Table des matières

Chapitre 1

Rappels sur les lois usuelles de probabilités

Une loi de probabilité est une fonction mathématique qui modélise théoriquement une expérience aléatoire. En biologie, ces lois jouent un rôle clé en permettant de quantifier et de prédire la variabilité des processus biologiques. Elles offrent aux biologistes les outils nécessaires pour analyser des données, poser des hypothèses et prendre des décisions éclairées. Par conséquent, ces lois contribuent à une meilleure compréhension des phénomènes aléatoires observés dans la nature.

1.1 Variable aléatoire

Définition 1.1.1

Une variable aléatoire X est une application de l'ensemble fondamental Ω dans $\mathbb{R}$, définie comme suit

$$\begin{array}{rccc} X: & \Omega & \to & \mathbb{R} \\ & \omega & \mapsto & X(\omega) \end{array}$$

On distingue deux types de variables aléatoires :

1.1.1 Variable aléatoire discrète

Définition 1.1.2

Une variable aléatoire X est dite discrète si elle peut prendre un nombre fini de valeurs isolées.

Loi de probabilité

Soit X une variable aléatoire sur Ω, $X(\Omega) = \{x_1, x_2, \ldots, x_n\}$. La loi de probabilité de X est donnée par

x_i	x_1	x_2	$\ldots$	x_n
$p(X = x_i)$	$p(X = x_1)$	$p(X = x_2)$	$\ldots$	$p(X = x_n)$

Fonction de répartition

La fonction de répartition de la variable aléatoire X est donnée par

$$F_X(x) = p(X \leq x) = \sum_{x_i \leq x} p(X = x_i) = \begin{cases} 0 & \text{si} \quad x < x_1 \\ p_1 & \text{si} \quad x_1 \leq x < x_2 \\ p_1 + p_2 & \text{si} \quad x_2 \leq x < x_3 \\ \vdots & \vdots \quad \vdots \\ 1 & \text{si} \quad x \geq x_n \end{cases}$$

Espérance mathématique

L'espérance mathématique de la v.a X est donnée par

$$E(X) = \sum_{i=1}^{n} x_i p(X = x_i).$$

Variance et écart-type

La variance de la v.a X est donnée par

$$V(X) = E\left[(X - E(X))^2\right] = \sum_{i=1}^{n} (x_i - E(X))^2 p(X = x_i),$$

ou

$$V(X) = E\left(X^2\right) - (E(X))^2 = \sum_{i=1}^{n} x_i^2 p(X = x_i) - \left(\sum_{i=1}^{n} x_i p(X = x_i)\right)^2.$$

L'écart-type de la v.a X est donné par

$$\sigma_X = \sqrt{V(X)}.$$

Exemple 1.1.1

Soit l'expérience aléatoire suivante

"on lance un dé à six faces et on regarde le résultat".

On considère le jeu suivant

— *Si le résultat est pair, on gagne* 2 *DA.*

— *Si le résultat est* 1*, on gagne* 3 *DA.*

— *Si le résultat est* 3 *ou* 5*, on perd* 4 *DA.*

La variable aléatoire X représente le gain à ce jeu.

— *L'ensemble fondamentale* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

— $X\{\Omega\} = \{-4, 2, 3\}$.

— *Déterminons la loi de probabilité de X*

x_i	-4	2	3
$p(X = x_i)$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{1}{6}$

— *Déterminons la fonction de répartition de X*

$$F_X(x) = p(X \leq x) = \sum_{x_i \leq x} p(X = x_i) = \begin{cases} 0 & si \quad x < -4 \\ \frac{2}{6} & si \quad -4 \leq x < 2 \\ \frac{5}{6} & si \quad 2 \leq x < 3 \\ 1 & si \quad x \geq 3 \end{cases}$$

— *Calculons l'espérance mathématique* $E(X)$

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} x_i p(X = x_i) \\ &= -4\left(\frac{2}{6}\right) + 2\left(\frac{3}{6}\right) + 3\left(\frac{1}{6}\right) = \frac{1}{6}. \end{aligned}$$

— $V(X) = E(X^2) - (E(X))^2 = 8.8$ *et* $\sigma_X = \sqrt{V(X)} = 2.97$.

1.1.2 Variable aléatoire continue

Définition 1.1.3

Une variable aléatoire X est continue si elle peut prendre toutes les valeurs comprises dans un intervalle.

Densité de probabilité

Soit X une variable aléatoire continue. On dit que $f : \mathbb{R} \to \mathbb{R}^+$ *est une densité de probabilité de X si*

— $f(x) \geq 0, \forall x \in \mathbb{R}$.

— f *est continue sur* $\mathbb{R}$.

— $\int_{-\infty}^{+\infty} f(x)dx = 1$.

Fonction de répartition

La fonction de répartition d'une variable aléatoire continue X est définie par

$$F_X(x) = p(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Propriété 1.1.1

— $F_X(x)$ *est positive.*

— $\lim_{x \to -\infty} F_X(x) = 0$ *et* $\lim_{x \to \infty} F_X(x) = 1$.

Remarque 1.1.1

— $p(X = x) = 0$.

— $p(X \leq x) = p(X < x)$.

— *la probabilité de la v.a* $X \in [a, b]$ *est donnée par*

$$\begin{aligned} p(a \leq X \leq b) &= p(a < X \leq b) \\ &= p(a \leq X < b) \\ &= p(a < X < b) \\ &= \int_a^b f(t)dt \\ &= F_X(b) - F_X(a). \end{aligned}$$

Espérance mathématique

L'espérance mathématique d'une variable aléatoire continue X est donnée par

$$E(X) = \int_{-\infty}^{+\infty} x f(x)dx.$$

Variance et écart-type

La variance de la v.a.c X est donnée par

$$V(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx,$$

ou

$$V(X) = E\left(X^2\right) - (E(X))^2.$$

1.2 Lois usuelles de probabilités

1.2.1 Lois discrètes

Loi de Bernoulli

• *Une expérience aléatoire ayant deux résultats possibles, succès et échec, est appelée une expérience de Bernoulli.*

• *La loi de probabilité associée est la suivante*

la variable aléatoire $X = 1$ *en cas de succès avec la probabilité* p*, et* $X = 0$ *en cas d'échec avec la probabilité* $q = 1 - p$

$$p(X = x) = \begin{cases} p & si \quad x = 1, \\ q & si \quad x = 0. \end{cases}$$

• *Notation*

$$X \sim B(p).$$

• $E(X) = p$.

• $V(X) = pq$.

• $\sigma_X = \sqrt{V(X)} = \sqrt{pq}$.

Exemple 1.2.1

On lance une pièce de monnaie bien équilibrée en air. Si on obtient face, il y a succès.

- *E.a est une expérience de Bernoulli*

$$\Omega = \{face, pile\}, \ A = \{face\}, \ \overline{A} = \{pile\}.$$

- *Loi de probabilité*

La probabilité de succès est $p(A) = \frac{1}{2}$ *et celle d'échec est* $p\left(\overline{A}\right) = \frac{1}{2}$.

La loi de probabilité est donc

$$\begin{cases} \frac{1}{2} & si \quad x = face, \\ \frac{1}{2} & si \quad x = pile. \end{cases}$$

- $E(X) = p = \frac{1}{2}$.
- $V(X) = pq = \frac{1}{4}$.
- $\sigma_X = \sqrt{V(X)} = \frac{1}{2}$.

Loi Binomiale

- *La loi binomiale de paramètre n et p modélise le nombre de succès obtenus lors de la répétition de n expériences de Bernoulli identiques et indépendantes.*

- *La probabilité de k succès est donnée par la formule*

$$p(X = k) = C_n^k p^k q^{n-k}; \ k = 1, 2, \ldots, n; \ C_n^k = \frac{n!}{k!(n-k)!}.$$

- *Notation*

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

- $E(X) = np$.
- $V(X) = npq$.
- $\sigma_X = \sqrt{V(X)}$.

Exemple 1.2.2

On jette un dé équilibré 5 *fois et on s'intéressera au résultat "avoir le chiffre* 2*".*

— *Quelle est la probabilité d'obtenir deux fois le chiffre* 2?

— *Quelle est la probabilité d'obtenir au moins trois fois le chiffre* 2?

— *Calculer* $E(X)$, $V(X)$ *et* σ_X.

Solution

La variable aléatoire X : "avoir le chiffre 2*", suit une loi binomiale* $X \sim B(5,p)$.

$$\Omega = \{1,2,3,4,5,6\}, \quad A = \{2\}, \quad \overline{A} = \{1,3,4,5,6\}$$

La probabilité de succès $p = p(A) = \frac{1}{6}$.

La probabilité d'échec $q = p\left(\overline{A}\right) = \frac{5}{6}$.

— *La probabilité d'obtenir deux fois le chiffre* 2

$$p(X=2) = C_5^2\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{5-2} = \frac{5!}{2!(5-2)!}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.16.$$

— *La probabilité d'obtenir au moins trois fois le chiffre* 2 *est*

$$\begin{aligned} p(X \geq 3) &= 1 - p(X < 3) \\ &= 1 - (p(X=0) + p(X=1) + p(X=2)) \\ &= 1 - (0.4 + 0.4 + 0.16) = 0.036. \end{aligned}$$

— *Les valeurs de l'espérance, la variance et l'écart-type sont*

$$E(X) = np = \frac{5}{6},\ \ V(X) = npq = \frac{25}{36},\ \ \sigma_X \sqrt{V(X)} = \frac{5}{6}.$$

Loi de Poisson

• *La loi de Poisson modélise les événements rares, c'est-à-dire les événements ayant une faible probabilité de réalisation.*

• *Loi de probabilité*

$$p(X = k) = e^{-\lambda}\frac{\lambda^k}{k!},\ \ \lambda > 0.$$

• *Notation*

$$X \sim P(\lambda)$$

• $E(X) = \lambda$.

• $V(X) = \lambda$.

• $\sigma_X = \sqrt{V(X)} = \sqrt{\lambda}$.

Exemple 1.2.3

Une centrale téléphonique reçoit en moyenne 5 *appels par minute. Quelle est la probabilité que la centrale reçoive exactement deux appels*?

La variable aléatoire X représente le nombre d'appels dans la centrale , $X \sim P(5)$,

$$p(X = 2) = e^{-\lambda}\frac{\lambda^k}{k!},\ \ = e^{-5}\frac{5^2}{2!} = 0.084.$$

1.2.2 Lois continues

Loi Normale

• *Une variable aléatoire X suit une loi normale ou loi de Gauss-Laplace de paramètres m et σ si sa densité de probabilité est donnée par*

$$f(x) = \frac{1}{\sqrt{2\Pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}.$$

• *Notation*

$$X \sim N(m, \sigma).$$

Loi log normale

• *Une variable aléatoire X suit une loi log normale si* $\ln(X)$ *suit une loi normale* $N(m, \sigma)$,

$$f(x) = \frac{1}{x\sqrt{2\Pi}\sigma} e^{-\frac{1}{2}\left(\frac{\ln(x)-m}{\sigma}\right)^2}, \quad x > 0.$$

• *Notation*

$$X \sim LN(m, \sigma).$$

Loi normale centrée réduite

• *Une loi normale centrée réduite est notée* $N(0, 1)$.

$$f(x) = \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}x^2}.$$

• *Si* $X \sim N(m, \sigma)$, *alors* $Y = \frac{X-m}{\sigma} \sim N(0, 1)$.

Loi de Khi-deux

• *Soient $X_1, X_2, \ldots, X_n$ n variables aléatoires indépendantes $\sim N(0,1)$.*

Alors $Y = X_1^2 + X_2^2 + \ldots + X_n^2$ suit une loi de Khi-deux à n degrés de liberté.

$$f(y) = \frac{y^{\frac{n}{2}-1} e^{\frac{-y}{2}}}{2^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)},$$

avec

$$\Gamma(n) = \int_0^{+\infty} e^x x^{n-1} dx.$$

• *Notation*

$$Y \sim \chi_n^2.$$

• $E(Y) = n$.

• $V(Y) = 2n$.

Loi de Student

Soient X et Y deux variables aléatoires indépendantes telles que $X \sim N(0,1)$ et $Y \sim \chi_n^2$. Alors, $T = \dfrac{X}{\sqrt{\frac{Y}{n}}}$ suit une loi de Student à n degrés de liberté.

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

• *Notation*

$$T \sim t_n.$$

• $E(T) = 0, n > 1$.

• $V(T) = \dfrac{n}{n-2}, n > 2$.

Chapitre 2

Échantillonnage et estimation

2.1 Échantillonnage

2.1.1 Notion d'échantillonnage

Définition 2.1.1

On considère une population Ω de taille N. On appelle ***échantillon*** *un sous-ensemble de cette population. Un échantillon de taille n est donc une liste de n individus $(\omega_1, \omega_2, \ldots, \omega_n)$ extraits de la population mère.*

Exemple 2.1.1

On considère une population constituée de 5 étudiants et on s'intéresse au temps hebdomadaire consacré par chaque étudiant à l'étude des statistiques. $\Omega = \{A, B, C, D, E\}$ et $N = 5$.

Étudiant	*Temps d'étude(h)*
A	7
B	3
C	6
D	10
E	4

Définition 2.1.2

On appelle ***échantillonnage*** *le prélèvement d'échantillons. Le rapport t de l'effectif n de l'échantillon sur l'effectif N de la population dans laquelle il a été prélevé, est appelé* ***taux d'échantillonnage*** *ou* ***fraction de sondage*** *i.e.t* $= \frac{n}{N}$.

Exemple 2.1.2

On prélève des échantillons de taille 2, *alors on aura* $t = \frac{2}{5}$.

Définition 2.1.3

On appelle ***échantillonnage aléatoire*** *un prélèvement de n individus dans une population mère tel que toutes les combinaisons possible de n individus aient la même probabilités d'être prélevés.*

Il existe d'autres formes d'échantillonnage, on ne s'intéressera néanmoins qu'à des échantillonnage aléeatoires.

Remarque 2.1.1

On cherche à décrire un caractère C qualitatif ou quantitatif présent dans une population Ω *à travers l'étude des résultats obtenus sur un échantillon de taille n.*

Exemple 2.1.3

1. *Étant donnée une population, on peut s'intéresser aux caractère quantitatif tels que le poids, la taille,...etc.*
2. *Étant donnée une population, on peut s'intéresser aux caractère qualitatif tels que la couleur des yeux, la couleur des cheveux,...etc*
3. *Le caractère étudié dans l'exemple initial est le temps hebdomadaire consacré à l'étude des statistiques.*

Définition 2.1.4

Soit C un caractère quantitatif défini sur une population mère Ω. *C est la réalisation d'une*

variable aléatoire X définie sur Ω :

$$X : \Omega \to \mathbb{R}$$
$$\omega_i \mapsto X(\omega_i) = x_i$$

*On appelle $n-$**échantillon de valeur de** X la liste des valeurs $(x_1, x_2, \ldots, x_n)$ observées prises par X sur un échantillons $(w_1, w_2, \ldots, w_n)$ de la population Ω. Les coordonnées peuvent être considérées comme les valeurs des réalisations d'un vecteur de variables aléatoires $(X_1, X_2, \ldots, X_n)$ appelé $n-$**échantillon de** X où les X_i sont de même loi, indépendantes.*

Définition 2.1.5

*On appelle **statistique** toute variable aléatoire qui s'écrit à l'aide des variables aléatoires $X_1, X_2, \ldots, X_n$.*

Exemple 2.1.4

X_i , $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ *sont des statistiques.*

*Si on extrait plusieurs échantillons de taille n fixée, les résultats que l'on va pouvoir déduire sont variables car ils dépendent de l'échantillon considéré. On parle de **fluctuations d'échantillonnage**. Comment dans ce cas tirer des conclusions valables sur la population mère ? On va pour cela étudier les loi de probabilité qui régissent ces fluctuations.*

2.1.2 Distributions d'échantillonnage

Moyenne d'échantillon - Variance d'échantillon

Définition 2.1.6

On considère une population Ω dont les éléments possèdent un caractère quantitatif C qui est la réalisation d'une variable aléatoire X qui suit une loi de probabilité d'espérance μ et d'écart type σ. On suppose que la famille est de taille infinie ou que l'échantillonnage se fait avec remise.

On prélève un échantillon $(X_1, X_2, \ldots, X_n)$ *de* X *de valeurs* $(x_1, x_2, \ldots, x_n)$*. La moyenne* $\bar{x}$ *de l'échantillon est donnée par*

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

il s'agit de la valeur prise par la variable aléatoire

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La variable aléatoire $\overline{X}$ *appelé* ***moyenne d'échantillon****.*

De la même manière la variance v de l'échantillon $(x_1, x_2, \ldots, x_n)$ *est donnée par*

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

Il s'agit de la valeur prise par la variable aléatoire

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

On définit la variable aléatoire S^2*, appelé* ***variance d'échantillon****, par*

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

Paramètre descriptifs de la distribution

Proposition 2.1.1

(1) *Quelle que soit la loi de X, on a*

a) $E\left(\overline{X}\right) = \mu.$

b) $E(V) = \frac{n-1}{1}\sigma^2.$

d) $Var\left(\overline{X}\right) = \dfrac{\sigma^2}{n}$.

e) $E\left(S^2\right) = \sigma^2$.

(2) *Si* $X \sim N(\mu, \sigma)$, *on a*

a) *si* σ *est connu, alors* $\dfrac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$.

b) *si* σ *est inconnu, alors* $\dfrac{\overline{X} - \mu}{\sqrt{\frac{V}{n}}} \sim T_{n-1}$.

c) $\dfrac{nV}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

2.1.3 Proportion d'échantillon

Il arrive que le caractère à estimer ne soit pas quantitatif mais qualitatif. Dans ce cas, on recherche la proportion p des individus présentant ce caractère.

La proportion p sera estimée à l'aide des résultats obtenus sur un n−échantillon.

Définition 2.1.7

La proportion f obtenue dans un n−échantillon est la valeur observée d'une variable aléatoire F, fréquence d'apparition de ce caractère dans un échantillon de taille n, appelée ***proportion d'échantillon*** *ou* ***fréquence statistique****.*

On peut écrire

$$F = \frac{K}{n},$$

où K est la variable aléatoire qui compte le nombre d'apparitions du caractère considéré dans un échantillon de taille n.

Par définition, $K \sim B(n,p)$. *Soit*

$$E(K) = np, \quad Var(K) = npq.$$

t.q $q = 1 - p$.

Proposition 2.1.2

$E(F) = p$ *et* $Var(K) = \frac{pq}{n}$.

Remarque 2.1.2

Pour $n \geq 30$, $np \geq 15$ *et* $nq \geq 15$, *on peut approcher F par une loi normale* $N\left(p, \sqrt{\frac{pq}{n}}\right)$.

2.2 Estimation

2.2.1 Les estimateurs

Estimer un paramètre c'est en recherche une valeur approchée à partir des résultats obtenus sur un échantillon.

Exemple 2.2.1

Estimer la taille moyenne d'une population à partir de la moyenne empirique obtenue sur un échantillon de cette population.

Définition 2.2.1

un ***estimateur*** $\bar{\theta}$ *du paramètre inconnu* θ *est une fonction qui fait correspondre à une suite d'observations une valeur approchée* $\bar{\theta}$ *de* θ, *appelée estimation*

$$\bar{\theta} : (x_1, x_2, \ldots, x_n) \mapsto \bar{\theta} = f(x_1, x_2, \ldots, x_n).$$

Un estimateur $\bar{\theta}$ *est donc une variable aléatoire, on peut en calculer son espérance* $E(\bar{\theta})$ *et sa variance* $Var(\bar{\theta})$. *Ces quantité vont permettre de déterminer la qualité d'un estimateur du paramètre* θ *à estimer. Un paramètre peut en effet avoir plusieurs estimateurs. Dans le cas de la taille moyenne d'une population, on peut choisir la moyenne arithmétique, la médiane, etc*

Définition 2.2.2

On dit que $\bar{\theta}$ *est un* ***estimateur sans biais*** *si la moyenne de sa distribution d'échantillonnage*

est égale à la valeur θ du paramètre à estimer

$$E(\bar{\theta}) = \theta.$$

Sinon, on parle d'estimateur biais.

Pour comparer les estimateurs biaisés, on introduit le ***biais*** *d'un estimateur $\bar{\theta}$, qui est donné par*

$$Biais(\bar{\theta}) = E(\bar{\theta}) - \theta.$$

Remarque 2.2.1

L'absence de biais n'est pas suffisante pour assurer de l'efficacité d'un estimateur. Le paramètre θ peut d'ailleurs présenter plusieurs estimateurs sans biais. Dans ce cas, c'est la variance des estimateurs qui permet de les comparer. Si cette variance est élevée, l'estimateur peut prendre des valeurs très éloignées de la valeur effective du paramètre θ.

Définition 2.2.3

On dit qu'un estimateur sans biais est ***efficace*** *si sa variance est la plus petite parmi les variance des estimateurs sans biais. Si $\bar{\theta}_1$ est un estimateur sans biais de θ, on dit que $\bar{\theta}_1$ est efficace si pour tout estimateur sans biais $\bar{\theta}_2$*

$$E(\bar{\theta}_1) = E(\bar{\theta}_2),$$

et

$$Var(\bar{\theta}_1) < Var(\bar{\theta}_2).$$

Définition 2.2.4

Un estimateur $\bar{\theta}$ est ***convergent*** *si sa distribution tend à se concentrer autour de la valeur θ à estimer, en d'autre termes si sa variance tend vers zéro lorsque la taille de l'échantillon augmente :*

$$\lim_{n\to\infty} Var(\bar{\theta}) = 0.$$

Estimateurs usuels

(*A*) *Cas d'un caractère quantitatif :*

Soit X une variable aléatoire de moyenne μ et d'écart-type σ définie sur une population mère Ω. Soit $(X_1, X_2, \ldots, X_n)$ un $n-$échantillon de X.

Propriété 2.2.1

On a les résultats suivants :

(1) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *est un estimateur sans biais et convergent de* μ*(ie.* $E\left(\overline{X}\right) = \mu$*).*

(2) $V = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ *est un estimateur biais de la variance* σ^2*.*

(3) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \frac{n}{n-1}V$ *est un estimateur sans biais et convergent de la variance* σ^2*.*

(*B*) *Cas d'un caractère qualitatif*

On considère un caractère qualitatif d'une population dont on cherche à estimer la proportion p.

Propriété 2.2.2

La proportion d'échantillon F est un estimateur sans biais et convergent de la proportion p.

2.2.2 Intervalles de confiance

Plutôt que de déterminer une valeur approchée d'un paramètre θ obtenue à l'aide d'un estimateur $\bar{\theta}$, on va rechercher un intervalle dans lequel on sait avec une probabilité satisfaisante que la valeur de θ s'y trouve.

Définition 2.2.5

Soit X une variable aléatoire dont la loi dépend d'un paramètre θ. ***Les intervalles de confiance*** *de* ***risque*** *α pour le paramètre θ, issue des différents $n-$échantillons $(x_1, x_2, \ldots, x_n)$, sont les intervalles $[a(x_1, x_2, \ldots, x_n); b(x_1, x_2, \ldots, x_n)]$ tels qu'une proportion α de ces intervalles contiennent θ.*

Remarque 2.2.2

(1) *La quantité* $1-\alpha$ *est appelée* ***niveau de confiance*** *de l'intervalle* $[a,b]$ *:*

$$P(a \leq \bar{\theta} \leq b) = 1-\alpha.$$

(2) *Dans la pratique, on ne dispose bien souvent que d'un seul échantillon qui fournit un intervalle de confiance* $[a,b]$.

(3) *Le paramètre à estimer est souvent l'espérance ou la variance dans le cas d'un caractère quantitatif, la proportion dans le cas d'un caractère qualitatif.*

Dans la suite on va rechercher des intervalle de confiance $[a,b]$ *symétriques, c'est à dire tels que :*

$$P(\bar{\theta} < a) = \frac{\alpha}{2},$$

et

$$P(\bar{\theta} > b) = \frac{\alpha}{2}.$$

On détermine ensuite les variables aléatoires A_n *et* B_n *en fonction de* $\bar{\theta}$ *telles que :*

$$P(A_n \leq \bar{\theta} \leq B_n) = 1-\alpha.$$

Un intervalle de confiance $[a,b]$ *de risque* α *pour* θ*, issu d'un* $n-$*échantillon* $(x_1, x_2, \ldots, x_n)$ *de valeurs de* X*, s'obtient alors en calculant :*

$$a = A_n(x_1, x_2, \ldots, x_n), \quad b = B_n(x_1, x_2, \ldots, x_n).$$

Intervalle de confiance pour une moyenne

On se place dans le cas où X *suit une loi normale de paramètres* μ *et* σ *ou bien dans le cas où l'on ne connait pas forcément la loi de* X *mais pour laquelle on dispose d'un échantillon de taille* $n > 30$.

Dans le premier cas $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, *dans le second cas* $\overline{X}$ *suit approximativement cette même loi.*

On considère un $n-$*échantillon* $(x_1, x_2, \ldots, x_n)$ *de valeurs de X. On note*

$$m = \frac{x_1 + x_2 + \ldots + x_n}{n},$$

et

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - m)^2 .$$

Cas.1 σ *connu*

On sait que $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, *soit encore*

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1) .$$

Donc

$$p\left(-t_{1-\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < t_{1-\frac{\alpha}{2}}\right) = 1 - \alpha,$$

alors

$$p\left(\overline{X} - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Donc

$$Ic = \left[m - t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}; m + t_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right].$$

Cas.2 σ *inconnu*

$$Ic = \left[m - t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}; m + t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}\right],$$

où $t_{1-\frac{\alpha}{2},n-1}$ *est le quantile d'ordre* $1 - \frac{\alpha}{2}$ *de la loi de Student à* $n-1$ *degrés de liberté.*

Remarque 2.2.3

Si $n > 30$, $t_{1-\frac{\alpha}{2},n-1} = t_{1-\frac{\alpha}{2}}$.

Intervalle de confiance pour une variance

On se place dans le cas où X suit une loi normale de paramètres μ et σ.

Cas.1 *μ connu*

$$Ic = \left[\frac{nv}{\chi^2_{1-\frac{\alpha}{2}}(n)}; \frac{nv}{\chi^2_{\frac{\alpha}{2}}(n)} \right],$$

où $\chi^2_{1-\frac{\alpha}{2}}(n)$ et $\chi^2_{\frac{\alpha}{2}}(n)$ sont les quantiles d'ordre $1-\frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi de chi-deux à n degrés de liberté et

$$v = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 .$$

Cas.2 *μ inconnu*

$$Ic = \left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \right],$$

où $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ et $\chi^2_{\frac{\alpha}{2}}(n-1)$ sont les quantiles d'ordre $1-\frac{\alpha}{2}$ et $\frac{\alpha}{2}$ de la loi de chi-deux à $n-1$ degrés de liberté.

Intervalle de confiance pour une proportion

On a vu dans la partie précédente que la proportion d'échantillon F peut être approchée par une loi normale $N\left(p, \sqrt{\frac{pq}{n}}\right)$ tq : $p = 1-q$.

On en déduit :

$$Ic = \left[f - t_{1-\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}; f + t_{1-\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} \right].$$

où f est la proportion de l'échantillon analysé.

Chapitre 3

Modèles linéaires

3.1 Notions de base

Observation

Lors d'une expérimentation, on peut être amené à differents mements à mesurer differents paramètres liés au phénomène ou processus étudié. Ainsi des medecins scolaires notent pour chaque élève d'une école, leur taille et leur poids. À chaque élève est associée une observation. L'ensemble des élèves est parfois appelé aussi une population et chaque élève est un individu.

Exemple 3.1.1 *On a mesuré les hauteurs en centimètres des plantes de blé d'un même champ dont on a tiré un echantillon chaque semaine. Les observations sont données dans le tableau suivant :*

âge en semaine x_i	1	2	3	4	5	6	7	8
hauteur en cm y_i	5	12	15	22	35	35	41	48

Les variables observées simultanément sont donc :

X : âge en semaines.

Y : hauteur en centimètres.

Relation fonctionnelle

Étant données deux variables statistiques X et Y, on dit qu'il existe une relation fonctionnelle de X vers Y si à chaque valeur de la variable X est associée une valeur et une seule de la variable Y.

Modèle

S'il existe une relation fonctionnelle entre les variable X et Y et si f est la fonction donnant pour toute valeur x de X la valeur correspondante y de Y, $(Y = f(x))$ on dit que f est un modèle du phénomène étudié.

Valeur observée

Étant donnée une variable statistique X on appelle valeur obsevée de cette variable, toute valeur de cette variable relevée au cours d'une observation.

Valeur théorique

S'il existe une relation fonctionnelle entre les variable X et Y de X vers Y et si f un modèle de cette relation $(Y = f(x))$. À toute valeurs x_i de X est associée une valeur $y_i = f(x_i)$. y_i est la valeur théorique ou valeur expliquée par le modèle.

Variable statistique centrée réduite

Soit X une variable statistique et $\overline{X}$ sa moyenne, σ_X son écart-type.

On appelle varible statistique centrée la varible $X - \overline{X}$.

On appelle varible statistique centrée réduite la varible $\dfrac{X - \overline{X}}{\sigma_X}$.

Varible explicative et variable expliquée

S'il existe une relation fonctionnelle entre les variable X et Y de X vers Y et si f un modèle de cette relation $(Y = f(X))$, on dit que X est la variable explicative et Y est la variable expliquée.

Nuage de points

Soit $\mathbb{R}$ un repère cartésien du plan.

Si au cours de n observations $\omega_1, \omega_2, \ldots, \omega_n$ on relève les valeurs $X(\omega_i)$ et $Y(\omega_i)$ prises par deux variables statistiques X et Y, on appelle nuage de points l'ensemble des points du plan dont les cordonnées dans le repère $\mathbb{R}$ sont les couples $(X(\omega_i), Y(\omega_i))$.

Point moyen

On appelle point moyen d'un nuage de points, le point dont les cordonnées sont les moyennes arithmetiques des cordonnées des points du nuage.

3.2 L'ajustement linéaires

Principe d'ajustement

Lorsqu'on procède à un ajustement, on recherche à partir des observations une fonction notée f telle que les deux variables X et Y soient liées par la relation $Y = f(X)$.

Intérêt de l'ajustement

Connaissant l'équation d'ajustement de tendance générale, il est alors aisé de prévoire les résultas des autres observations. Par exemple le résultat de notre experience à une date ulterieur si notre experience dépond du temps.

3.2.1 régression linéaire simple

Dans ce cas le nuage statistique est approché par une droite, la fonction f est donc donnée par $y = f(x) = ax + b$. Le nombre a est appelé pente de la droite.

La méthode des points moyens (méthode de Mayer)

Après le classement ordonné croissant selon les x_i, on partage le nuage en deux sous nuages égaux (ou égaux à une unité près si le nombre d'obsevations est impair). On considèr pour chaque sous nuage un point appelé point moyen.

Les cordonnées du premier point sont $(\bar{x_1}, \bar{y_1})$.

$\bar{x_1}$: Moyenne arithmétique des abscisses x_i du premier sous nuage.

$\bar{y_1}$: Moyenne arithmétique des ordonnées y_i du premier sous nuage.

On note ce point $G_1(\bar{x_1}, \bar{y_1})$.

On procède en suite de même pour le reste des valeurs du second sous nuage, on obtient un second point moyen noté $G_2(\bar{x_2}, \bar{y_2})$. La droite de Mayer est l'unique droite qui passe par ces deux points, son équation est de la forme $y = ax + b$.

Or $G_1(\bar{x_1}, \bar{y_1})$ appartient à la droite, ces cordonnées vérifient donc l'équation et en conséquence : $\bar{y_1} = a\bar{x_1} + b$.

Il en est de même pour $G_2(\bar{x_2}, \bar{y_2})$ et donc $\bar{y_2} = a\bar{x_2} + b$.

Pour déterminer a et b, on résoud le système de deux équations à deux inconues donnés par :

$$\begin{cases} \bar{y_1} = a\bar{x_1} + b \\ \bar{y_2} = a\bar{x_2} + b \end{cases}$$

En retranchant membre à membre la première équation de la deuxième, on trouve

$$\bar{y_2} - \bar{y_1} = a(\bar{x_2} - \bar{x_1}),$$

alors

$$a = \frac{\bar{y_2} - \bar{y_1}}{\bar{x_2} - \bar{x_1}}.$$

Pour trouver b, il suffit de roporter la valeur de a dans l'une des deux équations précédentes.

Exemple 3.2.1

Le premier sous nuage de la distribution des couples (x_i, y_i) *est l'ensemble des points* $(1,5)$, $(2,12)$, $(3,15)$, $(4,22)$.

On a

$$\bar{x_1} = \frac{1+2+3+4}{4} = 2.5, \quad \bar{y_1} = \frac{5+12+15+22}{4} = 13.5$$

d'où

$$G_1(\bar{x_1}, \bar{y_1}) = G_1(2.5, 13.5).$$

De même, le deuxième sous nuage de la distribution des couples (x_i, y_i) *est l'ensemble des points* $(5,34)$, $(6,35)$, $(7,41)$, $(8,48)$.

On a

$$\bar{x_2} = \frac{5+6+7+8}{4} = 6.5, \quad \bar{y_2} = \frac{34+35+41+48}{4} = 39.5$$

d'où $G_2(\bar{x_2}, \bar{y_2}) = G_2(6.5, 39.5)$.

Pour déterminer a et b, on résoud le système

$$\begin{cases} 13.5 = 2.5a + b \\ 39.5 = 6.5a + b \end{cases}$$

d'où $a = 6.5$, *en portant cette valeur dans la première équation on obtient* $b = -2.75$, *alors* $Y = 6.5x - 2.75$.

La méthode des moindres carrés

Les points du nuage sont numérotés $M_1, M_2, \ldots, M_n$.

Lorsque les points du nuage semblent à peut-près alignés, on essaie de trouver l'équation d'une droite (Δ) *approchant mieux le nuage.*

Le point $G(\bar{x}, \bar{y})$ *s'appelle point moyen du nuage.*

Soit (Δ) *la droite d'équation* $Y = ax + b$.

Chercher l'équation $Y = ax + b$ *de la droite d'ajustement* (Δ) *de y en x par la méthode des moindes carrés revient à chercher a et b, t.q*

$$\begin{cases} a = \dfrac{COV(X,Y)}{V(X)} \\ b = \overline{Y} - a\overline{X} \end{cases}$$

D'autre part $r = \dfrac{COV(X,Y)}{\sigma_X \sigma_Y}$ *est le coefficient de corrélation de X et Y.*

Relation entre le coefficient de corrélation et l'ajustement mathématique

1. *Le coefficient de corrélation mesure la qualité d'ajustement affine.*
2. $-1 \leq r \leq 1$.
3. *Si* $r = 0$ *alors il n y a pas de corrélation entre X et Y (X et Y sont independantes.)*
4. *Si* $0 < r < 1$ *alors il y a une corrélation positive faible, moyenne ou forte entre X et Y.*
5. *Si* $-1 < r < 0$ *alors il y a une corrélation négative faible, moyenne ou forte entre X et Y.*
6. $|r| = 1$ *si et seulement si* $Y = ax + b$.

Remarque 3.2.1

Une relation statistique, détectée par le coefficient de corréltion ou par un grapphique, ne montre jamais de relation causale entre deux variables. La causalité ne peut être déduite que d'une analyse non statistique des données.

Conséquence importante

La droite d'ajustement toujours passe par le point moyen $(\bar{x}, \bar{y})$.

Exemple 3.2.2

On calcule dans un premier temps les résultas intermédiaires en utilisant le tableau suivant :

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	5	5	1	25
2	12	24	4	144
3	15	45	9	225
4	22	88	16	484
5	34	170	25	1156
6	35	210	36	1225
7	41	287	49	1681
8	48	384	64	2304
$\bar{x} = 4.5$	$\bar{y} = 26.5$	$\sum x_i y_i = 1213$	$\sum x_i^2 = 204$	$\sum y_i^2 = 7244$

On cherche la droite d'ajustement de X en Y. On a

$$V(X) = \left(\frac{1}{8}\sum_{i=1}^{8} x_i^2\right) - \bar{x}^2 = \frac{204}{8} - (4.5)^2 = 5.25,$$

$$V(Y) = \left(\frac{1}{8}\sum_{i=1}^{8} y_i^2\right) - \bar{y}^2 = \frac{7244}{8} - (26.5)^2 = 203.25,$$

$$COV(X,Y) = \left(\frac{1}{8}\sum_{i=1}^{8} x_i y_i\right) - \bar{x}\bar{y} = \frac{1213}{8} - 4.5 \times 26.5 = 32.375,$$

alors

$$a = \frac{COV(X,Y)}{V(X)} = 6.166,$$

et

$$b = \bar{y} - a\bar{x} = -1.25,$$

d'où

$$Y = 6.166x - 1.25.$$

D'autre part

$$\sigma_X = \sqrt{V(X)} = 2.29, \quad \sigma_Y = \sqrt{V(Y)} = 14.25,$$

d'où

$$r = \frac{COV(X,Y)}{\sigma_X \sigma_Y} = 0.99,$$

et par suite il y a une corrélation positive forte entre X et Y.

3.2.2 régression linéaire multiple

La régression linéaire multiple est une généralisation, à p variables explicatives, de la régression linéaire simple.

Nous sommes toujours dans le cadre de la régression mathématique : nous cherchons à prédire, avec le plus de précision possible, les valeurs prises par une variable y, dite endogène, à partir d'une série de variables explicatives $x_1, x_2, \ldots, x_p$.

Dans le cas de la régression linéaire multiple, la variable endogène et les variables exogènes sont toutes quantitatives (continues), et le modèle de prédiction est linéaire.

Equation de régression et objectifs

Nous disposons de n observations ($i = 1, 2, \ldots, n$). L'équation de régression s'écrit

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \ldots + a_p x_{i,p} + \epsilon_i.$$

où ϵ_i est l'erreur du modèle, elle exprime, ou résume, l'information manquante dans l'explication linéaire des valeurs de y à partir des x_j. $a_0, a_1, \ldots, a_p$ sont les coefficients (paramètres) du modèle à estimer.

Notation matricielle

Nous pouvons adopter une écriture condensée qui rend la lecture et la manipulation de l'ensemble plus facile. Les équations suivantes

$$\begin{cases} y_1 &= a_0 + a_1 x_{1,1} + a_2 x_{1,2} + \ldots + a_p x_{1,p} + \epsilon_1 \\ y_2 &= a_0 + a_1 x_{2,1} + a_2 x_{2,2} + \ldots + a_p x_{2,p} + \epsilon_2 \\ &\vdots \\ y_n &= a_0 + a_1 x_{n,1} + a_2 x_{n,2} + \ldots + a_p x_{n,p} + \epsilon_n \end{cases}$$

peuvent être résumées avec la notation matricielle $Y = Xa + \epsilon$, avec

— *Y est un vecteur colonne de dimension $(n, 1)$.*
— *X est la matrice des observations de dimension $(n, p + 1)$.*
— *a est un vecteur des coefficients de dimension $(p + 1, 1)$.*
— *ϵ est le vecteur des erreurs de dimension $(n, 1)$.*

La matrice X est égale à

$$\begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{pmatrix}.$$

*Les paramètres de la régression linéaire multiple peuvent être estimés en utilisant **la méthode des moindres carrés**. L'estimation des coefficients $a_0, a_1, \ldots, a_p$ se fait en résolvant le système d'équations obtenu par cette minimisation.*

La solution en notation matricielle est donnée par la formule :

$$\bar{a} = \left(X^t X\right)^{-1} X^t Y.$$

où $\bar{a}$ est le vecteur des coefficients estimés, X est la matrice des variables explicatives, et Y est le vecteur des observations de la variable dépendante.

Exemple 3.2.3

Supposons que nous voulons modéliser le prix de vente d'une maison Y en fonction de deux variables explicatives :

1. *X_1 : La surface de la maison en m^2.*
2. *X_2 : Le nombre de pièces.*

   Nous avons 5 observations (maisons) avec les informations suivantes :

Maison(i)	*Surface X_1*	*Nombre de pièces X_2*	*Prix de vente Y*
1	50	3	200
2	60	4	250
3	80	5	300
4	100	6	400
5	120	7	500

L'équation de régression multiple est :

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \epsilon_i,$$

où :

— *y_i : Le prix de vente de la maison i.*

— *$x_{i,1}$: la surface de la maison i.*

— *$x_{i,2}$: Le nombre de pièces de la maison i.*

— *a_0, a_1, a_2 : Les coefficients à estimer.*

Nous pouvons écrire cela sous forme matricielle :

$$Y = Xa + \epsilon.$$

Où

$$Y = \begin{pmatrix} 200 \\ 250 \\ 300 \\ 400 \\ 500 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 50 & 3 \\ 1 & 60 & 4 \\ 1 & 80 & 5 \\ 1 & 100 & 6 \\ 1 & 120 & 7 \end{pmatrix}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ a_1 \end{pmatrix}.$$

La solution en notation matricielle est donnée par la formule :

$$\bar{a} = \left(X^t X\right)^{-1} X^t Y,$$

où

$$X^t = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 80 & 100 & 120 \\ 3 & 4 & 5 & 6 & 7 \end{pmatrix}.$$

On a

$$X^t X = \begin{pmatrix} 1 & 50 & 3 \\ 1 & 60 & 4 \\ 1 & 80 & 5 \\ 1 & 100 & 6 \\ 1 & 120 & 7 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 80 & 100 & 120 \\ 3 & 4 & 5 & 6 & 7 \end{pmatrix} = \begin{pmatrix} 5 & 410 & 25 \\ 410 & 36900 & 2230 \\ 25 & 2230 & 135 \end{pmatrix}.$$

$$\det\left(X^t X\right) = \begin{vmatrix} 5 & 410 & 25 \\ 410 & 36900 & 2230 \\ 25 & 2230 & 135 \end{vmatrix}$$

$$= 5[36900\times135-(2230)^2]-410[410\times135-25\times2230]+25[410\times2230-25\times36900] = 2000.$$

$$\left(X^tX\right)^{-1} = \frac{1}{det\,(X^tX)}\left(com\left(X^tX\right)\right)^t$$

$$= \frac{1}{2000}\begin{pmatrix} 8600 & 400 & -8200 \\ 400 & 50 & -900 \\ -8200 & -900 & 16400 \end{pmatrix}^t = \begin{pmatrix} 4.3 & 0.2 & -4.1 \\ 0.2 & 0.025 & -0.45 \\ -4.1 & -0.45 & 8.2 \end{pmatrix}.$$

$$X^tY = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 50 & 60 & 80 & 100 & 120 \\ 3 & 4 & 5 & 6 & 7 \end{pmatrix}\begin{pmatrix} 200 \\ 250 \\ 300 \\ 400 \\ 500 \end{pmatrix} = \begin{pmatrix} 1650 \\ 149000 \\ 9000 \end{pmatrix}.$$

Alors

$$\bar{a} = \left(X^tX\right)^{-1}X^tY = \begin{pmatrix} 4.3 & 0.2 & -4.1 \\ 0.2 & 0.025 & -0.45 \\ -4.1 & -0.45 & 8.2 \end{pmatrix}\begin{pmatrix} 1650 \\ 149000 \\ 9000 \end{pmatrix} = \begin{pmatrix} -5 \\ 5 \\ -15 \end{pmatrix}.$$

Pour calculer le prix de vente prédit des maisons, on va utiliser l'équation de régression multiple

$$\hat{y}_i = -5 + 5x_{i,1} - 15x_{i,2},$$

alors, on obtient le tableau suivant

Maison	*Surface* X_1	*Nombre de pièce* X_2	*Prix observé* Y	*Prix estimé* $\hat{Y}$
1	50	3	200	200
2	60	4	250	235
3	80	5	300	320
4	100	6	400	405
5	120	7	500	490

Remarquons que :

— *Pour la Maison* 1, *le prix estimé correspond exactement au prix observé* (200).

— *Pour les Maisons* 2, 3, 4 *et* 5, *les prix estimés sont relativement proches des prix observés, avec une différence maximale de* 20 *unités (Maison* 3*).*

Cela montre que le modèle de régression multiple fournit des prédictions raisonnablement proches des valeurs réelles, indiquant une bonne qualité d'ajustement dans l'ensemble. Toutefois, il y a une légère marge d'erreur pour certaines observations, ce qui est attendu dans tout modèle statistique.