithme, racine carrée, etc.), formant une loi log-normale.

Loi du χ^2

Loi du χ^2 : Une variable χ^2 avec n degrés de liberté (DDL) est une somme des n carrés de variables normales centrées réduites indépendantes.

$$X^2 = \sum_{i=1}^{n} u_i^2, \quad \text{avec } u = \frac{X^2 - n}{\sqrt{2n}}$$

Densité de probabilité de forme non symétrique, avec une courbe spécifique pour chaque valeur de n.

Avec $n = 1$, le χ^2 est le carré d'une variable normale centrée réduite (par exemple, pour $\alpha = 0{,}05$, $u = 1{,}96 \Rightarrow \chi^2 = 3{,}84$).

Pour $n > 30$, la distribution du χ^2 tend vers une loi normale.

Loi t de Student

Loi t de Student : La loi t de Student à n degrés de liberté (DDL) est la distribution d'une variable

$$t = \frac{U}{\sqrt{X^2/n}},$$

où :

- U est une variable normale centrée réduite,
- X^2 suit une loi du χ^2 à n DDL.

Lorsque $n > 30$, la loi t est pratiquement confondue avec la loi normale.

Loi de Fisher-Snedecor

Loi F : Caractérisée par deux degrés de liberté n_1 et n_2.

Si Y_1 et Y_2 sont deux variables indépendantes suivant des lois χ^2 avec n_1 et n_2 DDL respectivement, alors

- la variable $F = \frac{Y_1/n_1}{Y_2/n_2}$ suit une loi F avec n_1 et n_2 DDL.
- L'inverse de F, $(1/F)$ suit également une loi F mais avec les DDL inversés (n_2 et n_1).

Exercice

Calculer la probabilité d'avoir au moins un résultat positif dans un bilan de 10 paramètres, en l'absence de maladie, avec une spécificité de 95%.

- **Données :**
  - $n = 10$: nombre de tests (paramètres)
  - $p = 0{,}05$: probabilité d'un faux positif (erreur)
  - $q = 1 - p = 0{,}95$: probabilité d'un test correct
- **Calcul de la probabilité d'aucun test positif :**

$$P(K = 0) = C_{10}^{0} \cdot p^{0} \cdot q^{10} = 0{,}95^{10} \approx 0{,}599$$

- **Probabilité d'au moins un test positif :**

$$P(K > 0) = 1 - P(K = 0) = 1 - 0{,}599 = 0{,}401$$

Interprétation : Il y a 40,1% de probabilité d'obtenir au moins un test faussement positif dans ce bilan de 10 paramètres, en l'absence de maladie.

Table of Contents

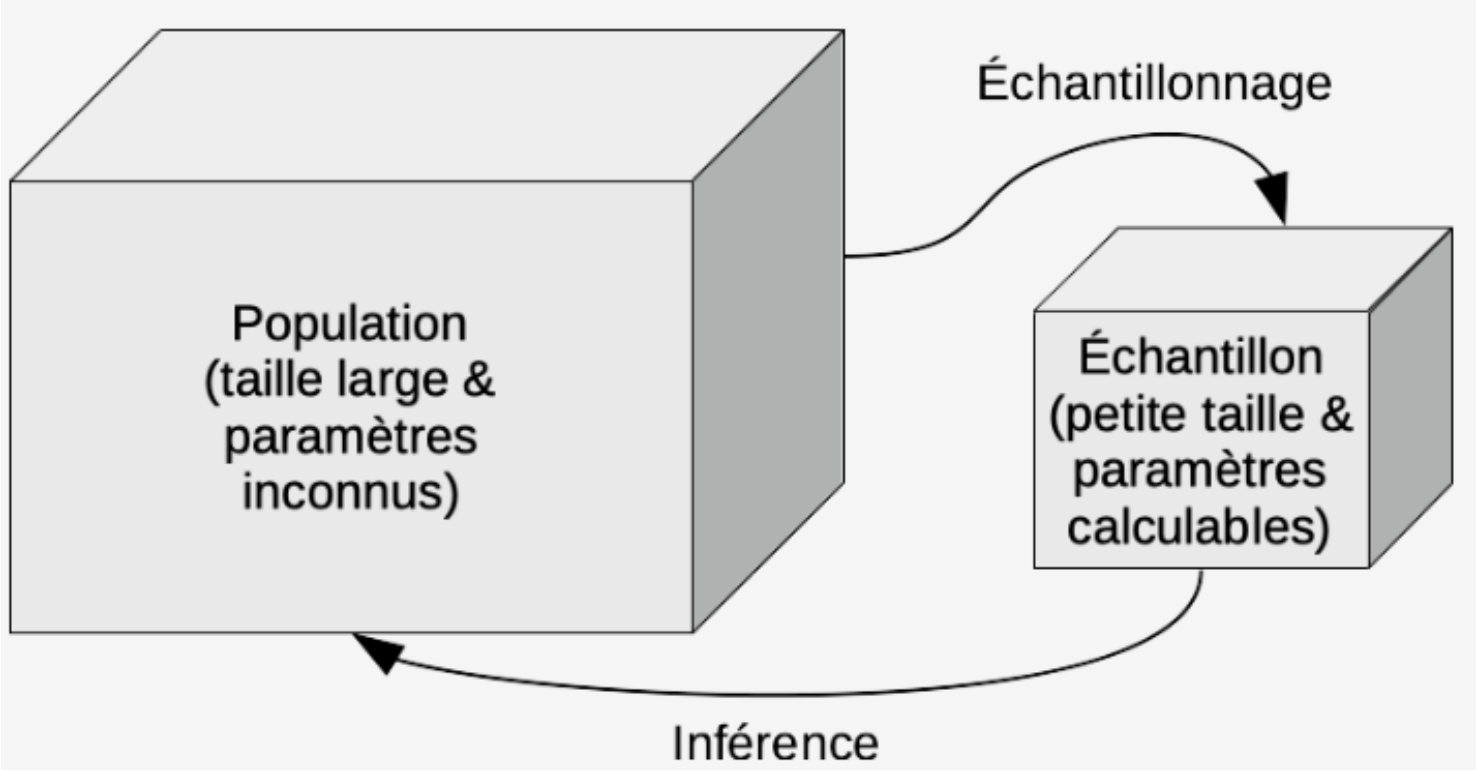
Population :

- En statistique : ensemble des valeurs que peut prendre une variable.
- En biologie : ensemble des individus d'une même espèce pouvant se reproduire entre eux.

Échantillon : Un petit sous-ensemble de la population de taille finie déterminée.

Échantillonnage : Le processus qui mène à la sélection des individus dans l'échantillon.

Inférence : Retirer un maximum d'information sur la population à partir de l'échantillon.


Échantillonnage
Population
(taille large &
paramètres
inconnus)
Échantillon
(petite taille &
paramètres
calculables)
Inférence

Travail Préliminaire

1. Comprendre bien la question posée, en termes biologiques (bibliographie, hypothèses, ...).
2. Valider les méthodes de collecte de données, vérifier que l'échantillonnage réalisé est représentatif, identifier la population étudiée pour éviter les surgénéralisations.
3. Réaliser une analyse exploratoire des données (graphiques, tableaux, nettoyage des données, ...).

Test d'hypothèse

- H_0 : L'hypothèse nulle est l'affirmation de base ou de référence qu'on cherche à rejeter.
- H_1 : L'hypothèse alternative représente une autre affirmation qui doit nécessairement être vraie si H_0 est fausse.

Réalité / Décision	**H_0 est vraie**	**H_0 est fausse**
H_0 acceptée	Bonne décision	Erreur de $2^{ème}$ espèce
H_0 rejetée	Erreur de $1^{ère}$ espèce	Bonne décision

Erreur de $1^{ère}$ espèce = $\alpha = P(H_0 \text{ rejetée} / \text{vraie})$.

Erreur de $2^{ème}$ espèce = $\beta = P(H_0 \text{ acceptée} / \text{fausse})$.

α : niveau de signification du test ou seuil souvent égal à 5%.

Exemples

Pour chaque situation suivante, formulez l'hypothèse nulle (H_0) et l'hypothèse alternative (H_1).

Efficacité d'un vaccin : Un chercheur souhaite vérifier si un nouveau vaccin réduit le taux d'infection par rapport à l'absence de vaccination.

H_0 : Le vaccin ne réduit pas le taux d'infection ; le taux d'infection est identique pour les groupes vaccinés et non vaccinés.

H_1 : Le vaccin réduit le taux d'infection par rapport au groupe non vacciné.

Plante médicinale et glycémie : Des patients souffrant de diabète sont traités avec une plante médicinale, et le chercheur souhaite vérifier si cette plante réduit la glycémie par rapport aux patients non traités.

H_0 : La plante médicinale n'a aucun effet sur la glycémie ; la glycémie moyenne est identique entre les groupes traité et non traité.

H_1 : La plante médicinale réduit la glycémie par rapport au groupe non traité.

Pollution et biodiversité d'un lac : La biodiversité d'un lac est mesurée avant et après une exposition prolongée à une source de pollution pour voir s'il y a une diminution du nombre d'espèces.
H_0 : La pollution n'a pas d'effet sur la biodiversité du lac ; le nombre d'espèces est le même avant et après l'exposition à la pollution.
H_1 : La pollution diminue la biodiversité du lac.

Hormone de croissance et taille des grenouilles : Un biologiste souhaite déterminer si l'injection d'une hormone de croissance augmente significativement la taille moyenne des grenouilles.

H_0 : L'injection d'hormone de croissance n'augmente pas la taille moyenne des grenouilles ; la taille moyenne est la même pour les groupes traité et non traité.

H_1 : L'injection d'hormone de croissance augmente la taille moyenne des grenouilles.

Test du χ^2

Conditions d'application :

- Échantillon représentatif : échantillonnage aléatoire et individus indépendants les uns des autres.
- Les observations doivent être indépendantes.
- L'effectif théorique par classe doit au moins égal à 5.

Comparaison des répartitions

Soient plusieurs échantillons d'effectifs $n_1, n_2, n_3, \ldots, n_m$ et soient $K_1, K_2, K_3, \ldots, K_m$ les effectifs portant un caractère A.

Les proportions des individus portant le caractère A dans chaque échantillon :

$$p_1 = \frac{K_1}{n_1},\ p_2 = \frac{K_2}{n_2},\ \ldots,\ p_m = \frac{K_m}{n_m}.$$

Les proportions des individus ne portant pas le caractère A :

$$q_1 = 1 - p_1,\ q_2 = 1 - p_2,\ \ldots,\ q_m = 1 - p_m.$$

Les proportions $p_1, p_2, \ldots, p_m$ et $q_1, q_2, \ldots, q_m$ permettent de comparer la présence et l'absence du caractère A entre les différents échantillons.

H_0 : Les échantillons proviennent de la même population.
Estimation du pourcentage du caractère A dans la population

P_0 :

$$p_0 = \frac{K_1 + K_2 + \cdots + K_m}{n_1 + n_2 + \cdots + n_m}.$$

Calcul des effectifs théoriques pour chaque échantillon C_i :

$$\begin{aligned} C_1 &= n_1 \cdot p_0, \\ C_2 &= n_2 \cdot p_0, \\ &\vdots \\ C_m &= n_m \cdot p_0. \end{aligned}$$

Calcul du χ^2 observé

$$\chi^2 = \sum \frac{(K_i - C_i)^2}{C_i}$$

- K_i : Effectifs expérimentaux $(i = 1, 2, \ldots, m)$
- C_i : Effectifs théoriques $(i = 1, 2, \ldots, m)$

Au seuil de signification $\alpha = 5\%$, et pour un $ddl = m - 1$:

- Si $\chi^2 < \chi^2_\alpha$, H_0 est **acceptée** : *Les échantillons proviennent de la même population.*
- Si $\chi^2 \geq \chi^2_\alpha$, H_0 est **rejetée** : *Les échantillons ne proviennent pas de la même population.*

Note : χ^2_α est lu dans une table de χ^2 pour $\alpha = 5\%$ et les *ddl* correspondants.

Test t de Student

Conditions d'application :

- Échantillon représentatif.
- Les observations doivent être indépendantes.
- Distribution de la population :
  - Normale : le test est exact.
  - Approximativement Normale : le test est approx. exact.
  - Non Normale : le test reste approx. exact si la taille de l'échantillon est grande (grâce au théorème central limite).

Comparaison des moyennes

Comparer la moyenne $\bar{X}$ d'un grand échantillon expérimental ($n \geq 30$) à une population de moyenne μ_0 et écart-type σ connus.

$$\bar{X} \sim \mathcal{N}(\mu_0, \frac{\sigma}{\sqrt{n}})$$

La transformation t qui correspond à la valeur critique de Student suit une loi Normale Centrée Réduite :

$$t = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \quad \text{avec} \quad t \sim \mathcal{N}(0, 1)$$

Hypothèses :

- $H_0 : \mu = \mu_0$ (pas de différence entre la moyenne de l'échantillon et celle de la population).
- $H_1 : \mu \neq \mu_0$ (différence significative entre les deux moyennes).

Seuil de décision ($\alpha = 5\%$) :

- Si $|t| < 1{,}96$, la différence n'est pas significative $\Rightarrow H_0$ est acceptée.
- Si $|t| \geq 1{,}96$, la différence est significative $\Rightarrow H_0$ est rejetée.

Test F

Conditions d'application :

- Échantillon représentatif
- Les observations doivent être indépendantes.
- Les résidus doivent suivre une loi normale.
- **Homoscédasticité :** même variance intragroupe.

Remarque : Le respect de ces conditions est crucial pour que le test ANOVA soit valide et que ses résultats soient interprétables de manière fiable.

Comparaison des variances

Comparer deux variances σ_1^2 et σ_2^2 en posant l'hypothèse nulle :

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Statistique de test :

$$F_{\text{obs}} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} = \frac{\text{SCE}_1/(n_1 - 1)}{\text{SCE}_2/(n_2 - 1)}$$

où :

$$\text{SCE}_i = \sum_{j=1}^{n_i} (x_{ij} - \bar{X}_i)^2$$

et n_1, n_2 sont les tailles des échantillons respectifs.

Décision :

- On compare F_{obs} avec la valeur critique F_α issue des tables de Fisher-Snedecor (selon $n_1 - 1$ et $n_2 - 1$ degrés de liberté).
- Si $F_{\text{obs}} > F_\alpha$, on rejette H_0.
- Sinon, H_0 est acceptée.

Conclusion : Si H_0 est rejetée, cela indique que les deux variances sont significativement différentes au seuil α.

Remarque : L'égalité des variances est une des conditions de l'analyse statistique paramétrique telle que ANOVA.

Exercice

Un chercheur analyse le poids de 200 patients participant à un programme de santé, avec un poids moyen de 72 kg et un écart-type de 8 kg. La distribution est proche de la normale. Le chercheur pose deux questions :

1. Quel est l'intervalle de confiance à 95 % pour le poids moyen ?
2. En testant l'hypothèse que le poids moyen est de 70 kg, quelle est la *p-valeur* ?

Intervalle de confiance : Pour un intervalle de confiance à 95 %, on utilise :

$$IC = \bar{X} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

avec :

- $\bar{X} = 72$ kg,
- $Z_{\alpha/2} = 1.96$ pour un niveau de confiance de 95 %,
- $\sigma = 8$ kg, écart-type,
- $n = 200$, taille de l'échantillon.

Calculons :

$$IC = 72 \pm 1.96 \times \frac{8}{\sqrt{200}}$$

$$IC = 72 \pm 1.96 \times \frac{8}{14.14} = 72 \pm 1.11$$

L'intervalle de confiance à 95 % pour le poids moyen est donc $[70.89, 73.11]$ kg.

Test d'hypothèse : Hypothèse nulle (H_0) : Le poids moyen est de 70 kg.
Calculons la statistique Z :

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Avec un score Z de 3.54, la *p-valeur* est inférieure à 0.001, indiquant un résultat statistiquement significatif.

Conclusion : L'intervalle de confiance pour le poids moyen est $[70.89, 73.11]$ kg. Le test statistique montre que le poids moyen diffère significativement de 70 kg, avec une *p-valeur* très faible ($p < 0.001$).

Table of Contents

Estimation de la Prévalence

- **Mesure du risque de maladie dans la population**
- Proportion de malades présents (M^+) dans la population (N) à un instant donné :

$$P = \frac{M^+}{N}$$

- C'est une probabilité.
- Intègre la notion de **durée de la maladie** :
  - Augmentation de la durée de la maladie $\Rightarrow$ augmentation du nombre de M^+ $\Rightarrow$ augmentation de la prévalence.
- Intègre la notion de **vitesse d'apparition** des nouveaux cas M^+ :
  - Augmentation de la vitesse d'apparition des cas $\Rightarrow$ augmentation de la prévalence.

Estimation de l'Incidence

- **Quantifie la production de nouveaux cas de maladie** dans la population pendant un certain intervalle de temps.
- Calcul de l'incidence :

$$I = \frac{\text{nombre de nouveaux cas pendant } t}{N \times t}$$

- Représente le **taux d'apparition de la maladie** dans la population.
- Permet d'évaluer la vitesse à laquelle de nouveaux cas apparaissent.

Risque Relatif

- Considérons une maladie (**M**) qui peut être présente (M+) ou absente (M-), et un facteur de risque (**F**) qui peut être présent (F+) ou absent (F-).
- Risque d'apparition de la maladie chez les exposés au facteur F : $P(M+/F+)$
- Risque d'apparition de la maladie chez les non-exposés au facteur F : $P(M+/F-)$
- **Risque Relatif (RR)** : mesure l'influence du facteur F sur la survenue de la maladie.

$$RR = \frac{P(M+/F+)}{P(M+/F-)}$$

- Le risque relatif varie de 0 à l'infini.

Interprétation du Risque Relatif

- **Si RR > 1 :**
  - $P(M+/F+) > P(M+/F-)$
  - La présence du facteur F favorise la maladie : il s'agit d'un **facteur de risque**.
- **Si RR < 1 :**
  - $P(M+/F+) < P(M+/F-)$
  - La présence du facteur F favorise la non-maladie : il s'agit d'un **facteur protecteur**.
- **Si RR $= 1$:**
  - $P(M+/F+) = P(M+/F-)$
  - Le facteur F n'a **pas d'effet** sur la maladie.

Exercice

Une étude est menée pour évaluer l'association entre le tabagisme (*facteur* $F+$) et le développement du cancer du poumon (*maladie* $M+$) dans une population de 1200 personnes. Les résultats montrent :

- 400 individus fument ($F+$).
- Parmi les fumeurs, 50 ont un cancer du poumon ($M+$).
- Parmi les non-fumeurs ($F-$), 20 ont un cancer du poumon ($M+$).

1. Quelle est la prévalence du cancer du poumon dans la population ?
2. Quel est le risque relatif (RR) associé au tabagisme ?

1. **Prévalence :**

$$P(M+) = \frac{\text{Nb total de cas de cancer}}{\text{Population totale}} = \frac{50+20}{1200} = \frac{70}{1200} \approx 5.83\%.$$

2. **Risque relatif (RR) :**

$$RR = \frac{P(M+/F+)}{P(M+/F-)} = \frac{50/400}{20/800} = \frac{0.125}{0.025} = 5.$$

Conclusion : Les fumeurs ont un risque 5 fois plus élevé de développer un cancer du poumon par rapport aux non-fumeurs.

QCM : Biais en épidémiologie

Question : Qu'est-ce qu'un **biais de mesure** dans une étude épidémiologique ?

- A. Un biais qui résulte de la manière dont les participants sont choisis, affectant ainsi la représentativité de l'échantillon.
- B. Un biais qui survient lorsque le diagnostic est incorrect ou que les données sont mal collectées, faussant les résultats.
- C. Un biais causé par des facteurs externes qui influencent les résultats de l'étude.
- D. Un biais lié à la taille de l'échantillon, qui n'est pas suffisamment grande pour rendre les conclusions fiables.

Réponse: B.

Question : Qu'est-ce qu'un **biais de sélection** dans une étude épidémiologique ?

- A. Un biais lié à la manière dont les participants sont répartis dans les groupes d'étude.
- B. Un biais qui survient lorsque les participants ne sont pas représentatifs de la population cible, ce qui peut fausser les conclusions.
- C. Un biais qui survient lorsque les données sont collectées de manière incorrecte.
- D. Un biais causé par une taille d'échantillon trop grande.

Réponse: B.

Question : Qu'est-ce qu'un **biais de confusion** dans une étude épidémiologique ?

- A. Un biais causé par des facteurs confondants, comme des variables non contrôlées, qui peuvent influencer les résultats.
- B. Un biais qui résulte d'une taille d'échantillon trop petite.
- C. Un biais qui survient lorsque les mesures sont mal prises.
- D. Un biais dû à une analyse statistique incorrecte.

Réponse : A.

Table of Contents

Conclusion

Les **tests statistiques** permettent de vérifier des hypothèses sur des échantillons en utilisant des lois spécifiques :

- **Test χ^2** pour comparer des répartitions.
- **Test t** pour comparer des moyennes.
- **Test F** pour comparer des variances.

Les **indicateurs épidémiologiques** permettent d'évaluer et de comparer l'état de santé des populations :

- **Prévalence :** proportion de malades à un instant donné.
- **Incidence :** fréquence de nouveaux cas.
- **Risque relatif :** influence d'un facteur ($RR > 1$: risque, $RR < 1$: protection).

Une interprétation rigoureuse des résultats est essentielle pour guider les décisions.