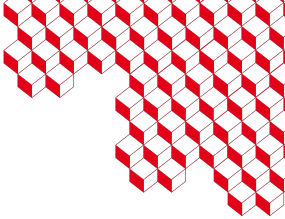


Statistique I

M2 Radiophysique médicale, INSTN, 2023

Clément GAUCHY (clement.gauchy@cea.fr) Blog: clgch.github.io
CEA SACLAY

Sommaire

Résumé du cours

- Contexte et définition de la statistique
- Statistique descriptive
- Statistique inférentielle
- Intervalles de confiance
- Rappels de métrologie

Définition de la statistique

Définition: "Le mot statistique désigne à la fois un ensemble de données d'observations et l'activité qui consiste dans leur receuil, leur traitement et leur interprétation." (Encyclopedia Universalis)

Étymologie: "De l'allemand Statistik, forgé par l'économiste Gottfried Achenwall, dérivé de l'italien statista (« homme d'État, statiste »), la statistique représentant pour lui l'ensemble des connaissances que doit posséder un homme d'État." (Wiktionary)

Objectif de la statistique

La statistique concerne le receuil, le traitement, et l'analyse de données. Le point fondamental étant que **toutes données est entachée d'incertitudes**, et fait donc intervenir la théorie des probabilités. L'origine de ces incertitudes est multiple:

- Les phénomènes observés ne sont pas explicable par un modèle ou un lien logique (On ne sait pas prévoir les cours de la bourse)
- Les mesures sont entachées d'erreur
- La physique étudié est probabiliste (Physique quantique)

$\Longrightarrow$ La statistique requiert donc de modéliser les données à l'aide de la théorie des probabilité.

Deux classes de méthodes statistiques

1. **Statistique descriptive**: Elle a pour but de résumer l'information contenue dans les données de façon synthétique et efficace, en utilisant des réprésentation graphiques, des indicateurs quantitatifs. $\Longrightarrow$ Elle permet de dégager les caractéristiques des données étudiées et d'aider à la décision pour une modélisation plus poussée de celles-ci. Les probabilités n'ont qu'un role marginal.

2. **Statistique inférentielle**: Elle a pour but de faire des prévisions et de prendre des décisions au vu des observations. Il faut pour cela faire des modèles probabilistes du phénomène étudié pour savoir gérer les risques d'erreurs.

Variable aléatoire

On va représenter les données via une modélisation probabiliste.

Une variable aléatoire noté $X : \Omega \to \mathcal{X}$ est une grandeur dépendant d'une expérience aléatoire.

Je suis un atome radioactif, $X =$ le temps avant de devenir un isotope stable.

Une réalisation x est une valeur prise par X.

La fonction de répartition est défini par $F(x) = \mathbb{P}(X \leq x)$.

Statistique descriptive: position

Moyenne: ⚠ Sensible aux valeurs extrêmes de l'échantillon ⚠

Médiane: Valeur séparant la moitié inférieure et la moitié supérieure d'un échantillon (statistique d'ordre). insensible aux valeurs extrêmes.

Mode: Pic de l'histogramme, pic de fréquence

Statistique descriptive: dispersion

Variance: Moyenne du carré des distances à la moyenne de chaque valeur des échantillons.

Ecart-type: Racine carré de la variance (homogène avec les unités des observations !)

Coefficient de variation: écart-type divisé par la moyenne ⚠ instable si la moyenne est proche de 0 ⚠

Quantiles: Statistique de rang (k-ième plus grande/plus petite valeur de l'échantillon)

Introduction à l'inférence statistique

- On dispose d'un n-échantillon $(x_1, x_2, \ldots, x_n)$ de X de loi inconnue
- L'inférence statistique consiste à approcher certaines caractéristiques de X (moyenne, variance) à partir du n-échantillon
- Dans certains cas, on pourra supposer X appartenant à une certaine famille de lois de probabilités $\mathcal{M} = \{p_\theta, \theta \in \Theta\}$ paramétrée par θ. On parle d'estimation paramétrique de θ et $\mathcal{M}$ est appellé le modèle statistique.

Introduction à l'inférence statistique

- On dispose d'un n-échantillon $(x_1, x_2, \ldots, x_n)$ de X de loi inconnue
- L'inférence statistique consiste à approcher certaines caractéristiques de X (moyenne, variance) à partir du n-échantillon
- Dans certains cas, on pourra supposer X appartenant à une certaine famille de lois de probabilités $\mathcal{M} = \{p_\theta, \theta \in \Theta\}$ paramétrée par θ. On parle d'estimation paramétrique de θ et $\mathcal{M}$ est appellé le modèle statistique.
- **Hypothèse 1:** Le n-échantillon a été tiré de façon à assurer la représentativité de la loi X

tirages équiprobables et indépendants les uns des autres

Introduction à l'inférence statistique

- On dispose d'un n-échantillon $(x_1, x_2, \ldots, x_n)$ de X de loi inconnue
- L'inférence statistique consiste à approcher certaines caractéristiques de X (moyenne, variance) à partir du n-échantillon
- Dans certains cas, on pourra supposer X appartenant à une certaine famille de lois de probabilités $\mathcal{M} = \{p_\theta, \theta \in \Theta\}$ paramétrée par θ. On parle d'estimation paramétrique de θ et $\mathcal{M}$ est appellé le modèle statistique.
- **Hypothèse 1:** Le n-échantillon a été tiré de façon à assurer la représentativité de la loi X

  tirages équiprobables et indépendants les uns des autres

- **Hypothèse 2:** Chaque valeur observée y_i est une réalisation de X. Le n-échantillon est donc aléatoire :

  l'estimateur est une variable aléatoire

Introduction à l'inférence statistique

- On dispose d'un n-échantillon $(x_1, x_2, \ldots, x_n)$ de X de loi inconnue
- L'inférence statistique consiste à approcher certaines caractéristiques de X (moyenne, variance) à partir du n-échantillon
- Dans certains cas, on pourra supposer X appartenant à une certaine famille de lois de probabilités $\mathcal{M} = \{p_\theta, \theta \in \Theta\}$ paramétrée par θ. On parle d'estimation paramétrique de θ et $\mathcal{M}$ est appellé le modèle statistique.
- **Hypothèse 1:** Le n-échantillon a été tiré de façon à assurer la représentativité de la loi X

  tirages équiprobables et indépendants les uns des autres

- **Hypothèse 2:** Chaque valeur observée y_i est une réalisation de X. Le n-échantillon est donc aléatoire :

  l'estimateur est une variable aléatoire

- Le n-échantillon est une réalisation des variables $(X_1, X_2, \ldots, X_n)$ de même loi que la loi parente X et indépendantes, i.i.d. (indépendantes et identiquement distribuées)
- Par exemple, l'espérance de X est estimée par la moyenne empirique :

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

  On remarque bien que $\bar{X}_n$ est aléatoire car les X_i sont des variables aléatoires.

Statistique - Estimateur

- Sur un n-échantillon, une statistique S_n est une fonction des n variables aléatoires (i.i.d.)
- Un estimateur est défini par une statistique $S_n(X_1, X_2, \ldots, X_n)$ souvent noté $\widehat{\theta}_n$. C'est donc une variable aléatoire. Une estimation s_n est une réalisation de S_n
- Soit S_n un estimateur de θ par S_n. Il est sans biais si

$$\mathbb{E}(S_n) = \theta$$

- ou asymptotiquement sans biais si :

$$\lim_{n \to \infty} \mathbb{E}(S_n) = \theta$$

- et son erreur quadratique moyenne est :

$$\mathbb{E}[(S_n - \theta)^2] = \underbrace{(\mathbb{E}(S_n) - \theta)^2}_{\texttt{biais}^2} + \underbrace{\mathrm{Var}(S_n)}_{\texttt{variance}}$$

Sommaire

Estimateur des moments (EMM)

La méthode des moments s'appuie sur un théorème fondamental de la théorie des probabilités.

Théorème (Loi forte des grands nombres)

Soit X une variable aléatoire réelle suivant une loi de probabilité ayant pour densité f_X. Considérant un échantillon $(X_i)_{i \leq i \leq N}$ de réalisations indépendantes de même loi que X, si ϕ est une fonction mesurable tel que $\mathbb{E}[|\phi(X)|] < +\infty$ alors

$$\frac{1}{N} \sum_{i=1}^{N} \phi(X_i) \xrightarrow[N \to +\infty]{} \mathbb{E}[\phi(X)] = \int_{\mathbb{R}} \phi(x) f_X(x) dx$$

Estimation de la moyenne

- A partir d'un n-échantillon i.i.d. de X, l'espérance μ de X est estimée par

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

- $\bar{X}_n$ est un estimateur de l'espérance μ sans biais et convergent :

$$\begin{aligned}
\mathbb{E}(\bar{X}_n) &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mu \\
\mathrm{Var}(\bar{X}_n) &= \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\mathrm{Var}[\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \texttt{ car les } X_i \texttt{ sont indépendants} \\
&= \frac{1}{n^2} n\mathrm{Var}(X) = \frac{\mathrm{Var}(X)}{n} \to 0 \texttt{ lorsque } n \to \infty
\end{aligned}$$

- Retenir : la variance de la moyenne empirique $\bar{X}_n$ est n fois plus petite que la variance de X

$$\sigma^2_{\bar{X}_n} = \frac{1}{n}\sigma^2_X$$

Estimation de la variance

- Variance de X estimée par la variance empirique de l'échantillon :

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

- S_n^2 est un estimateur biaisé de la variance. Démonstration :

$$\begin{aligned}
\mathbb{E}(S_n^2) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - \mathbb{E}(\bar{X}_n^2) = \mathbb{E}(X^2) - \mathbb{E}(\bar{X}_n^2) \\
&= \mathrm{Var}(X) + [\mathbb{E}(X)]^2 - [\mathrm{Var}(\bar{X}_n) + [\mathbb{E}(\bar{X}_n)]^2] = \mathrm{Var}(X) - \mathrm{Var}(\bar{X}_n) \\
&= \mathrm{Var}(X) - \frac{1}{n}\mathrm{Var}(X) = \frac{n-1}{n}\mathrm{Var}(X) \quad \texttt{estimateur biaisé}
\end{aligned}$$

- L'estimateur $S_n'^2$ sans biais et convergent comme S_n^2

$$S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \texttt{estimateur non biaisé}$$

Estimation d'une probabilité

- On note p la probabilité pour que les valeurs de X soient inférieures à un seuil donné :

$$\mathbb{P}(X < \texttt{seuil}) = p$$

- On est dans le cas de l'estimation du paramètre p de la loi de Bernoulli $1_{X<\texttt{seuil}}$ qui vaut 1 si $X < \texttt{seuil}$ et 0 sinon. On a donc :

$$\mathbb{E}(1_{X<\texttt{seuil}}) = p \Rightarrow \mathbb{P}(X < \texttt{seuil}) = \mathbb{E}(1_{X<\texttt{seuil}})$$

- La probabilité p peut donc être estimée par la statistique P_n définie par la moyenne empirique de la variable aléatoire $1_{X<\texttt{seuil}}$ (loi de Bernoulli de paramètre p)

$$P_n = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i<\texttt{seuil}}$$

- L'estimateur est sans biais :

$$\begin{aligned}
\mathbb{E}(P_n) &= \mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} 1_{X_i<\texttt{seuil}}] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(1_{X_i<\texttt{seuil}}) \\
&= \frac{1}{n}\sum_{i=1}^{n} p = p
\end{aligned}$$

Estimation d'une probabilité (suite)

- Calcul de la variance de P_n

$$\begin{aligned}
\mathrm{Var}(P_n) &= \mathbb{E}[P_n^2] - [\mathbb{E}(P_n)]^2 = \mathbb{E}[(\frac{1}{n}\sum_{i=1}^{n} 1_{X_i<\texttt{seuil}})^2] - p^2 \\
&= \frac{1}{n^2}\sum_{i,j}\mathbb{E}(1_{X_i<\texttt{seuil}}1_{X_j<\texttt{seuil}}) - p^2 \\
&= \frac{1}{n^2}[\sum_{i}\underbrace{\mathbb{E}(1_{X_i<\texttt{seuil}})}_{1_{X_i<\texttt{seuil}}=(1_{X_i<\texttt{seuil}})^2} + \sum_{i\neq j}\underbrace{\mathbb{E}(1_{X_i<\texttt{seuil}})\mathbb{E}(1_{X_j<\texttt{seuil}})}_{\texttt{indépendance}}] - p^2 \\
&= \frac{1}{n^2}[n\times p + n(n-1)p^2] - p^2 = \frac{p(1-p)}{n}
\end{aligned}$$

- Remarque : la précision relative de l'estimateur donnée par le coefficient de variation :

$$\frac{\sqrt{\mathrm{Var}(P_n)}}{\mathbb{E}(P_n)} = \sqrt{\frac{1-p}{np}}$$

- Si $p = 10^{-\alpha},\ \alpha > 0$, estimer p avec une précision relative de 10% nécessite $n = 10^{\alpha+2}$ tirages de X.

Maximum de vraisemblance

- Pour un n-échantillon $(X_i)_{1\leq i\leq n}$, on se donne un modèle paramétrique pour la loi de probabilité $p_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ où le paramètre θ est inconnu
- Principe du Maximum de Vraisemblance
  - on dispose d'un n-échantillon $(x_i)_{1\leq i\leq n}$
  - la valeur la plus probable de θ est la valeur pour laquelle la probabilité d'observer $(X_i = x_i)_{1\leq i\leq n}$ est la plus forte.
- La vraisemblance est considérée comme fonction de θ. Une notation faisant apparaître le conditionnement du paramètre aux données :

$$\mathcal{L}(\theta|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- Estimateur de θ par Maximum de Vraisemblance

$$\theta_{MV} = \arg\max_\theta \mathcal{L}(\theta|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

- Cas des v.a. réelles: Lorsque les variables X_i sont i.i.d. de densité de probabilité $f_\theta(x)$, la vraisemblance se définit par:

$$\boxed{\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i)}$$

- Il sera plus simple de traiter le logarithme de la vraisemblance :

$$\theta_{MV} = \arg\max_\theta \ln \mathcal{L}(\theta|x_1, x_2, \ldots, x_n) \rightarrow \theta_{MV} = \arg\max_\theta \sum_{i=1}^n \ln f_\theta(x_i)$$

Exemple - Loi normale moyenne et variance inconnues

- Echantillon $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ i.i.d. tel que $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. La vraisemblance est :

$$\mathcal{L}(\mu, \sigma^2 | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x_i - \mu)^2}{2\sigma^2}$$

Exercice: calcul de l'EMV des paramètres μ et σ

Exemple - Loi normale moyenne et variance inconnues

- Echantillon $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ i.i.d. tel que $X_1 \sim \mathcal{N}(\mu, \sigma^2)$. La vraisemblance est :

$$\mathcal{L}(\mu, \sigma^2 | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x_i - \mu)^2}{2\sigma^2}$$

  Exercice: calcul de l'EMV des paramètres μ et σ

- Estimateur en minimisant l'opposé du logarithme de la vraisemblance (plus simple)

$$(\mu_{MV}, \sigma^2_{MV}) = \arg \min_{\mu, \sigma^2} [n \log \sqrt{2\pi\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}]$$

- Solution: par calcul, on retrouve les estimateurs empiriques usuels :

$$\mu_{MV} = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \sigma^2_{MV} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_{MV})^2$$

Exemple - Loi de Bernoulli de paramètre inconnu

- Echantillon $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ i.i.d. d'une loi de Bernoulli $X_1 \sim \mathcal{B}(1, \theta)$.
- Les x_i sont dans $\{0, 1\}$, $\mathbb{P}(X_1 = 1) = \theta$ et $\mathbb{P}(X_1 = 0) = 1 - \theta$. On a donc :

$$\mathbb{P}(X_1 = x) = p_\theta(x) = \theta^x (1 - \theta)^{1-x}$$

- La vraisemblance de θ conditionnellement à l'échantillon D est :

$$\mathcal{L}(\theta | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{k=1}^{n} \theta^{x_k} (1 - \theta)^{1 - x_k}$$

- Considérons la fonction log-vraisemblance (plus facile à manipuler) :

$$\begin{aligned}
\ell(\theta | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= \ln \mathcal{L}(\theta | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\
&= \sum_k [x_k \log \theta + (1 - x_k)) \log(1 - \theta)] \\
\frac{\partial}{\partial \theta} \ell(\theta | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) &= \sum_{k=1}^{n} [\frac{x_k}{\theta} - \frac{1 - x_k}{1 - \theta}] = \frac{1}{\theta(1 - \theta)} \sum_{k=1}^{n} (x_k - \theta) \\
\frac{\partial}{\partial \theta} \ell(\hat{\theta} | X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = 0 \Rightarrow \hat{\theta} &= \frac{1}{n} \sum_{k=1}^{n} x_k
\end{aligned}$$

- On retrouve l'estimateur *classique* de la moyenne empirique
- On vérifie la condition du maximum $\frac{\partial^2}{\partial \theta^2} \ell(\hat{\theta} | x_1, x_2, \ldots, x_n)) < 0$

Score et information de Fisher

- Soit une v.a. réelle X de loi p_θ ayant f_θ comme densité de probabilité.
- on note $\ell_\theta = \ln f_\theta$, la fonction

$$s_\theta = \frac{\partial \ell_\theta}{\partial \theta} ,$$

  est appelé score ou également fonction score.
- Remarques:

$$s_\theta = \frac{\dot{f}_\theta}{f_\theta} ,$$

  où $\dot{f}_\theta = \frac{\partial f_\theta}{\partial \theta}$.

$$\mathbb{E}_{X \sim P_\theta}[s_\theta(X)] = 0 \text{ démonstration en exercice}$$

Score et information de Fisher

- Soit une v.a. réelle X de loi p_θ ayant f_θ comme densité de probabilité.
- on note $\ell_\theta = \ln f_\theta$, la fonction

$$s_\theta = \frac{\partial \ell_\theta}{\partial \theta} ,$$

est appelé score ou également fonction score.

- Remarques:

$$s_\theta = \frac{\dot{f}_\theta}{f_\theta} ,$$

où $\dot{f}_\theta = \frac{\partial f_\theta}{\partial \theta}$.

$$\mathbb{E}_{X \sim P_\theta}[s_\theta(X)] = 0 \quad \texttt{démonstration en exercice}$$

- La variance de s_θ est appelé information de Fisher en θ et notée I_θ:

$$I_\theta = \mathbb{E}[s_\theta(X)^2] - \underbrace{\mathbb{E}[s_\theta(X)]^2}_{=0} = \mathbb{E}[s_\theta(X)^2]$$

- Exercice: Montrer que $I_\theta = \mathbb{E}[(\frac{\partial s_\theta}{\partial \theta})^2] = -\mathbb{E}[\frac{\partial^2}{\partial \theta^2} \ln f_\theta(X)]$.

Efficacité d'un estimateur - Borne de Cramèr-Rao

- Information de Fisher sur θ de l'échantillon est (si elle existe)

$$I_n(\theta) = \mathbb{E}[(\frac{\partial}{\partial\theta} \ln \mathcal{L}(\theta|X_1, X_2, \ldots, X_n))^2]$$

- Si le domaine de définition des X_i ne dépend pas de θ on montre (voir planche précédente)

$$I_n(\theta) = -\mathbb{E}[\frac{\partial^2}{\partial\theta^2} \ln \mathcal{L}(\theta|X_1, X_2, \ldots, X_n)]$$

- Pour tout estimateur T_n non biasé, l'inégalité de Cramèr-Rao donne une borne de sa variance

$$\mathrm{Var}(T_n) \geq \frac{1}{I_n(\theta)}$$

- Un estimateur est efficace si la borne de Cramèr-Rao est atteinte

$$\mathrm{Var}(T_n) = \frac{1}{I_n(\theta)}$$

Principales propriétés

- L'estimateur par maximum de vraisemblance est asymptotiquement sans biais et efficace

$$(\theta_{MV} - \theta) \to \mathcal{N}(0, \mathrm{Var}(\theta_{MV}) = \frac{1}{I_n(\theta)})$$

- Remarque : Si les tirages sont i.i.d., par linéarité de l'espérance et des dérivées, on a

$$I_n(\theta) = nI_1(\theta) \Rightarrow (\theta_{MV} - \theta) \to \mathcal{N}(0, \mathrm{Var}(\theta_{MV}) = \frac{1}{nI_1(\theta)})$$

- Dans le cas multi dimensionnel pour θ, l'information de Fisher est la matrice des covariances. Et si les tirages sont i.i.d. :

$$(I_n(\theta))_{i,j} = -n \times \mathrm{Cov}[\frac{\partial}{\partial\theta_i} \ln \mathcal{L}(\theta; X) \frac{\partial}{\partial\theta_j} \ln \mathcal{L}(\theta; X)]$$

- Exercice : utiliser ces propriétés pour l'exemple précédent $(X_i)_{1 \leq i \leq n} \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.

$$I_n(\mu, \sigma^2) = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

- Noter que les covariances sont nulles. Les variances des estimateurs sont donc :

$$\mathrm{Var}(\mu_{MV}) = \frac{\sigma^2}{n}, \quad \mathrm{Var}(\sigma^2_{MV}) = \frac{2\sigma^4}{n}$$

- Exercice (plus facile) : pour la loi de Bernoulli, on (re)trouve la variance $\theta(1-\theta)/n$ de l'estimateur $\hat{\theta}$ du paramètre θ.

Sommaire

Problématique et définition

On a un modèle statistique $\mathcal{M} = \{p_\theta, \theta \in \Theta\}$. On a des données $(X_i)1 \leq i \leq n$ i.i.d. tel que $X_1 \sim p_{\theta_*}$. On possède un estimateur $\widehat{\theta}_n$ du paramètre θ

Même si $\widehat{\theta}_n$ possède de bonnes propriétes (sans biais, variance minimale,...). Il est très peu probable que $\widehat{\theta}_n = \theta_*$

Il est naturel de proposer une estimation **ensembliste** de la valeur de θ plutôt qu'une simple estimation ponctuelle $\widehat{\theta}_n$

Un intervalle de confiance de seuil $\alpha \in (0, 1)$ pour le paramètre θ est un intervalle aléatoire I tel que $\mathbb{P}(\theta_* \in I) = 1 - \alpha$

Une définition subtile !

⚠ Il est facile de mal interpréter la notion d'intervalle de confiance !

Dans l'écriture $\mathbb{P}(\theta_* \in I)$, θ_* est une valeur inconnue mais **déterministe** (paradigme *fréquentiste*), ce sont les bornes de I qui sont aléatoires !

En pratique, à partir d'un échantillon $(X_i)_{1 \leq i \leq n}$ on détermine une statistique $\varepsilon(X_1, \ldots, X_n)$ tel que $I = [\widehat{\theta}_n - \varepsilon(X_1, \ldots, X_n), \widehat{\theta}_n + \varepsilon(X_1, \ldots, X_n)]$.

Difficulté: Trouver une statistique $\varepsilon(X_1, \ldots, X_n)$ tel que la loi de probabilité de $\widehat{\theta}_n \pm \varepsilon(X_1, \ldots, X_n)$ ne dépende pas de θ_*.

Exemple: moyenne d'une Gaussienne - variance connue

On dispose d'un n-échantillon $(X_i)_{1\leq i\leq n}$ i.i.d. tel que $X_1 \sim \mathcal{N}(\mu_*, \sigma^2)$.

On suppose σ connu, on cherche à estimer la moyenne μ et disposer d'un intervalle de confiance de seuil $\alpha \in (0,1)$.

$$\bar{X}_n \sim \mathcal{N}(\mu_*, \sigma^2/n) \iff U = \sqrt{n}\left(\frac{\bar{X}_n - \mu_*}{\sigma}\right) \sim \mathcal{N}(0,1)$$

$$\mathbb{P}(|\bar{X}_n - \mu_*| < \varepsilon) = \mathbb{P}(|U| < \frac{\sqrt{n}\varepsilon}{\sigma}) = 1 - \mathbb{P}(|U| > \frac{\sqrt{n}\varepsilon}{\sigma})$$

On note $u_{(1+\alpha)/2}$ le quantile de niveau $(1+\alpha)/2$ de la loi $\mathcal{N}(0,1)$, alors $\varepsilon = \frac{\sigma}{\sqrt{n}} u_{(1+\alpha)/2}$.

$$\boxed{\mathbb{P}(\bar{X}_n - \tfrac{\sigma}{\sqrt{n}} u_{(1+\alpha)/2} < \mu_* < \bar{X}_n + \tfrac{\sigma}{\sqrt{n}} u_{(1+\alpha)/2}) = 1 - \alpha}$$

⚠ Dans la majorité des cas, σ est inconnu...

Théorème de Fisher

- Soient $X_1, X_2, \ldots, X_n$ n variables indépendantes de loi normale $\mathcal{N}(\mu, \sigma^2)$
- Rappelons les estimateurs de la moyenne et et de la variance

$$\bar{X}_n = \frac{1}{n}\sum_i^n X_i$$

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

- On a les propriétés suivantes :
  1. $\bar{X}_n \sim \mathcal{N}(\mu, \sigma^2/n)$
  2. nS_n^2/σ^2 suit une loi du χ^2 à $n-1$ degrés de liberté notée χ^2_{n-1}
  3. $\bar{X}_n, S_n^2$ indépendants
  4. $\frac{\bar{X}_n - \mu}{S_n/\sqrt{n-1}}$ suit une loi de Student notée $\mathcal{T}_{n-1}$

Loi du χ^2 ("chi-deux")

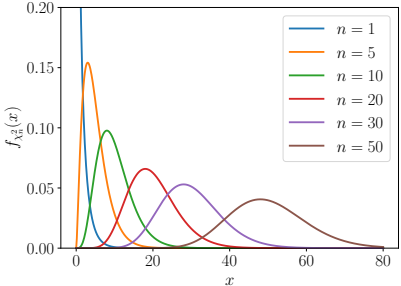
- Rôle important en statistique (test du χ^2)
- n variables aléatoires $X_i \sim \mathcal{N}(0,1)$ indépendantes,
- $\chi^2_n = \sum_{i=1}^n X_i^2$ variable du χ^2 à n degrés de liberté.

$$\mathbb{E}(\chi^2_n) = n \qquad \mathrm{Var}(\chi^2_n) = 2n$$

- Densité de probabilité

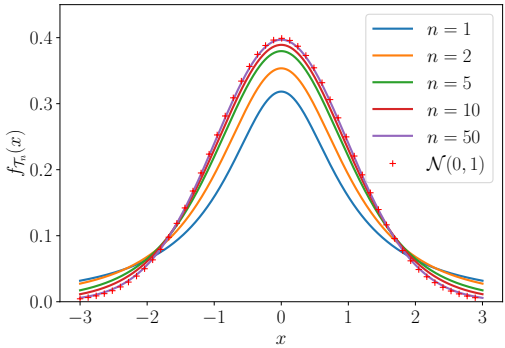
$$f_{\chi^2_n}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{n/2-1}\ 1_{x\geq 0}$$

$$\Gamma(x) = \int_0^x e^{-t} t^{x-1} dt$$



Loi de Student

- Soient $X \sim \mathcal{N}(0,1)$ et χ^2_n une variable du chi-deux à n degrés de liberté indépéndante de X
- La variable $\mathcal{T}_n = \frac{X}{\sqrt{\chi^2_n / n}}$ suit une loi de Student à n degrés de liberté
- La loi de Student tend rapidement vers la loi normale centrée réduite.
- Pour $n > 50$ elle peut-être approchée par la loi normale avec précision.



Exemple: moyenne d'une Gaussienne - variance inconnue

- Par le théorème de Fisher, on a :

$$\frac{\bar{X}_n - \mu_*}{S_n/\sqrt{n-1}} \sim \mathcal{T}_{n-1} \quad \texttt{loi de Student à n-1 degrés de liberté}$$

- Pour déterminer les bornes de l'intervalle de confiance, il suffit donc de remplacer les quantiles de la loi normale par ceux de la loi de Student à $n-1$ degrés de liberté.
- Notons $t_{n-1,\delta}$ les quantiles de la loi de Student $\mathcal{T}_{n-1}$ de niveau δ

$$\begin{aligned} \mathbb{P}\left(\frac{\bar{X}_n - \mu_*}{S_n/\sqrt{n-1}} < t_{n-1,(1+\alpha)/2}\right) &= \frac{1+\alpha}{2} \\ \mathbb{P}\left(\frac{\bar{X}_n - \mu_*}{S_n/\sqrt{n-1}} < t_{n-1,(1-\alpha)/2}\right) &= \frac{1-\alpha}{2} \end{aligned}$$

- La loi de Student étant symétrique, les quantiles en $(1+\alpha)/2$ et $(1-\alpha)/2$ sont symétriques (comme dans le cas de la loi normale).
- L'intervalle de confiance de niveau α est donc :

$$\boxed{\mathbb{P}(\bar{X}_n - t_{n-1,\frac{1+\alpha}{2}}\frac{S_n}{\sqrt{n-1}} < \mu_* < \bar{X}_n + t_{n-1,\frac{1+\alpha}{2}}\frac{S_n}{\sqrt{n-1}}) = 1-\alpha}$$

Intervalle de confiance pour la moyenne - cas général

Dans le cas général où la loi de X est inconnue, il n'y a pas de méthode pour calculer les bornes de l'intervalle de confiance

Grâce au TCL (théorème central limite), on sait que l'estimateur $\bar{X}_n$ tend vers une loi normale (somme de variables aléatoires i.i.d.)

On utilise donc cette propriété pour approcher les bornes de l'intervalle

$$\mathbb{P}(\bar{X}_n - t_{n-1,\frac{1+\alpha}{2}} \frac{S_n}{\sqrt{n-1}} < \mu_* < \bar{X}_n + t_{n-1,\frac{1+\alpha}{2}} \frac{S_n}{\sqrt{n-1}} \simeq 1-\alpha$$

L'approximation est d'autant plus précise que n est grand. Et compte tenu de la convergence de la loi de Student vers la loi normale, on pourra utiliser les quantiles de la loi normale dès que $n >\sim 100$

$$n >\sim 100 \Rightarrow \quad \mathbb{P}(\bar{X}_n - u_{\frac{1+\alpha}{2}} \frac{S_n}{\sqrt{n-1}} < \mu_* < \bar{X}_n + u_{\frac{1+\alpha}{2}} \frac{S_n}{\sqrt{n-1}}) \simeq 1-\alpha$$

Sommaire

Définition

"La **métrologie** est la science de la mesure. Elle définit les principes et les méthodes permettant de garantir et maintenir la confiance envers les mesures résultant des processus de mesure." (Wikipédia)

Définition

"La **métrologie** est la science de la mesure. Elle définit les principes et les méthodes permettant de garantir et maintenir la confiance envers les mesures résultant des processus de mesure." (Wikipédia)

La variabilité des valeurs obtenus lors d'un mesurage traduit **l'incertitude de mesure**.

Erreur de mesure

En physique, on effectue un ensemble d'opérations dans le but de déterminer une grandeur physique nommé **mesurande**.

Pendant l'opération de mesurage, il se produit forcément des erreurs de mesure.

Si on cherche à déterminer expérimentalement une grandeur x. On présentera le résultat sous la forme

$$x_m \pm \Delta x \ ,$$

où x_m est la valeur mesurée et Δx la demi-largeur d'un intervalle de confiance.

Pour déterminer Δx on va déterminer **l'incertitude type** u_x.

Deux types d'incertitudes

L'incertitude de type A est issue d'un traitement statistique des mesures.

L'incertitude de type B concerne les causes d'erreurs due au protocole expérimental et au matériel de mesure.

Incertitude de type B

On note q la **résolution** de l'appareil de mesure (exemple: plus petite graduation d'un appareil analogique).

la valeur vraie x_v de la grandeur à mesurer est donc dans l'interval suivant

$$x_v \in [x_{\text{mes}} - q/2, x_{\text{mes}} + q/2] \ ,$$

On peut également définir la **tolérance** t de l'instrument de mesure tel que:

$$x_v \in [x_{\text{mes}} - t, x_{\text{mes}} + t] \ ,$$

On considère que toutes les valeurs dans l'intervalle sont équiprobables $\Longrightarrow$ modélisation de l'incertitude par une loi uniforme.

Incertitude de type B

On note X_v une v.a. tel que $X_v \sim \mathcal{U}([x_{\text{mes}} - q/2, x_{\text{mes}} + q/2])$. Dans ce cas, on définit l'incertitude-type de type B u_B tel que:

$$u_B^2 = \text{Var}(X_v) = \frac{q^2}{12} = \frac{4t}{12} = \frac{t}{3}$$

Incertitude de type B

Dans le cas d'un écart entre deux mesures x_1 et x_2, On modèlise X_1 et X_2 comme deux variables aléatoires indépendantes tel que $X_1 \sim \mathcal{U}([x_1 - q/2, x_1 + q/2])$ et $X_2 \sim \mathcal{U}([x_2 - q/2, x_2 + q/2])$ et on a:

$$u_B^2 = \mathrm{Var}(X_1 - X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \frac{q^2}{6}$$

Incertitude de type A

On répète n fois l'expérience, ce qui donne un n-échantillon $(x_i)_{1\leq i\leq n}$ et on estima la valeur vraie par la moyenne empirique !

$$\widehat{\sigma_n}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X}_n)^2$$

L'incertitude de type A u_A est donné par

$$u_A = \frac{\widehat{\sigma_n}}{\sqrt{n}}$$

Propagations des incertitudes

L'incertitude-type totale est obtenu par cumul quadratique

$$u_T^2 = u_A^2 + u_B^2$$

Quand on doit calculer une incertitude-type sur une grandeur physique y tel que $y = f(x_1, \ldots, x_p)$, on utilise la formule.

$$\boxed{u_y^2 = \sum_{i=1}^{p} \left(\frac{\partial f}{\partial x_i} u_{x_i} \right)^2}$$

Construire un intervalle de confiance

Comment déterminer Δx à partir de u_x ?

$$\boxed{\Delta x = k u_x}$$

Le choix de k se fait à l'aide de la notion d'intervalle de confiance ! (Hypothèse Gaussienne)

Références du cours

- Notes de cours ENSIMAG 1ere année de O. Gaudoin, *Principes et méthodes statistiques*, Chapitres 1 à 4, `https://membres-ljk.imag.fr/Olivier.Gaudoin/PMS.pdf`
- Notes de cours de L. Pietri, classe de PC du Lycée Joffre à Montpellier `http://pcjoffre.fr/Data/cours/A1_incertitude.pdf`
- Site web Culture Sciences Chimie `https://culturesciences.chimie.ens.fr/thematiques/chimie-experimentale/les-incertitudes-de-type-a-et-b-en-chimie-application-a-un-dosage`