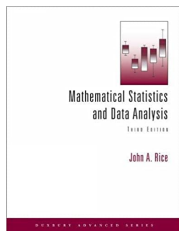


Introduction aux statistiques pour l'Ingénieur

Stéphane Canu

asi.insa-rouen.fr/enseignants/~scanu

scanu@insa-rouen.fr

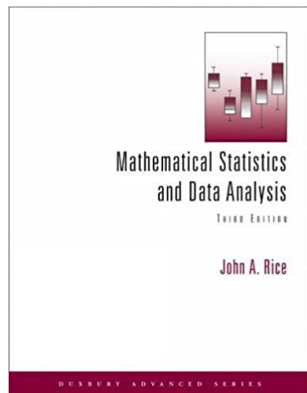


ITI 3, INSA Rouen Normandie, Février 2024

Lecture road map

1 Échantillonnage

- Échantillon, vraisemblance et statistique
 - Échantillon
 - Vraisemblance
 - Statistique
 - Deux statistiques : la moyenne et la variance
- Moyenne et variance : le cas normal
 - Définition
 - Le cas de deux échantillons gaussien



<https://moodle.insa-rouen.fr/course/view.php?id=93>

Échantillon

Définition naïve (après l'observation)

Une partie (représentative) d'un ensemble, utilisée pour analyser ou comprendre la nature de l'ensemble entier.

Exemple de « partie » (échantillon observé):

$$s_n = (\textit{pile}, \textit{pile}, \textit{face}, \dots, \textit{pile} \dots, \textit{face})$$

L'ensemble entier : le résultat du tirage suit une loi de Bernouilli de paramètre p

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

comprendre sa nature

Quelle est la valeur du paramètre p ?

Définition mathématique d'un échantillon (avant l'observation)

Une liste de variables aléatoires

$$S_n = (X_1, \dots, X_i, \dots, X_n)$$

Échantillon

L'ensemble entier : loi parente de paramètre θ

$$X \sim \mathcal{L}(\theta)$$

comprendre sa nature

Quelle est la valeur du paramètre θ ?

Définition mathématique d'un échantillon

Une liste de variables aléatoires **i.i.d de loi parente $\mathcal{L}(\theta)$**

$$\mathcal{S}_n = (X_1, \dots, X_i, \dots, X_n)$$

i.i.d = indépendant et identiquement distribué

Contre-exemples :

- non indépendance : suite de mots
- non identiquement distribué : temps entre 2 pannes d'un logiciel

Échantillon

L'ensemble entier : loi parente de paramètre θ

$$X \sim \mathcal{L}(\theta) \quad \text{de densité ou probabilité } f(x, \theta) \text{ ou } \mathbb{P}(x, \theta)$$

Définition mathématique d'un échantillon

Une liste de variables aléatoires i.i.d de loi parente $\mathcal{L}(\theta)$

$$\mathcal{S}_n = (X_1, \dots, X_i, \dots, X_n)$$

i.i.d = indépendant et identiquement distribué

\mathcal{S}_n est une variable aléatoire de densité (ou probabilité)

$$f(\mathcal{S}_n, \theta) = \prod_{i=1}^n f(x_i, \theta) \quad \text{ou} \quad \mathbb{P}(\mathcal{S}_n, \theta) = \prod_{i=1}^n \mathbb{P}(x_i, \theta)$$

Exemples :

Loi normale

Loi exponentielle

Loi de Bernouilli

Loi de Poisson

Exemples de loi de probabilité d'un Échantillon

- Loi normale

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

$$\begin{aligned} f(s_n, \theta) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x_i - \mu)^2}{2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2}{2\sigma^2} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp -\frac{t_2(x_1, \dots, x_i, \dots, x_n) - 2\mu t_1(x_1, \dots, x_i, \dots, x_n) + n\mu^2}{2\sigma^2} \end{aligned}$$

$$\text{avec } t_1(x_1, \dots, x_i, \dots, x_n) = \sum_{i=1}^n x_i \text{ et } t_2(x_1, \dots, x_i, \dots, x_n) = \sum_{i=1}^n x_i^2$$

$$f(s_n, \theta) = f(t_1, t_2, \mu, \sigma)$$

- Loi de Bernouilli
- Loi exponentielle
- Loi de Poisson

Une fonction de deux variables

La densité : θ est fixé, la variable est x

$$f(x, \theta) \text{ ou } \mathbb{P}(x, \theta)$$

$$f(s_n, \theta) = \prod_{i=1}^n f(x_i, \theta) \quad \text{ou} \quad \mathbb{P}(s_n, \theta) = \prod_{i=1}^n \mathbb{P}(x_i, \theta)$$

La vraisemblance : x est fixé, la variable est θ

$$\begin{aligned} \mathcal{L} : \mathbb{R} &\rightarrow \mathbb{R} \\ \theta &\mapsto f(s_n, \theta) \end{aligned}$$

Statistique

Définition naïve d'une statistique

C'est une caractéristique ou une mesure qui décrit un aspect des données d'un échantillon. Cette mesure peut être utilisée pour résumer ou interpréter un ensemble de données.

Exemple : la moyenne

Définition mathématique d'une statistique

Une statistique T est une fonction de l'échantillon, donc une variable aléatoire

$$\begin{aligned} T : \Omega^n &\rightarrow \mathbb{R} \\ \mathcal{S}_n &\mapsto T(X_1, \dots, X_i, \dots, X_n) \end{aligned}$$

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

Moyenne et variance : la moyenne

si X_1, X_2, \dots, X_n est un échantillon de n réalisations i.i.d. d'une loi parente d'espérance μ et la variance σ^2 .

La Moyenne

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

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

$$\begin{aligned} \mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \mu \end{aligned}$$

$$\begin{aligned} V(\bar{X}) &= V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n V(X_i) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

Le théorème central limite (De Moivre, 1733)

La loi de la moyenne

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$

converge vers

$$\xrightarrow[n \rightarrow \infty]{(d)}$$

la loi normale

$$\mathcal{N}(0, 1)$$

Moyenne et variance : la variance (1)

Espérance μ est connue (c'est rare)

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$$

Propriété : $\mathbb{E}(S_n^2) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}((X_i - \mu)^2) = \sigma^2$

Espérance μ est inconnue (c'est le plus souvent le cas)

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

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

Moyenne et variance : le cas normal

Que se passe-t-il quand la loi parente est la loi normale ?

La loi du χ^2

Soit $Y \sim \mathcal{N}(0, 1)$ une variable aléatoire normale centrée réduite. Soit Y_1, Y_2, \dots, Y_n un échantillon de n réalisations i.i.d. de cette variable aléatoire.

Definition (La loi du χ^2)

On appelle loi du χ^2 à n degrés de libertés la loi de la variable aléatoire Z_n

$$Z_n = \sum_{i=1}^n Y_i^2$$

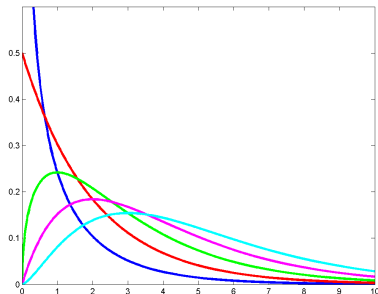


Figure: Exemples de loi du chi 2 pour 1 (bleu), 2 (rouge), 3 (vert), 4 (violet) et 5 (bleu ciel) degrés de liberté

le cas de la moyenne

si X_1, X_2, \dots, X_n est un échantillon de n réalisation i.i.d. d'une variable aléatoire normale $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et le variance σ^2 .

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

$$\begin{aligned} \mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) & V(\bar{X}) &= V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) & &= \frac{1}{n^2} \sum_{i=1}^n V(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu & &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 & &= \frac{\sigma^2}{n} \\ &= \mu & & & & \end{aligned}$$

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \qquad \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \sim \mathcal{N}(0, 1)$$

et

$$n \frac{(\bar{X} - \mu)^2}{\sigma^2} \sim \chi_1^2$$

La moyenne se concentre autour de l'espérance

si X_1, X_2, \dots, X_n est un échantillon de n réalisation i.i.d. d'une variable aléatoire normale $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et la variance σ^2 .

$$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi_n^2$$

puisque $Y_i = \frac{X_i - \mu}{\sigma}$ suit une loi normale centrée réduite.

Il est moins évident c'est de montrer que : $\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$ Lorsque

l'on remplace le paramètre μ par son estimation $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ on perd un degré de liberté. En effet on a la décomposition suivante :

$$\underbrace{\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma^2}}_{\chi_n^2} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} + \underbrace{n \frac{(\bar{X} - \mu)^2}{\sigma^2}}_{\chi_1^2}$$

Qui permet de conclure en invoquant le théorème de Cochran sur l'additivité des degrés de liberté.

La loi de Student : définition

- Soit $N \sim \mathcal{N}(0, 1)$ une variable aléatoire normale centrée réduite.
- Soit X_n la variable aléatoire distribuée suivant une loi du χ^2 à n ddl
 - ▶ C'est le cas par exemple, si N_1, N_2, \dots, N_n un échantillon de n réalisation i.i.d. une variable aléatoire normale centrée réduite quand $X_n = \sum_{i=1}^n N_i^2$
- supposons que N et X_n sont indépendantes (*i.e.* $\text{cov}(Y, X_n) = 0$)

Definition (La loi de student)

On appelle loi de student à n degrés de libertés la loi de la variable aléatoire T_n

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}}$$

$$N \sim \mathcal{N}(0, 1)$$

$$X_n \sim \chi_n^2$$

Le cas de la moyenne d'un échantillon gaussien

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ une variable aléatoire normale d'espérance μ et de variance σ^2 . Soit X_1, X_2, \dots, X_n un échantillon de n réalisations i.i.d. de cette variable aléatoire. La moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ de cet échantillon suit aussi une loi normale

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

car $\mathbb{E}(\bar{X}) = \mu$ et $V(\bar{X}) = \frac{\sigma^2}{n}$:

$$\begin{aligned}\mathbb{E}(\bar{X}) &= \mathbb{E}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(X_i) \\ &= \frac{1}{n} \sum_{i=1}^n \mu = \mu\end{aligned}$$

$$\begin{aligned}V(\bar{X}) &= V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n V(X_i) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sigma^2 = \frac{\sigma^2}{n}\end{aligned}$$

Le cas de la moyenne d'un échantillon gaussien

Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ une variable aléatoire normale d'espérance μ et de variance σ^2 . Soit X_1, X_2, \dots, X_n un échantillon de n réalisations i.i.d. de cette variable aléatoire. La moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ de cet échantillon suit aussi une loi normale

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

On peut donc construire la variable normale centrée réduite

$$Y = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1). \text{ Or } Z_{n-1} = \sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi_{n-1}^2$$

On peut construire une variable aléatoire suivant une loi de Student

$$T_{n-1} = \frac{Y}{\sqrt{\frac{Z_{n-1}}{n-1}}} = \frac{\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}}{\sqrt{\frac{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma^2}}{n-1}}} = \frac{\bar{X} - \mu}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}} = \frac{\bar{X} - \mu}{\frac{S_{n-1}}{\sqrt{n}}}$$

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

Conclusion

- Échantillon

- ▶ Une liste de variables aléatoires $\mathcal{S}_n = (X_1, \dots, X_i, \dots, X_n)$
- ▶ i.i.d de loi parente $\mathcal{L}(\theta)$
- ▶ paramètre θ

- Vraisemblance

- ▶ fonction de densité (ou probabilité) de l'échantillon
- ▶ vue comme une fonction du paramètre θ
- ▶ importance de l'hypothèse i.i.d.

- Statistique

- ▶ c'est une variable aléatoire
- ▶ apparaît « naturellement » dans l'écriture de la vraisemblance
- ▶ suit une certaine loi