Statistiques inférentielles

Cadre Données

- Ici X est une VAR attachée à une population $\mathscr{P}$. On pose $E(X) = \mu$, $V(X) = \sigma^2$.
- Le quantile d'ordre t de la loi $\mathscr{N}(0,1)$ est noté u_t.

- **Quantiles u_t utiles :**
  - $u_{0.975} =$ **1.96** (risque α =5%)
  - $u_{0.995} =$ **2.58** (risque α =1%)
  - $u_{0.999} =$ **3.29** (risque α =0.2%)
  - u_t : (risque à $2(100-t)$%)
  - $u_{1-\frac{\alpha}{2}}$: risque à α.

I Loi faible des grands nombres

[Inégalité de Markov] **Thm.1**

Si X est une variable aléatoire positive admettant un moment d'ordre 1 :

$$\forall \lambda > 0 \quad P(X \geq \lambda) \leq \frac{E(X)}{\lambda}$$

Rem.1 | Majore la probabilité qu'une VAR prenne de très grandes valeurs.

[Lemme] **Prop.1**

Si $X_1 \dots X_n$ sont n VAR deux à deux non corrélées admettant un moment d'ordre 2, et de même variance $\sigma^2 > 0$:

a. $\overline{X} := \dfrac{X_1 + \cdots + X_n}{n}$ admet un moment d'ordre 2.

b. $E(\overline{X}) = \dfrac{E(X_1) + \cdots + E(X_n)}{n}$ et $V(\overline{X}) = \dfrac{\sigma^2}{n}$.

Rem.2 | Si les X_i sont un n-échantillon d'une loi d'espérance μ, $E(\overline{X}) = \mu$.

[Loi faible des grands nombres] **Thm.2**

Si $X_1 \dots X_n$ sont n VAR deux à deux non corrélées et admettant un moment d'ordre 2, et de même variance σ^2, alors :

a. $\overline{X} := \dfrac{X_1 + \cdots + X_n}{n}$ admet un moment d'ordre 2.

b. Pour tout $\varepsilon > 0$:

$$P(|\overline{X} - E(\overline{X})| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \underset{n\to+\infty}{\to} 0.$$

II Vocabulaire de l'estimation

A) Fondamental

Les données du problème :

a. Une population $\mathscr{P}$ théorique. Ex : les femmes françaises entre 20 et 30 ans.

b. Une VAR X définie sur $\mathscr{P}$ décrivant un caractère quantitatif sur $\mathscr{P}$. Ex : la taille d'une femme dans $\mathscr{P}$.

c. La **loi de X** s'appelle **distribution du caractère X**. Ex : distribution des tailles.

Rem.3 | Cette loi est en général inconnue, ou partiellement inconnue.

[Moyenne, variance théoriques] **Déf.1**

a. L'**espérance de X**, μ s'appelle **moyenne théorique du caractère**. Ex : la taille moyenne des femmes françaises entre 20 et 30 ans est $E(X)$.

b. La **variance de X**, σ^2 s'appelle **la variance théorique du caractère**. Ex : variance des tailles.

En général le but est d'inférer la valeur de $\mu = E(X)$ (voire de σ^2) à partir d'échantillons de mesures de X sur une sous-population.

B) Grandeurs empiriques

Avec les notations précédentes, si $X_1, \dots, X_n$ est un n-échantillon de la loi de X :

a. $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ s'appelle **moyenne empirique**. C'est une VAR, et même un estimateur (*cf.* II D)).

b. La VAR définie par :

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

s'appelle **variance empirique**. C'est aussi un estimateur.

c. La loi de $\overline{X}$ s'appelle **distribution d'échantillonnage**. Celle de S_n^2 s'appelle aussi **distribution d'échantillonnage** (voir **Déf.2**). Bien entendu, ces VAR n'ont pas pour autant les mêmes lois !

C) Principe de l'inférence

[Principe de l'inféference statistique] **Thm.3**

Les valeurs mesurées du caractère X sur un échantillon de n individus prélevés convenablement[a] dans $\mathscr{P}$ sont les réalisations d'un n-échantillon de la loi de X.

a. On dit échantillon bernoullien ou représentatif.

Exple.1 | Par ce principe,si on a prélevé un échantillon représentatif de 30 femmes françaises entre 20 et 30 ans et mesuré leur tailles, la série statistique x des 30 valeurs observées : $x = (x_1, \dots, x_n)$, est une réalisation d'un 30-échantillon $(X_1, \dots, X_n)$ de la loi de X.

D) Estimateur

[Estimateur, distribution d'échantillonnage] **Déf.2**

Estimateur : toute fonction T des variables $X_1, \dots, X_n$ (n-échantillon de la loi de X). C'est donc aussi une variable aléatoire. Sa loi s'appelle **distribution d'échantillonage**.

Exple.2 |
La moyenne empirique, ainsi que la la variance empirique sont des estimateurs.

Rem.4 |

a. Tout le problème est de savoir de quoi ces estimateurs fournissent des estimations (voir **Exple. 4**).

b. Suivant les besoins, on doit : **1)** fabriquer des estimateurs spécifiques, et **2)** vérifier que ceux-ci fournissent une estimation du paramètre à estimer.

Exple.3 | Pour estimer les coefficients d'une droite de régression (qui ne serait pas la même à partir d'un autre échantillon de mesures), on **1)** construit l'estimateur des moindres carrés et **2)** on calcule son espérance.

[Biais d'un estimateur] **Déf.3**

Si θ est le paramètre à estimer dans la population $\mathscr{P}$, et si T est un estimateur fabriqué à partir d'un n-échantillon de X, T est dit **sans biais** si $E(T) = \theta$.

Exple.4

a. La moyenne empirique est un estimateur sans biais de la moyenne théorique (**Déf. 1**) puisque $E(\overline{X}) \overset{\text{Prop. 1b.}}{=} \mu$.

b. La variance empirique est un estimateur biaisé de la variance théorique (**Déf. 1**) puisqu'un calcul montre que $E(S_n^2) = \dfrac{n-1}{n}\sigma^2 < \sigma^2$.

III Estimation ponctuelle

[Estimation ponctuelle] **Déf.4**

Toute réalisation d'un estimateur est une estimation ponctuelle.

Exple.5

a. La taille moyenne $\overline{x}$ calculée sur un échantillon représentatif $(x_1 \dots x_{30})$ des tailles de 30 femmes françaises est une estimation ponctuelle fournie par l'estimateur *moyenne empirique* $T = \dfrac{X_1 + \cdots + X_{30}}{30}$. Comme c'est un estimateur (*cf.* II B)) sans biais, il fournit une estimation de la moyenne théorique μ (**Exple 4 a.**)

b. La variance s_x^2 des tailles calculée sur ce même échantillon représentatif de 30 femmes françaises est une estimation ponctuelle fournie par l'estimateur «variance empirique» S_{30}^2. Comme c'est un estimateur biaisé, il ne fournit pas tout à fait une estimation de la variance théorique σ^2, mais de $\dfrac{29\sigma^2}{30}$ (**Exple 4b.** : on a sous-estimation structurelle).

A) Estimateur correct

[Estimateur correct] **Déf.5**

Un estimateur $T = T(X_1, \dots, X_n)$ est dit correct si sa variance tend vers 0 quand $n \to +\infty$.

Exple.6 D'après **Prop.1 b.**, la moyenne empirique $\overline{X}$ est un estimateur correct (de μ, *cf.* **Exple.4a**).

Rem.5 Entre deux estimateurs non biaisés d'un paramètre θ, on préfèra celui de plus faible variance, car cela siginifie que sa distribution d'échantillonnage (c-à-d. sa loi) est plus concentrée autour de son espérance, qui est justement le paramètre à estimer (en effet on parle d'estimateurs sans biais).

IV Estimation par intervalle de confiance

A) Intervalle de confiance (IC)

Les estimations ponctuelles ne suffisent pas à renseigner sur la précision de l'estimation

[Intervalle de confiance au risque α] **Déf.6**

Soit $\alpha \in]0,1[$. Si θ est le paramètre à estimer au sein de la population $\mathscr{P}$, un intervalle de confiance de θ au risque α est un intervalle $I = [A_-, A_+]$ où $A_\pm$ sont des estimateurs (donc des VAR) tels que $P(A_- \leq \theta \leq A_+) \geq 1 - \alpha$.

Rem.6 Les intervalles de confiance (c-à-dire leurs extrémités) sont construits à partir des mesures effectuées sur un échantillon. Donc, si on prélève un autre échantillon (c-à-d. si on refait une série de mesures), on change d'intervalle. Ainsi, les extrémités $A_\pm$ sont bien des VAR.

B) Obtention d'IC

[IC pour un estimateur de loi normale, et de variance connue] **Thm.4**

Si T suit une loi $\mathscr{N}(m, s^2)$, et si a est une réalisation de T, l'intervalle :

$$I = [a - u_{1-\frac{\alpha}{2}} s, a + u_{1-\frac{\alpha}{2}} s]$$

est un intervalle de confiance au seuil $1 - \alpha$ de m.

Exple.7

Cas typique : contrôle de qualité (pièces d'usine). Dans de telles situations, la loi de X est supposée normale, m est justement inconnue, mais σ^2 est une donnée (variabilité liée aux machines). La moyenne empirique $\overline{X}$ suit une loi normale par stabilité des lois normales. Sa variance est $s^2 = \sigma^2/n$. On applique le théorème à $T = \overline{X}$.

[IC de la moyenne théorique si la loi de X est approximativement normale] **Thm.5**

Si $\overline{X}$ vérifie les conditions d'application du théorème central-limite :

$$I = \left[\overline{x} - u_{1-\frac{\alpha}{2}} \frac{s_x}{\sqrt{n}}, \overline{x} + u_{1-\frac{\alpha}{2}} \frac{s_x}{\sqrt{n}}\right]$$

est un intervalle de confiance au seuil $1 - \alpha$ de μ, s_x^2 est la variance calculée sur l'échantillon considéré, et n la taille de l'échantillon.

Rem.7 Cet intervalle de confiance **reste valable même si la loi de $\overline{X}$ est normale mais de variance inconnue** pour peu que n soit assez grand :

a. En pratique $n \geq 30$ est accepté.

b. Sinon, on remplace le $u_{1-\frac{\alpha}{2}}$ de la loi normale par le $t_{1-\frac{\alpha}{2}}$ de loi loi de Student à $n-1$ degrés de liberté.

V Tests statistiques

Test de conformité à la moyenne

Soit X une grandeur attachée à une population $\mathscr{P}$, de moyenne théorique μ (c'est-à-dire $E(X) = \mu$).

T1> Test de l'hypothèse (H_0) de conformité :

$$(H_0) \qquad \mu = \mu_0$$

a. On fixe un risque α (souvent $\alpha = 5\%$).

b. Sous réserve que la loi de l'estimateur «moyenne empirique» est approximativement normale sous (H_0) :

1) On calcule les quantités empiriques $\overline{x}$ et s_x qui sont des estimations de la moyenne et de l'écart-type théoriques sur l'échantillon de taille n prélevé.

2) On calcule $u = \dfrac{\overline{x} - \mu_0}{\frac{s_x}{\sqrt{n}}}$

c. Si $u \notin \left[-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right]$: il y avait une probabilité $\leq \alpha$ (donc trop faible) que cela se produise sous (H_0) : on peut rejetter (H_0) au risque α.

d. Si $u \in \left[-u_{1-\frac{\alpha}{2}}, u_{1-\frac{\alpha}{2}}\right]$: on n'a pas d'élément permettant de douter de H_0. Faute d'indices supplémentaires, on conserve (H_0).