جامعة مولاي إسماعيل
UNIVERSITÉ MOULAY ISMAÏL

الكلية المتعددة التخصصات
FACULTÉ POLYDISCIPLINAIRE

Licence fondamentale en Sciences Économiques et Gestion
Module échantillonnage et estimation/Semestre 3

Chapitre 3.
Estimation… de l'échantillon à la population

Professeur **Mohamed AIT HOU**

Année universitaire 2023/2024

Plan

- Introduction

I • Estimation: de l'intervalle de pari à l'intervalle de confiance

II • Propriétés des estimateurs

III • Méthodes d'estimation

IV • Intervalle de confiance sur une moyenne

V • Intervalles de confiance sur une variance

VI • Intervalle de confiance sur une proportion

Objectif

Comprendre la problématique de l'estimation;

Savoir utiliser des estimateurs pour des estimations ponctuelles;

Maîtriser toutes les étapes de la construction d'un intervalle de confiance;

Savoir appliquer les résultats concernant les moyennes à des proportions;

Les étapes de construction d'un intervalle de confiance

1. **Spécifier la variable étudiée sur la population et fixer le paramètre à estimer**

2. **Choisir le meilleur estimateur du paramètre**

3. **Calculer une estimation ponctuelle du paramètre sur l'échantillon**

4. **Construire l'intervalle de confiance et calculer la marge d'erreur**

5. **Interpréter les résultats d'un point de vu managérial**

Introduction

Dans un pays X, le Ministère de l'Économie, des Finances et de l'industrie a mis en place la déclaration de revenus pré-remplie par l'administration. Les premiers contrôles effectués sur un échantillon de 780 déclarations d'imposition pré-remplies montrent que:

• 12% de ces déclarations présentent au moins une erreur,

• le montant moyen des erreurs, calculé en dirham près, est de 3500 dh avec un écart type de 800 dh.

Comment estimer dans la population mère composée de toutes les déclarations (la taille de cette population mère est N ≅ 34millions):

-la valeur de la proportion p de déclarations erronées,

- la valeur du montant moyen m des erreurs,

- la va leur de l'écart type e du montant des erreurs?

Structure du chapitre et problème concret d'estimation

Problème concret

Dans un échantillon de taille n = 780.

-La proportion de déclaration erronées vaut $f_{780} = 12\%$

-Erreur moyenne vaut $\bar{X}_{780} = 3500dh$

-Ecart type des erreurs vaut $S_{780} = 800dh$

-À combien peut-on estimer la valeur de ces trois paramètres dans la population mère?

Solution concrète

-L'écart type de l'erreur est estimé à 800,51 dh.

-Il y a 95% de chances pour que l'intervalle de 10% à 14% contienne la proportion inconnue de déclarations erronées.

-il y a 95% de chance pour que l'intervalle de 3443 à 3557 dh contienne l'erreur moyenne.

Section 1: De l'intervalle de pari à l'intervalle de confiance

Monde réel

A- Modélisation

C- Traduction

Section 3: Modélisation Simulation Traduction

Monde artificiel

B- Simulation

Section 2: Propriétés des estimateurs

Modèle – énoncé

Soit X_S la variable statistique parente qui...

Soit X la variable aléatoire parente qui …

Construire une estimation ponctuelle de e = σ

Construire un **intervalle de confiance à 95%:**

-sur la proportion $p = \pi$

- sur la moyenne $m = \mu$

Modèle – solution

$\hat{\sigma} = 800{,}51$

$ic_{95\%}(\pi) = [0{,}10; 0{,}14]$

$ic_{95\%}(\mu) = [3443; 3557]$

I. Estimation: de l'intervalle de pari à l'intervalle de confiance

1-1. Estimateur et estimation ponctuelle

Définition 1

Soit Θ l'ensemble des valeurs possibles d'un paramètre inconnu θ également appelé la cible.

Un *estimateur* de θ est une statistique d'échantillon $T_n = f\left(X_1, X_2, ..., X_n\right)$ dont la réalisations appartiennent à Θ.

Définition 2

Soit $T_n = f\left(X_1, X_2, ..., X_n\right)$ un estimateur d'un paramètre inconnu θ. Une *estimation ponctuelle* de θ est la réalisation $t_n = f\left(x_1, x_2, ..., x_n\right)$ de T_n (obtenue après le tirage de l'échantillon). Cette estimation ponctuelle est notée θ.

1-2. De l'intervalle de pari à l'intervalle de confiance

Lorsque la taille de l'échantillon est supérieure à 30 (n > 30), les intervalles de pari bilatéraux de niveau (1 - α) sur la moyenne empirique sont de la forme:

$$p\left(\mu - z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}} \leq \bar{X}_n \leq \mu + z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Où $z_{(1-\frac{\alpha}{2})}$ est le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite et μ et σ sont l'espérance et l'écart type de la variable aléatoire parente.

1-3. De l'intervalle de pari à l'intervalle de confiance

L'intervalle $p\left(\mu - z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}} \le \bar{X}_n \le \mu + z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ s'écrit par soustraction de μ:

$$p\left(-z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}} \le \bar{X}_n - \mu \le +z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

D'où en soustrayant $\bar{X}_n$ dans chaque membre de chaque inégalité:

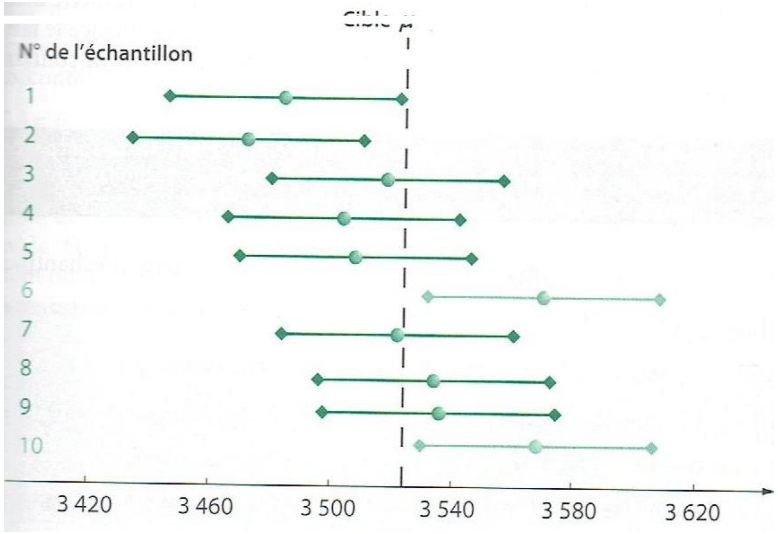
$$p\left(-\bar{X}_n - z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}} \le -\mu \le -\bar{X}_n + z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Soit:

$p\left(\bar{X}_n - z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + z_{(1-\frac{\alpha}{2})}.\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ cette expression s'appelle ***un intervalle de confiance sur μ de niveau (ou seuil) 1 - α.***

1-3. De l'intervalle de pari à l'intervalle de confiance

Huit réalisations de l'intervalle de confiance encadrant la cible



NB:

- Chaque borne de l'intervalle de confiance est une variable aléatoire. Ainsi, il ne s'agit pas d'un intervalle réel mais d'un intervalle aléatoire.
- Ce n'est pas μ qui se « déplace », mais d'un intervalle de confiance qui peut prendre une infinité de réalisations différentes.
- Le paramètre μ est inconnu mais fixe.

1-3. De l'intervalle de pari à l'intervalle de confiance

Définition:

Soit un paramètre inconnu θ.

1. Un *intervalle de confiance bilatéral sur θ de niveau ou de seuil (1 - α)* est un intervalle aléatoire [A; B] tel que $p[A \leq \theta \leq B] = 1-\alpha$ où A et B sont deux statistique d'échantillon.
2. Un *intervalle de confiance unilatéral sur θ de niveau ou de seuil (1 - α)* est un intervalle aléatoire [A; +∞[ou]-∞; A] tel que $p[A \leq \theta] = 1-\alpha$ ou $p[\theta \leq A] = 1-\alpha$
3. Notation des intervalles de confiance de niveau (1 - α) sur θ: $IC_{(1-\alpha)}(\theta)$.
4. Une réalisation d'un intervalle de confiance est un intervalle réel [a ; b] où $a \in A(\Omega) \quad et \quad b \in B(\Omega)$.
5. Cette réalisation est notée $ic_{(1-\alpha)}(\theta)$.

1-3. De l'intervalle de pari à l'intervalle de confiance

Définition:

Soit $T_n = f(X_1, X_2, ..., X_n)$ un estimateur d'un un paramètre inconnu θ.

Soit $IC_{(1-\alpha)}(\theta)$ un *intervalle de confiance bilatéral sur* θ *de niveau ou de seuil (1-*α*)*

1. L'écart type de T_n noté SE_n (= σ_{Tn}) est appelé *l'erreur standard*.
2. La demi-amplitude de l'intervalle ME_n est appelée la *marge d'échantillonnage*.

2-1. Biais et erreur quadratique moyenne

Pour que T_n soit un bon estimateur de θ, il faut qu'il y ait une probabilité élevée que T_n prenne une valeur proche de θ. Cela est réalisé, en particulier, lorsque T_n remplit les deux conditions suivantes:

- $E(T_n) = \theta$ (1)
- $\sigma_{T_n} = SE_n$ faible (2)

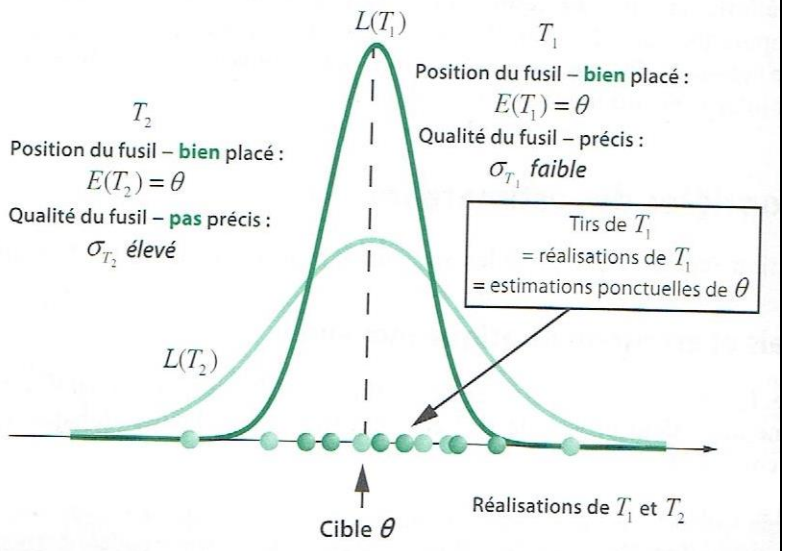
Définition

Soit $T_n = f(X_1, X_2, ..., X_n)$ un estimateur d'un paramètre inconnu θ.

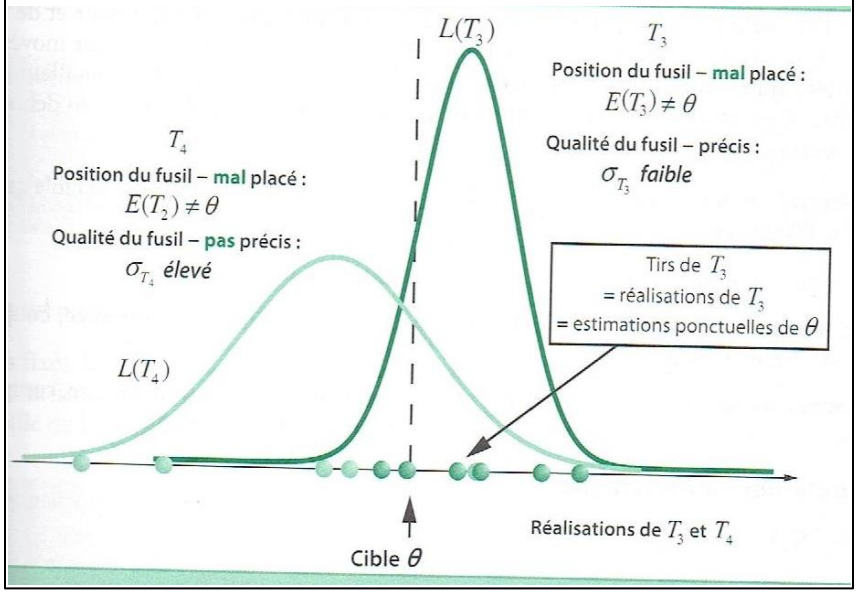
1. Le biais de T_n est définie par : $B(T_n) = E(T_n) - \theta$
2. T_n est un estimateur *non biaisé* de θ si $E(T_n) = \theta$, autrement dit si le biais est nul.
3. T_n est un estimateur *asymptomatique non biaisé* de θ si $\lim_{n \to \infty} E(T_n) = \theta$

2-1. Biais et erreur quadratique moyenne

Deux estimateurs non biaisés de la même cible, l'un précis, l'autre non



Deux estimateurs biaisés de la même cible, l'un précis, l'autre non



2-1. Biais et erreur quadratique moyenne

Définition et proposition

Soit $T_n = f\left(X_1, X_2, ..., X_n\right)$ un estimateur d'un paramètre inconnu θ.

L' ***erreur quadratique moyenne*** ou ***fonction de risque*** est définie par:

$$R_{T_n}\left(\theta\right) = E\left[\left(T_n - \theta\right)^2\right]$$

Cette erreur vaut : $R_{T_n}\left(\theta\right) = Var\left(T_n\right) + \left[B\left(T_n\right)\right]^2$ où $B\left(T_n\right)$ le biais de T_n .

Démonstration:

En développant la formule de définition de l'erreur quadratique moyenne:

$$R_{T_n}\left(\theta\right) = E\left[\left(T_n - \theta\right)^2\right] = E\left[T_n^2 - 2\theta T_n + \theta^2\right] = E\left[T_n^2\right] - 2\theta E\left(T_n\right) + \theta^2$$

2-1. Biais et erreur quadratique moyenne

En ajoutant puis en soustrayant $\left[E\left(T_n\right)\right]^2$ à la relation ci-dessus, on obtient:

$$R_{T_n}\left(\theta\right)=\left[E\left[T_n^2\right]-\left[E\left(T_n\right)\right]^2\right]+\left[E\left(T_n\right)\right]^2-2\theta E\left(T_n\right)+\theta^2$$

$$R_{T_n}\left(\theta\right)=Var\left(T_n\right)+\left[E\left(T_n\right)-\theta\right]^2=Var\left(T_n\right)+\left[B\left(T_n\right)\right]^2$$

Ainsi, l'erreur quadratique moyenne dépend à la fois du biais de l'estimateur et de son écart type. Un estimateur de θ est d'autant meilleur que son erreur quadratique moyenne est faible. Si l'estimateur est non biaisé, il est clair qu'il est alors d'autant meilleur que son écart type est faible. Ceci traduit les deux conditions (1) et (2) évoquées en début de sous-section.

D'où, un bon tireur, avec un bon fusil, a plus de chances d'atteindre la cible qu'un mauvais tireur avec un mauvais fusil.

2-1. Biais et erreur quadratique moyenne

Cas de la moyenne empirique

On sait que: $E\left(\bar{X}_n\right) = \mu$

De plus, $Var\left(\bar{X}_n\right) = \frac{\sigma^2}{n}$ tend vers 0 quand la taille de l'échantillon augmente, et ceci indépendamment de la valeur de μ.

Or, $E\left[\left(\bar{X}_n - \mu\right)^2\right] = Var\left(\bar{X}_n\right) + \left[B\left(\bar{X}_n\right)\right]^2 = \frac{\sigma^2}{n} + 0$, d'où la proposition:

Proposition:

1. $\bar{X}_n$ est un estimateur non biaisé de l'espérance μ de la variable aléatoire parente.
2. L'erreur quadratique moyenne de $\bar{X}_n$ tend vers 0 quant la taille de l'échantillon tend vers l'infini.

2-1. Biais et erreur quadratique moyenne

Cas de la variance empirique et la variance empirique corrigée

On a: $E\left(S_n^2\right)=\frac{n-1}{n}\sigma^2$ (chapitre précédent). Si l'on choisit S_n^2 comme estimateur de la variance de la variable aléatoire parente σ^2, alors:

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

Cette relation signifie que la variance empirique est un estimateur biaisé de σ^2. Le biais vaut $-\frac{\sigma^2}{n}$, ce qui veut dire que la variance empirique vise systématiquement légèrement à gauche de σ^2. quand la taille de l'échantillon augmente, ce biais tend vers 0.

2-1. Biais et erreur quadratique moyenne

Proposition

S_n^2 est un estimateur asymptomatique non biaisé de σ^2.

Par ailleurs, le biais de S_n^2 étant connu, un estimateur de σ^2 corrigé de ce biais peut être construit. En effet, en posant $S_{Cn}^2 = \frac{n}{n-1} S_n^2$, on obtient $E\left(S_{Cn}^2\right) = \frac{n}{n-1} E\left(S_n^2\right) = \sigma^2$

Définition:

1. La *variance empirique corrigée* est la variable aléatoire $S_{Cn}^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$

2.L'écart type empirique corrigé S_{Cn} est la racine carrée de la variance empirique corrigée.

2-1. Biais et erreur quadratique moyenne

Proposition:

1.La variance empirique corrigée S_{Cn}^2 est un estimateur sans biais de la variance σ^2 de la variable aléatoire parente.

2.L'erreur quadratique moyenne de S_{Cn}^2 tend vers 0 quand n tend vers l'infini.

3. $Var\left(S_{Cn}^2\right) \cong \frac{1}{n}\left[\mu_{C4} - \frac{n-3}{n-1}\sigma^4\right]$ où μ_{C4} est le moment d'ordre 4.

La variance empirique corrigée S_{Cn}^2 est un estimateur sans biais de la variance σ^2, mais S_{Cn} est un estimateur biaisé de σ. D'où: $E\left(\sqrt{S_{Cn}^2}\right) \neq \sqrt{E\left(S_{Cn}^2\right)}$.

2-2. Estimateur convergent

Jusqu'à maintenant, nous nous sommes intéressés à la précision d'un estimateur en cherchant des estimateurs sans biais et avec une faible variance. L'idée était de minimiser l'erreur quadratique moyenne. Ce faisant, nous avons cherché, sans l'exprimer directement, à construire des estimateurs convergent, c'est-à-dire tel que:

$$\lim_{n\to\infty} T_n = \theta$$

La notion de convergence qui est « multiforme » mérite d'être précisée.

Le schéma suivant présente la définition de quatre « type » de convergence.

.

2-2. Estimateur convergent

Convergence presque sûre $T_n \xrightarrow{ps} X$

Si l'ensemble $\left\{\omega / \lim_{n\to\infty} T(\omega) \neq X(\omega)\right\}$

a une probabilité nulle

Convergence en moyenne d'ordre k

$T_n \xrightarrow{\text{moyenne d'ordre k}} X$

Si: $\lim_{n\to\infty} E\left[\left|T_n - X\right|^k\right] = 0$

Convergence en probabilité: $T_n \xrightarrow{p} X$

La suite (T_n) converge en probabilité vers la variable X si la probabilité que l'écart entre T_n et X soit inférieur à une valeur η aussi petite que l'on veut tend vers 1: $\lim_{n\to\infty} p\left(\left|T_n - X\right| < \eta\right) = 1$

Convergence en loi $T_n \xrightarrow{L} X$

La suite (T_n) converge en loi vers la variable X si la suite F_{T_n} des fonctions de répartition des Ti tend vers la fonction de répartition F_x de X: $\lim_{n\to\infty} F_{T_n}(x) = F_X(x)$ en tout x où F_x est continue.

2-2. Estimateur convergent

Définition

T_n est estimateur ***convergent*** (ou consistant) de θ si T_n converge en probabilité vers θ.

Si un estimateur est asymptomatique sans biais et si sa variance tend vers zéro quand la taille de l'échantillon tend vers l'infini, alors son erreur quadratique moyenne tend également vers zéro. En d'autres termes, ces deux conditions sont suffisantes pour que l'estimateur soit convergent.

Proposition

1. $\bar{X}_n$ est un estimateur convergent de l'espérance μ de la variable aléatoire parente.
2. S_n^2 est un estimateur convergent da la variance σ^2 de la variable aléatoire parente.
3. S_{Cn}^2 est un estimateur convergent de la variance σ^2 de la variable aléatoire parente.

3-1. Estimateur du maximum de vraisemblance

Plusieurs méthodes d'estimation existe. Les plus courantes sont:

- Maximum de vraisemblance;
- Moments

Cas d'une variable aléatoire parente discrète

Supposant que la loi de la variable aléatoire parente *X* est connue et qu'elle dépend du paramètre θ. La probabilité que *X* prenne la réalisation x est $p(x,\theta)$, *où* p est une application connue.

L'expérience qui permet de construire l'échantillon est effectuée et la réalisation: $n-eas = (x_1, x_2,\ldots x_n)$ de l'échantillon $(X_1, X_2,\ldots,X_n)$ est observée.

Comme $(X_1, X_2,\ldots,X_n)$ est un $n-EAS = (x_1, x_2,\ldots x_n)$, les variables aléatoires Xi sont indépendantes: $p(n-eas) = p(x_1,\theta).p(x_2,\theta)....p(x_n,\theta) = \prod_{i=1}^{n} p(x_i,\theta) = L(x_1, x_2,..., x_n)$

3-1. Estimateur du maximum de vraisemblance

Cas d'une variable aléatoire parente discrète

La fonction $L(x_1, x_2, \ldots x_n, \theta)$ également notée $L(n-eas, \theta)$ s'appelle la ***vraisemblance***.

Exemple 1: EMV de l'espérance d'une variable de Bernoulli

Soit X une variable aléatoire parente qui suit une loi de Bernoulli de paramètre θ.

Donc:

- $L(X) = \begin{pmatrix} 0 & 1 \\ 1-\theta & \theta \end{pmatrix}$
- D'où, $p(x_i, \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$ pour $x_i = 0$ ou $x_i = 1$.

$$L(n-eas, \theta) = \theta^{x_1}(1-\theta)^{1-x_1}.\theta^{x_2}(1-\theta)^{1-x_2}.....\theta^{x_n}(1-\theta)^{1-x_n}$$

$$Or, \quad \theta^{x_1}.\theta^{x_2}....\theta^{x_n} = \theta^{x_1+x_2+\ldots+x_n} = \theta^{n\bar{x}}$$

$$et \quad (1-\theta)^{1-x_1}.(1-\theta)^{1-x_2}.....(1-\theta)^{1-x_n} = (1-\theta)^{(1+1+\ldots+1)-(x_1+x_2+\ldots+x_n)} = (1-\theta)^{n(1-\bar{x})}$$

3-1. Estimateur du maximum de vraisemblance

Donc: $L(n-eas,\theta)=\theta^{x_1+x_2+...+x_n}.(1-\theta)^{(1+1+...+1)-(x_1+x_2+...+x_n)}=\theta^{n\bar{x}}.(1-\theta)^{n(1-\bar{x})}$

Recherche du maximum de vraisemblance:

Il faut chercher la valeur de θ qui rend la ***vraisemblance*** maximale. Le logarithme étant une fonction croissante, il est ici plus rapide de chercher à maximiser le logarithme de la ***vraisemblance***.

Lorsque le paramètre inconnu θ est strictement compris entre 0 et 1, $L(n-eas,\theta)$ est strictement positive. L'examen des valeurs 0 et 1 est intéressant.

• Si θ est nul, la réalisation $\bar{x}$ de la moyenne empirique est égale à 0 puisque chaque X_i ne peut prendre que la valeur 0. Il en résulte que: $L(n-eas,0) = 0^0.1^n = 1$

• Si θ = 1, la réalisation $\bar{x}$ de la moyenne empirique est forcément égale à 1. ce ci conduit à $L(n-eas,1) = 1^n.\ 0^0 = 1$.

3-1. Estimateur du maximum de vraisemblance

Pour la suite des calculs, les hypothèses sont donc: $\theta \neq 0$ et $\theta \neq 1$.

Calcul de la dérivée du logarithme de la vraisemblance ($\theta \neq 0$ et $\theta \neq 1$)

On utilisant la propriété caractéristique des logarithmes $lna^x = xlna$, on obtient:

$$\ln\left[L(x_1, x_2, ..., x_n, \theta)\right] = n\bar{x}\ln\theta + n\left(1-\bar{x}\right)\ln\left(1-\theta\right)$$

$$\frac{\partial \ln\left[L(x_1, x_2, ..., x_n, \theta)\right]}{\partial\theta} = n\bar{x}\frac{d\ln\theta}{d\theta} + n\left(1-\bar{x}\right)\frac{d\ln\left(1-\theta\right)}{d\theta}$$

D'une part, $\frac{d\ln\theta}{d\theta} = \frac{1}{\theta}$ et d'autre part $\frac{d\ln\left(1-\theta\right)}{d\theta} = \frac{u'}{u} = \frac{-1}{1-\theta}$

D'où, $\frac{\partial \ln\left[L(x_1, x_2, ..., x_n, \theta)\right]}{\partial\theta} = n\bar{x}\frac{1}{\theta} + n\left(1-\bar{x}\right)\frac{(-1)}{1-\theta}$

En mettant les deux fractions au même dénominateur, on obtient:

3-1. Estimateur du maximum de vraisemblance

$$\frac{\partial \ln\left[L(x_1, x_2, ..., x_n, \theta)\right]}{\partial \theta} = \frac{n\bar{x}(1-\theta) - n\left(1-\bar{x}\right)\theta}{\theta(1-\theta)} = \frac{n\bar{x} - n\bar{x}\theta - n\theta + n\bar{x}\theta}{\theta(1-\theta)} = \frac{n\left(\bar{x}-\theta\right)}{\theta(1-\theta)}$$

La ***vraisemblance*** admet donc un point critique en $\theta = \bar{x}$ qui annule la dérivée de son logarithme.

Calcul de la dérivée seconde du logarithme de la vraisemblance $(\theta \neq 0)$

$$\frac{\partial\left[n\bar{x}\frac{1}{\theta} + n\left(1-\bar{x}\right)\frac{(-1)}{1-\theta}\right]}{\partial\theta} = n\bar{x}\frac{d\left[\frac{1}{\theta}\right]}{d\theta} - n\left(1-\bar{x}\right)\frac{d\left[\frac{1}{1-\theta}\right]}{d\theta} = n\bar{x}\frac{(-1)}{\theta^2} - n\left(1-\bar{x}\right)\frac{1}{(1-\theta)^2}$$

En mettant les deux fraction au même dénominateur, la dérivée seconde s'écrit:

$$\frac{\partial^2\left[\ln(L)\right]}{\partial^2\theta} = \frac{-n\bar{x}(1-\theta)^2 - n\left(1-\bar{x}\right)\theta^2}{\theta^2(1-\theta)^2} = \frac{-n\bar{x}\left(1-2\theta+\theta^2\right) - n\theta^2 + n\bar{x}\theta^2}{\theta^2(1-\theta)^2}$$

3-1. Estimateur du maximum de vraisemblance

Soit, $$\frac{\partial^2\left[\ln(L)\right]}{\partial^2\theta}=\frac{-n\left(\theta^2-2\theta\bar{x}+\bar{x}\right)}{\theta^2(1-\theta)^2}$$

Le signe de la dérivée seconde du logarithme de la fonction de ***vraisemblance*** et celui de $-n\left(\theta^2-2\theta\bar{x}+\bar{x}\right)$. Le discriminant de ce polynôme du second degré en θ vaux $\Delta=4\bar{x}.\left(\bar{x}-1\right)$ qui est toujours négatif, car $0<\bar{x}<1$. Le polynôme $\left(\theta^2-2\theta\bar{x}+\bar{x}\right)$ est donc de signe positif constant et la dérivée seconde de la fonction de vraisemblance est toujours négative.

La valeur de θ qui maximise la vraisemblance est bien $\theta=\bar{x}$.

L'estimateur du maximum de vraisemblance de θ et donc $\bar{X}_n$

3-1. Estimateur du maximum de vraisemblance

Proposition

L'EMV du paramètre π d'une variable de Bernoulli est la fréquence empirique $F_n = \bar{X}$

Cas d'une variable aléatoire parente continue

En suppose que pour θ fixé, la densité de probabilité $f_X(x,\theta)$ de la variable aléatoire parente X est inconnue.

La réalisation de $n-eas = (x_1, x_2,\ldots x_n)$ a cette fois une probabilité nulle puisque toutes les probabilités ponctuelles sont nulle pour une variable aléatoire continue. Le raisonnement porte alors sur l'intervalle des réalisations:

$$[n-eas] = \left[(x_1 \le X_1 \le x_1 + h);(x_2 \le X_2 \le x_2 + h);\ldots;(x_n \le X_n \le x_n + h)\right]$$

Si h est très petit, $p(x_i \le X_i \le x_i + h) \cong h f_X(x_i,\theta)$, et comme $(X_1, X_2,\ldots,X_n)$ est $n-EAS$, alors:

3-1. Estimateur du maximum de vraisemblance

$$p\left[n-eas\right] \cong h^n f_X\left(x_1,\theta\right).f_X\left(x_2,\theta\right).....f_X\left(x_n,\theta\right)$$

On pose $L\left(n-eas,\theta\right)= f_X\left(x_1,\theta\right).f_X\left(x_2,\theta\right).....f_X\left(x_n,\theta\right)$. La fonction $L\left(n-eas,\theta\right)$ est appelée ***vraisemblance***.

On a $p\left[n-eas\right] \cong h^n L\left(n-eas,\theta\right)$

Si l'événement $[n - eas]$ s'est produit, c'est qu'il est vraisemblable que sa probabilité était importante. En d'autres termes, la valeur θ qui rend la ***vraisemblance*** maximale est à nouveau cherchée.

Exemple 2: EMV de l'espérance d'une variable gaussienne

Soit X une variable aléatoire parente qui suit une loi de Gauss d'espérance inconnue θ:

$X \sim N(\theta;\sigma)$.

Soit $f_X\left(x,\theta\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left[\frac{x-\theta}{\sigma}\right]^2\right]$ (fonction de densité de probabilité de la loi de Laplace – Gauss).

3-1. Estimateur du maximum de vraisemblance

La vraisemblance est donc: $L(x_1, x_2, ..., x_n, \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right]$

Cette fonction est maximale quand $\sum_{i=1}^{n}(x_i-\theta)^2$ est minimale.

Or, $\frac{1}{n}\sum_{i=1}^{n}(x_i-\theta)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \left(\frac{1}{n}\sum_{i=1}^{n}x_i-\theta\right)^2 = S_n^2 + \left(\frac{1}{n}\sum_{i=1}^{n}x_i-\theta\right)^2$

Donc $\sum_{i=1}^{n}(x_i-\theta)^2$ est minimale pour $\theta = \frac{1}{n}\sum_{i=1}^{n}x_i$

L'estimateur du maximum de vraisemblance de θ est finalement $\bar{X}_n$ (que l'écart type σ soit connu ou inconnu, puisqu'il n'intervient pas dans la démonstration).

3-1. Estimateur du maximum de vraisemblance

Définition

Soit $(X_1, X_2,\ldots,X_n)$ un $n - EAS$ issu d'une variable aléatoire parente X de loi connue.

1.Si la variable aléatoire X est discrète de loi connue $p[X = x_i] = p(x_i,\theta)$, alors la fonction de vraisemblance de X est définie par:

$$L(x_1, x_2,..., x_n,\theta) = L(n-eas,\theta) = p(x_1,\theta).p(x_2,\theta).....p(x_n,\theta)$$

2.Si la variable aléatoire X est continue de densité de probabilité connue $f_X(x_n,\theta)$, alors: $L(x_1, x_2,..., x_n,\theta) = L(n-eas,\theta) = f_X(x_1,\theta).f_X(x_2,\theta).....f_X(x_n,\theta)$

3.Soit $\theta = g(x_1, x_2,..., x_n)$, la valeur, si elle existe, qui maximise la fonction de vraisemblance. θ Est l'estimation du maximum de vraisemblance de θ et la valeur aléatoire correspondante $T_n = g(x_1, x_2,..., x_n)$ est l'EMV de θ .

3-1. Estimateur du maximum de vraisemblance

Proposition

Si la variable aléatoire parente X suit une loi de Gauss d'espérance μ et d'écart type σ, alors:

1. L'EMV de l'espérance μ est la moyenne empirique $\overline{X}_n$, que l'écart type soit connu ou non.
2. L'EMV de la variance σ^2 est $S_n^{*2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ si l'espérance μ est connue.
3. L'EMV de la variance σ^2 est $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$ si l'espérance μ est inconnue.

IV. Intervalles de confiance sur une moyenne

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

L'hypothèse de travail est qu'une étude de grande envergure à montré que l'écart type de la variable statistique parente est de 9 km/h. en d'autres termes, $e = 9$.

Il faut noter que la construction d'un intervalle de confiance est très proche de celle d'un intervalle de pari.

Modélisation

1. Population mère

P est la population constituée de tous les véhicules susceptibles de passer devant les points de contrôle. N = 39026000 véhicules immatriculées.

2. Variable statistique parente (VSP)

Variable statistique parente qui associe à chaque véhicule de P sa vitesse au point de contrôle.

La vitesse moyenne m est inconnue.

L'écart type e est connu: $e = 9$.

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

3. Variable aléatoire parente (VAP)

soit ε l'expérience aléatoire parente qui consiste à choisir au hasard un véhicule dans P (véhicule qui se présentent au hasard).

X, la variable aléatoire parente, associe au véhicule tiré au hasard sa vitesse.

L'espérance μ de X est inconnue et l'écart type de X est connu ($\sigma = e = 9$).

4. Echantillon aléatoire simple (EAS)

l'expérience ε est répétée 50 fois.

X_i (i *varie de 1 à 50*) est la variable aléatoire qui associe au $i^{ème}$ véhicule sa vitesse.

• Par construction, les 50 variables aléatoire X_i suivent la même loi que X.

• L'expérience est répétée sans remise, mais le taux de sondage $\frac{50}{39026000}$ est extrêmement faible. Les variables aléatoire qui constituent l'échantillon peuvent être supposées indépendantes . Donc $\left(X_1, X_2, ..., X_{50}\right)$ est un $50 - EAS$.

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

5. Du problème concret au modèle – énoncé

Il s'agit de construire une fourchette de valeurs ayant une probabilité de 95% de contenir la vitesse moyenne m dans la population mère de tous les véhicules. Comme $m = \mu$, le problème revient à construire un intervalle de confiance de niveau 95% sur μ avec $\sigma = 9$.

Simulation

1. Statistique d'échantillon et loi de probabilité

•*Choix de l'estimateur et de la statistique d'échantillon*

Il n'y a qu'un seul paramètre inconnue qui est l'espérance μ de la variable parente.

$\bar{X}_{50}$ est un estimateur sans biais et convergent de μ . La statistique d'échantillon $T_n = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ est pertinente. En effet la valeur de σ est connue, et il est donc possible avec cette statistique d'aboutir à une réalisation de l'intervalle.

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

•*Loi de probabilité de* T_n

Échantillon gaussien de taille quelconque

La variable aléatoire parente est supposée gaussienne, donc la moyenne empirique $\bar{X}_n$ suit une loi normale d'espérance μ et d'écart type $\sigma/\sqrt{n}$. Par conséquent:

$$T_{50} = \frac{\bar{X}_{50} - \mu}{9/\sqrt{50}} \rightarrow N[0;1]$$

Échantillon non gaussien mais de grande taille

Dans ce cas, le TCL permet d'affirmer que $T_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{L} N[0;1]$

Et donc $T_n = \frac{\bar{X}_{50} - \mu}{9/\sqrt{50}} \approx N[0;1]$

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

2. Construction de l'intervalle de confiance

Cet intervalle n'est pas unique. En général, le choix consiste à construire un intervalle « symétrique » dans le sens où on laisse 2,5% à droite et 2,5% à gauche de l'intervalle. En effet, compte tenu de la symétrie de la densité gaussienne standard et du fait que celle-ci présente une densité maximale autour de 0, ce choix conduit à l'intervalle le plus petit possible.

Or, pour une loi gaussienne standard, $\boldsymbol{p[Z \leq 1,96] = 97,5\%}$.

On en déduit que: $$p\left[-1,96 \leq \frac{\bar{X}_{50} - \mu}{9/\sqrt{50}} \leq 1,96\right] = 95\%$$

En multipliant les trois membres par $9/\sqrt{50}$, on obtient:

$$p\left[-1,96 \times 9/\sqrt{50} \leq \bar{X}_{50} - \mu \leq 1,96 \times 9/\sqrt{50}\right] = 95\%$$

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

2. Construction de l'intervalle de confiance

Soit en soustrayant $\bar{X}_{50}$ à chaque membre:

$$p\left[-\bar{X}_{50}-1,96\times 9/\sqrt{50}\le -\mu\le -\bar{X}_{50}+1,96\times 9/\sqrt{50}\right]=95\%$$

En fin en multipliant chaque membre par(-1) et en changeant le sens des inégalités, on obtient:

$$p\left[\bar{X}_{50}-1,96\times 9/\sqrt{50}\le \mu\le \bar{X}_{50}+1,96\times 9/\sqrt{50}\right]=95\%$$

Résultat qui peut s'écrire sous la forme:

$$IC_{95\%}(\mu)=\left[\bar{X}_{50}\pm 1,96\times 9/\sqrt{50}\right]$$

Donc pour un intervalle de confiance de niveau ($1-\alpha$):

$$IC_{95\%}(\mu)=\left[\bar{X}_{n}\pm z_{(1-\alpha/2)}\,\sigma/\sqrt{n}\right]$$

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

3. Réalisation de l'intervalle de confiance

Soit le tableau suivant qui représente cinquante relevés de vitesse en km/h:

124	105	94	93	106	101	90	91	118	115
100	96	112	109	89	102	95	89	88	108
101	96	105	107	100	101	91	114	98	101
103	92	103	94	95	115	97	105	99	103
99	103	99	105	122	114	82	100	119	99

D'après les données du tableau ci-dessus, $\bar{x} = 101,74$ km/h, donc $ic_{95\%}(\mu) = \left[101,74 \pm 1,96 \times 9/\sqrt{50}\right]$

4-1. variance connue, échantillon gaussien ou non gaussien mais de grande taille

3. Réalisation de l'intervalle de confiance

D'après les données du tableau ci-dessus, $\bar{x} = 101,74$ km/h, donc:

$$ic_{95\%}(\mu) = \left[101,74 \pm 1,96 \times 9/\sqrt{50}\right]$$

Soit en arrondissant la borne de gauche par valeur inférieur et la borne de droite par valeur supérieur pour agrandir l'intervalle:

$$ic_{95\%}(\mu) = [99;105]$$

Traduction

Il y a une probabilité de 95% pour que l'intervalle de 99 km/h à 105 km/h contienne la vitesse moyenne de tous les véhicules .

4-2. Variance inconnue, échantillon gaussien ou non gaussien mais de grande taille

Reprenons les mêmes données de l'exemple précédent.

Modélisation

Même procédure à l'exception du fait que l'écart type est inconnu.

Simulation

1. Statistique d'échantillon et loi de probabilité

Choix des estimateur et de la statistique d'échantillon

$\bar{X}_{50}$ est un estimateur sans biais et convergent de l'espérance μ. L'écart type étant inconnu, il convient d'en faire une estimation ponctuelle par l'écart type empirique corrigé. La statistique :

$$T_n = \frac{\bar{X}_n - \mu}{S_{Cn}/\sqrt{n}}$$ est la statistique qui convient.

4-2. Variance inconnue, échantillon gaussien ou non gaussien mais de grande taille

1. Statistique d'échantillon et loi de probabilité

Loi de probabilité T_n

Échantillon gaussien de taille quelconque

La variable aléatoire parente X suit une loi de Gauss, ce qui permet de démontrer que la statistique T_n suit une ***loi de Student*** à $n-1$ degrés de liberté. L'idée de la démonstration est d'introduire une variable aléatoire Z qui suit une loi normale centrée réduite et une variable aléatoire K qui suit une ***loi de khi-deux*** à $n-1$ degrés de liberté.

$$T_n = \frac{\bar{X}_n - \mu}{S_{Cn}/\sqrt{n}} = \frac{\dfrac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\dfrac{S_{Cn}/\sqrt{n}}{\sigma/\sqrt{n}}} = \frac{Z}{\dfrac{S_{Cn}}{\sigma}} \quad \text{avec} \quad Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

4-2. Variance inconnue, échantillon gaussien ou non gaussien mais de grande taille

1. Statistique d'échantillon et loi de probabilité

Or, la variable aléatoire parente X est gaussienne, donc la moyenne empirique T_n Suit une loi normale d'espérance μ et d'écart type $\sigma/\sqrt{n}$. En conséquence, $Z \to N(0;1)$.

Par ailleurs: $\frac{S_{Cn}}{\sigma} = \sqrt{\frac{S_{Cn}^2}{\sigma^2}} = \sqrt{\frac{(n-1)S_{Cn}^2}{\sigma^2} \times \frac{1}{n-1}} = \sqrt{\frac{K}{n-1}}$ en posant $K = \frac{(n-1)S_{Cn}^2}{\sigma^2}$

Or, $K = \frac{(n-1)S_{Cn}^2}{\sigma^2} = \frac{nS_{Cn}^2}{\sigma^2}$ et $\frac{nS_{Cn}^2}{\sigma^2} \to \chi^2_{n-1}$

Donc, $T_n = \frac{\bar{X}_n - \mu}{S_{Cn}/\sqrt{n}} = \frac{Z}{\sqrt{\frac{K}{n-1}}}$ avec $Z \to N(0;1)$ et $K \to \chi^2_{n-1}$

4-2. Variance inconnue, échantillon gaussien ou non gaussien mais de grande taille

1. Statistique d'échantillon et loi de probabilité

En effet Z et K sont indépendantes car lorsqu'une variable aléatoire est gaussienne, alors sa moyenne empirique est indépendante de sa variance empirique.

On peut ainsi utiliser le fait que: $$T_n = \frac{\bar{X}_n - \mu}{S_{Cn}/\sqrt{n}} \rightarrow St_{n-1}$$

Ici: $$T_{50} = \frac{\bar{X}_{50} - \mu}{S_{C50}/\sqrt{50}} \approx St_{49}$$

Échantillon non gaussien de grande taille

Le TCL a pour conséquence que: $$T_n = \frac{\bar{X}_n - \mu}{S_{Cn}/\sqrt{n}} \approx St_{n-1}$$ (pour n assez grand)

Soit ici: $$T_{50} = \frac{\bar{X}_{50} - \mu}{S_{C50}/\sqrt{50}} \approx St_{49}$$

4-2. Variance inconnue, échantillon gaussien ou non gaussien mais de grande taille

2. Construction de l'intervalle de confiance

Le choix est fait de construire un intervalle « symétrique » dans le sens où on laisse 2,5% à droite et 2,5% à gauche de l'intervalle. Comme la loi de Student est symétrique par rapport à l'axe des ordonnées, on a: $p\left[-t_{0,975} \le \frac{\bar{X}_{50} - \mu}{S_{C50}/\sqrt{50}} \le t_{0,975}\right] = 95\%$ où $t_{0,975}$ est le quantile d'ordre 97,5%, de la loi de Student à 49 degrés de liberté.

On multiplie les trois membres par $S_{C50}/\sqrt{50}$, puis on soustrait $\bar{X}_{50}$ à chaque membre et, en fin, on multiplie chaque membre par ***(-1)*** en changeant le sens des inégalités.

$$p\left[\bar{X}_{50} - t_{0,975} \times S_{C50}/\sqrt{50} \le \mu \le \bar{X}_{50} + t_{0,975} \times S_{C50}/\sqrt{50}\right] = 95\%$$

Résultat qui s'écrit sous la forme: $IC_{95\%}(\mu) = \left[\bar{X}_{50} \pm t_{0,975} \times S_{C50}/\sqrt{50}\right]$

4-2. Variance inconnue, échantillon gaussien ou non gaussien mais de grande taille

2. Construction de l'intervalle de confiance

L'intervalle qui vient d'être déterminé avec $\alpha = 0,05\ (5\%)$ pour un niveau de confiance ($1 - \alpha$):

$$IC_{95\%}(\mu) = \left[\bar{X}_n \pm t_{(1-\alpha/2)} S_{Cn} / \sqrt{n} \right]$$

Il faut donc déterminer le quantile d'ordre 97,5%, de la loi de Student à 49 degrés de liberté.

Détermination de la valeur de $t_{0,975}$, deux solutions se présentent:

•Soit utiliser Excel: $t_{0,975}$ = **LOI.STUDENT.INVERSE.N(0,975;49) = 2,0096**

•Soit prendre le premier ***ddl*** inférieur à ***49*** présent dans le table de manière à maximiser l'intervalle. Selon la table, pour ***40 ddl***:

$t_{0,975}$ = ***2,021.*** L'intervalle est bien maximiser avec cette méthode.

4-2. Variance inconnue, échantillon gaussien ou non gaussien mais de grande taille

3. Réalisation de l'intervalle de confiance

$$ic_{95\%}(\mu) = \left[\bar{x}_{50} \pm t_{0,975} \times S_{C50} / \sqrt{50}\right]$$

D'après les données du tableau de la section précédente: $\bar{x} = 101,74$ et $s_{C50} = 9,2688$

D'où:

$$ic_{95\%}(\mu) = \left[101,74 \pm 2,0096 \times 9,2688 / \sqrt{50}\right]$$

$$ic_{95\%}(\mu) = [99;105]$$

Traduction:

Compte tenu des arrondis sur les bornes finales de l'intervalle, la conclusion est la même que la section (4.1): il y a une probabilité de 95% que l'intervalle de 99 km/h à 105 km/h contienne la vitesse moyenne de tous les véhicules.

5-1. Intervalles de confiance sur la variance

Contenance de 100 bouteilles d'une chaîne A en centilitres

75,1	74,9	75,1	74,8	75,0	75,2	75,2	75,0	74,7	74,6
75,2	75,1	75,5	75,1	74,8	75,0	75,1	75,2	74,7	74,8
75,1	75,1	75,0	75,1	74,9	75,0	75,2	74,9	74,8	75,4
75,2	75,0	74,9	75,1	74,6	75,3	75,1	75,1	74,9	74,9
74,8	74,9	74,9	75,2	75,4	75,2	75,0	75,0	75,3	75,0
74,7	74,9	74,9	74,9	75,3	74,6	74,7	75,3	75,6	74,9
74,7	74,8	74,9	74,7	74,7	74,9	75,4	74,9	75,3	75,2
74,6	74,8	74,8	75,0	75,1	74,8	74,6	75,2	75,2	75,3
75,2	74,8	75,2	75,1	75,1	74,8	75,0	75,2	75,2	75,0
74,8	74,9	74,7	74,9	75,1	75,0	75,1	75,1	75,1	75,0

5-1. Intervalles de confiance sur la variance

Modélisation

1. Population mère

P_A toutes les bouteilles de boisson remplies par la chaîne A. sa taille N_A et de plusieurs milliers par jours.

2. Variable statistique parente

X_{SA} est la VSP qui associe à chaque bouteille de P_A sa contenance. La contenance moyenne m_A est inconnue. L'écart type e_A est également inconnu.

3. Variable aléatoire parente

ε_A est l'expérience aléatoire qui consiste à choisir une bouteille dans la production totale P_A.

La VAP X_A associe à la bouteille choisie sa contenance. μ_A et σ_A de X_A sont inconnus.

X_A est gaussienne.

5-1. Intervalles de confiance sur la variance

4. Échantillon aléatoire simple

L'expérience ε_A est répétée 100 fois.

X_{Ai} (i *varie de 1 à 100*) est la variable aléatoire qui associe à la $i^{ème}$ bouteille sa contenance.

• Par construction, les 100 variables aléatoire X_{Ai} suivent la même loi que X_A.

• L'expérience est répétée sans remise, mais le taux de sondage (100 sur plusieurs milliers) est très inférieur à 10%. Les variables aléatoire qui constituent l'échantillon peuvent être supposées indépendantes . Donc (X_{A1}, X_{A2},…,X_{A100}) est un $100 - EAS$ gaussien.

5. Du problème concret au modèle – énoncé

Il s'agit d'estimer au seuil de 95% l'écart type de e_A de la production totale. Le problème revient à construire un intervalle de confiance au seuil de 95% sur σ_A.

5-1. Intervalles de confiance sur la variance

Simulation

1. Statistique d'échantillon et loi de probabilité

- *Choix des estimateurs et de la statistique d'échantillon*
- L'échantillon est de grande taille et est gaussien,
- D'espérance μ_A inconnu.

On choisit comme estimateur de σ^2_A la variance empirique corrigée S^2_{Cn} .

Sachant que $\frac{nS^2_{Cn}}{\sigma^2} \rightarrow \chi^2_{n-1}$ (Chapitre I)

Par définition: $nS_n^2 = (n-1)S_{Cn}^2$. On en déduit que: $\frac{nS_n^2}{\sigma^2} = \frac{(n-1)S_{Cn}^2}{\sigma^2}$

Statistique d'échantillon pour la variance, échantillon gaussien: $T_n = \frac{(n-1)S^2_{Cn}}{\sigma^2} \rightarrow \chi^2_{n-1}$

5-1. Intervalles de confiance sur la variance

- Loi de probabilité

Ici : $$T_{100} = \frac{99S^2_{C100}}{\sigma^2_A} \rightarrow \chi^2_{n-1}$$

2. Construction de l'intervalle de confiance

Pour ce fait, on cherche k_1 et k_2 telles que:

$$p\left[k_1 \leq \frac{99S^2_{CA100}}{\sigma^2_A} \leq k_2\right] = 95\% \Leftrightarrow p\left[\frac{99S^2_{C100}}{k_2} \leq \sigma^2_A \leq \frac{99S^2_{C100}}{k_1}\right] = 95\% \quad \text{soit,} \quad IC_{95\%}\left(\sigma^2_A\right) = \left[\frac{99S^2_{C100}}{k_2}; \frac{99S^2_{C100}}{k_1}\right]$$

3. Réalisation de l'intervalle de confiance

les couple de valeurs (k_1,k_2) ne sont pas uniques. Une solution courante consiste à retenir sous la courbe de densité de la loi khi-deux une aire de 2,5% à gauche de k_1 et une aire de 2,5% à droite de k_2.

5-1. Intervalles de confiance sur la variance

D'après la table de khi-deux, $k_1 = k_{2,5\%} = 73,361$ et $k_2 = k_{97,5\%} = 128,422$

Avec Excel:

$k_1 = k_{2,5\%}$ = LOI.KHIDEUX.INVERSE(0,025;99) = 73,361

$k_2 = k_{97,5\%}$ = LOI.KHIDEUX.INVERSE(0,975;99) = 128,422

D'après les données du tableau (début de la section) on a: $S^2_{CA100} = 0,0462$

D'où on obtient: $ic_{95\%}\left(\sigma_A^2\right) = \left[\frac{99.0,0462}{128,422}; \frac{99.0,0462}{73,361}\right] \Rightarrow ic_{95\%}\left(\sigma_A^2\right) = \left[0,0356; 0,0624\right]$

Finalement, $ic_{95\%}\left(\sigma_A\right) = \left[0,1887; 0,2498\right]$

5-1. Intervalles de confiance sur la variance

L'hypothèse de normalité de la variable aléatoire parente est très importante dans ce cas de figure. L'intervalle de confiance est donné par la formule suivante:

$$IC_{(1-\alpha)}\left(\sigma^2\right) = \left[\frac{(n-1).S_{Cn}^2}{k_{\left(1-\frac{\alpha}{2}\right)}} ; \frac{(n-1).S_{Cn}^2}{k_{\left(\frac{\alpha}{2}\right)}}\right]$$

Traduction

Au seuil de 95%, l'écart type de la contenance des bouteilles de la chaîne A est compris entre $0,19\ cl$ et $0,25\ cl$. Il y a donc une probabilité d'au moins 95% que l'écart type des machines anciennes de la chaîne A ne soit plus proche de $0,1\ cl$. **La précision des machines s'est nettement dégradée.**

VI. Intervalles de confiance sur une proportion

6-1. Loi faible des grands nombres et estimateur d'une proportion

Proposition: Loi faible des grands nombres

Soit $X_1, X_2,\ldots,X_n$ une suite de n variables aléatoires indépendantes, d'espérances $\mu_1, \mu_2,\ldots,\mu_n$ finies et de variances $\sigma^2_1, \sigma^2_2,\ldots,\sigma^2_n$ finies .

Si $\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i\right)=\mu$ et si $\lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\right)=0$, alors $\bar{X}_n \xrightarrow{p} \mu$

Ce théorème s'applique à des variables quantitatives continues ou discrètes.

Il aurait donc pu servir pour démontrer que dans le cas d'un $n-EAS$, la moyenne empirique est un estimateur convergent de l'espérance de la variable parente.

Il s'agi donc ici, d'utiliser ce théorème appliqué à un $n-EAS$ où la variable parente suit une loi de Bernoulli, pour démontrer la proposition ci-dessous.

VI. Intervalles de confiance sur une proportion

6-1. Loi faible des grands nombres et estimateur d'une proportion

On pour tout i, $E(X_i)=\pi$ finie et $Var(X_i)=\sigma_i^2=\pi(1-\pi)$ finie, les variables sont indépendantes. Par ailleurs:

- $\frac{1}{n}\sum_{i=1}^{n}E(X_i)=\frac{1}{n}n\pi=\pi$ (ce qui montre que $F_n=\frac{1}{n}\sum_{i=1}^{n}X_i$ est un estimateur non biaisé de π).
- Et $\lim_{n\to\infty}\left[\frac{1}{n^2}\sum_{i=1}^{n}Var(X_i)\right]=\lim_{n\to\infty}\left[\frac{1}{n^2}n\pi(1-\pi)\right]=\lim_{n\to\infty}\left[\frac{\pi(1-\pi)}{n}\right]=0$

On a donc, d'après la loi faible des grands nombres: $F_n \xrightarrow{p} \pi$

Proposition

La fréquence empirique est un estimateur sans biais et convergent de la probabilité π.

Dans le cas particulier des estimations de proportions, la variable aléatoire parente X suit forcément une loi de Bernoulli. La situation est donc toujours celle d'un échantillon non gaussien. Pour la même raison l'écart type de X est forcément inconnu puisque $\sigma=\sqrt{\pi(1-\pi)}$

6-2. Intervalle de confiance sur une proportion

Pour n suffisamment grand, $IC_{(1-\alpha)}(\pi) = \left[F_n \pm z_{(1-\alpha/2)} \cdot \sqrt{F_n(1-F_n)} \Big/ \sqrt{n} \right]$

L'erreur standard vaut $SE_n = \sqrt{F_n(1-F_n)} \Big/ \sqrt{n}$, or on a $F_n \xrightarrow{p} \pi$

L'erreur standard SE_n, est fonction du paramètre π à estimer (et de la taille de l'échantillon).

Cette erreur est maximale lorsque son carré est maximum.

Pour n fixé assez grand, SE_n^2 atteint son maximum pour $\pi(1-\pi)$ maximum

Pour n fixé assez grand, L'erreur standard de l'estimation de π est maximale pour $\pi = \frac{1}{2}$

Donc l'erreur standard maximale est: $SE_n(\max) = \sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)} \Big/ \sqrt{n} = \frac{1}{2\sqrt{n}}$

La marge d'erreur maximale vaut toujours: $ME_n(\max) = z_{1-\alpha/2} \frac{1}{2\sqrt{n}}$