Probabilités et Statistiques 2

M. BOUTAHAR

1. Variables aléatoires

Définitions

Soit Ω un espace (ou ensemble), soit $\mathcal{F}$ une famille de sous-ensembles de Ω.

Definition

On dit que $\mathcal{F}$ est une **algèbre** sur Ω si les conditions suivantes sont satisfaites:
i. $\emptyset$ et Ω sont dans $\mathcal{F}$,
ii. si A est dans $\mathcal{F}$, alors A^c est dans $\mathcal{F}$, où A^c est le complémentaire de A dans Ω.
iii. si A et B sont dans $\mathcal{F}$, alors $A \bigcup B$ est dans $\mathcal{F}$.

Exemple.

$\Omega = \{1, 2, 3\}$,

$\mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$
est une algèbre.

Definition

On dit que $\mathcal{F}$ est une **tribu** (ou σ-**algèbre**) sur Ω si les conditions de la définition précédente sont satisfaites avec iii) remplacée par:
iii'. si $(A_n, n \geq 1)$ est dans $\mathcal{F}$, alors $\bigcup_{n=1}^{\infty} A_n$ est dans $\mathcal{F}$.

Exemple.

$\Omega = \{1, 2, 3\}$,

$\mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$

est une algèbre.

Definition

On dit que $\mathcal{F}$ est une **tribu** (ou σ-**algèbre**) sur Ω si les conditions de la définition précédente sont satisfaites avec iii) remplacée par:

iii'. si $(A_n, n \geq 1)$ est dans $\mathcal{F}$, alors $\bigcup_{n=1}^{\infty} A_n$ est dans $\mathcal{F}$.

Espace de probabilité

Definition

On appelle **loi de probabilité** sur $(\Omega, \mathcal{F})$, toute application $P : \Omega \to [0, 1]$, qui vérifie:

i. $P(\Omega) = 1$,

ii. $\forall (A_n, n \geq 1) \subset \mathcal{F}$ telle que $A_k \bigcap A_l = \emptyset$ si $k \neq l$,

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

Espace de probabilité

Definition

On appelle **loi de probabilité** sur $(\Omega, \mathcal{F})$, toute application $P : \Omega \to [0,1]$, qui vérifie:

i. $P(\Omega) = 1$,

ii. $\forall (A_n, n \geq 1) \subset \mathcal{F}$ telle que $A_k \bigcap A_l = \emptyset$ si $k \neq l$,

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

1.Loi de probabilité discrète

Espace de probabilité

Definition

On appelle **loi de probabilité** sur $(\Omega, \mathcal{F})$, toute application $P : \Omega \to [0,1]$, qui vérifie:

i. $P(\Omega) = 1$,

ii. $\forall (A_n, n \geq 1) \subset \mathcal{F}$ telle que $A_k \bigcap A_l = \emptyset$ si $k \neq l$,

$$P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n).$$

1.Loi de probabilité discrète

Definition

$(\Omega, \mathcal{F}, P)$ est appelé **espace de probabilité** si:

- Ω est un espace,
- $\mathcal{F}$ est une tribu,
- P est une **loi de probabilité**

Variable aléatoire

Definition

Une **variable aléatoire** réelle X est une application de $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ **mesurable** : c'est à dire pour tout ouvert $C \subset \mathbb{R}$, on a $X^{-1}(C) = \{\omega, X(\omega) \in C\} \in \mathcal{F}$.($\mathcal{B}$ est la tribu borélienne contenant les ouverts de $\mathbb{R}$).

Definition

$(\Omega, \mathcal{F}, P)$ est appelé **espace de probabilité** si:

- Ω est un espace,
- $\mathcal{F}$ est une tribu,
- P est une **loi de probabilité**

Variable aléatoire

Definition

Une **variable aléatoire** réelle X est une application de $(\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B})$ **mesurable** : c'est à dire pour tout ouvert $C \subset \mathbb{R}$, on a $X^{-1}(C) = \{\omega, X(\omega) \in C\} \in \mathcal{F}$.($\mathcal{B}$ est la tribu borélienne contenant les ouverts de $\mathbb{R}$).

1.Loi de probabilité discrète

Definition

$(\Omega, \mathcal{F}, P)$ est appelé **espace de probabilité** si:

- Ω est un espace,
- $\mathcal{F}$ est une tribu,
- P est une **loi de probabilité**

Variable aléatoire

Definition

Une **variable aléatoire** réelle X est une application de $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ **mesurable** : c'est à dire pour tout ouvert $C \subset \mathbb{R}$, on a $X^{-1}(C) = \{\omega, X(\omega) \in C\} \in \mathcal{F}$.($\mathcal{B}$ est la tribu borélienne contenant les ouverts de $\mathbb{R}$).

Definition

$(\Omega, \mathcal{F}, P)$ est appelé **espace de probabilité** si:
- Ω est un espace,
- $\mathcal{F}$ est une tribu,
- P est une **loi de probabilité**

Variable aléatoire

Definition

Une **variable aléatoire** réelle X est une application de $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ **mesurable** : c'est à dire pour tout ouvert $C \subset \mathbb{R}$, on a $X^{-1}(C) = \{\omega, X(\omega) \in C\} \in \mathcal{F}$.($\mathcal{B}$ est la tribu borélienne contenant les ouverts de $\mathbb{R}$).

Definition

$(\Omega, \mathcal{F}, P)$ est appelé **espace de probabilité** si:
- Ω est un espace,
- $\mathcal{F}$ est une tribu,
- P est une **loi de probabilité**

Variable aléatoire

Definition

Une **variable aléatoire** réelle X est une application de $(\Omega, \mathcal{F}) \rightarrow (\mathbb{R}, \mathcal{B})$ **mesurable** : c'est à dire pour tout ouvert $C \subset \mathbb{R}$, on a $X^{-1}(C) = \{\omega, X(\omega) \in C\} \in \mathcal{F}$.($\mathcal{B}$ est la tribu borélienne contenant les ouverts de $\mathbb{R}$).

Definition

Soit X une variable aléatoire à valeurs dans $E \subset \mathbb{R}$, si E est dénombrable, alors on dit que X est une variable aléatoire **discrète.**

Soit X une variable discrète à valeur dans $E = \{x_1, x_2, ...\}$.

Posons $p_k = P(\omega,\ X(\omega) = x_k) = P(X = x_k)$. On a

$$\begin{aligned} P(\Omega) &= \sum_{k=1}^{\infty} P(\omega, X(\omega) = x_k) \\ &= \sum_{k=1}^{\infty} P(X = x_k) \\ &= \sum_{k=1}^{\infty} p_k = 1 \end{aligned}$$

2. Espérance mathématique

Soit g une application de $E = \{x_1, x_2, ...\}$. dans $\mathbb{R}$.

Definition

On définit l'espérance mathématique de $g(X)$ par:

$$E(g(X)) = \sum_{k=1}^{\infty} g(x_k) p_k$$

En particulier:

-Pour $g(x) = x$, on définit l'espérance mathématique de X par:

2. Espérance mathématique

Soit g une application de $E = \{x_1, x_2, ...\}$. dans $\mathbb{R}$.

Definition

On définit l'espérance mathématique de $g(X)$ par:

$$E(g(X)) = \sum_{k=1}^{\infty} g(x_k)p_k$$

En particulier:

-Pour $g(x) = x$, on définit l'espérance mathématique de X par:

1.Loi de probabilité discrète

$$
\begin{aligned}
E(X) &= \sum_{k=1}^{\infty} x_k P(X = x_k) \\
&= \sum_{k=1}^{\infty} x_k p_k
\end{aligned}
$$

-Pour $g(x) = x^2$, on définit l'espérance mathématique de X^2 par:

$$E(X^2) = \sum_{k=1}^{\infty} x_k^2 p_k.$$

$$E(X) = \sum_{k=1}^{\infty} x_k P(X = x_k) = \sum_{k=1}^{\infty} x_k p_k$$

-Pour $g(x) = x^2$, on définit l'espérance mathématique de X^2 par:

$$E(X^2) = \sum_{k=1}^{\infty} x_k^2 p_k.$$

1.Loi de probabilité discrète

On définit alors:

-La variance de X par

$$
\begin{aligned}
var(X) &= E(X^2) - (E(X))^2 \\
&= \sum_{k=1}^{\infty} x_k^2 p_k - \left(\sum_{k=1}^{\infty} x_k p_k \right)^2 .
\end{aligned}
$$

-L'écart-type de X par

$$\sigma(X) = \sqrt{\mathrm{var}(X)}.$$

Si E est fini $E = \{x_1, x_2, ..., x_r\}$. Alors

$$E(g(X)) = \sum_{k=1}^{r} g(x_k)p_k$$

$$E(X) = \sum_{k=1}^{r} x_k p_k$$

$$var(X) = \sum_{k=1}^{r} x_k^2 p_k - \left(\sum_{k=1}^{r} x_k p_k\right)^2.$$

Propriétés de l'espérance et la variance

Soit X et Y deux variables aléatoires discrètes à valeurs dans E.

Posons $p_k = P(X = x_k), q_k = P(Y = y_k)$.

Soient a,b et c des constantes réelles.

Soit

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

la covariance entre X et Y.

Theorem

$$E(c) = c, \tag{1}$$

$$E(aX + bY) = aE(X) + bE(Y), \ \textit{(linéairité)} \tag{2}$$

$$var(aX) = a^2 var(X), \tag{3}$$

$$var(aX + bY) = a^2 var(X) + 2abcov(X, Y) + b^2 var(Y). \tag{4}$$

3. Inégalités classiques

3.1. Inégalité de Markov

Theorem

Soit X une variable aléatoire à valeurs dans un espace dénombrable E avec $E \subset \mathbb{R}$, alors pour tout $a > 0$,

$$P(|f(X)| \geq a) \leq \frac{E(|f(X)|)}{a}. \tag{5}$$

Preuve: On a

$$\begin{aligned} E(|f(X)|) &= \sum_{x_k \in E} |f(x_k)|\, p_k \\ &= \sum_{x_k \in A} |f(x_k)|\, p_k + \sum_{x_k \in A^c} |f(x_k)|\, p_k, \end{aligned}$$

où $A = \{x_k \in E, |f(x_k)| \geq a\}$, et A^c est le complémentaire de A dans E. Donc

$$\begin{aligned} E(|f(X)|) &\geq \sum_{x_k \in A} |f(x_k)|\, p_k \\ &\geq \sum_{x_k \in A} a p_k, \end{aligned}$$

de plus $\sum_{x_k \in A} p_k = P(X \in A) = P(|f(X)| \geq a)$.$\square$

3.2. Inégalité de Bienaymé-Tchebyshev

Theorem

Soit X une variable aléatoire à valeurs dans E, alors pour tout $\varepsilon > 0$

$$P(|X - E(X)| \geq \varepsilon) \quad \leq \quad \frac{\sigma_X^2}{\varepsilon^2}. \tag{6}$$

Preuve : On applique l'inégalité de Markov (5) avec $f(t) = |t - E(X)|^2$, $a = \varepsilon^2$.$\square$

4. Indépendance

4.1.Variables indépendantes

Definition

Soit X et Y deux variables aléatoires. On dit que X et Y sont indépendantes si

$$P(\{X \in A\} \cap \{Y \in B\}) = P(X \in A)P(Y \in B). \quad (7)$$

4.2 Distributions marginales

Soit $E_1 \times E_2$ et le couple aléatoire $X = (X_1, X_2)'$, où chaque X_i est à valeurs dans E_i. La **distribution marginale** de la variable X_i est définie par $P_i(A) = P(X_i \in A), \forall A \in \mathcal{F}$.

4. Indépendance

4.1.Variables indépendantes

Definition

Soit X et Y deux variables aléatoires. On dit que X et Y sont indépendantes si

$$P(\{X \in A\} \cap \{Y \in B\}) = P(X \in A)P(Y \in B). \qquad (7)$$

4.2 Distributions marginales

Soit $E_1 \times E_2$ et le couple aléatoire $X = (X_1, X_2)'$, où chaque X_i est à valeurs dans E_i. La **distribution marginale** de la variable X_i est définie par $P_i(A) = P(X_i \in A), \forall A \in \mathcal{F}$.

Theorem

Soit X et Y deux variables aléatoires discrètes à valeurs dans E_1 et E_2, et indépendantes, soit $f : E_1 \to \mathbb{R}$, et $g : E_2 \to \mathbb{R}$, alors

$$E(f(X)g(Y)) = E(f(X))E(g(Y)) \tag{8}$$

Preuve : Posons $Z = (X, Y)$ et $h(Z) = f(X)g(Y)$, alors Z est à valeurs dans $E = E_1 \times E_2$, de plus d'après l'hypothèse d'indépendance $P((X, Y) = (x, y)) = P(X = x)P(Y = y)$. Donc

$$\begin{aligned} E(h(Z)) &= \sum_{z_k \in E} h(z_k)P(Z = z_k) \\ &= \sum_{x_k, y_l} f(x_k)g(y_l)P(X = x_k)P(Y = y_l) \\ &= \sum_{x_k} f(x_k)P(X = x_k) \sum_{y_l} g(y_l)P(Y = y_l). \square \end{aligned}$$

Theorem

Soit X_1 et X_2 deux variables aléatoires indépendantes, alors

$$var(X_1 + X_2) = var(X_1) + var(X_2). \qquad (9)$$

Preuve :

$$var(X_1 + X_2) = var(X_1) + var(X_2) + 2\operatorname{cov}(X_1, X_2),$$

où $\operatorname{cov}(X_1, X_2) = E(X_1X_2) - EX_1EX_2$ est la covariance de X_1 et X_2, mais d'après (8) on a $E(X_1X_2) = EX_1EX_2.\Box$

Remarque: Si deux variables aléatoires X et Y sont indépendantes alors elles sont non **covariées** $(\operatorname{cov}(X, Y) = 0)$.

1.Loi de probabilité discrète

Theorem

Soit X_1 et X_2 deux variables aléatoires indépendantes, alors

$$var(X_1 + X_2) = var(X_1) + var(X_2). \tag{9}$$

Preuve :

$$var(X_1 + X_2) = var(X_1) + var(X_2) + 2\,\mathrm{cov}(X_1, X_2),$$

où $\mathrm{cov}(X_1, X_2) = E(X_1X_2) - EX_1EX_2$ est la covariance de X_1 et X_2, mais d'après (8) on a $E(X_1X_2) = EX_1EX_2$.$\square$

Remarque: Si deux variables aléatoires X et Y sont indépendantes alors elles sont non **covariées** $(\mathrm{cov}(X, Y) = 0)$.

5.Lois usuelles

5.1. Loi de Bernoulli

Definition

On dit que X suit une loi de Bernoulli de paramètre p, noté $X \rightsquigarrow B(p)$, si $E = \{x_1, x_2\}$ avec $P(X(\omega) = x_1) = p$, et $P(X(\omega) = x_2) = 1 - p$.

Exemple : On lance une pièce de monnaie, X est le résultat obtenu, $P(X = pile) = P(X = face) = 1/2$.

5.Lois usuelles

5.1. Loi de Bernoulli

Definition

On dit que X suit une loi de Bernoulli de paramètre p, noté $X \rightsquigarrow B(p)$, si $E = \{x_1, x_2\}$ avec $P(X(\omega) = x_1) = p$, et $P(X(\omega) = x_2) = 1 - p$.

Exemple : On lance une pièce de monnaie, X est le résultat obtenu, $P(X = pile) = P(X = face) = 1/2$.

5.2. Loi Binomiale

Definition

Soit $Y_i, 1 \leq i \leq n$, une suite de variables aléatoires de Bernoulli indépendantes à valeurs dans $\{0, 1\}$ avec $P(Y_k(\omega) = 1) = p$, alors la variable $X = \sum_{k=1}^{n} Y_k$ suit une loi Binomiale, noté $X \rightsquigarrow B(n, p)$, dans ce cas

$$E = \{0, 1, ..., n\}$$

et

$$P(X(\omega) = k) = C_n^k p^k (1 - p)^{n-k}, 0 \leq k \leq n,$$

$$C_n^k = \frac{n!}{k!(n - k)!}.$$

5.3. Loi de Poisson

Definition

On dit que X suit une loi de Poisson de paramètre λ, noté $X \rightsquigarrow \mathcal{P}(\lambda)$, si $E = \mathbb{N}$ et $P(X(\omega) = k) = e^{-\lambda}\frac{\lambda^k}{k!}$, λ est un réel positif.

Exemple: $N(t)$ représente le nombre de clients qui arrivent dans une agence (ou le nombre de voitures qui arrivent sur un péage d'autoroute) pendant la période d'observation $[0, t]$, alors on montre que

$$P(N(t) = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!},$$

avec λ l'intensité d'arrivage.

5.3. Loi de Poisson

Definition

On dit que X suit une loi de Poisson de paramètre λ, noté $X \rightsquigarrow \mathcal{P}(\lambda)$, si $E = \mathbb{N}$ et $P(X(\omega) = k) = e^{-\lambda}\frac{\lambda^k}{k!}$, λ est un réel positif.

Exemple: $N(t)$ représente le nombre de clients qui arrivent dans une agence (ou le nombre de voitures qui arrivent sur un péage d'autoroute) pendant la période d'observation $[0, t]$, alors on montre que

$$P(N(t) = k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!},$$

avec λ l'intensité d'arrivage.

2.Loi de probabilité continue

Soit X une variable à valeurs dans $E \subset \mathbb{R}$, si E est non dénombrable, alors on dit que X est une variable **continue**.

1. Densité de probabilité, moments

Soit X une variable aléatoire à valeurs dans $\mathbb{R}$, et soit f_X une fonction de $\mathbb{R}$ dans $\mathbb{R}^+$ intégrable et telle que

$$\int_{\mathbb{R}} f_X(x)dx = 1. \tag{10}$$

Definition

On dit que X admet la densité de probabilité f_X si pour tout $A \subset \mathbb{R}$

$$P(X \in A) = \int_A f_X(x)dx.$$

2.Loi de probabilité continue

Soit X une variable à valeurs dans $E \subset \mathbb{R}$, si E est non dénombrable, alors on dit que X est une variable **continue**.

1. Densité de probabilité, moments

Soit X une variable aléatoire à valeurs dans $\mathbb{R}$, et soit f_X une fonction de $\mathbb{R}$ dans $\mathbb{R}^+$ intégrable et telle que

$$\int_{\mathbb{R}} f_X(x)dx = 1. \tag{10}$$

Definition

On dit que X admet la densité de probabilité f_X si pour tout $A \subset \mathbb{R}$

$$P(X \in A) = \int_A f_X(x)dx.$$

Definition

Soit $g : \mathbb{R} \to \mathbb{R}$, telle que $\int_{\mathbb{R}} |g(x)| \, f_X(x)dx < \infty$, alors on définit l'espérance mathématique de $g(X)$ par

$$E(g(X)) = \int_{\mathbb{R}} g(x) f_X(x) dx. \tag{11}$$

En particulier: -Pour $g(x) = x$, on définit l'espérance mathématique de X par

$$E(X) = \int_{\mathbb{R}} x f_X(x) dx.$$

Definition

Soit $g : \mathbb{R} \to \mathbb{R}$, telle que $\int_{\mathbb{R}} |g(x)| \, f_X(x)dx < \infty$, alors on définit l'espérance mathématique de $g(X)$ par

$$E(g(X)) = \int_{\mathbb{R}} g(x) f_X(x)dx. \tag{11}$$

En particulier: -Pour $g(x) = x$, on définit l'espérance mathématique de X par

$$E(X) = \int_{\mathbb{R}} x f_X(x)dx.$$

2.Loi de probabilité continue

-Pour $g(x) = x^2$, on définit l'espérance mathématique de X^2 par:

$$E(X^2) = \int_{\mathbb{R}} x^2 f_X(x) dx.$$

On définit alors la variance de X par

$$\begin{aligned} var(X) &= E(X^2) - (E(X))^2 \\ &= \int_{\mathbb{R}} x^2 f_X(x) dx - \left(\int_{\mathbb{R}} x f_X(x) dx \right)^2 \end{aligned}$$

et l'écart-type de X par

$$\sigma(X) = \sqrt{\mathrm{var}(X)}.$$

Couple aléatoire

Soit $X = (X_1, X_2)$ un couple aléatoire à valeurs dans $\mathbb{R}^2$, et soit f_X une fonction de $\mathbb{R}^2$ dans $\mathbb{R}^+$ intégrable et telle que

$$\int_{\mathbb{R}^2} f_X(x)dx = 1. \tag{12}$$

Definition

On dit que X admet la densité de probabilité f_X si pour tout $A \subset \mathbb{R}^2$

$$P(X \in A) = \int_A f_X(x)dx.$$

2.Loi de probabilité continue

Couple aléatoire

Soit $X = (X_1, X_2)$ un couple aléatoire à valeurs dans $\mathbb{R}^2$, et soit f_X une fonction de $\mathbb{R}^2$ dans $\mathbb{R}^+$ intégrable et telle que

$$\int_{\mathbb{R}^2} f_X(x)dx = 1. \tag{12}$$

Definition

On dit que X admet la densité de probabilité f_X si pour tout $A \subset \mathbb{R}^2$

$$P(X \in A) = \int_A f_X(x)dx.$$

2.Loi de probabilité continue

Soit $g : \mathbb{R}^2 \to \mathbb{R}$, telle que $\int_{\mathbb{R}^2} |g(x)| f_X(x)dx < \infty$, alors on définit l'espérance mathématique de $g(X)$ par

$$E(g(X)) = \int_{\mathbb{R}^2} g(x) f_X(x)dx, \qquad (13)$$

en particulier: • Si $g(x_1, x_2) = g_i(x) = x_i, 1 \leq i \leq 2$, alors on définit l'espérance mathématique du couple X par

$$E(X) = \begin{pmatrix} E(X_1) \\ \\ E(X_2) \end{pmatrix} = \begin{pmatrix} \int_{\mathbb{R}^2} x_1 f_X(x)dx \\ \\ \int_{\mathbb{R}^2} x_2 f_X(x)dx \end{pmatrix}.$$

• Si $g(x_1, x_2) = x_1 x_2$, on définit

$$E(X_1 X_2) = \int_{\mathbb{R}^2} x_1 x_2 f_X(x)dx.$$

2.Indépendance

2.1. Distributions marginales

Definition

Soit X et Y deux variables aléatoires à valeurs dans $\mathbb{R}$ tels que $(X, Y)'$ admette la densité $f_{X,Y}(x, y)$. Alors on définit les **densités marginales** de X et de Y par

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y)dy$$

et

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x, y)dx.$$

2.2.Variables indépendantes

Definition

Soit X et Y deux variables aléatoires à valeurs dans $\mathbb{R}$. On dit que X et Y sont indépendantes si

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B). \tag{14}$$

En utilisant les mêmes arguments que le théorème 12 on obtient le théorème

Theorem

Soit X et Y deux variables aléatoires à valeurs dans $\mathbb{R}$ indépendantes, soit $f : \mathbb{R} \to \mathbb{R}$, et $g : \mathbb{R} \to \mathbb{R}$EndExpansion, alors

$$E(f(X)g(Y)) = E(f(X))E(g(Y)) \tag{15}$$

Theorem

(admi) Soit $X = (X_1, X_2)$ *de densité* $f_X(x)$*, et tel que les* $X_i, 1 \leq i \leq 2$*, soient des variables aléatoires indépendantes. On a alors*

$$f_X(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2), \tag{16}$$

où f_{X_i} *est la densité marginale de la variable* X_i*.*

Réciproquement, si f_X *a la forme (16) où les* $f_{X_i}, 1 \leq i \leq 2$ *sont des densités de probabilité, alors les* $X_i, 1 \leq i \leq 2$*, sont indépendantes et les* f_{X_i} *sont les densités des* X_i*.*

3. Fonction de répartition

Cas d'une variable aléatoire

Definition

Soit X une variable aléatoire de densité $f(x)$, la fonction de **répartition** de X est définie par : $F : \mathbb{R} \to [0,1]$ et

$$F(t) = P(X \leq t) = \int_{-\infty}^{t} f(x)dx.$$

Remarque : F est une fonction non décroissante telle que $F(-\infty) = 0, F(+\infty) = 1$ et $F'(t) = f(t)$.

Cas d'un couple aléatoire

Soit $X = (X_1, X_2)$ un couple aléatoire de densité $f(x)$, la fonction de répartition de X est définie par : $F : \mathbb{R}^2 \to [0, 1]$ et si $t = (t_1, t_2)$ alors

$$F(t) = P(X_1 \leq t_1, X_2 \leq t_2) = \int_{-\infty}^{t_1} \int_{-\infty}^{t_2} f(x_1, x_2) dx_1 dx_2.$$

On a aussi

$$f(t_1, t_2) = \frac{\partial^2 F(t_1, t_2)}{\partial t_1 \partial t_2}.$$

4. Lois usuelles

4.1 Loi uniforme

Definition

On dit que X suit une loi uniforme sur $[a, b]$, noté $X \rightsquigarrow U[a, b]$, si elle admet la densité

$$f(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{sinon.} \end{cases} \tag{17}$$

On montre que

$$E(X) = (a+b)/2, V(X) = (b-a)^2/12.$$

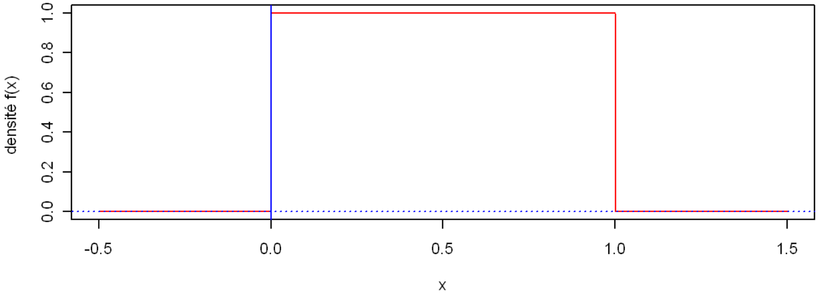


Figure: Densité de la loi uniforme U[0,1]

4.2. Loi de Gauss

Definition

On dit que X suit une loi de Gauss de moyenne m et de variance σ^2, noté $X \rightsquigarrow N(m, \sigma^2)$, si elle admet la densité

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \forall x \in \mathbb{R}. \tag{18}$$

On montre que $E(X) = m$, $V(X) = \sigma^2$.

2.Loi de probabilité continue

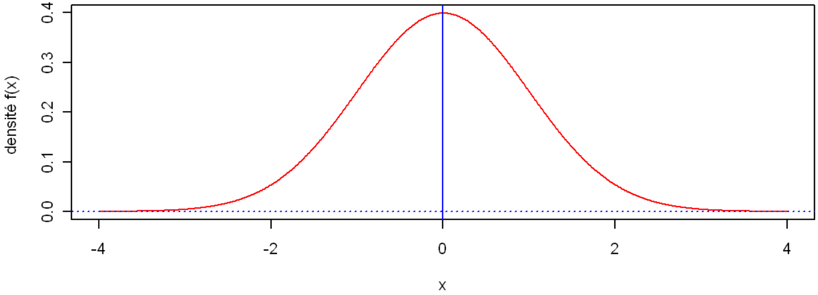


Figure: Densité de la loi Gauss N(0,1)

2.Loi de probabilité continue

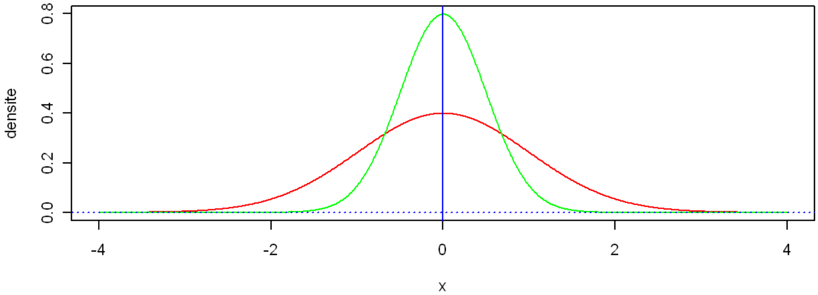


Figure: Densités Gaussiennes, rouge: $N(0,1)$, vert: $\mathrm{N}(0,\frac{1}{2})$.

4.3. Loi exponentielle

Definition

On dit que X suit une loi exponentielle de paramètre $\lambda > 0$, noté $X \rightsquigarrow E(\lambda)$, si elle admet la densité

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{sinon.} \end{cases} \tag{19}$$

On montre que $E(X) = \frac{1}{\lambda}$, $V(X) = \frac{1}{\lambda^2}$.

2.Loi de probabilité continue

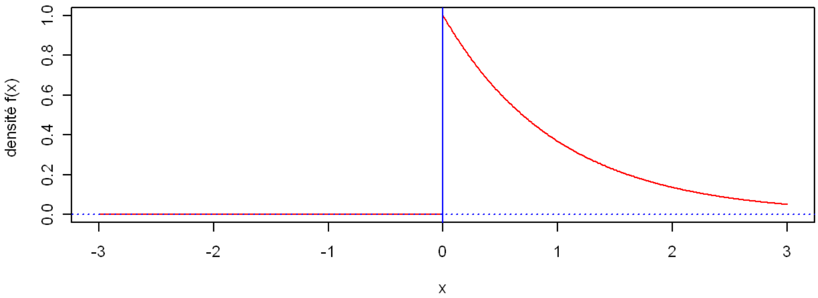


Figure: Densité de la loi exponentielle E(1)

4.4. Loi gamma

Definition

On dit que X suit une loi gamma de paramètres $\alpha > 0$ et $\beta > 0$, noté $X \rightsquigarrow \Gamma(\alpha, \beta)$, si elle admet la densité

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{si } x \geq 0 \\ 0 & \text{sinon,} \end{cases} \tag{20}$$

où $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.
On montre que $E(X) = \frac{\alpha}{\beta}$, $V(X) = \frac{\alpha}{\beta^2}$.

5. Transformation desvariables aléatoires

5.1 Cas discret

Exemple : Soit X la variable aléatoire définie par

x_i	0	1	-1	2
$P(X(\omega) = x_i$	0.25	0.2	0.25	0.3

On définit la variable aléatoire $Y = X(X+1)$, alors Y prends ses valeurs dans $\{0, 2, 6\}$ avec la loi de probabilité suivante:

y_i	0	2	6
$P(Y(\omega) = y_i)$	0.5	0.2	0.3

.

2.Loi de probabilité continue

Theorem

Si X est une variable aléatoire discrète à valeurs dans E, et g une application, alors $Y = g(X)$ est une variable aléatoire discrète à valeurs dans $g(E)$ telle que:

- *Si g est injective alors Y a la même loi de probabilité que X : $P(Y(\omega) = y_i) = P(X(\omega) = x_i), y_i = g(x_i)$.*
- *Si g est non injective alors Y a loi de probabilité suivante:*

$$P(Y(\omega) = y_i) = \sum_{x_i \text{ tels que } y_i = g(x_i)} P(X(\omega) = x_i).$$

5.2. Cas continu

Soit $Y = X^2$, où $X \rightsquigarrow N(0,1)$ donc Y est à valeurs dans $\mathbb{R}^+$ donc $f_Y(y) = 0 \; \forall y \leq 0$. Si $y \geq 0$, on a $F_Y(y) = P(Y \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{y}}^{\sqrt{y}} e^{-\frac{x^2}{2}} = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{y}} e^{-\frac{x^2}{2}}$ donc $f_Y(y) = F_Y'(y) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{y^2}{2}}$, c'est à dire que $Y \rightsquigarrow \mathcal{X}^2(1)$. Si la transformation g n'est pas bijective, on utilise souvent la fonction de répartition pour calculer la loi de $Y = g(X)$. Cependant si g est bijective on a le théorème

Theorem

Soit X une variable aléatoire de densité f_X, et soit g une application de $\mathbb{R}$ dans $\mathbb{R}$. On définit la transformée de X par g: $Y = g(X)$. On suppose que g est continûment dérivable sur $\mathbb{R}$, et que sa dérivée g' ne s'annule pas. Alors Y admet une densité f_Y donnée par

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|}, \quad x = g^{-1}(y). \tag{21}$$

Exemple $Y = g(X) = aX + b, a \neq 0$. On a
$y = g(x) = ax + b$ donc $x = \frac{y-b}{a} = g^{-1}(y)$, donc

$$f_Y(y) = \frac{f_X(x)}{|g'(x)|} = \frac{f_X(\frac{y-b}{a})}{|a|}.$$

6. Suite de variables aléatoires

6.1. Loi de Khi-deux

Definition

On dit que X suit une loi Khi-deux à n degrés de liberté , noté $X \rightsquigarrow \mathcal{X}^2(n)$, si $X \rightsquigarrow \Gamma(\frac{n}{2}, \frac{1}{2})$.
$X = \sum_{k=1}^{n} Z_k^2$ où $(Z_k, 1 \leq k \leq n)$ est une suite de variables aléatoires indépendantes et identiquement distribuées (i.i.d.) suivant une loi gaussienne $N(0, 1)$.

6.2. Loi de Student

Definition

On dit que X suit une loi Student à n degrés de liberté , noté $X \rightsquigarrow t(n)$, si $X = \frac{Z}{\sqrt{Y/n}}$, $Z \rightsquigarrow N(0,1)$, $Y \rightsquigarrow \mathcal{X}^2(n)$, et Z et Y sont indépendantes.

6.3. Loi de Fisher

Definition

On dit que X suit une loi Fisher à (n_1, n_2) degrés de liberté , noté $X \rightsquigarrow F(n_1, n_2)$, si $X = \frac{Y_1/n_1}{Y_2/n_2}$, $Y_i \rightsquigarrow \mathcal{X}^2(n_i), i = 1, 2$ et Y_1 et Y_2 sont indépendantes.

6.2. Loi de Student

Definition

On dit que X suit une loi Student à n degrés de liberté , noté $X \rightsquigarrow t(n)$, si $X = \frac{Z}{\sqrt{Y/n}}$, $Z \rightsquigarrow N(0,1)$, $Y \rightsquigarrow \mathcal{X}^2(n)$, et Z et Y sont indépendantes.

6.3. Loi de Fisher

Definition

On dit que X suit une loi Fisher à (n_1, n_2) degrés de liberté , noté $X \rightsquigarrow F(n_1, n_2)$, si $X = \frac{Y_1/n_1}{Y_2/n_2}$, $Y_i \rightsquigarrow \mathcal{X}^2(n_i), i = 1, 2$ et Y_1 et Y_2 sont indépendantes.

4.1 Généralités, Définitions

- **La statistique**: C'est l'étude des phénomènes préalablement exprimés sous forme numérique.

- **L'unité statistique (ou individu)** :C'est l'élément de l'ensemble que l'on veut étudier, exemple: un étudiant est une unité statistique lorsque l'on étudie une classe.

- **Le caractère:** C'est l'aspect de l'unité statistique que l'on retient dans l'analyse.

4.Description d'une série statistique

4.1 Généralités, Définitions

- **La statistique**: C'est l'étude des phénomènes préalablement exprimés sous forme numérique.

- **L'unité statistique (ou individu)** :C'est l'élément de l'ensemble que l'on veut étudier, exemple: un étudiant est une unité statistique lorsque l'on étudie une classe.

- **Le caractère:** C'est l'aspect de l'unité statistique que l'on retient dans l'analyse.

4.1 Généralités, Définitions

- **La statistique**: C'est l'étude des phénomènes préalablement exprimés sous forme numérique.

- **L'unité statistique (ou individu)** :C'est l'élément de l'ensemble que l'on veut étudier, exemple: un étudiant est une unité statistique lorsque l'on étudie une classe.

- **Le caractère:** C'est l'aspect de l'unité statistique que l'on retient dans l'analyse.

4.Description d'une série statistique

- **L'échantillon:** C'est un sous-ensemble d'une population statistique.

- **La variable statistique:** C'est l'expression numérique du caractère observé sur les unités statistiques considérées, exemple, taille, poids, note en proba-stat.... de l'étudiant.

4.Description d'une série statistique

- **L'échantillon:** C'est un sous-ensemble d'une population statistique.

- **La variable statistique:** C'est l'expression numérique du caractère observé sur les unités statistiques considérées, exemple, taille, poids, note en proba-stat.... de l'étudiant.

4.Description d'une série statistique

- La variable statistique x est dite **discrète** lorsqu'elle ne peut prendre que des valeurs isolées :

$$x_1, x_2, ..., x_k, \qquad \text{où } x_1 < x_2 < ... < x_k.$$

Exemple : le nombre de chevaux fiscaux d'une voiture; le nombre d'enfants dans une famille,...

- La variable statistique x est dite **continue** lorsqu'elle peut prendre n'importe quelle valeur d'un intervalle $[a, b]$.
Exemple : la durée en minute d'une conversation téléphonique.

4.Description d'une série statistique

- La variable statistique x est dite **discrète** lorsqu'elle ne peut prendre que des valeurs isolées :

$$x_1, x_2, ..., x_k, \quad \text{où } x_1 < x_2 < ... < x_k.$$

  Exemple : le nombre de chevaux fiscaux d'une voiture; le nombre d'enfants dans une famille,...

- La variable statistique x est dite **continue** lorsqu'elle peut prendre n'importe quelle valeur d'un intervalle $[a, b]$.
  Exemple : la durée en minute d'une conversation téléphonique.

4.Description d'une série statistique

Dans ce cas l'intervalle des valeurs possibles est divisé en k classes

$$[a_0, a_1[, [a_1, a_2[, ..., [a_{k-1}, a_k],$$

avec

$$a_0 = a < a_1 < ... < a_{k-1} < a_k = b.$$

a_{h-1} et a_h sont les frontières de la h-ième classe, $x_h = \frac{a_{h-1}+a_h}{2}$ est le centre de celle-ci.

4.Description d'une série statistique

La série statistique: C'est à la fois :

- **L'ensemble des valeurs** $\{x_1, x_2, ..., x_k\}$ (resp. des classes de valeurs

$$\{[a_0, a_1[, [a_1, a_2[, ..., [a_{k-1}, a_k]\})$$

de la variable x.

- **Le nombre d'observation** associé à chaque valeur x_i (resp. à chaque classe), qu'on appelle **effectif**.

4.Description d'une série statistique

La série statistique: C'est à la fois :

- **L'ensemble des valeurs** $\{x_1, x_2, ..., x_k\}$ (resp. des classes de valeurs

$$\{[a_0, a_1[, [a_1, a_2[, ..., [a_{k-1}, a_k]\})$$

de la variable x.

- **Le nombre d'observation** associé à chaque valeur x_i (resp. à chaque classe), qu'on appelle **effectif**.

2. Distribution des fréquences

2.1 Cas discret

- Si dans une série statistique résultant de n observations, on a trouvé n_i fois la valeur x_i, n_i représente **l'effectif** de cette valeur et le rapport $f_i = \frac{n_i}{n}$ représente sa **fréquence.**

- L 'expression $\sum_{i=1}^{j} n_i$ est appelée **effectif cumulé** des valeurs de x inférieures ou égales à x_j.

- L 'expression $\sum_{i=1}^{j} f_i$ est appelée **fréquence cumulée** des valeurs de x inférieures ou égales à x_j.

2. Distribution des fréquences

2.1 Cas discret

- Si dans une série statistique résultant de n observations, on a trouvé n_i fois la valeur x_i, n_i représente **l'effectif** de cette valeur et le rapport $f_i = \frac{n_i}{n}$ représente sa **fréquence.**
- L 'expression $\sum_{i=1}^{j} n_i$ est appelée **effectif cumulé** des valeurs de x inférieures ou égales à x_j.
- L 'expression $\sum_{i=1}^{j} f_i$ est appelée **fréquence cumulée** des valeurs de x inférieures ou égales à x_j.

2. Distribution des fréquences

2.1 Cas discret

- Si dans une série statistique résultant de n observations, on a trouvé n_i fois la valeur x_i, n_i représente **l'effectif** de cette valeur et le rapport $f_i = \frac{n_i}{n}$ représente sa **fréquence.**

- L 'expression $\sum_{i=1}^{j} n_i$ est appelée **effectif cumulé** des valeurs de x inférieures ou égales à x_j.

- L 'expression $\sum_{i=1}^{j} f_i$ est appelée **fréquence cumulée** des valeurs de x inférieures ou égales à x_j.

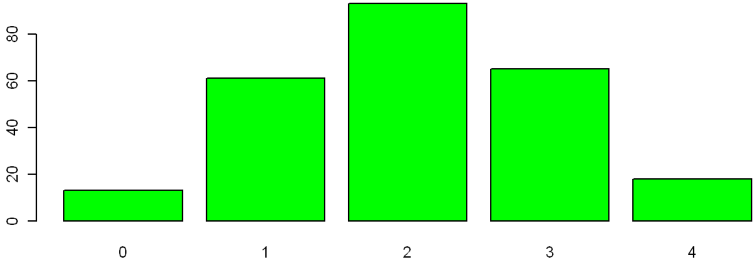
4.Description d'une série statistique

Exemple : L'unité statistique (ou individu) étant la famille de quatre enfants dont l'aîné a moins de 16 ans, on s'intéresse au nombre x de garçons qui la compose. L'étude statistique suivante porte sur un échantillon de 250 familles.

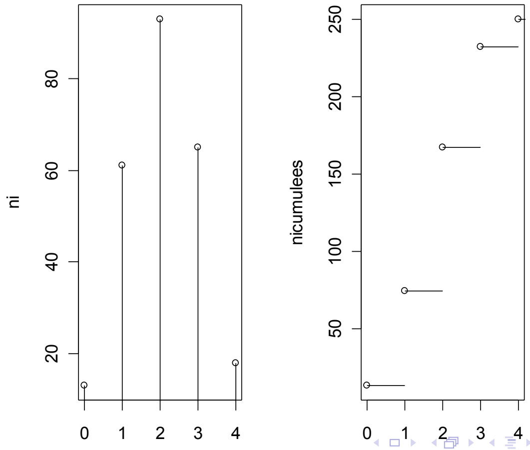
Valeurs de x	0	1	2	3	4
Effectif n_i	13	61	93	65	18
Fréquence f_i	0.052	0,244	0,372	0,260	0,072
Effectif cumulé	13	74	167	232	250
Fréquence cumulée	0.052	0,296	0,668	0,928	1

4.Description d'une série statistique

Valeurs de x	0	1	2	3	4
Effectif n_i	13	61	93	65	18



4.Description d'une série statistique



2.2 Cas continu

- On appelle **effectif** de la hième classe la nombre n_h d'observations dont les valeurs appartiennent à cette classe. Le nombre $f_h = \frac{n_h}{n}$ représente la **fréquence** de la hième classe.
- L 'expression $\sum_{i=1}^{h} n_i$ est appelée **effectif cumulé** des h premières classes.
- L 'expression $\sum_{i=1}^{h} f_i$ est appelée **fréquence cumulée** des h premières classes.

2.2 Cas continu

- On appelle **effectif** de la hième classe la nombre n_h d'observations dont les valeurs appartiennent à cette classe. Le nombre $f_h = \frac{n_h}{n}$ représente la **fréquence** de la hième classe.

- L 'expression $\sum_{i=1}^{h} n_i$ est appelée **effectif cumulé** des h premières classes.

- L 'expression $\sum_{i=1}^{h} f_i$ est appelée **fréquence cumulée** des h premières classes.

4.Description d'une série statistique

Exemple : On s'intéresse à la durée x de service d'un guichet de la poste qui peut servir au plus un client à la fois. On a relevé la durée de service de 1000 clients consécutifs. L'unité de temps est la seconde, les résultats sont résumés par le tableau suivant:

Classes des valeurs de x	[0,30[	[30,60[	[60,90[	[90,120[	[120,150[	[150,180[	[180,240[
Effectifs n_h	369	251	148	98	65	43	26
Fréquences f_h	0,369	0,251	0,148	0.098	0.065	0.043	0.026
Effectifs cumulés	369	620	768	866	931	974	1000
Fréquences cumulées	0,369	0,620	0,768	0,866	0,931	0,974	1

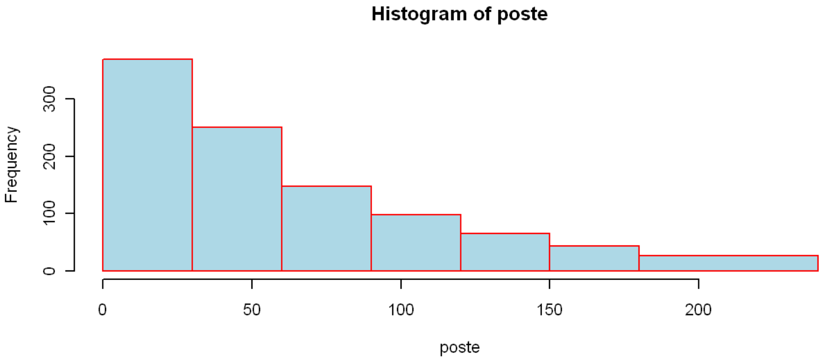


Figure: Effectifs des 7 classes

2.3 Distribution des fréquences

L'ensemble des k valeurs x_i (resp. des k classes) et les effectifs n_i correspondants, constitue la distribution des fréquences de la variable x.

Représentation graphique des distributions des fréquences

On place en ordonnée les effectifs. Dans le cas discret on obtient un diagramme en bâtons. Dans le cas continu, on construit un histogramme des fréquences, l'aire de chaque classe devant être proportionnelle à son effectif correspondant.

3. Caractéristiques de tendance centrale

3.1.Définition: Ce sont des indicateurs qui permettent de synthétiser l'ensemble de la série statistique en faisant ressortir une position centrale de la valeur du caractère étudié.

- **La médiane :** C'est la valeur x_m de la variable x qui partage les éléments de la série statistique, préalablement classés par ordre de valeurs croissantes, en deux groupes d'effectifs égaux : 50 % des valeurs observées sont inférieures à x_m et 50 % sont supérieures.

4.Description d'une série statistique

Dans le cas continu, elle est donnée par

$$x_m = a_{j-1} + \frac{a_j - a_{j-1}}{n_j}\left(\frac{n}{2} - \sum_{i=1}^{j} n_i\right),$$

sachant que x_m appartient à $[a_{j-1}, a_j[$.

- **Le mode :**

  Variable discrète : c'est la valeur pour laquelle la fréquence est maximale.

  Variable continue : la classe modale est celle pour laquelle la fréquence est maximale.

  Lorsqu'il y a une seule valeur ou classe modale, la distribution est dite **unimodale.**

- **La moyenne.**

  -La moyenne $\overline{x}$ de k valeurs distinctes $x_1, ..., x_k$, et d'effectifs $n_1, ..., n_k$:

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

-La moyenne $\overline{x}$ de n valeurs distribuées en k classes $[a_0, a_1[, ..., [a_{k-1}, a_k[$ et d'effectifs $n_1, ..., n_k$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, \, x_i = \frac{a_{i-1} + a_i}{2}.$$

- **La moyenne.**

  -La moyenne $\overline{x}$ de k valeurs distinctes $x_1, ..., x_k$, et d'effectifs $n_1, ..., n_k$:

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

  -La moyenne $\overline{x}$ de n valeurs distribuées en k classes $[a_0, a_1[, ..., [a_{k-1}, a_k[$ et d'effectifs $n_1, ..., n_k$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i, x_i = \frac{a_{i-1} + a_i}{2}.$$

4.Caractéristiques de dispersion

Deux séries de même nature peuvent présenter des caractéristiques de tendance centrale voisines, mais être différentes quant à la dispersion de leurs valeurs par rapport à la tendance centrale.

4.1. La dispersion en termes d'intervalles

- **Intervalle de variation ou étendue :** C'est la différence entre les valeurs extrêmes de la variable ou entre les centres des classes extrêmes $x_k - x_1$.

- **L'intervalle interquartile :** Si l'on peut diviser $x_1, ..., x_k$ en quatre classes $[x_1, Q_1[, [Q_1, Q_2[, [Q_2, Q_3[, [Q_3, x_k]$ ayant toutes la même fréquence $\frac{n}{4}$, alors Q_1, Q_2, et Q_3 sont appelés respectivement premier, deuxième et troisième quartiles. Dans le cas continu

$$Q_1 = a_{j-1} + \frac{a_j - a_{j-1}}{n_j}\left(\frac{n}{4} - \sum_{i=1}^{j} n_i\right),$$

$$Q_3 = a_{j-1} + \frac{a_j - a_{j-1}}{n_j}\left(\frac{3n}{4} - \sum_{i=1}^{j} n_i\right).$$

$[Q_1, Q_3]$ est l'intervalle interquartile.

4.2. La dispersion en termes d'écarts

Les écarts ne traduiront la dispersion que si l'on considère leurs valeurs absolues ou leurs puissances paires.

- **Les écarts absolus :**

  -L' écart absolu par rapport à la médiane

  $e_M = \frac{1}{n}\sum_{i=1}^{k} n_i \left|x_i - M\right|$, M est la médiane de la série.

  -L' écart absolu par rapport à la moyenne arithmétique

  $e_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{k} n_i \left|x_i - \overline{x}\right|$.

- **Variance, écart-type:**

  -Variance: $V(x) = \widehat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2$.

  -Ecart-type : $\widehat{\sigma}_x$.

4.2. La dispersion en termes d'écarts

Les écarts ne traduiront la dispersion que si l'on considère leurs valeurs absolues ou leurs puissances paires.

- **Les écarts absolus :**

  -L' écart absolu par rapport à la médiane
  $e_M = \frac{1}{n}\sum_{i=1}^{k} n_i \left|x_i - M\right|$, M est la médiane de la série.

  -L' écart absolu par rapport à la moyenne arithmétique
  $e_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{k} n_i \left|x_i - \overline{x}\right|$.

- **Variance, écart-type:**

  -Variance: $V(x) = \widehat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \overline{x})^2$.

  -Ecart-type : $\widehat{\sigma}_x$.

4.2. La dispersion en termes d'écarts

Les écarts ne traduiront la dispersion que si l'on considère leurs valeurs absolues ou leurs puissances paires.

- **Les écarts absolus :**

  -L' écart absolu par rapport à la médiane

  $e_M = \frac{1}{n}\sum_{i=1}^{k} n_i \left| x_i - M \right|$, M est la médiane de la série.

  -L' écart absolu par rapport à la moyenne arithmétique

  $e_{\overline{x}} = \frac{1}{n}\sum_{i=1}^{k} n_i \left| x_i - \overline{x} \right|$.

- **Variance, écart-type:**

  -Variance: $V(x) = \widehat{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2$.

  -Ecart-type : $\widehat{\sigma}_x$.

4.2. La dispersion en termes d'écarts

Les écarts ne traduiront la dispersion que si l'on considère leurs valeurs absolues ou leurs puissances paires.

- **Les écarts absolus :**

  -L' écart absolu par rapport à la médiane

  $e_M = \frac{1}{n} \sum_{i=1}^{k} n_i \left| x_i - M \right|$, M est la médiane de la série.

  -L' écart absolu par rapport à la moyenne arithmétique

  $e_{\overline{x}} = \frac{1}{n} \sum_{i=1}^{k} n_i \left| x_i - \overline{x} \right|$.

- **Variance, écart-type:**

  -Variance: $V(x) = \widehat{\sigma}_x^2 = \frac{1}{n} \sum_{i=1}^{k} n_i (x_i - \overline{x})^2$.

  -Ecart-type : $\widehat{\sigma}_x$.

Theorem

pour tout $a \in \mathbb{R}$, *on a*

$$V(a+x) = V(x), V(ax) = a^2 V(x),$$

et

$$V(x) = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - (\overline{x})^2.$$

Exemple 1. La distribution des salaires annuels dans une population contenant $n = 14890442$ actifs, (le nombre d'individus N_i touchant le salaire s_i exprimé en euros) est distribuée comme suit:

s_i (en euro)	N_i
10000	11583265
20000	2045372
30000	683240
40000	349740
50000	183280
100000	36250
200000	7975
500000	1015
1000000	305

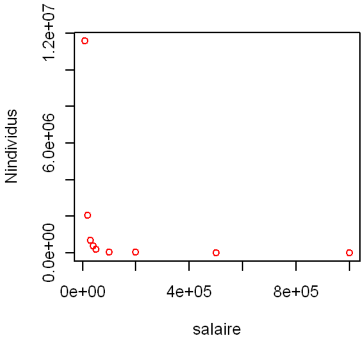


Figure: Distribution des salaires

4.Description d'une série statistique

Le salaire moyen est donné par

$$\overline{s} = \frac{1}{n}\sum_{i=1}^{9} N_i s_i = 13862.81 \text{ euros },$$

la variance et l'écart-type sont donnés par:

$$V(s) = \frac{1}{n}\sum_{i=1}^{9} N_i (s_i - \overline{s})^2 = 133496156, \widehat{\sigma}_s = 11554.05,$$

le premier, deuxième et troisième quartiles sont

$$Q_1 = 10435.56, Q_2 = 14338.53 \text{ et } \; Q_3 = 18241.5.$$

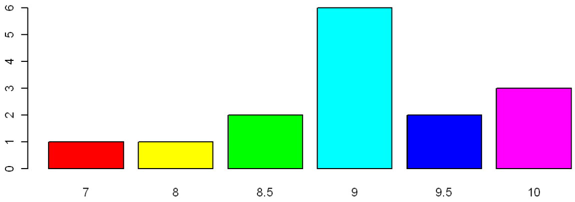
Remarque. Le salaire médian qui est égal à $s_m = Q_2$ est différent du salaire moyen $\overline{s}$.

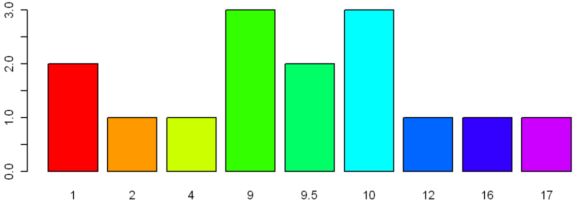
4.Description d'une série statistique

Exemple 2. Dans une classe constituée de deux groupes G1 et G2, la note obtenue dans l'examen de statistique est distribuée comme suit:

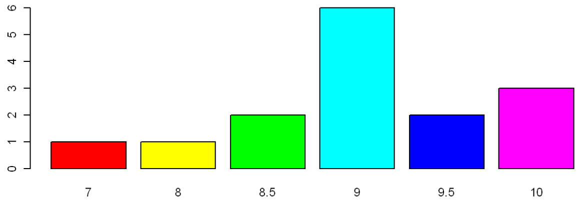
Groupe 1	9	10	9	9	8.5	9.5	10	10	7	9	9	9	8.5
	9.5	8											
Groupe 2	2	10	1	9	4	9.5	10	10	7	9	9	17	12
	9.5	16											

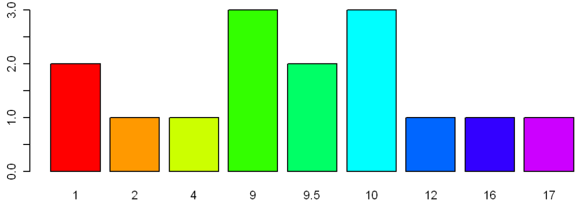
4.Description d'une série statistique





4.Description d'une série statistique





4.Description d'une série statistique

La moyenne de la classe est égale à $\overline{x} = 9$.

La moyenne du groupe 1 est égale à $\overline{x}_{G1} = 9$, et la moyenne du groupe 2 est aussi égale à $\overline{x}_{G2} = 9$. Cependant les deux groupes, même s'ils ont la même moyenne, sont complètement différents. En effet dans le groupe 1, la plupart des notes tournent autour de la moyenne globale 9, les étudiants ont presque le même niveau. Le groupe 2 est plus hétérogène, il y a des étudiants faibles (notes:1,2 et 4) et des étudiants forts (notes: 12,16 et 17). Si on calcule la variance de chaque groupe on trouve

$V_{G1} = 0.6428571$, $V_{G2} = 19.10714$. $V_{G2} > V_{G1}$, en conclusion plus la variance est grande plus l'échantillon est hétérogène.

4.Description d'une série statistique

5. Moments

Definition

Le moment d'ordre r de la variable x par rapport à l'origine x_0 est défini par

$$m_r(x_0) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - x_0)^r.$$

Si $x_0 = 0$ les moments sont dits simples et notés m_r.
Si $x_0 = \overline{x}$ les moments sont dits centrés et notés μ_r.

Theorem

$$\mu_2 = m_2 - m_1^2, \mu_3 = m_3 - 3m_1m_2 + 3m_1^3,$$

$$\mu_4 = m_3 - 4m_1m_3 + 6m_1^3m_2 - 3m_1^4.$$

5. Moments

Definition

Le moment d'ordre r de la variable x par rapport à l'origine x_0 est défini par

$$m_r(x_0) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - x_0)^r.$$

Si $x_0 = 0$ les moments sont dits simples et notés m_r.
Si $x_0 = \overline{x}$ les moments sont dits centrés et notés μ_r.

Theorem

$$\mu_2 = m_2 - m_1^2, \mu_3 = m_3 - 3m_1m_2 + 3m_1^3,$$

$$\mu_4 = m_3 - 4m_1m_3 + 6m_1^3m_2 - 3m_1^4.$$

5.Liaisons entre variables statistiques

Definition

Une série double à deux indices est définie par l'observation de n couples de valeurs tels que:

- Cas discret: x prends les valeurs $x_1, ..., x_p$ et y prends les valeurs $y_1, ..., y_r$, et à chaque couple de valeurs $(x_i, y_j)_{1 \leq i \leq p; 1 \leq j \leq r}$ on associe:
  - Son effectif $n_{i,j}$ qui est le nombre de fois où $(x, y) = (x_i, y_j)$.
  - Sa fréquence $f_{i,j} = \frac{n_{i,j}}{n}$.

- Cas continu : Les valeurs observées pour x et pour y sont regroupées respectivement en p et r classes $[a_0, a_1[,[a_1, a_2[,...,[a_{p-1}, a_p]$, resp. $[b_0, b_1[$, $[b_1, b_2[$,..., $[b_{r-1}, b_r]$, et $n_{i,j}$ est le nombre d'observations pour lesquelles $a_{i-1} \leq x \leq a_i$, et $b_{j-1} \leq y \leq b_j$.

Exemple : sur 200 personnes de même âge et de même sexe on en dénombre 25 dont la taille y est comprise entre 165 et 167 cm et le poids x entre 60 et 65 kg ; 30 dont la taille entre 167 et 169 et le poids entre 60 et 65 kg ; le reste est tel que $55 \leq x \leq 60$ et $160 \leq y \leq 165$.On peut représenter les données par le tableau

y(taille) / x(poids)	$160-165$	165-167	$167-169$	Totaux partiels
$55-60$	145	0	0	145
$60-65$	0	25	30	55
Totaux partiels	145	25	30	200

Présentation sous forme d'un tableau à double entrée(ou matrice):

y / x	y_1	$\dots$	y_j	$\dots$	y_r	Totaux partiels
x_1	$n_{1,1}$	$\dots$	$n_{1,j}$	$\dots$	$n_{1,r}$	$n_{1,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	$n_{i,1}$	$\dots$	$n_{i,j}$	$\dots$	$n_{i,r}$	$n_{i,\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$n_{p,1}$	$\dots$	$n_{p,j}$	$\dots$	$n_{p,r}$	$n_{p,\bullet}$
Totaux partiels	$n_{\bullet,1}$	$\dots$	$n_{\bullet,j}$	$\dots$	$n_{\bullet,r}$	n

$$n_{i,\bullet} = \sum_{j=1}^{r} n_{i,j},\ n_{\bullet,j} = \sum_{i=1}^{p} n_{i,j},\ n = \sum_{i=1}^{p}\sum_{j=1}^{r} n_{i,j}$$

6.La distribution marginale

Definition

La distribution **marginale** de x se lit sur la dernière colonne du tableau; elle concerne les effectifs de x considérés isolément: le nombre d'observations pour lesquelles $x = x_i$ (resp. x appartient à la classe $[a_{i-1}, a_i[$) est

$$n_{i,\bullet} = \sum_{j=1}^{r} n_{i,j}.$$

De même, les effectifs de la distribution marginale de y se lisent dans la dernière ligne du tableau.

On définit:

- **Les moyennes marginales** $\overline{x}$ et $\overline{y}$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i,\bullet}x_i \text{ et } \overline{y} = \frac{1}{n}\sum_{j=1}^{r} n_{\bullet,j}y_j.$$

- **Les variances marginales**

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i,\bullet}(x_i-\overline{x})^2 \text{ et } \sigma_y^2 = \frac{1}{n}\sum_{j=1}^{r} n_{\bullet,j}(y_j-\overline{y})^2.$$

5.Liaisons entre variables statistiques

On définit:

- **Les moyennes marginales** $\overline{x}$ et $\overline{y}$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_{i,\bullet} x_i \text{ et } \overline{y} = \frac{1}{n}\sum_{j=1}^{r} n_{\bullet,j} y_j.$$

- **Les variances marginales**

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{p} n_{i,\bullet}(x_i - \overline{x})^2 \text{ et } \sigma_y^2 = \frac{1}{n}\sum_{j=1}^{r} n_{\bullet,j}(y_j - \overline{y})^2.$$

7.La distribution conditionnelle

A chaque valeur fixée x_i correspond la distribution conditionnelle de y sachant $x = x_i$ d'effectifs $n_{i,1}, \ldots, n_{i,r}$.
On définit:

- **La moyenne conditionnelle de y sachant** $x = x_i$:

$$\overline{y}_i = \frac{1}{n_{i,\bullet}} \sum_{j=1}^{r} n_{i,j} y_j .$$

- **La variance conditionnelle de y sachant** $x = x_i$:

$$\widehat{\sigma}_{x_i}^2(y) = \frac{1}{n_{i,\bullet}} \sum_{j=1}^{r} n_{i,j} (y_j - \overline{y}_i)^2 .$$

7.La distribution conditionnelle

A chaque valeur fixée x_i correspond la distribution conditionnelle de y sachant $x = x_i$ d'effectifs $n_{i,1}, \ldots, n_{i,r}$.
On définit:

- **La moyenne conditionnelle de y sachant $x = x_i$** :

$$\overline{y}_i = \frac{1}{n_{i,\bullet}} \sum_{j=1}^{r} n_{i,j} y_j.$$

- **La variance conditionnelle de y sachant $x = x_i$** :

$$\widehat{\sigma}^2_{x_i}(y) = \frac{1}{n_{i,\bullet}} \sum_{j=1}^{r} n_{i,j} (y_j - \overline{y}_i)^2.$$

De même on définit :

- **la moyenne conditionnelle de x sachant** $y = y_j$:

$$\overline{x}_j = \frac{1}{n_{\bullet,j}} \sum_{i=1}^{p} n_{i,j} x_i.$$

- **la variance conditionnelle de x sachant** $y = y_j$:

$$\widehat{\sigma}^2_{y_j}(x) = \frac{1}{n_{\bullet,j}} \sum_{i=1}^{p} n_{i,j} (x_i - \overline{x}_j)^2.$$

De même on définit :

- **la moyenne conditionnelle de x sachant $y = y_j$** :

$$\overline{x}_j = \frac{1}{n_{\bullet,j}} \sum_{i=1}^{p} n_{i,j} x_i.$$

- **la variance conditionnelle de x sachant $y = y_j$** :

$$\widehat{\sigma}^2_{y_j}(x) = \frac{1}{n_{\bullet,j}} \sum_{i=1}^{p} n_{i,j} (x_i - \overline{x}_j)^2.$$

6.1 Coefficient de corrélation linéaire

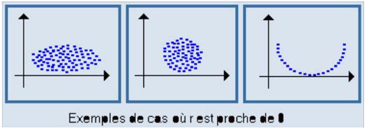
Definition

C'est un nombre sans dimension destiné à mesurer l'intensité de liaison entre les variations de x et celles de y. Il a pour expression

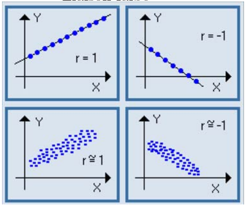
$$r = r(x,y) = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y},$$

où

$$cov(x,y) = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{r} n_{i,j}(x_i - \overline{x})(y_j - \overline{y}).$$

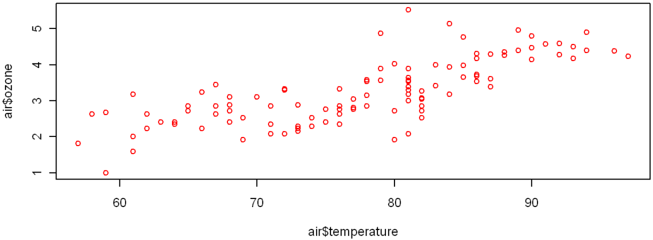


Liaison faibe



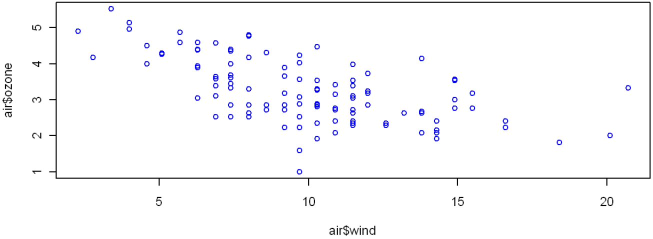
Liaison forte

6.L'ajustement, régression simple



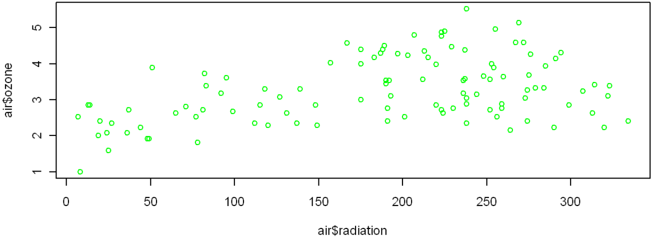
r=0.7531038

6.L'ajustement, régression simple



r=-0.5989278

6.L'ajustement, régression simple



r=0.422013

6.2 L'ajustement

Definition

Etant donné un ensemble de points de coordonnées $(x_i, y_i), 1 \leq i \leq n$ formant un nuage dans le plan XOY, l'ajustement consiste à choisir une courbe continue qui résume le nuage. Cette courbe a pour but d'expliquer comment y varie en fonction de x.

Ajustement par la méthode des moindres carrés

Le type d'ajustement étant préalablement choisi, la courbe d'ajustement de y en x, d'équation $y = f(x)$, est dite des moindres carrés si f est telle que $\sum_{i=1}^{n} d_i^2$ soit minimale où $d_i = y_i - f(x_i)$

l'ajustement linéaire: droite des moindres carrés

Lorsque le nuage est rectiligne on choisit $f(x) = ax + b$. La droite des moindres carrés est donnée par $y = \widehat{a}x + \widehat{b}$, où

$$\widehat{a} = \frac{\mathrm{cov}(x,y)}{\sigma_x^2}, \widehat{b} = \overline{y} - \widehat{a}\overline{x},$$

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}),$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \ \overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \ \overline{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$$

Qualité d'ajustement

Pour juger de la bonne qualité de l'ajustement on définit

$$R^2 = \frac{\sigma_y^2 - SSR/n}{\sigma_y^2}, SSR = \sum_{i=1}^{n} e_i^2,$$

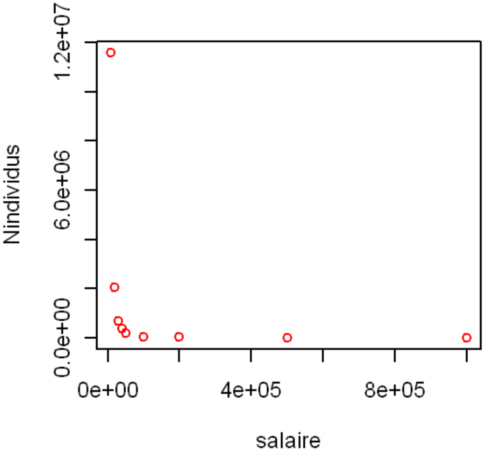
SSR est la somme des carrés des résidus, avec $e_i = y_i - (\widehat{b} + \widehat{a}x_i)$ le résidu qui correspond à l'observation y_i.

R^2 est un réel toujours inférieur à 1.

- Si R^2 est proche de 1 alors l'ajustement est de bonne qualité.
- Si R^2 est proche de 0 alors l'ajustement est de mauvaise qualité.

6.L'ajustement, régression simple

Exemple On reprend la distribution des salaires,

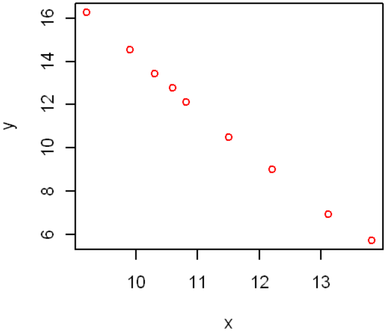


6.L'ajustement, régression simple

Le graphique nous suggère que le nombre d'individus décroît lorsque le salaire croit, le rythme de décroissance est exponentiel. On peut donc déduire que le coefficient de corrélation linéaire est négatif; en effet on trouve $r = -0.3005368$.

Pour expliquer N en fonction de s, nous allons faire une transformation logarithmique. En effet en posant $y = \ln(N)$ et $x = \ln(s)$,

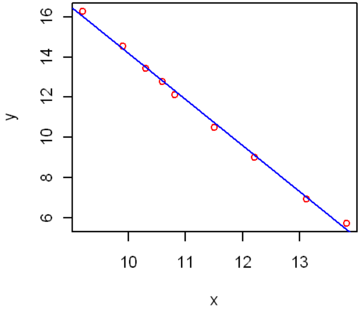
6.L'ajustement, régression simple



Le graphique montre qu'il y a une relation linéaire en y et x.

6.L'ajustement, régression simple

Le coefficient de corrélation linéaire est $r = -0.998564$. Un ajustement linéaire de y en x est bien justifié. On trouve $\widehat{b} = 37.241$ et $\widehat{a} = -2.305$. La droite des moindres carrés est donnée par $y = -2.305x + 37.241$.



La somme des carrés des résidus $SSR = 0.2012$ est très faible.

Le coefficient $R^2 = 0.9971$ est proche de 1, on peut donc affirmer que l'ajustement est de bonne qualité.

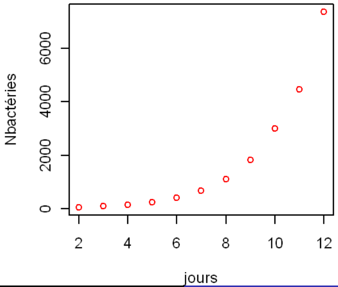
En résumé, on peut donc supposer que le nombre d'individus N est lié au salaire s par la relation

$$N(s) = e^{-2.305 \ln(s)+37.241} = 1.491286 \times (10^{16})s^{-2.305}.$$

Exemple 2. En l'absence de mortalité, on souhaite décrire l'évolution dans le temps de la croissance d'une population de bactéries. Des numérations faites tous les jours à partir du 2^e donne les résultats suivants:

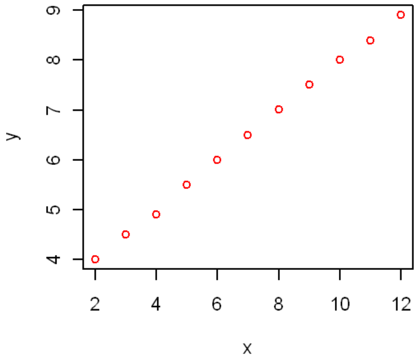
6.L'ajustement, régression simple

t_i (jours)	2	3	4	5	6	7	8	9	10	11	12
N_i (bactéries)	55	90	135	245	403	665	1100	1810	3000	4450	7350



Le nombre de bactéries croit de manière rapide (exponentielle). On peut donc déduire que le coefficient de corrélation linéaire entre le nombre de bactéries N et la variable temps t est positif; en effet on trouve $r = 0.8596725$. Pour expliquer N en fonction de t, nous allons faire une transformation logarithmique seulement de la variable N (car c'est la variable qui a des valeurs très grandes). En effet en posant $y = \ln(N)$, $x = t$ (jours) on obtient

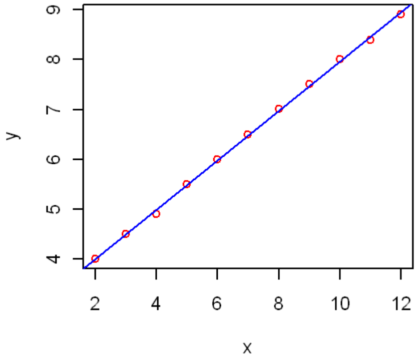
6.L'ajustement, régression simple



Le coefficient de corrélation linéaire est $r = 0.9916918$. Un ajustement linéaire de y en x est bien justifié.

6.L'ajustement, régression simple

On trouve $\widehat{b} = 3.014$ et $\widehat{a} = 0.494$. La droite des moindres carrés est donnée par $y = 0.494x + 3.014$.



La somme des carrés des résidus $SSR = 0.04499$ est très faible.

Le coefficient $R^2 = 0.9993$ est très proche de 1, on peut donc affirmer que l'ajustement est de très bonne qualité.

En résumé, on déduit que l'évolution du nombre de bactéries en fonction des jours suit l'équation

$$N(t) = e^{0.494t+3.014} = 20.36871e^{0.494t}.$$

L'estimation a pour but de chercher, à travers l'examen d'un échantillon, une information quantitative à propos d'un paramètre.

Definition

L' estimation est dite ponctuelle lorsque l'on se propose de substituer à la valeur θ un nombre unique $\hat{\theta}$ construit à partir d'un échantillonnage.

Exemple: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est un estimateur ponctuel de la moyenne m.

Definition

L' estimation est dite par intervalle de confiance lorsque l'on se propose de construire à partir de l'échantillon un intervalle $[a, b]$ qui peut contenir θ avec une probabilité $1 - \alpha$

$$P(a \leq \theta \leq b) = 1 - \alpha,$$

α fixé à l'avance. On dit alors que $[a, b]$ est un intervalle de confiance au niveau α pour θ.

7.1 Estimation ponctuelle des paramètres

Estimateur d'une proportion

Soit une population de taille N constituée de deux catégories A (atteints d'une maladie M) et B (non atteints de la maladie M) d'individus dans des proportions p et q respectivement, avec p et q sont inconnues. Comme $p + q = 1$ il n'y a qu'un seul paramètre pour caractériser la structure de cette population. On identifie par exemple p. Pour cela on prélève dans la population un échantillon aléatoire de taille n; si n_1 est le nombre d'individu présentant le caractère A. $f = \frac{n_1}{n}$ est la proportion du caractère A dans l'échantillon (ou effectif relatif).

7.Estimation statistique

L'entier n_1 est l'observation d'une variable aléatoire X et f est l'observation de la variable aléatoire $Y = \frac{X}{n}$.On suppose l'échantillon non exhaustive (tirage avec remise), donc X suit une loi binomiale $B(n, p)$.

Donc

$$
\begin{aligned}
E(Y) &= p \\
\sigma(Y) &= \sqrt{\frac{p(1-p)}{n}}.
\end{aligned}
$$

C'est à dire que la variable aléatoire fluctue autour de p avec un écart type qui tend vers 0 lorsque n tend vers l'infini.

Dans ce cas on dit que Y est un estimateur consistant de p et que l'observation f de Y est une estimation consistante de p et on écrit

$$\widehat{p} = f = \frac{n_1}{n}.$$

Remarque. Si l'échantillon est exhaustive (tirage sans remise) on montre que

$$\begin{aligned} E(Y) &= p \\ \sigma(Y) &= \sqrt{\frac{N-n}{N-1}\frac{p(1-p)}{n}}. \end{aligned}$$

7.2 Estimation par intervalle de confiance

Soit $\hat{\theta}$ un estimateur de θ, la construction d'un intervalle de confiance est basée sur un **pivot** $pv = f(\hat{\theta})$, dont la loi ne dépend pas de θ.

1.Intervalle de confiance d'une proportion

On suppose que n$\geq$ 100 et que p n'est pas proche de 0 ni de 1 $(0.1 \leq p \leq 0.9)$. Dans ce cas X est assimilée à une variable de Gauss, il en est de même pour Y.

Les deux variables

$$Z = \frac{Y-p}{\sqrt{\frac{p(1-p)}{n}}}. \text{ et } T = \frac{Y-p}{\sqrt{\frac{f(1-f)}{n}}}.$$

sont assimilées à une loi centré réduite de Gauss N(0,1).

On fixe α (0.01 ou 0.05) et on lit dans la table de Gauss $u_{1-\alpha/2}$ tel que $P(|N(0,1)| \leq u_{1-\alpha/2}) = 1-\alpha$, c'est dire que

$P(|T| \leq u_{1-\alpha/2}) = 1-\alpha$ ou encore

$$P\left(Y - u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \leq p \leq Y + u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}\right) = 1-\alpha$$

Par conséquent l'intervalle de confiance de p au niveau de signification α est donné par

$$I_p = \left[f - u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}}, f + u_{1-\alpha/2}\sqrt{\frac{f(1-f)}{n}} \right].$$

2.Intervalle de confiance pour la moyenne d'une population gaussienne

On suppose que $X_i \rightsquigarrow N(m, \sigma^2), 1 \leq i \leq n$.

•σ connu

Le pivot est donné par $pv = \sqrt{n}\frac{\overline{X}-m}{\sigma}$, avec $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightsquigarrow N(m, \sigma^2/n)$, donc $pv \rightsquigarrow N(0,1)$.

On lit alors dans la table de Gauss le quantile $u_{1-\frac{\alpha}{2}}$ tel que $P(N(0,1) \in [-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}]) = 1-\alpha$ donc $P(pv \in [-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}]) = 1-\alpha$, par conséquent $P(m \in [\overline{X} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]) = 1-\alpha$.

Donc l' intervalle de confiance au niveau α pour m est donné par

$$I_\alpha = \left[\overline{x} - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]. \qquad (22)$$

$\bullet\sigma$ inconnu

L'intervalle ci-dessus est inutilisable car il dépend de σ, donc il faut estimer ce dernier, le pivot est donné par $pv = \sqrt{n}\frac{\overline{X}-m}{S}$, où $S^2 = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$, qui est un estimateur empirique et sans biais de la variance σ^2.
On a $\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \mathcal{X}^2(n-1)$, et donc $pv \rightsquigarrow t(n-1)$, loi de Student à (n-1) degrés de liberté. On lit alors dans la table de Student $t_{1-\frac{\alpha}{2}}$ tel que $P(t(n-1) \in [-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]) = 1-\alpha$;

l'intervalle de confiance au niveau α pour m est alors donné par

$$I'_\alpha = \left[\overline{x} - t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \overline{x} + t_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right]. \qquad (23)$$

3.Intervalle de confiance pour la variance d'une population gaussienne

On a le pivot $pv = \frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \mathcal{X}^2(n-1)$, loi de Khi-deux à (n-1) degrés de liberté. On lit alors dans la table de Khi-deux x_a et x_b tels que $P(x_a \leq \mathcal{X}^2(n-1) \leq x_b) = 1-\alpha$.

L'intervalle de confiance au niveau α pour σ^2 est donné par

$$I_v = \left[\frac{(n-1)s^2}{x_b}, \frac{(n-1)s^2}{x_a}\right].$$

Definition

Les tests d'hypothèses servent à valider une caractéristique d'une population à travers l'examen d'un échantillon, ils sont aussi utilisés pour comparer des caractéristiques entre plusieurs populations.

8.1 Test de conformité

1. Test de conformité d'une proportion: Il s'agit de comparer une proportion théorique p_0 à une proportion observée. On effectue le test suivant

$$\begin{cases} H_0 : \quad p_0 \text{ est la proportion des individus de la population} \\ \quad \text{de la catégorie A} \\ \text{contre} \\ H_1 : \text{ non } H_0 \end{cases}$$

ou encore

$$\begin{cases} H_0 : & p = p_0 \\ \text{contre} & \\ H_1 : & p \neq p_0 \end{cases}$$

On prélève un échantillon aléatoire non exhaustive de taille n dans lequel la proportion des individus de la catégorie A est $f = \frac{n_1}{n}$.

Sous H_0 : $\frac{Y - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}.} \rightsquigarrow N(0, 1)$.

On fixe α (0.01 ou 0.05) et on lit dans la table de Gauss $u_{1-\alpha/2}$ tel que

$$P(|N(0,1)| > u_{1-\alpha/2}) = \alpha.$$

l'intervalle critique au seuil α est donné par

$$I_c =]-\infty, p_0 - u_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}] \bigcup [p_0 + u_{1-\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}, +\infty[$$

Décision : si $f \in I_c$ on rejette H_0.

2. Test de conformité d'une moyenne

On attribue la valeur m_0 pour moyenne du caractère dans une population et on veut juger le bien fondé de cette hypothèse en se basant sur un échantillon. On effectue alors le test d'hypothèse:

$$\begin{cases} H_0 : \quad m = m_0 \\ \text{contre} \\ H_1 : \quad m \neq m_0 \end{cases}$$

où m est la moyenne théorique.

On considère alors $\sqrt{n}\frac{\overline{X}-m_0}{S}$ qui suit, sous H_0, une loi de Student à (n-1) degrés de liberté.

Soit le quantile $t_{1-\frac{\alpha}{2}}$ tel que
$P(t(n-1) \in [-t_{1-\frac{\alpha}{2}}, t_{1-\frac{\alpha}{2}}]) = 1 - \alpha$.

Si $\left|\sqrt{n}\frac{\overline{X}-m_0}{S}\right| \leq t_{1-\frac{\alpha}{2}}$ alors on ne peut pas refuser H_0 au seuil de signification α.

3. Test de conformité d'une variance

C'est la comparaison d'une variance théorique et d'une variance observée: on attribue la valeur σ_0^2 pour variance du caractère dans une population et on veut juger le bien fondé de cette hypothèse en se basant sur un échantillon. On effectue alors le test d'hypothèse:

$$\left\{\begin{array}{ll} H_0 : & \text{la variance du caractère dans la population est } \sigma_0^2 \\ \text{contre} & \\ H_1 : & \text{non } H_0. \end{array}\right.$$

Sous H_0, $\frac{(n-1)S^2}{\sigma_0^2} \rightsquigarrow \mathcal{X}^2(n-1)$ donc

si $\frac{(n-1)S^2}{\sigma_0^2} \in [x_a, x_b]$ alors on ne peut pas refuser l'hypothèse H_0.

8.2 Tests d'homogénéité

1. Comparaison de deux proportions:

Il s'agit de comparer deux proportions observées. Soit deux échantillons aléatoire E_1 et E_2 extraits des deux populations P_1 et P_2 constituées des individus des deux seules catégories A et B. Soit $f_1 = \frac{k_1}{n_1}$ et $f_2 = \frac{k_2}{n_2}$ les proportions des individus de la catégorie A dans E_1 et E_2 respectivement. En général $f_1 \neq f_2$; il s'agit d'expliquer cette différence.

On effectue le test d'hypothèse suivant:

$$\begin{cases} H_0 : \text{Dans } P_1 \text{ et } P_2 \text{ il y a la même proportion } \mathrm{p}_0 \\ \quad (\text{inconnue}) \text{ d'individus de la catégorie A} \\ \text{contre} \\ H_1 : \text{ non } H_0 \end{cases}$$

On considère $D = Y_1 - Y_2$, avec f_i est l'observation de la variable aléatoire Y_i.

Sous H_0 $\frac{Y - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}.} D \rightsquigarrow N(0, \sigma_D^2)$ où

$$\sigma_D^2 = p_0(1 - p_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right). \tag{24}$$

Mais p_0 est inconnue et on peut l'estimer par

$$\frac{k_1 + k_2}{n_1 + n_2} = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2},$$

ainsi σ_D^2 est estimée par

$$\begin{aligned}\widehat{\sigma}_D^2 &= \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}\left(1 - \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \\ &= \frac{(n_1 f_1 + n_2 f_2)\left(n_1(1 - f_1) + n_2(1 - f_2)\right)}{(n_1 + n_2) n_1 n_2}.\end{aligned}$$

On fixe α (0.01 ou 0.05) et on lit dans la table de Gauss $u_{1-\alpha/2}$ tel que $P(|N(0,1)| > u_{1-\alpha/2}) = \alpha$;

Décision :Si $\frac{|D|}{\widehat{\sigma}_D} = \frac{|f_1 - f_2|}{\widehat{\sigma}_D} \geq u_{1-\alpha/2}$ on rejette H_0.

2.Comparaison des moyennes de deux échantillons indépendants

Soient deux échantillons aléatoires d'observations provenant de deux populations P_1 et P_2 et de moyenne $\overline{X}$ et $\overline{Y}$ respectivement. En général $\overline{X} \neq \overline{Y}$,il s'agit d'expliquer cette différence.

On suppose que $X \rightsquigarrow N(m_x, \sigma_x^2)$, $Y \rightsquigarrow N(m_y, \sigma_y^2)$, X et Y sont deux échantillons indépendants, on désire effectuer le test

$$\begin{cases} H_0 : \quad m_x = m_y \\ \text{contre} \\ H_1 : \quad m_x \neq m_y. \end{cases}$$

On utilise $D = \overline{X} - \overline{Y}$, qui est un bon indicateur de l'écart entre les deux moyennes.

- Cas σ_x^2 et σ_y^2 connues:

Sous H_0 : $\dfrac{D}{\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}} \rightsquigarrow N(0,1)$, on fixe α (0.01 ou 0.05) et on lit dans la table de Gauss $u_{1-\alpha/2}$ tel que $P(|N(0,1)| > u_{1-\alpha/2}) = \alpha$, l'intervalle critique au seuil α est donné par

$$I_m =]-\infty, -u_{1-\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}]\bigcup[u_{1-\alpha/2}\sqrt{\frac{\sigma_x^2}{n_x}+\frac{\sigma_y^2}{n_y}}, +\infty[,$$

décision : si $d = \overline{x} - \overline{y} \in I_m$ on rejette H_0.

- Cas σ_x^2 et σ_y^2 inconnues:

On suppose que les échantillons sont grands $(n_x \geq 30$ et $n_y \geq 30)$: On peut remplacer σ_x^2 (resp . σ_y^2) par $S_x^2 = \frac{1}{n_x-1}\sum_1^{n_x}(X_i - \overline{X})^2$ (resp. S_y^2), et donc l'intervalle critique au seuil α est donné par

$$I_m =]-\infty, -u_{1-\alpha/2}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}] \bigcup [u_{1-\alpha/2}\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}, \infty[, \tag{25}$$

décision : si $d = \overline{x} - \overline{y} \in I_m$ on rejette H_0.

3. Comparaison des moyennes de deux séries appariées

Soient deux séries de n observations $X_1, ..., X_n$ de moyenne $\overline{X}$ et $Y_1, ..., Y_n$ de moyenne $\overline{Y}$. Ici l'observation X_i est appariée à Y_i. On pose

$D_i = X_i - Y_i, S_D^2 = \frac{1}{n-1}\sum_1^n (D_i - \overline{D})^2, \overline{D} = \frac{1}{n}\sum_{i=1}^n D_i.$

Pour tester

$$\begin{cases} H_0 : \quad m_x = m_y \\ \text{contre} \\ H_1 : \quad m_x \neq m_y, \end{cases}$$

on considère $T = \sqrt{n}\frac{\overline{D}}{S_D}$.

Si n est grand, disons supérieur à 30, T est assimilée à une variable centrée réduite de Gauss; si n est petit et si D suit une loi de Gauss, T suit une loi de Student à $(n-1)$ degrés de liberté.

4. Comparaison des variances de deux échantillons indépendants

On veut tester

$$\begin{cases} H_0 : \quad \sigma_x^2 = \sigma_y^2 \\ \text{contre} \\ H_1 : \quad \sigma_x \neq \sigma_y. \end{cases}$$

On calcule les rapports $\frac{S_x^2}{S_y^2}$ et $\frac{S_y^2}{S_x^2}$ et on considère le plus grand parmi les deux:

• Si $\frac{S_x^2}{S_y^2} > \frac{S_y^2}{S_x^2}$, on décidera à l'aide du rapport $\frac{S_x^2}{S_y^2}$, comme suit: on a sous H_0 : $\frac{S_x^2}{S_y^2} \rightsquigarrow F(n_x - 1, n_y - 1)$, loi de Fisher-Snédécor à $(n_x - 1, n_y - 1)$ degrés de liberté, donc l'intervalle critique au seuil α est donné par

$$I_\alpha = [f_\alpha, \infty[, \text{ avec } P(F(n_x - 1, n_y - 1) > f_\alpha) = \alpha, \qquad (26)$$

décision : si $\frac{s_x^2}{s_y^2} \in I_\alpha$ on rejette H_0.

• Si $\frac{S_x^2}{S_y^2} \leq \frac{S_y^2}{S_x^2}$, on décidera à l'aide du rapport $\frac{S_y^2}{S_x^2}$, comme suit: on a sous H_0 : $\frac{S_y^2}{S_x^2} \rightsquigarrow F(n_y - 1, n_x - 1)$, loi de Fisher-Snédécor à $(n_y - 1, n_x - 1)$ degrés de liberté, donc l'intervalle critique au seuil α est donné par

$$I_\alpha = [f_\alpha, +\infty[, \text{ avec } P(F(n_y - 1, n_x - 1) > f_\alpha) = \alpha,$$

décision : si $\frac{s_y^2}{s_x^2} \in I_\alpha$ on rejette H_0.