Espérance et estimation

Espérance mathématique :

L'espérance mathématique d'une fonction $u(x)$ d'une variable aléatoire x est définie comme étant :

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

ou, pour une variable discrète, $E(u(x)) = \sum_{i=1}^{j} u(x_i)f(x_i)$

Moyenne

La moyenne μ d'une variable aléatoire x n'est pas autre chose que son moment du premier ordre :

$$\mu = E(x) = \int_{-\infty}^{+\infty} xf(x)dx$$

Moments

Le moment d'ordre n d'une distribution d'une variable aléatoire x s'exprime par :

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

Variance et écart-type

La variance d'une variable aléatoire x est le carré de son écart type :

$$var(x) = \sigma^2 = E((x-\mu)^2) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx = \alpha_2^2 - \mu^2$$

Estimation d'une variance

n événements indépendants d'une variable aléatoire x sont mesurés : $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ sans biais de la moyenne.

$E(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}nE(x) = E(x) \quad \forall n$

$var(\bar{x}) = E((\bar{x} - E(x))^2) = E((\frac{x_1 + x_2 + \ldots + x_n}{n} - E(x))^2) = \frac{1}{n^2}E((x_1 - E(x))^2 + (x_2 - E(x))^2 + \ldots + (x_n - E(x))^2) = \frac{\sigma^2}{n}$

Estimation d'une variance

n événements indépendants d'une variable aléatoire x (de moyenne inconnue) sont mesurés :

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$ est un estimateur sans biais de la variance.

Démonstration :

$E(s^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - \bar{x})^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - E(x) + E(x) - \bar{x})^2)$

$E(s^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - E(x))^2 + \sum_{i=1}^{n}(E(x) - \bar{x})^2 + 2\sum_{i=1}^{n}(x_i - E(x))(E(x) - \bar{x}))$

$E(s^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - E(x))^2 + \sum_{i=1}^{n}(E(x) - \bar{x})^2 - 2n(E(x) - \bar{x})^2)$

$$E(s^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - E(x))^2 - \sum_{i=1}^{n}(E(x) - \bar{x})^2)$$
$$E(s^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - E(x))^2 - E(\sum_{i=1}^{n}(\bar{x} - E(x))^2))$$
$$E(s^2) = \frac{1}{n-1}E(\sum_{i=1}^{n}(x_i - E(x))^2 - \sum_{i=1}^{n}E((\bar{x} - E(x))^2))$$
$$E(s^2) = \frac{1}{n-1}(n\sigma^2 - n\frac{\sigma^2}{n}) = \sigma^2 \quad \forall n > 1$$

Explication du facteur $(n-1)$: pour $n = 1$, on a évidemment $\bar{x} = x_1$ et $var(\bar{x}) = \frac{1}{n-1}E(x_1^2 - x_1^2) = \frac{0}{0}$, c'est-à-dire indéterminée.
Si la moyenne théorique $E(x)$ d'une vairable aléatoire est connue, un meilleur estimateur de sa variance est obtenue par :

$$s'^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - E(x))^2 = \frac{1}{n}(\sum_{i=1}^{n}x_i^2 + nE(x)^2 - 2n\bar{x}E(x))$$

Propagation des erreurs et des incertitudes

Lors d'un mesurage on réalise une unique mesure d'une grandeur X_i, le résultat du mesurage ne sera pas égal à la valeur vraie de X_i. On aura $\hat{X}_i = X_{i,\text{vrai}} + \delta X_i$ où δX_i représente l'erreur commise lors de cette unique mesure.

- Si l'erreur sur la mesure est aléatoire, alors $E(\delta X_i) = 0$.
- X_i étant une variable aléatoire, on a $\sigma_{X_i}^2 = \overline{(\hat{X}_i - X_{i,\text{vrai}})^2} = \overline{\delta X_i}^2 = E(\delta X_i^2)$

De la même manière, si l'on dispose d'une mesure de chaque X_i, il est possible de calculer, à partir de $f(X_1 \dots X_n)$, un estimateur de X que l'on peut exprimer comme $\hat{X} = X_{\text{vrai}} + \delta X$ avec

$$\delta X = \sum_i \left(\frac{\partial f}{\partial X_i}\right)_{X_{i,\text{vrai}}} \delta X_i$$

$$\delta X^2 = \sum_i \left(\frac{\partial f}{\partial X_i}\right)^2_{X_{i,\text{vrai}}} \delta X_i^2 + \left[\sum_i \left(\frac{\partial f}{\partial X_i}\right)_{X_{i,\text{vrai}}} \delta X_i\right] . \left[\sum_{j\neq i} \left(\frac{\partial f}{\partial X_j}\right)_{X_{j,\text{vrai}}} \delta X_j\right]$$

Or
– X est une variable aléatoire, donc $\sigma_X^2 = E(\delta X_i^2)$.
– Si de plus les grandeurs X_i sont indépendantes, $E(\delta X_i \delta X_{j\neq i}) = 0$ d'où :

$$E\left[\left(\frac{\partial f}{\partial X_i}\right)_{X_{i,\text{vrai}}} \delta X_i \left(\frac{\partial f}{\partial X_{j\neq i}}\right)_{X_{j,\text{vrai}}} \delta X_j\right] = 0$$

On a donc

$$E(\delta X) = \sum_i \left(\frac{\partial f}{\partial X_i}\right)^2_{X_{i,\text{vrai}}} E(\delta X_i^2)$$

Soit :

$$\sigma_X = \sqrt{\sum_i \left(\frac{\partial f}{\partial X_i}\right)^2_{X_{i,\text{vrai}}} \times \sigma_{X_i}^2}$$

Dans la pratique, comme on ne dispose ni des $X_{i,\text{vrai}}$ ni des $\sigma_{X_i}^2$, on utilise les meilleurs estimateurs dont on dispose :

$$u_X = \sqrt{\sum_i \left(\frac{\partial f}{\partial X_i}\right)^2_{\hat{X}_i} u_{X_i}^2}$$

Coefficients de régression linéaire

Expressions des coefficients

La méthode des moindres carrés consiste à chercher l'équation de la droite $\hat{y} = \hat{a}x + \hat{b}$ par mis tous les $ax + b$ possibles, qui minimise la grandeur $C = \sum_i^N (y_i - \hat{y}_i)^2$. Cela revient alors à chercher les valeurs de $\hat{a}$ et $\hat{b}$ telles que :

$$\left(\frac{\partial C}{\partial a}\right)_{\hat{a}} = 0 \qquad \text{et} \qquad \left(\frac{\partial C}{\partial b}\right)_{\hat{b}} = 0$$

En explicitant la condition précédente, on obtient alors le système d'équations suivant :

$$\begin{cases} \left(\dfrac{\partial C}{\partial a}\right)_{\hat{a}} = -2\displaystyle\sum_i^N (y_i - \hat{a}x_i - \hat{b})x_i = 0 \\ \left(\dfrac{\partial C}{\partial b}\right)_{\hat{b}} = -2\displaystyle\sum_i^N (y_i - \hat{a}x_i - \hat{b}) = 0 \end{cases}$$

En distribuant les sommes,

$$\begin{cases} \displaystyle\sum_i^N x_i y_i - \hat{a}\sum_i^N x_i^2 - \hat{b}\sum_i^N x_i = 0 \\ \displaystyle\sum_i^N y_i - \hat{a}\sum_i^N x_i - N\hat{b} = 0 \end{cases}$$

On peut alors exprimer les coefficients de régression linéaire en exploitant les relations précédentes par substitution :

$$\begin{cases} \displaystyle\sum_i^N N x_i y_i - N\hat{a}\sum_i^N x_i^2 - N\hat{b}\sum_i^N x_i = 0 \\ N\hat{b} = \displaystyle\sum_i^N y_i - \hat{a}\sum_i^N x_i \end{cases} \quad \text{d'où} \quad \begin{cases} \displaystyle\sum_i^N N x_i y_i - N\hat{a}\sum_i^N x_i^2 - \left(\sum_i^N y_i - \hat{a}\sum_i^N x_i\right)\sum_i^N x_i = 0 \\ N\hat{b} = \displaystyle\sum_i^N y_i - \hat{a}\sum_i^N x_i \end{cases}$$

$$\begin{cases} \hat{a} = \dfrac{N\displaystyle\sum_i^N x_i y_i - \sum_i^N x_i \sum_i^N y_i}{N\displaystyle\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2} \\ \hat{b} = \dfrac{1}{N}\left[\displaystyle\sum_i^N y_i - \left(\dfrac{N\displaystyle\sum_i^N x_i y_i - \sum_i^N x_i \sum_i^N y_i}{N\displaystyle\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2}\right)\sum_i^N x_i\right] \end{cases}$$

On obtient alors :

$$\hat{a} = \frac{N\displaystyle\sum_i^N x_i y_i - \sum_i^N x_i \sum_i^N y_i}{N\displaystyle\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2} \qquad \text{et} \qquad \hat{b} = \frac{N\displaystyle\sum_i^N x_i^2 - \sum_i^N x_i \sum_i^N x_i y_i}{N\displaystyle\sum_i^N x_i^2 - \left(\sum_i^N x_i\right)^2}$$

On peut ausi exprimer les coefficients $\hat{a}$ et $\hat{b}$ sous une autre forme, plus simple à mémoriser et à exploiter, en faisant apparaitre dans le système d'équations les valeurs moyennes des échantillons $\overline{x} = \dfrac{1}{N}\sum_i^N x_i$ et $\overline{y} = \dfrac{1}{N}\sum_i^N y_i$.

Reprenons les expressions

$$\begin{cases} \sum_i^N x_i y_i - \hat{a}\sum_i^N x_i^2 - \hat{b}\sum_i^N x_i = 0 \\ \sum_i^N y_i - \hat{a}\sum_i^N x_i - N\hat{b} = 0 \end{cases}$$

L'utilisation des valeurs moyennes $\overline{x}$ et $\overline{y}$ donne :

$$\begin{cases} \sum_i^N x_i y_i - \hat{a}\sum_i^N x_i^2 - \hat{b}\sum_i^N x_i = 0 \\ \hat{b} = \dfrac{1}{N}\sum_i^N y_i - \hat{a}\dfrac{1}{N}\sum_i^N x_i = \overline{y} - \hat{a}\overline{x} \end{cases}$$

Dans l'hypothèse où les points de coordonnées (x_i, y_i) sont normalement répartis autour de la droite, on a

$$\sum_i^N x_i(y_i - \overline{y}) = \sum_i^N x_i(y_i - \overline{y}) - \underbrace{\overline{x}\sum_i^N (y_i - \overline{y})}_{=0} = \sum_i^N (x_i - \overline{x})\sum_i^N (y_i - \overline{y})$$

et

$$\sum_i^N x_i(x_i - \overline{x}) = \sum_i^N x_i(x_i - \overline{x}) - \underbrace{\overline{x}\sum_i^N (x_i - \overline{x})}_{=0} = \sum_i^N (x_i - \overline{x})^2$$

On obtient alors une autre expression équivalente des coefficients de régression linéaire :

$$\hat{a} = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^N (x_i - \overline{x})^2} \qquad \text{et} \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}$$

Dans cet ouvrage, le choix a été fait de travailler avec ces expressions plus simples à manipuler.

Incertitude-type associée à la pente

$$\hat{a} = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^N (x_i - \overline{x})^2}$$

De même que précédemment, on peut remarquer que :

$$\sum_i^N (x_i - \overline{x})(y_i - \overline{y}) = \sum_i^N (x_i - \overline{x}) y_i + \underbrace{\overline{y}\sum_i^N (x_i - \overline{x})}_{=0} = \sum_i^N (x_i - \overline{x}) y_i$$

On obtient alors une expression nouvelle expression de $\hat{a}$:

$$\hat{a} = \frac{\sum_i^N (x_i - \overline{x}) y_i}{\sum_i^N (x_i - \overline{x})^2}$$

Si on considère $u_{x_i} = 0$ et u_{y_i} identique pour chaque y_i, l'incertitude-type associée à la pente s'obtient par composition d'incertitude :

$$u_{\hat{a}}^2 = \frac{\sum_i^N (x_i - \overline{x})^2 u_y^2}{\left[\sum_i^N (x_i - \overline{x})^2\right]^2} = \frac{u_y^2}{\sum_i^N (x_i - \overline{x})^2}$$

$$u_{\hat{a}} = \frac{u_y}{\sqrt{\sum_i^N (x_i - \overline{x})^2}}$$

Incertitude-type associée à l'ordonnée à l'origine

$$\hat{b} = \overline{y} - \hat{a}\overline{x} = \frac{1}{N}\sum_i^N y_i - \hat{a}\overline{x}$$

En composant les incertitudes :

$$u_{\hat{b}}^2 = \frac{1}{N^2} N u_y^2 + \frac{u_y^2}{\sum_i^N (x_i - \overline{x})^2} \overline{x}^2$$

D'où :

$$u_{\hat{b}} = u_y \sqrt{\frac{1}{N} + \frac{\overline{x}^2}{\sum_i^N (x_i - \overline{x})^2}}$$

Cas linéaire

Dans le cas d'une modélisation linéaire, la méthode des moindres carrés consiste à chercher l'équation de la droite $\hat{y} = \hat{a}x$ par mis tous les ax possibles, qui minimise la grandeur $C = \sum_{i=1}^N (y_i - \hat{y}_i)^2$. Cela revient alors à chercher la valeur de $\hat{a}$ telle que :

$$\left(\frac{\partial C}{\partial a}\right)_{\hat{a}} = 0$$

En explicitant la condition précédente, on obtient alors l'équation suivant :

$$\left(\frac{\partial C}{\partial a}\right)_{\hat{a}} = -2\sum_{i=1}^N (y_i - \hat{a}x_i)x_i = 0$$

d'où

$$\hat{a} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

Chaque expérience réalisée, permet de placer sur le graphe un point de coordonnées (x_i, y_i). L'ensemble de ces points est supposé normalement distribué par rapport à la droite de régression linéaire. Pour chaque point, la distance à la droite $r_i = y_i - \hat{y}_i$ (appelée "le résidu") traduit l'erreur que commise lors de chaque mesure.

On quantifie alors la dispersion des points expérimentaux par rapport au modèle selon la relation :

$$s_{stat} = \sqrt{\frac{1}{d}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

Remarques :

- Le coefficient d correspond au nombre de degrés de liberté de l'échantillon. Dans le cas d'une modélisation linéaire, de la forme $Y = aX$, $d = N - 1$
- Dans le cas où l'on suppose que les x_i sont parfaitement déterminés, c'est à dire $u_{x_i} = 0$ alors on identifie $s_{stat} = \hat{u}_y$

L'incertitude type associée à la pente a découle de $\hat{u}_y$ par simple propagation d'incertitudes.

Fonctions tableurs (Excel et LibreOffice)

MOYENNE(données)
détermine la moyenne arithmétique sur une série de données.

ECARTYPE(données)
détermine l'écart-type expérimental (pou N-1) pour une série de données.

FREQUENCE(données ; classes)
données est une plage ou une matrice contenant des données numériques.
classes est une unique colonne de plage ou une matrice contenant des nombres en ordre croissant qui représentent la limite supérieur de chaque catégorie.
FREQUENCE renvoie une unique colonne de matrice, où le premier élément est le nombre de valeurs dans données qui sont inférieures ou égales à la première valeur dans classes, la seconde valeur est le nombre de valeurs dans données qui sont supérieures ou égales à la première valeur mais inférieures ou égales à la seconde valeur dans classes, et ainsi de suite. La matrice renvoyée est plus longue d'un élément que classes ; le dernier élément contient le nombre de valeurs dans données qui sont supérieures à la dernière valeur de classes.
Pour renvoyer une matrice, **FREQUENCE** doit être saisie comme une formule de matrice, en pressant *Cntrl-Maj-Entrée* à la place de *Entrée* (ou en cochant la case à cocher Matrice si vous utilisez l'assistant Formules).

LOI.STUDENT.INV(probabilité ;degré_de_liberté)
permet de déterminer le coefficient de Student en indiquant la probabilité retenue et le nombre de degré de liberté.

DROITEREG(valeursy ; valeursx ; type_linéaire ; stats)
permet de calculer le coefficient directeur d'une droite ainsi que l'ordonnée à l'origine avec pour chacune de ces valeurs un ecart-type.
valeursy est une unique colonne ou ligne de plage spécifiant les coordonnées y dans un ensemble de points de données.
valeursx est une unique ligne ou colonne de plage correspondante spécifiant les coordonnées x. Si **valeursx** est omis, il est par défaut 1, 2, 3, ..., n. S'il y a plus d'un ensemble de variables valeursx peut être une plage avec des lignes et colonnes multiples correspondantes.
DROITEREG trouve une ligne droite $y = a + bx$ qui correspond le mieux aux données, en utilisant une régression linéaire. Avec plus d'un ensemble de variables, la ligne droite est de la forme $y = a + b_1x_1 + b_2x_2... + b_nx_n$.
Si **type_linéaire** est **FAUX** la ligne droite trouvée est forcée pour passer à travers l'origine (la constante a est zéro ; $y = bx$). Si omis, **type_linéaire** est par défaut **VRAI** (la ligne n'est pas forcée à travers l'origine).
DROITEREG renvoie une table (matrice) de statistiques comme ci-dessous et doit être saisie comme une formule de matrice (par exemple, en utilisant *Cntrl-Maj-Entrée* à la place de *Entrée*). Si **stats** est omis ou **FAUX** seule la ligne supérieure des statistiques est renvoyée. Si **VRAI** la table entière est renvoyée.

b_n	b_{n-1}	...	b_1	a
s_n	s_{n-1}	...	s_1	s_a
r^2	s_y			
F	df			
ss_{reg}	ss_{resid}			

b_1 à b_n sont la pente de la ligne ; a est l'intersection de l'axe y.
s_1 à s_n sont les estimateurs d'écart-type pour la pente de la ligne ; s_a est la valeur d'écart-type pour l'intersection de l'axe y.
r^2 est le coefficient de détermination ; s_y est l'estimateur de l'écart-type pour l'estimation y.
F est la statistique F (valeur F-observée) ; df est le nombre de degrés de liberté.
ss_{reg} est la somme de régression des carrés ; ss_{resid} est la somme des carrés résiduels.

LOI.NORMALE(x ; μ ; σ ; mode)

La distribution normale est une famille souvent rencontrée de distributions de probabilité continue avec les paramètres μ (moyenne) et σ (écart type).

Si mode est **0**, **LOI.NORMALE** calcule la fonction de densité de probabilité de la distribution normale :

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

Si le mode est **1**, **LOI.NORMALE** calcule la fonction de distribution cumulative de la distribution normale :

$$\int_{-\infty}^{t} \frac{\mathrm{e}^{-\frac{(t-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} dt$$

LOI.NORMALE.STANDARD(x)

La distribution normale standard est une distribution normale avec une moyenne $\mu = 0$ et un écart type $\sigma = 1$.

LOI.NORMALE.STANDARD calcule la fonction de distribution cumulative d'une distribution normale standard. C'est l'équivalent de **LOI.NORMALE(x ; 0 ; 1 ; 1)**.