Semestre S6 (Statistique inférentielle)

Chapitre 1

Statistique inférentielle

Introduction

Le mot statistique désigne à la fois un ensemble de données d'observations et l'acivité qui consiste dans leur recueil, leur traitement et leur interprétation.

Les raisonements par lesquels on peut tirer, à partir des observations, des conclusions concernant les lois de probabilité régissant les phénomènes étudiés

sont appelés inférence ou induction statistique et sont codifiés par la statistique mathématique.

1 Echantillonnage

Si on veut étudier une population, il est le plus souvent non envisageable d'étudier tous les éléments de cette population (problème de temps et de coût).

Le choix d'un échantillon s'impose et c'est l'étude de celui qui va permettre de conclure pour toute la population.

Ceci est l'objet de la statistique inférentielle que nous pouvons décomposer en 3 étapes:

- ⁕ choix d'un échantillon (échantillonnage).
- ⁕ étude de cet échantillon.
- ⁕ conclure pour toute la population.

Remarques

⊛ Pour que les conclusions de la théorie de l'échantillonnage soient valides, les échantillons doivent être choisis de telle façon qu'ils soient représentatifs d'une population.

⊛ Une façon d'obtenir un échantillon représentatif d'une population est de procéder à un échantillonnage aléatoire, qui garantit que chaque élément de la population

a la même probabilité d'appartenir à l'échantillon. En d'autres termes, des tirages équiprobables et indépendants les uns aux autres, dans ces conditions deviennent

des variables aléatoires ainsi que les résumés numériques usuels (moyenne, variance,...).

⊛ L'échantillonnage peut-être:

a) exhaustif ou sans remise: prélèvement de n individus en une seule fois, ou successivement sans remise, dans ce cas, la composition de l'urne est modifiée

à chaque tirage (les tirages ne sont pas indépendants).

b) non exhaustif ou avec remise: lorsque chaque individu prélevé est remis dans la population (urne) avant le tirage de l'individu suivant (les tirages sont indépendants).

Re*marque* :

Lorsque la taille de l'échantillon est suffisamment petite par rapport à celle de la population, on peut assimiler l'échantillonnage (tirage) sans remise à l'échantillonnage

(tirage) avec remise.

Dans la statistique inférentielle, on peut distinguer 2 problèmes:

1.1 Problème d'échantillonnage:

Connaissant la valeur d'un paramètre (moyenne, écart-type, proportion,...) de la population, on cherche des renseignements sur la valeur que peut prendre

ce paramètre dans l'échantillon.

1.1.1 Problème d'estimation:

On connait la valeur d'un paramètre dans un échantillon et on cherche des renseignements dans la population.

Ainsi la théorie de l'échantillonnage vise à:

• étudier les relations qui existent entre la population et les échantillons de cette population.

• permet d'estimer des quantités inconnues: moyenne, variance, ... de la population.

• utile pour déterminer si des différences observées entre deux échantillons de la population sont dues à des variations aléatoires, ou si au contraire elles sont significatives.

1.1.2 Exemple

On présente quelques situations pouvant faire l'objet d'une étude statistique:

1- Etude épidémiologique de la population d'une grande ville.

2- Contrôle de qualité d'un lot de pièces fabriquées par la même machine.

3- Etude du débit annuel d'une rivière sur une période de 50 ans.

De l'observation de ces phénomènes on pourrait, dans le cadre d'un modèle statistique adéquat, tirer des informations pour:

1- Permettre de lancer des compagnes de vaccination pour l'exemple 3.

2- S'assurer de la fiabilité de la machine pour l'exemple 2.

3- Construire un barrage ou un pont pour l'exemple 1.

1.2 Techniques d'échantillonnage

1.2.1 ✠ Le modèle de l'urne:

On numérote chaque élément de la population. On note ce numéro sur un papier et on le met dans l'urne. On prélève, au hasard, un nombre de papiers égal

à la taille de l'échantillon désiré.

Si le tirage est fait avec remise, cette méthode devient impraticable pour une grande taille de la population.

1.2.2 ✠ Table de nombres aléatoires:

Cette table est formée de nombres entiers compris entre 0 et 9 choisis au hasard, avec remise, selon le modèle de l'urne.

→ numéroter tous les individus de la population de 1 à N.

→ choisir au hasard un nombre, dans la table, qui servira de point de départ.

→ choisir un sens de déplacement (vers la gauche, droite, haut, bas).

→ à partir du point de départ choisi, en considérant des blocs de 1,2,...chiffres, noter les nombres entre 1 et N, en éliminantceux qui se répètent si le tirage se fait sans remise.

Exemple Un club sportif compte 76 membres. on veut faire un sondage parmi les membres de ce club en choisissant un échantillon aléatoire simple sans remise de taille 15 ($n = 15$) .

Solution

1- on numérote tous les membres du club de 1 à 76.

2- on choisit dans la table un point de départ, par exemple à l'intersection de la ligne 9 ($i = 9$) et la colonne 5 ($j = 5$) .

3- on lit, en éliminant les nombresqui sont >76: $11, 12, 60, 75, 29, 09, 74, 48, 41, 17, 49, 08, 45, 56, 04.$

(si $i = 6, j = 10$, vers le bas: $57, 53, 52, 25, 38, 45, 28, 33, 51, 37, 48, 24, 09, 43, 13$).

l'échantillon choisi est formé des membres du club portant les numéros précédents.

Remarquons, qu'en population humaine, les listes électorales ou l'annuaire téléphonique peuvent être utilisés comme bases de sondages.

1.2.3 ✠ Echantillon périodique ou systématique:

On choisit un premier élément au hasard et une période fixe.

Exemple: Dans une chaîne de production de pièces on prélève une pièce puis la 20ème, 40ème.....jusqu'à obtention de la taille désirée de l'échantillon.

1.2.4 ✠ Echantillonnage par degrés ou par grappes:

C'est un tirage où l'échantillonnage n'a pas lieu directement parmi les éléments de la population mais en plusieurs temps (tirage en cascade).

Exemple: On s'intéresse à une population d'écoliers, on tire au hasard en premier lieu les villes, puis dans les villes tirées, les écoles, puis en fin des écoliers.

1.2.5 ✠ Echantillon stratifié:

On subdivise la population en sous groupes ou strates dans lesquels la variable d'intérêt ne varie pas beaucoup. On prélève de chaque strate un échantillon aléatoire

et on regroupe tous ces échantillons pour former l'échantillon désiré. Ceci permet de réduire la taille de l'échantillon mais elle suppose une connaissance préalable

de la population et de ses sous groupes.

Exemple: Pour étudier le salaire moyen d'un fonctionnaire, on peut subdiviser la population en strates formées des ouvriers, agents,cadres moyens, cadres supérieurs etc...

1.3 Echantillon aléatoire

1.3.1 Position du problème

Supposons qu'on dispose d'un lot de pièces éléctroniques pour vente. Il est naturel de supposer que dans ce lot il y a des pièces défectueuses.

On veut connaître le poucentage de ces pièces défectueuses. La seule manière est de tester toutes les pièces et de compter le nombre de pièces défectueuses.

Naturellement, cette procédure n'est pas réalisable (problème de temps et de coût).

On se propose de répondre à cette question, en se basant sur les résultats que donnera un échantillon aléatoire.

Donc, on va tester n pièces au hasard.

Soit $X_i, i = 1, 2, ..., n$ une suite de v.a de Bernoilli telle que

$$X_i = \left\{ \begin{array}{cc} 1 & \text{si la pièce est défectueuse} \\ 0 & \text{sinon} \end{array} \right\}$$

1.3.2 Définition

On appelle échantillon aléatoire ou n-échantillon toute suite de v.a $(X_1, X_2, ..., X_n)$ indépendantes et de même loi de probabilité.

1.3.3 Remarque

En statistique descriptive, un échantillon est défini comme un sous ensemble de la population.

1.3.4 Exemples

1) On lance un dé n fois. On pose X_i le chiffre obtenu au $i^{ème}$ lancé; $i = 1, 2, ..., n$

les X_i constituent une suite de v.a indépendantes et de même loi de probabilité.

$$P(X_i = k) = \frac{1}{6} \quad \forall k = 1, 2, ..., 6, \ \forall i = 1, 2, ..., n$$

donc $(X_1, X_2, ..., X_n)$ est bien un échantillon.

2) On tire n jetons avec remise d'une boite contenant 5 jetons numérotés de 1 à 5.

On note X_i la v.a représentant le numéro du $i^{ème}$ jeton tiré $(i = 1, 2, ..., n)$.

Les tirages se font avec remise, donc ils sont indépendants et

$$P(X_i = k) = \frac{1}{5} \quad \forall k = 1, 2, ..., 5, \ \forall i = 1, 2, ..., n$$

donc les X_i ont la même loi de probabilité, alors $(X_1, X_2, ..., X_n)$ est bien un échantillon.

1.3.5 Remarque

Si tous les tirages se font sans remise, les X_i ne sont pas indépendants et dans ce cas $(X_1, X_2, ..., X_n)$ n'est pas un échantillon.

1.4 Définition de la statistique

Toute fonction mesurable $T = \Phi(X_1, X_2, ..., X_n)$ de l'échantillon $(X_1, X_2, ..., X_n)$ ne dépendant d'aucun paramètre inconnu est appelée statistique.

1.4.1 Exemple

1)

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$$

$\overline{X}$ est appelé moyenne empirique de l'échantillon, S^2 est appelé variance empirique de l'échantillon.

2 Distribution d'échantillonnage

Prenons tous les échantillons de taille n d'une population. pour chaque échantillon, on peut calculer une statistique: moyenne, écart-type,...

ce qu'on appelle distribution d'échantillonnage. Si on utilise la moyenne comme statistique; on dit que c'est une distribution d'échantillonnage de la moyenne.

2.1 Distribution d'échantillonnage de la moyenne

Soit une population de taille suffisamment grande (de sorte que l'échantillonnage peut être considéré comme avec remise).

On prélève un 1^{er} échantillon de taille n : $x_1^{(1)}, x_2^{(1)}, ..., x_n^{(1)}$ sa moyenne est $\overline{x}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(1)}$

si on prélève un $2^{ème}$ échantillon de taille n : $x_1^{(2)}, x_2^{(2)}, ..., x_n^{(2)}$ sa moyenne est $\overline{x}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(2)}$

on peut répéter les prélèvements et obtenir $\overline{x}^{(1)}, \overline{x}^{(2)}, ...$ qui sont en général des valeurs différentes, dépendants de l'échantillon et peuvent être considérées comme

des réalisations de la v.a: $\overline{X} = \frac{X_1+X_2+...+X_n}{n}$

la v.a.r $\overline{X}$ possède une loi de probabilité appellée distribution d'échantillonnage de la moyenne et tel que:

$$E\left(\overline{X}\right) = m, \quad Var\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

en effet

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

$$\begin{aligned} E\left(\overline{X}\right) &= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n}\left(nm\right) = m \\ E\left(\overline{X}\right) &= m \end{aligned}$$

$$\begin{aligned} Var\left(\overline{X}\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) \qquad (X_i \text{ sont indépendants}) \\ &= \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \\ Var\left(\overline{X}\right) &= \frac{\sigma^2}{n} \end{aligned}$$

2.1.1 Remarques

⁕ Quand $n \longrightarrow +\infty$, $Var\left(\overline{X}\right) \longrightarrow 0$
donc

$$Var\left(\overline{X}\right) = E\left[\left(\overline{X} - E\left(\overline{X}\right)\right)^2\right] = E\left[\overline{(X} - m)^2\right] \longrightarrow 0, \text{ quand } n \longrightarrow +\infty$$

et on dit dans ce cas que $\overline{X}$ converge en moyenne quadratique vers m.

⁕ Si la distribution de la population est normale, alors la distribution d'échantillonnage de la moyenne est aussi normale, car $\overline{X}$ est une combinaison

linéaire de v.a indépendantes de loi $N\left(m, \sigma^2\right)$, i.e

$$X_i \sim N\left(m, \sigma^2\right) \quad \forall i = 1, ..., n \quad \Longrightarrow \overline{X} \sim N\left(m, \frac{\sigma^2}{n}\right)$$

⁕ Si la distribution de la population est quelquonque et n assez grand ($n \geq 30$), le théorème central limite (T.C.L) permet d'ffirmer que $\frac{\overline{X} - m}{\frac{\sigma}{n}} \backsim N(0, 1)$.

(T.C.L: $X_1, X_2, ..., X_n$ une suite de v.a indépendantes et de même loi (i.i.d), alors la v.a $Y_n = \sum_{i=1}^{n} X_i$ vérifie:

$$\frac{Y_n - E(Y_n)}{\sigma(Y_n)} \longrightarrow N(0, 1) \quad \text{quand } n \longrightarrow +\infty$$

2.2 Distribution d'échantillonnage de la variance

De la même manière que pour les moyennes, on considère les variances des échantillons prélevés:

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{(1)} - \overline{x}^{(1)}\right)^2, \quad S_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^{(2)} - \overline{x}^{(2)}\right)^2$$

et aisi de suite, ces valeurs peuvent être considérées comme des réalisations de la v.a.r S^2 (S^2 : variance empirique de l'échantillon)

où

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

la loi de probabilité de cette v.a est appelée distribution d'échantillonnage de la variance avec

$$E\left(S^2\right) = \frac{n-1}{n}\sigma^2$$

en effet

$$
\begin{aligned}
E\left(S^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(\left(X_i-m\right)-\left(\overline{X}-m\right)\right)^2\right) \\
&= E\left\{\frac{1}{n}\sum_{i=1}^{n}\left[\left(X_i-m\right)^2-2\left(X_i-m\right)\left(\overline{X}-m\right)+\left(\overline{X}-m\right)^2\right]\right\} \\
&= E\left\{\frac{1}{n}\left[\sum_{i=1}^{n}\left(X_i-m\right)^2-2\left(\overline{X}-m\right)\sum_{i=1}^{n}\left(X_i-m\right)+\sum_{i=1}^{n}\left(\overline{X}-m\right)^2\right]\right\} \\
&= E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-m\right)^2-\frac{2}{n}\left(\overline{X}-m\right)\sum_{i=1}^{n}\left(X_i-m\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\overline{X}-m\right)^2\right\} \\
&= E\left\{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-m\right)^2-2\left(\overline{X}-m\right)^2+\left(\overline{X}-m\right)^2\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}E\left(X_i-m\right)^2-E\left(\overline{X}-m\right)^2 = \frac{1}{n}n\sigma^2-Var\left(\overline{X}\right) = \sigma^2-\frac{\sigma^2}{n} \\
E\left(S^2\right) &= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

2.2.1 Remarque

On remarque que $E\left(S^2\right) \neq \sigma^2$, mais il suffit de modifier la variance S^2 pour avoir l'égalité d'où l'introduction de la quasi-variance de l'échantillon $\overset{\wedge}{\sigma}^2$ ou S^{*2}

$$\overset{\wedge}{\sigma}^2 = S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

on a

$$E\left(S^{*2}\right) = \sigma^2$$

en effet

$$
\begin{aligned}
E\left(S^{*2}\right) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right) \\
&= E\left\{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-m\right)^2-\frac{2}{n-1}\left(\overline{X}-m\right)\sum_{i=1}^{n}\left(X_i-m\right)+\frac{1}{n-1}\sum_{i=1}^{n}\left(\overline{X}-m\right)^2\right\} \\
&= E\left\{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-m\right)^2-\frac{2n}{n-1}\left(\overline{X}-m\right)^2+\frac{n}{n-1}\left(\overline{X}-m\right)^2\right\} \\
&= \frac{n}{n-1}\sigma^2-\frac{n}{n-1}Var\left(\overline{X}\right) = \frac{n}{n-1}\sigma^2-\frac{n}{n-1}\frac{\sigma^2}{n} \\
E\left(S^{*2}\right) &= \sigma^2
\end{aligned}
$$

donc $E\left(S^{*2}\right)=\sigma^2$ mais $E\left(\sqrt{S^{*2}}\right)\neq\sqrt{E\left(S^{*2}\right)}$ (à voir).

$\odot$ si la population est normale, alors:

$$Var\left(S^{*2}\right)=\frac{2}{n-1}\sigma^4$$

3 Rappel

3.1 la loi du Khi-deux χ^2

3.1.1 Définition

Une variable aléatoire X suit la loi du $\chi^2(n)$ (khi-deux de degrè de liberté n) $\left(X\backsim\chi^2(n)\right)$ si sa densité de probabilité est définie par:

$$f(x)=\left\{\begin{array}{ll}\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}e^{-\frac{x}{2}}x^{\frac{n}{2}-1} & \text{si } x>0\\ 0 & \text{sinon}\end{array}\right\}$$

avec

$$\Gamma(\alpha)=\int_0^{+\infty}e^{-t}\ t^{\alpha-1}dt\qquad(\alpha\geq 0)$$

$$\text{si}\quad\alpha=\frac{1}{2},\quad\Gamma\left(\frac{1}{2}\right)=\int_0^{+\infty}e^{-t}\ t^{-\frac{1}{2}}dt\ =2\int_0^{+\infty}e^{-u^2}\ du=2\sqrt{\pi}$$

en effet

$$\begin{array}{rcl}\text{on pose}\ \ t & = & u^2, dt=2udu\\ \int_0^{+\infty}e^{-t}\ t^{-\frac{1}{2}}dt & = & \int_0^{+\infty}e^{-u^2}\left(u^2\right)^{-\frac{1}{2}}2u\ du=\int_0^{+\infty}e^{-u^2}du\end{array}$$

et comme on a:

$$\int_0^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx=1\qquad(\text{ d.d.p de la loi } N(0,1)$$

alors

$$\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\left(\frac{x}{\sqrt{2}}\right)^2} dx = 1 \qquad \text{en posant } u = \frac{x}{\sqrt{2}},\ dx = \sqrt{2}du$$

$$= \int_0^{+\infty} e^{-\left(\frac{x}{\sqrt{2}}\right)^2} dx = \sqrt{2\pi} \Longrightarrow \int_0^{+\infty} e^{-u^2}\sqrt{2}du = \sqrt{2\pi}$$

$$\Longrightarrow \int_0^{+\infty} e^{-u^2} du = \sqrt{\pi}$$

autres propriétés de la fonction Γ

- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha) \qquad (\alpha \geq 0)$

$$\text{en effet} \quad : \quad \Gamma(\alpha+1) = \int_0^{+\infty} e^{-t}t^{\alpha}dt,$$

$$\text{en intégrant par parties} \quad : \quad U = t^{\alpha}; dU = \alpha t^{\alpha-1},\ \ dV = e^{-t}; V = -e^{-t}$$

$$\Gamma(\alpha+1) = \left[-t^{\alpha}e^{-t}\right]_0^{+\infty} + \alpha \int_0^{+\infty} e^{-t}t^{\alpha-1}dt = \alpha\Gamma(\alpha)$$

- $\forall n \in \mathbb{N} : \Gamma(n+1) = n!, \Gamma(1) = 1, \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$

$$\Gamma(n+1) = n\int_0^{+\infty} e^{-t}t^{n-1}dt = n\left\{\left[\left[-t^{n-1}e^{-t}\right]_0^{+\infty} + (n-1)\int_0^{+\infty} e^{-t}t^{n-2}dt\right]\right\}$$

en faisant des intégrations par parties succéssives, on obtient

$$\Gamma(n+1) = n(n-1)(n-2)\cdots 1\int_0^{+\infty} e^{-t}\ tdt = n(n-1)(n-2)\cdots 1\Gamma(1)$$

$$\text{or } \Gamma(1) = 1, \text{ donc} \quad \Gamma(n+1) = n!$$

3.1.2 Définition

Si $X_1, X_2, ..., X_n$ des v.a.r indépendantes suivant toutes la loi $N(0,1)$, alors la somme des carrés de ces v.a.r

$$X_1^2 + X_2^2 + ... + X_n^2 = \sum_{i=1}^{n} X_i^2 \backsim \chi_n^2$$

On en déduit immédiatement que si X et Y 2 v.a indépendantes telles que

$$\left\{ \begin{array}{c} X \backsim \chi_p^2 \\ Y \backsim \chi_q^2 \end{array} \right\} \Longrightarrow X+Y \backsim \chi_{p+q}^2$$

3.1.3 Définition

On dit qu'une v.a X suit la loi Gamma de paramètre r ($X \backsim \gamma(r)$) si sa densité de probabilité est donnée par

$$f(x) = \frac{1}{\Gamma(r)} e^{-x} x^{r-1} 1_{]0,+\infty[}$$

avec

$$E(X) = \frac{1}{\Gamma(r)} \int_0^{+\infty} x^r e^{-x} dx = \frac{\Gamma(r+1)}{\Gamma(r)} = r$$

et

$$\begin{aligned} Var(X) &= E\left(X^2\right) - E^2(X) = \frac{1}{\Gamma(r)} \int_0^{+\infty} x^{r+1} e^{-x} dx - r^2 \\ &= \frac{\Gamma(r+2)}{\Gamma(r)} - r^2 = \frac{(r+1)\Gamma(r+1)}{\Gamma(r)} - r^2 = (r+1)r - r^2 = r \end{aligned}$$

3.1.4 Propriété

Si $X \backsim \gamma(r) \Longrightarrow 2X \backsim \chi_{2r}^2$

On en déduit donc par transformation les propriétés de la loi du χ^2

$$E\left(\chi_n^2\right) = n, \qquad Var\left(\chi_n^2\right) = 2n$$

3.1.5 Théorème

Si la distribution de la population est normale, alors la variable aléatoire

$$\frac{nS^2}{\sigma^2} \backsim \chi_{n-1}^2$$

en effet

$$\begin{aligned}
\frac{nS^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(X_i-m\right)-\left(\overline{X}-m\right)^2\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left\{\left(X_i-m\right)^2+\left(\overline{X}-m\right)^2-2\left(\overline{X}-m\right)\left(X_i-m\right)\right\}\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(X_i-m\right)^2+n\left(\overline{X}-m\right)^2-2\left(\overline{X}-m\right)n\left(X_i-m\right)\right] \\
&= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(X_i-m\right)^2-n\left(\overline{X}-m\right)^2\right] = \sum_{i=1}^{n}\left(\frac{X_i-m}{\sigma^2}\right)^2-\left(\frac{\overline{X}-m}{\frac{\sigma}{n}}\right)^2
\end{aligned}$$

donc, on a:

$$\left\{\begin{array}{c}\sum_{i=1}^{n}\left(\frac{X_i-m}{\sigma^2}\right)^2 \backsim \chi_n^2 \\ \left(\frac{\overline{X}-m}{\frac{\sigma}{n}}\right)^2 \backsim \chi_1^2\end{array}\right\} \Longrightarrow \frac{nS^2}{\sigma^2} \backsim \chi_{n-1}^2$$

car

$$\begin{aligned}
X_i &\backsim N(0,1) \Longrightarrow X_i^2 \backsim \chi_1^2 \\
nX_i &\backsim N(0,1) \Longrightarrow nX_i^2 \backsim \chi_n^2 \qquad i=1,...n
\end{aligned}$$

3.2 Distribution des fréquences

Considérons une population de taille N dont les éléments possèdent un certain caractère avec une fréquence p. On prélève avec remise dans cette population des échantillons

de taille n et on mesure pour chacun d'eux la fréquence F avec laquelle les éléments possèdent cette même propriété.

soit S_n: v.a représentant le nombre d'éléments de l'échantillon possédant la propriété considérée.

$$S_n \backsim \beta(n,p)$$

$F=\frac{S_n}{n}$ est la proportion d'individus ayant ce caractère dans l'échantillon, alors

$$E(F)=p, \qquad Var(F)=\frac{p(1-p)}{n}$$

en effet

$$E(F)=E\left(\frac{S_n}{n}\right)=\frac{1}{n}E(S_n)=\frac{1}{n}np=p$$

$$\begin{aligned}
Var(F) &= Var\left(\frac{S_n}{n}\right)=\frac{1}{n^2}np(1-p) \\
&= \frac{p(1-p)}{n}=\frac{pq}{n}, \qquad q=1-p
\end{aligned}$$

3.2.1 Remarques

1) Si le tirage se fait sans remise, alors:

$$E(F) = p, \qquad Var(F) = \frac{p(1-p)}{n}\frac{N-n}{N-1}$$

en effet
dans ce cas:

$$\begin{aligned} S_n &\backsim H(N,n,p) \\ E(S_n) &= np, \quad Var(S_n) = npq\frac{N-n}{N-1} \end{aligned}$$

et donc

$$E(F) = p$$

$$\begin{aligned} Var(F) &= Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2}Var(S_n) \\ &= \frac{p(1-p)}{n}\frac{N-n}{N-1} \end{aligned}$$

$\frac{N-n}{N-1}$ est appelé coefficient d'exhaustivité

2) Pour une taille n de l'échantillon assez grande (en pratique pour $n \geq 30$), on a

$$\frac{F - E(F)}{\sigma(F)} = \frac{F-p}{\sqrt{\frac{pq}{n}}} \backsim N(0,1)$$

3.3 Distribution des différences de moyennes

Soient deux populations P_1 et P_2 de moyennes m_1 et m_2 et de variances σ_1^2, σ_2^2 (respectivement). On s'intérresse dans de nombreux problèmes à la différence $m_1 - m_2$.

On extrait de P_1 un échantillon de taille n_1 $(x_{11}, x_{12}, ..., x_{1n_1})$ et de P_2 un échantillon de taille n_2 $(x_{21}, x_{22}, ..., x_{2n_1})$

en notant par

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i}, \qquad \overline{X}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}$$

on a:

$$1) \qquad E\left(\overline{X}_1 - \overline{X}_2\right) = m_1 - m_2$$

en effet

$$E\left(\overline{X}_1 - \overline{X}_2\right) = E\left(\overline{X}_1\right) - E\left(\overline{X}_2\right) = m_1 - m_2$$

$$2) \qquad Var\left(\overline{X}_1 - \overline{X}_2\right) = \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2$$

en effet

$$\begin{aligned} Var\left(\overline{X}_1 - \overline{X}_2\right) &= Var\left(\overline{X}_1\right) + Var\left(\overline{X}_2\right) \qquad \text{car } \overline{X}_1, \overline{X}_2 \text{ son indépendantes} \\ &= \frac{1}{n_1}\sigma_1^2 + \frac{1}{n_2}\sigma_2^2 \end{aligned}$$

ceci pour un tirage avec remise (non exhaustif), dans le cas d'un tirage sans remise (exhaustif), on aura:

$$Var\left(\overline{X}_1 - \overline{X}_2\right) = \frac{\sigma_1^2}{n_1}\frac{N_1 - n_1}{N_1 - 1} + \frac{\sigma_2^2}{n_2}\frac{N_2 - n_2}{N_2 - 1}$$

où N_1 est la taille de P_1 et N_2 est la taille de P_2.

3) Si n_1, n_2 sont assez grands, on peut dire que

$$\left\{ \begin{array}{l} \overline{X}_1 \backsim N\left(m_1, \frac{\sigma_1^2}{n_1}\right) \\ \overline{X}_2 \backsim N\left(m_2, \frac{\sigma_2^2}{n_2}\right) \end{array} \right\} \Longrightarrow \overline{X}_1 - \overline{X}_2 \backsim N\left(m_1 - m_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

et on conclut donc que

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \backsim N(0,1) \qquad (\text{ T.C.L })$$

3.3.1 Exemple

On choisit au hasard 6 nombres parmi les nombres entiers de 1 à 9, chacun de ces nombres a la même probabilité d'être choisi.

Calculer la moyenne et l'écart-type de la distribution d'échantillonnage des moyennes dans les 2 cas:

a) tirage sans remise.

b) tirage avec remise.

Solution

a) la moyenne de la population est

$$m = \frac{1 + 2 + ... + 9}{9} = 5$$

la variance est

$$\sigma^2 = \frac{1}{9}\left[(1-5)^2 + (2-5)^2 + ... + (9-5)^2\right] = 6.67 \Longrightarrow \sigma = 2.58$$

il y a $C_9^6 = 84$ façons de choisir 6 nombres parmi 9 nombres. Chacun de ces 84 échantillons possibles a une moyenne $\overline{x} = \frac{1}{6}\sum_{i=1}^{6} x_i$ où $i = 1, 2, ..., 6$ représente un des 9 nombres.

par exemple l'échantillon $(3, 8, 7, 2, 5, 1)$ a pour moyenne $\overline{x} = 4.33$. On obtient ainsi 84 moyennes et la moyenne de la distribution d'échantillonnage des moyennes est

$$E\left(\overline{X}\right) = m = 5$$

la variance de la distribution d'échantillonnage des moyennes est

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{6.67}{6}\left(\frac{9-6}{9-1}\right) = 0.417 \Longrightarrow \sigma\left(\overline{X}\right) = 0.645$$

b) Il y a $9^6 = 531441$ façons de choisir 6 nombres parmi les 9 nombres. chacun de ces échantillons a une moyenne $\overline{x} = \frac{1}{6}\sum_{i=1}^{6} x_i$ où $i = 1, 2, ..., 6$ représente un des 9 nombres.

par exemple l'échantillon $(4, 3, 4, 5, 7, 8)$ a pour moyenne $\overline{x} = 5.17$. On obtient de cette manière 531441 moyennes et la moyenne de la distribution d'échantillonnage des moyennes est

$$E\left(\overline{X}\right) = m = 5$$

la variance de la distribution d'échantillonnage des moyennes est

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n} = \frac{6.67}{6} = 1.11 \Longrightarrow \sigma\left(\overline{X}\right) = 1.05$$

Chapitre 2

ESTIMATION

Introduction

En général, nous cherchons des informations sur une population d'effectif relativement important (N) à partir de l'étude d'un échantillon (application: domaine économique, social, industriel,...)

car le plus souvent, il n'est pas possible d'étudier toute la population; cela prendrait trop de temps , reviendrait trop cher ou serait aberrant comme (par exemple: dans le cas d'un contôle de qualité entrainant

la destruction des pièces. Donc, on cherche à avoir une idée sur (estimer) la valeur d'un paramètre θ inconnu d'un caractère défini sur cette population (moyenne, variance, proportion).

En d'autres termes, l'estimation consiste à donner des valeurs approchées aux paramètres d'une population $\left(m, \sigma^2, ...\right)$ à l'aide d'un échantillon de cette population.

Pour résoudre ce problème, il existe 2 types de réponses:

⁕ L'estimation ponctuelle: à partir de l'information fournie par l'échantillon, donne une valeur unique au paramètre inconnu θ.

⁕ L'estimation par intervalle de confiance: consiste à construire un intervalle à l'intérieur duquel, le paramètre se trouve avec une probabilité donnée.

4 Principes généraux

Soit une population dont la distribution dépend d'un paramètre inconnu θ et soit un échantillon $(X_1, X_2, ..., X_n)$ extrait de cette population.

Une statistique $T = \Phi\left(X_1, X_2, ..., X_n\right)$, utilisée pour estimer θ, est dite estimateur qu'on note $\overset{\wedge}{\theta}\left(\overset{\smile}{\theta}\right)$. la valeur prise par T est dite estimation de θ.

Evidemment, l'estimateur doit avoir certaines qualités pour être accéptable.

4.0.2 Exemple

Les variables aléatoires $\overline{X}, S^2, F$ sont appelées estimateurs de m, σ^2, p (resp).

4.0.3 Remarque

Il est possible que le même paramètre soit estimé par des estimateurs différents.

4.0.4 Exemple

Pour une distribution symétrique, la médiane de l'échantillon est égalemnt une estimation de m.

Afin de choisir entre plusieurs estimateurs possibles d'un même paramètre il faut définir les qualités exigées d'un estimateur.

4.1 Qualités d'un estimateur

Soit $T = \Phi(X_1, X_2, ..., X_n)$ un estimateur du paramètre inconnu θ.

On définit l'erreur d'estimation par: $T - \theta$ qui est une v.a

$$T - \theta = (T - E(T)) + (E(T) - \theta)$$

$T - E(T)$: représente les fluctuations alétoires de T autour de sa valeur moyenne.

$E(T) - \theta$: erreur systématique due au fait que T varie autour de sa valeur centrale $E(T)$ et non autour de θ.

la quantité $E(T) - \theta$ s'appelle le **biais**.

Il est donc souhaitable d'utiliser des estimateurs sans biais.

GRAPHE

4.1.1 Estimateur sans biais

Définition1:

On dit que T est un estimateur sans biais de θ si

$$E(T) = \theta$$

Exemple

$\divideontimes$ $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $E(\overline{X}) = m \Longrightarrow \overline{X}$: **estimateur sans biais de** m

$\divideontimes$ $S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$, $E(S^2) = \frac{n-1}{n}\sigma^2 \Longrightarrow S^2$: estimateur biaisé de σ^2

$\divideontimes$ $S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n}{n-1}S^2$, $E(S^{*2}) = \frac{n}{n-1}E(S^2) = \sigma^2 \Longrightarrow S^{*2}$: estimateur sans biais de

$\divideontimes$ $F = \frac{S_n}{n}$, $E(F) = p \Longrightarrow F$ estimateur sans biais de p

Remarque Si

$E(T) \longrightarrow \theta, n \longrightarrow +\infty$; T est un estimateur asymptotiquement sans biais de θ

On mesure généralement la précision d'un estimateur T par:

4.1.2 L'erreur quadratique moyenne (M.S.E) $E\left[(T-\theta)^2\right]$

On peut écrire:

$$\begin{aligned} E\left[(T-\theta)^2\right] &= E\left[(T-E(T))+(E(T)-\theta))^2\right] \\ &= E\left[\Big(T-E(T))^2+(E(T)-\theta)^2+2(T-E(T))(E(T)-\theta)\Big)\right] \\ &= E\left[(T-E(T))^2\right]_{=V(T)}+E\left[(E(T)-\theta)^2\right]_{=cste}+2E(T-E(T))(E(T)-\theta)_{=0} \end{aligned}$$

donc

$$E\left[(T-\theta)^2\right]=Var(T)+(E(T)-\theta)^2$$

On conclut que de 2 estimateurs sans biais, le plus précis est celui de variance minimale, i.e

si T_1, T_2 2 estimateurs de θ tel que $E(T_1)=\theta,\ E(T_2)=\theta P$

$$\Longrightarrow \left\{ \begin{array}{l} E\left[(T_1-\theta)^2\right]=Var(T_1) \\ E\left[(T_2-\theta)^2\right]=Var(T_2) \end{array} \right\}$$

si

$$Var(T_1)<Var(T_2)\Longrightarrow E\left[(T_1-\theta)^2\right]<E\left[(T_2-\theta)^2\right]$$

donc T_1 est plus précis que T_2.

4.1.3 Estimateur efficace

Définition On dit que T est l'estimateur le plus efficace de θ si:

- $\rightarrowtail\ E(T)=\theta$
- $\rightarrowtail\ Var(T)\leq Var(T')$ où T' est un autre estimateur sans biais de θ

4.1.4 Estimateur convergent

Définition Un estimateur T est dit convergent si sa variance tend vers 0 quand la taille de l'échantillon augmente.

Exemple $\overline{X}$ est un estimateur convergent car

$$Var\left(\overline{X}\right)=\frac{\sigma^2}{n}\longrightarrow 0,\ \text{quand } n\longrightarrow+\infty$$

4.1.5 Théorème de convergence en probabilité d'un estimateur

T estimateur de θ, on dit qu'il converge en probabilité vers le paramètre inconnu θ.

$$\begin{aligned} T \xrightarrow{P} \theta, n \quad &\longrightarrow \quad +\infty \iff P\left[|T-\theta| \leq \varepsilon\right] \longrightarrow 1, \text{ quand } n \longrightarrow +\infty \\ &\iff \quad P\left[\theta-\varepsilon \leq T \leq \theta+\varepsilon\right] \longrightarrow 1, \text{ quand } n \longrightarrow +\infty \end{aligned}$$

si

$$\begin{aligned} 1) \;\; E(T) \quad &= \quad \theta \;\text{ ou }\; (E(T)) \longrightarrow \theta, \text{ quand } n \longrightarrow +\infty \\ 2) \;\; Var(T) \quad &\longrightarrow \quad 0, \text{ quand } n \longrightarrow +\infty \end{aligned}$$

A présent, on va essayer de déterminer une borne inférieure pour les variances des estimateurs sans biais pour θ.

4.2 Quantité d'information de Fisher $I_n(\theta)$

On appelle quantité d'information de Fisher $I_n(\theta)$ apportée par un échantillon sur le paramètre θ la quantité suivante positive ou nulle (si elle existe).

$$\begin{aligned} I_n(\theta) \quad &= \quad nE\left[(\ln f(X,\theta))'_\theta\right]^2 = nE\left[\frac{\partial \ln f(X,\theta)}{\partial \theta}\right]^2 \\ &= \quad -nE\left[\frac{\partial^2 \ln f(X,\theta)}{\partial \theta^2}\right] \qquad \text{où } f(X,\theta): \text{ densité de probabilité de } X \end{aligned}$$

4.2.1 Remarque

Si la densité f n'est pas une fonction de θ, alors $I_n(\theta) = 0$.

4.2.2 Inégalité de Cramer-Rao et efficacité

sous des conditions très générales, la variance d'un estimateur est toujours plus grande que $\frac{1}{I_n(\theta)}$.

$$Var(T) \geq \frac{1}{I_n(\theta)} \quad \text{où } T \text{ estimateur de } \theta$$

4.2.3 Remarque

Si

$Var(T) = \dfrac{1}{I_n(\theta)}$; T est un estimateur efficace car il a la plus petite variance.

Exemple on admet que le montant X des dépôts de chaque client sur un compte d'épargne, on obeit à une loi normale $N\left(m,\sigma^2\right);m,\sigma^2$ inconnus et l'on propose de les

estimer en tirant au hasard n clients parmi les titulaires de compte d'épargne.

Soit X_i v.a: montant des dépôts du client $i, i=1,2,...,n$.

1) Donner les estimateurs naturels de m et σ^2 et montrer que l'estimateur de m est sans biais et convergent en probabilité.

2) On a tiré au hasard 10 clients et obtenu les montants suivants: $4700, 6900, 13500, 22000, 10000, 12000, 14000, 18000, 8600, 9000$.

Donner les estimations de m et σ associés à ces expériences.

3) Montrer que l'estimateur de m est efficace i.e $I_n(m)=\frac{1}{Var\left(\overline{X}\right)}=\frac{n}{\sigma^2}$

Solution

1) estimateur naturel de m :

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n}X_i, \quad E\left(\overline{X}\right)=m, \quad Var\left(\overline{X}\right)=\frac{\sigma^2}{n}\longrightarrow 0, \quad \text{quand } n\longrightarrow+\infty$$

donc $\overline{X}$ convergent

estimateur de σ^2 :

$$S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2, \qquad S^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

2) Les estimateurs correspondants aux réalisations x_i des v.a X_i sont:

$$\overline{x} = \frac{x_1+x_2+...+x_{10}}{10}=\frac{4700+...+9000}{10}=11870$$

$$s^* = \sqrt{\frac{1}{9}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}=5227.5$$

3) on a:

$$I_n(m)=nE\left[\left(\ln f(X,m)\right)_m^{'}\right]^2=nE\left[\frac{\partial \ln f(X,m)}{\partial m}\right]^2$$

$$\ln f(X,m)=\ln\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{X-m}{\sigma}\right)^2}\right]=\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\frac{1}{2}\left(\frac{X-m}{\sigma}\right)^2$$

$$\frac{\partial \ln f(X,m)}{\partial m}=\frac{\partial}{\partial m}\left[\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\frac{1}{2}\left(\frac{X-m}{\sigma}\right)^2\right]=\frac{X-m}{\sigma^2}$$

$$\begin{aligned} I_n(m) &= nE\left[\frac{X-m}{\sigma^2}\right]^2 = \frac{n}{\sigma^4}E(X-m)^2 = \frac{n}{\sigma^4}Var(X) \\ &= \frac{n}{\sigma^4}\sigma^2 = \frac{n}{\sigma^2} = \frac{1}{V\left(\overline{X}\right)} \Longrightarrow \overline{X} \text{ efficace} \end{aligned}$$

5 A) Estimation ponctuelle:

5.1 Introduction:

soit $(X_1, ..., X_n)$ un échantillon prélevé d'une population de loi de probabilité P_θ (ou F_θ : fonction de répartition) dépendant d'un ou plusieurs paramètres noté θ, il s'agit d'évaluer (estimer)

la valeur de θ sur la base de l'observation de l'échantillon.

5.1.1 Exemple:

Si la loi de probabilité de la population est $N\left(m, \sigma^2\right)$ et si m, σ^2 sont inconnues alors $\theta = \left(m, \sigma^2\right) \in \Theta = \mathbb{R} \times \mathbb{R}^+$ (Θ espace des paramètres).

5.2 Définition:

Un estimateur d'un paramètre θ est une statistique dont la valeur est une estimation du paramètre θ.

l'estimation d'un paramètre par une valeur unique est unique est appelée estimation ponctuelle.

5.2.1 Exemple:

Soit $(X_1, ..., X_n)$ un échantillon tel que: $E(X_i) = m, Var(X_i) = \sigma^2$

On a: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ est une estimation ponctuelle de la moyenne m.

on a aussi $\frac{x_1+x_2}{2}$ est aussi une estimation ponctuelle de la moyenne m.

5.3 Estimation ponctuelle de la moyenne:

Soit une population de distribution d'esperance m, de variance σ^2.

$\overline{X}$: la moyenne empirique de l'échantillon, on sait que $E\left(\overline{X}\right) = m, V\left(\overline{X}\right) = \frac{\sigma^2}{n} \longrightarrow_{n\to+\infty} 0$.

ce qui montre que pour n assez grand, les valeurs de $\overline{X}$ se rapprochent de m , ce qui permet d'estimer m par $\overline{X}$.

5.4 Estimation ponctuelle de la variance:

A première vue, on pourrait estimer la variance inconnue σ^2 par S^2, mais on a vu que $E\left(S^2\right) \neq \sigma^2$, donc plutôt estimer σ^2 par S^{*2} car $E\left(S^{*2}\right) = \sigma^2$

$\left(V\left(S^{*2}\right) \longrightarrow_{n\to+\infty} 0\text{, pour un caractère normal}\right)$.

5.5 Estimation ponctuelle de la proportion:

On sait que $E(F) = p, \quad Var(F) = \frac{pq}{n} \longrightarrow_{n\to+\infty} 0$

où F: proportion d'individus ayant un caractère dans l'échantillon $F = \frac{S_n}{n}$

donc on estime p par F.

5.5.1 Remarques:

1) En pratique, on tire des échantillons de grande taille pour que la variance de l'estimateur $\longrightarrow 0$ quand $n \longrightarrow +\infty$, dans le but d'avoir une bonne précision de l'estimateur.*

2) Jusqu'à présent, on a présenté les estimateurs par la méthode des moments (moyenne, variance,...) mais cette méthode a ses limites. On va montrer à travers un exemple que cette méthode n'est pas toujours la plus pertinente.

5.5.2 Exemple

La population suit une loi uniforme sur $[0,\theta]$; $X_i \longrightarrow U_{[0,\theta]}$ de densité de probabilité

$$f(x) = \left\{ \begin{array}{cc} \frac{1}{\theta}; & 0 \leq x \leq \theta \\ 0 & \text{sinon} \end{array} \right\}$$

on sait que

$$E(X_i) = \frac{\theta}{2}, V(X_i) = \frac{\theta^2}{12}$$

donc

$$E\left(\overline{X}\right) = E(X_i) = \frac{\theta}{2}$$

Si θ est inconnu, son estimateur sans biais, déterminé par la méthode des moments à partir d'un échantillon $(X_1, X_2, ..., X_n)$ est $T(\theta) = 2\overline{X}$ puisque

$$E[T(\theta)] = \theta \Longrightarrow E[T(\theta)] = E\left(2\overline{X}\right) = \theta$$

donc $2\overline{X}$ est un estimateur sans biais de θ et de même $2\overline{X}$ est un estimateur convergent car

$$Var\left(2\overline{X}\right) = 4Var\left(\overline{X}\right) = \frac{4}{n}\frac{\theta^2}{12} \longrightarrow_{n\to+\infty} 0$$

Après expérience, on considère la réalisation suivante d'un échantillon de taille 10 de la v.a X : $1.04 - 0.95 - 0.29 - 0.12 - 1.13 - 0.51 - 0.89 - 1.96 - 1.36$

ces données numériques conduisent à l'estimation de θ suivante:

$$\widehat{\theta}_1 = 2\overline{x} = 2\frac{\sum_{i=1}^{10} x_i}{10} = \frac{1.04 + ... + 1.36}{5} = 1.89$$

donc θ est estimée par $\widehat{\theta}_1 = 1.89$

Toute fois, puisque les réalisations de X sont comprises entre 0 et θ, il est évident que

$$x_i < \theta \quad \text{donc} \quad \theta \geq \text{ tous les } x_i \Longrightarrow \theta \geq \sup_{1 \leq i \leq 10} x_i$$

donc $\widehat{\theta} \geq 1.96$, ainsi la valeur $\widehat{\theta}_2 = 1.96$ est " plus vraisemblable que $\widehat{\theta}_1 = 1.89$. $\left(\widehat{\theta} = \max_{1 \leq i \leq n} x_i\right)$.

D'où la recherche d'autres méthodes pour estimer le paramètre:(méthode du maximum de vraisemblance, méthode de Bayes, méthode des moindres-carrés).

6 Méthode du Maximum de Vraisemblance (MV)

6.1 Objectifs:

Les estimateurs proposés reposent sur l'idée intuitive qu'un moment inconnu d'une v.a ou d'une population peut-être estimé ou approché par le moment équivalent calculé sur un échantillon. Ainsi, le moment d'ordre 1 (espérance mathématique) ou la moyenne dans une population est habituellemnt estimé par la moyenne des observations issues d'un échantillon. Cette méthode des moments ne convient pas dans toutes les applications pratiques et dans certains cas, elle ne fournit pas les meilleures estimations. C'est pour cette raison que le staticien **Ronald Fisher** proposa, dans les années 1920, la méthode du maximum de vraisemblance. Cette méthode consiste à rechercher l'estimation du paramètre inconnu qui rend le plus probable ou le plus vraisemblable l'échantillon observé. puisque il s'agit de trouver un maximum, cette méthode fait appel à la notion de dérivée en mathématique, tout au moins pour le cas où la loi de probabilité est une fonction dérivable.

Ces estimateurs ont de bonnes propriétés statistiques.

6.2 Fonction de vraisemblance:

On appelle fonction de vraisemblance, notée $L(x_1, ..., x_n)$ de l'échantillon $(X_1, ..., X_n)$ la loi (distribution) de probabilité du vecteur aléatoire $(X_1, ..., X_n)$

(vecteur aléatoire: c'est un vecteur dont toutes ses composantes sont des variables aléatoires).

6.2.1 Cas discret:

Si les X_i sont des variables aléatoires discrètes $(i = 1, ..., n)$, alors:

$$\begin{aligned} L(x_1, ..., x_n, \theta) &= P[(X_1 = x_1) \cap (X_2 = x_2) \cap ... \cap (X_n = x_n)] \\ &= P(X_1 = x_1) P(X_2 = x_2) ... P(X_n = x_n) \quad \text{car les } X_i \text{ sont indépendantes} \\ &= \prod_{i=1}^{n} P_\theta(X_i = x_i) \end{aligned}$$

Exemple Soit un échantillon $(X_1, ..., X_n)$ de loi de probabilité de Poisson: $L(X) = P(\lambda)$

$$P(X = x_i) = e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

$$\begin{aligned} L(x_1, ..., x_n, \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!} \\ &= e^{-n\lambda} \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} (x_i!)} \end{aligned}$$

6.2.2 Cas continu:

Si les X_i sont des variables aléatoires continues $(i = 1, ..., n)$, alors:

$$\begin{aligned} L(x_1, ..., x_n, \theta) &= f_{X_1, ..., X_n}(x_1, ..., x_n, \theta) \\ &= f_{X_1}(x_1) ... f_{X_n}(x_n) \end{aligned}$$

ou bien

$$L(x_1, ..., x_n, \theta) = f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

Exemple Soit un échantillon $(X_1, ..., X_n)$ de loi de probabilité normale; $X \rightarrow N\left(m, \sigma^2\right)$

$$\begin{aligned} L\left(x_1, ..., x_n, \sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)^2} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}\ \sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2} \end{aligned}$$

6.3 Estimateur du MV

L'estimation du MV est la valeur $\widehat{\theta}$ qui, sustituée à θ dans la fonction $L(x_1, ..., x_n, \theta)$, rend celle-ci maximale.

6.3.1 Définition:

L'estimateur noté $\widehat{\theta} = T(X_1, ..., X_n)$ de θ tel que $L_{\widehat{\theta}}(X_1, ..., X_n)$ soit maximum est appelé estimateur du MV du paramètre θ.

$$L_{\widehat{\theta}}(X_1, ..., X_n) = \max_{\theta \in \Theta}\ L_\theta(X_1, ..., X_n)$$

Il s'agit de chercher le maximum sur θ de $L_\theta(X_1, ..., X_n)$ c'est à dire résoudre le système suivant:

$$\left\{ \begin{array}{c} \frac{dL(x_1,...,x_n,\theta)}{d\theta} = 0 \\ \frac{d^2L(x_1,...,x_n,\theta)}{d\theta^2} < 0 \end{array} \right\}$$

ce qui revient à résoudre:

$$\left\{ \begin{array}{c} \frac{d \ln\ L(x_1,...,x_n,\theta)}{d\theta} = 0 \\ \frac{d^2 \ln\ L(x_1,...,x_n,\theta)}{d\theta^2} < 0 \end{array} \right\}$$

Remarque: • L'estimation du MV est $\widehat{\theta} = T(x_1, ..., x_n)$.

- L'estimateur du MV est $\widehat{\theta} = T(X_1, ..., X_n)$.
- $\widehat{\theta}$ solution de $L'_\theta = 0$ est aussi solution de $(\ln\ L)'_\theta = 0$.
- Une condition nécessaire pour qu'une fonction à plusieurs variables admette un extremum (max ou min) est que les dérivées partielles d'ordre 1 par rapport aux variables soient nulles.

6.3.2 Exemple:

Soit $X \longrightarrow P(\lambda)$, Donner une estiamtion du MV de λ

$$P(X = x_i) = e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}$$

on a:

$$L(x_1,...,x_n;\lambda) = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n}(x_i!)}$$

$$\begin{aligned}\ln L(x_1,...,x_n;\lambda) &= -n\lambda + \ln\left(\lambda^{\sum_{i=1}^{n} x_i}\right) - \ln\left[\prod_{i=1}^{n}(x_i!)\right] \\ &= -n\lambda + \sum_{i=1}^{n} x_i \ln\lambda - \ln\left[\prod_{i=1}^{n}(x_i!)\right]\end{aligned}$$

$$\begin{aligned}\frac{d\ln\ L(x_1,...,x_n,\lambda)}{d\lambda} &= -n + \frac{\sum_{i=1}^{n} x_i}{\lambda} = 0 \\ &\Longrightarrow \quad \widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i\end{aligned}$$

on vérifie que c'est bien un max; f" $\left(\widehat{\lambda}\right) < 0$?

$$\frac{d^2\ln\ L(x_1,...,x_n,\lambda)}{d\lambda^2} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}$$

au point $\lambda = \widehat{\lambda}$:

$$= -\frac{\sum_{i=1}^{n} x_i}{\frac{1}{n^2}\left(\sum_{i=1}^{n} x_i\right)^2} = -\frac{n^2}{\sum_{i=1}^{n} x_i} < 0$$

donc $\widehat{\lambda}$ est bien le max de $L(x_1,...,x_n,\lambda)$.
Si $n = 6$; $(0,2,2,3,1,2)$

$$\widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{6} x_i = \frac{1}{6}(10) = 1.67$$

6.3.3 Propriétés des estimateurs du MV:

$$1) \quad E(T) = \theta \quad \text{ou } E(T) \longrightarrow_{n\to+\infty} \theta$$

$$2) \quad T \longrightarrow^{P}_{n\to+\infty} \theta \qquad ((1) + Var(T) \longrightarrow_{n\to+\infty} 0)$$

$$\text{la loi de } \; T \longrightarrow_{n\to+\infty} N\left(\theta, \frac{1}{I_n(\theta)}\right); \qquad V(T) = \frac{1}{I_n(\theta)}$$

$V(T) \longrightarrow_{n\to+\infty} \dfrac{1}{I_n(\theta)}$ ceci exprime que l'estimateur est asymptotiquement efficace.

6.3.4 Remarque:

Certains estimateurs obtenus par la méthode des momemts sont les mêmes que ceux obtenus par la méthode du MV.B) Estimation par intervalle:

6.4 Introduction:

Soit θ un paramètre inconnu que l'on veut estimer. On connaît une estimation ponctuelle $\widehat{\theta}$ de θ qui est en général légèrement différente de θ, puisque $\widehat{\theta}$ dépend de l'échantillon tiré. Or, aucun échantillon, aussi "représentatif" soit-il, ne pourraonner une estimation égale à la vraie valeur de θ. Il paraît donc raisonnable de completer l'estimation ponctuelle par une fourchette, c-à-d la donnée d'un intervalle réel $[a, b]$ qui a une forte probabilité de contenir θ, appelé par Neyman intervalle de confiance de θ (IC). Plus l'intervalle est large, plus la probabilité qu'il contienne θ est grande mais moins bonne est la précision de l'estimation. On doit donc trouver un compromis entre la précision et la fiabilité de l'estimation. pour trouver cet intervalle, on se fixe à l'avance un coefficient (proche de 1) appelé niveau de confiance, noté $(1-\alpha)$, la valeur α mesure la probabilité que $[a, b]$ ne contienne pas la vraie valeur de θ, (en général, α vaut 5%, ou 10%, ou 0.1%). On cherche ensuite les bornes de l'intervalle appelées limites de confiance de telle façon que $P(a \leq \theta \leq b) = 1-\alpha$, où $a = a(X_1, ..., X_n)$, $b = b(X_1, ..., X_n)$, $(X_1, ..., X_n)$: échantillon.

6.5 Définition:

La valeur α est appelée seuil de signification et $(1-\alpha)$ est appelée niveau de confiance.

6.6 Intervalle aléatoire:

On appelle intervalle aléatoire tout intervalle dont une extrémité au moins est une variable aléatoire.

6.6.1 Exemple:

1) Soit la v.a $X \longrightarrow N(0,1)$, l'intervalle $]-\infty, X]$ est un intervalle aléatoire.

$$\begin{aligned} P(2 \in]-\infty, X]) &= P(2 \leq X) \\ &= 1 - P(X < 2) = 1 - \Phi(2) \quad \text{où } \Phi : \text{fonction de répartition de la v.a } X \\ &\quad \text{la valeur } \Phi(2) \text{ est lue sur la table de la loi normale standart.} \\ &= 1 - 0.9772 = 0.0228 \end{aligned}$$

ou bien:

$$P(X \geq 2) = \frac{1}{2} - P(0 \leq X \leq 2) = \frac{1}{2} - 0.4772 = 0.0228$$

2) Soit la v.a $X \longrightarrow N(\mu, 1)$, trouver la probabilité que l'intervalle aléatoire $[X-1, X+1]$ contienne μ.

$$\begin{aligned} P(\mu \in [X-1, X+1]) &= P(X-1 \leq \mu \leq X+1) \\ \text{or } X-1 &\leq \mu \leq X+1 \iff \left\{ \begin{array}{c} X-1 \leq \mu \\ et \\ X+1 \geq \mu \end{array} \right\} \iff \left\{ \begin{array}{c} X-\mu \leq 1 \\ et \\ X-\mu \geq -1 \end{array} \right\} \\ \text{donc} &: P(X-1 \leq \mu \leq X+1) = P(-1 \leq X-\mu \leq 1); \quad Z = X-\mu \longrightarrow N(0,1) \\ &= P(-1 \leq Z \leq 1) = \Phi(1) - \Phi(-1) \\ &= \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1 = 2(0.8413) - 1 = 0.6826 \end{aligned}$$

6.6.2 Détermination d'un intervalle de confiance:

Le principe de l'estimation par IC est de proposer un encadrement d'un paramètre inconnu d'une population dont la loi, elle, est connue.

Les différents types de IC sont:

* Intervalle bilatéral symétrique; si $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$.
* Intervalle bilatéral ; si $\alpha_1 \neq 0, \alpha_2 \neq 0$ et $\alpha_1 + \alpha_2 = \alpha$.
* Intervalle unilatéral à gauche; si $\alpha_2 = 0$.
* Intervalle unilatéral à droite; si $\alpha_1 = 0$.

Graphe

Alors, d'une façon générale on a :

$$\begin{aligned} P(\theta \in [a,b]) &= 1 - \alpha \iff P(a \leq \theta \leq b) = 1 - \alpha \\ \text{donc } P(\theta \notin [a,b]) &= \alpha \end{aligned}$$

On choisit dans la plus part des cas un intezrvalle de probabilité à risques symétriques $\frac{\alpha}{2}$ c-à-d

$$P(\theta < a) = \frac{\alpha}{2}, \qquad P(\theta > b) = \frac{\alpha}{2}$$

6.7 Intervalle de confiance d'une moyenne:

On distingue 2 cas.

6.7.1 1) σ connu (et $n \geq 30$) :

Si $X \longrightarrow N\left(m, \sigma^2\right)$ ou $X \longrightarrow$loi quelquonque avec n grand $(n \geq 30)$.

On sait que $\overline{X} \longrightarrow N\left(m, \frac{\sigma^2}{n}\right)$ (si $X \longrightarrow$loi quelquonque, on applique T.C.L) et on sait que $E\left(\overline{X}\right) = m$,(estimateur sans biais).

donc IC de m est symétrique par rapport à $\overline{X}$.

graphe

Remarque: Lorsque l'estimateur $\widehat{\theta}$ est sans biais, il est naturel de construire un intervalle centré sur l'estimation ponctuelle obtenue pour θ.

Donc, en partageant le risque α en 2 parties égales et en lisant sur la table $N(0,1)$, on peut chercher $U_{\frac{\alpha}{2}}$ vérifiant ce qui suit:

on a:

$$\overline{X} \longrightarrow N\left(m, \frac{\sigma^2}{n}\right) \Longrightarrow Z = \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \longrightarrow N(0,1)$$

où Z est une v.a libre

Définition: On dit qu'une v.a est libre si sadistribution de probabilité ne dépend pas des paramètres du modèle.

Et donc,

$$\begin{aligned} P\left(-U_{\frac{\alpha}{2}} \leq Z \leq U_{\frac{\alpha}{2}}\right) &= 1-\alpha \\ &\Longleftrightarrow P\left(-U_{\frac{\alpha}{2}} \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq U_{\frac{\alpha}{2}}\right) = P\left(\overline{X} - U_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + U_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \end{aligned}$$

d'où:

$$IC_\alpha = \left[\overline{x} - U_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + U_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$$

et on a :

$$\begin{aligned} P\left(-U_{\frac{\alpha}{2}} \leq Z \leq U_{\frac{\alpha}{2}}\right) &= 1-\alpha \\ &\Longleftrightarrow \Phi\left(U_{\frac{\alpha}{2}}\right) - \Phi\left(-U_{\frac{\alpha}{2}}\right) = 1-\alpha \\ &\Longleftrightarrow \Phi\left(U_{\frac{\alpha}{2}}\right) - \left(1 - \Phi\left(U_{\frac{\alpha}{2}}\right)\right) = 1-\alpha \\ &\Longleftrightarrow 2\Phi\left(U_{\frac{\alpha}{2}}\right) - 1 = 1-\alpha \\ &\Longleftrightarrow \Phi\left(U_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} \\ &\Longleftrightarrow U_{\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \text{ puis lire sur la table } N(0,1) \text{ la valeur } U_{\frac{\alpha}{2}} \end{aligned}$$

6.7.2 Exemple:

Si $\alpha = 5\% = 0.05$ donc $1 - \alpha = 0.95$

$$\begin{aligned} \Phi\left(U_{\frac{\alpha}{2}}\right) &= \frac{1+0.95}{2} = 0.975 \\ &\Longrightarrow \quad U_{\frac{\alpha}{2}} = \Phi^{-1}(0.975) = 1.96 \end{aligned}$$

Donc:

$$\alpha = 5\% \longrightarrow U_{\frac{\alpha}{2}} = 1.96$$

et on obtient:

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

6.7.3 Remarque:

On a:

$$\begin{aligned} P\left(U_{\alpha_1} \leq Z \leq U_{\alpha_2}\right) &= 1 - \alpha \\ &\Longleftrightarrow \quad P\left(U_{\alpha_1} \leq \frac{\overline{X} - m}{\frac{\sigma}{\sqrt{n}}} \leq U_{\alpha_2}\right) = 1 - \alpha \\ &\Longleftrightarrow \quad P\left(\overline{X} - U_{1-\alpha_2}\frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X} - U_{\alpha_1}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \end{aligned}$$

donc:

$$IC_\alpha = \left[\overline{x} - U_{1-\alpha_2}\frac{\sigma}{\sqrt{n}}, \overline{x} - U_{\alpha_1}\frac{\sigma}{\sqrt{n}}\right]$$

où U_{α_1} et $U_{1-\alpha_2}$ sont des valeurs lues dans la table de la loi $N(0,1)$.
Graphe

6.7.4 Remarque: σ connu et $n < 30$

La méthode ne s'applique que si nous supposons que tous les X_i sont de loi normale (loi de la population normale) et donc IC_α de m est le même que dans le cas σ connu et $n \geq 30$.

6.7.5 2) σ inconnu (et $n < 30$) :

Rappel: Loi de Student Soient deux variables aléatoires indépendantes U et X telles que $U \longrightarrow N(0,1)$, $X \longrightarrow \chi^2_n$.

On définit alors la v.a T à n degrès de liberté:

$$T_n = \frac{U}{\sqrt{\frac{X}{n}}}; \quad T_n \longrightarrow St(n)$$

où sa densité de probabilité est donnée par:

$$\begin{aligned} f(t) &= \frac{1}{\sqrt{n}\beta\left(\frac{1}{2},\frac{n}{2}\right)\left(1+\frac{t^2}{n}\right)^{\frac{n+1}{2}}} \\ \text{où} &: \beta(n,p)=\frac{\Gamma(n)\Gamma(p)}{\Gamma(n+p)} \end{aligned}$$

et

$$E(T_n)=0 \quad \text{si} \quad n>1$$

$$Var(T_n)=\frac{n}{n-2} \quad \text{si} \quad n>2$$

Avec la remarque suivante:

- si $n \longrightarrow +\infty; T_n \longrightarrow^{Loi} N(0,1)$

- si $n = 1$, la loi de student est la loi de Cauchy de d.d.p
  $f(t) = \frac{1}{\pi(1+t^2)}$

Donc dans ce cas, on doit supposer que la loi de la population est normale et puisque σ est inconnu, on l'estime par S^* qui est un estimateur sans biais.
on a:

$$S^{*2}=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2, \qquad S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$$

alors;

$$S^{*2}=\frac{n}{n-1}S^2$$

et

$$\begin{aligned} X &\longrightarrow N\left(m,\sigma^2\right) \Longrightarrow \overline{X} \longrightarrow N\left(m,\frac{\sigma^2}{n}\right) \\ &\Longrightarrow \frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}=\frac{\overline{X}-m}{\frac{S^*}{\sqrt{n}}} \end{aligned}$$

or

$$\begin{aligned} S^{*2} &= \frac{n}{n-1}S^2 \Longrightarrow S^*=S\frac{\sqrt{n}}{\sqrt{n-1}} \\ &\Longrightarrow \frac{S^*}{\sqrt{n}}=\frac{S}{\sqrt{n-1}} \end{aligned}$$

D'où:

$$\begin{aligned}\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} &= \frac{\overline{X}-m}{\frac{S^*}{\sqrt{n}}} \\ &= \frac{\overline{X}-m}{\frac{S}{\sqrt{n-1}}} = T\end{aligned}$$

et on montre que la v.a $T = \frac{\overline{X}-m}{\frac{S}{\sqrt{n-1}}}$ suit la loi de Student à $(n-1)$ degrès de liberté.

En effet:

$$\begin{aligned}\overline{X} &\longrightarrow N\left(m, \frac{\sigma^2}{n}\right) \Longrightarrow \frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} \longrightarrow N(0,1) \\ \text{et} \quad \frac{nS^2}{\sigma^2} &\longrightarrow \chi^2_{n-1}\end{aligned}$$

donc:

$$\begin{aligned}T &= \frac{\overline{X}-m}{\frac{S}{\sqrt{n-1}}} = \frac{\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{n-1}\sqrt{\frac{n}{\sigma^2}}}} \\ &= \frac{\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{nS^2}{\sigma^2}\frac{1}{n-1}}} = \frac{\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{nS^2}{\sigma^2}}{n-1}}} \\ \text{où} \quad \frac{\overline{X}-m}{\frac{\sigma}{\sqrt{n}}} &\longrightarrow N(0,1) \quad \text{et} \quad \frac{nS^2}{\sigma^2} \longrightarrow \chi^2_{n-1} \\ &\Longrightarrow T \longrightarrow St(n-1)\, d.d.l\end{aligned}$$

Et on remarque que la v.a T est une statistique libre. Et du fait que la distribution de Student est symétrique, alors on peut partager le risque α en 2 parties égales, on obtient alors:

$$\begin{aligned}P\left(-t_{\frac{\alpha}{2}} \leq \frac{\overline{X}-m}{\frac{S}{\sqrt{n-1}}} \leq t_{\frac{\alpha}{2}}\right) &= 1-\alpha \\ F\left(t_{\frac{\alpha}{2}}\right) - F\left(-t_{\frac{\alpha}{2}}\right) &= 1-\alpha \quad \text{or la loi de Student est symétrique} \\ &\Longrightarrow 2F\left(t_{\frac{\alpha}{2}}\right) - 1 = 1-\alpha \\ &\Longrightarrow F\left(t_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} \Longrightarrow t_{\frac{\alpha}{2}} = F^{-1}\left(1-\frac{\alpha}{2}\right)\end{aligned}$$

Donc:

$$IC_\alpha = \left[\overline{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}}\ ,\quad \overline{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n-1}}\right]$$

Remarque: 1) Il est possible de déterminer graphiquement ces intervalles; c'est ce qu'on appelle construction d'abaques. Ce principe est basé sur la construction d'une carte qui nous donne, par simple lecture, l'intervalle cherché.

2) La quantité

$$W = 2F^{-1}\left(1 - \frac{\alpha}{2}\right)\sqrt{\frac{1}{n\,(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

est appelé "étendue de Student". C'est une statistique utile dans certains problèmes de test.

6.7.6 σ inconnu (et $n \geq 30$) :

La loi de la population est quelquonque, mais comme $n \geq 30$, donc la loi de Student converge vers la loi normale (Quand $n \longrightarrow +\infty$), et donc on remplace $t_{\frac{\alpha}{2}}$ par $u_{\frac{\alpha}{2}}$.

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

6.7.7 Remarques:

1) Si on considère que le tirage se fait sans remise (exhaustif), alors

$$Var\left(\overline{X}\right) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

et donc IC_α de la moyenne dans le cas où σ connu, la population est de loi normale ou $n \geq 30$ si la polulation n'est pas de loi normale.

$$IC_\alpha = \left[\overline{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \;\;,\;\; \overline{x} + u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

2) Un intervalle de confiance de la moyenne est toujours centré sur une estimation de cette moyenne issue d'un échantillonnage aléatoire.

6.7.8 Exemple:

Dans un pays, l'écart-type du poids de nouveaux nés= $450g$. on a observé sur 700 nouveaux nés une moyenne de poids de $3365g$.

Donner IC_α , au niveau de confiance de 95%, puis au seuil 1% du poids moyen des nouveaux nés de ce pays.

Solution:

Soit la v.a X : poids des nouveaux nés, $E(X) = m, V(X) = \sigma^2$.

On se trouve dans le cas où σ^2 est connue et comme $n = 700 > 30$

donc

$$\begin{aligned} \sigma &= 450, n = 700 > 30, \quad \text{la population a une distribution quelquonque} \\ 1-\alpha &= 0.95 \Longrightarrow \alpha = 0.05 = 5\% \end{aligned}$$

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

$$\overline{x} = 3365, \quad \alpha = 0.05 \Longrightarrow \frac{\alpha}{2} = 0.025$$

on a:

$$\begin{aligned} P\left(Z > u_{\frac{\alpha}{2}}\right) &= \frac{\alpha}{2} \iff 1 - P\left(Z \leq u_{\frac{\alpha}{2}}\right) = \frac{\alpha}{2} \\ &\iff P\left(Z \leq u_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} = 0.975 \end{aligned}$$

on lit sur la table $N(0,1)$ la valeur $u_{\frac{\alpha}{2}} = \Phi^{-1}(0.975) = 1.96$

$$\alpha = 5\% \longrightarrow u_{\frac{\alpha}{2}} = 1.96$$

ainsi:

$$\begin{aligned} IC_{5\%} &= \left[3365 - 1.96\left(\frac{450}{\sqrt{700}}\right) \ , \ 3365 + 1.96\left(\frac{450}{\sqrt{700}}\right)\right] \\ &= [3331, 3398] \end{aligned}$$

Si

$$\begin{aligned} \alpha &= 1\% = 0.01, \quad P\left(Z > u_{\frac{\alpha}{2}}\right) = P\left(Z > u_{\frac{0.01}{2}}\right) = \frac{\alpha}{2} = \frac{0.01}{2} \\ &\iff P\left(Z \leq u_{\frac{0.01}{2}}\right) = 1 - \frac{0.01}{2} = 0.995 \end{aligned}$$

on lit sur la table $N(0,1)$ la valeur $u_{\frac{\alpha}{2}} = \Phi^{-1}(0.995) = 2.58$
alors:

$$\begin{aligned} IC_{1\%} &= \left[3365 - 2.58\left(\frac{450}{\sqrt{700}}\right) \ , \ 3365 + 2.58\left(\frac{450}{\sqrt{700}}\right)\right] \\ &= [3321, 3409] \end{aligned}$$

On remarque $IC_{5\%} \subset IC_{1\%}$.

c'est-à-dire l'estimation au seuil 5% est plus précise, ce qui est normal puisque la probabilité d'erreur est plus grande dans ce cas.

6.8 Intervalle de confiance de la variance

On suppose que la population suit la loi normale $N\left(m, \sigma^2\right)$, avant de déterminer un intervalle de confiance IC_α pour la variance, il faut d'abord chercher un estimateur sans biais pour cette variance.

On considère alors 2 situations.

6.8.1 i) m connu (cas peu fréquent)

Dans ce cas, on prend pour estimateur de σ^2 la statistique définie par:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$$

qui est le meilleur estimateur de σ^2 car:

$$\begin{aligned} E\left(s^2\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i - m\right)^2 = \frac{1}{n}\sum_{i=1}^{n} V\left(X_i\right) \\ &= \sigma^2 \quad \text{du fait de l'indépendance des } X_i \\ &\quad \text{donc } s^2 \text{ est sans biais} \end{aligned}$$

et

$$\begin{aligned} \frac{ns^2}{\sigma^2} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - m)^2 = \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 \longrightarrow \chi^2_n \\ \text{car } \frac{X_i - m}{\sigma} &\longrightarrow N(0,1) \end{aligned}$$

donc

$$\begin{aligned} V\left(\frac{ns^2}{\sigma^2}\right) &= \frac{n^2}{\sigma^4}V\left(s^2\right) = 2n \\ &\Longrightarrow \quad V\left(s^2\right) = \frac{\sigma^4}{n^2}2n = \frac{2}{n}\sigma^4 \longrightarrow_{n\to+\infty} 0 \end{aligned}$$

6.8.2 ii) m inconnu

Dans ce cas, on prend pour estimateur de σ^2 l'estimateur sans biais

$$S^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \quad E\left(S^{*2}\right) = \sigma^2, \quad V\left(S^{*2}\right) = \frac{2}{n-1}\sigma^4 \longrightarrow_{n\to+\infty} 0$$

et on sait que

$$\frac{n-1}{\sigma^2}S^{*2} = \frac{n}{\sigma^2}S^2 \longrightarrow \chi^2_{n-1}$$

De i) et ii) on remarque que la distribution de l'estimateur de σ^2 est une χ^2 que soit m connu ou non.

L'objectif est de déterminer un intervalle $\left[S_1^2,\ S_2^2\right]$ tel que

$$P\left(S_1^2 \leq \sigma^2 \leq S_2^2\right) = 1 - \alpha$$

On a:

$$\begin{aligned} P\left(\chi^2_{\alpha_1} \leq \frac{ns^2}{\sigma^2} \leq \chi^2_{1-\alpha_2}\right) &= 1-\alpha \\ &\iff P\left(\frac{\chi^2_{\alpha_1}}{ns^2} \leq \frac{1}{\sigma^2} \leq \frac{\chi^2_{1-\alpha_2}}{ns^2}\right) = 1-\alpha \\ &\iff P\left(\frac{ns^2}{\chi^2_{1-\alpha_2}} \leq \sigma^2 \leq \frac{ns^2}{\chi^2_{\alpha_1}}\right) = 1-\alpha \end{aligned}$$

donc

$$IC_\alpha = \left[\frac{ns^2}{\chi^2_{1-\alpha_2}} \quad , \quad \frac{ns^2}{\chi^2_{\alpha_1}}\right]$$

graphe

6.8.3 Remarques:

1) Si $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, alors

$$IC_\alpha = \left[\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad , \quad \frac{ns^2}{\chi^2_{\frac{\alpha}{2}}}\right]$$

graphe

2) Si m est connu, la χ^2 est de n d.d.l
et si m est inconnu, la χ^2 est de $(n-1)$ d.d.l

6.8.4 Exemple:

Un échantillon de 30 composants extrait d'une population a donné un écart-type $s = 10$. Trouver IC à 95% de l'écart-type pour l'ensemble de la population.

Solution

On a $n = 30$, $s^2 = 100$, m inconnu, $1-\alpha = 0.95 \Longrightarrow \alpha = 0.05$, $\frac{\alpha}{2} = 0.025$, $1-\frac{\alpha}{2} = 0.975$

$$IC_\alpha = \left[\frac{ns^2}{\chi^2_{1-\frac{\alpha}{2}}} \quad , \quad \frac{ns^2}{\chi^2_{\frac{\alpha}{2}}}\right]$$

avec

$$\begin{aligned} \chi^2_{0.025} &= 16.047, \quad \chi^2_{1-0.025} = \chi^2_{0.975} = 45.722 \\ &\quad \text{ces valeurs sont lues sur la table du } \chi^2_{n-1} \text{ d.dl} \end{aligned}$$

donc

$$IC_{5\%} = \left[\frac{30\,(100)}{45.722} \quad , \quad \frac{30\,(100)}{16.047}\right] = [65.64 \quad , \quad 187.5]$$

6.9 Intervalle de confiance d'une proportion

Dans la population, une proportion p d'individus possède un certain caractère. On cherche un intervalle de confiance IC_α pour p à partir de la valeur de F,

où F v.a qui représente la proportion d'individus possédant le caractère dans l'échantillon.

(on suppose que le tirage est avec remise et $n \geq 30$), donc on peut approximer la loi de F qui est binomiale par une loi normale.

$$F \longrightarrow N\left(p,\ \frac{pq}{n}\right) \Longrightarrow U = \frac{F-p}{\sqrt{\frac{pq}{n}}} \longrightarrow N(0,1)$$

alors:

$$\exists\ p_1, p_2\ /\ P(p_1 \leq p \leq p_2) = 1-\alpha$$

graphe

on a

$$P(u_{\alpha_1} \leq U \leq u_{1-\alpha_2}) = 1-\alpha$$

$$\begin{aligned} P\left(u_{\alpha_1} \leq \frac{F-p}{\sqrt{\frac{pq}{n}}} \leq u_{1-\alpha_2}\right) &= 1-\alpha \\ &\iff P\left(u_{\alpha_1}\sqrt{\frac{pq}{n}} \leq F-p \leq u_{1-\alpha_2}\sqrt{\frac{pq}{n}}\right) = 1-\alpha \\ &\iff P\left(F-u_{1-\alpha_2}\sqrt{\frac{pq}{n}} \leq p \leq F-u_{\alpha_1}\sqrt{\frac{pq}{n}}\right) = 1-\alpha \end{aligned}$$

donc

$$p_1 = F-u_{1-\alpha_2}\sqrt{\frac{pq}{n}} \ \text{ et } \ p_2 = F-u_{\alpha_1}\sqrt{\frac{pq}{n}}$$

on remarque que p se trouve dans les bornes, donc il faut le remplacer.

1) On remplace p par f et q par $1-f$ (car F est un estimateur sans biais de p et f est une valeur de F), alors

$$P\left(F-u_{1-\alpha_2}\sqrt{\frac{f(1-f)}{n}} \leq p \leq F-u_{\alpha_1}\sqrt{\frac{f(1-f)}{n}}\right) = 1-\alpha$$

ainsi

$$IC_\alpha = \left[f-u_{1-\alpha_2}\sqrt{\frac{f(1-f)}{n}}\ ,\ f-u_{\alpha_1}\sqrt{\frac{f(1-f)}{n}}\right]$$

Sous les conditions:

$$np_1(1-p_1) > 3 \ \text{ et } \ np_1(1-p_1) > 3$$

qui permettent l'approximation vers la loi normale.

$\divideontimes$ Si $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ (intervalle bilateral symétrique):

$$IC_\alpha = \left[F - u_{\frac{\alpha}{2}} \sqrt{\frac{f\,(1-f)}{n}} \;\;,\;\; F + u_{\frac{\alpha}{2}} \sqrt{\frac{f\,(1-f)}{n}} \right]$$

2) Méthode par excès:

$$\begin{aligned} p \quad &\leq \quad \frac{1}{2} \Longrightarrow pq \leq \frac{1}{4} \\ &\Longrightarrow \quad \sqrt{\frac{pq}{n}} \leq \sqrt{\frac{1}{4n}} = \frac{1}{2\sqrt{n}} \end{aligned}$$

donc

$$IC_\alpha = \left[f - u_{1-\alpha_2} \frac{1}{2\sqrt{n}} \;\;,\;\; f - u_{\alpha_1} \frac{1}{2\sqrt{n}} \right]$$

$$IC_\alpha = \left[f - u_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}} \;\;,\;\; f + u_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}} \right] \qquad \text{si} \;\; \alpha_1 = \alpha_2 = \frac{\alpha}{2}$$

avec:

$$P\left(p \in \left[f \pm u_{\frac{\alpha}{2}} \frac{1}{2\sqrt{n}} \right] \right) \geq 1 - \alpha$$

6.9.1 Exemple:

On veut estimer, dans une population de plusieurs millions, le pourcentage p de sujets porteurs d'un certain virus. On examine un échantillon de 12600 personnes. Le virus est détecté chez 630 d'entre elles.

Déterminer IC_α , $\alpha = 1\%$, pour p.

Solution

On a

$$n = 12600, \quad f = \frac{630}{12600} = 0.05, \quad 1 - f = 0.95$$

$u_{\frac{\alpha}{2}} = u_{0.005}, \Phi\left(u_{\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2} = 0.995 \Longrightarrow u_{\frac{\alpha}{2}} = 2.58$ cette valeur est lue sur la table $N(0,1)$

$$\begin{aligned} p_1 \quad &= \quad 0.05 - 2.58 \sqrt{\frac{0.05\,(0.95)}{12600}} = 0.045 \\ p_2 \quad &= \quad 0.05 + 2.58 \sqrt{\frac{0.05\,(0.95)}{12600}} = 0.055 \end{aligned}$$

$$IC_{1\%} = [0.045 \;\;,\;\; 0.055]$$

avec

$$\begin{aligned} np_1\,(1-p_1) \quad &= \quad 541.485 > 3 \\ np_2\,(1-p_2) \quad &= \quad 654.885 > 3 \end{aligned}$$

6.10 Intervalle de confiance de la différence de 2 moyennes

On extrait un échantillon de taille n_i de la population P_i $(i = 1, 2)$ de moyenne m_i et d'écart-type σ_i.

Dans le cas de distributions normales et d'indépendance des 2 échantillons on a:

$$\begin{aligned} \overline{X}_1 - \overline{X}_2 &\longrightarrow N\left(m_1 - m_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \\ \text{car } \overline{X}_1 &\longrightarrow N\left(m_1,\ \frac{\sigma_1^2}{n_1}\right), \quad \overline{X}_2 \longrightarrow N\left(m_2,\ \frac{\sigma_2^2}{n_2}\right) \end{aligned}$$

donc:

$$\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \longrightarrow N\,(0,\ 1)$$

graphe???

ainsi IC_α , si le tirage se fait avec remise, est alors:

$$\begin{aligned} (\overline{x}_1 - \overline{x}_2) - u_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} &\leq (m_1 - m_2) \leq (\overline{x}_1 - \overline{x}_2) + u_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ \text{ici } \alpha_1 &= \alpha_2 = \frac{\alpha}{2} \end{aligned}$$

dans le cas général: $\alpha_1 \neq 0,\ \alpha_2 \neq 0,\ \alpha_1 + \alpha_2 = \alpha$

$$(\overline{x}_1 - \overline{x}_2) - u_{1-\alpha_2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq (m_1 - m_2) \leq (\overline{x}_1 - \overline{x}_2) - u_{\alpha_1}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

graphe??

Chapitre 3

TESTS D'HYPOTHESES

7 Introduction:

D'une manière générale, il s'agit, à partir de l'étude d'un ou plusieurs échantillons, de préciser comment prendre des décisions concernant l'ensemble de la population. Naturellement, comme on ne dispose pas de renseignements sur l'ensemble de la population, on risque de se tromper en prenant la décision et il importe de contrôler au maximum tout risque d'erreur. On commence par formuler une hypothèse, appelée hypothèse nulle (H_0) et on est amené à l'accepter ou la rejeter; rejeter H_0 c'est accepter H_1 (hypothèse alternative).

Donc on est face de 2 hypothèses H_0 et H_1 dont une seule est vraie. Les règles de décision qui permettent de faire un choix entre H_0 et H_1 sont appelées tests statistiques.

Définition

Un test est un mécanisme qui permet de trancher entre 2 hypothèses au vu des résultats d'un échantillon.

Soient H_0 et H_1 ces hypothèses, dont une seule est vraie. la décision aboutira à choisir H_0 ou H_1 et pour chaque décision, il y a un risque d'erreur.

7.1 Risque d'erreur

Il y a 4 possibilités qui peuvent se présenter et si on note par:

α =P(rejeter H_0 sachant que H_0 est vraie)

on appelle α : risque d'erreur de 1ère espèce.

β =P(accepter H_0 sachant que H_0 est fausse)

on appelle β: risque d'erreur de 2ème espèce.

On remarque que dans les 2 cas, il a été commis une erreur.

état/décision	H_0		H_1	
H_0 est vraie	$1-\alpha$	(Décision correcte)	α	(Erreur 1ère espèce)
H_0 est fausse	β	(Erreur 2ème espèce)	$1-\beta$	(Décision correcte)

Dans la pratique, on essaie de contrôler et maitriser α qu'on appelle le seuil de signification du test (en général, α est choisi égal à 5%, 1%, 0.1%) et ainsi on définit une région critique au seuil α.

Graphe??

L'idéal serait de rendre α, β les plus petits possibles, mais en diminuant α on agrandit la région d'acceptation de H_0, donc, le plus souvent, on augmente β (probabilité d'accepter H_0 alors que H_0 est fausse). En général, n étant fixé, quand $\alpha \searrow, \beta \nearrow$ et inversement. La seule façon pour que $\alpha \searrow$ et $\beta \searrow$ en même temps est que $n \nearrow$ (augmenter la taille de l'échantillon) ce qui n'est pas toujours possible. En fait, la plupart du temps, les erreurs des 2 types n'ont pas la même importance et on essaie de limiter la plus grave.

7.1.1 Exemple:

Si on ne se rappelle pas si l'examen commence à 9h ou à 10h, le risque de la 1ère espèce est d'arriver à 10h alors que l'examen a commencé à 9h et le risque de la 2ème espèce est d'arriver à 9h alors que l'examen commence à 10h.

Si on note par H_0: l'examen commence à 9h (c'est là qu'on a l'erreur la plus grave).

H_1: l'examen commence à 10h.

⊙ On appelle $1-\beta$ la puissance du test.

⊙ Un test est dit sans biais si $1-\beta > \alpha$.

⊙ Un test est dit convergent si $1-\beta \longrightarrow_{n\to+\infty} 1$.

7.1.2 Remarque:

H_0 doit être simlpe: elle ne mentionne qu'une seule valeur du paramètre.

H_1 mentionne, en général, plusieurs valeurs du paramètre.

L'hypothèse nulle H_0 est l'hypothèse que l'on désire contrôler; elle consiste à dire qu'il n'existe pas de différence entre les paramètres comparés ou que la différence observée n'est pas significative et elle est due aux fluctuations d'échantillonage.

7.1.3 Test unilatéral et test bilatéral

Avant d'appliquer tout test statistique, il s'agit de bien définir le problème posé. En effet, selon les hypothèses posées, on applique soit un test bilatéral, soit un test unilatéral.

Le test bilatéral s'applique quand on cherche une différence entre 2 estimations ou entre une estimation et une valeur donnée sans se préoccuper du signe ou du sens de la différence. Dans ce cas, la zone de rejet de l'hypothèse principale se fait de part et d'autre de la distribution de référence.

Exemple Pour un secteur donné, on affirme que le salaire mensuel moyen m est de 40000 dinars. Pour vérifier cette assertion, on tire un échantillon de salaires du secteur même et on effectue un test sur la moyenne m.

soit: $H_0 : m = 40000$ contre $H_1 : m \neq 40000$

H_0 est une hypothèse simple, alors que H_1 est une hypothèse multiple. Il s'agit d'un test bilatéral.

Le test unilatéral s'applique quand on cherche à savoir si une estimation est supérieure (ou inférieure) à une autre ou à une valeur donnée. la zone de rejet de l'hypothèse principale est située d'un seul côté de la distribution de probabilité de référence.

Exemple Pour le même secteur, on affirme que la variance des salaires est au moins égale à 120. Pour vérifier ceci, on tire un échantillon de salaires du secteur même et on effectue un test sur la variance σ^2.

soit: $H_0 : \sigma^2 \geq 120$ contre $H_1 : \sigma^2 < 120$

H_0, H_1 sont des hypothèses multiples, il s'agit d'un test unilatéral à gauche.
Remarque
$H_0 : m = 40000$ contre $H_1 : m > 40000$ (test unilateral à droite).
Un test unilatéral est plus puissant qu'un test bilatéral.

8 Tests paramétriques sur un échantillon

La question posée revient à comparer la caractéristique d'un échantillon (moyenne, variance, proportion) à une valeur de référence connue a priori (test paramétrique).

On peut distinguer 2 types de tests: les tests de conformité à la population et les tests d'homogénéité de 2 échantillons entre eux.

9 Tests de conformité

Ces tests sont destinés à vérifier si un échantillon peut-être considéré comme extrait d'une population donnée ou représentatif de cette population, vis-à-vis d'un paramètre (moy, var, fréq) observé. Ceci implique que la loi théorique du paramètre est connue au niveau de la population.

9.1 Test sur la moyenne:

Soit une population sur laquelle on étudie une variable aléatoire réelle X, de moyenne inconnue m et de variance σ^2. On cherche à tester l'hypothèse $m = m_0$ (m_0 : valeur connue et fixée) contre une hypothèse alternative à préciser; en se basant sur l'observation d'un échantillon $X_1, ..., X_n$.

9.1.1 Test bilatéral:

On teste

$$H_0 : m = m_0 \qquad / \qquad H_1 : m \neq m_0$$

on sait déterminer IC_α de m, et on sait que

$$P\left(m \in IC_\alpha\right) = 1 - \alpha$$

$\oplus$ si $m_0 \notin IC_\alpha$ on rejette H_0 (on accepte H_1)

$\oplus$ si $m_0 \in IC_\alpha$ on accepte H_0 (on dit que l'écart n'est pas significatif au seuil α)

9.1.2 Cas où: $n \geq 30$, σ^2 connue:

On a:

$$\begin{aligned} m_0 \quad \in \quad IC_\alpha &\iff \overline{x} - u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq m_0 \leq \overline{x} + u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \\ &\iff -u_{\frac{\alpha}{2}} \leq \frac{\overline{x} - m_0}{\frac{\sigma}{\sqrt{n}}} \leq u_{\frac{\alpha}{2}} \\ \text{c'est à dire } Z \quad = \quad & \frac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}}, \quad Z \longrightarrow N(0,1) \quad \text{sous} \quad H_0 \end{aligned}$$

on accepte H_0 si $-u_{\frac{\alpha}{2}} \leq Z \leq u_{\frac{\alpha}{2}}$ (zone d'acceptation), on la rejette si $Z > u_{\frac{\alpha}{2}}$ ou $Z < -u_{\frac{\alpha}{2}}$ (zone de rejet ou zone critique).

en d'autres termes; ici on rejette l'égalité soit si l'écart est trop grand ou trop petit et le risque de rejeter à tord est partagé en 2; $\frac{\alpha}{2}$ de chaque côté.

On applique la même démarche pour les autres IC_α et on obtient le tableau suivant:

Test bilatéral sur la moyenne

cas	loi utilisée	zone de rejet; $H_1 : m \neq m_0$
σ^2 connue, $X \longrightarrow N$	$Z = \left(\overline{X} - m_0\right) / \frac{\sigma}{\sqrt{n}} \longrightarrow N(0,1)$ sous H_0	$z > u_{\frac{\alpha}{2}}$ ou $z < -u_{\frac{\alpha}{2}}$
σ^2 connue, $n \geq 30$	$Z = \left(\overline{X} - m_0\right) / \frac{\sigma}{\sqrt{n}} \backsimeq N(0,1)$ sous H_0	$z > u_{\frac{\alpha}{2}}$ ou $z < -u_{\frac{\alpha}{2}}$
σ^2 inconnue, $X \longrightarrow N$ et $n < 30$	$T_{n-1} = \left(\overline{X} - m_0\right) / \frac{s}{\sqrt{n-1}} \longrightarrow$Student sous H_0	$t_{n-1} > t_{\frac{\alpha}{2}}$ ou $t_{n-1} < -t_{\frac{\alpha}{2}}$
σ^2 inconnue, $n \geq 30$	$Z = \left(\overline{X} - m_0\right) / \frac{s}{\sqrt{n-1}} \backsimeq N(0,1)$ sous H_0	$z > u_{\frac{\alpha}{2}}$ ou $z < -u_{\frac{\alpha}{2}}$

où; z est la réalisation de la v.a Z (et t est la réalisation de la v.a T_{n-1}) .

Graphe??

9.1.3 Tests unilatéraux:

On teste les hypothèses

1) $H_0 : m = m_0$ / $H_1 : m > m_0$ (test unilatéral à droite)

on a

$$\begin{aligned} P\left(\text{rejeter } H_0 / \ H_0 \text{ vraie}\right) &= \alpha \\ P\left(\overline{X} > \overline{x}_c \ / \ H_0 \text{ vraie}\right) &= \alpha \\ P\left(\frac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}} > \frac{\overline{x}_c - m_0}{\frac{\sigma}{\sqrt{n}}} \ / \ H_0 \text{ vraie}\right) &= \alpha \\ P\left(Z > u_\alpha \ / \ H_0 \text{ vraie}\right) &= \alpha, \ \ Z = \frac{\overline{X} - m_0}{\frac{\sigma}{\sqrt{n}}} \longrightarrow N(0,1), \ u_\alpha = \frac{\overline{x}_c - m_0}{\frac{\sigma}{\sqrt{n}}} \end{aligned}$$

2) $H_0 : m = m_0 \ \ / \ \ H_1 : m < m_0$ (test unilatéral à gauche)

On procède de la même manière et on obtient le tableau suivant:

Tests unilatéraux sur la moyenne

cas	loi utilisée
σ^2 connue, $X \longrightarrow N$	$Z = \left(\overline{X} - m_0\right) / \frac{\sigma}{\sqrt{n}} \longrightarrow N(0,1)$ sous H_0
σ^2 connue, $n \geq 30$	$Z = \left(\overline{X} - m_0\right) / \frac{\sigma}{\sqrt{n}} \backsimeq N(0,1)$ sous H_0
σ^2 inconnue, $X \longrightarrow N$ et $n < 30$	$T_{n-1} = \left(\overline{X} - m_0\right) / \frac{s}{\sqrt{n-1}} \longrightarrow Student$ sous H_0
σ^2 inconnue, $n \geq 30$	$Z = \left(\overline{X} - m_0\right) / \frac{s}{\sqrt{n-1}} \backsimeq N(0,1)$ sous H_0

zone de rejet; $H_1 : \ m > m_0$	zone de rejet; $H_1 : \ m < m_0$
$z > u_\alpha$	$z < -u_\alpha$
$z > u_\alpha$	$z < -u_\alpha$
$t_{n-1} > t_\alpha$	$t_{n-1} < -t_\alpha$
$z > u_\alpha$	$z < -u_\alpha$

9.1.4 Exemple:

Faire le test suivant

$$\begin{aligned} H_0 \ &: \ m = 25 \ \ / \ \ H_1 : m > 25 \\ n \ &= \ 35, \ \overline{x} = 25.8, \ s = 3 \ \text{ et } \ \alpha = 10\% \end{aligned}$$

Solution
On se trouve dans le cas où $n \geq 30$, σ inconnu donc on utilise la statistique

$$Z = \frac{\overline{X} - m_0}{s/\sqrt{n-1}} \longrightarrow N(0,1) \quad \text{sous } H_0$$

$$z = \frac{\overline{x} - m_0}{s/\sqrt{n-1}} = \frac{25.8 - 25}{3/\sqrt{34}} = 1.55$$

on a

$$\begin{aligned} P(Z > u_\alpha) &= \alpha \Longrightarrow 1 - P(Z \leq u_\alpha) = \alpha \\ &\Longrightarrow P(Z \leq u_\alpha) = 1 - \alpha \\ &\Longrightarrow \Phi(u_\alpha) = 1 - \alpha \\ &\Longrightarrow u_\alpha = \Phi^{-1}(1-\alpha) = \Phi^{-1}(0.9) \\ &\Longrightarrow u_\alpha = 1.29 \end{aligned}$$

finalement, on trouve que $z > u_\alpha$ $(1.55 > 1.29)$ donc on rejette H_0.

Sur la distribution d'échantillonnage de $\overline{X}$, la valeur critique $\overline{x}_c$ est

$$\overline{x}_c = m_0 + u_\alpha \frac{s}{\sqrt{n-1}} = 25 + 1.29\frac{3}{\sqrt{34}} = 25.66$$

on rejette H_0 si $\overline{X} > 25.66$

9.2 Test sur la variance

Soit une population sur laquelle on étudie une v.a.r X de variance inconnue σ^2. On cherche à tester l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ où σ_0^2 est une valeur connue et fixée, en se basant sur l'observation d'un échantillon $X_1, ..., X_n$.

On sait que la statistique $\frac{nS^2}{\sigma^2} \longrightarrow \chi^2_{n-1}$ si $X \longrightarrow N$

Sous l'hypothèse que $X \longrightarrow N$ on teste $H_0 : \sigma^2 = \sigma_0^2$ au seuil α. En utilisant IC_α pour la variance, on obtient le tableau suivant pour le test bilatéral et les tests unilatéraux :

test sur la variance

loi utilisée	zone de rejet;	zone de rejet;	zone de rejet;
	$H_1 : \sigma^2 \neq \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$
$k_{n-1} = \frac{nS^2}{\sigma_0^2} \rightarrow \chi^2_{n-1}$ sous H_0	$k_{n-1} > \chi^2_{1-\frac{\alpha}{2}}$ ou $k_{n-1} < \chi^2_{\frac{\alpha}{2}}$	$k_{n-1} > \chi^2_{1-\alpha}$	$k_{n-1} < \chi^2_{\alpha}$

9.2.1 Exemple:

Tester:

$$\left\{ \begin{array}{l} H_0 : \sigma^2 = 2 \\ H_1 : \sigma^2 \neq 2 \end{array} \right\}$$

$$\text{si } n = 26,\ s^2 = 5.76 \text{ et } \alpha = 0.05$$

Solution

on a:

$$\begin{array}{rcl} k_{n-1} & = & \dfrac{n\ s^2}{\sigma^2} = \dfrac{(26)\,(5.76)}{4} = 37.44 \\ \chi^2_{\frac{\alpha}{2}} & = & \chi^2_{0.025} = 13.12, \quad \chi^2_{1-\frac{\alpha}{2}} = \chi^2_{0.975} = 40.62 \\ \text{comme } k_{n-1} & \not< & \chi^2_{\frac{\alpha}{2}} \text{ et } k_{n-1} \not> \chi^2_{1-\frac{\alpha}{2}} \\ 37.44 & \not< & 13.12 \text{ et } 37.44 \not> 40.62 \end{array}$$

alors, on accepte H_0 au seui $\alpha = 0.05$.

9.3 Test sur la proportion

Soient une variable qualitative prenant 2 modalités (succès=1, échec=0) observée sur une population et un échantillon extrait de cette population. Le but est de savoir si un échantillon de fréquence observée F estimateur de p appartient à une population de référence connue de fréquence p_0 (H_0 vraie) ou à une autre population inconnue de fréquence p (H_1 vraie) .

Si $n \geq 30$, on sait que la v.a

$$F \longrightarrow N\left(p,\ \frac{p\,(1-p)}{n}\right) \Longrightarrow \frac{F-p}{\sqrt{\frac{p(1-p)}{n}}} \longrightarrow N\,(0,1)$$

On teste $H_0 : p = p_0$ / $H_1 : p \neq p_0$ au seuil α. En utilisant IC_α pour la proportion, on obtient le tableau suivant pour le test bilatéral et les tests unilatéraux :

loi utilisée	zone de rejet;	zone de rejet;	zone de rejet;
	$H_1 : p \neq p_0$	$H_1 : p < p_0$	$H_1 : p > p_0$
$Z = (F - p_0)\,/\sqrt{\frac{p_0(1-p_0)}{n}} \backsimeq N\,(0,1)$ sous H_0	$z < -u_{\frac{\alpha}{2}}$ ou $z > u_{\frac{\alpha}{2}}$	$z < -u_\alpha$	$z > u_\alpha$

10 Tests d'homogénéité

Ces tests sont destinés à comparer 2 populations à l'aide d'un nombre équivalent d'échantillons (appelé aussi test d'égalité) et sont les couramment utilisés. Dans ce cas, la loi théorique du paramètre étudié (moyenne, variance, proportion) est inconnue au niveau des populations étudiées.

10.1 Test sur 2 moyennes

Ce test sert à comparer 2 populations relativement à une caractèristique. par exemple, on veut comparer la population de lait de 2 races de vâches.

et on veut tester l'hypothèse

$$H_0 : m_1 = m_2 \ \ / \ \ \left\{ \begin{array}{ccc} & H_1 : m_1 \neq m_2 & \\ H_1 : m_1 < m_2, & \text{ou} & H_1 : m_1 > m_2 \end{array} \right\}$$

si on suppose que $\overline{X}_1$ et $\overline{X}_2$ suivent la loi normale (ce qui est le cas si les populations sont de lois normales ou si $n_1 \geq 30$ et $n_2 \geq 30$), alors la v.a.r

$$D = \overline{X}_1 - \overline{X}_2 \longrightarrow N\left(m_1 - m_2, \quad \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) \quad \text{si } \overline{X}_1 \, et \, \overline{X}_2 \quad \text{sont indépendantes}$$

$\oslash$ Si σ_1^2 , σ_2^2 sont connues, alors:

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \longrightarrow N(0,1)$$

$$\text{sous } H_0, \quad Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \longrightarrow N(0,1)$$

$\oslash$ Si σ_1^2 , σ_2^2 sont inconnues et ($n_1 < 30$ et (ou) $n_2 < 30$) et les populations sont de loi normale:

Si on suppose que $\sigma_1^2 = \sigma_2^2$, on a:

$$\left\{ \begin{array}{l} \overline{X}_1 \longrightarrow N\left(m_1, \quad \frac{\sigma_1^2}{n_1}\right) \text{ et } \frac{n_1 S_1^2}{\sigma^2} \longrightarrow \chi^2_{n_1-1} \\ \overline{X}_2 \longrightarrow N\left(m_2, \quad \frac{\sigma_2^2}{n_2}\right) \text{ et } \frac{n_2 S_2^2}{\sigma^2} \longrightarrow \chi^2_{n_2-1} \end{array} \right\}$$

$$\Longrightarrow \left\{ \begin{array}{c} 1) \quad \frac{n_1 S_1^2}{\sigma^2} + \frac{n_2 S_2^2}{\sigma^2} = \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} \longrightarrow \chi^2_{n_1+n_2-1} \\ 2) \quad \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \longrightarrow N(0,1) \end{array} \right\}$$

on a alors, par définition de la variable de Student

$$T = \frac{\frac{\left(\overline{X}_1 - \overline{X}_2\right) - (m_1 - m_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2(n_1 + n_2 - 2)}}} \quad \text{suit la loi de Student à } (n_1 + n_2 - 2) \text{ d.d.l}$$

et on obtient le tableau du test bilatéral et les tests unilatéraux suivant:

Test sur 2 moyennes

cas	loi utilisée
P_1, P_2 normales, σ_1^2, σ_1^2 connues	$Z = \left(\overline{X}_1 - \overline{X}_2\right) / \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \longrightarrow N(0,1)$ sous H_0
$n_1 \geq 30$, $n_2 \geq 30$ et σ_1^2, σ_2^2 connues	$Z = \left(\overline{X}_1 - \overline{X}_2\right) / \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \backsimeq N(0,1)$ sous H_0
$n_1 \geq 30$, $n_2 \geq 30$ et σ_1^2, σ_2^2 inconnues	$Z = \left(\overline{X}_1 - \overline{X}_2\right) / \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \backsimeq N(0,1)$ sous H_0
$n_1 < 30$ et(ou) $n_2 < 30$,	$T = \left(\overline{X}_1 - \overline{X}_2\right) / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \sqrt{\frac{n_1 S_1^2 + n_2 S_2^2}{(n_1 + n_2 - 2)}}$
P_1, P_2 normales, $\sigma_1^2 = \sigma_2^2$ inconnue	$\rightarrow T_{n_1+n_2-2}$ sous H_0

zone de rejet :$H_1 : m_1 \neq m_2$	zone de rejet: $H_1 : m_1 < m_2$	zone de rejet :$H_1 : m_1 > m_2$
$z > u_{\frac{\alpha}{2}}$ ou $z < -u_{\frac{\alpha}{2}}$	$z < -u_\alpha$	$z > u_\alpha$
$z > u_{\frac{\alpha}{2}}$ ou $z < -u_{\frac{\alpha}{2}}$	$z < -u_\alpha$	$z > u_\alpha$
$z > u_{\frac{\alpha}{2}}$ ou $z < -u_{\frac{\alpha}{2}}$	$z < -u_\alpha$	$z > u_\alpha$
$t > t_{\frac{\alpha}{2}}$ ou $t < -t_{\frac{\alpha}{2}}$	$t < -t_\alpha$	$t > t_\alpha$

10.2 Test sur 2 proportions

Chaque individu de 2 populations P_1 et P_2 peut posséder ou non un certain caractère. Ce caractère est présent en proportion $p_1$et p_2 dans P_1 et P_2.

Le problème est de savoir si la différence entre les 2 fréquences observées est réelle ou explicable par les fluctuations d'échantillons. Ainsi au seuil de signification α, on teste:

$$\left\{ \begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{array} \right\} \text{ ou } \left\{ \begin{array}{l} H_0 : p_1 - p_2 = 0 \\ H_1 : p_1 - p_2 \neq 0 \end{array} \right\}$$

De la population P_i $(i = 1, 2)$ on extrait un échantillon de taille n_i, il lui correspond une fréquence f_i. Si les n_i sont grands $(n_i \geq 30)$, alors la v.a

$$F_i \longrightarrow N\left(p_i, \ \frac{p_i\left(1-p_i\right)}{n_i}\right); \ i = 1, 2$$

et la v.a

$$F_1 - F_2 \longrightarrow N\left(p_1 - p_2, \ \frac{p_1\left(1-p_1\right)}{n_1} + \frac{p_2\left(1-p_2\right)}{n_2}\right)$$

sous H_0 :

$$Z = \frac{F_1 - F_2}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}}} = \frac{F_1 - F_2}{\sqrt{p\left(1-p\right)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \backsim N\left(0, 1\right)$$

on accepte H_0 si la valeur de $f_1 - f_2$ est telle que:

$$\begin{array}{rcl} -u_{\frac{\alpha}{2}} k & \leq & f_1 - f_2 \leq u_{\frac{\alpha}{2}} k \\ \text{avec } \ k & = & \sqrt{p\left(1-p\right)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}} \end{array}$$

dans la pratique, p n'est pas connu, il faut le remplacer par f où

$$f = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

Pour un test bilatéral, on compare z_{obs} à z_{seuil} qui est lue sur la table $N\left(0, 1\right)$ pour α fixé.

Si $|z_{obs}| > z_{seuil}$ H_0 est rejeté au risque α; les deux échantillons sont extraits de deux populations ayant des probabilités de succès respectivement p_1 et p_2.

Si $|z_{obs}| < z_{seuil}$ H_0 est accepté au risque α; les deux échantillons sont extraits de deux populations ayant même probabilité de succès p.

Le tableau suivant résume le test bilatéral et les tests unilatéraux.

Test sur 2 proportions

loi utilisée	zone de rejet; $H_1 : p_1 \neq p_2$
$Z = (F_1 - F_2) / \sqrt{p(1-p)} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \backsimeq N(0,1)$ sous H_0	$z < -u_{\frac{\alpha}{2}}$ ou $z > u_{\frac{\alpha}{2}}$

zone de rejet; $H_1 : p_1 < p_2$	zone de rejet; $H_1 : p_1 > p_2$
$z < -u_\alpha$	$z > u_\alpha$

10.2.1 Exemple

Au cours de 2 livraisons différentes, on observe 48 articles défectueux parmi les 800 constituants la 1ère livraison et 32 articles défectueux parmi les 400 constituants la 2ème livraison. Les 2 pourcentages d'articles défectueux observés sont-ils différents d'une manière significative au seui de 5%?

Solution

On teste $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$

on a

$$z = \frac{f_1 - f_2}{\sqrt{p(1-p)}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ sous } H_0 \text{ et } f = \frac{800\,(48/800) + 400\,(32/400)}{800 + 400} = 0.067$$

$$z = \frac{0.06 - 0.08}{\sqrt{(0.067)(0.933)\left(\frac{1}{800} + \frac{1}{400}\right)}} = -1.31$$

avec la remarque que $np > 5, nq > 5$, on peut donc appliquer l'approximation normale.

Au seuil 5%, $\alpha = 5\% \longrightarrow u_{\frac{\alpha}{2}} = 1.96$ donc la zone d'acceptation de H_0 est $[-1.96, 1.96]$

par conséquent $z \in [-1.96, 1.96]$, d'où la différence observée n'est pas significative.

10.3 Test sur 2 variances

10.3.1 Rappel (loi de Fisher)

Définition Soient X, Y deux v.a indépendantes telles que:

$$X \longrightarrow \chi^2_p, \quad Y \longrightarrow \chi^2_q$$

alors:

$$F(p,q) = \frac{X/p}{Y/q} \longrightarrow \text{loi de Fisher à } p \text{ et } q \text{ d.d.l}$$

Soit X une v.a observée sur 2 populations P_1, P_2 suivant une loi normale de variances inconnues σ_1^2, σ_2^2. On teste au seuil de signification α:

$$\left\{ \begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{array} \right\}$$

On extrait de chaque population P_i un échantillon de taille n_i $(i = 1, 2)$.

Comme

$$\frac{n_i S_i^2}{\sigma_i^2} \longrightarrow \chi^2_{n_i-1} \qquad (i=1,2)$$

La statistique associeé au test de comparaison de 2 variances correspond au rapport des 2 variances estimées.

$$\text{sous } H_0 : F_{obs} = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} = \frac{\frac{n_1 S_1^2}{n_1-1}}{\frac{n_2 S_2^2}{n_2-1}} = F_{n_1-1,n_2-1} \qquad \text{suit une loi de Fisher à } (n_1-1, n_2-1) \text{ d.d.l}$$

on peut affirmer que ce rapport ne doit pas différer significativement de 1.

En pratique, on met toujours au numérateur la plus grande des 2 quantités $\frac{n_1 S_1^2}{n_1-1}$ et $\frac{n_2 S_2^2}{n_2-1}$.

Décision:

On compare F_{obs} à F_{seuil} lue sur la table de loi de Fisher pour un risque d'erreur α fixé et (n_1-1, n_2-1) d.d.l

Si $F_{obs} > F_{seuil}$, l'hypothèse H_0 est rejetée au risque d'erreur α; les 2 échantillons sont extraits de 2 populations ayant des variances statistiquement différentes.

Si $F_{obs} \leq F_{seuil}$, l'hypothèse H_0 est acceptée; les 2 échantillons sont extraits de 2 populations ayant même variance.

En d'autres termes, la région critique est déterminée par

$$P\left(F_{obs} > F_\alpha\right) = \alpha \qquad \text{avec} \quad F_\alpha > 1$$

10.3.2 Remarque

Pour que l'application de ce test soit justifiée, il est impératif que $X \longrightarrow N\left(m, \sigma^2\right)$ et que les 2 échantillons soient indépendants.

10.3.3 Exemple

Avec les données suivantes, faire le test des 2 variances.

$$\alpha = 0.05,\ n_1 = 16,\ s_1 = 5.888,\ n_2 = 11,\ s_2 = 6.292$$

on a:

$$\begin{aligned} \frac{n_1 s_1^2}{n_1-1} &= 36.979, \quad \frac{n_2 s_2^2}{n_2-1} = 43.548 \\ F_{obs} &= \frac{43.548}{36.979} = 1.17 \end{aligned}$$

et puisque $P\left(F_{10,15} > F_\alpha\right) = 0.05$ alors de la table de Fisher, on lit $F_\alpha = 2.54$

comme $F_{obs} = 1.17 < 2.54 = F_\alpha$, on accepte H_0 donc la différence entre les 2 variances n'est pas significative.

10.4 Conclusion

On peut récapituler les tests d'ypothèses de la façon suivante:

1) Définir H_0 (hypothèse à contrôler).

2) Choisir un test statistique ou une statistique pour contrôler H_0.

3) Définir la distribution de la statistique sous H_0.

4) Définir le niveau de signification (seuil) du test ou région critique α.

5) Calculer, à partir des données fournies par l'échantillon, la valeur de la statistique.

6) Prendre une décision concernant H_0 et faire une interpretation.

10.5 Probabilité de signification (p-value)

Il existe 2 strategies pour prendre une décision en ce qui concerne un test statistique:

La 1ère: fixe (à priori) la valeur du seuil de signification α.

la 2ème: établit la valeur de la probabilité critique α_{obs} à posteriori qui est appelée p-value .

10.5.1 Définition

La p-value est la probabilité d'observer une valeur statistique de test au moins aussi "extrême" que celle qui a été calculée à partir de l'échantillon, lorsque H_0 est vraie.

Cette valeur correspond au plus petit niveau de probabilité pour lequel H_0 est rejetée. On compare la p-value à α au lieu de comparer la valeur de la statistique de test à une valeur critique (lue sur la table).

Si la p-value $< \alpha$, alors on rejette H_0.