

# Bases statistiques pour la biologie

Cyril Dalmasso  
cyril.dalmasso@genopole.cnrs.fr

Ecole doctorale GAO  
Université d'Evry Val d'Essonne

20-22 janvier 2014

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités**
  - Généralités
  - Variables aléatoires réelles
  - Distributions usuelles
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités**
  - Généralités
    - Variables aléatoires réelles
    - Distributions usuelles
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire

# Expérience aléatoire, espace fondamental

## Définitions

- Une **expérience aléatoire** est une expérience dont l'issue ne peut être prédite avec certitude.
- On appelle **espace fondamental** (ou **univers**) l'ensemble de toutes les issues possibles. On le note  $\Omega$ .
- Un **événement**  $A$  est un sous-ensemble de l'espace fondamental (c'est à dire :  $A \subset \mathcal{P}(\Omega)$ ).
- Un **événement élémentaire** est l'un des résultats observables à l'issue de l'expérience aléatoire.
- Un **événement composé** est un événement dont la réalisation peut se produire à partir de plusieurs événements élémentaires.

# Événements / Ensembles

## Définitions

- $\Omega$  est l'**événement certain**
- $\emptyset$  est l'**événement impossible**
- $\bar{A}$  (ou  $A^c$ ) est l'**événement contraire** de  $A$
- $A \cap B$  est l'événement 'A et B'
- $A \cup B$  est l'événement 'A ou B'
- $A \cap B = \emptyset$  signifie que A et B sont deux **événements incompatibles**
- $A \subset B$  signifie que A implique B

## Remarque

Les relations entre événements peuvent être représentées par des diagrammes de Venn.

# Probabilité

## Définition axiomatique (Kolmogorov-1933)

Une **probabilité** est une application  $\mathbb{P} : \Omega \rightarrow [0, 1]$  telle que :

- pour tout  $A \in \Omega$ , on a  $\mathbb{P}(A) \geq 0$ ,
- $\mathbb{P}(\Omega) = 1$ ,
- Si  $A \cap B = \emptyset$ , alors  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ .

## Remarque

Une probabilité est une mesure.

## Définition

On appelle **espace probabilisé** le triplet  $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$

# Probabilité

## Propriétés élémentaires

- i)  $\mathbb{P}(\bar{A}) = 1 - \mathbb{P}(A)$
- ii)  $\mathbb{P}(\emptyset) = 0$
- iii)  $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
- iv)  $0 \leq \mathbb{P}(A) \leq 1$
- v)  $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

## Remarque

$\mathbb{P}(A) = 0$  ne signifie pas que  $A$  est un événement impossible.

# Probabilité conditionnelle

## Définition

Soit  $A$  et  $B$  deux événements tels que  $\mathbb{P}(B) \neq 0$ . La **probabilité conditionnelle** de  $A$  par rapport à  $B$ , notée  $\mathbb{P}(A|B)$  ou  $P_B(A)$  (probabilité de  $A$  sachant  $B$ ), est donnée par :

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

# Formule des probabilités totales

## Définition

Soit  $E$  un ensemble.  $A_1, A_2, \dots, A_n$  constituent une partition de  $E$  si :

- $\forall i \in \{1, \dots, n\}; A_i \neq \emptyset$
- $\forall i \neq j; A_i \cap A_j = \emptyset$
- $A_1 \cup A_2 \cup \dots \cup A_n = E$

## Formule des probabilités totales

Si les événements  $B_1, B_2, \dots, B_n$  forment une partition de  $\Omega$

$$\begin{aligned}\mathbb{P}(A) &= \mathbb{P}(B_1)\mathbb{P}(A|B_1) + \mathbb{P}(B_2)\mathbb{P}(A|B_2) + \dots + \mathbb{P}(B_n)\mathbb{P}(A|B_n) \\ &= \sum_{i=1}^n \mathbb{P}(B_i)\mathbb{P}(A|B_i)\end{aligned}$$

# Formule de Bayes

## Formule de Bayes

- Si  $\mathbb{P}(A) \neq 0$  et  $\mathbb{P}(B) \neq 0$  alors

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}$$

- Si les événements  $A_1, A_2, \dots, A_n$  forment une partition de  $\Omega$  alors

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_{j=1}^n \mathbb{P}(A_j)\mathbb{P}(B|A_j)}$$

## Exemple

Un examen systématique de dépistage est institué pour détecter une maladie  $M$ . On sait que le risque d'avoir cette maladie est de 0.001. L'examen donne des faux positifs avec probabilité 0.1 et des faux négatifs avec une probabilité de 0.3. Un individu subit un examen qui se révèle négatif. Quelle est la probabilité qu'il soit malade ?

# Indépendance

## Définition

Deux événements  $A$  et  $B$  sont dits indépendants si et seulement si

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

On note  $A \perp\!\!\!\perp B$

## Remarques

- Si  $A$  et  $B$  sont indépendants, alors  $\mathbb{P}(A|B) = \mathbb{P}(A) \Leftrightarrow \mathbb{P}(B|A) = \mathbb{P}(B)$
- Attention à ne pas confondre indépendance et incompatibilité.

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités**
  - Généralités
  - Variables aléatoires réelles**
  - Distributions usuelles
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire

# Types de variables

- Variable **quantitative** : variable/caractère à laquelle on peut associer un nombre
  - **discrète** : ne peut prendre qu'un nombre fini ou dénombrable de valeurs
  - **continue** : peut prendre toutes les valeurs d'un intervalle de l'ensemble des nombres réels
- Variable **qualitative** : variable/caractère dont les modalités ne sont pas quantifiables
  - **ordinaire** : variable dont les modalités peuvent être ordonnées
  - **nominale** : variable dont les modalités ne peuvent pas être ordonnées

# Variables aléatoires réelles

## Définitions

- Une **variable aléatoire réelle**  $X$  est une application qui à tout élément  $\omega$  de  $\Omega$  associe un nombre réel  $x$

$$X : \Omega \rightarrow \mathbb{R}$$

$$\omega \mapsto x$$

- On appelle **domaine de variation** (ou **support**) de  $X$  l'ensemble  $D_X \subseteq \mathbb{R}$  des valeurs que peut prendre la variable aléatoire  $X$ .

## Remarques

- On note généralement  $X$  la variable aléatoire et  $x$  sa réalisation (c'est à dire  $x = X(\omega_i)$  où  $\omega_i \in \Omega$ ).

# Variables aléatoires réelles

## Définition

- Une variable aléatoire **discrète** est une variable aléatoire dont le domaine de variation contient un nombre fini ou une infinité dénombrable de valeurs.
- Une variable aléatoire **continue** est une variable aléatoire dont le domaine de variation contient une infinité non dénombrable de valeurs.

# Loi de probabilité

## Définitions

- Soit  $X$  une variable aléatoire **discrète** telle que  $\Omega_X = x_1, \dots, x_N$ . La loi de probabilité de  $X$  est définie/caractérisée par sa **fonction de probabilité** qui donne, pour tout  $i \in 1, \dots, N$

$$p_i = \mathbb{P}(X = x_i)$$

- Soit  $X$  une variable aléatoire **continue**. On appelle **densité** de probabilité la fonction  $f(x)$  définie par :

$$f(x) = \lim_{\delta \rightarrow 0} \frac{\mathbb{P}(X \in [x; x + \delta])}{\delta}$$

*Remarque* : Pour  $\delta$  proche de 0,  $f(x)dx \approx \mathbb{P}(X \in [x; x + \delta])$

# Propriétés

## Cas discret

- $\sum_{i=1}^N p_i = 1$
- $\forall A \subset \Omega, \mathbb{P}(A) = \sum_{i: x_i \in A} p_i$

## Cas continu

- $f(x) \geq 0 \forall x \in \mathbb{R}$
- $\int_{-\infty}^{+\infty} f(x) dx = 1$

## Remarque

Toute fonction vérifiant ces deux propriétés est une densité de probabilité.

# Fonction de répartition

## Définition

On appelle **fonction de répartition** la fonction  $F$  définie par :

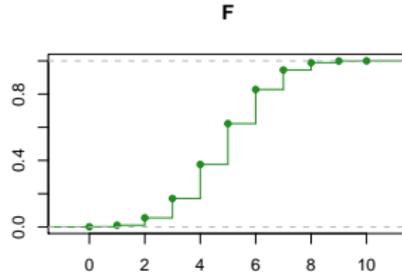
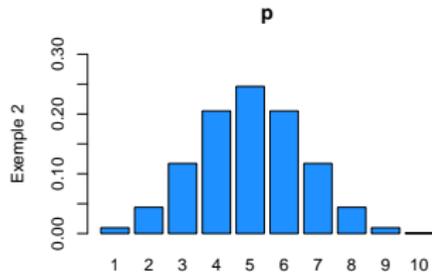
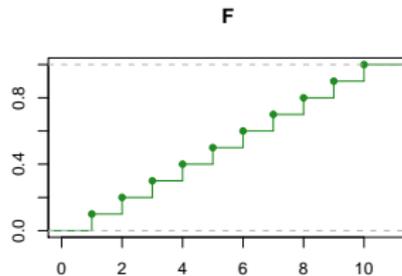
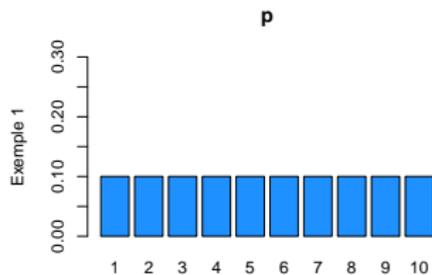
$$F : D_X \longrightarrow [0, 1]$$
$$x_j \longmapsto \mathbb{P}(X \leq x_j)$$

## Propriétés

- i)  $F(x) \in [0, 1]$
- ii)  $F$  est une fonction croissante
- iii)  $\lim_{x \rightarrow -\infty} F(x) = 0$
- iv)  $\lim_{x \rightarrow +\infty} F(x) = 1$
- v)
  - Pour une variable aléatoire discrète,  $F$  est une fonction en escaliers
  - Pour une variable aléatoire continue,  $F$  est une fonction continue

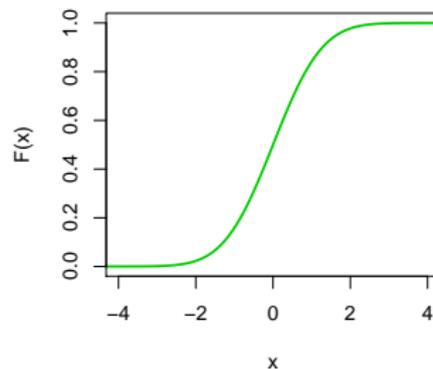
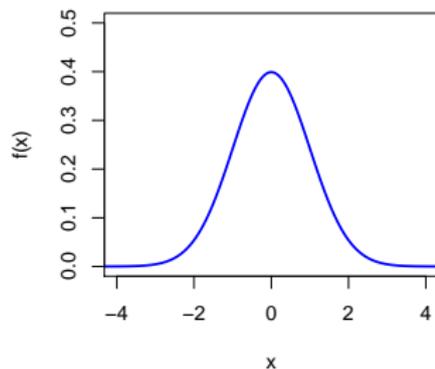
# Fonction de répartition

## Représentation graphique - lois discrètes



# Fonction de répartition

## Représentation graphique - lois continues



# Fonction de répartition

Liens entre fonction de répartition et fonction de probabilité : cas discret

- $F(x) = \sum_{x_i \leq x} \mathbb{P}(x_i)$
- $\mathbb{P}(x_i) = F(x_i) - F(x_{i-1})$

# Fonction de répartition

Liens entre fonction de répartition et densité : cas continu

- $F(x) = \int_{-\infty}^x f(t)dt \Leftrightarrow f(x) = F'(x)$
- Pour tout intervalle  $[a, b] \subset \mathbb{R}$  :

$$\mathbb{P}(X \in [a, b]) = F(b) - F(a) = \int_a^b f(x)dx$$

## Remarque

Il est équivalent de spécifier  $f(x)$  ou  $F(x)$

# Espérance

## Définition

L'**espérance**  $E(X)$  d'une variable aléatoire  $X$  est définie par :

$$E(X) = \sum_{i=1}^N x_i p_i \quad (\text{cas discret})$$

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \quad (\text{cas continu})$$

## Remarques

- L'espérance ne fait pas nécessairement partie de  $D_x$
- L'espérance n'est pas toujours définie

# Espérance

## Théorème de transfert

Soit  $h$  une fonction telle que  $h(x)$  est définie.

$$E(h(X)) = \sum_{i=1}^N h(x_i) p_i \quad (\text{cas discret})$$

$$E(h(X)) = \int_{-\infty}^{+\infty} h(x) f(x) dx \quad (\text{cas continu})$$

# Espérance

## Propriétés (linéarité de l'espérance)

- $E(aX + b) = aE(X) + b$
- $E(X + Y) = E(X) + E(Y)$
- Attention :  $E(XY) \neq E(X)E(Y)$

# Variance et écart-type

## Définitions

- La **variance**  $Var(X)$  d'une variable aléatoire  $X$  est définie par :

$$Var(X) = E \left[ (X - E(X))^2 \right]$$

- L'**écart-type** est défini par

$$\sigma = \sqrt{Var(X)}$$

# Variance et écart-type

## Théorème (de König-Huygens)

Pour toute variable aléatoire réelle  $X$ , on a :

$$\text{Var}(X) = E(X^2) - E(X)^2$$

## Propriétés

- Comme l'espérance, la variance n'existe pas toujours
- $\text{Var}(X) \geq 0$
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

# Quantiles

## Définition

Le **p-quantile** (parfois appelé **p-fractile**) d'une variable aléatoire  $X$  est la valeur  $q_p$  telle que :

$$\mathbb{P}(X \leq q_p) = p; p \in [0, 1]$$

$$\Leftrightarrow q_p = F^{-1}(p)$$

## Quantiles particuliers

- **Médiane** : quantile d'ordre  $i/2$
- **Quartiles** : quantile d'ordre  $i/4$
- **Déciles** : quantile d'ordre  $i/10$
- **Centiles** : quantile d'ordre  $i/100$

# Couples aléatoires

## Fonction de répartition

Pour un couple  $(X, Y)$ , la fonction de répartition conjointe  $F(x, y)$  est donnée par :

$$F(x, y) = \mathbb{P}(X \leq x \cap Y \leq y)$$

## Propriétés

- $F(x, y) \in [0, 1]$
- La fonction  $F$  est monotone croissante par rapport à  $x$  et à  $y$ .

# Couples aléatoires

## Distribution conjointe

- Pour un couple  $(X, Y)$  de variables aléatoires discrètes, la fonction de probabilité est donnée par :

$$p(x, y) = \mathbb{P}(X = x \cap Y = y)$$

- Pour un couple  $(X, Y)$  de variables aléatoires continues, la fonction de densité est donnée par :

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

# Couples aléatoires

## Remarque

La fonction de répartition  $F(x, y)$  se déduit de  $f(x, y)$  par une double intégration :

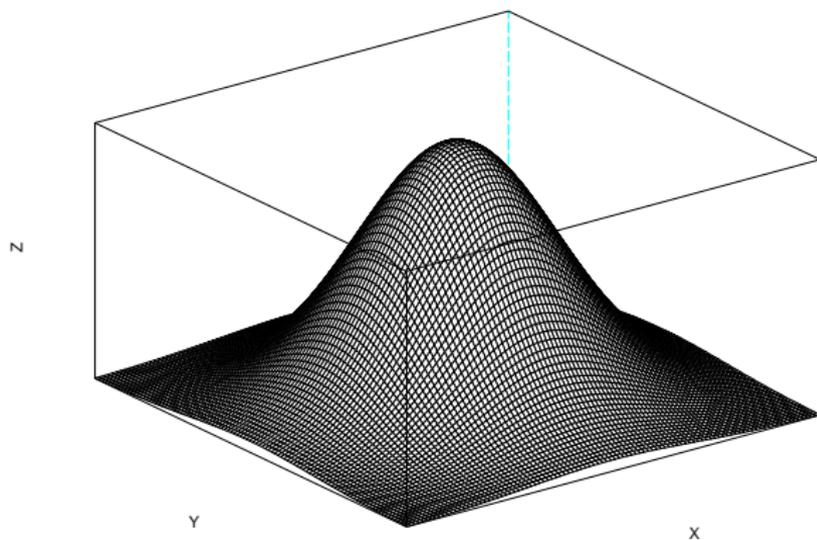
$$F(x, y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) du dv$$

## Propriétés

- $\forall (x, y) \in \mathbb{R}^2 \quad f(x, y) \geq 0$
- $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) dx dy$

# Couples aléatoires

## Exemple de densité jointe



# Couples aléatoires

## Distributions marginales

- Pour un couple  $(X, Y)$  de variables aléatoires discrètes, les fonctions de probabilité marginales sont données par :

$$p_X(x) = \sum_j p(x, y_j)$$

$$p_Y(y) = \sum_i p(x_i, y)$$

# Couples aléatoires

## Distributions marginales

- Pour un couple  $(X, Y)$  de variables aléatoires continues, les densités marginales sont données par :

$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{+\infty} f(x, y) dx$$

# Couples aléatoires

## Distributions conditionnelles

- Pour un couple  $(X, Y)$  de variables aléatoires discrètes, les fonctions de probabilité conditionnelles sont données par :

$$p_X(x|y) = \frac{p(x, y)}{p_Y(y)} \quad \text{avec } p_Y(y) > 0$$

$$p_Y(y|x) = \frac{p(x, y)}{p_X(x)} \quad \text{avec } p_X(x) > 0$$

# Couples aléatoires

## Distributions conditionnelles

- Pour un couple  $(X, Y)$  de variables aléatoires continues, les densités conditionnelles sont données par :

$$f_X(x|y) = \frac{f(x, y)}{f_Y(y)}$$

$$f_Y(y|x) = \frac{f(x, y)}{f_X(x)}$$

# Covariance

## Définition

Pour un couple de variables aléatoires  $(X, Y)$ , la **covariance** est définie par :

$$\begin{aligned} \text{Cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

## Remarques

- $X \perp\!\!\!\perp Y \Rightarrow \text{Cov}(X, Y) = 0$  mais  $\text{Cov}(X, Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$
- $\text{Var}(X) = \text{Cov}(X, X)$

# Coefficient de corrélation

## Définition

Pour un couple de variables aléatoires  $(X, Y)$ , le **coefficient de corrélation** est défini par :

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

## Propriété

$$\rho_{X,Y} \in [-1; 1]$$

## Exemple

Dans un élevage de souris, chaque femelle donne naissance de manière équiprobable à une portée de 1, 2 ou 3 souriceaux. Deux femelles vont mettre bas. Les deux portées auront respectivement  $n$  et  $m$  petits, de manière indépendante. On considère deux variables aléatoires :  $X$ , le minimum de  $n$  et  $m$ , et  $Y$  le maximum. La distribution conjointe du couple  $(X, Y)$  est donnée dans le tableau suivant :

|     | $Y$ | 1   | 2   | 3   |
|-----|-----|-----|-----|-----|
| $X$ |     |     |     |     |
| 1   |     | 1/9 | 2/9 | 2/9 |
| 2   |     | 0   | 1/9 | 2/9 |
| 3   |     | 0   | 0   | 1/9 |

- 1 Donner la distribution marginale de  $X$ .
- 2 Donner la distribution conditionnelle de  $X$  sachant  $Y = 2$ .
- 3 Les variables aléatoires  $X$  et  $Y$  sont-elles indépendantes ?

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités**
  - Généralités
  - Variables aléatoires réelles
  - Distributions usuelles**
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire

# Loi uniforme discrète

## Définition

La loi uniforme discrète sur  $\{1, \dots, n\}$  est la loi d'une variable aléatoire  $X$  qui peut prendre les valeurs  $1, \dots, n$  de manière équiprobable.

## Notation

$$X \sim U(\{1, \dots, n\})$$

## Fonction de probabilité

$$\mathbb{P}(X = k) = \frac{1}{n}; \forall k \in \{1, \dots, n\}$$

# Loi uniforme discrète

## Espérance et variance

$$E(X) = \frac{n+1}{2}$$

$$\text{Var}(X) = \frac{n^2-1}{12}$$

## Exemple type

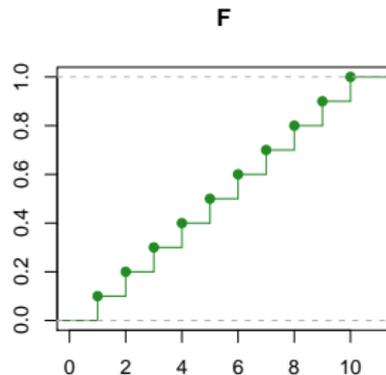
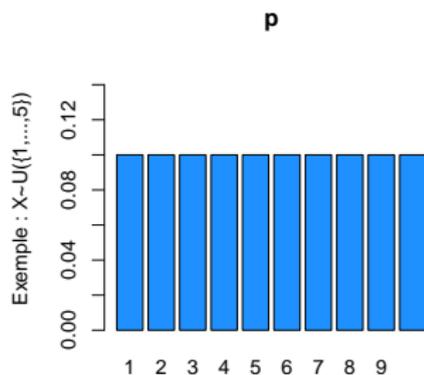
Lancé d'un dé équilibré.

## Fonctions R

sample

# Loi uniforme discrète

## Graphiques



# Loi de Bernoulli

## Définition

La loi de Bernoulli de paramètre  $p$  est la loi d'une variable aléatoire discrète  $X$  qui prend la valeur 1 avec probabilité  $p$  et la valeur 0 avec probabilité  $1 - p$ . L'expérience associée est appelé une **épreuve de Bernoulli**.

## Notation

$$X \sim \mathcal{B}(p)$$

## Fonction de probabilité

$$\mathbb{P}(X = x) = \begin{cases} p & \text{si } x = 1 \\ 1 - p & \text{si } x = 0 \\ 0 & \text{sinon} \end{cases}$$

# Loi de Bernoulli

## Espérance et variance

$$E(X) = p$$

$$\text{Var}(X) = p(1 - p)$$

## Exemple type

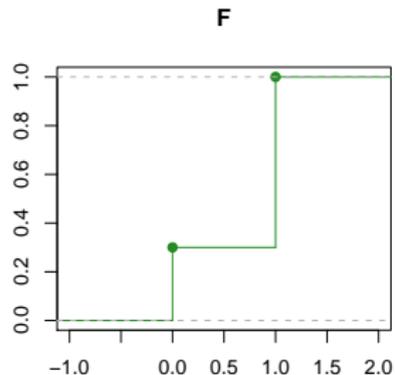
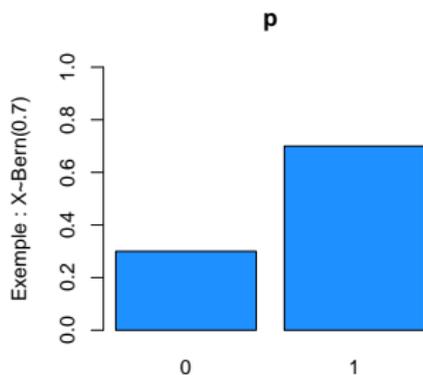
Pile ou face

## Fonctions R

`pbinom dbinom qbinom rbinom`

# Loi de Bernoulli

## Graphiques



# Loi binomiale

## Définition

La loi binomiale de paramètres  $n$  et  $p$  est la loi de la somme  $X$  de  $n$  variables aléatoires  $Y_i$  indépendantes telles que  $Y_i \sim \mathcal{B}(p)$ .

## Notation

$$X \sim \mathcal{B}(n, p)$$

## Fonction de probabilité

$$\mathbb{P}(X = k) = C_n^k p^k (1 - p)^{(n-k)}; k = 1, \dots, n$$

# Loi binomiale

## Espérance et variance

$$E(X) = np$$

$$\text{Var}(X) = np(1 - p)$$

## Exemple type

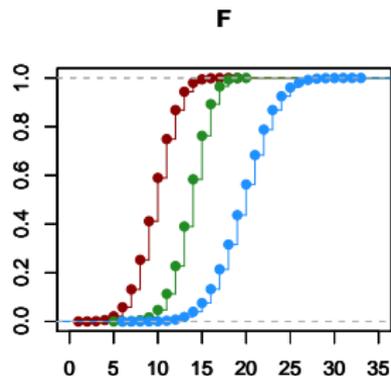
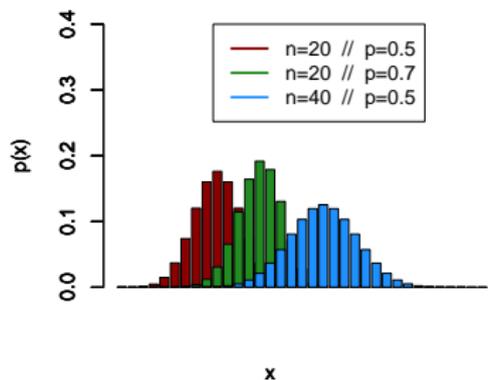
Comptage du nombre de succès sur  $n$  épreuves de Bernoulli.

## Fonctions R

`pbinom dbinom qbinom rbinom`

# Loi binomiale

## Graphiques



# Loi de Poisson

## Définition

La loi de Poisson (parfois appelée loi des événements rares) de paramètre  $\lambda > 0$  est définie par la fonction de probabilité qui suit.

## Notation

$$X \sim \mathcal{P}(\lambda)$$

## Fonction de probabilité

$$\mathbb{P}(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}; k = 1, \dots, n$$

# Loi de Poisson

## Espérance et variance

$$E(X) = \lambda$$

$$\text{Var}(X) = \lambda$$

## Exemple type

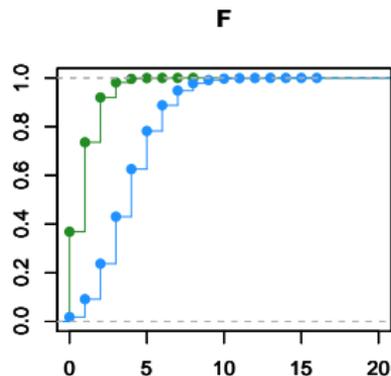
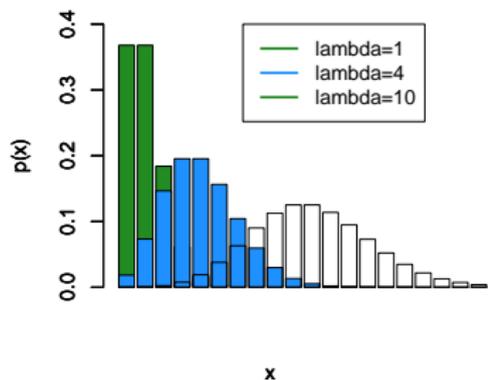
Comptage du nombre d'événements au cours d'un intervalle de temps.

## Fonctions R

ppois dpois qpois rpois

# Loi de Poisson

## Graphiques



# Loi de Poisson

## Théorème 1

Lorsque  $n$  tend vers l'infini et que, simultanément,  $p_n$  devient petit de sorte que  $\lim_{n \rightarrow \infty} np_n = \lambda > 0$ , la loi binomiale de paramètres  $n$  et  $p_n$  converge vers la loi de Poisson de paramètre  $\lambda$ .

En pratique, l'approximation peut être faite lorsque  $n > 30$  et  $np < 5$  ou  $n > 50$  et  $p < 0.1$ .

## Théorème 2

Si  $X_1$  et  $X_2$  sont deux variables aléatoires indépendantes telles que  $X_1 \sim \mathcal{P}(\lambda_1)$  et  $X_2 \sim \mathcal{P}(\lambda_2)$ , alors

$$Y = X_1 + X_2 \sim \mathcal{P}(\lambda_1 + \lambda_2)$$

# Loi géométrique

## Définition

La loi géométrique de paramètre  $p \in [0, 1]$  est la loi de la variable aléatoire  $Y$  qui compte le nombre de répétitions indépendantes d'une épreuve de Bernoulli (de paramètre  $p$ ) jusqu'au premier succès.

## Notation

$$X \sim \mathcal{G}(p)$$

## Fonction de probabilité

$$\mathbb{P}(X = k) = p(1 - p)^{k-1}; k = 1, \dots, n$$

# Loi géométrique

## Espérance et variance

$$E(X) = \frac{1}{p}$$

$$\text{Var}(X) = \frac{1-p}{p^2}$$

## Exemple type

Comptage du nombre d'expériences nécessaires pour obtenir un premier succès en répétant une épreuve de Bernoulli.

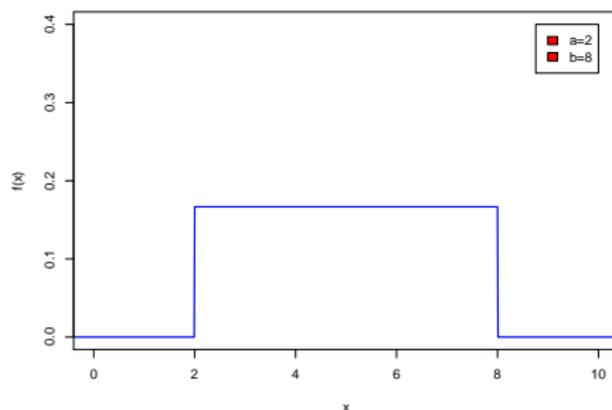
## Fonctions R

pgeom dgeom qgeom rgeom

# Loi uniforme

## Densité

$$f(x) = \frac{1}{b-a} 1_{[a,b]}(x)$$



# Loi uniforme

## Espérance et variance

$$E(X) = \frac{a + b}{2}$$

$$\text{Var}(X) = \frac{(b - a)^2}{12}$$

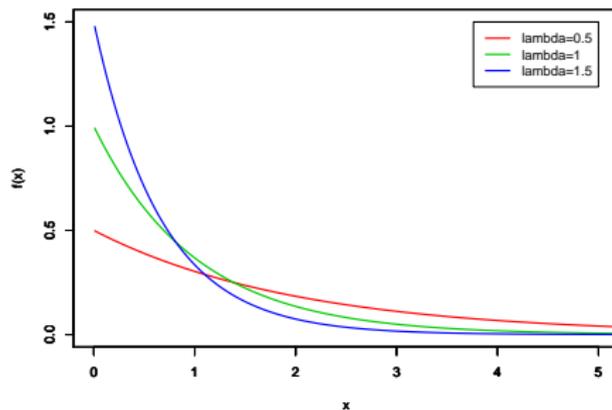
## Fonctions R

punif dunif qunif runif

# Loi exponentielle

## Densité

$$f(x) = \lambda e^{-\lambda x} 1_{x>0}$$



# Loi exponentielle

## Espérance et variance

$$E(X) = \frac{1}{\lambda}$$

$$\text{Var}(X) = \frac{1}{\lambda^2}$$

## Fonctions R

$p_{\text{exp}}$   $d_{\text{exp}}$   $q_{\text{exp}}$   $r_{\text{exp}}$

# Loi normale

## Définition

- Une variable aléatoire  $X$  suit une loi normale (ou loi de Gauss-Laplace) de paramètres  $\mu$  et  $\sigma^2$  si :

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

- On note :  $X \sim N(\mu, \sigma^2)$

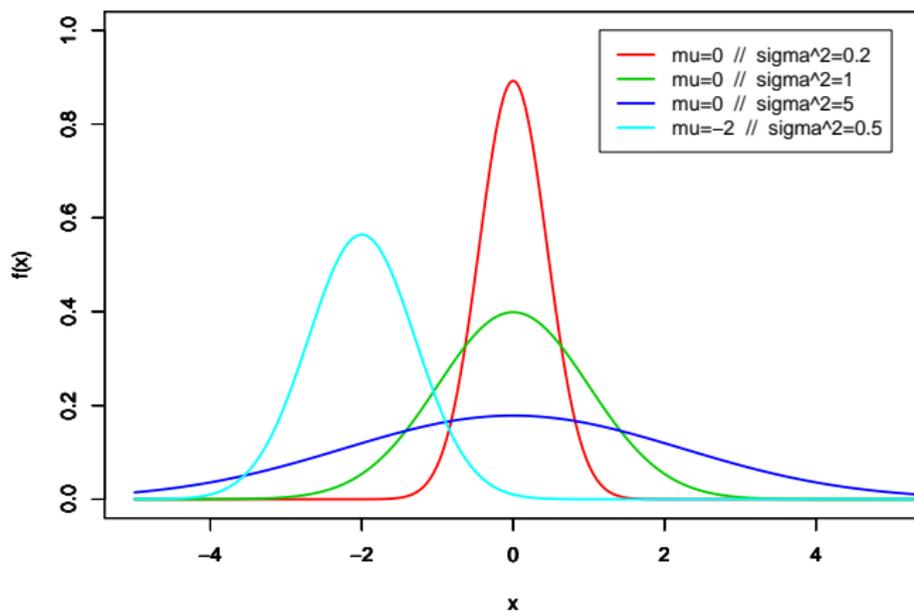
## Fonction de répartition

$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}} dx$$

Il n'existe pas de forme analytique de la fonction de répartition  $F$ .

# Loi normale

## Densité



# Loi normale

## Espérance et variance

Si  $X$  est une variable aléatoire réelle telle que  $X \sim N(\mu, \sigma^2)$ , alors :

- $E(X) = \mu$
- $Var(X) = \sigma^2$

## Fonctions R

`pnorm` `dnorm` `qnorm` `rnorm`

# Loi normale

## Stabilité par combinaisons linéaires

- Si  $X$  est une variable aléatoire telle que  $X \sim N(\mu, \sigma^2)$ , alors :

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

- Si  $X$  et  $Y$  sont deux variables aléatoires indépendantes telles que  $X \sim N(\mu_1, \sigma_1^2)$  et  $Y \sim N(\mu_2, \sigma_2^2)$ , alors :

$$X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

# Loi normale centrée réduite

## Définition

On appelle **loi normale centrée réduite** la loi  $N(0, 1)$ .

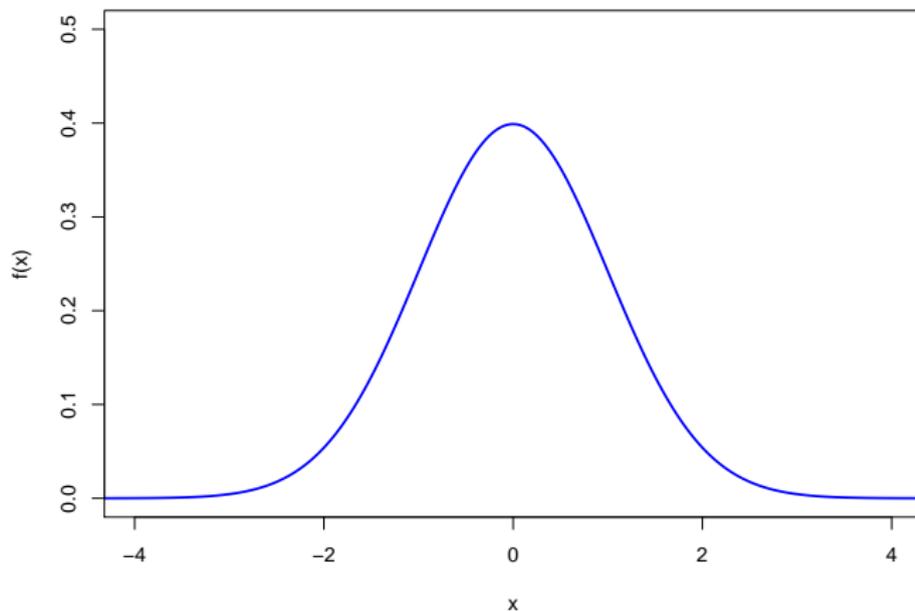
## Propriété

Si  $X \sim N(\mu, \sigma^2)$ , alors  $Y = \frac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$

## Notations

Par convention, on note  $\phi$  la densité d'une  $N(0, 1)$  et  $\Phi$  sa fonction de répartition.

# Loi normale centrée réduite



# Loi normale centrée réduite

## Propriétés

- $\phi$  est une fonction paire
- $\Phi(x) = 1 - \Phi(-x)$
- $\mathbb{P}(|X| \leq x) = \mathbb{P}(-x \leq X \leq x) = 2(\Phi(x) - 1/2)$
- $\mathbb{P}(|X| \geq x) = \mathbb{P}((X \leq -x) \cap (X \geq x)) = 2(1 - \Phi(x))$

# Théorème de la limite centrale

## Théorème

Soit  $X_1, \dots, X_n$   $n$  variables aléatoires indépendantes et identiquement distribuées d'espérance  $\mu$  et de variance  $\sigma^2$  :

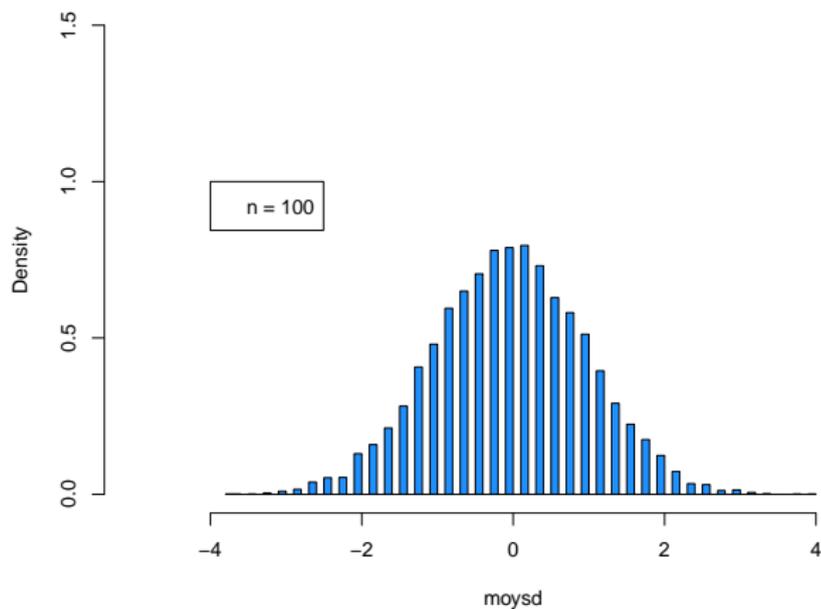
$$Y = \sum_{i=1}^n X_i \xrightarrow{\mathcal{L}} N(n\mu, n\sigma^2)$$
$$\Leftrightarrow Y = \frac{\frac{1}{n} \sum_{i=1}^n X_i - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{\mathcal{L}} N(0, 1)$$

## Remarque

Ce théorème est appelé théorème de la limite centrale (TLC) ou théorème de la limite centrée (TLC) ou théorème central limite (TCL)

# Théorème de la limite centrale

## Illustration



# Théorème de la limite centrale

## Exemple

Une compagnie aérienne fournit des réservations sur le vol d'un appareil de 500 places. La probabilité qu'un passager ayant effectué la réservation pour ce vol ne se présente pas à l'embarquement est de 10%. Si la compagnie aérienne accorde 550 réservations sur ce vol, quel est la probabilité pour que certains passagers se retrouvent sans place ?

# Théorème de la limite centrale

## Approximations de lois

$$X \sim B(n, p)$$

$$\implies$$

$$X \sim P(\mu)$$

$$n \geq 30; p \leq 0.1$$

$$\mu = np$$

$$\Downarrow \mu \geq 20$$

$$\sigma^2 = \mu$$

$$X \sim B(n, p)$$

$$\implies$$

$$N(\mu, \sigma^2)$$

$$n \geq 30; np \geq 5; n(1-p) \geq 5$$

$$\mu = np; \sigma^2 = np(1-p)$$

$$\Uparrow n \geq 30$$

$$X = Y_1 + \dots + Y_n$$

avec  $Y_1, \dots, Y_n$  iid

# Satellites de la loi normale - Loi du $\chi^2$

## Définition

Soient  $X_1, \dots, X_n$   $n$  variables aléatoires indépendantes et identiquement distribuées de loi normale centrée réduite.

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

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

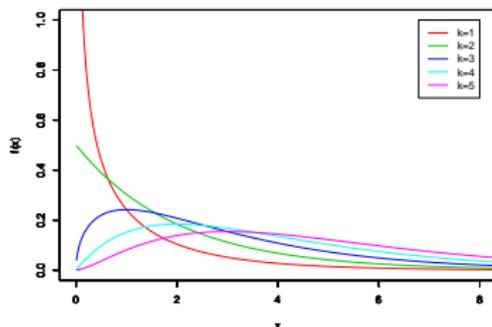
## Propriétés

- 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  $E(Y) = n$  et  $\text{Var}(Y) = 2n$

# Satellites de la loi normale - Loi du $\chi^2$

## Densité

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



## Fonctions R

pchisq dchisq qchisq rchisq

# Satellites de la loi normale - Loi de Student

## Définition

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

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

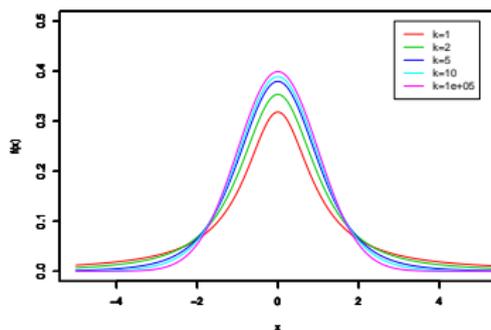
## Propriétés

- $E(T) = 0$
- $Var(T) = \frac{n}{n-2}$  si  $n > 2$

# Satellites de la loi normale - Loi de Student

Densité

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



Fonctions R

pt dt qt rt

# Satellites de la loi normale - Loi de Fisher-Snedecor

## Définition

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

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

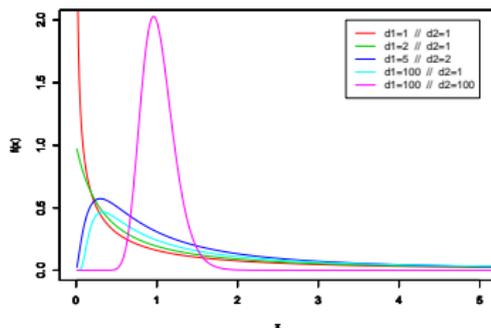
## Remarques

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

## Satellites de la loi normale - Loi de Fisher-Snedecor

## Densité

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



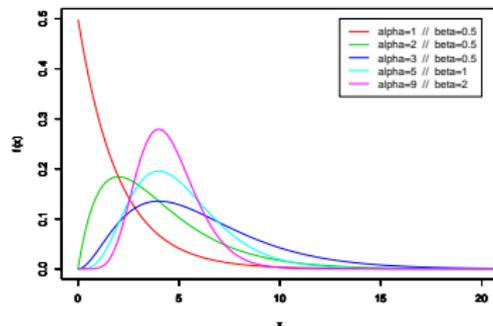
## Fonctions R

pf df qf rf

# Loi Gamma

## Densité

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}$$



## Remarque

Les lois du  $\chi^2$  et les lois exponentielles sont des lois Gamma particulières

# Distributions usuelles

## Lois discrètes

| Lois de probabilité discrètes                                      | Fct de probabilité   | $E(X)$          | $Var(X)$           | Genèse  |
|--|--|-----------------|--------------------|---|
| Uniforme $X \sim U(\{1, 2, \dots, n\})$<br>$n \in \mathbb{N}$      | $\frac{1}{n}$  | $\frac{n+1}{2}$ | $\frac{n^2-1}{12}$ | -   |
| Bernoulli $X \sim B(p)$<br>$0 \leq p \leq 1$                       | $p^x(1-p)^{1-x}$<br>si $x = 0, 1$                              | $p$             | $p(1-p)$           | Lancer d'une pièce de monnaie<br>avec $\mathbb{P}(\text{pile}) = p$   |
| Binomiale $X \sim B(n, p)$<br>$n$ entier $> 0$ , $0 \leq p \leq 1$ | $C_n^x p^x(1-p)^{n-x}$<br>si $x=0,1,\dots,n$                   | $np$            | $np(1-p)$          | Loi de la somme de $n$ v.a.<br>indépendantes de loi $B(p)$  |
| Poisson $X \sim P(\lambda)$<br>$\lambda > 0$                       | $\frac{e^{-\lambda} \lambda^x}{x!}$<br>si $x = 0, 1, 2, \dots$ | $\lambda$       | $\lambda$          | Limite de la loi binomiale lorsque<br>$n \rightarrow \infty$ , $np_n \rightarrow \lambda > 0$<br>et $p_n \rightarrow 0$ |
| Géométrique $X \sim G(p)$<br>$0 \leq p \leq 1$                     | $p(1-p)^{x-1}$<br>si $x = 1, 2, \dots$                         | $\frac{1}{p}$   | $\frac{1-p}{p^2}$  | Nombre de lancers nécessaire<br>pour l'obtention du premier <i>pile</i><br>avec $\mathbb{P}(\text{pile}) = p$           |
| Binomiale négative $X \sim BN(n, p)$<br>$0 \leq p \leq 1$          | $C_{x-1}^{n-1} p^n(1-p)^{x-n}$<br>si $x = n, n+1, \dots$       | $\frac{n}{p}$   | $\frac{n(1-p)}{p}$ | Loi de la somme de $n$ v.a.<br>indépendantes de loi $G(p)$  |

# Distributions usuelles

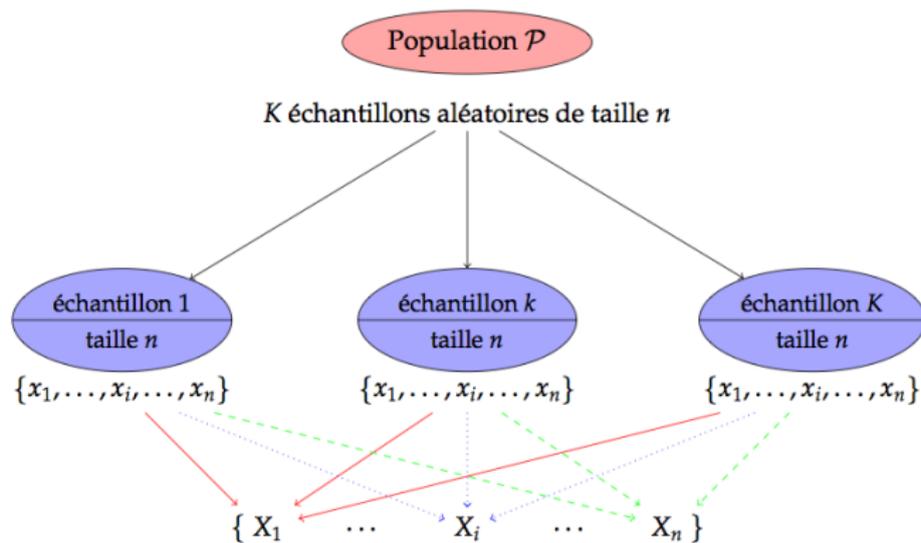
## Lois continues

| Lois de probabilité continues  | Fct de densité  | Fct de répartition             | $E(X)$                 | $Var(X)$                               |
|--|---|--------------------------------|------------------------|--|
| Uniforme $X \sim U[a, b]$<br>$a < b$   | $\frac{1}{b-a} 1_{[a,b]}(x)$  | $\frac{x-a}{b-a} 1_{[a,b]}(x)$ | $\frac{a+b}{2}$        | $\frac{(b-a)^2}{12}$                   |
| Exponentielle $X \sim Exp(\lambda)$<br>$\lambda > 0$                                 | $\lambda e^{-\lambda x} 1_{x>0}$  | $1 - e^{-\lambda x}$           | $\frac{1}{\lambda}$    | $\frac{1}{\lambda^2}$                  |
| Normale $X \sim N(\mu, \sigma^2)$<br>$\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$ | $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}}$                               | -                              | $\mu$                  | $\sigma^2$                             |
| Khi-deux $X \sim \chi_\nu^2$<br>$\nu > 0$  | $\frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{(\nu-2)/2} e^{-x/2}$  | -                              | $\nu$                  | $2\nu$                                 |
| Student $X \sim \chi_\nu^2$<br>$\nu > 0$   | $\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{n\pi} \Gamma(\frac{\nu}{2}) (1 + \frac{x^2}{n})^{\frac{\nu+1}{2}}}$ | -                              | 0<br>si $\nu \geq 2$   | $\frac{\nu}{\nu-2}$<br>si $\nu \geq 3$ |
| Cauchy $X \sim C(\mu, \sigma)$<br>$\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+$    | $\frac{\sigma}{\pi [(x-\mu)^2 + \sigma^2]}$   | -                              | n'existe pas           | n'existe pas                           |
| Gamma $X \sim \Gamma(\alpha, \beta)$<br>$\alpha > 0, \beta > 0$                      | $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-x\beta}$<br>si $x \geq 0$                           | -                              | $\frac{\alpha}{\beta}$ | $\frac{\alpha}{\beta^2}$               |

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation**
- 5 Tests d'hypothèses
- 6 Modèle linéaire

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation**
  - **Estimateur**
    - Estimation ponctuelle
    - Intervalles de confiance
- 5 Tests d'hypothèses
- 6 Modèle linéaire

# Echantillonnage



# Echantillonnage

## Définition

On appelle **échantillon aléatoire** (ou  $n$ -échantillon) le vecteur aléatoire  $(X_1, \dots, X_n)$

## Remarque

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.

# Modèle statistique

## Définition

On appelle **modèle statistique** la donnée du triplet  $(\Omega, \mathcal{A}, (\mathcal{L}_\theta)_{\theta \in \Theta})$  où :

- $\Omega$  est l'univers
- $\mathcal{A}$  est l'ensemble des parties de  $\Omega$
- $(\mathcal{L}_\theta)_{\theta \in \Theta}$  est une famille de lois de probabilité indicée par un vecteur de paramètres  $\theta \in \Theta$

## Remarque

En général, on suppose  $X \sim \mathcal{L}_\theta$  et on cherche à obtenir de l'information sur  $\theta$ .

# Statistique

## Définition

On appelle **statistique** toute fonction du n-échantillon  $X_1, \dots, X_n$  :

$$\begin{aligned} T &: \mathbb{R}^n \longrightarrow \mathbb{R} \\ (X_1, \dots, X_n) &\longmapsto T(X_1, \dots, X_n) \end{aligned}$$

## Remarque

$t = T(x_1, \dots, x_n)$  est une réalisation de la variable aléatoire  $T$ .

# Statistique

## Exemples

- Somme :  $S_n = \sum_{i=1}^n X_i$

- Moyenne empirique (ou moyenne expérimentale / observée) :

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

- Variance empirique (ou variance expérimentale / observée) :

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

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

# Estimation

## Position du problème

- Soit  $X$  une variable aléatoire d'intérêt de loi  $\mathcal{L}_\theta$ ,  $\theta \in \Theta$ .
- Soit  $x_1, \dots, x_n$  une observation du  $n$ -échantillon  $X_1, \dots, X_n$ .
- Comment estimer  $\theta$  à partir de  $x_1, \dots, x_n$  ?

# Estimation

On distingue...

- Estimation ponctuelle
- Estimation par intervalles de confiance

# Estimation ponctuelle

Un **estimateur ponctuel** est une statistique dont la réalisation (pour un échantillon donnée) constitue une estimation de l'un des paramètres  $\theta$  de la distribution (ou de l'une des fonctions permettant de la caractériser).

## Définition

On appelle estimateur de  $\theta$  toute statistique  $Z$  à valeurs dans l'espace des paramètres  $\Theta$ .

## Notation

On note généralement  $\hat{\theta}$  l'estimateur de  $\theta$ .

# Qualités d'un estimateur

## Définitions

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

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta$$

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

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

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

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

# Qualités d'un estimateur

## Définitions (suite)

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

$$EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + [b(\hat{\theta})]^2$$

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

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

## Remarque

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

# Estimateurs usuels

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

# Estimateurs usuels

## Exercice

Soit  $X_1, \dots, X_n$  un  $n$ -échantillon de loi quelconque tel que  $E(X_i) = \mu$  et  $V(X_i) = \sigma^2$ .

- 1 Montrer que  $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$  est un estimateur sans biais et convergent de  $\mu$ .
- 2 Montrer que  $S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$  est un estimateur biaisé de  $\sigma^2$ .
- 3 En déduire un estimateur sans biais de  $\sigma^2$ .

# Moyenne empirique

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

## Propriétés

Soit  $X_1, \dots, X_n$  un  $n$ -échantillon tel que  $E(X_i) = \mu$  et  $V(X_i) = \sigma^2$

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

## Corollaire

$\bar{X}_n$  est un estimateur sans biais et convergent de  $\mu$ .

# Moyenne empirique

## Loi forte des grands nombres

- Si  $X_1, \dots, X_n$  sont  $n$  variable aléatoires indépendantes et identiquement distribuées telles que :

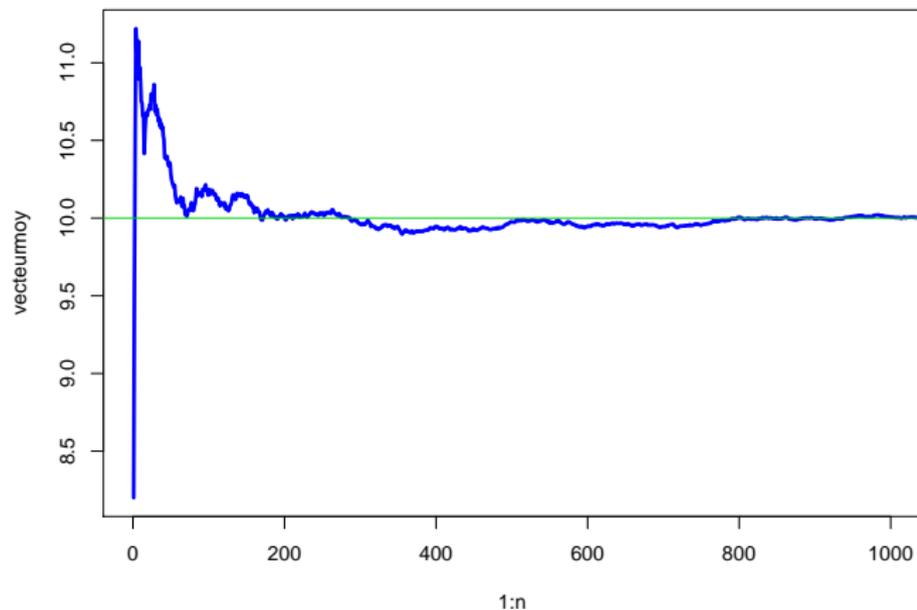
$$E(|X_1|) < \infty \quad \text{et} \quad E(X_1) = \mu$$

- Alors

$$\overline{X_n} \xrightarrow{p.s.} \mu$$

# Moyenne empirique

## Loi forte des grands nombres



# Variance empirique et variance empirique corrigée

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

## Propriétés

Soit  $X_1, \dots, X_n$  un  $n$ -échantillon tel que  $E(X_i) = \mu$  et  $V(X_i) = \sigma^2$

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

## Corollaire

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

# Construction d'estimateurs

## Méthodes

- Méthode des moments
- Maximum de vraisemblance

# Méthode des moments

## Définitions

Soit  $X$  une variable aléatoire réelle.

- On appelle **moment d'ordre**  $k$  la quantité :

$$m_k = E(X^k)$$

- On appelle **moment centré d'ordre**  $k$  la quantité :

$$\mu_k = E((X - E(X))^k)$$

## Remarque

- Espérance = moment d'ordre 1
- Variance = moment centré d'ordre 2

# Méthode des moments

## Construction d'estimateurs par la méthode des moments

Si  $\theta$  comporte  $d$  paramètres réels indépendants, l'estimateur des moments de  $\theta$  est obtenu par la résolution du système d'équations :

$$\overline{x^k} = E(X^k); 1 \leq k \leq d$$

(où  $E(X^k)$  dépend des paramètres  $\theta$ )

## Remarques

- (+) L'estimateur des moments est (en général) facilement accessible
- (-) La construction de l'estimateur des moments nécessite de grandes tailles d'échantillons

# Méthode des moments

## Exemple

Pour estimer la taille  $l$  d'une population d'oiseaux, on en capture respectivement  $X_1, \dots, X_n$  pendant  $n$  journées successives en les relâchant le soir. On suppose que chaque oiseau a la même probabilité inconnue  $p$  de se faire capturer chaque jour et ce indépendamment des autres oiseaux.

- 1 Proposer un modèle.
- 2 Déterminer un estimateur de  $(l, p)$  par la méthode des moments.

# Maximum de vraisemblance

## Définition

Soit  $X$  une va de densité  $f(x, \theta)$  où  $\theta$  est un vecteur de  $k$  paramètres  $\theta = (\theta_1, \dots, \theta_k)$ . Soit  $X_1, \dots, X_n$  un n-échantillon de  $X$ . La fonction de **vraisemblance** est définie par :

$$L(X, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

## Remarque

Alors que la densité est, pour une valeur de  $\theta$  donnée, une fonction de l'observation  $x$ , la vraisemblance est, pour une observations  $x$  donnée, une fonction du paramètre  $\theta$  :  $L_x(\theta) = f_\theta(x)$

# Maximum de vraisemblance

## Estimateur du maximum de vraisemblance

L'**estimateur du maximum de vraisemblance** (EMV) est la valeur de  $\theta$  qui maximise la fonction de vraisemblance  $L_X(\theta)$  :

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}(L_X(\theta))$$

## Remarque

En pratique, on maximise la log-vraisemblance :

$$l(\theta) = \ln(L(\theta)) = \sum_{i=1}^n \ln(f(x_i, \theta))$$

# Maximum de vraisemblance

## Propriétés de l'EMV

L'estimateur du maximum de vraisemblance est :

- convergent
- asymptotiquement normal
- asymptotiquement de variance minimale

## Remarques

- Soit  $\gamma = g(\theta)$ . L'EMV de  $\gamma$  est  $g(\hat{\theta})$  où  $\hat{\theta}$  est l'EMV de  $\theta$ .
- L'EMV n'existe pas toujours et n'est pas nécessairement unique.

# Maximum de vraisemblance

## Exemple

- On suppose que la durée de vie d'une cellule suit une distribution exponentielle de paramètre  $\lambda$  (dont la densité est donnée par :  
 $f(x) = \lambda e^{-\lambda x} 1_{\mathbb{R}_+}(x)$ )
- Pour estimer le paramètre  $\lambda$ , on réalise des mesures sur 200 cellules.
- Déterminer un estimateur du paramètre  $\lambda$  par la méthode du maximum de vraisemblance.

# Estimation par intervalle

## Définition

Un **intervalle de confiance** de niveau  $1 - \alpha$  du paramètre  $\theta$  est un intervalle  $[a, b]$  tel que :

$$\mathbb{P}(\theta \in [a, b]) = 1 - \alpha$$

## Remarque

Ce sont les bornes  $a$  et  $b$  de l'intervalle qui sont aléatoires.



# Estimation par intervalle

## Exemple

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  $N(\mu, \sigma^2)$  dont l'espérance  $\mu$  est la valeur que l'on cherche à déterminer, et dont l'écart-type  $\sigma$  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.

- 1 Déterminer à partir de ces mesures un intervalle de confiance pour  $\mu$  avec un coefficient de sécurité de 95%.
- 2 Quelle taille d'échantillon est nécessaire pour avoir un intervalle plus petit que 0.1. (en mg/litre) ?
- 3 Recalculez l'intervalle de confiance en supposant que la variance est inconnue.

# Estimation par intervalle

## Famille gaussienne, variance connue

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

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

est un intervalle de confiance de niveau  $(1 - \alpha)$  pour  $\mu$

# Estimation par intervalle

## Famille gaussienne, variance inconnue

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

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

est un intervalle de confiance de niveau  $(1 - \alpha)$  pour  $\mu$

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses**
- 6 Modèle linéaire

# Tests d'hypothèses

## Définitions

- Un **test statistique** est une procédure permettant de trancher entre deux hypothèses au vue des observations.
- Une **hypothèse statistique** est un énoncé portant sur les caractéristiques d'une population (paramètre ou forme d'une distribution)

# Exemple

## Exemple

- On désire étudier la durée de vie d'une fleur  $F$ . On admet que la durée de vie de cette fleur est une variable aléatoire gaussienne dont l'espérance est de 77 jours dans des conditions normales. On suppose que l'écart-type est égal à 10 jours.
- Un spécialiste propose une alimentation qui - selon lui - augmente la durée de vie moyenne de cette fleur. Pour s'en assurer un laboratoire soumet 10 fleurs au régime proposé. A la fin de l'expérience, les durées de vie de ces 10 fleurs sont les suivantes (en jours) : 94, 73, 85, 82, 84, 95, 71, 86, 82, 68.
- Proposer un test au niveau 5% permettant de déterminer si le régime proposé par le spécialiste a un effet significatif ou pas.

# Démarche

- 1 Choisir les hypothèses à tester ( $H_0$  et  $H_1$ )
- 2 Fixer le niveau du test  $\alpha$
- 3 Choisir une statistique de test
- 4 Déterminer la règle de décision (région de rejet  $\Gamma$ )
- 5 Calculer la statistique (et la p-valeur)
- 6 Conclure

# Risques d'erreurs

## Résultats possibles

| Décision \ Réalité | Ne pas rejeter $H_0$<br>(conclure $H_0$ ) | Rejeter $H_0$<br>(conclure $H_1$ ) |
|--------------------|---|------------------------------------|
| $H_0$ vraie        | OK  | Erreur de type I                   |
| $H_1$ vraie        | Erreur de type II                         | OK                                 |

## Définitions

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

# Asymétrie

## Hypothèse nulle et hypothèse alternative

- L'**hypothèse nulle** (notée  $H_0$ ) est l'hypothèse privilégiée. C'est celle qui est supposée vraie par défaut (vérité établie) et qui sera conservée en cas de doutes (trop importants).
- L'**hypothèse alternative** (notée  $H_1$ ) contredit l'hypothèse nulle. C'est l'hypothèse que l'on cherche à montrer.

# Hypothèse simple/composite

## Exemple d'hypothèses simples

- $H : \theta = \theta_0$

## Exemple d'hypothèses composites

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

# Tests unilatéraux / bilatéraux

## Test unilatéral (à droite)

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta > \theta_0$$

## Test bilatéral

$$H_0 : \theta = \theta_0$$

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

# 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.

# Règle de décision

## Choix du niveau du test

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

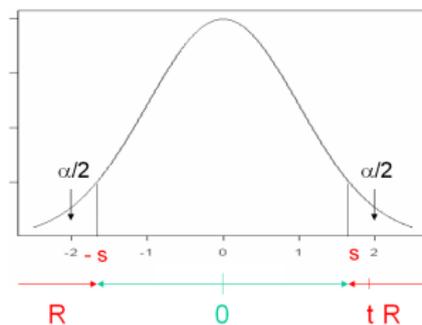
# Région de rejet

## Définition

La **Région de rejet** est l'ensemble  $R$  des valeurs (de la statistique de test) pour lesquelles l'hypothèse nulle est rejetée.

## Démarche de Neyman Pearson

Maximiser la puissance tout en contrôlant  $\alpha$ .



# Risques d'erreurs

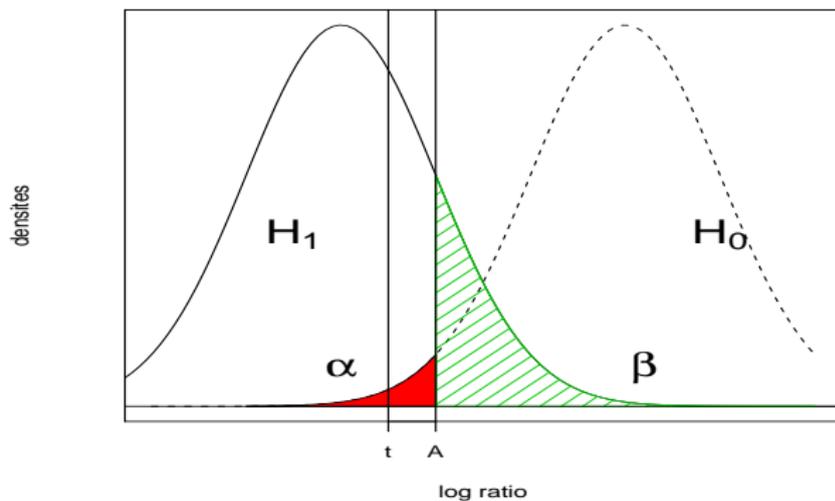


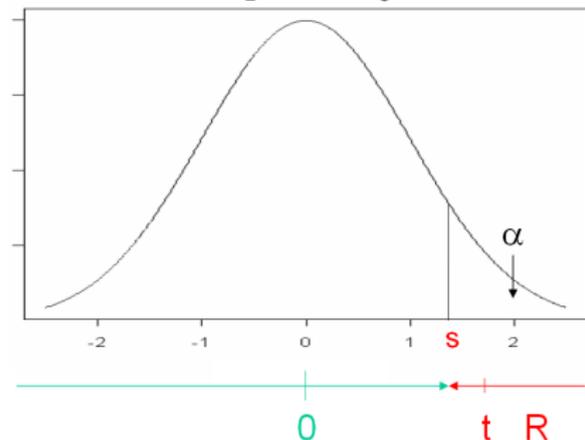
FIGURE : Erreur de type I et II et région critique de la forme  $\Gamma = ] - \infty, A]$

# Tests unilatéraux / bilatéraux

## Test unilatéral (à droite)

$$H_0 : \theta = \theta_0$$

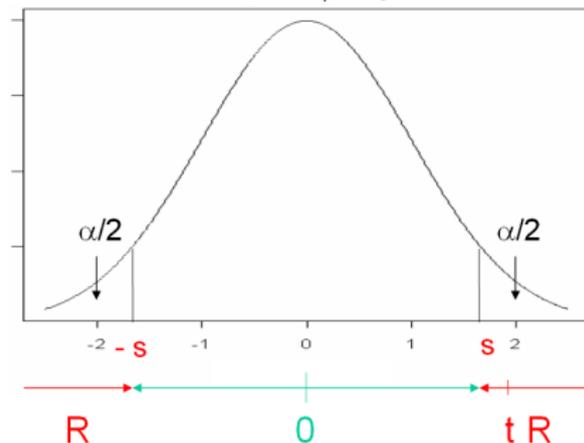
$$H_1 : \theta > \theta_0$$



## Test bilatéral

$$H_0 : \theta = \theta_0$$

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



# Degré de signification (p-valeur)

## Définition

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

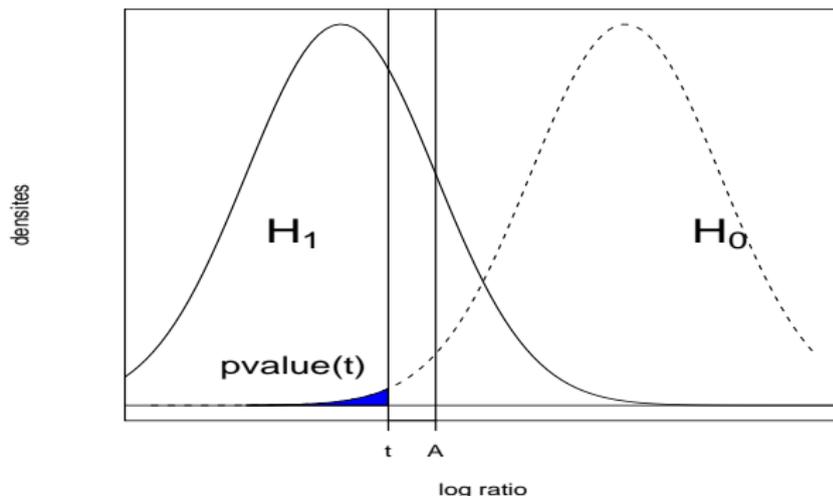
$$p = \min\{\alpha \mid T \in \Gamma_\alpha\}$$

| Test unilatéral à droite      | Test unilatéral à gauche      | Test bilatéral                    |
|-------------------------------|-------------------------------|-----------------------------------|
| $p = \mathbb{P}(T > t   H_0)$ | $p = \mathbb{P}(T < t   H_0)$ | $p = \mathbb{P}( T  >  t    H_0)$ |

## 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
- **En pratique, on rejette  $H_0$  lorsque  $p < \alpha$**

# 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)

## 5 Tests d'hypothèses

- Démarche
- Tests sur l'espérance d'un échantillon
- Comparaison de deux échantillons
- Tests du  $\chi^2$
- Tests non paramétriques

# Test sur l'espérance d'un échantillon gaussien

## Cas 1 : variance connue (test z)

- Présupposés :  $X_1, \dots, X_n$  iid avec  $X_i \sim N(\mu, \sigma_0^2)$ ,  $\sigma_0^2$  connu.
- Hypothèse nulle :  $H_0 : \mu = \mu_0$
- Statistique de test :

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \underset{H_0}{\sim} N(0, 1)$$

# Test sur l'espérance d'un échantillon gaussien

## Exemple

- On désire étudier la durée de vie d'une fleur  $F$ . On admet que la durée de vie de cette fleur est une variable aléatoire gaussienne dont l'espérance est de 77 jours dans des conditions normales. On suppose que l'écart-type est égal à 10 jours.
- Un spécialiste propose une alimentation qui - selon lui - augmente la durée de vie moyenne de cette fleur. Pour s'en assurer un laboratoire soumet 10 fleurs au régime proposé. A la fin de l'expérience, les durées de vie de ces 10 fleurs sont les suivantes (en jours) : 94, 73, 85, 82, 84, 95, 71, 86, 82, 68.
- Proposer un test au niveau 5% permettant de déterminer si le régime proposé par le spécialiste a un effet significatif ou pas.

# Test sur l'espérance d'un échantillon gaussien

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

- Présupposés :  $X_1, \dots, X_n$  iid avec  $X_i \sim N(\mu, \sigma^2)$ ,  $\sigma^2$  inconnu.
- Hypothèse nulle :  $H_0 : \mu = \mu_0$
- Statistique de test :

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{S^{*2}}{n}}} \underset{H_0}{\sim} t_{n-1}$$

## Fonction R

t.test

# Test sur l'espérance d'un échantillon gaussien

## Exemple

- On note  $X$  la température intérieure (en  $^{\circ}\text{C}$ ) d'une espèce de crabes du Pacifique, prise à une température ambiante de  $24.3^{\circ}\text{C}$ . On suppose que  $X$  suit une loi normale  $N(\mu; \sigma^2)$  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  $\sum_i x_i^2 = 13255.53$ .

- 1 Mettre en place un test statistique de niveau  $\alpha = 5\%$  pour déterminer si cette espèce de crabes possède sa propre température intérieure ou si cette dernière est la même que la température ambiante.
- 2 On suppose qu'en réalité la température moyenne des crabes est  $25^{\circ}\text{C}$ . Quelle est la puissance du test construit pour détecter une telle différence ?

# Test sur l'espérance d'un échantillon de loi quelconque

## Cas 1 : n grand

- Présupposés :  $X_1, \dots, X_n$  iid avec  $X_i \sim \mathcal{L}$  inconnue.
- Hypothèse nulle :  $H_0 : \mu = \mu_0$
- Statistique de test :
  - Si  $\sigma_0^2$  connue

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \xrightarrow[H_0]{\mathcal{L}} N(0, 1) \text{ (TLC)}$$

- Si  $\sigma_0^2$  inconnue

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{S^{*2}}{n}}} \xrightarrow[H_0]{\mathcal{L}} N(0, 1) \text{ (TLC)}$$

## Fonction R

t.test

# Test sur l'espérance d'un échantillon de loi quelconque

## Exemple

- 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 l'espérance d'un échantillon de loi quelconque

Cas 2 :  $n$  petit

Test du signe

$$Z_n = \sum 1_{(X_i - \mu_0) > 0} \underset{H_0}{\sim} B(n, p)$$

où  $p = P(X_i > \mu_0)$

Test des signes et rangs de wilcoxon

$$W_n = \sum R_i 1_{(X_i - \mu_0) > 0} \underset{H_0}{\sim} \text{loi tabulée}$$

où  $R_i = \text{rang de } |X_i - \mu_0|$

# Test sur un pourcentage - n grand

## Cas 1 : n grand

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

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

## Remarque

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

# Test sur un pourcentage - $n$ grand

## Exemple

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.

- 1 Le pourcentage d'anomalies chromosomique est-il significativement différent d'il y a 10 ans.
- 2 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

## Cas 2 : n petit

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

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

## Fonction R

binom.test

# Test sur un pourcentage - $n$ petit

## exemple

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 ?

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses**
  - Démarche
  - Tests sur l'espérance d'un échantillon
  - Comparaison de deux échantillons**
  - Tests du  $\chi^2$
  - Tests non paramétriques
- 6 Modèle linéaire

# 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

- Présupposés :

$X_{11}, \dots, X_{n_1}$  iid avec  $X_{1i} \sim N(\mu_1, \sigma_1^2)$

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

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

- Hypothèse nulle :  $H_0 : \mu_1 = \mu_2$
- Statistique de test :

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

## Fonction R

t.test

# Comparaison des espérances de deux échantillons gaussiens

## Exemple

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.

- 1 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

- Présupposés :  
 $X_1, \dots, X_{n_1}$  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 :

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^{*2}}{n_1} + \frac{S_2^{*2}}{n_2}}} \xrightarrow[H_0]{\mathcal{L}} N(0, 1) \text{ (TLC)}$$

Fonction R

t.test

# Comparaison des espérances de deux échantillons - $n_1$ et $n_2$ grands

## Exemple

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<sup>me</sup> 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

$$\sum x_i = 3780$$

$$\sum x_i^2 = 84884$$

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

$$\sum y_i = 2043$$

$$\sum y_i^2 = 46586$$

Que peut-on conclure ?

# Comparaison de deux pourcentages - $n_1$ et $n_2$ grands

- Présupposés :  
 $X_1, \dots, X_{n_1}$  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 :

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

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 chi 2 (voir plus loin)

# Comparaison de deux pourcentages - $n_1$ et $n_2$ grands

## Exemple

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 de deux pourcentages - n petit

Test exact de Fisher

Fonction R

`fisher.test`

# 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

- Présupposés :

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

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

Soit  $D_i = X_i - Y_i$

- Hypothèse nulle :  $H_0 : \mu_d = 0$

- Statistique de test

- $n$  petit

$$\frac{\bar{D} - \mu_d}{\sqrt{\frac{S_d^{*2}}{n}}} \underset{H_0}{\sim} t_{n-1}$$

- $n$  grand, loi quelconque

$$\frac{\bar{D} - \mu_d}{\sqrt{\frac{S_d^{*2}}{n}}} \underset{H_0}{\xrightarrow{\mathcal{L}}} N(0, 1)$$

# Comparaison des espérances de deux échantillons appariés

## Exemple

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 observation normalement distribuées. Le nouveau médicament est-il efficace ?

# Comparaison des variances de deux échantillons gaussiens

- Présupposés :

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

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

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

- Statistique de test :

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

## Fonction R

var.test

# Comparaison des variances de deux échantillons gaussiens

## Exemple

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.

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

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses**
  - Démarche
  - Tests sur l'espérance d'un échantillon
  - Comparaison de deux échantillons
  - Tests du  $\chi^2$**
  - Tests non paramétriques
- 6 Modèle linéaire

# Rappels - Loi du $\chi^2$

## Définition

Soient  $X_1, \dots, X_n$   $n$  variables aléatoires indépendantes et identiquement distribuées de loi normale centrée réduite.

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

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

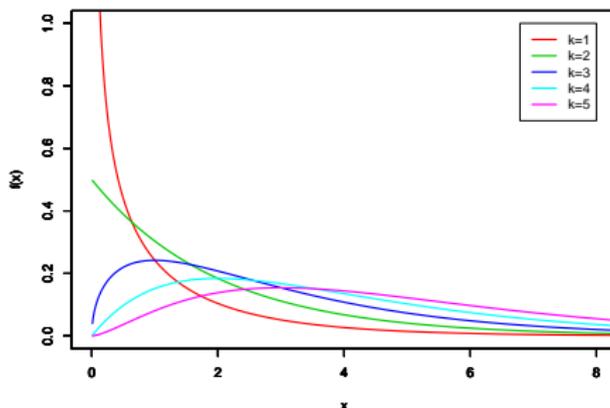
## Propriétés

- 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  $E(Y) = n$  et  $Var(Y) = 2n$

Rappels - Loi du  $\chi^2$ 

## Densité

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



# Test d'adéquation

## Exemple

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

- $X$  est une variable aléatoire discrète à valeurs dans  $E = \{x_1, x_2, \dots, x_k\}$ .
- 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 :

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

# Test d'adéquation

## Remarques

L'approximation de la distribution de  $D_{k,n}$  par  $\chi^2_{(k-1)}$  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  $np_i \geq 5$ )

## Fonction R

chisq.test

# 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 inconnu (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.

# Test d'adéquation à une famille de lois

## Exemple

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%) ?

# Test d'indépendance

## Exemple

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</i> \ <i>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 :

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

# 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
- $\frac{N_{i\bullet} \cdot N_{\bullet j}}{n} \geq 5$  pour tout couple  $(i, j)$

## Fonction R

chisq.test

# Test d'homogénéité

## Exemple

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<br>légère | Réaction<br>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

## Proposition

Un test d'homogénéité peut toujours s'écrire comme un test d'indépendance

## Fonction R

chisq.test

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses**
  - Démarche
  - Tests sur l'espérance d'un échantillon
  - Comparaison de deux échantillons
  - Tests du  $\chi^2$
  - **Tests non paramétriques**
- 6 Modèle linéaire

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire**
  - Analyse de la variance à un facteur - ANOVA1
  - Analyse de la variance à deux facteurs - ANOVA2
  - Régression
  - Vue d'ensemble

# Le modèle linéaire

## Modèle

$$Y = X\theta + \epsilon$$

où  $\epsilon \sim N(0, \sigma^2 I_p)$

## Remarque

Le modèle linéaire est linéaire en ses paramètres (et non en  $X$ )

# Les données

## Exemple 1

Etude de la durée de survie de patients atteints d'un cancer du poumon en fonction de l'âge, du sexe et du statut tabagique

## Exemple 2

Etude du rendement de champs de maïs en fonction du type d'engrais utilisé

## Exemple 3

Etude de la tension artérielle en fonction de l'âge du patient

# Le modèle linéaire

## Objectifs

- Expliquer les variations de  $Y$  en fonction de  $X$
- Prédire les valeurs de  $Y$  à partir des valeurs de  $X$

## Intérêt

- Simplicité des algorithmes d'estimation et des tests
- Utilisable dans la plupart des situations

# Le modèle linéaire

## Types de modèles linéaires

On distingue 3 cas :

- X qualitative (ANOVA)
- X quantitatif (REGRESSION)
- X des 2 types (ANCOVA)

# Théorème de Cochran

## Théorème

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

- $\mu$  est estimé sans biais par la moyenne  $\bar{X}$ ;  $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- $\sigma^2$  est estimé sans biais par la variance empirique corrigée :  

$$S^{*2} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$
;  $Q^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 \sim \sigma^2 \chi_{n-1}^2$
- $Q^2$  et  $\bar{X}$  sont indépendants.

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire**
  - Analyse de la variance à un facteur - ANOVA1
  - Analyse de la variance à deux facteurs - ANOVA2
  - Régression
  - Vue d'ensemble

# Notations

$$X_{i+} = \sum_{j=1}^{n_j} X_{ij}$$

$$X_{+j} = \sum_{i=1}^{n_i} X_{ij}$$

$$X_{++} = \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} X_{ij}$$

$$X_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_j} X_{ij} = \frac{1}{n_i} X_{i+}$$

$$X_{\bullet j} = \frac{1}{n_j} \sum_{i=1}^{n_i} X_{ij} = \frac{1}{n_j} X_{+j}$$

$$X_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^{n_i} \sum_{j=1}^{n_j} X_{ij} = \frac{1}{n} X_{++}$$

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Exemple

- On souhaite comparer trois traitements contre l'asthme : le traitement 2 est un nouveau traitement, que l'on souhaite mettre en compétition avec les traitements classiques 1 et 3.
- On répartit par tirage au sort les patients venant consulter dans un centre de soin et on leur affecte l'un des trois traitements. On mesure sur chaque patient la durée, en jours séparant de la prochaine crise d'asthme.

| Traitement 1           | Traitement 2           | Traitement 3           |
|------------------------|------------------------|------------------------|
| 26 ; 27 ; 35 ; 36 ; 38 | 29 ; 42 ; 44 ; 44 ; 45 | 26 ; 26 ; 30 ; 30 ; 33 |
| 38 ; 41 ; 42 ; 45 ; 50 | 48 ; 48 ; 52 ; 56 ; 56 | 36 ; 38 ; 38 ; 39 ; 46 |
| 65                     | 58 ; 58 ; 60 ; 61 ; 63 | 47 ; 51 ; 51 ; 56 ; 75 |
|                        | 63 ; 69                |                        |

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Problème des tests multiples

- Lors du test d'une seule hypothèse  $H_0$  (au niveau  $\alpha$ ), le risque de commettre une erreur de type I est :

$$\mathbb{P}(P < \alpha | H_0) \leq \alpha$$

- Lors du test de  $n$  hypothèses  $H_0^1; \dots; H_0^n$  (au niveau  $\alpha$  pour chacun des tests), le risque de commettre (au moins) une erreur de type I est :

$$\mathbb{P}\left(\bigcup_{i=1}^n \{P_i < \alpha | H_0^i\}\right) \leq \sum_{i=1}^n \mathbb{P}(P_i < \alpha | H_0^i) = n\alpha$$

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Test global

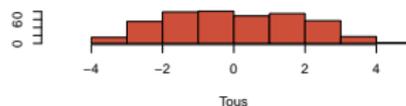
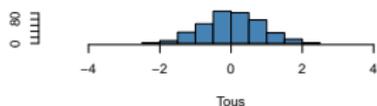
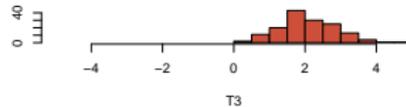
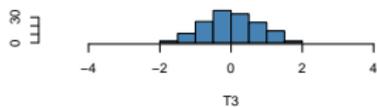
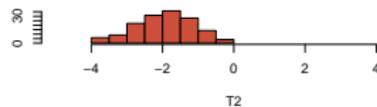
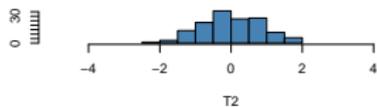
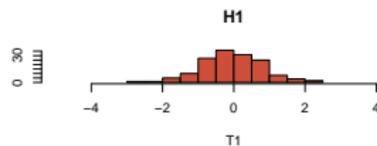
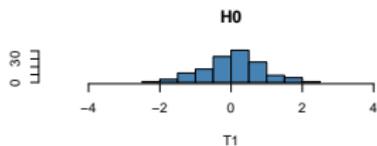
Pour comparer  $k$  espérances, au lieu de comparer toutes les espérances deux à deux (ce qui conduirait à un problème de tests multiples important), on effectue un test global des hypothèses nulle et alternative suivantes :

$$H_0 : \mu_1 = \dots = \mu_k$$

$$H_1 : \exists(i, j) | \mu_i \neq \mu_j$$

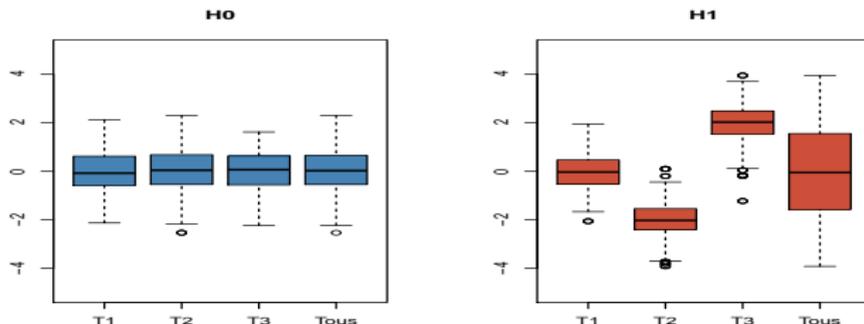
# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Idée générale



# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Idee générale



- 1 Trouver un estimateur de la variance  $\sigma^2$  valable sous  $H_0$  et sous  $H_1$
- 2 Trouver un estimateur de la variance  $\sigma^2$  valable uniquement sous  $H_0$
- 3 Comparer les deux variances estimées

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Modèle

- Observations :  $Y_{ij}$   $i = 1, \dots, k, j = 1, \dots, n_i$
- Modèle :  $Y_{ij}$  indépendantes avec

$$Y_{ij} \sim N(\mu_i, \sigma^2)$$

## Remarque

L'hypothèse d'homogénéité des variances ( $\sigma_1^2 = \dots = \sigma_k^2 = \sigma^2$ ) est appelée hypothèse d'**homoscédasticité**.

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Autre formulation

- Observations :  $Y_{ij}$   $i = 1, \dots, k, j = 1, \dots, n_i$
- Modèle :  $Y_{ij}$  indépendantes avec

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

$$\sum_{i=1}^k n_i \alpha_i = 0$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

## Remarque

- La distribution des  $\epsilon_{ij}$  ne dépend pas du groupe
- $\mu$  s'interprète comme l'effet 'global',  $\alpha_i$  comme l'effet spécifique du groupe  $i$  et  $\epsilon_{ij}$  comme le résidu (bruit gaussien)

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Sommes des carrés

$$\text{Résiduels : } SCR(i) = \sum_{j=1}^{n_i} (Y_{ij} - Y_{i\bullet})^2 \quad \text{avec } SCR(i) \sim \sigma^2 \chi_{n_i-1}^2$$

$$SCR = \sum_{i=1}^k SCR(i) \quad \text{avec } SCR \sim \sigma^2 \chi_{n-k}^2$$

$$\text{Totaux : } SCT = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - Y_{\bullet\bullet})^2 \quad \text{avec } SCT \underset{H_0}{\sim} \sigma^2 \chi_{n-1}^2$$

$$\text{Factoriels : } SCF = \sum_{i=1}^k n_i (Y_{i\bullet} - Y_{\bullet\bullet})^2 \quad \text{avec } SCF \underset{H_0}{\sim} \sigma^2 \chi_{k-1}^2$$

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Carrés moyens

Résiduel :  $CMR = \frac{1}{n - k} SCR$  estimateur sans biais de  $\sigma^2$

Total :  $CMT = \frac{1}{n - 1} SCT$  estimateur sans biais (sous  $H_0$ ) de  $\sigma^2$

Factoriel :  $CMF = \frac{1}{k - 1} SCF$  estimateur sans biais (sous  $H_0$ ) de  $\sigma^2$

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

Théorème fondamental (décomposition de la variance)

En utilisant les notations introduites précédemment, on a :

1

$$SCT = SCF + SCR$$

2

$$SCF \perp\!\!\!\perp SCR$$

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

Statistique de test

$$F = \frac{SCF/(k-1)}{SCR/(n-k)} = \frac{CMF}{CMR} \underset{H_0}{\sim} \mathcal{F}(k-1, n-k)$$

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Présentation des résultats

| Source  | Degrés de liberté | Sommes des carrés | Carrés moyens         | F           |
|---------|-------------------|-------------------|-----------------------|-------------|
| Facteur | $k - 1$           | SCF               | $CMF = SCF / (k - 1)$ | $CMF / CMR$ |
| Résidus | $n - k$           | SCR               | $CMR = SCR / (n - k)$ |             |
| Total   | $n - 1$           | SCT               |                       |             |

# Comparaison des espérances de $k$ échantillons ( $k > 2$ )

## Exemple

| Source  | Degrés de liberté | Sommes des carrés | Carrés moyens | F     |
|---------|-------------------|-------------------|---------------|-------|
| Facteur | 2                 | 1426.84           | 713.42        | 5.467 |
| Résidus | 40                | 5219.44           | 130.49        |       |
| Total   | 42                | 6646.28           |               |       |

## Fonctions R

```
anova(lm(Y ~ as.factor(X)))
```

# Comparaison de modèles emboîtés

## Modèles emboîtés

Deux modèles sont dits **emboîtés** si l'un peut être vu comme un cas particulier de l'autre.

## Exemple

Dans l'ANOVA1, on compare les deux modèles suivants :

- $\mathcal{M}_1$ , le modèle sous  $H_0 : Y = \mu + \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  (1 paramètre)
- $\mathcal{M}_k$ , le modèle sous  $H_1 : Y = \mu_i + \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  ( $k$  paramètres)

où  $\mathcal{M}_1$  est emboîté dans  $\mathcal{M}_k$

# Comparaison de modèles emboîtés

## Ligne associée à un modèle

A chaque modèle peut correspondre une ligne d'un tableau de synthèse :

|                      |       |                    |                   |
|----------------------|-------|--------------------|-------------------|
| modèle $\mathcal{M}$ | $ddl$ | $SCR(\mathcal{M})$ | $CM(\mathcal{M})$ |
|----------------------|-------|--------------------|-------------------|

## Remarque

Tous les tests de choix de modèle dans le contexte du modèle linéaire reviennent à comparer deux telles lignes.

# Comparaison de modèles emboîtés

## Tableau de synthèse

Soit un modèle  $\mathcal{M}_1$  emboîté dans un modèle  $\mathcal{M}_2$  (i.e.  $SC_1 > SC_2$ ) :

| Source                          | Degrés de liberté | Sommes des carrés  | Carrés moyens       | F         |
|---------------------------------|-------------------|--------------------|---------------------|-----------|
| $\mathcal{M}_1 - \mathcal{M}_2$ | $n_1 - n_2$       | $SC = SC_1 - SC_2$ | $CM = SC/n$         | $CM/CM_1$ |
| $\mathcal{M}_2$                 | $n_2$             | $SC_2$             | $CM_2 = SC_2/(n_2)$ |           |
| $\mathcal{M}_1$                 | $n_1$             | $SC_1$             | $CM_1 = SC_1/(n_1)$ |           |

## Question

Une augmentation du nombre de paramètres du modèle (i.e. de sa complexité) fait-elle significativement diminuer la variance estimée ?

# Comparaison de modèles emboîtés

## Cas de l'ANOVA1

On compare le modèle à 1 paramètre  $\mathcal{M}_1$  au modèle à  $k$  paramètres  $\mathcal{M}_k$  (avec  $\mathcal{M}_1$  emboîté dans  $\mathcal{M}_k$ ).

Par différence on obtient une ligne mesurant le gain apporté par  $\mathcal{M}_k$  en comparaison de  $\mathcal{M}_1$  :

| Gain de $\mathcal{M}_k/\mathcal{M}_1$ | $ddl_1 - ddl_k$ | $SCR(\mathcal{M}_1) - SCR(\mathcal{M}_k)$ | $\frac{SCR(\mathcal{M}_1) - SCR(\mathcal{M}_k)}{ddl_1 - ddl_k}$ |
|---------------------------------------|-----------------|---|---|
| modèle $\mathcal{M}_k$                | $ddl_k$         | $SCR(\mathcal{M}_k)$                      | $CM(\mathcal{M}_k)$   |
| modèle $\mathcal{M}_1$                | $ddl_1$         | $SCR(\mathcal{M}_1)$                      | $CM(\mathcal{M}_1)$   |

## Remarque

La ligne 'gain' correspond à la ligne associée au facteur dans le tableau de l'analyse de la variance.

## Ligne associée à un modèle

### Exemple

Dans le cas des données relatives aux traitements de l'asthme, la vraie question était : 'le nouveau traitement 2 est-il supérieur aux deux traitements classiques 1 et 3?'

### Fonctions R

```
anova(reslm1, reslm2)
```

# Test de Student : un cas particulier de l'ANOVA

## Equivalence entre test de Student et ANOVA

- Test de Student :

$$\Gamma = \left\{ T = \frac{|Y_{1\bullet} - Y_{2\bullet}|}{\sqrt{CMR\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > q_{t_{n_1+n_2-2}; 1-\alpha/2} \right\}.$$

- ANOVA :

$$\Gamma = \left\{ F = \frac{CMF}{CMR} > q_{\{\mathcal{F}_{1;n_1+n_2-2}; 1-\alpha\}} \right\}.$$

## Remarque

$T^2 = F \Rightarrow$  Les deux régions de réjets précédentes sont identiques

# Validité du modèle

## Normalité

- Test de Stephens
- Test de Shapiro-Wilk

## Homoscédasticité

- Test de Bartlett
- Test de Levene

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire**
  - Analyse de la variance à un facteur - ANOVA1
  - Analyse de la variance à deux facteurs - ANOVA2**
  - Régression
  - Vue d'ensemble

# Exemples

## Exemple 1

On veut mesurer l'effet de trois traitements, U, V, W contre les arthralgies chroniques. On sélectionne 9 patients que l'on soumet successivement à chacun des traitements, en laissant une période jugée assez longue entre deux traitements pour que l'on puisse considérer qu'il n'y a pas d'effet résiduel d'un traitement sur le suivant. On mesure, selon une échelle adéquate, l'amélioration apportée dans l'arthralgie, que l'on reporte dans le tableau qui suit :

|   | P1   | P2   | P3   | P4   | P5   | P6   | P7   | P8   | P9   |
|---|------|------|------|------|------|------|------|------|------|
| U | 5.20 | 6.10 | 7.40 | 6.80 | 5.80 | 4.70 | 5.40 | 5.40 | 4.70 |
| V | 7.70 | 7.00 | 7.50 | 8.10 | 5.70 | 7.30 | 7.10 | 6.30 | 6.40 |
| W | 7.40 | 7.30 | 8.20 | 7.80 | 6.30 | 6.30 | 5.90 | 7.00 | 6.60 |

# Exemples

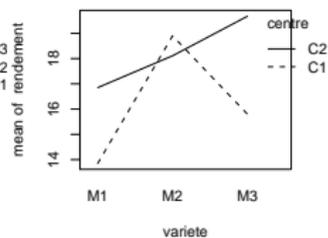
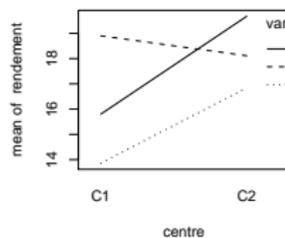
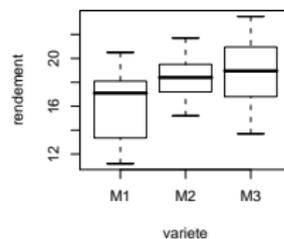
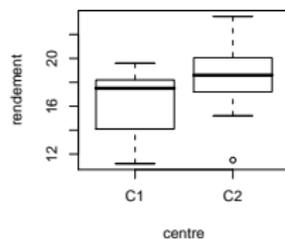
## Exemple 2

On mesure le rendement de trois variétés de maïs M1, M2, et M3, dans deux centres expérimentaux, C1 et C2 de l'INRA. Le tableau suivant résume les données recueillies :

|           | Variété M1 |      | Variété M2 |      |      | Variété M3 |      |      |      |
|-----------|------------|------|------------|------|------|------------|------|------|------|
| Centre C1 | 11.2       | 16.5 | 19.6       | 18.2 | 18.9 | 13.7       | 17.9 | 17.5 | 14.1 |
|           | 20.5       | 18.1 | 17.3       | 21.7 | 15.2 | 23.5       | 23.3 | 19.3 | 19.6 |
| Centre C2 | 17.7       | 18.1 | 19.4       | 18.6 | 16.5 | 17.3       | 15.5 | 21.3 | 19.4 |
|           | 15.2       | 11.5 | 19.6       | 17.1 | 17.6 | 21.4       | 20.6 | 16.3 | 18.6 |

# Représentations graphiques

## Exemple 2



# Cas général

## Les données

|       |          |     |          |          |
|-------|----------|-----|----------|----------|
|       | $B_1$    | ... | $B_b$    |          |
| $A_1$ | $n_{11}$ | ... | $n_{1b}$ | $n_{1+}$ |
| ...   | ...      | ... | ...      | ...      |
| $A_a$ | $n_{a1}$ | ... | $n_{ab}$ | $n_{a+}$ |
|       | $n_{+1}$ | ... | $n_{+b}$ | $n_{++}$ |

## Questions d'intérêt

- Y a-t-il un effet lié au facteur A ?
- Y a-t-il un effet lié au facteur B ?
- Y a-t-il une interaction entre les deux facteurs ?

# Plans d'expériences

## Définitions

- La **matrice d'incidence** indique le nombre d'observations dans chaque case du tableau de données
- Un plan d'expérience est dit **complet** (ou factoriel d'ordre 2) si l'on dispose d'au moins une mesure pour toutes les combinaisons possibles
- Un plan d'expérience est dit **avec répétitions** si on a au moins deux expériences par combinaison
- Un plan est dit **équilibré** si le nombre d'expériences pour chaque combinaison  $ij$  vérifie : 
$$n_{ij} = \frac{n_i + n_j}{n}$$
- Un plan est dit **équirépété** si on fait le même nombre d'expériences pour chaque combinaison  $ij$

# Modèles à deux facteurs

## Modèle général

Notons  $jk$  la combinaison des facteurs  $A$  et  $B$  aux niveaux  $j$  et  $k$ .

$$Y_{ijk} = \mu_{jk} + \epsilon_{ijk}$$

- $\mu_{jk}$  est l'espérance de la combinaison  $jk$  (effet fixe ou effet principal)
- $\epsilon_{ijk}$  représente un aléa individuel.

## Hypothèses

Les  $\epsilon_{ijk}$  sont supposés indépendants et gaussiens de même variance  $\sigma^2$  et d'espérance nulle.

# Modèles à deux facteurs

## Modèle général reparamétré

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}$$

- $\mu_{j\bullet} = \frac{1}{n_{j+}} \sum_k n_{jk} \mu_{jk}$  est l'effet moyen pour  $A = j$ ,
- $\mu_{\bullet k} = \frac{1}{n_{k+}} \sum_j n_{jk} \mu_{jk}$  est l'effet moyen pour  $B = k$ ,
- $\mu = \frac{1}{n} \sum_{jk} n_{jk} \mu_{jk}$  est l'effet moyen des deux facteurs,
- $\alpha_j = \mu_{j\bullet} - \mu$  représente l'effet du facteur  $A$  lorsqu'il est au niveau  $j$ ,
- $\beta_k = \mu_{\bullet k} - \mu$  représente l'effet du facteur  $B$  lorsqu'il est au niveau  $k$ ,
- $\gamma_{jk} = \mu_{jk} - \alpha_j - \beta_k$  représente l'effet supplémentaire dû à l'interaction des niveaux  $j$  et  $k$  des facteurs  $A$  et  $B$ .

# Modèles à deux facteurs

## Exemple 2

Dans l'exemple des cultures INRA, nous testons le modèle suivant :

$$\textit{rendement} = \textit{centre} + \textit{variete} + \textit{centre} * \textit{variete}$$

# Modèles à deux facteurs

## Modèle additif

$$Y_{ijk} = \mu + \alpha_j + \beta_k + \epsilon_{ijk}$$

## Remarques

Ce modèle est utilisé lorsque l'interaction peut être négligée ou lorsque l'estimation de l'interaction ne permet pas les tests (une seule répétition par combinaison d'effet).

# Modèles à deux facteurs

## Exemple 1

Dans l'exemple de l'arthralgie, l'interaction entre les facteurs ne peut pas être estimée (car une seule répétition par combinaison de facteur). Le modèle choisi est donc un modèle simplifié :

$$\textit{amelioration} = \textit{traitement} + \textit{patient}$$

# Sommes de carrés

## Somme des carrés totaux

$$SCT = \sum_{ijk} (Y_{ijk} - Y_{\dots})^2$$

$$= \sum_{ijk} Y_{ijk}^2 - nY_{\dots}^2$$

$$\text{avec } SCT \underset{H_{0AB}}{\sim} \sigma^2 \chi_{n-1}^2$$

# Sommes de carrés

## Somme des carrés liés aux facteurs

- Somme des carrés liés au facteur  $A$  :

$$\begin{aligned}
 SCA &= \sum_{j=1}^a n_{j+} (Y_{\bullet j \bullet} - Y_{\bullet \bullet \bullet})^2 \\
 &= \sum_{j=1}^a n_{j+} Y_{\bullet j \bullet}^2 - n Y_{\bullet \bullet \bullet}^2
 \end{aligned}
 \quad \text{avec } SCA \underset{H_{0A}}{\sim} \sigma^2 \chi_{a-1}^2$$

- Somme des carrés liés au facteur  $B$  :

$$\begin{aligned}
 SCB &= \sum_{k=1}^b n_{+k} (Y_{\bullet \bullet k} - Y_{\bullet \bullet \bullet})^2 \\
 &= \sum_{k=1}^b n_{+k} Y_{\bullet \bullet k}^2 - n Y_{\bullet \bullet \bullet}^2
 \end{aligned}
 \quad \text{avec } SCAB \underset{H_{0B}}{\sim} \sigma^2 \chi_{b-1}^2$$

# Sommes de carrés

Somme des carrés liés à l'interaction des facteurs

$$SCAB = \sum_{jk} n_{jk} (Y_{\bullet jk} - Y_{\bullet j\bullet} - Y_{\bullet\bullet k} + Y_{\bullet\bullet\bullet})^2$$

$$\text{avec } SCB \underset{H_{0AB}}{\sim} \sigma^2 \chi_{(a-1)(b-1)}^2$$

# Sommes de carrés

## Somme des carrés résiduels

- Modèle général :

$$SCR = \sum_{ijk} (Y_{ijk} - Y_{jk\bullet})^2$$

$$SCR \sim \sigma^2 \chi_{n-ab}^2$$

- Modèle additif (sans interaction) :

$$SCR = SCT - SCA - SCB$$

$$SCR \sim \chi_{n-a-b+1}^2$$

# Analyse de la variance

Théorème (décomposition de la variance)

$$SCT = SCR + SCA + SCB + SCAB - 2 \sum_{jk} n_{jk} (Y_{j\bullet} - Y_{\bullet\bullet})(Y_{\bullet l} - Y_{\bullet\bullet})$$

# Analyse de la variance

## Proposition

Lorsque le plan d'expérience est orthogonal, la formule de décomposition de la variance s'écrit :

$$SCT = SCR + SCA + SCB + SCAB$$

# Test de l'absence d'interaction

## Hypothèses testées

$$\begin{cases} H_{0AB} : \gamma_{jk} = 0, \forall j, l \\ H_{1AB} : \overline{H_{0AB}} \end{cases}$$

## Statistique de test

$$F_{AB} = \frac{(n - ab)SCAB}{((a - 1)(b - 1))SCR} \underset{H_{0AB}}{\sim} \mathcal{F}_{(a-1)(b-1), (n-ab)}$$

# Test de l'effet d'un facteur

## Hypothèses testées (facteur A)

$$\begin{cases} H_{0A} : \alpha_j = 0, \forall j \\ H_{1A} : \overline{H_{0A}} \end{cases}$$

## Statistique de test

- Modèle général :

$$F_A = \frac{(n - ab)SCA}{(a - 1)SCR} \underset{H_{0A}}{\sim} \mathcal{F}_{(a-1), (n-ab)}$$

- Modèle additif :

$$F_A = \frac{(n - a - b + 1)SCA}{(a - 1)SCR} \underset{H_{0A}}{\sim} \mathcal{F}_{(a-1), (n-a-b+1)}$$

# Tableau de synthèse

## Modèle général

| Source      | Degrés de liberté | Sommes des carrés | Carrés moyens             | F   |
|-------------|-------------------|-------------------|---------------------------|---|
| Facteur A   | a-1               | SCA               | $\frac{SCA}{a-1}$         | $F_A = \frac{(n-ab)SCA}{(a-1)SCR}$          |
| Facteur B   | b-1               | SCB               | $\frac{SCB}{b-1}$         | $F_B = \frac{(n-ab)SCB}{(b-1)SCR}$          |
| Interaction | (a-1)(b-1)        | SCAB              | $\frac{SCAB}{(a-1)(b-1)}$ | $F_{AB} = \frac{(n-ab)SCAB}{(a-1)(b-1)SCR}$ |
| Résiduelle  | n-ab              | SCR               | $\frac{SCR}{(n-ab)}$      |   |
| Total       | n-1               | SCT               |                           |   |

# Tableau de synthèse

## Modèle additif

| Source     | Degrés de liberté | Sommes des carrés | Carrés moyens           | F                                     |
|------------|-------------------|-------------------|-------------------------|---------------------------------------|
| Facteur A  | a-1               | SCA               | $\frac{SCA}{a-1}$       | $F_A = \frac{(n-a-b+1)SCA}{(a-1)SCR}$ |
| Facteur B  | b-1               | SCB               | $\frac{SCB}{b-1}$       | $F_B = \frac{(n-a-b+1)SCB}{(b-1)SCR}$ |
| Résiduelle | n-a-b+1           | SCR               | $\frac{SCR}{(n-a-b+1)}$ |                                       |
| Total      | n-1               | SCT               |                         |                                       |

# Tableau de synthèse

## Exemple 2

```
> anova(lm(rendement~centre*variete))
```

|                | Df | Sum Sq  | Mean Sq | F value | Pr(>F)  |   |
|----------------|----|---------|---------|---------|---------|---|
| centre         | 1  | 30.507  | 30.507  | 5.1350  | 0.03082 | * |
| variete        | 2  | 38.000  | 19.000  | 3.1981  | 0.05508 | . |
| centre:variete | 2  | 29.440  | 14.720  | 2.4777  | 0.10095 |   |
| Residuals      | 30 | 178.231 | 5.941   |         |         |   |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Tableau de synthèse

## Exemple 1

```
> anova(lm(amelioration~ patient+traitement))
```

|            | Df | Sum Sq  | Mean Sq | F value | Pr(>F)    |     |
|------------|----|---------|---------|---------|-----------|-----|
| patient    | 8  | 11.3807 | 1.4226  | 4.710   | 0.0041026 | **  |
| traitement | 2  | 9.5141  | 4.7570  | 15.750  | 0.0001657 | *** |
| Residuals  | 16 | 4.8326  | 0.3020  |         |           |     |

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Choix de modèle

## Interaction significative

- Si l'interaction est significative, les deux facteurs sont pertinents pour expliquer la variance. Les deux doivent être conservés.
- Les test sur les effets principaux (A et B) sont utiles pour l'interprétation mais ne permettent pas de conclure qu'un effet est absent puisque chaque effet est présent dans l'interaction.

## Interaction non significative

- Si l'interaction est non significative, il est possible de considérer un modèle additif. On préfère cependant garder l'estimation de la variance du modèle complet ( $SCR/(n - ab)$ ).
- Le modèle additif n'est donc utilisé que lorsque les tests sur l'interaction ne sont pas calculables.

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire**
  - Analyse de la variance à un facteur - ANOVA1
  - Analyse de la variance à deux facteurs - ANOVA2
  - **Régression**
  - Vue d'ensemble

# Régression linéaire simple

## Exemples

- Tension artérielle =  $f(\text{age})$
- Rendement de blé =  $f(\text{dose de fertilisant})$
- Concentration ozone =  $f(\text{température})$
- Effet d'un traitement =  $f(\text{dose})$
- Taux de DDT =  $f(\text{age du brochet})$

# Régression linéaire simple

## Le modèle

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

### Hypothèses :

- $E(\epsilon_i) = 0$
- $V(\epsilon_i) = \sigma^2$  (constante)
- $\epsilon_i \perp\!\!\!\perp \epsilon_j; \forall i \neq j$
- $\epsilon_i \sim N(0, \sigma^2)$  (hypothèse nécessaire pour les tests)

## Remarque

Le modèle est linéaire en ses paramètres (pas nécessairement en  $X$ )

# Régression linéaire simple

## Vocabulaire

- $X$  est une variable, aléatoire ou contrôlée, dite **explicative**
- $Y$  est une variable aléatoire dite **à expliquer**

## Remarque

Si  $X$  n'est pas aléatoire,  $cov(X, Y)$  peut être calculé mais n'a pas de sens

# Régression linéaire simple

## Questions d'intérêt

- Existe-t-il une relation entre  $X$  et  $Y$  ?
- Quelle est la forme de la relation ?
- Peut-on prédire  $Y$  à partir des valeurs de  $X$  ?

## Démarche

- Estimation des paramètres du modèle ( $\beta_0, \beta_1$ , et  $\sigma^2$ )
- Tests (paramètres / validité du modèle)
- Réalisation de prédictions

# Estimation des paramètres

## Estimateurs des moindres carrés

Les **estimateurs des moindres carrés ordinaires** (MCO) de  $\beta_0$  et  $\beta_1$  sont ceux qui minimisent la somme des carrés des résidus :

$$SCR = \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2$$

# Estimation des paramètres

## Proposition

Les estimateurs des moindres carrés ordinaires ont pour expressions :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

## Remarques

- Si les résidus suivent une distribution normale, les estimateurs des moindres carrés sont les mêmes que les estimateurs du maximum de vraisemblance
- La droite des moindres carrés passe par le barycentre du nuage de points  $(\bar{X}, \bar{Y})$

# Estimation des paramètres

## Propriétés

- $\hat{\beta}_0$  et  $\hat{\beta}_1$  sont des estimateurs sans biais de  $\beta_0$  et  $\beta_1$
- Les variances des estimateurs sont :

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- Leur covariance est :  $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- Si les résidus suivent une distribution normale,  $\hat{\beta}_0$  et  $\hat{\beta}_1$  suivent aussi une distribution normale

# Estimation des paramètres

## Résidus estimés

Les résidus estimés sont :

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i$$

où  $\hat{Y}_i$  est la valeur prédite par le modèle :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

## Propriété

$$\sum \hat{\epsilon}_i = 0$$

Remarque : Contrairement aux résidus  $\epsilon_i$ , les résidus estimés  $\hat{\epsilon}_i$  ne sont pas indépendants

# Estimation de la variance résiduelle $\sigma^2$

## Théorème

- La variance  $\sigma^2$  est estimée sans biais par :

$$s^{*2} = \frac{1}{n-2} \sum_{i=1}^n \hat{\epsilon}_i^2$$

- Sous l'hypothèse que les résidus sont normalement distribués :

$$(n-2)s^{*2} \sim \sigma^2 \chi_{n-2}^2$$

# Tests sur les paramètres du modèle

On suppose :  $\epsilon_j \sim N(0, \sigma^2)$

Hypothèses testées

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Statistique de test

$$\frac{\hat{\beta}_1}{\sqrt{\frac{s^{*2}}{\sum_{i=1}^n (x_i - \bar{x})^2}}} \underset{H_0}{\sim} t_{n-2}$$

# Tests sur les paramètres du modèle

On suppose :  $\epsilon_j \sim N(0, \sigma^2)$

Hypothèses testées

$$\begin{cases} H_0 : \beta_0 = 0 \\ H_1 : \beta_0 \neq 0 \end{cases}$$

Statistique de test

$$\frac{\hat{\beta}_0}{\sqrt{s^{*2} \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}} \underset{H_0}{\sim} t_{n-2}$$

# Prédiction

## Valeur prédite

Soit  $x_{n+1}$  une nouvelle observation. La valeur prédite par le modèle est :  
 $Y_{n+1} = \beta_0 + \beta_1 X_{n+1} + \epsilon_{n+1}$ . Cette valeur peut être approchée par :

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 X_{n+1}$$

## Remarque

Deux types d'erreurs entâchent cette prédiction :

- La non connaissance de  $\epsilon_{n+1}$
- L'incertitude sur l'estimation des paramètres  $\beta_0$  et  $\beta_1$

# Prédiction

## Proposition

Soit  $\hat{\epsilon}_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$  l'erreur de prévision. On a :

$$E(\hat{\epsilon}_{n+1}) = 0$$

$$\text{Var}(\hat{\epsilon}_{n+1}) = \sigma^2 \left( 1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

## Remarques

- La variance de l'erreur de prédiction est d'autant plus grande que  $x_{n+1}$  est éloigné de la moyenne  $\bar{x}$

# Prédiction

## Intervalle de confiance de la prédiction

$$IC_{1-\alpha}(y_{n+1}) = \left[ \hat{Y}_{n+1} \pm q_{t_{n-2}, 1-\frac{\alpha}{2}} \sqrt{s^{*2} \left( 1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)} \right]$$

## Remarque

Cet intervalle de confiance n'est valable que si les résidus suivent une distribution normale :

# Analyse de la variance

## Interprétation géométrique

# Analyse de la variance

## Théorème fondamental (Pythagore)

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$SCT$ 
 $SCR$ 
 $SCM$

## Sommes de carrés

- $SCR = (n - 2)s^{*2} \sim \sigma^2 \chi_{n-2}^2$
- Sous l'hypothèse  $\{H_0 : \beta_1 = 0\}$  :  $SCT \underset{H_0}{\sim} \sigma^2 \chi_{n-1}^2$
- Sous l'hypothèse  $\{H_0 : \beta_1 = 0\}$  :  $SCM \underset{H_0}{\sim} \sigma^2 \chi_1^2$

## Remarque

D'après le théorème de Cochran :  $SCR \perp\!\!\!\perp SCM$

# Analyse de la variance

## Coefficient d'ajustement

Le coefficient d'ajustement est défini par :

$$R^2 = \frac{SCM}{SCT}$$

## Remarque

Le coefficient d'ajustement peut être interprété comme le pourcentage de variance expliquée par le modèle

# Analyse de la variance

## Tableau de synthèse

| Source     | Degrés de liberté | Sommes des carrés | Carrés moyens       | F                          |
|------------|-------------------|-------------------|---------------------|----------------------------|
| Modèle     | 1                 | SCM               | $SCM$               | $F = \frac{(n-2)SCM}{SCR}$ |
| Résiduelle | n-2               | SCR               | $\frac{SCR}{(n-2)}$ |                            |
| Total      | n-1               | SCT               |                     |                            |

# Test de linéarité

Dans le cas de mesures répétées (plusieurs valeurs de  $y$  pour chaque valeur de  $x$ ), il est possible de tester l'hypothèse de linéarité.

# Validité du modèle

Examen visuel des résidus

# Régression linéaire multiples

## Exemples

- Concentration d'ozone= $f(\text{température, vent, nébulosité})$
- Vitesse de circulation coronarienne= $f(\text{poids, taux de cholestérol})$

# Régression linéaire multiple

## Le modèle

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_k X_{ki} + \epsilon_i$$

Hypothèses :

- $E(\epsilon_i) = 0$
- $V(\epsilon_i) = \sigma^2$  (constante)
- $\epsilon_i \perp \epsilon_j; \forall i \neq j$
- $\epsilon_i \sim N(0, \sigma^2)$  (hypothèse nécessaire pour les tests)

# Régression linéaire multiple

## Le modèle - écriture matricielle

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

$$\begin{pmatrix} Y_1 \\ \cdot \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \cdot \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Hypothèses :

- $E(\epsilon) = \mathbf{0}$
- $Var(\epsilon) = \sigma^2 \mathbf{I}_n$
- $\epsilon \sim N(0, \sigma^2 \mathbf{I}_n)$  (pour les tests)

# Estimation des paramètres

## Estimateurs des moindres carrés

L'**estimateurs des moindres carrés ordinaires** (MCO) de  $\beta$  est le vecteur aléatoire de  $\mathbb{R}^p$   $\hat{\beta}$  qui minimise la somme des carrés des résidus :

$$SCR(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2$$

## Théorème

Si la matrice  $\mathbf{X}'\mathbf{X}$  est inversible, l'estimateur des moindres carrés ordinaires de  $\beta$  est :

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

# Estimation des paramètres

## Propriétés

- L'estimateur  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$  est un estimateur sans biais de  $\beta$
- La variance de  $\hat{\beta}$  est :  $V(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$
- $\sigma^2$  est estimé sans biais par :

$$\hat{\sigma}^2 = CMR = \frac{SCR(\beta)}{n - (k + 1)} = \frac{\|\mathbf{Y} - \mathbf{X}\beta\|^2}{n - (k + 1)}$$

avec

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n - (k + 1)} \chi_{n-(k+1)}^2$$

# Analyse de la variance

## Notations

- Soit  $H_p$  un modèle à  $p$  paramètres.  $SCR(H_p)$  est la somme des carrés résiduels associée.
- Soit  $H_q$  un modèle à  $q$  paramètres emboîté dans  $H_p$  (donc  $q < p$ ).  $SCR(H_q)$  est la somme des carrés résiduels associée.
- Soit  $SCE = SCR(H_q) - SCR(H_p)$  la partie de la  $SCR$  expliquée par le passage du petit modèle  $H_q$  au grand modèle  $H_p$ .
- Soit  $n$  le nombre total de mesures.

# Comparaison de modèles emboîtés

## Théorème

Si les résidus  $\epsilon_i$  sont indépendants de loi  $N(0, \sigma^2)$ , alors sous  $H_q$  :

- $SCR(H_q) \sim \sigma^2 \chi_{n-q}^2$
- $SCR(H_p) \sim \sigma^2 \chi_{n-p}^2$
- $SCE \sim \sigma^2 \chi_{p-q}^2$
- $SCE \perp\!\!\!\perp SCR(H_p)$

# Comparaison de modèles emboîtés

## Tableau de synthèse

| Source         | Degrés de liberté | Sommes des carrés | Carrés moyens                    | F                     |
|----------------|-------------------|-------------------|----------------------------------|-----------------------|
| Gain $H_p/H_q$ | $p - q$           | $SCE$             | $CME = \frac{SCE}{p - q}$        | $F = \frac{CME}{CMR}$ |
| $H_p$          | $n - p$           | $SCR(H_p)$        | $CMR = \frac{SCR(H_p)}{(n - p)}$ |                       |
| $H_q$          | $n - q$           | $SCR(H_q)$        |                                  |                       |

- 1 Introduction
- 2 Statistique descriptive
- 3 Rappels de probabilités
- 4 Estimation
- 5 Tests d'hypothèses
- 6 Modèle linéaire**
  - Analyse de la variance à un facteur - ANOVA1
  - Analyse de la variance à deux facteurs - ANOVA2
  - Régression
  - **Vue d'ensemble**

# Vue d'ensemble

## Le modèle - écriture matricielle

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

### ANOVA

$\mathbf{X}$  qualitative :

$$\begin{pmatrix} 1 & 1_{A_1}(1) & \dots & 1_{A_k}(1) \\ 1 & 1_{A_1}(2) & \dots & 1_{A_k}(2) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1_{A_1}(n) & \dots & 1_{A_k}(n) \end{pmatrix}$$

### Régression

$\mathbf{X}$  quantitative :

$$\begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix}$$