

MSV31 – Statistiques

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

Vincent Runge  
[vincent.runge@univ-evry.fr](mailto:vincent.runge@univ-evry.fr)  
IBGBI – 4<sup>ème</sup> 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

# Plan

1 Introduction

2 Des statistiques pour quoi faire?

3 Statistiques descriptives

# Plan

## 1 Introduction

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

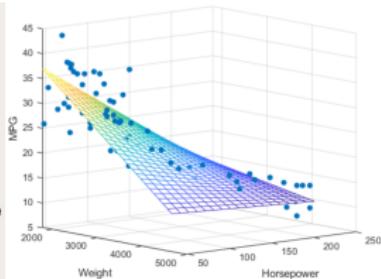
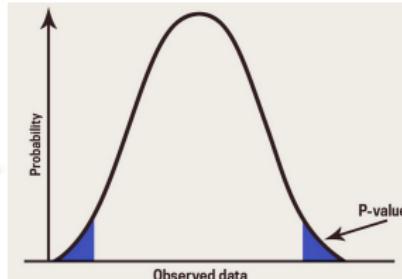
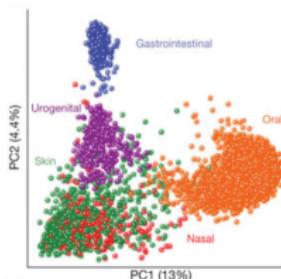
# Qu'est-ce que la statistique?

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?

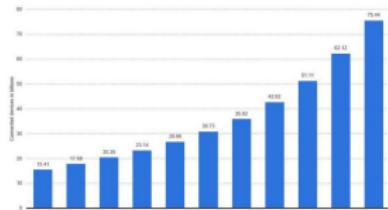
Statistique : ensemble des méthodes de **réduction** des données

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<sup>ème</sup> à IV<sup>ème</sup> 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 ( $\sim 800$ 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<sup>ème</sup> XVIII<sup>ème</sup> 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<sup>ème</sup> 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)
- ...

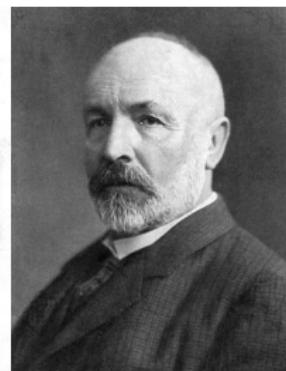


Figure: Augustin Louis Cauchy (1789 – 1857); Bernhard Riemann (1826 – 1866); Georg Cantor (1845 – 1918)

# Courte histoire des mathématiques

## ► XX<sup>ème</sup> 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
- ... ... ... ... ...

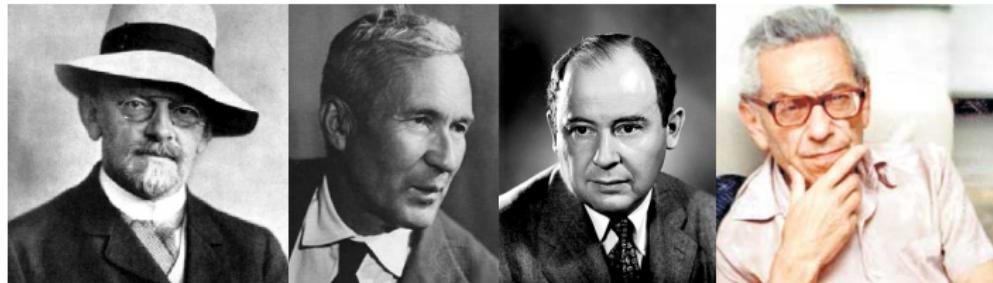


Figure: David Hilbert (1862–1943); Andrei Kolmogorov (1903-1987); John von Neumann (1903–1957); Paul Erdős (1913–1996)

Il reste beaucoup à faire...

1) congrès international des mathématiciens, tenu à Paris en août 1900,  
**David Hilbert** propose 23 problèmes pour le XX<sup>ème</sup> 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

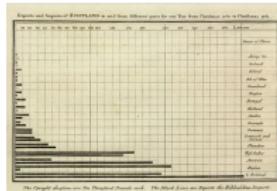
# Statistiques dans l'histoire des mathématiques

Et les statistiques ? ► 3 grandes périodes historiques

## ► ? – XIX<sup>ème</sup> siècle. Statistiques descriptives

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

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



- Karl Pearson impose (1893) l'écart-type (standard deviation)  $\sigma$

Statistiques descriptives → Analyse des données (XX<sup>ème</sup> 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

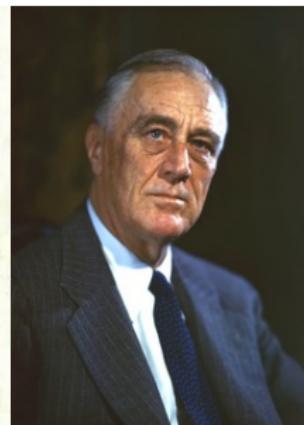
# Statistiques dans l'histoire des mathématiques

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

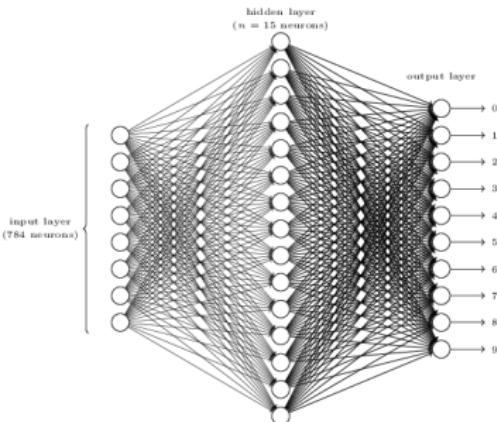
Influence de l'informatique

- 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

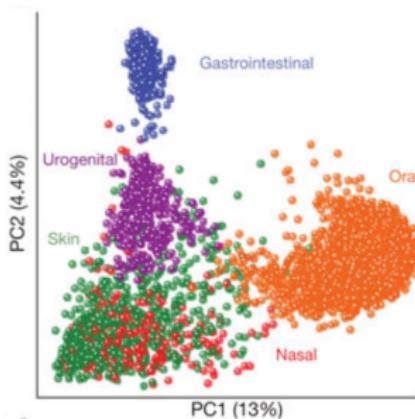


# Statistiques dans l'histoire des mathématiques

## ► 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

# 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

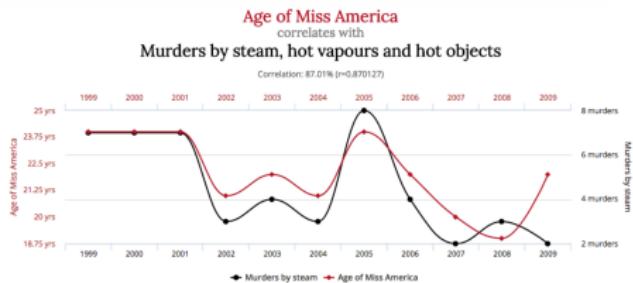
**Objectif :** Prendre une décision raisonnée

# 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

# 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

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

## Moyenne harmonique

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

## Moyenne quadratique

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

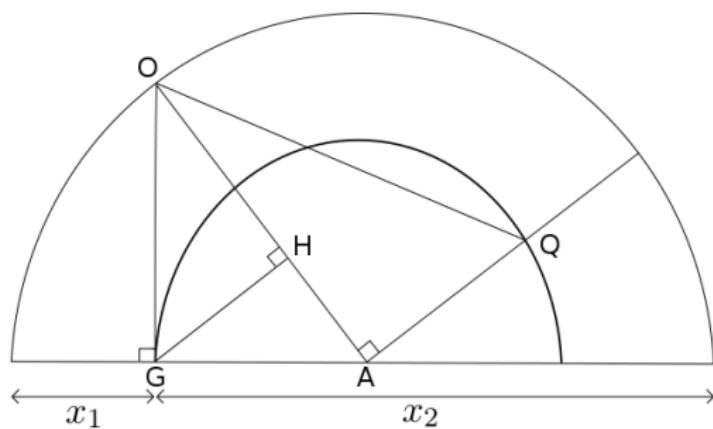
► **Théorème :** Si  $x_1, \dots, x_n > 0$ , alors

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

# 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**  $\sigma$  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  $\sigma_1^2$

Données de taille  $n_2$ , moyenne  $\bar{x}_2$  et variance  $\sigma_2^2$

Donnée  $n_1 + n_2$ , moyenne  $\bar{x}$  variance  $\sigma^2$

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

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  $\sigma^2$
- - : 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  $\gamma_2$ , 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  $q_\alpha$ , est défini comme étant "la plus petite valeur" de la série permettant d'avoir au moins  $\alpha n$  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 sur la série (2) de cet exemple.

# 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  $x_{(k)}$  dans  $[Q_1 - 1.5 \text{ } I/Q, Q_1]$  et  $x_{(b)}$  réalisant le maximum des  $x_{(k)}$  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)

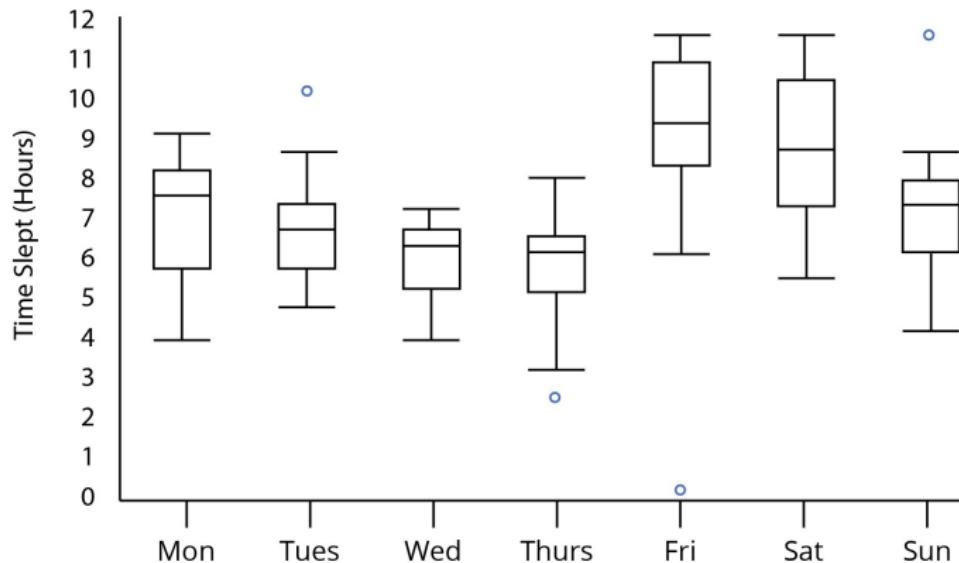


Figure: Nombre d'heures de sommeil pour 20 lycéens pendant une semaine.

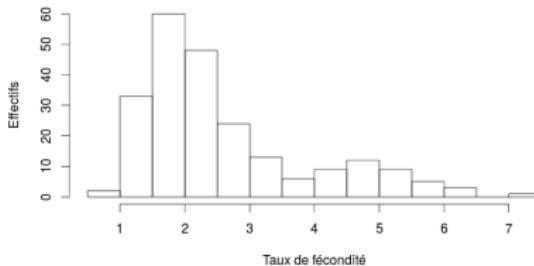
# Histogrammes

- ▶ 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  $n_i = \#\{x_k \mid x_k \in [a_{i-1}, a_i[\}$ , 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

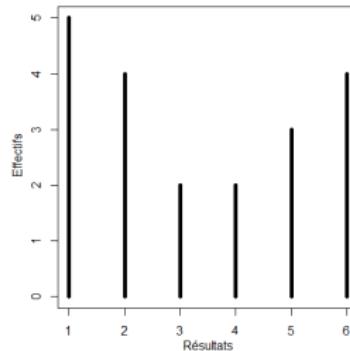


Figure: Résultat de 20 lancés d'un dé équilibré

# Fonction de répartition empirique

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

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

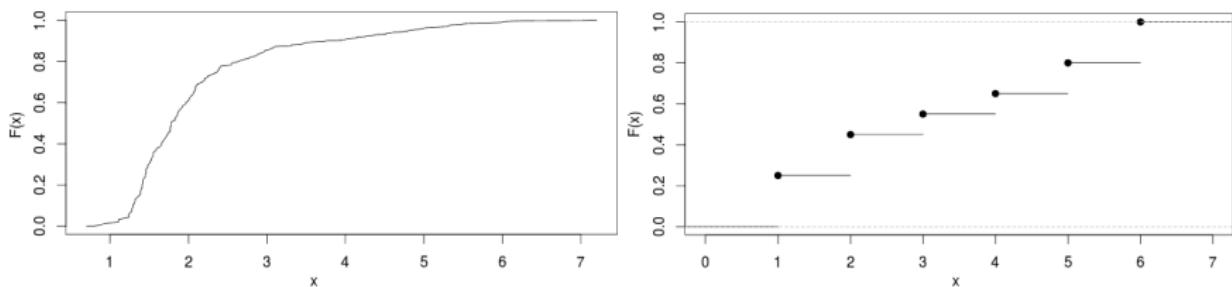


Figure: (gauche) fonction de répartition des taux de fécondité. (droite) fonction de répartition des 20 lancés de dé

# 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?