Statistique Inférentielle

S. Achchab

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes (ENSIAS)

2007-2008

Plan

Plan

Plan

Statistiques ?

Définition

Statistiques ?

Définition

- Statistiques : Ensemble de données.

Statistiques ?

Définition

- Statistiques : Ensemble de données.
- Statistique : Science qui consiste à collecter, analyser et interpréter des données.

Statistiques ?

Définition

- Statistiques : Ensemble de données.
- Statistique : Science qui consiste à collecter, analyser et interpréter des données.
- Statistique : Fonction de variables aléatoires.

Plan

Démarche statistique

Démarche statistique

Comporte deux grands aspects :

Démarche statistique

Démarche statistique

Comporte deux grands aspects :

- Aspect descriptif ou exploratoire

Démarche statistique

Démarche statistique

Comporte deux grands aspects :

- Aspect descriptif ou exploratoire
- Aspect inférentiel ou décisionnel

Plan

Définitions et notations

Définition

Unité statistique (ou individus) :
On appelle unité statistique l'élément de base de l'ensemble que l'on veut étudier.

Définitions et notations

Exemple

Définitions et notations

Exemple

- Si l'on s'intéresse à l'ensemble des petites et moyennes entreprises (PME). Une PME est considérée comme unité statistique.

Définitions et notations

Exemple

- Si l'on s'intéresse à l'ensemble des petites et moyennes entreprises (PME). Une PME est considérée comme unité statistique.
- Si l'on s'intéresse au revenu mensuel des ménages marocains, une famille est considérée comme unité statistique.

Définitions et notations

Définition

Population :
On appelle population l'ensemble des unités statistiques.

Définitions et notations

Exemple

Définitions et notations

Exemple

- Si l'on s'intéresse aux salariés d'une entreprise, alors l'ensemble des salariés forme la population.

Définitions et notations

Exemple

- Si l'on s'intéresse aux salariés d'une entreprise, alors l'ensemble des salariés forme la population.
- Si l'on s'intéresse au parc automobile marocain, ce parc est considéré comme population

Définitions et notations

Définition

On appelle modalité, l'aspect du caractère retenu dans le cadre de l'analyse.

Exemple

Définitions et notations

Définition

On appelle modalité, l'aspect du caractère retenu dans le cadre de l'analyse.

Exemple

- Pour une personne : la couleur des cheveux, la taille ...

Définitions et notations

Définition

On appelle modalité, l'aspect du caractère retenu dans le cadre de l'analyse.

Exemple

- Pour une personne : la couleur des cheveux, la taille ...
- Pour un ménage : le revenue mensuel, le nombre d'enfants ...

Définitions et notations

Définition

Définitions et notations

Définition

- **Caractère qualitatif :**
  C'est un caractère qui ne peut pas être mesuré, ni être repéré.

Définitions et notations

Définition

- **Caractère qualitatif :**
  C'est un caractère qui ne peut pas être mesuré, ni être repéré.
- **Caractère quantitatif :**
  Le caractère quantitatif peut être repérable ou mesurable, c'est-à-dire susceptible d'être soumise à une mesure donnant une valeur numérique.

Définitions et notations

Exemples

Définitions et notations

Exemples

- **Caractère qualitatif :** situation familiale : célibataire, marié, veuf...

Définitions et notations

Exemples

- **Caractère qualitatif :** situation familiale : célibataire, marié, veuf...
- **Caractère quantitatif :** Le revenu des ménages est mesurable en Dirhams, Taille d'un individu, poids ...

Définitions et notations

Définition

Variable statistique : C'est l'expression numérique ou littéraire retenue dans le cadre de l'analyse. On la note X_i.

Exemples

Définitions et notations

Définition

Variable statistique : C'est l'expression numérique ou littéraire retenue dans le cadre de l'analyse. On la note X_i.

Exemples

- Taille d'une entreprise : la variable exprimera les différentes tailles d'une entreprise : petite, moyenne, grande

Définitions et notations

Définition

Variable statistique : C'est l'expression numérique ou littéraire retenue dans le cadre de l'analyse. On la note X_i.

Exemples

- Taille d'une entreprise : la variable exprimera les différentes tailles d'une entreprise : petite, moyenne, grande
- Nombre d'enfants à charge des familles : la variable sera numérique : $0, 1, 2, \ldots, n$ enfants à charge.

Définitions et notations

Définition

Exemples

Dans une classe de 20 étudiants : 7 sont redoublants et 13 ne le sont pas.

$$n_1 = 7 \quad \text{et} \quad n_2 = 13.$$

Définitions et notations

Définition

- **Effectif :** C'est le nombre d'individus pouvant être rattachés à une variable. On le note n_i.

Exemples

Dans une classe de 20 étudiants : 7 sont redoublants et 13 ne le sont pas.

$$n_1 = 7 \quad \text{et} \quad n_2 = 13.$$

Définitions et notations

Définition

- **Effectif :** C'est le nombre d'individus pouvant être rattachés à une variable. On le note n_i.
- **Fréquence :** Elle peut être de deux types : soit absolue (c'est l'effectif lui-même n_i), soit relative (rapport d'un effectif à l'effectif total). On note la fréquence relative :

$$f = \frac{n_i}{N} \quad \text{où} \quad N = \sum_{i=1}^{p} n_i$$

Exemples

Dans une classe de 20 étudiants : 7 sont redoublants et 13 ne le sont pas.

$$n_1 = 7 \quad \text{et} \quad n_2 = 13.$$

Nature d'une variable : Variable quantitative

Définition

Nature d’une variable : Variable quantitative

Définition

- **Variable discrète :**
  Lorsque la variable ne peut prendre qu’une valeur numérique, on parle de variable discrète.

Nature d'une variable : Variable quantitative

Définition

- **Variable discrète :**
  Lorsque la variable ne peut prendre qu'une valeur numérique, on parle de variable discrète.
- **Variable continue :**
  On parle de variable continue lorsque celle-ci peut prendre plusieurs valeurs dans un intervalle donné. L'intervalle des valeurs possibles est divisé en n petits intervalles.

$$[\alpha, \beta] = [A_0, A_1[\cup[A_1, A_2[\cup \ldots \cup [A_{n-1}, A_n],$$

où $A_0 = \alpha < A_1 < \ldots < A_{n-1} < A_n = \beta$

Nature d'une variable : variable quntitative

Exemples

Nature d'une variable : variable quntitative

Exemples

- **Variable discrète :** Nombre d'automobiles possédées par les ménages.

Nature d’une variable : variable quntitative

Exemples

- **Variable discrète :** Nombre d’automobiles possédées par les ménages.
- **Variable continue :** Durée d’une conversation téléphonique, la variable pourrait être la suivante pour un effectif de 34 personnes :

Temps Ecoulés (en mn) X_i	Effectifs n_i
[0, 5[	10
[5, 15[	9
[15, 30[	15
Total	34

Nature d'une variable : Variable qualitative

Définition

Une variable qualitative peut être nominale ou ordinale.
Elle est nominale lorsque l'ensemble des modalités ne possède pas de structure particulière.
Une variable est ordinale lorsque l'ensemble des modalités est ordonné.

Définitions et notations

Exemples

Définitions et notations

Exemples

- **Variable nominale :** Dans un sondage, on s'intéresse à l'avis des consommateurs envers un produit : la variable X prend les modalités suivantes : " satisfait " ; " non satisfaits " et " ne sait pas ".

Définitions et notations

Exemples

- **Variable nominale :** Dans un sondage, on s'intéresse à l'avis des consommateurs envers un produit : la variable X prend les modalités suivantes : " satisfait " ; " non satisfaits " et " ne sait pas ".
- **Variable ordinale :**Si on prend la variable X " État des voitures ", X prend les modalités : " Neuf ", " Moyen " et " Médiocre ". X est une variable ordinale.

Plan

Plan

Statistique

Soit X_i la variable aléatoire égale à la valeur du caractère X pour le i^{eme} individu, on peut définir une deuxième fois un n-échantillon par toute suite $(X_1, X_2, \ldots, X_n)$ de variables aléatoires indépendantes et identiquement distribuées. La théorie de l'échantillonnage se propose d'étudier les propriétés du n-uple $(X_1, X_2, \ldots, X_n)$ et les caractéristiques le résumant.

Statistique

Définition

Une statistique T est une variable aléatoire fonction de $(X_1, X_2, \dots, X_n)$

$$T = f(X_1, X_2, \dots, X_n)$$

Statistique

Définition

La moyenne empirique

Elle est notée $\overline{X}$ et elle est définie par :

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

C'est la moyenne arithmétique des variables aléatoires de l'échantillon.

Statistique

Définition

La fréquence empirique

Soit un caractère qualitatif à deux modalités A et B que l'on peut coder par 0 et 1. On note p la proportion dans la population des individus qui possèdent la modalité A. Le nombre aléatoire Y_n d'individus de l'échantillon qui ont la modalité A, s'écrit :

$$Y_n = \sum_{i=1}^{n} X_i,$$

où les v.a X_i sont des v.a de Bernoulli de paramètre p. On pose $F_n = \dfrac{Y_n}{n}$ appelée fréquence empirique.

Statistique

Définition

La variance empirique

La variance empirique du caractère X dans l'échantillon est définie par :

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

Par ailleurs, une autre statistique utile en estimation est la variance empirique corrigée S^{*2} définie par :

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

Plan

On note $m_r = E[X^r]$ le moment d'ordre r de la variable aléatoire X, $r = E[(X - m)^r]$ le moment centré correspondant et $\sigma^2 = \mu_2$ la variance de X. Un premier résultat donne l'espérance et la variance de la moyenne empirique $\overline{X}$.

Proposition

Soit $(X_1, X_2, \ldots, X_n)$ un n-échantillon de la v.a. X. Si $E[X_2]$ existe, alors pour tout entier $n, \quad n > 1$, la moyenne empirique $\overline{X}$ vérifie :

$$E(\overline{X}) = m$$

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

Dans le cas particulier d'une fréquence empirique, on obtient :

$$E(F_n) = p$$

$$Var(F_n) = \frac{p(1-p)}{n};$$

Proposition

Soit X une variable aléatoire telle que $m_4 = E[X^4]$ existe. Alors, pour tout entier $n > 1$,

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

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

En outre, $Var[S^2]$ et $Var[S^{*2}]$ sont majorées par une expression de la forme $\frac{c}{n}$, où c est une constante dépendant de m_4.

Plan

Cas des grands échantillons

Définition

La suite infinie $(Y_n)_n$ de variables aléatoires converge en probabilité vers la v.a. Y si, pour tout $\epsilon > 0$,

$$Y_n \xrightarrow{P} Y \Leftrightarrow \lim_n p[|Y_n - Y| < \epsilon] = 1$$

Cas des grands échantillons

Théorème : (Lois des grands nombres)

Soit X une v.a. de moyenne m finie, alors pour toute suite infinie $(X_n)_n$ de v.a. deux à deux indépendantes et de même loi que X, la suite des v.a.

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

Converge en probabilité vers m lorsque n tend vers l'infini.

Cas des grands échantillons

Application

Grâce à la loi des grands nombres, on peut montrer que :

$$S_n^2 \xrightarrow{P} \sigma^2,$$

de même

$$S_n^{*2} \xrightarrow{P} \sigma^2,$$

Quand n tend vers l'infini et à condition que $E(X^2)$ existe.

Cas des grands échantillons

Théorème central limite

Soit X une v.a. de moment $E[X^2]$ fini, alors pour toute suite infinie $(X_n)_n$ de v.a. indépendantes et de même loi que X, la suite des v.a.

$$U_n = \frac{\overline{X}_n - m}{\sigma/\sqrt{n}},$$

tend en loi vers la loi normale centrée réduite $N(0, 1)$ lorsque n tend vers l'infini.

Cas des grands échantillons

Application : Cas d'une fréquence empirique

On prélève indépendamment et avec remise n individus d'une population séparée en deux sous-populations et de proportions p et $(1-p)$ (pièces défectueuses ou correctes dans une production industrielle par exemple). Soit F_n la fréquence empirique. On a

$$\begin{aligned} E[F_n] &= p \\ Var[F_n] &= \frac{p(1-p)}{n}. \end{aligned}$$

Et si n est grand,

$$F_n \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

Cas des grands échantillons

Exemple

Pour un échantillon de 400 pièces issues d'une fabrication où 10% sont défectueuses. On peut s'attendre à trouver dans 95% des cas un pourcentage de pièces défectueuses dans l'échantillon comprise entre : $10\% \pm 1.96\sqrt{\frac{0.1 \times 0.9}{400}}$ soit $9.7\% < F < 10.3\%$

Cas des grands échantillons

Corollaire

Soit X une v.a. de moment $E[X^2]$ fini, alors pour toute suite infinie $(X_n)_n$ de v.a indépendantes et de même loi que X, la moyenne empirique $\overline{X}_n$ tend vers la loi normale $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$

Cas des grands échantillons

Exemple

Soit un lot de 500 chocolats. Le poids d'un chocolat est une v.a. telle que $m = 5g$ et $\sigma = 0.5g$. Quelle est la probabilité qu'une boite de 50 chocolats issus de ce lot ait un poids total supérieur à $260g$?

Cas des grands échantillons

Proposition

Si X est une v.a. d'espérance m et d'écart-type σ, si g est une fonction numérique dérivable au voisinage de m et vérifiant $g'(m) \neq 0$, si enfin $\overline{X}_n$ désigne la moyenne empirique d'un n-échantillon de X, alors la suite des v.a.

$$g'(m)\frac{g(\overline{X}_n) - g(m)}{\sigma/\sqrt{n}},$$

tend en loi vers la loi normale $N(0, 1)$ lorsque n tend vers l'infini.

Plan

Cas des petits échantillons

Définition

Soit X une v .a. normale d'espérance m et d'écart-type σ . La densité associées est définie pour tout réel x par :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right].$$

De plus, on appelle échantillon normal tout échantillon de la v.a. X. Le résultat suivant sur la stabilité de la loi normale affirme que la somme de v.a. normales indépendantes est encore une v.a. normale.

Cas des petits échantillons

Proposition

Si $X_1, X_2, \ldots, X_n$ sont des v.a. indépendantes de lois respectives $N(m_i, \sigma_i^2)$, alors la variable aléatoire $Y = X_1 + \ldots + X_n$ suit la loi normale d'espérance mathématique $m = m_1 + \ldots + m_n$ et d'écart-type $\sigma = \sqrt{\sigma_1^2 + \ldots + \sigma_n^2}$.

Plan

Plan

Estimation

Définition

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (m, σ, *etc.*) à l'aide d'un échantillon de n observations issues de cette population. On suppose vérifiée l'hypothèse d'échantillonnage aléatoire simple.

Estimation

Exemple

D'après la loi des grands nombres :

$$\overline{X}_n \xrightarrow{P} m,$$

de même

$$S_n^2 \xrightarrow{P} \sigma^2,$$

et

$$F_n \xrightarrow{P} p,$$

Les variables aléatoires $\overline{X}_n$, S_n^2 et F_n sont appelées alors estimateurs de m, σ^2 et p respectivement.

Plan

Estimation

Qualité d'un estimateur

Soit θ le paramètre à estimer et T_n un estimateur, c'est-à-dire une fonction des X_i à valeurs dans un domaine acceptable pour θ

Estimation

Qualité d’un estimateur

Soit θ le paramètre à estimer et T_n un estimateur, c’est-à-dire une fonction des X_i à valeurs dans un domaine acceptable pour θ

- Convergence

Estimation

Qualité d'un estimateur

Soit θ le paramètre à estimer et T_n un estimateur, c'est-à-dire une fonction des X_i à valeurs dans un domaine acceptable pour θ

- Convergence
- Biais

Estimation

Qualité d'un estimateur

Soit θ le paramètre à estimer et T_n un estimateur, c'est-à-dire une fonction des X_i à valeurs dans un domaine acceptable pour θ

- Convergence
- Biais
- Précision

Estimation

Convergence

Il est souhaitable que si $n \longrightarrow \infty$, $T_n \longrightarrow \theta$

Estimation

Biais

L'erreur d'estimation $T_n - \theta$ qui est une variable aléatoire se décompose de façon élémentaire en :

$$T_n - \theta = \underbrace{T_n - E(T_n)}_{(1)} + \underbrace{E(T_n) - \theta}_{(2)}$$

Estimation

Biais

L'erreur d'estimation $T_n - \theta$ qui est une variable aléatoire se décompose de façon élémentaire en :

$$T_n - \theta = \underbrace{T_n - E(T_n)}_{(1)} + \underbrace{E(T_n) - \theta}_{(2)}$$

- (1) représente les fluctuations aléatoires de T_n autour de sa valeur moyenne

Estimation

Biais

L'erreur d'estimation $T_n - \theta$ qui est une variable aléatoire se décompose de façon élémentaire en :

$$T_n - \theta = \underbrace{T_n - E(T_n)}_{(1)} + \underbrace{E(T_n) - \theta}_{(2)}$$

- (1) représente les fluctuations aléatoires de T_n autour de sa valeur moyenne
- (2) biais

Estimation

Biais

Il est souhaitable d'utiliser des estimateurs sans biais tels que

$$E(T_n) = \theta$$

Estimation

Exemple

Estimation

Exemple

- $\overline{X}_n$ est un estimateur sans biais de m

Estimation

Exemple

- $\overline{X}_n$ est un estimateur sans biais de m
- S_n^2 est un estimateur avec biais de σ^2

Estimation

Exemple

- $\overline{X}_n$ est un estimateur sans biais de m
- S_n^2 est un estimateur avec biais de σ^2
- $S_n^{*2} = \dfrac{n}{n-1} S_n^2$ est un estimateur sans biais de σ^2

Estimation

Précision

On mesure généralement la précision d'un estimateur T_n par l'erreur quadratique moyenne :

$$E[(T_n - \theta)^2].$$

On peut démontrer que :

$$E[(T_n - \theta)^2] = V(T_n) + [E(T_n) - \theta]^2.$$

De deux estimateurs sans biais, le plus précis est donc celui de variance minimale.

Estimation

Exemple

Montrer que si m est connu, l'estimateur $T_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - m)^2$ est meilleur que S^{*2}.

Estimation

Définition

L'estimateur T_n est dit asymptotiquement sans biais si :

$$\lim_{n \to +\infty} E(T_n) = 0.$$

Estimation

Exemple

$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ est un estimateur asymptotiquement sans biais de σ^2.

Estimation

Définition

Un estimateur T_n du paramètre θ est dit convergent en moyenne quadratique (m.q) s'il vérifie :

$$\lim_{n \to +\infty} E[(T_n - \theta)^2] = 0.$$

Rappelons que la convergence en moyenne quadratique entraîne la convergence en probabilité, mais que la réciproque est fausse en général. En pratique, pour montrer le caractère convergent d'un estimateur, on essaie le plus souvent de montrer la convergence en moyenne quadratique.

Estimation

Proposition

Un estimateur T_n du paramètre θ est convergent en moyenne quadratique si et seulement si, il est asymptotiquement sans biais et si sa variance tend vers 0 lorsque n tend vers l'infini.

Preuve

Il suffit de voir :

$$E[(T_n - \theta)^2] = V(T_n) + [E(T_n) - \theta]^2.$$

Estimation

Exemple

Pour estimer la moyenne théorique $\theta = m$ d'une variable aléatoire X, on peut prendre $T_n = \overline{X}_n$. Cet estimateur est sans biais, si de plus sa variance $Var(X)$ est finie, la relation $Var(\overline{X}_n) = \dfrac{Var(X)}{n}$ montre que $Var(\overline{X}_n)$ tend vers 0 lorsque n tend vers l'infini. La proposition précédente permet alors d'affirmer que $\overline{X}_n$ est un estimateur convergent en moyenne quadratique de $m = E(X)$.

Estimation

Le critère de convergence en moyenne quadratique donné dans la proposition précédente suppose le calcul de l'espérance mathématique et la variance de l'estimateur T_n. Comme, en pratique ce calcul n'est pas toujours aisé, on propose une méthode permettant dans certains cas, de démontrer la convergence en probabilité de T vers .

Estimation

Proposition

Soit $(Y_n)_{n\geq 1}$ une suite infinie de v.a. qui converge en probabilité vers la v.a. Y et si g est une fonction d'une variable réelle, continue sur l'ensemble des valeurs possibles de Y et des Y_n, alors la suite des v.a. $g(Y_n)_{n\geq 1}$ converge en probabilité vers $g(Y)$ lorsque n tend vers l'infini, soit :

$$g(Y) = \lim_{n\to+\infty} g(Y).$$

Estimation

Exemple

L'estimateur S^2 (et aussi S^{*2}) est convergent en moyenne quadratique vers σ^2. Il est donc convergent en probabilité, or la fonction g définie par $g(u) = \sqrt{u}$ est continue sur $\mathbb{R}$. D'après la proposition précédente, $S = \sqrt{S^2}$ tend donc en probabilité vers σ lorsque n tend vers l'infini. S est donc un estimateur convergent du paramètre σ.

Estimation

Définition

Estimation

Définition

- Soit T_n et T_n' deux estimateurs sans biais de θ. T_n et dit plus efficace que T_n' si :

$$Var(T_n) \leq Var(T_n')$$

Estimation

Définition

- Soit T_n et T_n' deux estimateurs sans biais de θ. T_n et dit plus efficace que T_n' si :

$$Var(T_n) \leq Var(T_n')$$

- L'estimateur sans biais et de convergence minimale est appelé estimateur efficace.

Plan

Estimation de la moyenne

Proposition

Soit X_n une variable aléatoire dont on veut estimer la moyenne (ou espérance) $\mu = E(X)$ à partir d'un n-échantillon $(X_1, X_2, \ldots, X_n)$ de X.

On ne suppose rien sur la loi de X.

$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n}$, la moyenne empirique est un estimateur efficace de μ.

Estimation de la moyenne

Preuve

$\overline{X}_n$ est sans biais de μ, car $E(\overline{X}_n) = \mu$. De plus, à partir de $Var(\overline{X}_n) = \frac{Var(X)}{n} \stackrel{n \to +\infty}{\longrightarrow} 0$, on démontre que

$$\forall T, \text{un autre estimateur de} \mu, \; Var(T_n) > Var(\overline{X}_n)$$

Estimation de la variance d'une population Gaussienne

Proposition

Estimation de la variance d'une population Gaussienne

Proposition

- Si μ est connu

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

est un estimateur efficace de σ^2

Estimation de la variance d'une population Gaussienne

Proposition

- Si μ est connu

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

  est un estimateur efficace de σ^2

- Si μ est inconnu

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

  est un estimateur biaisé de σ^2, mais asymptotiquement sans biais.

Plan

Méthode du maximum de vraisemblance

Définition

La vraisemblance d'un échantillon $(X_1, X_2, \dots, X_n)$ de la v.a. X est la fonction L de $n+1$ variables définie par :

$$L(x_1, x_2, \dots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \dots f(x_n, \theta),$$

où $x_1, x_2, \dots, x_n$ sont des réalisations possibles de X, θ est le paramètre inconnu. On rappelle que $f(x, \theta)$ définie la loi de la v.a. X par la formule :

$$f(x, \theta) = \begin{cases} p[X = x], & \text{si } X \text{ est une v.a. discrète ;} \\ f_X(x), & \text{si } X \text{ est une v.a. continue.} \end{cases}$$

Méthode du maximum de vraisemblance

Remarque

Compte tenu de l'indépendance des v.a. $X_1, X_2, \ldots, X_n$, la vraisemblance $L(x_1, x_2, \ldots, x_n, \theta)$ n'est autre que la probabilité de réalisation de la suite des valeurs $x_1, x_2, \ldots, x_n$ lorsque le paramètre inconnu vaut θ.

Méthode du maximum de vraisemblance

Principe de la méthode

Étant donné un échantillon $x_1, x_2, \ldots, x_n$ à prendre comme estimation de θ la valeur de θ qui rend maximale la vraisemblance $L(x_1, x_2, \ldots, x_n, \theta)$. Il s'agit là d'un problème d'optimisation : on cherche une valeur θ_0 de θ qui maximise la fonction $L(x_1, x_2, \ldots, x_n, \theta)$.
La fonction de vraisemblance étant positive, on peut aussi se borner à rechercher le maximum sur l'ensemble des θ où cette fonction est strictement positive. Sur cet ensemble, il est alors commode et équivalent de chercher le maximum de la fonction :

Méthode du maximum de vraisemblance

Principe de la méthode

$$\log L(x_1, x_2, \dots, x_n, \theta) = \sum_{i=1}^{n} \log f(x_i, \theta),$$

une condition nécessaire sur θ est que

$$\frac{\partial \log L(x_1, x_2, \dots, x_n, \theta)}{\partial \theta} = 0,$$

pour $\theta = \theta_0$.
Et si elle remplie, une condition suffisante pour θ_0, alors on :

$$\frac{\partial^2 \log L(x_1, x_2, \dots, x_n, \theta)}{\partial^2 \theta} < 0,$$

pour $\theta = \theta_0$.

Méthode du maximum de vraisemblance

Principe de la méthode

L'estimateur du maximum de vraisemblance étant alors :

$$\widehat{\theta}_n = \theta_0(X_1, X_2, \ldots, X_n).$$

Méthode du maximum de vraisemblance

Exemple : Estimation de la moyenne et la variance d'une loi normale :

Dans ce cas :

$$\begin{aligned} X &\rightsquigarrow N(m, \sigma^2) \\ \theta &= (\theta_1, \theta_2), \end{aligned}$$

où $\theta_1 = m$ et $\theta_2 = \sigma^2$.
De plus :

$$f(x, \theta) = f(x, m, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right].$$

Méthode du maximum de vraisemblance

Exemple : Estimation de la moyenne et la variance d'une loi normale :

La vraisemblance de l'échantillon s'écrit :

$$L(x_1, x_2, \dots, x_n, \theta) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2\right],$$

et le logarithme népérien est

$$\log L(x_1, x_2, \dots, x_n, \theta) = n\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i - m}{\sigma}\right)^2 - n\log\sigma.$$

Méthode du maximum de vraisemblance

Exemple : Estimation de la moyenne et la variance d'une loi normale :

Les conditions nécessaires consistent en l'annulation des dérivées partielles premières par rapport à m et σ^2. Elles s'écrivent respectivement :

$$\begin{cases} \dfrac{\partial \log L}{\partial m} = -\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i - m) = 0, \\ \dfrac{\partial^2 \log L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{2\sigma^4}\displaystyle\sum_{i=1}^{n}(x_i - m)^2 = 0. \end{cases}$$

Méthode du maximum de vraisemblance

Exemple : Estimation de la moyenne et la variance d'une loi normale :

De ces deux relations, on conclut

$$\begin{cases} m_0 = \overline{x}, \\ \sigma^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2. \end{cases}$$

Plan

Plan

Méthode du maximum de vraisemblance

Définition

Il est souvent plus réaliste et plus intéressant de fournir un renseignement du type $a < \theta < b$ plutôt que d'écrire sèchement $\widehat{\theta} = c$.
Fournir un tel intervalle $[a, b]$ s'appelle donner une estimation par intervalle de confiance de θ.

Estimation par intervalle de confiance

Principe de la méthode

La méthode des intervalles de confiance est la suivante : Soit T un estimateur de θ, dont on connaît la loi de probabilité pour chaque valeur de θ.
Étant donné une valeur θ_0 de θ, on peut déterminer un intervalle de probabilité de niveau $1-\alpha$ pour T, c'est-à-dire deux bornes t_1 et t_2 telles que

$$p[t_1 < T < t_2] = 1 - \alpha,$$

pour $\theta_0 = \theta$.
Ces bornes dépendent évidemment de θ_0.
On choisira dans la plus parts des cas un intervalle de probabilité à risque symétrique $\frac{\alpha}{2}$ et $\frac{\alpha}{2}$.

Estimation par intervalle de confiance

Principe de la méthode

On adopte alors la règle de décision suivante :
Soit t la valeur observée de T :

- si t appartient à l'intervalle $[t_1, t_2]$, on conserve θ_0 comme valeur possible de θ ;
- si t n'appartient pas à l'intervalle $[t_1, t_2]$, on élimine θ_0.

On répète cette opération pour toutes les valeurs de θ.

Plan

Estimation par intervalle de confiance de m

L'écart-type σ est connu

$\overline{X}_n$ est le meilleur estimateur de m et $\overline{X}_n$ suit une loi $N\left(m, \frac{\sigma}{\sqrt{n}}\right)$. L'intervalle de probabilité de $\overline{X}_n$ à $1-\alpha$ est :

$$m - u_\alpha \frac{\sigma}{\sqrt{n}} < \overline{X}_n < m + u_\alpha \frac{\sigma}{\sqrt{n}},$$

d'où l'intervalle de confiance :

$$\overline{x}_n - u_\alpha \frac{\sigma}{\sqrt{n}} < m < \overline{x}_n + u_\alpha \frac{\sigma}{\sqrt{n}}.$$

Si $1-\alpha = 0.95$, on a $u_\alpha = 1.96$

Estimation par intervalle de confiance de m

L'écart-type σ est inconnu

On utilise le fait que $T = \dfrac{\overline{X}_n - m}{S_n^*/\sqrt{n}} = \dfrac{\overline{X}_n - m}{S_n/\sqrt{n-1}}$ suit une loi de student à $n-1$ degrés de liberté.

L'intervalle de probabilité pour t est

$$-t_\alpha < \frac{\overline{X}_n - m}{S}\sqrt{n-1} < t_\alpha.$$

D'où l'intervalle de confiance :

$$\overline{x}_n - t_\alpha \frac{S}{\sqrt{n-1}} < m < \overline{x}_n + t_\alpha \frac{S}{\sqrt{n-1}},$$

Estimation par intervalle de confiance de m

L'écart-type σ est inconnu

où bien

$$\overline{x}_n - t_\alpha \frac{S^*}{\sqrt{n}} < m < \overline{x}_n + t_\alpha \frac{S^*}{\sqrt{n}}.$$

Le théorème central-limite a pour conséquence que les intervalles précédents sont valables pour estimer la moyenne théorique m d'une loi quelconque à condition que n soit assez grand.

Plan

Estimation par intervalle de confiance de σ^2

m est connu

$T = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$ est le meilleur estimateur de σ^2 et $\frac{nT}{\sigma^2}$ suit la loi χ^2_n (chi deux à n degrés de liberté) comme somme de n carrés de loi $N(0,1)$ indépendantes.
Soient k_1 et k_2 les bornes de l'intervalle de probabilité d'une variable χ^2_n :

$$p\left[k_1 < \frac{nT}{\sigma^2} < k_2\right] = 1 - \alpha$$

L'intervalle de confiance est :

$$\frac{nt}{k_1} < \sigma^2 < \frac{nt}{k_2},$$

où t est la valeur prise par T.

Estimation par intervalle de confiance de σ^2

m est inconnu

On utilise $S^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$ et on sait que $\dfrac{nS^2}{\sigma^2}$ suit la loi χ^2_{n-1}.

Soit l_1 et l_2 les bornes de l'intervalle de probabilité ;

$$p\left[l_1 < \frac{nS^2}{\sigma^2} < l_2\right] = 1 - \alpha,$$

on a alors :

$$\frac{ns^2}{l_1} < \sigma^2 < \frac{ns^2}{l_2},$$

Plan

ENSIAS

Estimation par intervalle de confiance

Estimation du paramètre d'une loi binomiale

C'est le problème connu sous le nom d'intervalle de confiance pour une proportion p inconnu. Etant donnée une population infinie (ou fini si le tirage s'effectue avec remise) où une proportion p des individus possédant un certain caractère. Il s'agit de trouver un intervalle de confiance pour p à partir de f, proportion trouvée sur un échantillon de taille n.

Estimation par intervalle de confiance

Estimation du paramètre d'une loi binomiale

On sait que nF suit une loi binomiale $B(n, p)$; et si n est assez grand, on sait que :

$$nF \rightsquigarrow N(np, \sqrt{np(1-p)}),$$

donc que

$$F \rightsquigarrow N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

L'intervalle de probabilité symétrique est :

$$p - u_\alpha \sqrt{\frac{p(1-p)}{n}} < F < p + u_\alpha \sqrt{\frac{p(1-p)}{n}}.$$

Estimation par intervalle de confiance

Estimation du paramètre d'une loi binomiale

Posons $u = k$ pour simplifier les notations.
Les bornes de l'intervalle de probabilité sont données par :

$$y = p \pm k\sqrt{\frac{p(1-p)}{n}}.$$

Soit

$$(y-p)^2 = \frac{k^2 p(1-p)}{n},$$

ou

$$y^2 + p^2\left(1 + \frac{k^2}{n}\right) - 2py - \frac{k^2}{n}p = 0,$$

Estimation par intervalle de confiance

Estimation du paramètre d'une loi binomiale

Étant donné une valeur f observée, l'intervalle de confiance s'obtient en résolvant en p l'équation :

$$f^2 + p^2\left(1 + \frac{k^2}{n}\right) - 2pf - \frac{k^2}{n}p = 0,$$

ou

$$p^2\left(1 + \frac{k^2}{n}\right) - p\left(\frac{k^2}{n} + 2f\right) + f^2 = 0$$

Estimation par intervalle de confiance

Estimation du paramètre d'une loi binomiale

$$\begin{aligned}\Delta &= p\left(\frac{k^2}{n}+2f\right)^2-4\left(1+\frac{k^2}{n}\right)f\\ &= \frac{k^4}{n^2}+4f\frac{k^2}{n}-4f^2\frac{k^2}{n}.\end{aligned}$$

D'où

$$p=\frac{\left(2f+\frac{k^2}{n}\right)\pm\sqrt{\frac{k^4}{n^2}+4f\frac{k^2}{n}-4f^2\frac{k^2}{n}}}{2\left(1+\frac{k^2}{n}\right)}.$$

Estimation par intervalle de confiance

Estimation du paramètre d'une loi binomiale

Formule encombrante, mais dont on peut trouver une approximation en considérant que n est grand et en faisant un développement limité au premier ordre en $(\frac{1}{n})$; le premier terme

$$\frac{\left(2f + \frac{k^2}{n}\right)}{2\left(1 + \frac{k^2}{n}\right)} \approx f + o\left(\frac{1}{n^2}\right),$$

le second terme se réduit en simplifiant par n^2 :

$$\sqrt{\frac{k^2 + 4fnk^2 - 4f^2nk^2}{4(n+k^2)^2}} = \sqrt{\frac{k^2 + 4fnk^2 - 4f^2nk^2}{4n^2 + 8k^2n + 4k^4}}.$$

Estimation par intervalle de confiance

Estimation du paramètre d'une loi binomiale

Ce radical est équivalent au suivant (en écrivant que chaque terme est équivalent à celui de plus haut degré en n)

$$\sqrt{\frac{fnk^2 - f^2nk^2}{n^2}} = k\sqrt{\frac{f(1-f)}{n}}.$$

Donc, on a si n est grand, l'expression approchée suivante pour l'intervalle de confiance :

$$f - u_\alpha\sqrt{\frac{f(1-f)}{n}} < p < f + u_\alpha\sqrt{\frac{f(1-f)}{n}}.$$

Plan

L'estimateur Jack-Knife

Cette technique a été proposé pour diminuer le biais d'un estimateur

Définition

Soit T un estimateur calculé sur un échantillon de taille n. On note T_{-i} l'estimateur calculé sur le $(n-1)$-échantillon obtenu en enlevant l'observation i. On appelle pseudo-valeur T_i^* :

$$T_i^* = nT - (n-1)T_{-i}.$$

L'estimateur Jack-knife est alors la moyenne des pseudo-valeurs :

$$T_J = \frac{1}{n}\sum_{i=1}^{n} T_i^*,$$

L'estimateur Jack-Knife

Définition

ce qui donne

$$T_J = T - (n-1)\frac{1}{n}\sum_{i=1}^{n}(T_i - T).$$

La variance de l'estimateur de Jack-knife est alors donnée par :

$$S_J^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(T_i^* - T_J)^2}{n-1}$$

L'estimateur de Jack-knife

Réduction du biais

Supposons que $E(T) = \theta + \frac{a}{n}$, alors $E(T_J) = \theta$.

En effet :

$$
\begin{aligned}
E(T_J) &= E(T) - (n-1)[E(T_{-i}) - E(T)] \\
&= \theta + \frac{a}{n} - (n-1)[\theta + \frac{a}{n-1} - \theta - \frac{a}{n}] \\
&= \theta + \frac{a}{n} - a + \frac{n-1}{n}a = \theta
\end{aligned}
$$

L'estimateur de Jack-knife

Exercice :

A titre d'exercice, on peut vérifier que la méthode de Jack-knife appliquée à la variance S^2 donne l'estimateur S^{*2} et que appliquée à $\overline{X}$, on retrouve $\overline{X}$. Le calcul de Jack-knife est surtout utile pour des statistiques biaisées dont le biais est très difficile à calculé (coefficient de corrélation par exemple).

Plan

ENSIAS

Et c'est facile.