Cours de Biostatistique

3 ème année cycle ingénieur

Rym Jaroudi

ESSAI

Table of Contents

Organisation globale du module

- 3 séances de 3h + 1 séance de 1h30

Dates	Séances
Samedi **09/11/2024**	- **1ère séance : de 8h30 à 10h** - **2ème séance : de 10h10 à 11h40**
Samedi **16/11/2024**	- **3ème séance : de 8h30 à 10h** - **4ème séance : de 10h10 à 11h40**
Samedi **23/11/2024**	- **5ème séance : de 8h30 à 10h** - **6ème séance : de 10h10 à 11h40**
Samedi **07/12/2024**	- **7ème séance : de 8h30 à 10h**

- Module optionnel $\Rightarrow$ 1 examen en fin de semestre
- **MAX. ABSENCES = 2 séances !!!**

Objectifs du cours

1. Appliquer les concepts statistiques de base aux données biologiques et médicales.
2. Comprendre l'inférence statistique dans le contexte des études cliniques et épidémiologiques.
3. Explorer les méthodes de régression adaptées aux analyses de données en biologie et en santé.

Table of Contents

Concepts de base

Définition : La biostatistique est l'application de la statistique aux sciences de la vie. Elle permet de tirer des conclusions fiables à partir de données complexes, d'étudier l'incidence des maladies, l'efficacité des traitements, et de déterminer les facteurs de risque dans la santé publique.

Différences avec la statistique classique : En biostatistique, les données proviennent souvent d'observations naturelles, avec des défis comme les effets confondants, la variabilité biologique, et l'importance de la taille de l'échantillon.

Rôle des biostatisticiens : les biostatisticiens travaillent généralement en mode projet au sein d'équipes pluridisciplinaires en contribuant à la conception et au développement de méthodes statistiques utilisées dans

- essais cliniques : médicaments ou produit cosmétique
- études épidémiologiques : facteurs influençant la fréquence ou la distribution des maladies
- analyse de données omiques : génomique, transcriptomique, protéomique
- évaluation de produits alimentaire : propriétés sensorielles et organoleptiques

Opportunités et Débouchés

Carrières : ingénieur en entreprise ou en milieu académique

Secteurs d'activité : industrie pharmaceutique, biotechnologies, santé publique, cosmétiques ...

INRAE Inria

B M S
Laboratory of BioInformatics
bioMathematics & bioStatistics

Inserm cnrs

▸ inra ▸ inria

Applications typiques

Épidémiologie : analyse de la propagation des maladies, prévalence et incidence, modélisation de la transmission des infections (ex. COVID-19).
Essais cliniques : évaluation de l'efficacité et de la sécurité de traitements médicaux via des méthodologies strictes, incluant des randomisations, des groupes témoins, et des analyses d'effets secondaires.
Bioinformatique et génomique : analyse de données d'ADN (séquençage) et association entre gènes et traits phénotypiques, avec des méthodes avancées de réduction de dimensions.
Santé publique : identification des facteurs de risque et surveillance de la santé des populations par analyse de larges bases de données médicales.

Progrès récents en biostatistique

Méthodes d'analyse de survie (1960-1980) : David Cox et d'autres chercheurs développent l'analyse de survie, qui devient cruciale pour les études de santé à long terme. Le modèle de Cox permet d'étudier des données censurées et de mieux comprendre la durée jusqu'à des événements spécifiques (par exemple, décès, récidive).

Avancées en génomique et bioinformatique (1990-2000) : Avec le séquençage du génome humain et la croissance rapide des données génétiques, des méthodes statistiques spécifiques (réduction de dimension, tests d'association) permettent de relier gènes et maladies et d'aborder les données de haute dimension, qui caractérisent la biostatistique moderne.

Modèles bayésiens et hiérarchiques (2000-présent) : Grâce aux progrès informatiques, les méthodes bayésiennes, qui permettent d'incorporer des informations a priori, se généralisent. Elles sont aujourd'hui courantes dans les essais cliniques adaptatifs et l'épidémiologie, et facilitent la modélisation des phénomènes complexes de santé.
Apprentissage automatique et intelligence artificielle (2010-présent) : Le développement de l'IA et des méthodes de machine learning révolutionne la biostatistique en santé publique et en médecine. Ces outils permettent d'optimiser les diagnostics, de personnaliser les traitements et d'améliorer la santé publique via des modèles prédictifs et des analyses de données massives.

Types de variables

Données quantitatives

- **Continues** : Valeurs dans un intervalle
- **Discrètes** : Valeurs finies, dénombrables

Données qualitatives

- **Nominales** : Pas d'ordre, uniquement des catégories
- **Ordinales** : Ordre établi entre les catégories

Exercice d'application

Enoncé : Dans une étude sur le diabète, les variables suivantes sont recueillies : âge, sexe, IMC, glycémie, type de traitement, et gravité des symptômes. Identifier les types de données.
Réponse :

- âge : Quantitative continue
- Sexe : Qualitative nominale
- IMC : Quantitative continue
- Glycémie : Quantitative continue
- Type de traitement : Qualitative nominale
- Gravité des symptômes : Qualitative ordinale

Review Exercise

Question

For each of the following situations, answer questions a through e:

(a) What is the sample in the study?

(b) What is the population?

(c) What is the variable of interest?

(d) How many measurements were used in calculating the reported results?

(e) What measurement scale was used?

Situation A. A study of 300 households in a small southern town revealed that 20 percent had at least one school-age child present.

Situation B. A study of 250 patients admitted to a hospital during the past year revealed that, on the average, the patients lived 15 miles from the hospital.

Answer

Situation A

(a) 300 households **(b)** all households in the small southern town **(c)** number of school-aged children present **(d)** all that reported one or more children **(e)** nominal (categories: 0 children, 1 child, and so on)

Situation B

(a) 250 patients **(b)** all patients admitted to the hospital during the past year **(c)** distance patient lives from the hospital **(d)** 250 distances **(e)** ratio

Table of Contents

Erreurs Courantes en Biostatistique

- Utilisation de la mauvaise mesure de synthèse
- Interprétation incorrecte de la valeur p
- Mauvaise interprétation de l'intervalle de confiance (IC)
- Ignorance du calcul de taille d'échantillon
- Confusion entre corrélation et causalité
- ...

Exemple : Mesure de la Taille des Plantes dans une Population Asymétrique

Contexte : Supposons que l'on mesure la taille de plantes dans une région marécageuse avec quelques plantes exceptionnellement hautes.

- La majorité des plantes mesurent entre 15 et 20 cm, mais quelques plantes rares atteignent 150 cm.
- Objectif : Determiner la taille typique des plantes.

Erreur : Utiliser la **moyenne** donne une taille moyenne autour de 30 cm, en raison des plantes très hautes.
Solution : Utiliser la **médiane** pour une distribution asymétrique :

$$X = \{x_1, x_2, \ldots, x_n\}, \quad \text{où } X \text{ est asymétrique}$$

Formules :

- **Moyenne** :

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- **Médiane** : Valeur centrale de X lorsque les données sont triées.

Conclusion : La médiane (environ 18 cm) est une meilleure représentation de la taille typique des plantes.

Exemple : Effet d'un Nouveau Médicament sur la Croissance Cellulaire

Contexte : Un biologiste teste un nouveau médicament pour améliorer la croissance de cellules végétales.

- L'expérience compare les cellules traitées et non traitées (groupe témoin).
- Objectif : Interpreter l'analyse statistique donnant une valeur $p = 0,04$.

Erreur : $p = 0,04$ ne signifie pas que le médicament est efficace à 96 %. Cela n'indique pas non plus que le résultat est cliniquement significatif.

Solution : $p = 0,04$ signifie qu'il y a une probabilité de 4 % d'observer une différence aussi grande ou plus grande entre les groupes par hasard, si l'hypothèse nulle est vraie.

Formule : La **valeur p** est donnée par :

$$p = P(\text{Test Statistique} \geq t_{\text{observé}} \mid H_0)$$

où :

- $t_{\text{observé}}$ est la valeur observée de la statistique de test.
- H_0 est l'hypothèse nulle, indiquant qu'il n'y a pas d'effet réel du traitement.

Conclusion : La **valeur p** indique uniquement la probabilité d'observer le résultat actuel par hasard, sans lien avec l'efficacité clinique réelle du médicament.

Exemple : Estimation de la Fréquence d'un Allèle dans une Population

Contexte : Un généticien souhaite estimer la fréquence d'un allèle A dans une population.

- Dans un échantillon de 500 individus, 120 sont porteurs de l'allèle A.
- Objectif : Estimer la fréquence réelle de l'allèle A dans la population avec un intervalle de confiance (IC) à 95 %.

Erreur : Interpréter l'**IC** à 95 % comme une probabilité que la vraie fréquence se trouve dans cet intervalle spécifique.

Solution : Un **IC** de 95 % signifie que si l'on répète cette estimation plusieurs fois, environ 95 % des intervalles contiendront la fréquence réelle de l'allèle dans la population.

Formule : Calculer l'**IC** pour une proportion (la fréquence de l'allèle) avec la formule :

$$IC = \hat{p} \pm z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

où :

- $\hat{p} = \frac{120}{500} = 0{,}24$ est la fréquence observée de l'allèle A,
- $z = 1{,}96$ pour un niveau de confiance de 95 %,
- $n = 500$ est la taille de l'échantillon.

Conclusion : L'**intervalle de confiance** [0,2026 ; 0,2774] signifie qu'environ 95 % des intervalles ainsi calculés contiendraient la vraie fréquence de l'allèle A dans la population. Cela ne garantit pas que cette fréquence est dans cet intervalle spécifique.

Exemple : Essai Clinique sur l'Effet d'un Traitement sur le Niveau d'une Protéine Sanguine

Contexte : Un biologiste mène un essai clinique pour évaluer l'effet d'un médicament sur le niveau d'une protéine sanguine.

- Paramètres :
  - Différence minimale détectable (d) : 5 unités
  - Niveau de signification (α) : 0,05
  - Puissance ($1 - \beta$) : 80%
  - Écart-type estimé (σ) : 10 unités
- Objectif : Détecter une différence moyenne de 5 unités dans le niveau de la protéine entre le groupe traité et le groupe témoin.

Erreur : Ne pas calculer la **taille d'échantillon** nécessaire, ce qui réduit la puissance de l'étude et peut mener à des conclusions non fiables.

Solution : Une **taille d'échantillon** insuffisante peut ne pas détecter une différence significative même si le traitement est efficace.

Formule : Calculer la **taille d'échantillon** avec la formule :

$$n = \left(\frac{Z_{\alpha/2} + Z_{\beta}}{d} \right)^2 \sigma^2$$

où :

- $Z_{\alpha/2} = 1,96$ pour un niveau de confiance de 95%
- $Z_{\beta} = 0,84$ pour une puissance de 80%
- $d = 5$ est la différence minimale détectable
- $\sigma = 10$ est l'écart-type estimé

Conclusion : Pour cet essai clinique, 32 participants par groupe (traité et témoin) sont nécessaires pour détecter une différence moyenne de 5 unités dans le niveau de la protéine sanguine, avec une puissance de 80% et un niveau de confiance de 95%.

Exemple : Étude Épidémiologique de la Consommation de Café et des Maladies Cardiaques

Contexte : Un épidémiologiste observe une association entre la consommation de café et le taux de maladie cardiaque dans une étude observationnelle.

- Les données montrent que les personnes qui consomment plus de café présentent un taux de maladie cardiaque plus élevé.
- Objectif : Comprendre si la consommation de café est un facteur de risque de la maladie cardiaque.

Erreur : Interpréter la corrélation observée entre la consommation de café et la maladie cardiaque comme une relation de causalité. Conclure que boire du café cause directement un risque accru de maladie cardiaque.

Solution : Examiner les autres facteurs de confusion possibles et réaliser des études additionnelles pour établir la causalité.

- Par exemple, il se peut que les personnes qui consomment beaucoup de café aient également d'autres comportements à risque (ex. : fumer), qui augmentent le risque de maladie cardiaque.
- Utiliser des études expérimentales (ex. : essais contrôlés randomisés) ou des méthodes statistiques d'ajustement (ex. : régression) pour contrôler ces variables de confusion.

Formule : La corrélation entre deux variables X et Y est calculée par le coefficient de corrélation r :

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

où :

- $\bar{X}$ et $\bar{Y}$ sont les moyennes de X et Y,
- r varie de -1 (corrélation négative parfaite) à $+1$ (corrélation positive parfaite).

Cependant, r n'implique pas une causalité ; il indique seulement une association.

Conclusion : La corrélation observée entre la consommation de café et le risque de maladie cardiaque n'implique pas que le café cause la maladie cardiaque. D'autres facteurs de confusion doivent être examinés, et des études expérimentales sont nécessaires pour établir une relation de causalité.

Table of Contents

Conclusion

Pourquoi la Biostatistique est Cruciale :

- Permet des décisions cliniques basées sur des preuves solides.
- Essentielle pour analyser et interpréter les données en santé.

Risques des Erreurs :

- **Mauvaise pratique clinique :** Utilisation de traitements inefficaces ou dangereux.
- **Crédibilité en jeu :** La confiance dans les résultats scientifiques diminue.
- **Risques vitaux :** Dans les essais cliniques, des erreurs peuvent entraîner des risques graves pour les patients.

La biostatistique est une clé pour des recherches fiables et des décisions médicales sûres.