

MSV31 – Statistiques

- ▶ Statistiques **descriptives**
 - ▶ **Estimateurs**
 - ▶ **tests statistiques**

Vincent Runge
vincent.runge@univ-evry.fr
IBGBI – 4^{ème} étage
Cours L2 Bio. 2019 – 2020



Laboratoire de
Mathématiques
et Modélisation
d'Évry



Plan du cours

- 1 Introduction
- 2 Des statistiques pour quoi faire?
- 3 Statistiques descriptives
- 4 Estimateurs et estimations
- 5 Tests statistiques

Plan

1 Introduction

2 Des statistiques pour quoi faire?

3 Statistiques descriptives

4 Estimateurs et estimations

5 Tests statistiques

Plan

1 Introduction

- Qu'est-ce que la statistique?
- Courte histoire des mathématiques
- Statistiques dans l'histoire des mathématiques

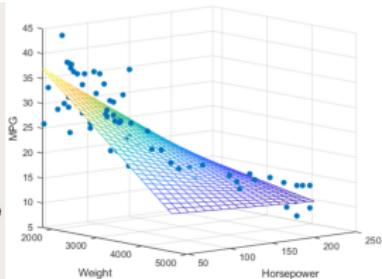
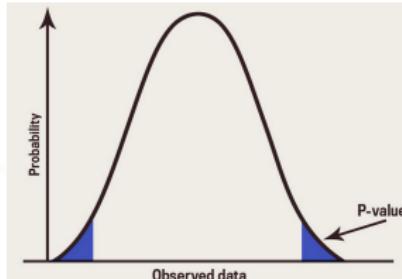
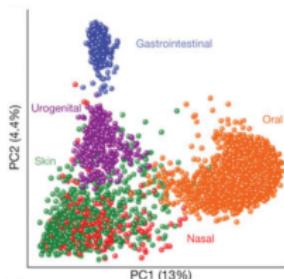
Qu'est-ce que la statistique?

Statistique : ensemble des méthodes de **réduction** des données

Définition V. Runge

Ses objectifs:

- Révéler la **structure** des données
- Permettre une prise de **décision** éclairée
- Rendre possible la **validation** d'une théorie scientifique



Qu'est-ce que la statistique?

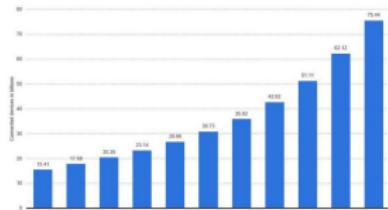
Statistique : ensemble des méthodes de **réduction** des données

Définition V. Runge

Compléments:

- ▶ La science statistique fait le lien entre **l'abstraction mathématique** et la **réalité physique** du monde
- ▶ Science en **hyper-croissance** ➡ numérisation du monde

Internet of Things - number of connected devices worldwide 2015-2025
Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (in billions)



Plan

1 Introduction

- Qu'est-ce que la statistique?
- **Courte histoire des mathématiques**
- Statistiques dans l'histoire des mathématiques

Courte histoire des mathématiques

- ▶ Les statistiques sont une branche **récente** des mathématiques
- ▶ Grèce antique (VI^{ème} à IV^{ème} siècles avant JC)
 - Arithmétique et géométrie
 - Naissance de la démonstration

Thalès, Euclide, Pythagore, Archimète



Figure: Les Éléments d'Euclide et la machine d'Anticythère

Courte histoire des mathématiques

► Mathématique arabes (~ 800 à 1500)

- Système décimal indo-arabe
- Développement de l'algèbre, des algorithmes, de la trigonométrie

Al-Khawarizmi, Al-Kashi



Figure: Al-Khawarizmi (780 (Ouzbékistan) – 850 (Bagdad)). Son nom a donné le mot *algorithme*. Le titre d'un de ses ouvrages est à l'origine du mot *algèbre*

Courte histoire des mathématiques

► Renaissance européenne

- Traduction des textes arabes
- Résolution d'équation et nombres complexes

► XVII^{ème} XVIII^{ème} siècles. Époque des Lumières

- Début de l'analyse (dérivée, intégrale, équation différentielle)
- Mathématisation de la physique (Newton)
- L'Encyclopédie (Jean le Rond d'Alembert)



Figure: Isaac Newton (1642– 1727); Leonhard Euler (1707 – 1783); Joseph-Louis Lagrange (1736 – 1813); Carl Friedrich Gauss (1777–1855)

Courte histoire des mathématiques

► XIX^{ème} siècle. Professionnalisation des mathématiques

- Mise en équation du monde (Navier-stokes, Maxwell)
- Équation de la chaleur et séries de Fourier
- Découverte par le calcul de Neptune (par Le Verrier)
- ...

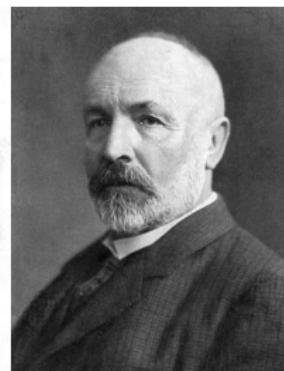


Figure: Augustin Louis Cauchy (1789 – 1857); Bernhard Riemann (1826 – 1866); Georg Cantor (1845 – 1918)

Courte histoire des mathématiques

► XX^{ème} siècle. Développement exponentiel des mathématiques

- Théorie des probabilités, théorie du chaos
- Analyse fonctionnelle, analyse numérique
- Optimisation, contrôle optimal, calcul variationnel
- Théorie de l'information, cryptographie
-

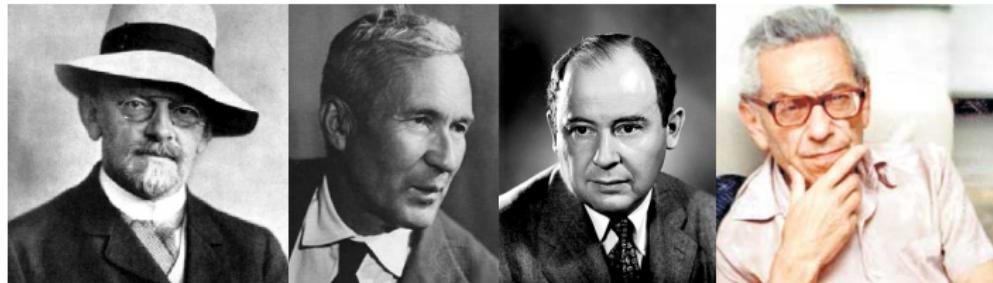


Figure: David Hilbert (1862–1943); Andrei Kolmogorov (1903–1987); John von Neumann (1903–1957); Paul Erdős (1913–1996)

Il reste beaucoup à faire...

1) congrès international des mathématiciens, tenu à Paris en août 1900,
David Hilbert propose 23 problèmes pour le XX^{ème} siècle

2) Prix du millénaire de l'**institut Clay** (10^6 \$ par problème)

- Hypothèse de Riemann
- Conjecture de Poincaré (résolu)
- Problème ouvert $P = NP$
- Conjecture de Hodge
- Conjecture de Birch et Swinnerton-Dyer
- Équations de Navier-Stokes
- Équations de Yang-Mills

Plan

1 Introduction

- Qu'est-ce que la statistique?
- Courte histoire des mathématiques
- Statistiques dans l'histoire des mathématiques

Statistiques dans l'histoire des mathématiques

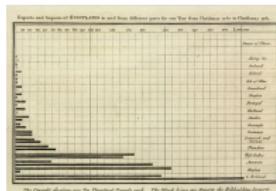
Et les statistiques ? ► 3 grandes périodes historiques

► ? – XIX^{ème} siècle. Statistiques descriptives

- "Statistique" = recensement bétails...
Chine, Égypte (XVIII^{ème} siècle avant JC)
- Statistique = Staatskunde (*service de l'État*) XVIII^{ème} siècle
= collecte de données tenue par les intendants de l'État

William Playfair (1759–1823)

Commercial and Political Atlas
(1786). Le 1^{er} histogramme connu



- Karl Pearson impose (1893) l'écart-type (standard deviation) σ

Statistiques descriptives → Analyse des données (XX^{ème} siècle)

Statistiques dans l'histoire des mathématiques

Et les statistiques ? ► 3 grandes périodes historiques

► 1880 – 1925. Début des statistiques inférentielles/mathématiques

- Rationalisation de la production (industrielle et agricole)
- Gestion des empires coloniaux, recensements,...
- Développement des sondages d'opinion

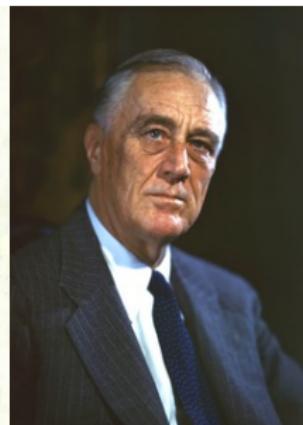
Statistiques dans l'histoire des mathématiques

Et les statistiques ? ► 3 grandes périodes historiques

► 1880 – 1925. Début des statistiques inférentielles/mathématiques

Élection USA (1936)

- Magazine hebdomadaire **Literary Digest**
- Sondage 2 300 000 réponses : Landon 55% vs Roosevelt 45%. En théorie erreur < 0.1%
- Roosevelt réélu avec 61% !



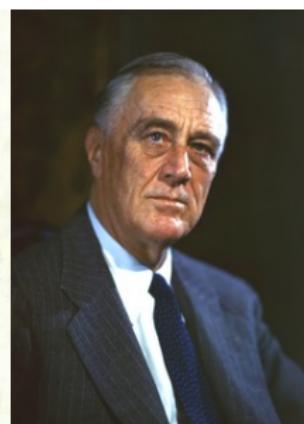
Statistiques dans l'histoire des mathématiques

Et les statistiques ? ► 3 grandes périodes historiques

► 1880 – 1925. Début des statistiques inférentielles/mathématiques

Élection USA (1936)

- Magazine hebdomadaire **Literary Digest**
- Sondage 2 300 000 réponses : Landon 55% vs Roosevelt 45%. En théorie erreur < 0.1%
- Roosevelt réélu avec 61% !
- ☞ Problème : Échantillon = lecteur du journal + listes de propriétaires de voitures et d'abonnés au téléphone



Statistiques dans l'histoire des mathématiques

Et les statistiques ? ► 3 grandes périodes historiques

► 1950 – aujourd'hui. Statistiques en grandes dimensions

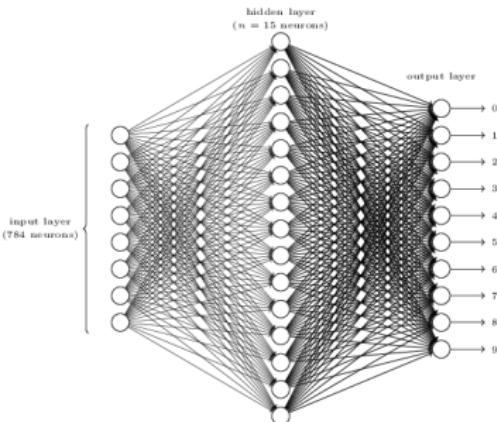
Influence de l'informatique

- Big data
 - Analyse exploratoire des données (multi-d)
 - Efficacité des algorithmes
- Apprentissage automatique
 - Machine Learning
 - Deep Learning
 - Intelligence artificielle

Statistiques dans l'histoire des mathématiques

► Statistiques en grandes dimensions. Exemple

Réseaux de neurones \approx Intelligence artificielle

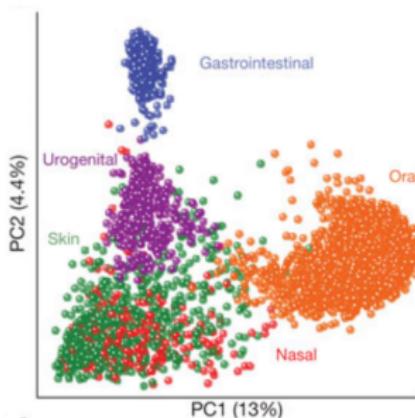


Statistiques dans l'histoire des mathématiques

► Analyse des données (cf statistique descriptive). Exemple

Analyse en composante principale (ACP)

Projeter au mieux les grandes dimensions sur la dimension 2 (pour visualiser, analyser,...).



De nombreux nouveaux métiers :

Computer scientist, data scientist, data analyst, bio-informaticien, astro-statisticien, machine learning engineer, data mining expert, AI Researcher...

Statut :

Au début peu considérée, la statistique est aujourd'hui en **très forte expansion** et **intégrée** au bagage standard du mathématicien

Aujourd'hui :

Plus de distinction nette entre mathématiques appliquées et mathématiques pures

Plan

1 Introduction

2 Des statistiques pour quoi faire?

3 Statistiques descriptives

4 Estimateurs et estimations

5 Tests statistiques

Plan

② Des statistiques pour quoi faire?

- Mission du statisticien
- Exemples
- Statistiques et probabilités
- Objectifs du cours!

Mission du statisticien

☞ Expérience statistique classique

- Planification de l'expérience
- Recueil des données
- Organisation des données (statistiques descriptives)
- Analyse des données (statistiques inférentielles)

Plan

② Des statistiques pour quoi faire?

- Mission du statisticien
- **Exemples**
- Statistiques et probabilités
- Objectifs du cours!

Exemples

☞ Tester une hypothèse

Statistiques routières

- Efficacité des images violentes pour les spots de prévention routière ?
- Baisse de la mortalité sur la route avec la limitation à 80km/h ?
- Efficacité du casque pour les cyclistes ?

Santé

- Dangérosité du glyphosate pour la santé ?
- Évaluer l'impact des OGM sur la santé
- Balance bénéfice/risque d'un vaccin

☞ La maîtrise des statistiques est **essentielle** au biologiste, à l'économiste, au sociologue, au physicien...

Biologie

Combien de souris sacrifier pour prouver une hypothèse? Ou quand s'arrêter de les sacrifier?

Économie

Mesurer l'effet économique des Jeux Olympiques?

Sociologie

Performance et mesure d'inégalité des systèmes d'éducation?

Physique fondamentale

Séparer le bruit des ondes gravitationnelles?

Plan

② Des statistiques pour quoi faire?

- Mission du statisticien
- Exemples
- **Statistiques et probabilités**
- Objectifs du cours!

Statistiques et probabilités

La statistique se base sur **une modélisation probabiliste** des données

- On observe une variable aléatoire X de loi **partiellement inconnue** plusieurs fois
- On veut en tirer des conclusions globales (intrinsèques) sur X

☞ Exemple : jeté de dés

On jette 100 fois un dé avec pour résultats

numéro	1	2	3	4	5	6
probabilité théorique	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$
occurrences sur 100 lancés	13	16	15	30	14	12

Le dé est-il équilibré?

☞ Autre exemple : Petits pois de Mendel

Plan

② Des statistiques pour quoi faire?

- Mission du statisticien
- Exemples
- Statistiques et probabilités
- Objectifs du cours!

Objectifs du cours!

Statistiques

= 4 types de problèmes

= **Statistique exploratoire, Estimation, Classification, Régression.**

Nos 3 objectifs :

► Statistiques descriptives

Objectif : "voir" un ensemble de données

Statistiques inférentielles

► Estimation statistique

Objectif : généraliser "au mieux" du "petit nombre au grand nombre"

► Tests statistiques

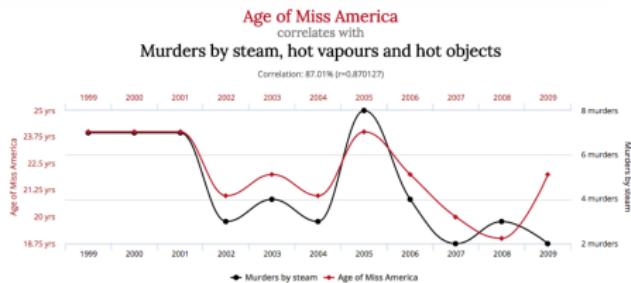
Objectif : Prendre une décision raisonnée

Références internet

- ➊ Corrélations bizarres :

<http://tylervigen.com/spurious-correlations>

- ➋ Chaîne Youtube : la statistique expliquée à mon chat
- ➌ MOOC FUN. Introduction à la statistique avec R



Plan

- 1 Introduction
- 2 Des statistiques pour quoi faire?
- 3 Statistiques descriptives
- 4 Estimateurs et estimations
- 5 Tests statistiques

C'est quoi les statistiques descriptives?

1) C'est déterminer le **cadre d'étude** :

- Population, échantillon
- Type de variable

2) construire des **résumés** (de 3 types) :

- Indicateurs de position
- Indicateurs de dispersion
- Représentations graphiques

☞ La statistique descriptive se concentre sur le cas de la dimension 1.
Une observation = un nombre.

Plan

3 Statistiques descriptives

- Cadre d'Etude
- Les différentes moyennes
- Indicateurs de dispersion
- Ordonner
- Représentations graphiques

Vocabulaire

Population : ensemble (grand, voire infini) d'individus ou d'objets de même nature

Echantillon : sous-ensemble de la population

Variable : une caractéristique de la population pouvant prendre différentes modalités

Modalité : toute valeur que peut prendre une variable

Série statistique : Ensemble des données recueillies pour une variable donnée

Compléments

- On utilise parfois le mot **caractère** à la place de **variable**
- Le **mode** est la modalité la plus présente

Types de variables

- Variable **quantitative** : les modalités sont *des nombres*
 - **discrète** : elles prennent un nombre fini ou dénombrable de valeurs
 - **continue** : elles prennent toutes les valeurs d'un intervalle des réels
- Variable **qualitative** : les modalités ne sont *pas des nombres*
 - **ordinale** : elles peuvent être ordonnées
 - **nominale** : elles ne peuvent pas être ordonnées

Exercice

Exercice 1 : les variables suivantes sont-elles quantitatives ou qualitatives. Discrètes, continues, ordinaires ou nominales ?

- Les marques de smartphone des étudiants de l'Université d'Evry
- La nationalité des touristes visitant le musée Picasso de Paris
- L'âge des utilisateurs du site www.arte.tv
- La taille des poissons pêchés par une équipe de biologistes marins
- Les notes sur 20 obtenues à ce cours de Statistiques
- Le niveau de satisfaction des utilisateurs d'un service de livraison (5 niveaux : faible – moyen – bon – très bon – excellent)

Exercice 2 : Définir la population, un échantillon (possible), la variable étudiée et les modalités de chaque exemple précédent.

Plan

3 Statistiques descriptives

- Cadre d'Etude
- **Les différentes moyennes**
- Indicateurs de dispersion
- Ordonner
- Représentations graphiques

Les différentes moyennes

Moyenne arithmétique

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

- ☞ La moyenne arithmétique a le désavantage d'être **sensible aux valeurs extrêmes** : on la dit alors **peu robuste**. Si dans la série $\{1, 2, 3\}$ une erreur de saisie est faite et 3 devient 30, la moyenne passera de 2 à 11! (mais la médiane est inchangée)

Les différentes moyennes

Moyenne géométrique

$$\bar{x}_G = \sqrt[n]{x_1 \times \dots \times x_n}$$

Moyenne harmonique

$$\bar{x}_H = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}}$$

Moyenne quadratique

$$\bar{x}_Q = \sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2}$$

► **Théorème :** Si $x_1, \dots, x_n > 0$, alors

$$\min(x_1, \dots, x_n) \leq \bar{x}_H \leq \bar{x}_G \leq \bar{x} \leq \bar{x}_Q \leq \max(x_1, \dots, x_n)$$

Les différentes moyennes

Théorème : Si $x_1, \dots, x_n > 0$, alors

$$\min(x_1, \dots, x_n) \leq \bar{x}_H \leq \bar{x}_G \leq \bar{x} \leq \bar{x}_Q \leq \max(x_1, \dots, x_n)$$

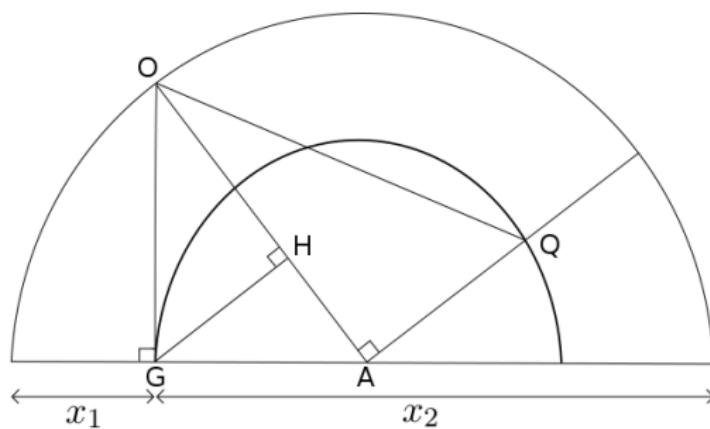


Figure: Moyennes de x_1 et x_2 : harmonique OH , géométrique OG , arithmétique OA et quadratique OQ

Exercice

► Calcul de rendement

- Votre banquier vous propose de placer 1000€ pendant 5 ans aux taux annuels progressifs de 1%, 2%, 3%, 4% et 5%. Il annonce que le pourcentage annuel moyen est de 3%. Le croyez-vous?

► Vitesse moyenne

- Un cycliste parcourt un kilomètre à l'aller et un autre au retour aux vitesses respectives de 30 km/h puis 20 km/h. Quelle est sa vitesse moyenne?

► Des câbles

- On sait que la résistance d'un câble est proportionnel à sa section. Par quel diamètre de câble faut-il remplacer 3 câbles de diamètres 3 cm, 5 cm, et 7 cm?

Plan

3 Statistiques descriptives

- Cadre d'Etude
- Les différentes moyennes
- Indicateurs de dispersion
- Ordonner
- Représentations graphiques

Variance et écart-type

Pour les données x_1, \dots, x_n la dispersion est mesurée par la **variance** Var

$$Var = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Théorème de König-Huygens

$$Var = \bar{x^2} - \bar{x}^2$$

avec

$$\bar{x^2} = \frac{1}{n} \sum_{i=1}^n x_i^2 \quad \text{et} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

On a par ailleurs l'**écart-type** σ défini par $\sigma^2 = Var$:

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Propriété de la variance

Données de taille n_1 , moyenne \bar{x}_1 et variance σ_1^2

Données de taille n_2 , moyenne \bar{x}_2 et variance σ_2^2

Données de taille $n_1 + n_2$, moyenne \bar{x} variance σ^2

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}$$

C'est la somme de la moyenne (pondérée) des variances et de la variance (pondérée) des moyennes.

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

On écrit aussi

$$\sigma^2 = \overline{Var(x_1, x_2)} + Var(\bar{x}_1, \bar{x}_2),$$

Propriété de la variance

On introduit la suite $\{x_k\}_{k=1 \dots (n_1+n_2)}$ dans laquelle les n_1 premiers nombres sont issus de la première collection. On calcule

$$\begin{aligned}
 (n_1 + n_2)\sigma^2 &= \sum_{k=1}^{n_1} (x_k - \bar{x})^2 + \sum_{k=n_1+1}^{n_1+n_2} (x_k - \bar{x})^2 \\
 &= \sum_{k=1}^{n_1} ((x_k - \bar{x}_1) + (\bar{x}_1 - \bar{x}))^2 + \sum_{k=n_1+1}^{n_1+n_2} ((x_k - \bar{x}_2) + (\bar{x}_2 - \bar{x}))^2 \\
 &= n_1\sigma_1^2 + 2 \sum_{k=1}^{n_1} (x_k - \bar{x}_1)(\bar{x}_1 - \bar{x}) + n_1(\bar{x}_1 - \bar{x})^2 \\
 &\quad + n_2\sigma_2^2 + 2 \sum_{k=n_1+1}^{n_1+n_2} (x_k - \bar{x}_2)(\bar{x}_2 - \bar{x}) + n_2(\bar{x}_2 - \bar{x})^2,
 \end{aligned}$$

en utilisant la définition des moyennes \bar{x}_1 et \bar{x}_2 chacune des deux sommes restantes s'annule et le résultat est démontré.

Analyse de la variance ANOVA (Hors programme)

Interprétation de la formule

$$\sigma^2 = \text{variance intrapopulation} + \text{variance interpopulation}$$

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1(\bar{x}_1 - \bar{x})^2 + n_2(\bar{x}_2 - \bar{x})^2}{n_1 + n_2}.$$

$$(n_1 + n_2)\sigma^2 = SCE = \text{somme des carrés des écarts}$$

$$SCE_{\text{total}} = SCE_{\text{residu}} + SCE_{\text{facteur}}$$

-
- ▶ Étudier $\frac{SCE_{\text{facteur}}}{SCE_{\text{residu}}}$ pour savoir si \bar{x}_1 et \bar{x}_2 sont "les mêmes" !
 - ▶ C'est une possibilité de test statistique...

Écart moyen

$$e = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

- + : facile à comprendre que σ^2
- - : il est peu utilisé : problème algébrique du calcul + problème statistique

Données de taille n_1 , moyenne \bar{x}_1 et écart moyen e_1

Données de taille n_2 , moyenne \bar{x}_2 et écart moyen e_2

Donnée $n_1 + n_2$, moyenne \bar{x} écart moyen = ?

Les moments

Moments centrés d'ordre r

$$m_r = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^r$$

- coefficient d'asymétrie de Fisher (skewness) :

$$\gamma_1 = \frac{m_3}{\sigma^3}$$

- coefficient d'aplatissement de Fisher (kurtosis) :

$$\gamma_2 = \frac{m_4}{\sigma^4} - 3$$

- Distribution étalée vers la droite si $\gamma_1 > 0$, vers la gauche si $\gamma_1 < 0$.
- Pour γ_2 , la soustraction par 3 correspond à une comparaison avec la série étalon (issue d'une loi normale pour laquelle $\gamma_2 = 0$).

Exercice : Transformation affine

Soient a et b des nombres réels. On a :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{et} \quad \text{Var}(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Montrer que pour les données $y_i = ax_i + b$ on obtient les formules:

- $\bar{y} = a\bar{x} + b$
- $\text{Var}(y) = a^2 \text{Var}(x)$

L'écart moyen ne possède pas de telle propriété

Remarque de vocabulaire

les moyennes

On parle de

- **Moyenne observée** pour \bar{x} , \bar{x}_G , \bar{x}_H ...
 - **Espérance** de la variable aléatoire X pour $\mathbb{E}[X]$
 - **Moyenne empirique** (pour un estimateur...voir plus loin...)

les variances

On parle de

- **Variance observée** pour Var en statistiques descriptives
 - **Variance** de la variable aléatoire X pour $V(X)$
 - **Variance empirique** (pour un estimateur...voir plus loin...)

► **DANGER!!!** Le vocabulaire de la statistique est très instable, on utilise souvent le même mot de "moyenne" pour des notions différentes... 

Plan

3 Statistiques descriptives

- Cadre d'Etude
- Les différentes moyennes
- Indicateurs de dispersion
- **Ordonner**
- Représentations graphiques

Quantiles

On note $x_{(1)}, x_{(2)}, \dots, x_{(n)}$, avec des indices entre parenthèses, les données triées par ordre croissant.

Par exemple $\{7, 3, 2, 4, 7\}$ on aura $x_{(1)} = 2$, $x_{(2)} = 3$, $x_{(3)} = 4$, $x_{(4)} = 7$ et $x_{(5)} = 7$

Le **quantile d'ordre** $\alpha \in]0, 1[$, noté q_α , est défini comme étant "la plus petite valeur" de la série permettant d'avoir au moins αn nombres inférieurs (ou égaux) à lui-même.

$$q_\alpha = \begin{cases} x_{(\lceil \alpha n \rceil)} & \text{si } \alpha n \notin \mathbb{N} \\ \frac{x_{(\alpha n)} + x_{(\alpha n + 1)}}{2} & \text{si } \alpha n \in \mathbb{N} \end{cases}$$

avec $\lceil \cdot \rceil$ désignant l'opération partie entière supérieure.

Quantiles

Principaux quantiles

- 4 parts égales, on a 3 **quartiles** (Q_1, Q_2, Q_3) = $(q_{\frac{1}{4}}, q_{\frac{2}{4}}, q_{\frac{3}{4}})$
- 10 parts égales, on a 9 **déciles** (D_1, \dots, D_9) = $(q_{\frac{1}{10}}, \dots, q_{\frac{9}{10}})$
- 100 parts égales, on a 99 **centiles** (C_1, \dots, C_{99}) = $(q_{\frac{1}{100}}, \dots, q_{\frac{99}{100}})$

La médiane

Indicateur de position centrale Me :

$$Me = Q_2 = D_5 = C_{50} = q_{\frac{1}{2}}$$

De nouveaux indicateurs de dispersion

Interquartiles

- Intervalle interquartile :

$$IIQ = [Q_1, Q_3]$$

- Ecart interquartile :

$$IQ = Q_3 - Q_1$$

Étendue

$$e = \max_{i=1,\dots,n} \{x_i\} - \min_{i=1,\dots,n} \{x_i\} = e_{(n)} - e_{(1)}$$

Exercice

Soient les deux séries statistiques triées suivantes

(1) : 1 2 3 11 12 13 21 22 23 31 32 33 **34**

(2) : 1 2 3 11 12 13 21 22 23 31 32 33

Déterminer les 3 quartiles de ces deux séries, l'écart interquartile et l'étendue. Voir que l'introduction du cas $\alpha n \in \mathbb{N}$ est justifiée par la série (2)

Plan

3 Statistiques descriptives

- Cadre d'Etude
- Les différentes moyennes
- Indicateurs de dispersion
- Ordonner
- Représentations graphiques

Boîte à moustaches (boxplot)

- Le long d'un axe gradué, on représente un **rectangle** délimité par les quartiles Q_1 et Q_3 et coupé en deux par un trait à hauteur de Me ;
- Cette boîte est complétée par des **moustaches** (des traits reliant la boîte) s'arrêtant (selon la définition choisie):
 - au minimum et au maximum des x_i
 - à D_1 et D_9 ou C_2 et C_{98}
 - à la valeur $x_{(a)}$ réalisant le minimum des $x_{(i)}$ dans $[Q_1 - 1.5 \text{ } I\!Q, Q_1]$ et $x_{(b)}$ réalisant le maximum des $x_{(i)}$ dans $[Q_3, Q_3 + 1.5 \text{ } I\!Q]$
- La dernière définition est souvent préférée. Les valeurs en dehors des moustaches sont repérées par **des croix ou des ronds**. Ce sont des valeurs dites extrêmes (**outliers**).

Boîte à moustaches (boxplot)

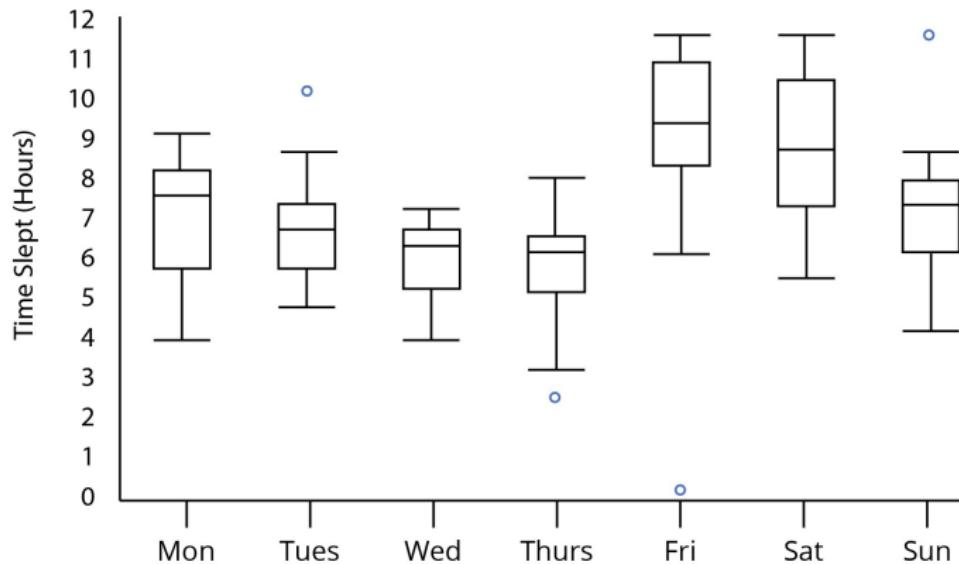


Figure: Nombre d'heures de sommeil pour 20 lycéens pendant une semaine.

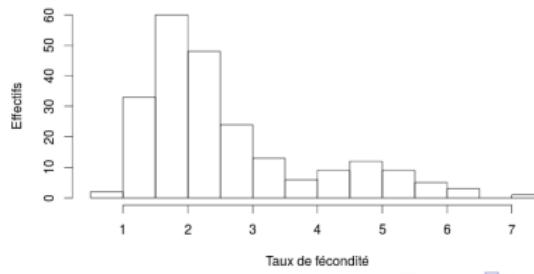
Histogrammes

- ▶ Rectangles adjacents, uniquement pour les **variables continues**
- ▶ Aire de chaque rectangle proportionnelle à l'effectif de la classe

En notant h_i la hauteur de la classe i ($[a_{i-1}, a_i[$) d'effectif n_i , défini par $n_i = \#\{x_k \text{ tels que } x_k \in [a_{i-1}, a_i[\}$, on obtient

$$h_i = \frac{n_i}{a_i - a_{i-1}}$$

Histogramme des différents taux de fécondité de tous les pays (225) estimés en 2013 (source: *World Factbook*)



Diagrammes en bâtons

- ▶ Bâtons adjacents, uniquement pour les **variables discrètes**
- ▶ Hauteur de chaque bâton proportionnelle au nombre d'occurrences de la modalité placée sous le bâton

Exemple : On jette 20 fois de suite un dé équilibré à 6 faces et on reporte le résultat sur le diagramme suivant

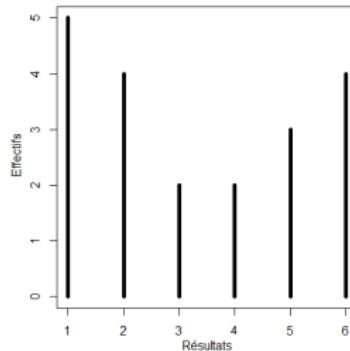


Figure: Résultat de 20 lancés d'un dé équilibré

Fonction de répartition empirique

- ▶ = Graphique des fréquences cumulées
- ▶ Pour n données,

$$F(x) = \frac{\text{nombre d'éléments} \leq x}{n}$$

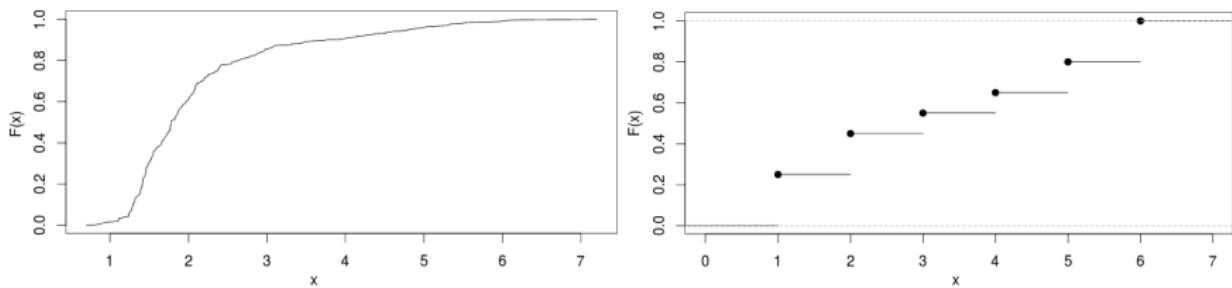


Figure: (gauche) fonction de répartition des taux de fécondité. (droite) fonction de répartition des 20 lancés de dé

Tableaux

- ▶ Tableau de contingence
- ▶ On croise deux variables d'une population

Cheveux Yeux \	BLONDS	CHÂTAINS	BRUNS	NOIRS	ROUX	TOTAL
BLEUS	2	5	2	1	0	10
VERTS	1	1	2	1	1	6
BRUNS	1	5	6	2	0	14
TOTAL	4	11	10	4	1	30

Figure: Répartition couleur des yeux et type de cheveux dans une classe de 30 élèves

- ☞ Objectif : détection d'éventuelles dépendances entre variables

Aller plus loin...

La statistique descriptive est **unidimensionnelle**

☞ On organise les réels x_1, \dots, x_n

L'analyse de donnée est **multidimensionnelle**

☞ On organise les données $(x_{11}, \dots, x_{1p}), \dots, (x_{n1}, \dots, x_{np})$

De nouvelles questions émergent:

- Comment mesurer la liaison entre 2 variables?
- Comment visualiser des données en dimension 10?

Plan

- 1 Introduction
- 2 Des statistiques pour quoi faire?
- 3 Statistiques descriptives
- 4 Estimateurs et estimations
- 5 Tests statistiques

Plan

4 Estimateurs et estimations

- Générer l'aleatoire
- L'échantillonnage et problématique
- Estimation ponctuelle
- Estimation par intervalle
- Le théorème central limite
- Propagation des incertitudes

Générer des données de loi uniforme $X \sim \mathcal{U}([0, 1])$

- Une variable aléatoire $X = (\Omega, \{\omega \in \Omega, P(X = \omega)\})$
 Ω est l'univers, l'ensemble des valeurs possibles pour X

Sources d'aléatoires (générer $X \sim \mathcal{U}([0, 1])$ loi uniforme)

- L'aléatoire dans **la nature** <https://www.random.org/>
Bruit atmosphérique généré par les éclairs! (capté par une simple radio)



ou l'instant d'une désintégration radioactive

- Les **algorithmes** (!) pseudo-aléatoires (très très utilisés)

Exemples d'algorithmes pseudo-aléatoires

► Générateur congruentiel linéaire (Lehmer 1948)

$$x_{n+1} = (ax_n + c) \bmod m$$

Set the seed x_0 using the current system time in microseconds

```
x0 <- as.numeric(Sys.time()) * 1000
```

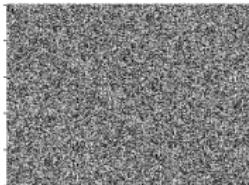


Figure: Le générateur d'UNIX et GCC. $a = 1103515245$, $c = 12345$, $m = 2^{31}$

Ici : seed = $x_0 = 99$, $2^{16} = 65536$ valeurs. $x_i/(2^{31}) \in [0, 1[$

► État de l'art : algorithme **Mersenne Twister**

(Makoto Matsumoto et Takuji Nishimura 1997)

utilisé par Python, Ruby, R, PHP, MATLAB, Stata, C++

Comparaisons

Characteristic	Pseudo-Random Number Generators	True Random Number Generators
Efficiency	Excellent	Poor
Determinism	Deterministic	Nondeterministic
Periodicity	Periodic	Aperiodic

Application	Most Suitable Generator
Lotteries and Draws	TRNG
Games and Gambling	TRNG
Random Sampling (e.g., drug screening)	TRNG
Simulation and Modelling	PRNG
Security (e.g., generation of data encryption keys)	TRNG
The Arts	Varies

source : <https://www.random.org/>

Plan

4 Estimateurs et estimations

- Générer l'aleatoire
- L'échantillonnage et problématique
- Estimation ponctuelle
- Estimation par intervalle
- Le théorème central limite
- Propagation des incertitudes

L'échantillonnage

Définitions

Échantillon aléatoire : variables aléatoires (X_1, \dots, X_n) avec $X_i = X$

Échantillon : réalisations x_1, \dots, x_n des variables X_1, \dots, X_n

Modélisation : $X_i \sim \mathcal{L}(\theta)$ (loi qui dépend d'un paramètre θ)

Estimateur : variable aléatoire $T = f(X_1, \dots, X_n)$

Estimation : réalisation t de T : $t = f(x_1, \dots, x_n)$

- ▶ Hypothèse : Les variables aléatoires X_1, \dots, X_n sont **indépendantes et identiquement distribuées** (on note i.i.d.). Elles ont toutes la même loi que la variable aléatoire X appelée variable aléatoire parente.

Problématique de l'estimation

Notre objectif

- Soit X une variable aléatoire de loi $\mathcal{L}(\theta)$
- Soit x_1, \dots, x_n une observation de l'échantillon X_1, \dots, X_n , copies de X
- Comment estimer θ à partir de x_1, \dots, x_n ?

Deux étapes :

- ▶ Estimation ponctuelle
- ▶ Estimation par intervalle de confiance

Plan

4 Estimateurs et estimations

- Générer l'aléatoire
- L'échantillonnage et problématique
- **Estimation ponctuelle**
- Estimation par intervalle
- Le théorème central limite
- Propagation des incertitudes

Exemples à connaître

- Moyenne empirique :

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

- Variance empirique : $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2$

- Variance empirique corrigée : $S^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n}{n-1} S^2$

Définition

Estimateur ponctuel

= estimateur dont la réalisation est une estimation du paramètre θ

Notation

On note généralement $\hat{\theta}$ ou $\hat{\theta}_n$ l'estimateur (=une variable aléatoire) de θ .

$$\hat{\theta} = \hat{\theta}_n = f(X_1, \dots, X_n)$$

Propriétés d'un estimateur

Définitions

- On appelle **biais** d'un estimateur la quantité :

$$b(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$

- Un estimateur est dit **sans biais** si

$$b(\hat{\theta}_n) = 0$$

- Un estimateur est dit **asymptotiquement sans biais** si

$$\lim_{n \rightarrow +\infty} b(\hat{\theta}_n) = 0$$

Propriétés d'un estimateur

Définitions

- On appelle **erreur quadratique moyenne** la quantité :

$$EQM(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2] = V(\hat{\theta}_n) + [b(\hat{\theta}_n)]^2$$

- Un estimateur est dit **consistant** (ou **convergent** en moyenne quadratique) si :

$$\lim_{n \rightarrow \infty} EQM(\hat{\theta}_n) = 0$$

Remarque

Pour montrer qu'un estimateur sans biais est consistant, il suffit de montrer que $\lim_{n \rightarrow \infty} Var(\hat{\theta}_n) = 0$

Exercice

Soit (X_1, \dots, X_n) un échantillon de **loi quelconque** tel que
 $\mathbb{E}[X_i] = \mu$ et $V(X_i) = \sigma^2$

- Montrer que $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais et consistant de μ
- Montrer que $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur biaisé de σ^2
- En déduire un estimateur sans biais de σ^2

Estimateur de la moyenne

Moyenne empirique

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

Propriétés

Soit (X_1, \dots, X_n) un échantillon aléatoire tel que $\mathbb{E}[X_i] = \mu$ et $V(X_i) = \sigma^2$

$$\mathbb{E}[\bar{X}] = \mu \quad \text{et} \quad V(\bar{X}) = \frac{\sigma^2}{n}$$

Remarque

\bar{X} est un estimateur **sans biais** et **consistant** de μ

Estimateur de la variance

Variance empirique et variance empirique corrigée

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{et} \quad S^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Propriétés

Soit (X_1, \dots, X_n) un échantillon tel que $\mathbb{E}[X_i] = \mu$ et $V(X_i) = \sigma^2$

$$\mathbb{E}[S^2] = \frac{n-1}{n} \sigma^2 \quad \text{et} \quad V(S^2) = \frac{n-1}{n^3} ((n-1)\mu_4 - (n-3)\sigma^2)$$

Remarque

$S^{*2} = \frac{n}{n-1} S^2$ est un estimateur **sans biais** et **consistant** de σ^2

Maximum de vraisemblance (Hors programme)

- ▶ Existe-t-il une méthode générale pour choisir l'estimateur des paramètres d'un loi?
- ☞ Oui! C'est la méthode du maximum de vraisemblance!
(*maximum likelihood*)

Pour une densité de probabilité $f = f(x, \theta)$.

Exemple distribution exponentielle: Exemple distribution normale:

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

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

$$L(x_1, \dots, x_n, \theta) = f(x_1, \theta) \times \cdots \times f(x_n, \theta)$$

Résoudre $\max_{\theta} L(x_1, \dots, x_n, \theta)$
 ou $\frac{d}{d\theta} L(x_1, \dots, x_n, \theta) = 0$ ou $\frac{d}{d\theta} \log(L(x_1, \dots, x_n, \theta)) = 0$

Maximum de vraisemblance (Hors programme)

Exemple avec la loi exponentielle:

Avec

$$f(x, \theta) = \begin{cases} \theta e^{-\theta x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

On obtient (par intégration par parties) :

$$\mathbb{E}[X] = \int_{x=0}^{+\infty} \theta x e^{-\theta x} dx = \frac{1}{\theta}$$

Une bonne idée (?) serait d'utiliser $\hat{\theta} = \frac{1}{\bar{X}}$. Mais est-ce un bon choix?

Maximum de vraisemblance (Hors programme)

Exemple avec la loi exponentielle : $f(x, \theta) = \begin{cases} \theta e^{-\theta x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$

$$L(x_1, \dots, x_n, \theta) = \prod_{i=1}^n \theta \exp(-\theta x_i) = \theta^n \exp\left(-\theta \sum_{i=1}^n x_i\right) = \theta^n \exp(-\theta n \bar{x})$$

Et on a :

$$\frac{d}{d\theta}(\log(L)) = \frac{d}{d\theta}(n \log(\theta) - \theta n \bar{x}) = \frac{n}{\theta} - n \bar{x}$$

Ainsi

$$\frac{d}{d\theta}(\log(L)) = 0 \iff \boxed{\theta = \frac{1}{\bar{X}}}$$

Un meilleur estimateur qui minimise l'erreur quadratique moyenne sera

$$\hat{\theta} = \frac{n-2}{n} \frac{1}{\bar{X}}$$

Plan

4 Estimateurs et estimations

- Générer l'aléatoire
- L'échantillonnage et problématique
- Estimation ponctuelle
- Estimation par intervalle**
- Le théorème central limite
- Propagation des incertitudes

Estimation par intervalle

Principe

- Trouver une variable aléatoire $B(\theta)$ de loi connue (indépendante de θ) pour laquelle le paramètre θ inconnu intervient
 - Recherche les quantiles $q_{\frac{\alpha}{2}}$ et $q_{1-\frac{\alpha}{2}}$ qui permettent d'écrire
$$P(q_{\frac{\alpha}{2}} < B(\theta) < q_{1-\frac{\alpha}{2}}) = 1 - \alpha$$
 - Transformer l'inégalité en une inégalité du type
$$P(B_1 < \theta < B_2) = 1 - \alpha$$
-
- ▶ Les bornes B_1 et B_2 de l'intervalle sont des variables aléatoires
 - ▶ $[B_1, B_2]$ est un **intervalle de probabilité** de niveau $1 - \alpha$ pour θ

Estimation par intervalle

Définition

Soient $B_1 = f_1(X_1, \dots, X_n)$ et $B_2 = f_2(X_1, \dots, X_n)$.

$[B_1, B_2]$ est un **intervalle de probabilité** de niveau $1 - \alpha$ du paramètre θ si :

$$P(B_1 < \theta < B_2) = 1 - \alpha$$

Définition

Un **intervalle de confiance** de niveau $1 - \alpha$ du paramètre θ est une réalisation $[b_1, b_2]$ de l'intervalle $[B_1, B_2]$.

Exercice

Pour déterminer la teneur en potassium d'une solution, on effectue des dosages à l'aide d'une technique expérimentale donnée.

On admet que le résultat d'un dosage est une variable aléatoire suivant une distribution normale $\mathcal{N}(\mu, \sigma^2)$ dont l'espérance μ est la valeur que l'on cherche à déterminer, et dont l'écart-type σ est de 1 mg/litre si l'on suppose que le protocole expérimental a été suivi scrupuleusement.

Les résultats pour cinq dosages indépendants sont les suivants (en mg/litre): 74.0, 71.6, 73.4, 74.3, 72.2.

- Déterminer à partir de ces mesures un intervalle de confiance pour μ avec un coefficient de sécurité de 95%.
- Quelle taille d'échantillon est nécessaire pour avoir un intervalle plus petit que 0.1. (en mg/litre) ?

Estimation par intervalle de la moyenne

Famille gaussienne, variance connue

Soit (X_1, \dots, X_n) un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. on suppose σ^2 connu.
L'intervalle :

$$IC_{(1-\alpha)} = \left[\bar{X} - q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

est un intervalle de probabilité de niveau $(1 - \alpha)$ pour μ

Avec $q_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

Estimation par intervalle de la moyenne

Famille gaussienne, variance connue

Soit (X_1, \dots, X_n) un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. on suppose σ^2 connu.
L'intervalle :

$$IC_{(1-\alpha)} = \left[\bar{X} - q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

est un intervalle de probabilité de niveau $(1 - \alpha)$ pour μ

Avec $q_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

► Que faire si la variance est inconnue?

Loi du χ^2

Définition

Soient X_1, \dots, X_n n variables aléatoires **indépendantes** et identiquement distribuées de loi normale centrée réduite: $X_i \sim \mathcal{N}(0, 1)$

La variable aléatoire $Y = X_1^2 + \dots + X_n^2$ suit une loi continue appelée loi du χ^2 à n degrés de liberté :

$$Y = \sum_{i=1}^n X_i^2 \sim \chi_n^2$$

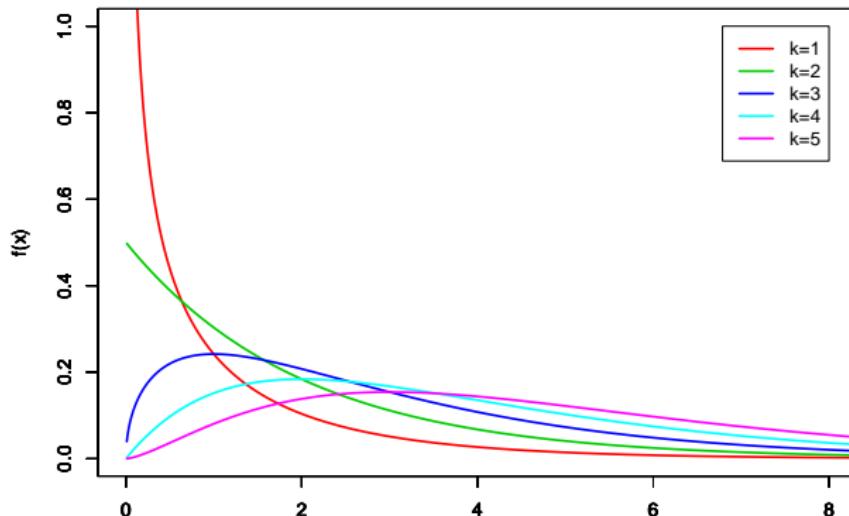
Propriétés

- Si $Y_1 \sim \chi_{n_1}^2$ et $Y_2 \sim \chi_{n_2}^2$ avec $Y_1 \perp\!\!\!\perp Y_2$, alors $Y = Y_1 + Y_2 \sim \chi_{n_1+n_2}^2$
- Si $Y \sim \chi_n^2$, alors $\mathbb{E}[Y] = n$ et $V(Y) = 2n$

Loi du χ^2

Densité

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)}x^{(n-2)/2}e^{-x/2}$$



Loi de Student

Définition

Soient X et Y deux variables aléatoires **indépendantes** telles que $X \sim \mathcal{N}(0, 1)$ et $Y \sim \chi_n^2$. La variable aléatoire $T = \frac{X}{\sqrt{Y/n}}$ suit une loi continue appelée loi de Student à n degrés de liberté :

$$T = \frac{X}{\sqrt{Y/n}} \sim t_n$$

Propriétés

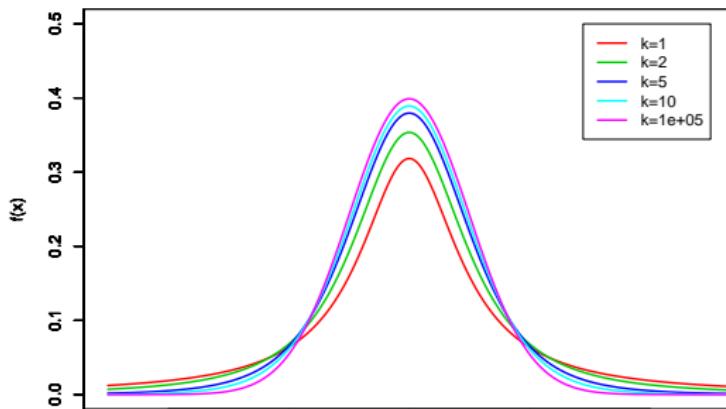
- $\mathbb{E}[T] = 0$
- $V(T) = \frac{n}{n-2}$ si $n > 2$

Loi de Student

Densité

C'est une loi "proche" de la loi normale centrée réduite

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}$$



Théorème de Cochran

Théorème

Soit (X_1, \dots, X_n) un échantillon de la variable aléatoire X où $X \sim \mathcal{N}(\mu, \sigma^2)$.

On sait que

- $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{Q^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2$
- Q^2 et \bar{X} sont **indépendants**.

Théorème de Cochran

Théorème

Soit (X_1, \dots, X_n) un échantillon de la variable aléatoire X où $X \sim \mathcal{N}(\mu, \sigma^2)$.

On sait que

- $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
- $\frac{Q^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2$
- Q^2 et \bar{X} sont **indépendants**.

Ainsi avec $U = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$ et $V = \frac{Q^2}{\sigma^2} \sim \chi_{n-1}^2$ on obtient

$$T = \frac{U}{\sqrt{\frac{V}{n-1}}} \sim t_{n-1} \quad \text{et} \quad \frac{U}{\sqrt{\frac{V}{n-1}}} = \frac{\bar{X} - \mu}{\frac{S^*}{\sqrt{n}}} \sim t_{n-1}$$

Estimation par intervalle de la moyenne

Famille gaussienne, variance inconnue

Soit (X_1, \dots, X_n) un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. on suppose σ^2 inconnu.
L'intervalle :

$$IC_{(1-\alpha)} = \left[\bar{X} - q_{1-\alpha/2}^{t_{n-1}} \frac{S^*}{\sqrt{n}}, \bar{X} + q_{1-\alpha/2}^{t_{n-1}} \frac{S^*}{\sqrt{n}} \right]$$

est un intervalle de probabilité de niveau $(1 - \alpha)$ pour μ

Avec $q_{1-\alpha/2}^{t_{n-1}}$ le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 1$ degrés de liberté.

Estimation par intervalle

Pour déterminer la concentration en glucose d'un échantillon sanguin, on effectue des dosages à l'aide d'une technique expérimentale donnée. On considère que le résultat de chaque dosage est une variable aléatoire normale. On effectue 10 dosages indépendants, qui donnent les résultats suivants (en g/l) :

0.96, 1.04, 1.08, 0.92, 1.04, 1.18, 0.99, 0.99, 1.25, 1.08

- Calculer une estimation de la concentration en glucose de cet échantillon.
- Calculer un interval de confiance de cette concentration de niveau 95%.

Estimation par intervalle de la variance

Famille gaussienne, variance inconnue

Soit (X_1, \dots, X_n) un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. On suppose σ^2 inconnu.
L'intervalle :

$$IC_{(1-\alpha)}(\sigma^2) = \left[\frac{(n-1)S^{*2}}{q_{1-\frac{\alpha}{2}}^{\chi_{n-1}^2}}, \frac{(n-1)S^{*2}}{q_{\frac{\alpha}{2}}^{\chi_{n-1}^2}} \right]$$

est un intervalle de probabilité de niveau $(1 - \alpha)$ pour σ^2 .

$$\frac{(n-1)S^{*2}}{\sigma^2} \sim \chi_{n-1}^2$$

Avec $q_{1-\frac{\alpha}{2}}^{\chi_{n-1}^2}$ le quantile d'ordre $1 - \alpha/2$ et $q_{\frac{\alpha}{2}}^{\chi_{n-1}^2}$ le quantile d'ordre $\alpha/2$ de la loi du Khi-2 à $n - 1$ degrés de liberté.

Plan

4 Estimateurs et estimations

- Générer l'aléatoire
- L'échantillonnage et problématique
- Estimation ponctuelle
- Estimation par intervalle
- **Le théorème central limite**
- Propagation des incertitudes

Le théorème central limite

Soient X_1, \dots, X_n n variables aléatoires **indépendantes et identiquement distribuées** d'espérance μ et de variance σ^2 :

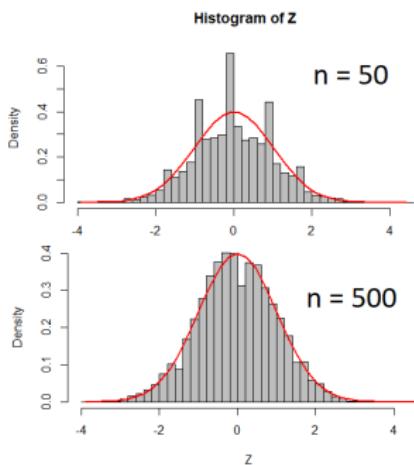
$$S = \sum_{i=1}^n X_i \xrightarrow{\mathcal{L}} \mathcal{N}(n\mu, n\sigma^2)$$

$$\Leftrightarrow Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

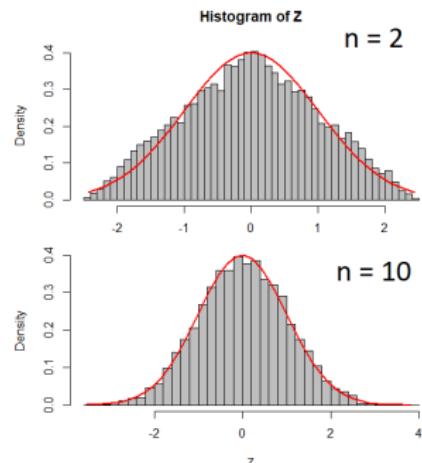
Le théorème central limite

$$\Leftrightarrow Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

$X_i \sim \mathcal{B}(m = 3, p = 0.5)$



$X_i \sim \mathcal{U}([0, 1])$



Exercice

► Jouer à pile ou face

Vous relevez 100 résultats pile-face lors du lancé successif de 100 pièces. La variable aléatoire associée $X = \frac{1}{n} \sum_{i=1}^n X_i$ avec $X_i \sim \mathcal{B}(p)$ donne une réalisation de $\bar{x} = 0.4$.

- Construire un intervalle de confiance à 95% autour de cette proportion observée.
- Pensez vous que la pièce soit équilibrée?

Réponse :

L'intervalle de confiance à construire avec une approximation de p par \bar{x} est :

$$\left[\bar{x} - 1.96 \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sqrt{\bar{x}(1-\bar{x})}}{\sqrt{n}} \right] = [0.304, 0.496]$$

Plan

4 Estimateurs et estimations

- Générer l'aléatoire
- L'échantillonnage et problématique
- Estimation ponctuelle
- Estimation par intervalle
- Le théorème central limite
- Propagation des incertitudes

Propagation des incertitudes

Il est courant de vouloir estimer la variance s_f^2 du résultat d'un calcul impliquant plusieurs mesures (variables aléatoires)

$$f = f(X, Y, Z, \dots)$$

Cela généralise le calcul de la variance pour $\bar{X} = \frac{X_1 + \dots + X_n}{n}$

Formule de la variance

Si $f = f(X, Y, Z, \dots)$ et les variables aléatoires X, Y, Z, \dots sont indépendantes, alors

$$s_f^2 = \left(\frac{\partial f}{\partial x} \right)^2 s_x^2 + \left(\frac{\partial f}{\partial y} \right)^2 s_y^2 + \left(\frac{\partial f}{\partial z} \right)^2 s_z^2 + \dots$$

Propagation des incertitudes

Function	Variance
$f = aA$	$\sigma_f^2 = a^2 \sigma_A^2$
$f = aA + bB$	$\sigma_f^2 = a^2 \sigma_A^2 + b^2 \sigma_B^2 + 2ab \sigma_{AB}$
$f = aA - bB$	$\sigma_f^2 = a^2 \sigma_A^2 + b^2 \sigma_B^2 - 2ab \sigma_{AB}$
$f = AB$	$\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 + 2 \frac{\sigma_{AB}}{AB} \right]$
$f = \frac{A}{B}$	$\sigma_f^2 \approx f^2 \left[\left(\frac{\sigma_A}{A} \right)^2 + \left(\frac{\sigma_B}{B} \right)^2 - 2 \frac{\sigma_{AB}}{AB} \right]$
$f = aA^b$	$\sigma_f^2 \approx \left(ab A^{b-1} \sigma_A \right)^2 = \left(\frac{fb \sigma_A}{A} \right)^2$

source :

https://en.wikipedia.org/wiki/Propagation_of_uncertainty

Plan

- 1 Introduction
- 2 Des statistiques pour quoi faire?
- 3 Statistiques descriptives
- 4 Estimateurs et estimations
- 5 Tests statistiques

Plan

5 Tests statistiques

- Exemples introductifs
- Principe
- Tests sur l'espérance d'un échantillon
- Tests sur la variance d'un échantillon
- Tests de comparaison de deux échantillons
- Tests du Khi-2

Exemple introductif 1

► Espérance de souris

- On suppose que l'espérance de vie de souris de laboratoire suit une variable aléatoire gaussienne dont l'espérance est de $\mu = 2$ ans dans des conditions normales. Son écart-type est une quantité connue égale à $\sigma = 0.5$ an (6 mois).
 - Un spécialiste propose une alimentation qui - selon lui - augmente la durée de vie moyenne de ces souris. Pour s'en assurer un laboratoire soumet 10 souris au régime proposé. À la fin de l'expérience, les durées de vie de ces 10 souris sont les suivantes (en années):
3.1, 2.0, 1.6, 3.2, 2.3, 1.7, 2.7, 3.2, 1.4, 2.2.
- Proposer un **test** au niveau 5% permettant de déterminer si le régime proposé par le spécialiste a un effet significatif ou non.

Exemple introductif 1

Résolution: ce qu'on peut faire...

- ▶ La moyenne observée est égale à $\bar{x} = 2.34$
- ▶ On fait l'**hypothèse** que $\mu = 2$
- ▶ On teste par le calcul $P(\bar{X} > 2.34) = ?$

Notre objectif est de s'apercevoir si, dans les conditions normales, il est absurde/peu probable d'observer les données qu'on a.

$Z \sim \mathcal{N}(0, 1)$, et on obtient $P(Z > \frac{2.34 - 2}{\frac{0.5}{\sqrt{10}}}) = 0.0157638$, que conclure?

Exemple introductif 2

► Une température de crabe

- On note X la température intérieure (en $^{\circ}\text{C}$) d'une espèce de crabes du Pacifique prise à une température ambiante de 24.3°C . On suppose que X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ dont on ne connaît pas les paramètres.
- On a mesuré cette température sur un échantillon de 21 crabes pris au hasard :

$24.6, 26.1, 25.1, 27.3, 24.0, 24.5, \dots$

On donne $\sum_i x_i = 526.9$ et $\sum_i x_i^2 = 13255.53$.

- Proposer un **test** au niveau 5% permettant de déterminer si cette espèce de crabes possède sa propre température intérieure ou si cette dernière est la même que la température ambiante.

Exemple introductif 2

Résolution: ce qu'on peut faire...

- ▶ La moyenne observée est égale à $\bar{x} \approx 25.09$ et l'estimation de la variance empirique corrigée $s \approx 1.33$
- ▶ On fait l'hypothèse que $\mu = 24.3^\circ C$
- ▶ On teste par le calcul $P(\bar{X} > 25.09) = ?$

$T \sim t_{20}$, et on obtient $P(T > \frac{25.09 - 24.3}{\sqrt{\frac{1.33}{21}}}) = 2.7220) = 0.006565$, que conclure?

Différence importante : On ne cherche pas ici à voir qu'un paramètre réel est plus grand que ce qu'on pensait, mais seulement qu'il diffère d'une valeur de référence.

Définitions

À la vue des deux exemples précédents on peut définir:

- Une **hypothèse statistique** est un énoncé ("littéraire") portant sur les caractéristiques d'une population (paramètre ou forme d'une distribution).
- Un **test statistique** est une procédure ("calculatoire") permettant de trancher entre deux hypothèses basées sur l'observation de données.

Plan

5 Tests statistiques

- Exemples introductifs
- **Principe**
- Tests sur l'espérance d'un échantillon
- Tests sur la variance d'un échantillon
- Tests de comparaison de deux échantillons
- Tests du Khi-2

Principe

- (a) Choisir un niveau de risque du test α
- (b) Choisir les hypothèses à tester (H_0 et H_1) selon le problème à résoudre
- (c) Déterminer la variable (aléatoire) de test (**Statistique de test**)
- (d) Déterminer la **région de rejet**
- (e) Calculer la **réalisation** et la **p-valeur** associée
- (f) Conclure

(a) Risques d'erreurs

Résultats possibles

Décision \ Réalité	Ne pas rejeter H_0 (conclure H_0)	Rejeter H_0 (conclure H_1)
H_0 vraie	OK	Erreur de type I
H_1 vraie	Erreur de type II	OK

Définitions

- **Risque de première espèce** : $\alpha = P(\text{rejeter } H_0 | H_0 \text{ vraie})$
(= probabilité de commettre une erreur de type I)
- **Risque de seconde espèce** : $\beta = P(\text{ne pas rejeter } H_0 | H_1 \text{ vraie})$
(= probabilité de commettre une erreur de type II)
- **Puissance** : $\mathbf{P} = 1 - \beta = P(\text{rejeter } H_0 | H_1 \text{ vraie})$
(probabilité de prendre la bonne décision en rejetant H_0)

(a) Focus sur α

Choix du niveau du test

- Le **niveau de signification** du test est le risque de première espèce α consenti.
- Le niveau de signification du test est souvent fixé à 0.05 ou 0.01, mais ce seuil est arbitraire et toute autre valeur peut être choisie.

(b) Les Hypothèses H_0 et H_1

Hypothèse nulle et hypothèse alternative

- L'**hypothèse nulle** (notée H_0) est l'hypothèse privilégiée. C'est celle qui est supposée vraie par défaut.
- L'**hypothèse alternative** (notée H_1) contredit l'hypothèse nulle. C'est l'hypothèse que l'on cherche à montrer.

► **Asymétrie** : On cherche à montrer que H_0 est fausse, c'est-à-dire très peu probable. Cette hypothèse sera conservée dans le cas contraire.

► **Ne pas rejeter, ce n'est pas accepter** : On ne peut jamais montrer la vérité d'une hypothèse avec certitude

(b) Hypothèse simple/composite

Exemple d'hypothèses simples

- $H : \theta = \theta_0$

Exemple d'hypothèses composites (i.e. plusieurs θ dans l'hypothèse)

- $H : \theta < \theta_0$
- $H : \theta > \theta_0$
- $H : \theta \neq \theta_0$
- $H : \theta \in [a, b]$

(b) Tests unilatéraux / bilatéraux

Cas très courants :

Test unilatéral (à droite)

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta > \theta_0$$

Test bilatéral

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$

(c) Statistique de test

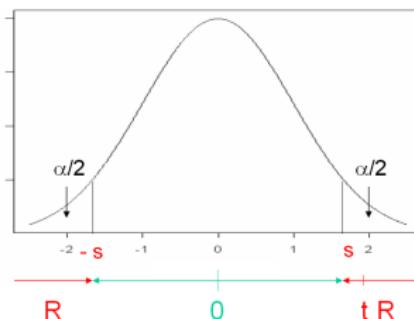
Définition

Une **statistique de test** est une statistique (dont la loi est connue sous H_0) qui permet de mesurer l'écart à l'hypothèse nulle.

(d) Région de rejet

Définition

La **Région de rejet** est l'ensemble R_α (dépend du choix de α) des valeurs (de la statistique de test) pour lesquelles l'hypothèse nulle est rejetée.



Risques d'erreurs

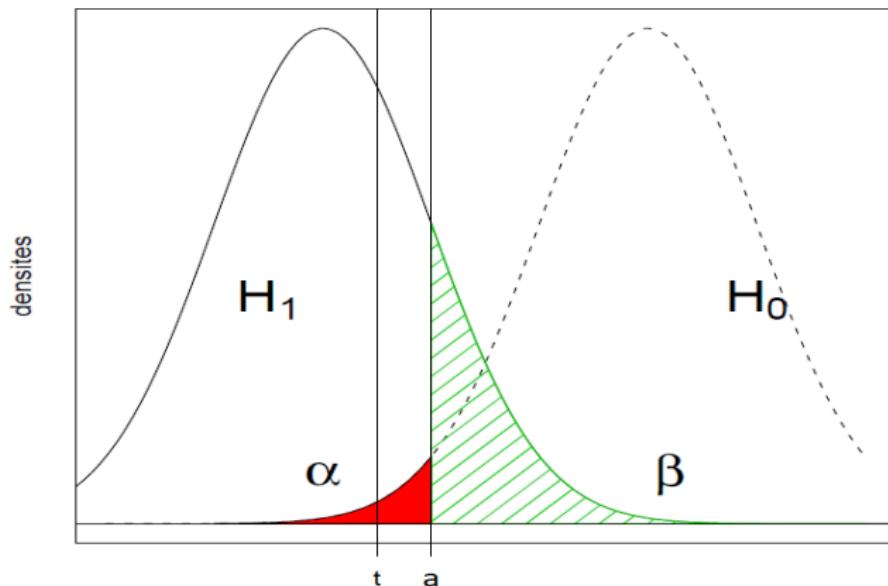


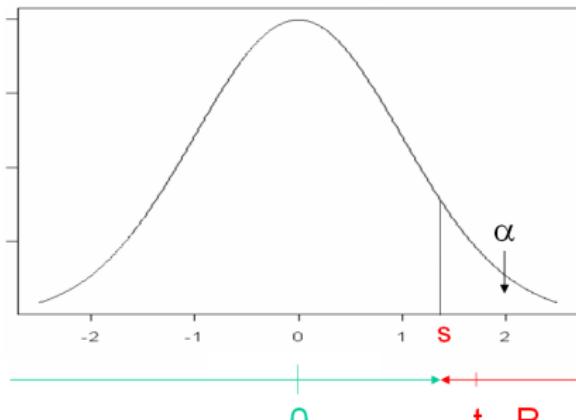
Figure: Erreur de type I et II et région critique de la forme $R_{alpha} =] - \infty, a]$

Tests unilatéraux / bilatéraux

Test unilatéral (à droite)

$$H_0 : \theta = \theta_0$$

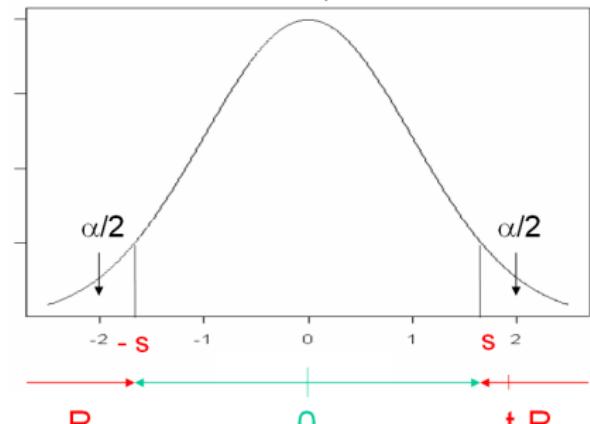
$$H_1 : \theta > \theta_0$$



Test bilatéral

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta \neq \theta_0$$



(e) Degré de signification (p-valeur)

Définition

Le **degré de signification** (ou **p-valeur**) est défini par :

$$p = \min\{\alpha | t \in R_\alpha\}$$

Test unilatéral à droite	Test unilatéral à gauche	Test bilatéral
$p = P(T > t H_0)$	$p = P(T < t H_0)$	$p = P(T > t H_0)$

Remarques

- La p-valeur est la probabilité d'obtenir une valeur de la statistique de test au moins aussi extrême que celle observée lorsque H_0 est vraie
- **Concrètement, on rejette H_0 lorsque $p < \alpha$**

(e) Degré de signification

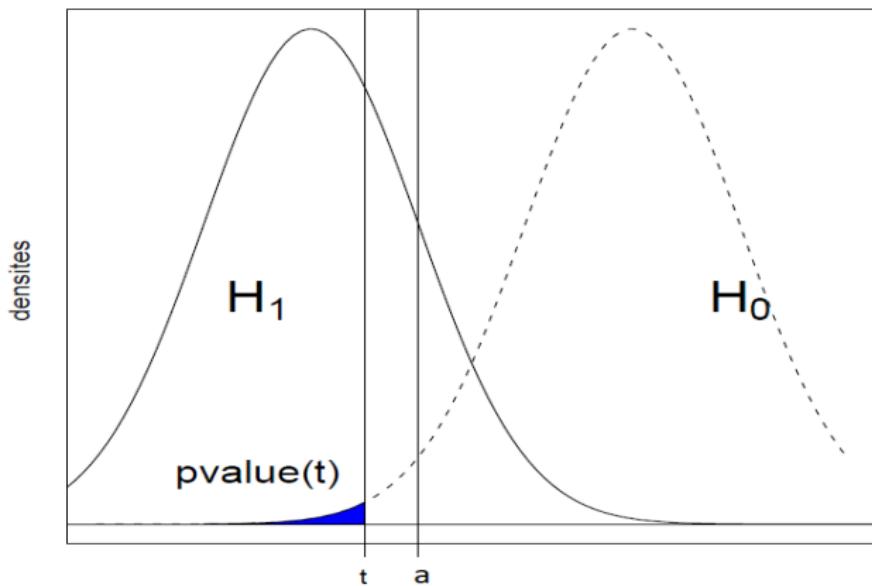


Figure: p-valeur associée à la réalisation t de la statistique de décision pour un test unilatéral à gauche.

Etude de la puissance

Remarque

- Le calcul de la puissance ne peut se faire que si l'on connaît la distribution de la statistique de test sous l'hypothèse alternative (H_1)
- L'étude de la puissance permet de déterminer, pour une alternative donnée, le nombre d'observations nécessaires pour conclure H_1 (avec une certaine puissance)
- L'étude de la puissance permet de déterminer, pour un nombre d'observations données, l'effet minimum pouvant être montré (avec une certaine puissance)

Plan

5

Tests statistiques

- Exemples introductifs
- Principe
- **Tests sur l'espérance d'un échantillon**
- Tests sur la variance d'un échantillon
- Tests de comparaison de deux échantillons
- Tests du Khi-2

Test sur l'espérance d'un échantillon gaussien

Cas 1 : variance connue (test z)

- Modélisation : X_1, \dots, X_n iid avec $X_i \sim \mathcal{N}(\mu, \sigma_0^2)$, σ_0^2 connue.
- Hypothèse nulle : $H_0 : \mu = \mu_0$, $H_1 : ?$ (plusieurs choix selon l'énoncé)
- Statistique de test :

$$\frac{\bar{X} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}} \underset{H_0}{\sim} \mathcal{N}(0, 1)$$

(Ex 1) Test sur l'espérance d'un échantillon gaussien

► Espérance de souris

- On suppose que l'espérance de vie de souris de laboratoire suit une variable aléatoire gaussienne dont l'espérance est de $\mu = 2$ ans dans des conditions normales. Son écart-type est une quantité connue égale à $\sigma = 0.5$ an (6 mois).
 - Un spécialiste propose une alimentation qui - selon lui - augmente la durée de vie moyenne de ces souris. Pour s'en assurer un laboratoire soumet 10 souris au régime proposé. À la fin de l'expérience, les durées de vie de ces 10 souris sont les suivantes (en années):
3.1, 2.0, 1.6, 3.2, 2.3, 1.7, 2.7, 3.2, 1.4, 2.2.
- Proposer un **test** au niveau 5% permettant de déterminer si le régime proposé par le spécialiste a un effet significatif ou non.

Test sur l'espérance d'un échantillon gaussien

Cas 2 : variance inconnue (**test de Student ou test t**)

- Modélisation : X_1, \dots, X_n iid avec $X_i \sim \mathcal{N}(\mu, \sigma^2)$, σ^2 inconnue.
- Hypothèse nulle : $H_0 : \mu = \mu_0$, $H_1 :$? (plusieurs choix selon l'énoncé)
- Statistique de test :

$$\frac{\bar{X} - \mu_0}{\frac{S^*}{\sqrt{n}}} \underset{H_0}{\sim} t_{n-1}$$

(Ex 2) Test sur l'espérance d'un échantillon gaussien

► Une température de crabe

- On note X la température intérieure (en $^{\circ}\text{C}$) d'une espèce de crabes du Pacifique prise à une température ambiante de 24.3°C . On suppose que X suit une loi normale $\mathcal{N}(\mu, \sigma^2)$ dont on ne connaît pas les paramètres.
- On a mesuré cette température sur un échantillon de 21 crabes pris au hasard : 24.6, 26.1, 25.1, 27.3, 24.0, 24.5, ...
On donne $\sum_i x_i = 526.9$ et $\sum_i x_i^2 = 13255.53$.

► Proposer un **test** au niveau 5% permettant de déterminer si cette espèce de crabes possède sa propre température intérieure ou si cette dernière est la même que la température ambiante.

► On suppose qu'en réalité la température moyenne des crabes est de 25°C . Quelle est la puissance du test construit pour détecter une telle différence ?

Test sur l'espérance d'un échantillon de loi quelconque - n grand

- Modélisation : X_1, \dots, X_n iid avec $X_i \sim \mathcal{L}$ inconnue.
- Hypothèse nulle : $H_0 : \mu = \mu_0$, $H_1 :$? (plusieurs choix selon l'énoncé)
- Statistique de test :
 - Si σ_0^2 connue

$$\frac{\bar{X} - \mu_0}{\frac{\sigma_0}{\sqrt{n}}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0, 1) \text{ (TLC)}$$

- Si σ_0^2 inconnue

$$\frac{\bar{X} - \mu_0}{\frac{s^*}{\sqrt{n}}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0, 1) \text{ (TLC)}$$

(Ex 3) Test sur l'espérance d'un échantillon de loi quelconque

- Le délai de survie, pour un certain type de cancer, peut être modélisé par une variable aléatoire de loi exponentielle. L'espérance de vie avec le traitement de référence est de 4 ans.
 - Un nouveau traitement est testé dans le cadre d'un essai clinique sur $n = 60$ patients. On observe un délai de survie moyen de 4.7 ans.
- Peut-on conclure que le nouveau traitement est significativement meilleur que le traitement de référence ?

Test sur un pourcentage - n grand

- Modélisation : X_1, \dots, X_n iid avec $X_i \sim \mathcal{B}(p)$, p inconnue.
- Hypothèse nulle : $H_0 : p = p_0$, , $H_1 : ?$ (plusieurs choix selon l'énoncé)
- Statistique de test ($\hat{p} = \bar{X}$) :

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0, 1) \text{ (TLC)}$$

Remarque

Tester un pourcentage revient à tester l'espérance d'une Bernoulli

(Ex 4) Test sur un pourcentage - n grand

- Le pourcentage d'anomalies chromosomiques dans les naissances d'une population donnée, était de 1% il y a 10 ans.
 - On effectue un dépistage systématique (obtention des caryotypes à partir de prélèvements de sang) sur 500 naissances tirées au sort dans la population actuelle. On observe 7 caryotypes anormaux.
- Le pourcentage d'anomalies chromosomique est-il significativement différent d'il y a 10 ans.
- On suppose que le pourcentage d'anomalies est en réalité passé de 1 à 1,2%. Sur l'observation des 500 naissances, quelle probabilité a-t-on de détecter cette différence ?

Test sur un pourcentage - n petit

- Modélisation : X_1, \dots, X_n iid avec $X_i \sim \mathcal{B}(p)$, p inconnue.
- Hypothèse nulle : $H_0 : p = p_0$
- Statistique de test :

$$n\hat{p} = \sum_{H_0} X_i \sim \mathcal{B}(n, p_0)$$

Remarque

Avec une loi discrète on ne peut plus trouver tous les quantiles exactement.

(Ex 5) Test sur un pourcentage - n petit

On croise des descendants directs du croisement [fleurs rouges \times fleurs blanches]. Sous l'hypothèse que le gène 'rouge' est dominant, la probabilité p d'obtenir une plante à fleurs blanches est de $1/4$ alors que sous l'hypothèse que le gène 'blanc' est dominant, la probabilité p d'obtenir une plante à fleurs blanche est de $3/4$.

Sur $n = 23$ croisements (supposés indépendants), on a observé 8 plantes à fleurs blanches.

- ① L'hypothèse admise jusqu'à présent est que le gène 'rouge' est dominant. Un généticien aimerait montrer qu'en réalité, c'est le gène 'blanc' qui est dominant. Tester cette hypothèse au niveau $\alpha = 5\%$.
- ② Quelle est la puissance du test construit ?

Plan

5

Tests statistiques

- Exemples introductifs
- Principe
- Tests sur l'espérance d'un échantillon
- **Tests sur la variance d'un échantillon**
- Tests de comparaison de deux échantillons
- Tests du Khi-2

Tests sur la variance d'un échantillon gaussien

- Hypothèses :

X_1, \dots, X_n iid avec $X_i \sim \mathcal{N}(\mu, \sigma^2)$

- Hypothèse nulle : $H_0 : \sigma^2 = \sigma_0^2$
- Statistique de test :

$$\frac{(n-1)S^{*2}}{\sigma_0^2} \underset{H_0}{\sim} \chi_{n-1}^2$$

Si $H_1 : \sigma^2 > \sigma_0^2$, alors on cherche u tel que $P\left(\frac{(n-1)S^{*2}}{\sigma_0^2} > u\right) = \alpha$

Si $H_1 : \sigma^2 < \sigma_0^2$, alors on cherche u tel que $P\left(\frac{(n-1)S^{*2}}{\sigma_0^2} < u\right) = \alpha$

Si $H_1 : \sigma^2 \neq \sigma_0^2$, alors on cherche u_1 et u_2 tels que

$$P\left(\frac{(n-1)S^{*2}}{\sigma_0^2} < u_1\right) = P\left(\frac{(n-1)S^{*2}}{\sigma_0^2} > u_2\right) = \frac{\alpha}{2}$$

(Ex 6) Test sur la variance

Une nouvelle technique de dosage du glucose sanguin vient d'être mise au point. Pour un même échantillon de sang, sept dosages effectués à l'aide de cette nouvelle technique ont donné les résultats suivants (en g/l) :

1.17 1.16 1.16 1.19 1.21 1.19 1.18

On admet que les sept mesures sont des variables aléatoires indépendantes et identiquement distribuées de loi $\mathcal{N}(\mu, \sigma^2)$ où σ^2 caractérise la précision du procédé. La technique utilisée jusqu'alors était caractérisée par un écart-type de 0.05 mg/l. Peut-on dire que la nouvelle technique est plus précise que l'ancienne ?

Plan

5

Tests statistiques

- Exemples introductifs
- Principe
- Tests sur l'espérance d'un échantillon
- Tests sur la variance d'un échantillon
- **Tests de comparaison de deux échantillons**
- Tests du Khi-2

Test sur l'espérance de deux échantillons indépendants

Indépendance des échantillons

Deux échantillons sont indépendants s'ils sont constitués indépendamment l'un de l'autre

Remarque

- Les sujets de l'échantillon 1 ne sont pas les mêmes que les sujets de l'échantillon 2
- Les effectifs des échantillons 1 et 2 ne sont pas nécessairement les mêmes

Comparaison des espérances de deux échantillons gaussiens

Test de Student

- Modélisation :

X_1, \dots, X_{n_1} iid avec $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$

Y_1, \dots, Y_{n_2} iid avec $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$

$\sigma_1^2 = \sigma_2^2 = \sigma^2$ inconnue.

- Hypothèse nulle : $H_0 : \mu_1 = \mu_2$

- Statistique de test :

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)S_1^{*2} + (n_2-1)S_2^{*2}}{n_1+n_2-2}}} \underset{H_0}{\sim} t_{n_1+n_2-2}$$

(Ex 7) Comparaison des espérances de deux échantillons gaussiens

On a prélevé une solution plusieurs fois en utilisant deux pipettes calibrées de même volume. On a pesé le contenu du volume délivré par la pipette. Les résultats des différents pipettages, qui sont supposés normalement distribués, sont exprimés en grammes.

Pipette 1	0.0987	0.0990	0.0996	0.0995	0.0998	0.0984
Pipette 2	0.1016	0.1008	0.1002	0.0995	0.0990	0.1023

On suppose que les variances sont les mêmes dans les deux groupes.

- Les quantités moyennes prélevées par chacune des deux pipettes sont-elles identiques ? (comparer les espérances)

Comparaison des espérances de deux échantillons - n_1 et n_2 grands

- Modélisation :
 X_1, \dots, X_{n_1} iid avec $X_i \sim \mathcal{L}^1$
 Y_1, \dots, Y_{n_2} iid avec $Y_i \sim \mathcal{L}^2$
- Hypothèse nulle : $H_0 : \mu_1 = \mu_2$
- Statistique de test :

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_1^{*2}}{n_1} + \frac{s_2^{*2}}{n_2}}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0, 1) \text{ (TLC)}$$

(Ex 8) Comparaison des espérances de deux échantillons - n_1 et n_2 grands

Dans le but d'étudier l'influence éventuelle de la lumière sur la croissance du poisson Lebistes Reticulus, on a élevé deux lots de ce poisson dans des conditions d'éclairage différentes. Au 95^e jour, on a mesuré (en mm) les longueurs x_i des poissons. On a obtenu les résultats suivants :

Lot 1 (180 individus) : éclairage à 400 lux

$$\sum x_i = 3780 \quad \sum x_i^2 = 84884$$

Lot 2 (90 individus) : éclairage à 3 000 lux.

$$\sum y_i = 2043 \quad \sum y_i^2 = 46586$$

Que peut-on conclure ?

Comparaison de deux pourcentages - n_1 et n_2 grands

- Modélisation :
 - X_1, \dots, X_{n_1} iid avec $X_i \sim \mathcal{B}(p_1)$
 - Y_1, \dots, Y_{n_2} iid avec $Y_i \sim \mathcal{B}(p_2)$
- Hypothèse nulle : $H_0 : p_1 = p_2 = p$
- Statistique de test :

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\hat{p}(1-\hat{p})}} \xrightarrow[H_0]{\mathcal{L}} \mathcal{N}(0, 1) \text{ (TLC)}$$

où

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Remarque

Ce test est équivalent au test du khi-2 (voir plus loin)

(Ex 9) Comparaison de deux pourcentages - n_1 et n_2 grands

Dans un groupe de 200 malades, on a constitué par tirage au sort une série soumise à un nouveau traitement A et une série soumise au traitement classique B. On a :

Traitement A : $n_A = 102$; 20 échecs soit $p_A = 19.6\%$

Traitement B : $n_B = 98$; 29 échecs soit $p_B = 29.6\%$

Au niveau $\alpha = 5\%$, les traitements A et B ont-ils un taux d'échecs significativement différent ?

Comparaison des espérances de deux échantillons appariés

Echantillons appariés

Deux échantillons sont appariés s'il existe une correspondance entre les observations du premier échantillon et les observations du second.

Exemple

Mesure avant traitement et après traitement (chez les mêmes sujets)

Comparaison des espérances de deux échantillons appariés

- Modélisation :

X_1, \dots, X_n iid avec $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$

Y_1, \dots, Y_n iid avec $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$

Soit $D_i = X_i - Y_i$

- Hypothèse nulle : $H_0 : \mu_d = 0$

- Statistique de test

- n petit

$$\frac{\bar{D} - \mu_d}{\frac{S_d^*}{\sqrt{n}}} \underset{H_0}{\sim} t_{n-1}$$

- n grand, loi quelconque

$$\frac{\bar{D} - \mu_d}{\frac{S_d^*}{\sqrt{n}}} \underset{H_0}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, 1)$$

(Ex 10) Comparaison des espérances de deux échantillons appariés

On veut comparer chez 10 malades la pression artérielle systolique moyenne après administration d'un nouveau médicament hypotenseur et après administration du traitement de référence. Le tableau suivant donne les résultats :

Malade	1	2	3	4	5	6	7	8	9	10
Référence	17	15	15	13	12	17	15	16	19	11
Nouveau traitement	16	11	12	13	14	11	13	13	17	10

On suppose les observations normalement distribuées. Le nouveau médicament est-il efficace ?

Comparaison des variances de deux échantillons gaussiens

- Modélisation :

X_1, \dots, X_{n_1} iid avec $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$

Y_1, \dots, Y_{n_2} iid avec $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$

- Hypothèse nulle : $H_0 : \sigma_1^2 = \sigma_2^2$

- Statistique de test :

$$\frac{S_1^{*2}}{S_2^{*2}} \underset{H_0}{\sim} \mathcal{F}(n_1 - 1, n_2 - 1)$$

Loi de Fisher

Définition

Soient Y_1 et Y_2 deux variables aléatoires indépendantes telles que $Y_1 \sim \chi^2_{n_1}$ et $Y_2 \sim \chi^2_{n_2}$. La variable aléatoire $Z = \overline{(Y_1/n_1)/(Y_2/n_2)}$ suit une loi continue appelée loi de Fisher à n_1 et n_2 degrés de liberté :

$$Z = \frac{Y_1/n_1}{Y_2/n_2} \sim \mathcal{F}(n_1, n_2)$$

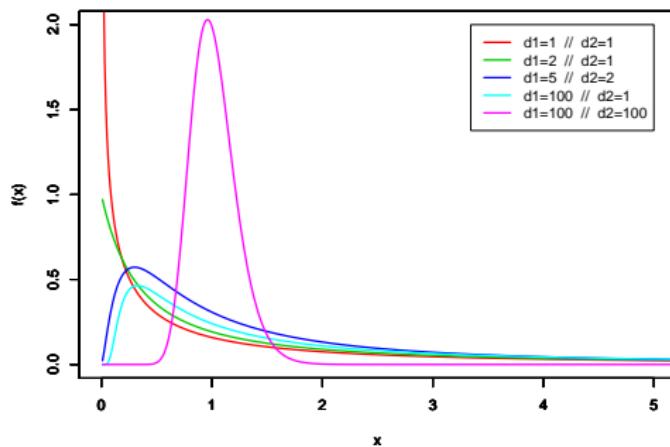
Remarques

- $Z_1 \sim \mathcal{F}(n_2, n_1) \Rightarrow Z_2 = 1/Z_1 \sim \mathcal{F}(n_1, n_2)$
- $T \sim t_n \Rightarrow Z = T^2 \sim \mathcal{F}(1, n)$

Loi de Fisher

Densité

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} n_1^{n_1/2} n_2^{n_2/2} \frac{x^{n_1/2-1}}{(n_2 + n_1 x)}$$



(Ex 11) Comparaison des variances de deux échantillons gaussiens

On a prélevé une solution plusieurs fois en utilisant deux pipettes calibrées de même volume. On a pesé le contenu du volume délivré par la pipette. Les résultats des différents pipettages, qui sont supposés normalement distribués, sont exprimés en grammes.

Pipette 1	0.0987	0.0990	0.0996	0.0995	0.0998	0.0984
Pipette 2	0.1016	0.1008	0.1002	0.0995	0.0990	0.1023

On suppose que les variances sont les mêmes dans les deux groupes.

- Les deux pipettes ont-elles la même précision de mesure ? (comparer les variances)

Plan

5

Tests statistiques

- Exemples introductifs
- Principe
- Tests sur l'espérance d'un échantillon
- Tests sur la variance d'un échantillon
- Tests de comparaison de deux échantillons
- Tests du Khi-2

Loi du χ^2 (Rappel)

Définition

Soient X_1, \dots, X_n n variables aléatoires **indépendantes** et identiquement distribuées de loi normale centrée réduite: $X_i \sim \mathcal{N}(0, 1)$

La variable aléatoire $Y = X_1^2 + \dots + X_n^2$ suit une loi continue appelée loi du χ^2 à n degrés de liberté :

$$Y = \sum_{i=1}^n X_i^2 \sim \chi_n^2$$

Propriétés

- Si $Y_1 \sim \chi_{n_1}^2$ et $Y_2 \sim \chi_{n_2}^2$ avec $Y_1 \perp\!\!\!\perp Y_2$, alors $Y = Y_1 + Y_2 \sim \chi_{n_1+n_2}^2$
- Si $Y \sim \chi_n^2$, alors $\mathbb{E}[Y] = n$ et $V(Y) = 2n$

(Ex 12) Test d'adéquation

Le tableau ci-dessous donne les résultats d'une des expériences de Mendel portant sur des pois : A est le phénotype 'graines sphériques', a le phénotype 'graines ridées' ; B le phénotype 'albumen jaune' et b est le phénotype 'albumen vert'.

AB	Ab	aB	ab
315	103	101	32

On se demande si la distribution observée est compatible avec la distribution théorique $9/16$, $3/16$, $3/16$, $1/16$.

Test d'adéquation

- Modélisation: X_1, \dots, X_n sont des variables aléatoires discrètes iid à valeurs dans les classes $E = \{x_1, x_2, \dots, x_k\}$.
- Hypothèses testées :

$$\begin{cases} H_0 : X \sim \mathcal{L} \\ H_1 : X \text{ ne suit pas la loi } \mathcal{L} \end{cases}$$

- Statistique de test :

$$D_{k,n} = \sum_{i=1}^k \frac{(N_i - np_i)^2}{np_i} \xrightarrow[H_0]{\mathcal{L}} \chi_{k-1}^2$$

avec $P(X = x_i) = p_i$ et N_i le nombre d'observations dans la classe x_i .

Test d'adéquation

Remarques

L'approximation de la distribution de $D_{k,n}$ par χ^2_{k-1} n'est valable que si :

- n est grand
- $np_i \geq 5$ pour tout i (si $np_i < 5$, on regroupe des classes de valeur pour se ramener à un cas où $np_i \geq 5$)

Test d'adéquation à une famille de lois

Degrés de liberté

Lorsque la loi théorique (c'est-à-dire les p_i) dépend d'un ou plusieurs paramètres inconnus (par exemple $\mathcal{N}(\mu, \sigma^2)$, $\mathcal{P}(\lambda)$), il est possible d'estimer ces paramètres à partir des mêmes données que celles utilisées pour le test d'adéquation. Dans ce cas, le nombre de degrés de liberté de $D_{k,n}$ est diminué d'autant que de paramètres estimés.

(Ex 13) Test d'adéquation à une famille de lois

On souhaite étudier la contamination du lait par des spores de clostridia. Pour cela on analyse des tubes de 1 ml de lait et, pour chaque tube, on compte le nombre X de spores présents. L'analyse est effectuée sur un échantillon de $n = 100$ tubes provenant du même lait.

Nombre de spores	0	1	2	3
Nombre de tubes	64	25	9	2

- ① Donner une estimation du nombre moyen de spores par ml de lait.
- ② Peut-on considérer au vu des observations que le nombre de spores contenu dans un ml de lait suit une loi de Poisson (répondre à l'aide d'un test de niveau 5%) ?

(Ex 14) Test d'indépendance

Le tableau ci-dessous indique le résultat de l'examen de 6800 sujets classés d'après la couleur de leurs yeux et celle de leurs cheveux :

<i>Yeux \ Cheveux</i>	Blonds	Bruns	Noirs	Roux	Total
Bleus	1768	807	189	47	2811
Gris ou verts	946	1387	746	53	3132
Bruns	115	438	288	16	857
Total	2829	2632	1223	116	6800

Existe-t-il une liaison entre ces deux caractères ? De manière équivalente, la répartition de la couleur des yeux est-elle la même quelle que soit la couleur des cheveux ou, réciproquement, la répartition de la couleur des cheveux est-elle la même quelle que soit la couleur des yeux ?

Test d'indépendance

- X et Y sont deux variables aléatoires discrètes à valeurs dans $E = \{x_1, \dots, x_k\}$ et $F = \{y_1, \dots, y_\ell\}$
- Hypothèses testées :

$$\begin{cases} H_0 : X \text{ et } Y \text{ sont indépendantes,} \\ H_1 : X \text{ et } Y \text{ ne sont pas indépendantes.} \end{cases}$$

- Statistique de test :

$$D_{k,\ell,n} = \sum_{i=1}^k \sum_{j=1}^\ell \frac{\left(N_{ij} - \frac{N_{i+} N_{+j}}{n}\right)^2}{\frac{N_{i+} N_{+j}}{n}} \xrightarrow[H_0]{\mathcal{L}} \chi^2_{(k-1)(\ell-1)}$$

Test d'indépendance

Remarques

L'approximation de la distribution de $D_{k,\ell,n}$ par un $\chi^2_{(k-1)(\ell-1)}$ n'est valable que si :

- n est grand
- $\frac{N_{i+}N_{+j}}{n} \geq 5$ pour tout couple (i,j)

(Ex 15) Test d'homogénéité

On veut comparer les réactions produites par deux vaccins BCG désignés par A et B. Un groupe de 348 enfants a été divisé par tirage au sort en 2 séries qui ont été vaccinées, l'une par A, l'autre par B. Les résultats figurent dans le tableau suivant :

Vaccin	Réaction légère	Réaction moyenne	Ulcération	Abcès	Total
A	12	156	8	1	177
B	29	135	6	1	171
Total	41	291	14	2	348

Existe-t-il une différence entre les deux vaccins ou, de manière équivalente, la répartition des réactions est-elle la même pour les deux vaccins ?

Test d'homogénéité

Indépendance : deux variables mesurées conjointement sont-elles indépendantes?

Homogénéité : Les différentes séries mesurées proviennent-elles de la même distribution?

Proposition

Un test d'homogénéité est identique à un test d'indépendance

Remerciements

☞ Un grand merci à Cyril Dalmasso

<http://www.math-evry.cnrs.fr/members/cdalmasso/welcome>

- Pour les slides (en particulier la section "Tests statistiques")
- Les exercices de TD