MSV31 – Statistiques
▶ Statistiques **descriptives**
▶ **Estimateurs**
▶ **tests** statistiques

Vincent Runge
vincent.runge@univ-evry.fr
IBGBI – 4$^{\text{ème}}$ étage

Cours L2 Bio. 2019 – 2020

Plan du cours

Plan

Plan

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

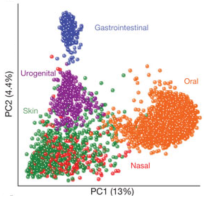
Qu'est-ce que la statistique?

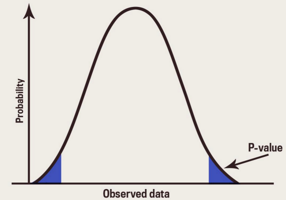
Statistique : ensemble des méthodes de **réduction** des données

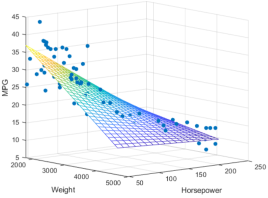
Définition V. Runge

Ses objectifs:

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





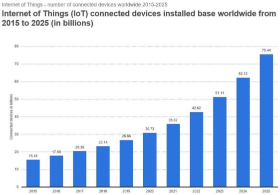


Qu'est-ce que la statistique?

> 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



Plan

Courte histoire des mathématiques

- Les statistiques sont une branche **récente** des mathématiques

- **Grèce antique** (VI$^{\text{ème}}$ à IV$^{\text{ème}}$ siècles avant JC)
  - Arithmétique et géométrie
  - Naissance de la démonstration

Thalès, Euclide, Pythagore, Archimède

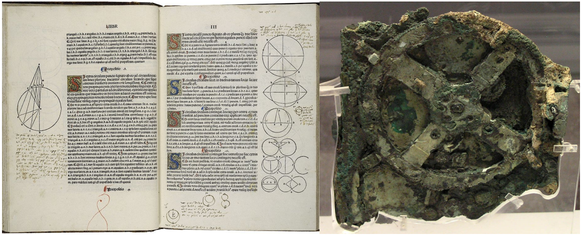


Figure: Les Éléments d'Euclide et la machine d'Anticythère

Courte histoire des mathématiques

- **Mathématique arabes** ($\sim$ 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$^{\text{ème}}$ XVIII$^{\text{ème}}$ 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$^{\text{ème}}$ 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)
  - ...

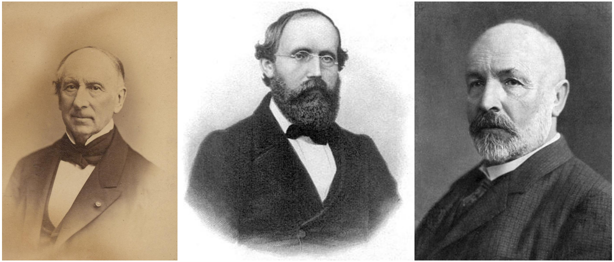


Figure: Augustin Louis Cauchy (1789 – 1857); Bernhard Riemann (1826 – 1866); Georg Cantor (1845 – 1918)

Courte histoire des mathématiques

- **XX$^{\text{ème}}$ 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$^{\text{ème}}$ 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

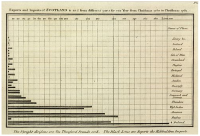
1. Introduction
   - Qu'est-ce que la statistique?
   - Courte histoire des mathématiques
   - Statistiques dans l'histoire des mathématiques

Statistiques dans l'histoire des mathématiques

Et les statistiques ? ▶ 3 grandes périodes historiques

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

William Playfair (1759–1823)
Commercial and Political Atlas
(1786). Le 1er histogramme connu



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

Statistiques descriptives → **Analyse des données** (XX$^{\text{ème}}$ 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**

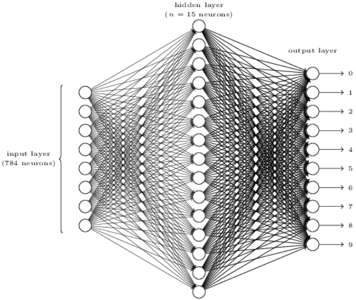
Influence de l'informatique

- ☐ 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 ≈ Intelligence artificielle

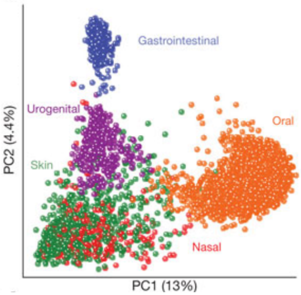


Statistiques dans l'histoire des mathématiques

- **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

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

- Dangerosité 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

Statistiques et probabilités

La statisique 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

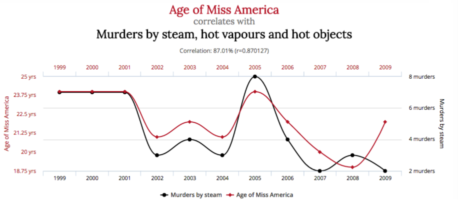
Objectif : généraliser "au mieux" du "petit nombre au grand nombre"

- Tests statistiques

Objectif : Prendre une décision raisonnée

Références internet

1. Corrélations bizarres :
   `http://tylervigen.com/spurious-correlations`
2. Chaîne Youtube : la statistique expliquée à mon chat
3. MOOC FUN. Introduction à la statistique avec R



FRANCE UNIVERSITÉ NUMÉRIQUE
Se former en liberté

Plan

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, ordinales 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

Les différentes moyennes

Moyenne arithmétique

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

☞ 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

$$\overline{x}_G = \sqrt[n]{x_1 \times \ldots \times x_n}$$

Moyenne harmonique

$$\overline{x}_H = \frac{n}{\frac{1}{x_1} + \cdots + \frac{1}{x_n}}$$

Moyenne quadratique

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

- **Théorème :** Si $x_1, ..., x_n > 0$, alors

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

Les différentes moyennes

Théorème : Si $x_1, ..., x_n > 0$, alors

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

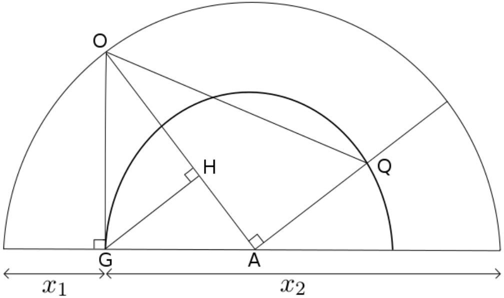


Figure: Moyennes de x_1 et x_2: harmonique OH, géométrique OG, arithmétique OA et quadratique OQ

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

Plan

Variance et écart-type

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

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

Théorème de König-Huygens

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

avec

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

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

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

Propriété de la variance

Données de taille n_1, moyenne $\overline{x_1}$ et variance σ_1^2
Données de taille n_2, moyenne $\overline{x_2}$ et variance σ_2^2
Données de taille $n_1 + n_2$, moyenne $\overline{x}$ variance σ^2

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1 + n_2} + \frac{n_1(\overline{x_1} - \overline{x})^2 + n_2(\overline{x_2} - \overline{x})^2}{n_1 + n_2}$$

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

$$\overline{x} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2}$$

On écrit aussi

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

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

$$(n_1+n_2)\sigma^2 = \sum_{k=1}^{n_1}(x_k-\overline{x})^2 + \sum_{k=n_1+1}^{n_1+n_2}(x_k-\overline{x})^2$$

$$= \sum_{k=1}^{n_1}((x_k-\overline{x_1})+(\overline{x_1}-\overline{x}))^2 + \sum_{k=n_1+1}^{n_1+n_2}((x_k-\overline{x_2})+(\overline{x_2}-\overline{x}))^2$$

$$= n_1\sigma_1^2 + 2\sum_{k=1}^{n_1}(x_k-\overline{x_1})(\overline{x_1}-\overline{x}) + n_1(\overline{x_1}-\overline{x})^2$$

$$+n_2\sigma_2^2 + 2\sum_{k=n_1+1}^{n_1+n_2}(x_k-\overline{x_2})(\overline{x_2}-\overline{x}) + n_2(\overline{x_2}-\overline{x})^2\,,$$

en utilisant la définition des moyennes $\overline{x_1}$ et $\overline{x_2}$ chacune des deux sommes restantes s'annule et le résultat est démontré.

Analyse de la variance ANOVA (Hors programme)

Interprétation de la formule

σ^2 = variance **intrapopulation** + variance **interpopulation**

$$\sigma^2 = \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{n_1+n_2} + \frac{n_1(\overline{x_1}-\overline{x})^2 + n_2(\overline{x_2}-\overline{x})^2}{n_1+n_2}.$$

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

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

- Étudier $\dfrac{SCE_{\text{facteur}}}{SCE_{\text{residu}}}$ pour savoir si $\overline{x_1}$ et $\overline{x_2}$ sont "les mêmes" !
- C'est une possibilité de test statistique...

Écart moyen

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

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

Données de taille n_1, moyenne $\overline{x_1}$ et écart moyen e_1
Données de taille n_2, moyenne $\overline{x_2}$ et écart moyen e_2
Donnée $n_1 + n_2$, moyenne $\overline{x}$ écart moyen $=$?

Les moments

Moments centrés d'ordre r

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

- 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 :

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

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

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

L'écart moyen ne possède pas de telle propriété

Remarque de vocabulaire

les moyennes

On parle de

- **Moyenne observée** pour $\overline{x}$, $\overline{x}_G$, $\overline{x}_H$...
- **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 $x_{(1)}, x_{(2)}, ..., x_{(n)}$, avec des indices entres 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.

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

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** $(D_1, \ldots, D_9) = (q_{\frac{1}{10}}, \ldots, q_{\frac{9}{10}})$
- 100 parts égales, on a 99 **centiles** $(C_1, \ldots, C_{99}) = (q_{\frac{1}{100}}, \ldots, q_{\frac{99}{100}})$

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

$$(1) \quad : \quad 1 \quad 2 \quad 3 \quad 11 \quad 12 \quad 13 \quad 21 \quad 22 \quad 23 \quad 31 \quad 32 \quad 33 \quad \mathbf{34}$$

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

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 $[Q_1 - 1.5\,IQ, Q_1]$ et $x_{(b)}$ réalisant le maximum des $x_{(i)}$ dans $[Q_3, Q_3 + 1.5\,IQ]$
- 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)

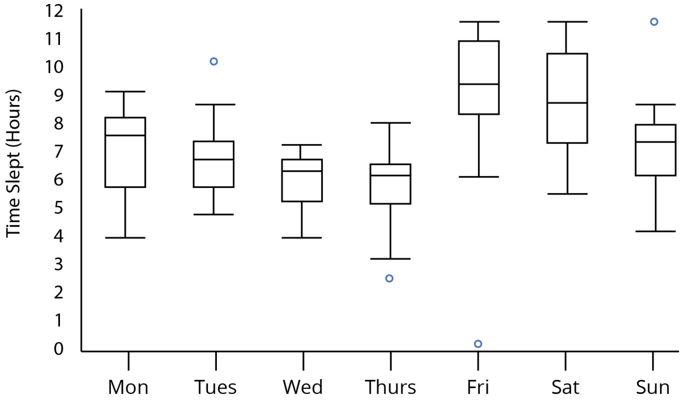


Figure: Nombre d'heures de sommeil pour 20 lycéens pendant une semaine.

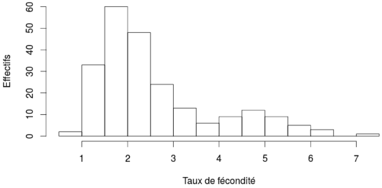
Histogrammes

- 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

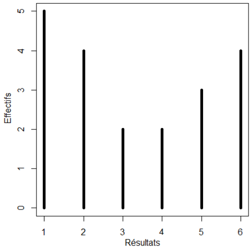
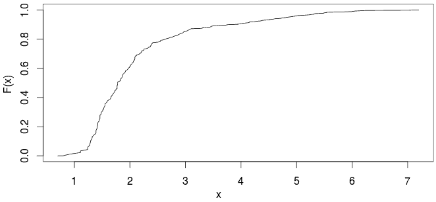


Figure: Résultat de 20 lancés d'un dé équilibré

Fonction de répartition empirique

- = Graphique des fréquences cumulées
- Pour n données,

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



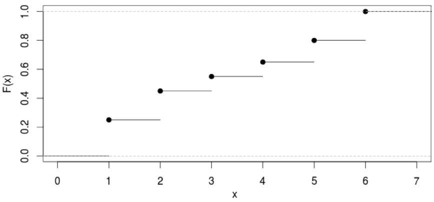


Figure: (gauche) fonction de répartition des taux de fécondité. (droite) fonction de répartition des 20 lancés de dé

Tableaux

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

Yeux \ Cheveux	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 $x_1, ..., x_n$
L'analyse de donnée est **multidimensionnelle**
☞ On organise les données $(x_{11}, ...x_{1p}), ..., (x_{n1}, ...x_{np})$
De nouvelles questions émergent:

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

Plan

Plan

4 Estimateurs et estimations
- Générer l'aléatoire
- 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)

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

Set the seed x_0 *using the current system time in microseconds*
$x_0 <-$ *as.numeric(Sys.time()) * 1000*

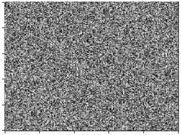


Figure: Le générateur d'UNIX et GCC. $a = 1103515245$, $c = 12345$, $m = 2^{31}$
Ici : seed $= x_0 = 99$, $2^{16} = 65536$ valeurs. $x_i/(2^{31}) \in [0, 1[$

- 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

L'échantillonnage

Définitions

Échantillon aléatoire : variables aléatoires $(X_1, ..., X_n)$ avec $X_i = X$
Échantillon : réalisations $x_1, ..., x_n$ des variables $X_1, ..., X_n$

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

Estimateur : variable aléatoire $T = f(X_1, ..., X_n)$
Estimation : réalisation t de T : $t = f(x_1, ..., x_n)$

▶ Hypothèse : Les variables aléatoires $X_1, ..., X_n$ 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 $x_1, ..., x_n$ une observation de l'échantillon $X_1, ..., X_n$, copies de X
- Comment estimer θ à partir de $x_1, ..., x_n$?

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 :

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

- Variance empirique : $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$
- Variance empirique corrigée : $S^{\star 2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n}{n-1} S^2$

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.

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

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

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

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 :

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

Remarque

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

Exercice

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

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

Estimateur de la moyenne

Moyenne empirique

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

Propriétés

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

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

Remarque

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

Estimateur de la variance

Variance empirique et variance empirique corrigée

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \quad \text{et} \quad S^{\star 2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Propriétés

Soit $(X_1, ..., X_n)$ 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

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

Maximum de vraisemblance (Hors programme)

▶ Existe-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:

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

Exemple distribution normale:

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

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

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 $\hat{\theta} = \frac{1}{\overline{X}}$. Mais est-ce un bon choix?

Maximum de vraisemblance (Hors programme)

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

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

Et on a :

$$\frac{d}{d\theta}(\log(L)) = \frac{d}{d\theta}\left(n \log(\theta) - \theta n \overline{x}\right) = \frac{n}{\theta} - n\overline{x}$$

Ainsi

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

Un meilleur estimateur qui minimise l'erreur quadratique moyenne sera

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

Plan

Estimation par intervalle

Principe

- Trouver une variable aléatoire $B(\theta)$ de loi connu 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 $B_1 = f_1(X_1, ..., X_n)$ et $B_2 = f_2(X_1, ..., X_n)$.
$[B_1, B_2]$ 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'ecart-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 $(X_1, ..., X_n)$ un échantillon de loi $\mathcal{N}(\mu, \sigma^2)$. on suppose σ^2 connu.
L'intervalle :

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

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

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

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 :

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

.

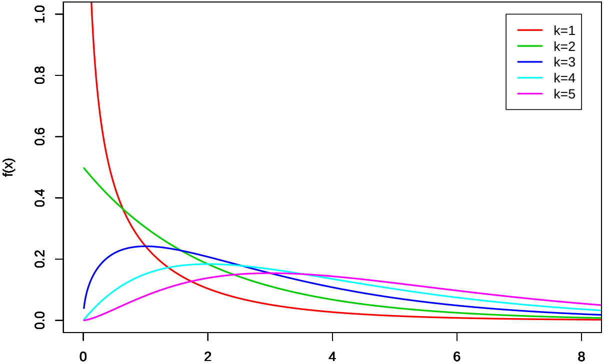
Propriétés

- Si $Y_1 \sim \chi^2_{n_1}$ et $Y_2 \sim \chi^2_{n_2}$ avec $Y_1 \perp\!\!\!\perp Y_2$, alors $Y = Y_1 + Y_2 \sim \chi^2_{n_1+n_2}$
- Si $Y \sim \chi^2_n$, 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 $Y \sim \chi^2_n$. 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

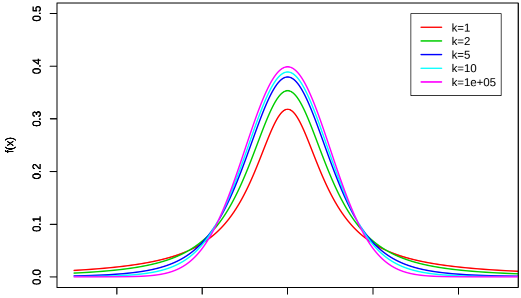
- $\mathbb{E}[T] = 0$
- $V(T) = \frac{n}{n-2}$ si $n > 2$

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

On sait que

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

Théorème de Cochran

Théorème

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

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

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

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

Estimation par intervalle de la moyenne

Famille gaussienne, variance inconnue

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

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

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

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

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

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

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

Plan

Le théorème central limite

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

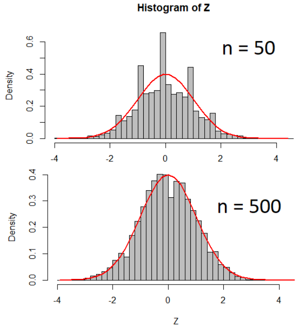
$$S = \sum_{i=1}^{n} X_i \xrightarrow{\mathcal{L}} \mathcal{N}\left(n\mu, n\sigma^2\right)$$

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

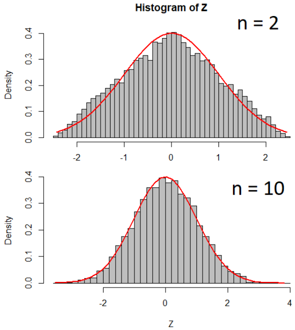
Le théorème central limite

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

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



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



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 $\overline{x} = 0.4$.

- 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 $\overline{x}$ est :

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

Plan

Propagation des incertitudes

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

$$f = f(X, Y, Z, ...)$$

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

Formule de la variance

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

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

Propagation des incertitudes

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

source :

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