Échantillonnage

Problème d'échantillonnage

On conna^la loi et les paramètres de la loi suivie par la population mère .

Type des Problèmes à résoudre :
Vérifier si l'échantillon donné appartient à cette population et avec quelle probabilité de se tromper.
C.à.d : Avec quel risque, et ce risque sera noté α.

Pour répondre on se fonde sur l'observation des échantillons.

Def : On appelle ***échantillon*** de taille n, ou n-***échantillon*** de la v.a. X, toute suite de v.a. $(X_1, X_2, ..., X_n)$ indépendantes et de même loi que X.

On note :

- $(X_1, X_2, ..., X_n)$ l'échantillon (la suite des v.a.i.i.d.)
- $(x_1, x_2, ..., x_n)$ une réalisation de l'échantillon.

Les caractéristiques de l'échantillon

1.1. ***Moyenne empirique ou expérimentale*** :

$$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$$

Propriétés de $\overline{X}$:

- $E(\overline{X}) = E\left(\frac{1}{n}\sum_{k=1}^{n} X_k\right) = E\left(X\right) = \mu_X$
- $V\left(\overline{X}\right) = \dfrac{\sigma_X^2}{n}$

1.2. *Variance Empirique ou Expérimentale*

- ***Variance empirique*** : $S^2 = \frac{1}{n} \sum_{k=1}^{n} \left(X_k - \overline{X} \right)^2$
- ***Variance empirique corrigée ou sans biais*** : $S'^2 = \frac{1}{n-1} \sum_{k=1}^{n} \left(X_k - \overline{X} \right)^2$

Propriétés de la variance empirique :

- $nS^2 = (n-1)S'^2$
- $E\left(S^2\right) = \dfrac{n-1}{n}\sigma_X^2$
- $E\left(S'^2\right) = \sigma_X^2$

2. Distributions d'échantillonnage

On appelle ***distributions d'échantillonnage*** les lois suivies par les caractéristiques de l'échantillon, c.à.d. la loi de la :

- Moyenne empirique $\overline{X}$
- Variance empirique S^2.
- Fréquence dans l'échantillon $f_n = \frac{X}{n}$

Pour étudier ces lois on doit d'abord définir les lois dites ***dérivées de la loi normale***, car elles se fondent sur la loi normale.

1. Lois dérivées de la loi normale

La loi du $\chi^2(\nu)$ (chi deux) ; ν degrés de liberté, noté d.d.l.

Soient $X_i \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right); i = 1, 2, \ldots, n$.

La somme des carrées des v.a. normales centrées et réduites,

$$U = \sum_{i=1}^{m} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(\nu)$$

La variable aléatoire U suit la loi du **chi deux** avec $\nu = n$ d.d.l.

On a :

$E(U) = \nu; \quad V(U) = 2\nu$

Loi de Student (ou t-distribution)

On suppose que :
$Z \sim \mathcal{N}(0,1)$ et $U \sim \chi^2(\nu)$

alors la v.a.

$$T = \frac{Z}{\sqrt{U\,/\,\nu}}$$

suit une loi de student $T \sim t(\nu)$.

$E(T) = 0; \quad V(T) = \frac{\nu}{\nu - 2}$

Loi de Fisher (loi f)

Soient $U_1 \sim \chi^2(\nu_1)$ et $U_2 \sim \chi^2(\nu_2)$ alors la v.a.

$$F = \frac{U_1/\nu_1}{U_2/\nu_2}$$

suit une loi de Fisher, notée $F \sim F(\nu_1, \nu_2)$.

Rq : La table de cette loi fournit les valeurs $F(\ p; \nu_1, \nu_2)$ pour $p \leq 0.5$

Si $p > 0.5$ on peut utiliser la même table car

$$F(p; \nu_1, \nu_2) = \frac{1}{F(1 - p; \nu_2, \nu_1)}$$

Lois suivies par les caractéristiques de l'échantillon ou distributions d'échantillonnage

On va étudier les cas de :

I Échantillons grands

II Échantillons petits mais la loi suivie dans l'échantillon est la loi normale.

Rappel :

Théorème 1. Th. central limite.
Soit la suite $(X_n)_n$ de v.a.i.i.d.

On pose : $S_n = \sum_{k=1}^{n} X_k$ *et* $Z = \dfrac{S_n - E(S_n)}{\sigma_{S_n}}$

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

I. Échantillons grands

1.1 Loi suivie par $\overline{X}$ quand la variance $V(X)$ de la population-mère est connue

$\overline{X} = \frac{1}{n} \sum_{k=1}^{n} X_k$; quand $n \to \infty$

On pose : $S_n = \frac{1}{n} \sum_{k=1}^{n} X_k$

On a : $E(S_n) = \mu$; $Var(S_n) = \sigma^2_{S_n} = \frac{\sigma^2}{n}$

En applIquant le **TCL** :

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

ou bien

$$\overline{X} \sim \mathcal{N}\,(\mu, \sigma^2_{\overline{X}}), \quad \sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

I. Échantillons grands (suite)

1.2 Loi suivie par $\overline{X}$ quand la variance σ^2 de la population-mère est inconnue

On utilisera la loi de Student.

On montre que

$$T = \sqrt{n-1}\frac{\overline{X}-\mu_X}{S} \sim t\left(\nu\right); \quad \nu = \mathrm{n}-1 \quad ; \mathrm{S}^2 = \frac{1}{\mathrm{n}}\sum_{\mathrm{k}=1}^{\mathrm{n}}\left(\mathrm{X_k}-\overline{\mathrm{X}}\right)^2$$

Ou bien

$$T = \sqrt{n}\frac{\overline{X}-\mu_X}{S'} \sim t\left(\nu\right); \quad \nu = \mathrm{n}-1; \quad \mathrm{S}'^2 = \frac{1}{\mathrm{n}-1}\sum_{\mathrm{k}=1}^{\mathrm{n}}\left(\mathrm{X_k}-\overline{\mathrm{X}}\right)^2$$

Rq : les démonstrations se feront en TD

On démontre aussi que $\overline{X}$ et S^2 sont des variables aléatoires indépendantes.

I. Échantillons grands (suite)

2. Loi suivie par la variance empirique S^2

On démontre que :

$$(n)\,\frac{S^2}{\sigma^2} \sim \chi^2(n-1)$$

Ou bien

$$(n-1)\,\frac{S'^2}{\sigma^2} \sim \chi^2(n-1)$$

Rq : Les démonstrations se feront en TD.

I. Échantillons grands (fin)

3. La loi de la fréquence f_n.

Soient les $v.a.$ $X_i \sim b\,(p)$ et $S_n = X = X_1 + X_2 + ... + X_n$.

On note $f_n = \frac{X}{n}$ la fréquence dans l'échantillon.

On a :

$E\,(f_n) = p$ et $V\,(f_n) = \frac{p(1-p)}{n}$

Par le T.C.L. on obtient : $\frac{f_n - E(f_n)}{\sqrt{V(f_n)}} = \frac{f_n - p}{\sqrt{\frac{p(1-p)}{n}}}$

et donc :

$$f_n \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \quad \sigma_{f_n}^2 = \frac{p(1-p)}{n}$$

II. Échantillons petits

- On sait résoudre le problème si la loi suivie dans l'échantillon est la loi normale.
  On utilise les lois dérivées de la loi normale.
- Si la loi n'est pas la loi normale on ne conna^pas toujours les lois des caractéristiques de l'échantillon.

Étapes à suivre

1. On définit le risque α
2. On identifie la loi suivie par la v.a. X dans l'échantillon
3. On se place dans les cas des grands ou petits échantillons.
4. On justifie la loi de $\overline{X}$
5. On établit une fourchette, une règle, dans laquelle $\overline{X}$ doit se trouver avec un risque $1 - \alpha$
6. On observe l'échantillon :
   - On calcule $\overline{x}$, (une réalisation de $\overline{X}$).
   - On calcule s^2, (une réalisation de S^2).

7. On vérifie si $\overline{x}$ appartient à l'intervalle, (la fourchette calculée).
8. Si oui, on décide que l'échantillon est conforme à la population-mère avec un risque α, sinon on décide que l'échantillon ne peut pas appartenir à la population en question.

Exemple

On prélève 50 pièces dans la production.
On avait réglé la machine à un diamètre de $\mu = 10$ cm et un écart-type de $\sigma^2 = 4$cm.

Pb : Est-ce que la machine produit toujours des pièces conformes aux normes ?

Étapes

1. On définit le risque $\alpha = 0.05$
2. $n = 50$, on utilisera le T.C.L.
3. On se place dans les cas des grands échantillons, σ^2 connu.
4. $\frac{\overline{X}-\mu}{\sigma}\sqrt{n} \sim \mathcal{N}(0,1)$

5. $P(a \leq \overline{X} \leq b) = P(\frac{a-\mu}{\sigma}\sqrt{n} \leq \frac{\overline{X}-\mu}{\sigma}\sqrt{n} \leq \frac{b-\mu}{\sigma}\sqrt{n}) = 1-\alpha$
   Donc $F(\frac{b-\mu}{\sigma}\sqrt{n}) - F(\frac{a-\mu}{\sigma}\sqrt{n}) = 2F(t) - 1 = 0.95$

   On a posé $t_1 = \frac{b-\mu}{\sigma}\sqrt{n}$ et $t_2 = \frac{a-\mu}{\sigma}\sqrt{n}$,$t_2 = -t_1$,
   (la loi normale étant symétrique)

   $F(t) = (1+0.95)/2$, par la table de la loi normale $t = 1.64$

   On obtient : $\mu - 1.64\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 1.64\frac{\sigma}{\sqrt{n}}$
   $\Rightarrow 9.34 \leq \overline{X} \leq 10.66$
6. On observe l'échantillon : Sur 50 pièces on a calculé $\overline{x} = 9.02$
7. Décision : La machine continue à produire des pièces de moyenne $\mu = 10$cm avec un risque de $\alpha = 0.05$

Faire la différence entre :

Modéle Théorique	et	Échantillon
paramètres de la loi		Il s'agit des v.a.
(valeurs certaines)		
$\mu = E(X)$		$\overline{X}$
$V(X)$		S^2
$E(X) = p$		f_n