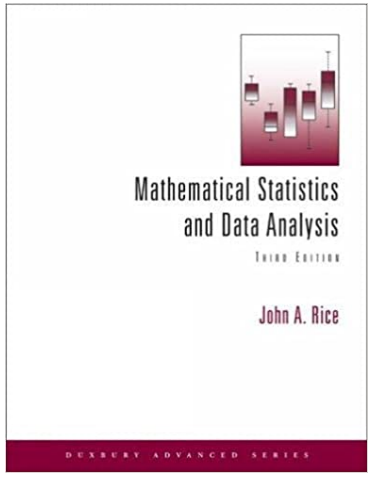
Introduction aux statistiques pour l'Ingénieur

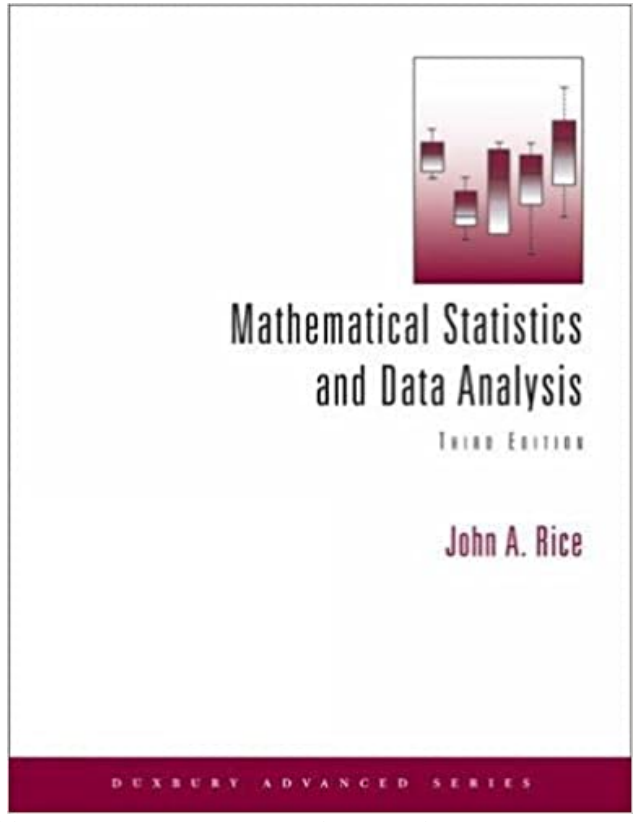
Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu
scanu@insa-rouen.fr



ITI 3, INSA Rouen Normandie, Janvier 2026

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):

$$s_n = (pile, pile, face, \ldots, pile \ldots, face)$$

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

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

Exemples d'échantillons

1. Quelle page d'accueil (A/B) génère le meilleur taux de conversion ?
   - Loi binomiale $x \in \{0, 1\}$
   - $s_n = (x_1 = 1, x_2 = 0, \ldots, x_i = 1, \ldots, x_n = 0)$
2. Comment détecter une attaque DoS : Combien de paquets arrivent à ce switch en une mili seconde ?
   - Loi de poisson : $x \in \mathbb{N}$
   - $s_n = (x_1 = 3, x_2 = 15, \ldots, x_i = 7, \ldots, x_n = 11)$
3. Comment dimensionner un load balanceur ? Quel ve être le temps entre deux arrivées de requêtes réseau ?
   - Loi exponentielle : $x \in \mathbb{R}^+$
   - $s_n = (x_1 = 0,329, x_2 = 1,511, \ldots, x_i = 0,271, \ldots, x_n = 0,111)$
4. Entraînement d'un modèle de Deep Learning sur GPU : Quel va être le temps d'exécution d'un algorithme complexe
   - Loi normale : $x \in \mathbb{R}$
   - $s_n = (x_1 = 12,2, x_2 = 15,11, \ldots, x_i = 27,1, \ldots, x_n = 14,53)$

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 $\mathcal{L}(\theta)$

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

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 $\mathcal{L}(\theta)$

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

i.i.d = indépendant et identiquement distribué
$\mathcal{S}_n$ est une variable aléatoire de densité (ou probabilité)

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

Exemples de loi de probabilité d'un Échantillon

1. Quelle page d'accueil (A/B) génère le meilleur taux de conversion ?
   - Loi binomiale $X \sim \mathcal{B}(p)$ $\quad \mathbb{P}(X = x) = p^x(1-p)^{1-x}$
   - $\mathbb{P}(s_n, p) = \prod_{i=1}^{n} \mathbb{P}(X = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$
2. Comment détecter une attaque DoS : Combien de paquets arrivent à ce switch en une mili seconde ?
   - Loi de poisson : $X \sim \mathcal{P}(\lambda)$ $\quad \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$
   - $\mathbb{P}(s_n, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} e^{-\lambda}$
3. Comment dimensionner un load balanceur ? Quel ve être le temps entre deux arrivées de requêtes réseau ?
   - Loi exponentielle : $X \sim \mathcal{E}(\mu)$ $\quad f(x) = \frac{1}{\mu} e^{-x/\mu}$
   - $f(s_n, \mu) = \prod_{i=1}^{n} \frac{1}{\mu} e^{-x_i/\mu}$
4. Entraînement d'un modèle de Deep Learning sur GPU : Quel va être le temps d'exécution d'un algorithme complexe
   - Loi normale : $X \sim \mathcal{N}(\mu, \sigma)$ $\quad f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$
   - $f(s_n, \theta = (\mu, \sigma^2)) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x_i-\mu)^2}{2\sigma^2}$

Exemples de loi de probabilité d'un Échantillon

1. Quelle page d'accueil (A/B) : Loi binomiale $X \sim \mathcal{B}(p)$

$$\begin{aligned}\mathbb{P}(s_n, p) &= \prod_{i=1}^{n} \mathbb{P}(X = x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \qquad = \mathbb{P}(t(s_n), p) \text{ avec } t(s_n) = \sum_{i=1}^{n} x_i\end{aligned}$$

2. Comment détecter une attaque DoS : Loi de poisson : $X \sim \mathcal{P}(\lambda)$

$$\begin{aligned}\mathbb{P}(s_n, \lambda) &= \prod_{i=1}^{n} \frac{\lambda^{k_i}}{k_i!} e^{-\lambda} \\ &= \frac{\lambda^{\sum_{i=1}^{n} k_i}}{\prod_{i=1}^{n} k_i!} e^{-n\lambda} \qquad = \mathbb{P}(t_1(s_n), t_2(s_n), p) \text{ avec } t_1(s_n) = \sum_{i=1}^{n} k_i\end{aligned}$$

3. Comment dimensionner un load balanceur : Loi exponentielle : $X \sim \mathcal{E}(\mu)$

$$\begin{aligned}f(s_n, \mu) &= \prod_{i=1}^{n} \frac{1}{\mu} e^{-x_i/\mu} \\ &= \frac{1}{\mu^n} e^{-\sum_{i=1}^{n} x_i/\mu} \qquad = f(t(s_n), \mu) \text{ avec } t(s_n) = \sum_{i=1}^{n} x_i\end{aligned}$$

4. Quel va être le temps d'exécution : Loi normale : $X \sim \mathcal{N}(\mu, \sigma)$

$$f(s_n, \theta = (\mu, \sigma^2)) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x_i-\mu)^2}{2\sigma^2}$$

Exemples de loi de probabilité d'un Échantillon

- Loi normale

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

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

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

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

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

$$\begin{array}{rccc} \mathcal{L} : & \mathbb{R} & \to & \mathbb{R} \\ & \theta & \mapsto & f(s_n, \theta) \end{array}$$

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

$$\begin{array}{rccl} T: & \Omega^n & \to & \mathbb{R} \\ & \mathcal{S}_n & \mapsto & T(X_1, \ldots, X_i, \ldots, X_n) \end{array}$$

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

Moyenne et variance : la moyenne

si $X_1, X_2, ..., X_n$ est un échantillon de n réalisation i.i.d. d'une loi parente d'espérance μ et le variance σ^2.

La Moyenne

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}(X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mu \qquad = \mu \end{aligned}$$

$$\begin{aligned} V(\overline{X}) &= V(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} V(X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} \sigma^2 \qquad = \frac{\sigma^2}{n} \end{aligned}$$

Le thèorème central limite (De Moivre, 1733)

La loi de la moyenne $\dfrac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$ converge vers $\xrightarrow[n\to\infty]{(d)}$ la loi normale $\mathcal{N}(0,1)$

Moyenne et variance : la variance (1)

- Espérance μ est connue (c'est rare) : $\mathcal{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$

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

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

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

Propriété :

$$\begin{aligned}\mathbb{E}(S_n^2) &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbb{E}\big((X_i-\overline{X})^2\big)\\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\mathbb{E}\big((X_i-\overline{X}-\mu+\mu)^2\big)\\ &= \sigma^2 - \tfrac{\sigma^2}{n} \qquad\qquad = \tfrac{n-1}{n}\sigma^2\end{aligned}$$

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

$$V(S_{n-1}) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right).$$

Moyenne et variance : le cas normal

Que ce passe t'il quand la loi parente est la loi normale ?

La loi de la moyenne	est	la loi normale
$\dfrac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$	$\sim$	$\mathcal{N}(0,1)$
$\overline{X}$	$\sim$	$\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$

La loi de la variance	est	la loi ???
S^2_{n-1}	$\sim$	???

La loi du χ^2

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

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

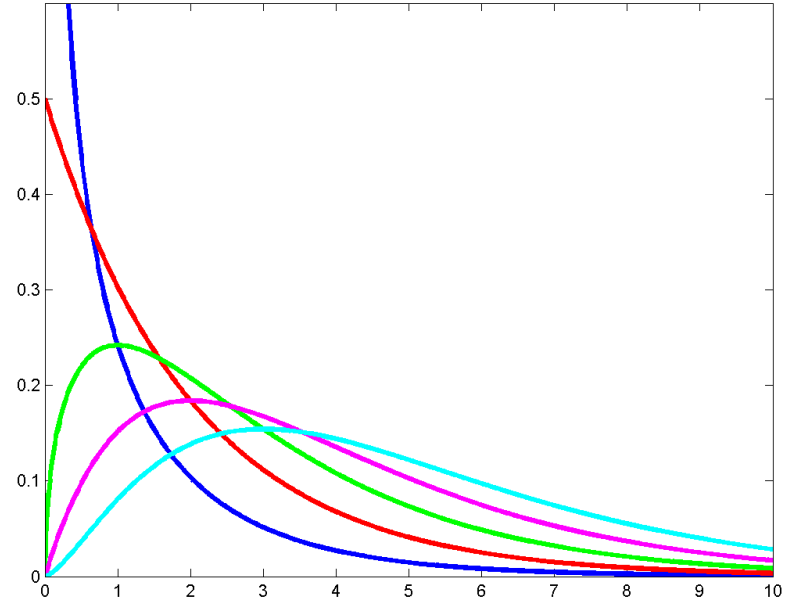


Figure: Exemples de loi du chi 2 pour 1 (bleu), 2 (rouge), 3 (vert), 4 (violet) et 5 (bleu ciel) degrés de liberté

le cas de la moyenne

si $X_1, X_2, ..., X_n$ 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.

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

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}(X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mu \qquad = \mu \end{aligned}$$

$$\begin{aligned} V(\overline{X}) &= V(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} V(X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} \sigma^2 \qquad = \frac{\sigma^2}{n} \end{aligned}$$

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

et

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

La moyenne se concentre autour de l'espérance

si $X_1, X_2, ..., X_n$ 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.

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

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

Il est moins évident c'est de montrer que : $\sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2_{n-1}$ Lorsque l'on remplace le paramètre μ par son estimation $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ on perd un degré de liberté. En effet on a la décomposition suivante :

$$\underbrace{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2}}_{\chi^2_n} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} + \underbrace{n\frac{(\overline{X} - \mu)^2}{\sigma^2}}_{\chi^2_1}$$

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

Moyenne et variance : le cas normal

Que ce passe t'il quand la loi parente est la loi normale ?

La loi de la moyenne	est	la loi normale
$\dfrac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}}$	$\sim$	$\mathcal{N}(0,1)$
$\overline{X}$	$\sim$	$\mathcal{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$

La loi de la variance	est	la loi du chi 2
$\dfrac{n-1}{\sigma^2} S^2_{n-1}$	$\sim$	χ^2_{n-1}

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 $N_1, N_2, ..., N_n$ 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 $(i.e.\ cov(Y, X_n) = 0)$

Definition (La loi de student)

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

$$T_n = \frac{N}{\sqrt{\frac{X_n}{n}}} \qquad N \sim \mathcal{N}(0,1) \qquad X_n \sim \chi^2_n$$

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 $X_1, X_2, ..., X_n$ un échantillon de n réalisation i.i.d. de cette variable aléatoire. La moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de cet échantillon suit aussi une loi normale

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

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

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbb{E}(X_i) \\ &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned} \qquad \begin{aligned} V(\overline{X}) &= V(\tfrac{1}{n}\textstyle\sum_{i=1}^{n} X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} V(X_i) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^{n} \sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

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 $X_1, X_2, ..., X_n$ un échantillon de n réalisation i.i.d. de cette variable aléatoire. La moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de cet échantillon suit aussi une loi normale

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

On peut donc construire la variable normale centrée réduite

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

On peut construire une variable aléatoire suivant une loi de Student

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

avec $S^2_{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

Conclusion

- Échantillon
  - Une liste de variables aléatoires $\mathcal{S}_n = (X_1, \ldots, X_i, \ldots, X_n)$
  - 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
  - apparait « naturellement » dans l'écriture de la vraisemblance
  - suit une certaine loi