الجمهوريـــة الجزائـــريـــة الديمقراطيـــة الشعبيـــة
République Algérienne Démocratique et Populaire
وزارة التعليـــــم العالـــي والبحـــث العلمـــي
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

N° Réf :……

Centre Universitaire de Mila

Institut des Sciences et de la Technologie
Département de Mathématique et Informatique

Mémoire préparé En vue de l'obtention du diplôme de licence
En : - Filière Mathématiques Fondamentales

Thème
STATISTIQUE INFERENTIELLE

Dirigé par : MEHAZZEM ALLAL Grade : Maître Assistant.

Préparé par : GHIMOUZ CHADIA
BOUMAALI ARIFA
BAOUCHE MERIEM
BOUTARIA SARA

Année universitaire : 2013/2014

Remerciements

Premièrement, on remercie le bon Dieu qui nous a donné la confiance en nous, la santé, la force et la volontépour pouvoir terminer ce travail.

D'une part on remercie bien l'enseignant Mr "MEHAZZEM ALLAL'' qui était chargé de l'encadrement de ce mémoire et d'autre part on le remercie pour ce que nous avons acquit de lui en matière d'orientation et conseils.

Espérant bien que nos enseignants d'université reçoivent nos remerciements avec satisfaction et joie pour ce qu'ils nous donné durant la période d'étude à l'université.

On remercie également les membres qui ont participé du début jusqu'à fin de ce travail notamment : Sara, Arifa, Meriem et Chadia pour leurs volontés et leurs efforts à fin d'arriver à ce travail. Sans oublier tout ceux qui nous ont bien participé avec nous durant nos études pratiques.

Merci infiniment.

Table des matières

Introduction

Les premières problèmes d'inférence statistique auxquels s'applique la théorie des distributions d'échantillonnage sont les problèmes d'estimation. Le but poursuivi est d'estimer, à partir d'un échantillon, la où les valeurs numériques d'un ou plusieurs paramètres de la population considérée, et de déterminer la précision de cette ou de ces estimations.

La théorie de l'échantillonnage est l'étude des liens existants entre les caractéristiques (en général la moyenne et l'écart-type) de la population et ceux des échantillons de cette population. Les échantillons de la population doivent être représentatifs.

La théorie de l'estimation est l'étude des liens qui existent entre les caractéristiques connues d'échantillons et ceux correspondants de la population. La notion de précision est très importante dans ce cas : on détermine un intervalle dans lequel se trouve le paramètre à estimer avec une probabilité fixée.

Chapitre 1

Introduction à la statistique inférentielle

1.1 Généralités sur l'inférence statistique

1.1.1 Définitions

Population, échantillon

- population = ensemble d'unités statistique

  (poulets, étudiants inscrits en AES en 1996, firmes commerciales ...)

  recensement = observer toutes les unités de la population

- échantillon = sous-ensemble de la population étudiée

  (joueurs de foot = population équipe de St-Etienne = échantillon)

  sondage = observer les unités de l'échantillon (il aboutit, on le verra plus tard, à une distribution expérimentale)

- en statistique, on décrit ces groupes d'unités (population ou échantillon) à l'aide de mesures ou caractéristiques (effectif, moyenne, écart-type, pourcentage...)

- mesures ou caractéristique utilisées pour décrire une population s'appellent PARAMETRES.

- mesures ou caractéristiques utilisées pour décrire un échantillon s'appellent réalisations (ou observations) de STATISTIQUE.

L'inférence statistique

C'est l'ensemble des méthodes permettant de tirer des conclusions sur un groupe déterminé à partir des données provenant d'un échantillon choisi dans cette population.

1.1.2 Les problèmes à résoudre

Question 1

Exemple : Le responsable de la diffusion d'un produit fait un sondage pour connaître la dépense moyenne par différentes catégories socioprofessionnelle de la population française pour ce type d'achat. Il fera ainsi une estimation de cette dépense moyenne. Il peut aussi vouloir connaître la précision de cette estimation.

Ainsi, les statistiques sont utilisées pour ESTIMER les paramètres.

Un premier problème qui se pose est donc de faire des :

- estimations ponctuelles
- estimations par intervalle de confance

Question 2

Exemple : En matière de contrôle de qualité, on souhaite lors de la réception d'échantillons de pièces mécaniques comparer le taux de déchets observés par rapport à la norme fixée de manière à refuser le lot si son le taux de déchets dépasse la norme.

Dans la plupart des situations réelles, la valeur du paramètre est inconnue, mais il arrive que l'on ait une idée du paramètre et qu'on puissc formuler une HYPOTHESE concernant la valeur de celui-ci. Les observations peuvent confirmer ou infirmer l'hypothése formulée. Il arrive souvent que la différence entre la valeur de la statistique d'échantillon et la valeur hypothétique du paramètre ne soit ni petite ni grande, de sorte que la décision à prendre ne s'impose pas d'elle même. Il faut donc définir les critères qui permettent la prise de décision.

Ce sont les TEST DE CONFORMITE

Question 3

Les personnes qui décident sont souvent intéressées à déterminer si deux populations données sont semblables ou nettement différentes par rapport à une caractéristique particulière.

Ex.1 : un médecin peut vouloir déterminer si la réponse à un certain médicament (expérimental) diffère d'une groupe à un autre.

Ex.2 : un acheteur peut vouloir comparer la durée de vie d'un certain produit provenant de 2 foournisseurs différents.

Ce sont les TEST DE COMPARAISON.

Question 4

D'autres problèmes peuvent se poser, par exemple de savoir si une population donnée suit une loi de probabilité particuliére connue.

Ce sont les TESTS D'AJUSTEMENT (analytique) qui permettent de vérifier la qualité de l'ajustement de la population étudiée à une loi normale, binomiale, de poisson ou encore uniforme.

Ils ont pour but d'établir s'il est plausible que l'échantillon (aléatoire) provienne d'une population dont la loi de probabilité aurait été celle spécifiée.

Question 5

Il est intéressant de savoir, dans certaines situation, si 2 caractères qualitatifs sont indépendants. Les TEST D'INDEPENDANCE.

Question 6

On peut vouloir savoir si plusieurs populations sont homogènes par rapport à un certain caractère. Les TEST D'HOMOGENEITE.

1.1.3 Echantillon, réalisation d'échantillon, statistique

On veut, à partir d'un échantillon de la population, déduire des informations sur cette population. Le problème qui se pose alors est le suivant : comment choisir une partie de la population que reproduit le plus fidèlement posssible ses caractéristiques. C'est problème de l'échantillonnage.

Prélévement d'une échantillon (échantillonnage)

1. Echantillonnage sur la base des méthodes empiriques

La Méthode des quotas (respect de la composition de la population pour certains critères) est la plus utilisée.

2. Echantillonnages aléatoires

- Quand la probabilité de sélection de chaque élément de la population est déterminée avant même que l'échantillon soit choisi.

- Il permet de juger objectivement la valeur des estimations.

Echantillonnage aléatoire simple - on tire au hasard et avec remise les unités dans la populatoin concernée.

Echantillonnage stratifié

- Subdiviser d'abord la population en sous-ensembels (strates) relativement homogènes.

- Extraire de chaque strate un échantillon aléatoire simple.

- Regrouper tous ces échantillons.

Echantillonnage par grappes

- Choisir un échantillon aléatoire d'unités qui sont elles-mêmes des sous-ensembles de la population (grappes).

(**Ex** : diviser la ville en quartiers ; un certain nombre de quartiers sont choisis pour faire partie de l'échantillon ; on fait l'enquête auprès de toutes les familles résidant dans ces quartiers).

Modélisation de l'échantillonnage aléatoire simple

Dans la suite, on traite le cas de l'échantillonnage aléatoire simple, car les concepts fondamentaux et les formules importantes découlent de cette méthode.

Ce type d'échantillonnage consiste extraire un échantillon de taille n dans une population de taille N par des tirages aléatoires équiprobables et indépendants (tirages avec remise). On introduit le modèle suivant :

Soit $\Omega = \{\omega_1, ..., \omega_N\}$ la population constituée d'éléments appelés unités d'observation.

Soit X le caractére que l'on voudrait étudier sur l'ensemble de cette population.

X_K , le résultat aléatoire du K ièm tirage, est une v.a que suit la même loi que X. On note x_K le résultat du K ièm tirage.

On note $(X_1, ..., X_n)$ les résultats aléatoires de ces tirages.

Définition 1 *: $(X_1, ..., X_n)$ sont n v.a indépendates et de même loi (celle de X); il est appelé n-échantillon ou échantillon de taille n de X. Aprés tirage au sort, $(X_1, ..., X_n)$ prend les valeurs $(x_1, ..., x_n)$.*

Définition 2 *: La réalisation unique $(x_1, ..., x_n)$ de l'échantillon $(x_1, ..., x_n)$ est l'ensemble des valeurs observées.*

Définition 3 *: Une statistique Y sur un échantillon $(x_1, ..., x_n)$ est une v.a., fonction mesurable des X_{ki} ; $Y = f\ (x_1, ..., x_n)$.*

Après réalisation, la v.a. Y (statistique) prend la valeur $f\ (x_1, ..., x_n)$.

Les statistique sont utilisées pour estimer les caractéristiques de la population totale. Les statistiques les plus utilisées sont la moyenne empirique, la variance empirique, la fréquence empirique.

1.2 Quelques statistiques classiques

Rappels

$$E(aX+b) = aE(X)+b$$
$$E(X+Y) = E(X)+E(Y)$$
$$V(aX+b) = a^2\ V(X)$$
$$V(X) = E(X^2) - [E(X)]^2 = E([X-E(X)]^2)$$
si X, Y indépendantes,
$$V(X+Y) = V(X)+V(Y)$$

1.2.1 La moyenne empirique et la variance empirique

Posons $E(X)=\mu$, $V(X)=\sigma^2$ (inconnues)

Définition 4 *:On appelle moyenne empirique de l'échantillon $(x_1,...,x_n)$ de X, la statistique*

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

Sa réalisation est $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (qui est la moyenne de l'échantillon) aussi appelée moyenne observée.
(on verra plus tard que $\overline{X}$ estimer l'espérance $E(X)$)

Propriétés :

$$\begin{cases} E(\overline{X}) = \mu \\ V(\overline{X}) = \frac{1}{n}\sigma^2 \end{cases}$$

Calculons

$$E(\overline{X}) = E(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum E(X) = \sum_{i=1}^{n} E(X) = \mu$$

$$V(\overline{X}) = V(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}V(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} V(X)$$
$$= \frac{nV(X)}{n^2} = \frac{1}{n}V(X) = \frac{1}{n}\sigma^2$$

Définition 5 *:On appelle variance empirique de l'échantillon $(x_1,...,x_n)$ de X, la statistique*

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

Sa réalisation est $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$(qui est la variance de l'échantillon), Aussi appelée variance observée.

- **Propriétés :**

$$\begin{cases} E\left(S^2\right) = \dfrac{n-1}{n}\sigma^2 \end{cases}$$

Calculons

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

Loi de probabilité des statistique $\overline{X}$ et S^2

- Théorème limite centrale (pour l'échantillon) (rappel) :

Soit X une a.v t.q. $E\left(X\right) = \mu,\ V\left(X\right) = \sigma^2 \neq 0$

Soit $\left(X_1, ...X_n\right)$ un n-échantillon de X

$\overline{X} = \frac{1}{n}\left(X_1 + ... + X_n\right)$

Alors $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N\left(0,1\right)$ pour $n \to \infty$

(loi approximtive)

(ou bien $\overline{\mathrm{X}} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$pour $\mathrm{n} \to \infty$)

- 2 cas à étudier :

-a) la taille n de l'échantillon est grande

-b) X suit une loi gaussienne

a) Taille n **grande**

(d'aprés le thm. limite centrale)

1) $\boxed{\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \text{ suit approximativement } N(0,1)}$

$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ pour $n \to \infty$

ou bien)

$\boxed{\overline{X} \text{ suit approximativement } N(\mu, \frac{\sigma}{\sqrt{n}})}$ (en pratique $n > 30$)

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

solution : L'échantillon étant grand ($n = 50 > 30$) et on peut appliquer la première formule :

$$\overline{X} \sim N(5, \frac{0.5}{\sqrt{50}}) \text{ approximativement}$$

on pose $T = N(5 \times 50, \frac{50 \times 0.5}{\sqrt{50}}) = N(250, 0.5\sqrt{50}) \sim$

calculons

$$\begin{aligned} P(T > 260) &= P(U > \tfrac{260-250}{0.5\sqrt{50}}) = P(U > 2.831) \\ &= 1 - P(U < 2.83) = 1 - 0.9977 \end{aligned}$$

b) Echantillon gaussien

soit $X \sim N(\mu, \sigma)$

(d'après l'additivité pour des v.a. suivant des lois normales)

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

ou bien

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

Attention ! ! !

C'est une loi exacte et non une approximation comme dans le cas d'un échantillon de grand taille où la loi n'est pas connue.

2) $\boxed{\frac{n}{\sigma^2} S^2 \sim X^2_{n-1}}$

3) $\boxed{\frac{\overline{X}-\mu}{\sqrt{S^2}/\sqrt{n-1}} \sim t_{n-1}}$

$$U = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
$$Y = \frac{nS^2}{\sigma^2} \sim X^2_{n-1}$$
et alors
$$Z = \frac{U}{\sqrt{Y/(n-1)}} \sim t_{n-1}$$
calculons Z : $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}.\frac{1}{\sqrt{\frac{nS^2}{\sigma^2(n-1)}}} = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n-1}}}$

Exercice : On prélève 25 pièces dans une production industrielle. Une étude préalable a montré que le diamètre de ces pièces suivait une loi gaussienne de moyenne $10mm$ et d'écart-type $2mm$. Entre quelles valeurs a-t-on 85% de chances de trouver l'écart-type de ces pièces?

Solution :

Pour commencer, il faut déterminer α et β t.q.

$$\begin{aligned} 0.85 = p(\alpha < \tfrac{nS^2}{\sigma^2} < \beta) &= p(\tfrac{nS^2}{\sigma^2} < \beta) - p(\tfrac{nS^2}{\sigma^2} < \alpha) \\ &= 1 - p(\tfrac{nS^2}{\sigma^2} > \beta) - \left[1 - p(\tfrac{nS^2}{\sigma^2} > \ \alpha)\right] \\ &= p(\tfrac{nS^2}{\sigma^2} > \alpha) - p(\tfrac{nS^2}{\sigma^2} > \beta) \end{aligned}$$

On sait que $\frac{nS^2}{\sigma^2} \sim X^2_{25-1} = X^2_{24}$ et alors cherche dans la table du X^2_n à 24 degrés de liberté les valeurs α et β comme suit :

$$\begin{cases} p\left(\frac{ns^2}{\sigma^2} > \alpha\right) = 0.90 \\ p\left(\frac{ns^2}{\sigma^2} > \beta\right) = 0.05 \end{cases} \quad \text{(choix du aux tables)}$$

on trouve :

$$\begin{cases} \alpha = 15.659 \\ \beta = 36.415 \end{cases}$$

et alors

$$p(15.659 < \tfrac{25S^2}{2^2} < 36.415) = 0.85$$
$$p\left(2.5054 < S^2 < 5.8264\right) = 0.85$$
$$p\left(1.58 < S < 2.41\right) = 0.85$$

Attention ! ! !

Il ne faut pas confondre l'écart-type de l'échantillon, notés, valeur observée de statistique S (les calculs ont été faits pour cette statistique S), avec le PARAMETRE écart-type sur la population, noté σ, de la loi normale qui était connu dans ce problème !

Fréquence empirique F

Soit une population comportant deux modalitée A et B. Soit π la proportion d'individus de la population possédant la modalité A. $1-\pi$ est donc la proportion des individus de la population possédant la modalité B.

On extrait de la population un échantillon de taille n. Soit K_n la v.a qui représente le nombre d'individus dans l'échantillon ayant la modalité A.

Définition 6 *La v.a* $F = \frac{K_n}{n}$ *s'appelle fréquence empirique.*

Sa réalisation f est la proportion d'individus dans l'échantillon ayant la modalité A.

- **Propriétés :**

$$\begin{cases} K \sim \beta(n,\pi) \text{ donc} \\ E(F) = \pi \\ V(F) = \frac{\pi(1-\pi)}{n} \end{cases}$$

- Loi de probabilité pour F

$$\boxed{F \sim N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)}$$

dés que $n > 30, \pi \in [0.1, 0.9]$ .On trouve aussi $n\pi > 5$; $n(1-\pi) > 5$

(loi approximative).

$$\boxed{\frac{F-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0.1)}$$

Chapitre 2

Les problèmes d'estimation

2.1 Généralités

2.1.1 Mise enplace de la problématique et notations

On s'intéresse à un caractére ζ de la population totale β à laquelle on n'a pas accés.

Un échantillon au sens statistique est nécessairement issu d'un tirage au sort, volontaire les problèmes d'estimationire (démographie, agronomie,...) ou bien involontaire (résultats de mesures) les observations $(x_1, x_2, ..., x_n)$ de l'échantillon noté E sont des réalisations de n variables aléatoires $(X_1, X_2, ..., X_n)$ indépendantes et de même loi :

$$X_i \sim \ell(\theta)$$

Dans cette expression θ représente le(s) paramétre(s) de la loi qui modélise la répartition du caractère ζ dans la population β.Comme cette population est inconnue, ce paramètre est en général, inconnu au moins patiellement. Le but de la théorie de l'estimation est de donner une valeur à ce paramètre à partir d'un échantillon

issu de la population. La variable aléatoire (ou plutôt le vecteur aléatoire $(X_1, X_2, ..., X_n)$ sera noté X).

2.1.2 Estimateur et estimation

Définition 7 *:*

Un estimateur T_n du paramétre θ de la population est une variable aléatoire (ou statistique), fonction des variables aléatoires $(X_1, ..., X_n)$:

$T_n = f\ (X_1, ..., X_n)$.

Il est censé décrire le paramétre θ de la population.

Une estimation est une réalisation de cette variable aléatoire obtenue à partir de l'échantillon E :

$Tn = f\ (x_1, ..., x_n)$

Recherche d'estimateur

Il existe des méthodes de recherche d'estimateur comme la méthode de substituyion (méthode des moments dans le cas paramétrique), la méthode du maximun de varaisemblance, ou encore la méthode de la distance minimale. Ces méthodes sont utiles quand on n'a aucune intuition sur l'estimateur.

On trouve des détails sur ces méthodes dans les livres de statistique mathématique. On peut aussi travailler par simultation.

2.2 L'estimation de la moyenne et de la variance

2.2.1 L'estimation de la moyenne

La meilleure estimation de la moyenne m d'une population, qui puisse être déduite d'un échantillon aléatoire et simple, est à premiére vue moyenne $\overline{x}$ de l'échantillon, ce qui peut s'écrire comme suit :

$$\widehat{m} = \overline{x},$$

L'accent circonflexe placé au-dessus du symbole m signifiant qu'il s'agit de la valeur estimée de ce paramètre.

Pour l'ensemble des échantillon qui peuvent être rencontrés, on doit en effet retrouver ainsi, en moyenne, la $\ll$ vraie$\gg$ de la population, puisque :

$$E(\overline{x}) = m.$$

La dispersion des différentes estimation possibles, autour de cette moyenne générale, est mesurée par l'erreur-standard de la moyenne :

$$\sigma_{\overline{x}} = \sigma/\sqrt{n}.$$

Pour des raisons de facilité de calcul notamment, on pourrait être tenté de substituer parfois médiane $\widetilde{x}$ à la moyenne $\overline{x}$ de l'échantillon dans le cas des distributions normales, on retrouve ainsi également, en moyenne, la$\ll$ vraie$\gg$ valeur m :

$$E(\widetilde{x}) = \widetilde{m} = m$$

Toute fois, pour un même effectif n, la dispersion des estimations ainsi obtenues est en général supérieure à celle qui concerne la moyenne de l'échantillon :
$\sigma_{\overline{x}} \approx \sigma\sqrt{\frac{\pi}{(2n)}} > \frac{\sigma}{\sqrt{n}}$

2.2.2 L'estimation de la variance

On pourait également croire, a priori, que la meilleure estimation de la variance σ^2 d'une population est la variance S^2 d'un échantillon aléatoire et simple extrait de cette population en obtiendrait ainsi, en moyenne, une valeur non pas égale, mais inférieure à la variance de la population. $E(S^2) = \frac{(n-1)}{n}\sigma^2$ l'erreur systématique qu'on commettrait en procédant de cette manière peut évidemment être corrigée en multipliant la variance de l'échantillon par le facteur $\frac{n}{(n-1)}$.

On obtient alors l'estimation : $\widehat{\sigma}^2 = \frac{nS^2}{(n-1)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ dont l'espérance mathématique est bien σ^2 :

$$E\left[nS^2/(n-1)\right] = [n/(n-1)]\, E\left(S^2\right) = \sigma^2.$$

L'erreur-standard de cette estimation est, dans le cas d'une population normalpe :

$$\sqrt{V\left[nS^2\right]/(n-1)} = [n/(n-1)]\sqrt{2(n-1)\sigma^4/n^2} = \sigma^2\sqrt{2/(n-1)}.$$

En eaison de l'existence du facteur correctif $n/(n-1)$, de nombreux auteurs définissent dès le départ la variance S^2 en divisant la somme des carrés des écarts par $n-1$, et non pas par n :

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

on notera que si la valeur :

$$\widehat{\sigma} = ns^2/(n-1)$$

est une bonne estimation de la variance, sa racine carrées :

$$\widehat{\sigma} = s\sqrt{n/(n-1)} = \sqrt{ns^2/(n-1)}$$

n'est pas une estimation absolument correcte de l'écart-type.

2.3 Principes généraux de l'estimation

2.3.1 Les principales qualités des estimateurs

Considérons une population quelconque, dont la distribution de probabilité est fonction d'un paramètre γ, et un échantillon de n valeurs observées, extrait de cette population. D'une façon générale, on appelle estimateur du paramètre γ, toute fonction des valeurs observées ou de ceraines de ces valeur, susceptible de servir à estimer γ :

$G = G\ (X_1, ..., X_n)$

On appelle estimations les valeurs numériques g de cette fonction.

Qualité d'un estimateur :

Pour un échantillon de taille n de la loi de bernoullide paramétre inconnu p la variable aléatoire égale à la somme des composantes divisée par n (la fréquence empirique), est un estimateur de p c'est une variable aléatoire qui prende ses valeurs dans [0, 1]. Si n est grend avec une forte probabilité des valeurs prochesde p d'aprés la loi des grands nombres. Quel que soit le modéle et le paramétre à estimer des valeurs proches de ce paramètre au moins pour de grands échantillon est la qualité principale que l'on attend d'un estimateur. En toute rigueur on doit considérer une suite d'stimateur (T_n), où pour tout n T_n est une variable aléatoire fonction de l'échantillon $(X_1, ..., X_n)$. par abus de langage, on appelle encore "estimateur" cette suite .

Définition 8 *: On dit que l'estimateur* (T_n) *est convergent si pour tout* $\varepsilon > 0$

$$\boxed{\lim_{n\to\infty} P[|T_n - \theta| > \varepsilon] = 0}$$

Un estimateur convergent s'écarte donc du paramètre avec une faible probabilité, si la taille de l'échantillon est assez grande .

L'exemple de base d'estimateur convergent est la moyenne empirique. Nous noterons $\overline{X}_n$ la moyenne empirique de l'échantillon $(X_1, ..., X_n)$: $\overline{X}_n = \frac{x_1+...+x_n}{n}$

La loi faible des grands nombres affirme que $\overline{X}_n$ est un estimateur convergent de l'espérence de X .

Si le paramètre θ s'exprime comme une fonction continue de E [X], alors l'image de $\overline{X}_n$ par cette fonction un estimateur convergent de θ .

2.3.2 La méthode du maximum de vraisemblance

La varaisemblance L $(x_1, ..., x_n; \theta)$ représente la probabilité d'observer le n-uple $(x_1, ..., x_n)$ pour une valeur fixée de θ. Dans la situation inverse ici où on a observé $(x_1, ..., x_n)$ sans connaître la valeur de θ, on va attribuer à θ la valeur qui paraît la plus vraisemblable, compt tenu de l'obsarvation dont on dispose, c'est-à-dire celle que va lui attribuer la plus

forte probabilité. On se fixe donc la règle suivante : à $(x_1, ..., x_n)$ fixé, on considére la vraisemblance L comme une fonction de θ et on attribue à valeur qui maximise cette fonction. D'où la définition suivante :

Définition 9 *:*

On appelle estimateur du maximum de vraisemblanc (emv) toute fonction $\widehat{\theta}_n$ *de* $(x_1, ..., x_n)$ *que vérifie :*

$$\boxed{L(x_1,...,x_n\ ;\widehat{\theta}_n)=\max_{\theta\in\Theta} L(x_1,...,x_n)}$$

Cette définition ne renseigne en aucune façon, ni sur l'existence, ni sur l'unicité d'une tel estimateur. La recherche de l'emv peut se faire directement par recherche du maximum de L, ou dans le cas particulier où la fonction L est deux fois dérivable par rapport à θ, comme solution de l'équation $\frac{\partial L}{\partial \theta} = 0$ qui vérifie aussi $\frac{\partial^2 \ln L}{\partial \theta^2} < 0$, cependant la varaisemblance se calculant à partir d'un produit, on préfère remplacer ce dernier problème par le problème équivalent pour la log-vraisemblance, puisque la fonction ln est strictement croissante, $\frac{\partial \ln L}{\partial \theta} = 0$ avec $\frac{\partial^2 \ln L}{\partial \theta^2} < 0$ et qui aura une expression généralement simplifiée.

Remarquons enfin que si $\widehat{\theta}_n$ est un emv du paramètre, alors $g(\widehat{\theta}_n)$ est une emv du paramètre g (θ) pour toute fonction g. Si par exemple la variance empirique modifiée S_n^2 qui est un estimateur sans biais de $\theta = V_\theta(X)$, est un emv pour un modèle statitique donné, alors S_n est un emv du paramètre g $(\theta\ ) = \sigma_\theta(X) = \sqrt{V_\theta(X)} = \sqrt{\theta}$. Notons cependant que S_n ne peut pas être aussi un estimateur sans car on aurait alors :

$V_\theta(S_n) = E_\theta(S_n^2) - E_\theta^2(S_n) = \theta - \theta = 0.$

2.3.3 Quelques autres méthodes classiques

Diverses autres méthodes d'estimation sont aussi fréquemment utilisées, mais lesestimations qu'elles fournissent ne possédent pas nécessairement les qualités générales d'efficacité ou d'efficacité asymptotique des estimateurs du maximum de vraisemblance. Dans

certaines conditions de normalité notamment, certaines de ces méthodes donnent cependant des résultats identiques à la méthode du maximum de vraisemblance.

D'une façon générale, en vue de l'estimation de k paramètres, la méthode des moments a pour principe d'égaler les k premiers moments estimés de la population exprimés en fonction des k paramètres, aux k premiers moments de l'échantillon. En termes de moments non centrés, on obtient ainssi un système d'équations du type :

$$\begin{cases} \widehat{\alpha}(\widehat{\gamma}_1, ..., \widehat{\gamma}_k) = a_1 \\ \widehat{\alpha}(\widehat{\gamma}_1, ..., \widehat{\gamma}_k) = a_k \end{cases}$$

Sous des conditions très générales, les estimateurs définis de cette manièreont tous des distributions asymptotiquement normales.

En autre, on peut être deux méthodes d'estimation plus particulières à savoir : la méthode du X^2 minimum, qui s'applique essentiellement à certains problèmes relatifs aux distributions de fréquence, et la méthode des moindres carrés, qui intervient notamment dans le dommaine de larégression.

2.3.4 Les méthodes robustes et non paramétriques

Le calcul de la moyenne ≪ rognée ≫ peut être concidéré comme une première méthode robuste d'estimation de la moyenne. D'une manière plus générale que précédemment, la moyenne ≪ rognée≫ peut être définie comme suit, pour des observations par ordre croissant :

$$\boxed{\overline{x}_\alpha = \frac{1}{n(1-\alpha)} \sum_{i=1+\alpha n/2}^{n-\alpha n/2} \mathrm{x}_i}$$

- α étant la proportion des observations qui sont écartées du calcul de la moyenne.

Une autre solution robuste consiste à se baser uniquement sur les valeurs d'un certain nombre de quantiles à partir des trois quartiles (ou ce qui est équivalent, du premier quartile, de la médiance et du troisième quartile), on peut par exemple estimer comme suit la moyenne, au moins pour des distribution symétriques :

$\widehat{m} = 0.3q_1 + 0.4q_2 + 0.3q_3 = 0.3q_1 + 0.4\widehat{x} + 0.3q_3$

La méthode du ≪ jackknife ≫ permet également d'obtenir des estimateurs robustes. Pour un paramètre quelconque, dont la valeur observée est g, on désingne par g_{-i} les n valeurs qui peuvent être calculées à partir des sous-échantillons obtenus en éliminant à tour de rôle chacune des n observations, et on en déduit les quantités :

$g_i = ng - (n-1)\, g_i$, qui sont applées pseudo-valeurs . La moyenne de ces quantités est un bon estimateur du paramétre γ :

$$\widehat{\gamma} = \frac{1}{n}\sum_{i=1}^{n} gi.$$

On peut en effet démontrer que, de cette façon, on élimine dans une largemesure le biais éventuel de l'estimation directe de γ par g g, et qu'on peut disposer en outre d'une estimation de la variance de la valeur estimée, à parti de la somme des carrés des écarts des pseudo-valeurs par rapport à leur moyenne :

$$\boxed{\widehat{\sigma_{\widehat{\gamma}}}^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} (g_i - \widehat{\gamma})^2.}$$

2.3.5 Les méthodes bayesiennes

Les méthodes dites bayesiennes ounéo-bayesiennes ont des fondements très dif-férents de la méthode du maximum de vraisembelance et des autres méthodes classiques d'estimation.

Les méthodes bayesiennes ont en effet pour principe de tenir compte, non seulement des information qui proviennent de l'échantillon, mais aussi de certaines informations qui seraient disponibles a priori, en dehors de l'échantillon observé.

Dans cette optique, on considére que le paramètre à estimer est, non pas une constante, mais bien une variable aléatoire, à propos de laquelle on possède certaines informations. On peut alors appliquer le théorème de BAYES ou sa transposition au cas de variables discontinues ou continues.

En vue de nous limiter à une présentation simple des principes de bas de cette méthode, nous n'envisageons que le cas d'un paramètre γ variant de façon discontinue, dont la distribution, connue a priori, est caractérisée par une série de probabilités P (g_i). On peut alors écrire :

$P(g_i/B) = P(B/g_i)P(g_i)/[P(B/g_1)P(g_1) + ... + P(B/g_m)P(g_m)]$.

2.4 Les intervalles de confiance

2.4.1 principese généraux

Soit $\widehat{\theta}$ l'estimateur d'un paramètre inconnu.

$\widehat{\theta}$ est une variable aléatoire dont la loi de probabilité notée $L(\left(\widehat{\theta}\right))$ supposée connue dépend de. Il est possible de trouver deux valeurs particulieres $t_1(\theta\ )$

et $t_2(\ \theta)$ telles que :

$1 - \alpha = prop\left[t_1\left(\widehat{\theta}\right) < \widehat{\theta} < t_2\left(\widehat{\theta}\right)\right]$

S'il est possible de réécpendent de $\left(\widehat{\theta}\right)$ et tel que :

Rire le système d'inégalités en isolant, on peut déterminer un intervalle

dont les limites dé

$$1 - \alpha = prob\left[g_1\left(\widehat{\theta}\right) < \theta < g_2\left(\widehat{\theta}\right)\right]$$

Ici, l'intervalle qui encadre θ est aléatoire et il possède la propriété de recouvrir la valeurθ dans $1 - \alpha$ des cas. La prise en compte d'un échantillon particulier, c'est-à-dire d'une valeur numérique particulière pour, $\widehat{\theta}$ et donc pour $g_1\left(\widehat{\theta}\right)$ et $g_2\left(\widehat{\theta}\right)$, permet d'obtenir une fourchette qui a de≪ grandes≫ chances de ≪contenir≫ la valeur inconnue si $1 - \alpha$ est élevé.

$1 - \alpha = prob\left[g_1\left(\widehat{\theta}\right) < \theta < g_2\left(\widehat{\theta}\right)\right] = prob\left[c_1 < \theta < c_2\right]$

$*$ $[c_1; c_2]$ ou $\left[g_1\left(\widehat{\theta}\right),\ g_2\left(\widehat{\theta}\right)\right]$ est appelé intervalle de confiance.

$*$ c_1, c_2 sont les limites de confiance.

$*$ $1-\alpha$: degré de confiance ou degré de certitude.

2.4.2 L'intervalles de confiance pour la moyenne

On a vu le meilleur estimateur de la moyenne m est la moyenne $\overline{X}$ de L'échantillon de plus cet estimateur suit lui-même une loi normale LG (m, $\sigma/\sqrt{n}$). On construit alors l'intervalle de confiance de m en lisant dans la table de la loi normale la valeur a correspondant à la probabilité $(1-\alpha)$

$$P(-a < \frac{\overline{X}-m}{\sigma/\sqrt{n}} < a) = 1-\alpha.$$

On en déduit l'intervalle de confiance suivant :

$$I = \left[\overline{X} - a \times \frac{\sigma}{\sqrt{n}},\ \overline{X}\ + a \times \frac{\sigma}{\sqrt{n}}\right].$$

On peut remarque que les bornes de cet intervalle sont aléatoires et dépendent de l'écart-type de la loi normale de X.

Si cet écart-type n'est pas connu, on considère l'estimateur s de cet écart-type $\frac{\overline{X}-m}{\sigma/\sqrt{n}}$ suit la loi normale LG(0, 1) pour n suffisamment grand,la variable :

$\frac{nS^2}{\sigma^2}$ suit la loi $X^2(n-1)$.

Par définition de loi de stuent, on en déuit que la variable $\frac{\overline{X}-n}{\frac{\sigma}{\sqrt{n}}} \times \frac{1}{\sqrt{\frac{nS^2}{(n-1)\sigma^2}}}$. Suit la loi de student à $(n-1)$ degrés de liberté.

Aprés simplification, on obtient comme variable de student la variable : $\frac{\overline{X}-m}{\frac{S}{\sqrt{n-1}}}$.

L'intervalle de confiance se construit alors en cherchant dans la table de la loi de student la valeur b telle que :

$P(-b < \frac{\overline{X}-m}{S/\sqrt{n-1}} < b) = 1-\alpha$.

L'intervalle de niveau de confiance égal à $(n-1)$ pour m est alors :

I = $\left[\overline{X} - b \times \frac{S}{\sqrt{n-1}},\ \overline{X} + b \times \frac{S}{\sqrt{n-1}}\right]$

2.4.3 L'intervalles de confiance pour la variance

Si la moyenne m de la loi est connue, le meilleur estimateur de la variance σ^2 est la statistique :

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

La variable $\frac{nT}{\sigma^2}$ suit une loi du chi-deux à n degrés de liberté. la table de cette loi fournit, à α fixé, deux valeurs a et b telles que : $\mathrm{P}\left(a < \frac{nT}{\sigma^2}\right) = 1 - \alpha$. On en déduit l'intervalle de confiance au risque α pour σ^2 : $I = \left[\frac{nT}{b}, \frac{nT}{a}\right]$.

Si la moyenne m de la loi n'est pas connue, le meilleur estimateur de la variance σ^2est la statistique :

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

La variable $(n-1)S^2/\sigma^2$ suit la loi du chi-deux $X^2(n-1)$. Dans la table de cette loi, à α fixé, nous trouvons les valeurs c et d telles que :

$$P(c < \frac{(n-1)\,S^2}{\sigma^2} < d) = 1 - \alpha$$

L'intervalle de confiance au seuil $(1-\alpha)$ pour σ^2 est alors :

$$I = \left[\frac{(n-1)\,S^2}{d}, \frac{(n-1)\,S^2}{c}\right]$$

2.4.4 L'intervalle de confiance pour proportion

Considérons une population dont l'effectif est important dans laquelle une proportion inconnue p d'individus possèdent un caractère particulier. On souhaite construire un intervalle de confiance pour cette proportion p.

On dispose d'un échantillon de taille n qui nous donne une estimation f de p. La variable nf suit alors une loi binomiale de paramètre n et p. Pour n suffisamment grand, la convergence en loi de la loi B(n, p) nous permet de concidérer que la variable :

$\frac{nf-np}{\sqrt{np(1-n)}}$

à α fixé, la table de la loi normale nous fournit la valeur a telle que (en négligeant les corrections de continuité) :

$$P(-a < \frac{np-nf}{\sqrt{np(1-p)}} < a) = 1 - \alpha$$

soit : $$P\ (f - a\sqrt{\frac{p(1-p)}{n}} < p < f + a\sqrt{\frac{p(1-p)}{n}}) = 1 - \alpha$$

L'intrevalle de confiance a des bornes qui dépendent du paramètre à estimer, trois solution sont usuelleement utilisées :

$*$ $P\,(1-p)$ est maximum pour p= 1/2. Quand on rempllace p par cette valeur, on obtien l'intervalle maximum à α fixé

$*$ On remplace p par son estimation f sur l'échantillon.

$*$ On détermine les bornes $p_1$et p_2 de l'intervalle de confiance, à α fixé solution de l'inéquation du second degré en p :

$(p-f)^2 < a^2 \times \frac{p(1-p)}{n}$

Bibliographie

1 - Jean-pierre Lecoutre - **Statistique et probabilités** - 4e édition, Dunod paris 2009.

2 - Thérèse Phan et Jean-pierre Rowenczyk - **Exercices et problèmes de statis tique et probabilités -** 2e édition, Dunod paris 2012.

3 - Frédéric Bertrand et Myriam Maumy-Bertrand - **Initiation à la statistique avec R -** Dunod paris 2010.

4 - Pierre Dagnelie - **Statistique théorique et appliquée** - 2e édition de boeck université Rue des Minimes 39, Bruxelles.

5 - B. L. Vander Waerden - Traduit par : DEGENNE, C. Guichat - **Statistique mathématique,** Dunod Paris 1967.

6 - Jean Jacques Droesbeke - **éléments de statistique**, édition de l'université de Bruxelles 1998.

7 - Yves Hébert - **Mathématiques probabilités et statistique**, Paris librairie Vui bert, Boulvard Saint, Germain 63, 1974.