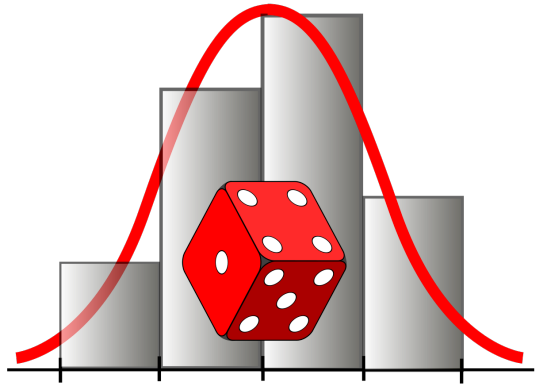
Estimation paramétrique ponctuelle & distribution d'échantillonnage



Module : Techniques d'estimation pour l'ingénieur

Plan

1. Introduction
   - Notion d'échantillonnage
   - Notion d'estimation paramétrique
2. Estimation paramétrique ponctuelle
   - Qualités des estimateurs
   - Méthodes d'estimation ponctuelle
     - Estimateur par méthode des moments (EMM)
     - Estimateur par la méthode du maximum de vraisemblance (EMV)
4. Distribution d'échantillonnage
   - Distribution échantillonnale de la moyenne $\overline{X_n}$
   - Distribution échantillonnale de la variance S_n^2
   - Distribution échantillonnale de la proportion F

esprit Se former autrement

Contexte général

En statistique, on décrit un échantillon ou une population à l'aide des mesures ou caractéristiques telles que la moyenne, l'écart-type, le pourcentage. De ce fait, s'il s'agit d'un :

-**Caractère quantitatif:** on estimera la moyenne μ et l'écart type σ d'une population.

-**Caractère qualitatif:** on estimera la proportion p de la population.

Les estimateurs ponctuels considerés pour l'estimation des paramètres inconnus : l'espérance μ , la variance σ^2 et la proportion p sont bien évidemment des variables aléatoires (statistiques) dont ils possèdent leurs propres lois (distributions).
On rappelle par la suite ces estimateurs ponctuels :

Distribution échantillonnale

A retenir

Soit $(X_1, \ldots, X_n)$ un échantillon de taille $n > 1$, qui suit une loi $\mathbb{P}$ (quelconque) d'espérance μ et de variance σ^2, et soit $(x_1, \ldots, x_n)$ une réalisation de cet échantillon.

- $\overline{X}_n := \frac{X_1 + \ldots + X_n}{n}$: **moyenne empirique** qui est un estimateur de m, son estimation $\overline{x}_n$ est la moyenne observée dans une réalisation de l'échantillon .
- $S_n^2 := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$: **variance empirique corrigée** qui est un estimateur de variance observée dans une réalisation de l'échantillon.
- $\hat{p}_n := \frac{X_1 + \ldots + X_n}{n}$: **proportion empirique** (notée aussi F) est un estimateur de la proportion p, avec $p \in]0, 1[$ représente la fréquence d'apparition d'un caractère qualitatif dans la population (ce caractère on le modèlise à chaque fois par la variable aléatoire $X_i \sim \mathcal{B}(p)$).

Loi de la moyenne empirique $\overline{X}_n$

On a défini une variable aléatoire qui à chaque $n-$échantillon associe sa moyenne échantillonnnale. On la note $\overline{X}_n$.
On cherche à caractériser la variable aléatoire $\overline{X}_n$ par :

- Sa moyenne.
- Sa variance.
- Sa distribution de probabilité.

Loi de la moyenne empirique $\overline{X_n}$

Espérance et variance de $\overline{X_n}$

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant X d'espérance μ et d'écart-type σ. La moyenne empirique de n échantillons aléatoires est définie par :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{avec} \quad \mathbb{E}[\overline{X}_n] = \mu \quad \text{et} \quad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}$$

En effet :

$$\begin{aligned} E[\overline{X_n}] &= E[\frac{1}{n}\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n}.E[\sum_{i=1}^{n} X_i] \\ &= \frac{1}{n}.\sum_{i=1}^{n} E[X_i] = \frac{1}{n}.n.\mu = \mu \end{aligned}$$

Loi de la moyenne empirique $\overline{X_n}$

Et :

$$
\begin{aligned}
V[\overline{X_n}] &= V[\frac{1}{n}\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}.V[\sum_{i=1}^{n} X_i] \\
&= \frac{1}{n^2}.\sum_{i=1}^{n} V[X_i] \\
&= \frac{1}{n^2}.n.\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}
$$

Pour déterminer la distribution de probabilité de $\overline{X}_n$, nous allons distinguer deux cas : celui des grands échantillons ($n \geq 30$) et celui des petits échantillons ($n < 30$).

Loi de la moyenne empirique $\overline{X_n}$

- **Cas des grands échantillons ($n \geq 30$)**

On commençe par annoncer le théorème fondamental suivant:

Le théorème central-limite (TCL)

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant la loi de X d'espérance μ et d'écart-type σ.

Alors, pour n est assez grand ($n \geq 30$), $Y = \sum_{i=1}^{n} X_i$ suit une loi normale de moyenne $n\,\mu$ et d'écart-type $\sqrt{n}\,\sigma$.

$$Y \sim \mathcal{N}(n\,m, \sqrt{n}\,\sigma)$$

En appliquant le théorème central limite, la loi normale est une bonne approximation de la loi de $\overline{X}_n$:

Loi de la moyenne empirique $\overline{X_n}$

Proposition

Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées suivant X d'une loi quelconque d'espérance μ et d'écart-type σ. La moyenne empirique de n échantillons aléatoires est défini par :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{avec} \quad \mathbb{E}[\overline{X}_n] = \mu \quad \text{et} \quad \mathbb{V}[\overline{X}_n] = \frac{\sigma^2}{n}$$

De plus, quand n est assez grand($n \geq 30$),

$$\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}}) \quad \text{donc} \quad Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

Loi de la moyenne empirique $\overline{X_n}$

Remarque

Si la variance σ^2 est inconnue, il suffit de l'estimer par

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2$$

On aura donc,

$$\overline{X}_n \sim \mathcal{N}(\mu, \frac{s_n}{\sqrt{n}}) \quad \text{donc} \quad Z = \frac{\overline{X}_n - \mu}{\frac{s_n}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

Loi de la moyenne empirique $\overline{X_n}$

Exemple

Soit un lot de 500 chocolats. Le poids d'un chocolat est une variable aléatoire d'espérance $\mu = 5g$ et de variance $\sigma^2 = 0.5g$.
Quelle est la probabilité qu'une boite de 50 chocolats issus de ce lot ait un poids moyen supérieur à $5.2g$?

Solution

On a $n = 50 > 30$, et $X \sim \mathcal{L}(\mu = 5; \sigma^2 = 0.5)$, alors par le TCL :

$$\overline{X_n} \sim \mathcal{N}(\mu\ ; \frac{\sigma}{\sqrt{n}})$$

Et par la suite :

$$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

$$\begin{aligned}\mathbb{P}(\overline{X}_n \geq 5.2) &= \mathbb{P}(Z \geq \frac{5.2 - 5}{\sqrt{\frac{0,5}{50}}}) \\ &= \mathbb{P}(Z \geq 2) \\ &= 0.022\end{aligned}$$

(Lecture sur la table de la loi Normale).

Loi de la moyenne empirique $\overline{X_n}$

- **Cas des petits échantillons ($n < 30$)**

Nous nous plaçons alors exclusivement dans le cas où la population est normale: X suit une **loi normale** de moyenne m et de variance σ^2.
Nous allons encore distinguer deux cas : celui où σ est connu et celui où σ est inconnu.

σ connu:

Proposition

X suit une loi normale $\mathcal{N}(\mu;\sigma)$ donc les variables X_i suivent toutes la même loi que X, $\forall 1 \leq i \leq n$,

$$X_i \sim \mathcal{N}(\mu;\sigma) \text{ alors } \overline{X}_n \sim \mathcal{N}(\mu;\frac{\sigma}{\sqrt{n}}) \text{ donc } Z = \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$$

Loi de la moyenne empirique $\overline{X_n}$

Exemple

Le responsable d'une entreprise a accumulé depuis des années les résultats à un test d'aptitude à effectuer un certain travail. Les résultats au test d'aptitude sont distribués suivant une loi normale de moyenne égale 150 et de variance 100. On fait passer le test à 25 individus de l'entreprise. Quelle est la probabilité que la moyenne de l'échantillon soit entre 146 et 154 ?

Solution

On a $n = 25 < 30$, et $X \sim \mathcal{N}(\mu = 150; \sigma^2 = 100)$, alors :

$$\overline{X_n} \sim \mathcal{N}(\mu; \frac{\sigma}{\sqrt{n}})$$

Et par la suite :

$$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

$$\begin{aligned}
\mathbb{P}(146 \leq \overline{X_n} \leq 154) &= \mathbb{P}(\frac{146 - 150}{\frac{10}{\sqrt{25}}} \leq Z \leq \frac{154 - 150}{\frac{10}{\sqrt{25}}}) \\
&= \mathbb{P}(-2 \leq Z \leq 2) \\
&= 1 - \mathbb{P}(Z \geq 2) - \mathbb{P}(Z \geq 2) \\
&= 1 - 2\mathbb{P}(Z \geq 2) = 1 - 2.(0.022) = 0.956
\end{aligned}$$

(Lecture sur la table de la loi Normale)

Loi de la moyenne empirique $\overline{X_n}$

σ inconnu:

La variance σ^2 est inconnue, il suffit d'utiliser l'estimateur :

$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$ d'estimation : $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2$

Proposition

Dans le cas où σ est inconnu, nous allons utiliser la statistique définie par:

$$T = \frac{\overline{X}_n - \mu}{\frac{S_n}{\sqrt{n}}} \sim \mathcal{T}_{(n-1)} \text{ la loi de Student de } (n-1) \text{ degrés de liberté .}$$

Pour le cas $n < 30$ et la loi $\mathbb{P}$ est quelconque alors le TCL ne s'applique pas et on ne peut pas déduire la loi de $\overline{X}_n$.

Loi de la moyenne empirique $\overline{X_n}$

Exemple

On détecte que l'âge auquel apparaissent les premiers mots de vocabulaire chez l'enfant suit une loi normale de moyenne 12 mois et de variance inconnue. On prend un échantillon de taille 25 alors donner la probabilité que l'âge moyen dans l'échantillon soit supérieur ou égal à 14 mois sachant que la variance sur l'échantillon est égale à $12,8$ mois.

Solution

On a $n = 25 < 30$, et $X \sim \mathcal{N}(\mu = 12; \sigma^2 = ?)$, et la variance mesurée sur l'échantillon est $s_n^2 = 12.8$, alors la statistique :

$$T = \frac{\overline{X_n} - \mu}{\frac{S_n}{\sqrt{n}}} \sim \mathcal{T}_{n-1}$$

$$\begin{aligned}
\mathbb{P}(\overline{X_n} \geq 14) &= \mathbb{P}(T \geq \frac{14 - 12}{\sqrt{\frac{12.8}{25}}}) \\
&= \mathbb{P}(T \geq 2.79) \quad , \quad T \sim \mathcal{T}_{24} \\
&= 0.005,
\end{aligned}$$

(lecture de la table de la loi de student à 24 ddl)

Loi de la moyenne empirique $\overline{X_n}$

Conclusion La loi de la moyenne empirique $\overline{X_n}$ est donnée par ce tableau récapitulatif :

$n \geq 30$ & **Population de loi quelconque de moyenne m et de variance σ^2**		
Variance σ^2	$\overline{X_n}$	**Ecart réduit**
connue	$\overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$	$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$
inconnue, on utilise l'estimation $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2$	$\overline{X_n} \sim \mathcal{N}(\mu, \frac{s_n}{\sqrt{n}})$	$Z = \frac{\overline{X_n} - m}{\frac{s_n}{\sqrt{n}}} \sim \mathcal{N}(0,1)$
$n < 30$ & **Population normale de moyenne μ et de variance σ^2**		
Variance σ^2	$\overline{X_n}$	**Ecart réduit**
connue	$\overline{X_n} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$	$Z = \frac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$
inconnue, on utilise l'estimateur $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$	$T = \frac{\overline{X_n} - \mu}{\frac{S_n}{\sqrt{n}}} \sim T_{n-1}$	$T_{n-1} \sim$ Student de $(n-1)$ ddl

Distribution échantillonnale de la variance S_n^2

Distribution échantillonnale de la variance S_n^2:

Nous nous plaçons alors exclusivement dans le cas où la population est normale: X suit une **loi normale** de moyenne μ et de variance σ^2. On appelle variance empirique, la statistique notée S_n^2, on cherhce à caractériser S_n^2. Nous allons distinguer deux cas : celui où m est connue et celui m est inconnue.

Proposition : μ connue

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une variable aléatoire X qui suit une loi normale $\mathcal{N}(\mu, \sigma)$. Soit la variance échantillonnale S_n^2 définie par:

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

$$\Longrightarrow Y = \frac{n\, S_n^2}{\sigma^2} \sim \chi_n^2 \quad \text{suit une loi de Khi-deux avec } n \text{ degrés de liberté}$$

Distribution échantillonnale de la variance S_n^2

Exemple

Une étude a montré que le taux des chômeurs tunisiens suit une loi normale d'espérance $\mu = 20$ et de variance $\sigma^2 = 42,67$. Un échantillon de taille 25 a été prélevé, alors donner la probabilité que la variance moyenne des chômeurs soit supérieur où égale à 22.39.

Distribution échantillonnale de la variance S_n^2

Solution

On a $n = 25$, et $X \sim \mathcal{N}(\mu = 20; \sigma^2 = 42.67)$, alors la statistique :

$$Y = \frac{n.S_n^2}{\sigma^2} \sim \chi_n^2$$

$$\begin{aligned}
\mathbb{P}(S_n^2 \geq 22.39) &= \mathbb{P}(\frac{n.S_n^2}{\sigma^2} \geq \frac{22.39 \times 25}{42.67}) \\
&= \mathbb{P}(Y \geq 13.12) \quad , \quad Y \sim \chi_{25}^2 \\
&= 0.975
\end{aligned}$$

(lecture de la table de la loi χ_{25}^2)

Distribution échantillonnale de la variance S_n^2

Proposition : m inconnue

Soit $X_1, \ldots, X_n$ un échantillon aléatoire d'une variable aléatoire X qui suit une loi normale $\mathcal{N}(\mu, \sigma)$. Soit la variance échantillonnale S^2 définie par:

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 \quad \text{avec} \quad \mathbb{E}(S_n^2) = \sigma^2$$

On a,

$$Y = \frac{(n-1)\, S_n^2}{\sigma^2} \quad \text{suit une loi du Khi-deux avec } (n-1) \text{ degrés de liberté}$$

Distribution échantillonnale de la variance S_n^2

Exemple

On fait l'hypothèse que la taille (en cm) des 3000 étudiants masculins d'ESPRIT est une variable aléatoire distribuée normalement de moyenne inconnue et de variance 100. Un échantillon de taille 25 est sélectionné de cette population. Quelle est la probabilité que la variance échantillonnale S_n^2 soit au plus égale 151,72 ?

Solution

On a $n = 25$, et $X \sim \mathcal{N}(\mu = ?; \sigma^2 = 100)$, alors la statistique :

$$Y = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\begin{aligned}
\mathbb{P}(S_n^2 \leq 151.72) &= \mathbb{P}\left(\frac{(n-1)S_n^2}{\sigma^2} \leq \frac{151.72 \times (25-1)}{100}\right) \\
&= \mathbb{P}(Y \leq 36.41) \quad , \quad Y \sim \chi^2_{24} \\
&= 1 - \mathbb{P}(Y \geq 36.41) \\
&= 1 - 0.05 = 0.95
\end{aligned}$$

(lecture de la table de la loi χ^2_{24})

Distribution échantillonnale de la variance S_n^2

Conclusion La loi de la variance empirique S_n^2 est donnée par ce tableau récapitulatif :

Cas possibles	S_n^2 estimateur de σ^2	Ecart réduit
μ connue	$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$	$Y = \frac{n.S_n^2}{\sigma^2} \sim \chi_n^2$
μ inconnue, $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$	$Y = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{n-1}^2$

Disribution échantillonnale de la proportion F

Soit $X_1, \ldots, X_n$ un échantillon aléatoire, tel que $X_i \sim \mathcal{B}(p)$, $\forall\ 1 \leq i \leq n$.
Soit F la fréquence d'apparition d'un caractère qualitatif dans un échantillon de taille n, donc

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

où X est le nombre de fois où le caractère apparaît dans l'échantillon de taille n. Par définition X suit $\mathcal{B}(n; p)$.

$$X \sim \mathcal{B}(n; p) \qquad \text{avec} \qquad E[X] = np \quad \text{et} \quad V[X] = np(1-p)$$

Disribution échantillonnale de la proportion F

Proposition

Soit $X_1, \ldots, X_n$ un échantillon aléatoire, telque $X_i \sim \mathcal{B}(p) \quad , \forall \ 1 \leq i \leq n$, donc

$$F = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{avec} \quad E[F] = p \qquad \text{et} \qquad V[F] = \frac{p(1-p)}{n}$$

Si $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, alors,

$$F \ \sim \ \mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}}) \quad \text{et} \quad Z = \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \ \sim \ \mathcal{N}(0,1)$$

Disribution échantillonnale de la proportion F

Proposition

1. L'espérance de la fréquence d'échantillon est égale à la probabilité théorique d'apparition dans la population.
2. Lorsque la taille de l'échantillon augmente, la variance de F diminue, ce qui est logique: plus on a d'informations, plus il est probable que la proportion observée dans l'échantillon soit proche de la proportion de la population.

Disribution échantillonnale de la proportion F

Exemple

Une étude affirme que 30% des citoyens de la population tunisienne effectuent régulièrement des achats en ligne. Prenons alors un échantillon de 100 personnes, quelle est la probabilité qu'au moins 40% parmi ces interrogés font des achats en ligne?

Disribution échantillonnale de la proportion F

Solution

En utilisant le résultat ci-dessus, (hypothèse $n > 30$, $np > 5$ et $n(1-p) > 5$ est déjà vérifiée), on a :

$$\begin{aligned}
\mathbb{P}(F \geq 0,4) &= \mathbb{P}\left(\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{0,4-0,3}{\sqrt{\frac{0,3\times 0,7)}{100}}}\right) \\
&= \mathbb{P}\left(\frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \geq 2,18\right) \\
&= 0,01463
\end{aligned}$$

(Selon la table $\mathcal{N}(0,1)$) .

Disribution échantillonnale de la proportion F

Conclusion La loi de la variance empirique F est donnée par ce tableau récapitulatif :

$n \geq 30 \quad \& \quad np \geq 5 \quad \& \quad nq = n(1-p) \geq 5$	
Loi de F	**Ecart réduit**
$F \sim \mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}})$	$Z = \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{N}(0,1)$