Statistiques Inférentielles 2

Contents

Partie 1

Introduction

La Statistique inférentielle à pour but de dégager à partir de renseignements sur un échantillon des renseignements sur une population entière.

Rappels

Moyenne

$$\overline{x} = \frac{1}{n} \sum r_i.x_i$$

Variance ($Unit^2$)

$$Varx = \overline{x^2} - \overline{x}^2 = \frac{1}{n}. \sum r_i.(x_i)^2 - (\overline{x})^2 = \frac{1}{n} \sum r_i.(x_i - \overline{x})^2$$

Ecart-type

$\sigma(x) = \sqrt{Var(x)}$

—

$S_n^2 = \frac{r_i.(C_i)^2}{n} - \overline{x}^2$ r sont le nombre d'occurence de l'échantillon et c la valeur dans l'échantillon

Procédé

Si on prends un échantillon que l'on soumets à une classification et une énumération dans des catégories propres. On peut calculer sur cet échantillon : une moyenne sur l'échantillon. A première vue, nous serions tentés d'extrapoller les résultats de cette moyenne à toute la population. Mais le résultat serait hautement incorrect et sera discuté ultérieurement. Pour obtenir 95% des résultats : On est entre (Moyenne - 2.Ecart-Type) et (Moyenne + 2.Ecart-Type).

Estimation des paramètres

Estimation ponctuelle, notion d'estimateur

une **estimation ponctuelle du paramètre** (estimateur) : paramètre dont la valeur sera calculée sur l'échantillon.

Notations

N : *taille de la population*
n : *taille de l'échantillon*

m : *moyenne pour la population*
$\overline{x}$: *moyenne pour l'échantillon*

p : *proportion pour la population*
f : *fréquence pour l'échantillon*

σ^2 : *variance pour la population*
S_n^2 : *variance pour l'échantillon*

Les paramètres n,$\overline{x}$ et S_n^2 sont des estimateurs. On constate que n et $\overline{x}$ sont des estimateurs **sans biais** tandis que S_n^2 est **biaisé** de θ^2. tandis que S_{n-1}^2 (**variance corrigée**) est un estimateur **sans biais**(non exhaustif).

Echantillons exhaustifs ou non exhaustifs

Un echantillon est **exhaustif** si un même individu ne peut être interrogé plus d'une fois. Sinon, il est **non exhaustif**. La méthode non exhaustive fournit des résultats numériquement proches de la méthode exhaustive tout en utilisant des formules plus simples.

La variable aléatoire "estimateur"

Un estimateur est une *variable aléatoire* qui dépends de l'échantillon choisi.

Non-Exhaustif

- $E(\overline{x}) = m$
- $Var(\overline{x}) = \frac{\sigma^2}{n}$
- $E(f) = p$
- $Var(f) = \frac{p.(1-p)}{n} . \frac{N-n}{N-1}$
-

Exhaustif

Qualités d'un estimateur

n estimateur est d'autant plus efficace que sa variance est petite. D'une manière générale, on décrit θ comme le paramètre à estimer et $\widehat{\theta}$ son estimateur.

$\widehat{\theta}$ doit être **consistant** (au plus l'échantillon de personnes interrogées grandit, au plus $\widehat{\theta}$ se rapproche de θ

$\widehat{\theta}$ doit aussi être **sans biais** (cad : $E(\widehat{\theta}) = \theta$.

Un estimateur est **absolument correct** si il est **sans biais** et **consistant**

Estimation par intervalle de confiance

Introduction

L'estimation par intervalle de confiance d'un paramètre θ est un procédé qui consiste à déterminer un intervalle dans lequel le paramètre θ à une certaine probabilité de se trouver. En général, on fixe l'intervalle de confiance à 95% soit un niveau de confiance de 0.95. à l'opposé du niveau de confiance se situe le niveau d'incertitude α $(1-\theta)$.

Il faut donc déterminer L_i et L_s tels que $\Pr\{L_i \leq \theta \leq L_s\}=1-\alpha$

Intervalle de confiance d'une proportion

> Sur un échantillon de 1 000 personnes, on constate 18% de fumeurs. On souhaite déterminer un intervalle dans lequel le pourcentage (vrai) de fumeurs pour l'ensemble de la population a 95 chances sur 100 de se trouver. (Echantillon non exhaustif).

Il faut donc trouver L_i et L_s tels que $\Pr\{L_i \leq p \leq L_s\}=0.95$

La variable aléatoire est une binomiale qui peut être, sur un échantillon de grande taille approximé par une loi normale d'espérance n.p et de variance n.p.(1-p).on s'interresse donc aux échantillons $n \geq 30$ et $n.p \geq 5$ et $n.(1-p) \geq 5$

n.f suit donc une loi normale

$$\frac{f-E(f)}{\sqrt{var(f)}} \approx X_G^*$$

Nous cherchons donc $-a$ et a tels que

$$Pr\{-a \leq X_G^* \leq a\} = 0.95 \Leftrightarrow Pr\{X_G^* \leq a\} = 0.975$$

0.975 = 1- $\frac{1-0.95}{2}$

à la lecture de la table, 0.975 nous donne un a equivalent à 1,96. On remplace ensuite X_G^* dans la formule, ce qui donne :

$$Pr\{f-1.96.\sqrt{\frac{p.(1-p)}{n}} \leq p \leq f+1.96.\sqrt{\frac{p.(1-p)}{n}}\}$$

On se retrouve cependant à estimer p qui contient p dans son expression. On peut soit

- remplacer p par son estimation ponctuelle f
- remplacer p par 0.5 qui maximise la quantité p.(1-p)

Si l'échantillon avait été **exhaustif**, l'interval recherché aurait été :

$$\left[f-1.96.\sqrt{\tfrac{N-n}{N-1}}.\sqrt{\tfrac{p.(1-p)}{n}}; f+1.96.\sqrt{\tfrac{N-n}{N-1}}.\sqrt{\tfrac{p.(1-p)}{n}}\right]$$

Que faire si Le niveau de confiance n'existe pas dans la table ? Imaginons un niveau de confiance de 0.99. On obtiens donc 1-1/2.(1-0,99) = 0,995 = a. Il n'existe pas dans la table d'entrées égales à 0.995. Nous alons donc chercher les deux valeurs qui encadrent 0.995 soit a^-=0.99492 à g^-=2.57 et a^+=0.99506 à g^+=2.58.

par interpolation linéaire,

$$a \approx g^- + \frac{a - a^-}{a^+ - a^-}.(g^+ - g^-) = g$$

Dans notre exemple, g est donc la valeur gauss = 2.575714

Intervalle de confiance d'une moyenne

Les échantillons de grande taille

Le résonnement utilisé pour le calcul de p à partir de f peut être étendu à $\overline{x}$. ce qui nous donne pour un échantillon non-exhaustif :

$$\left[\overline{x} - g.\frac{sigma}{\sqrt{n}}; \overline{x} + g.\frac{sigma}{\sqrt{n}}\right]$$

et pour un échantillon exhaustif :

$$\left[\overline{x} - g.\sqrt{\frac{N-n}{N-1}}.\frac{sigma}{\sqrt{n}}; \overline{x} + g.\frac{sigma}{\sqrt{n}}\right]$$

Les échantillons de petite taille ($n < 30$)

Si la population est **distribuée normalement**, c'est-à-dire dans l'exemple ci-dessus, si la durée des interventions étudiées *suit une loi normale*, on peut dire que la variable aléatoire « **moyenne standardisée des échantillons** » **suit une loi t de Student à (n − 1) degrés de liberté**.

Une variable aléatoire de Student est une variable aléatoire définie de $-\infty$ à $+\infty$ et symétrique par rapport à l'axe verticale.

Notée t_v, elle dépends de v (nombre de degrés de liberté). et son graph ressemble à celui de gauss plus applatit. mais il s'en approche en $v \to \infty$ et $E(t_v) = 0$, $Var(t_v) = \frac{v}{v-2}$ avec $(v > 2)$

v peut être calculé sur base du nombre de personnes interrogées pour former l'échantillon : si n = 20, v = 20-1

Intervalle de confiance d'une variance

SKIP

Marge d'erreur associée à un intervalle de confiance

De façon générale, la marge d'erreur vaut : * Pour l'intervalle de confiance d'une proportion : $a.\sqrt{\frac{p.(1-p)}{n}}$ * Pour un intervalle de confiance de moyenne : $a.\frac{\sigma}{\sqrt{n}}$

Pour diminuer cette marge d'erreur, nous pouvons soit **augmenter la taille de l'échantillon**, soit **augmenter le niveau d'incertitude**. La marge d'erreur maximale est obtenue lorsque $p = 0.5$

Partie 2

Introduction

Dans le cadre de l'évaluation d'un caractère mesurable. Un paramètre lié à ce caractère est inconnu (moyenne, proportion, variance) ou une loi de probabilité est ce caractère inconnu. Sur base d'une intuition ou d'une connaissance partielle de la réalité, on décide d'émettre une hypothèse concernant ce paramètre ou la loi de distribution. Il est ensuite judicieux d'être capable de porter un jugement sur cette hypothèse sur base d'échantillons obtenus. On opposera donc les résultats théoriques aux résultats issus de l'analyse des échantillons. Si Un écart trop important venait à être remarqué, l'hypothèse serait ainsi fausse et rejetée.

Hypothèse nulle, hypothèse alternative, seuil de signification

L'hypothèse nulle est celle que nous allons tester, notée H_0. Il s'agit pour les paramètres, d'une égalité et *d'affirmations positives pour les tests du khi-deux.*

L'hypothèse alternative est l'hypothèse que nous acceptons si on rejette H_0, notée H_1. Elle s'exprime sous la forme d'une différence($\neq$), d'une inégalité stricte et sous forme d'une affirmation négative

> Exemple :
> Un client à comme critère une résistance moyenne d'au moins égale à 13kg et un fabricant, lui produit un fil avec une résistance moyenne de 13kg.

Pour le client on a $H_0 : m = 13$ et $H_1 : m < 13$
Tandis que le fabricant aura $H_0 : m = 13$ et $H_1 : m \neq 13$

Le seuil de signification est la probabilité de rejeter H_0 au profit de H_1 alors que H_0 est **vraie**. C'est le risque de commettre une erreur. cette erreur est appelée **erreur de $1^{ère}$ espèce ou risque de la $1^{ère}$ espèce. et est noté α

Pour les exemples, nous choisirons un *seuil de signification* $\alpha = 0.05$. Cela signifie que si l'on rejette l'*hypothèse nulle*, on a 5% de chances de commettre une erreur.

> N.B. On pourrait envisager le risque ou erreur de 2^e espèce noté β qui serait le risque d'accepter L'*hypothèse nulle* alors que celle-ci est fausse.

Méthodologie (peu importe le type de tests)

- Fixer les Hypothèses $H_0 et H_1$
- Préciser le seuil de signification de α
- Préciser la loi de probabilité utilisée (différente en fonction de l'échantillon, type de test, ...)

- Déterminer le seuil de rejet, la zone de non refus
- Calculer la grandeur expérimentale
- Tirer les conclusions

Tests de valeur

Dans un test de valeur, on test l'hypothèse selon laquelle le paramètre (moy, proportion) est égal à une valeur déterminée. H_1 exprimera ensuite que le paramètre étudié est ($\neq$,$>$ ou $<$) de la valeur déterminée.

Dans le cadre de l'évaluation d'une modification sur la population à partir de renseignements obtenus sur un échantillon, on est en droit de se poser la question suivante: **Cette différence est-elle sufisament significative pour affirmer que le paramètre de la population à changé?

Test de valeur d'une proportion

Dans le cas de grands échantillons (n $\geq$ 30).

$$n \geq 30 \Rightarrow f \approx N(E(f), \sqrt{var(f)}$$

...

> L'appélation "test unilatéral gauche" vient du fait que le seuil de rejet se situe à gauche de l'axe de symétrie de la courbe de Gauss.
>
> Dans certains ouvrages,l'hypothèse est systématiquement le contraire de l'hypothèse alternative. (ex : $H_0 : p = 0.21$ est remplacé par $H_0 : p \geq 0.21$) ceci ne change ni le déroulement du test, ni sa conclusion.

Pour des Raisons de facilité de compréhension, dans les tests de valeur et d'égalité, nous exprimerons toujours l'hypothèse nulle sous forme d'une égalité.

Test de valeur d'une moyenne

a) Les échantillons de grande taille ($n \geq 30$) rappel : cfr Intervalle de confiance d'une moyenne (part 1)

Test d'indépendance

On peut sur base d'un échantillon de taille n sufisant déduire si deux variables X et Y sont indépendantes ou non.