Outils Statistiques pour la data intelligence (Module 1)

Introduction

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

Rappels

Quelques rappels de statistique descriptive :

- **Moyenne** : $\overline{x} = \frac{1}{n}\sum r_i.x_i$
- **Variance (Unités²)** : $Var(x) = \overline{x^2} - \overline{x}^2 = \frac{1}{n}.\sum r_i.(x_i)^2 - (\overline{x})^2 = \frac{1}{n}\sum r_i.(x_i - \overline{x})^2$
- **Ecart-Type** : $\sigma(x) = \sqrt{Var(x)}$

Estimation des paramètres

Estimation ponctuelle, notion d'estimateur

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

Notations

Variable	Estimateur
p	f
m	$\overline{x}$
σ^2	S_n^2

> 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 échantillon est dit **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

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

Exhaustif

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

Qualités d'un estimateur

Un bon estimateur se doit d'être le plus proche possible du paramètre qu'il estime. Il est d'autant plus efficace que sa *variance est petite*.

D'une manière générale, on décrit θ comme le paramètre à estimer et $\hat{\theta}$ son estimateur.

- $\hat{\theta}$ doit être **consistant** (au plus l'échantillon de personnes interrogées grandit, au plus $\hat{\theta}$ se rapproche de θ)
- $\hat{\theta}$ doit être **sans biais** (cad : $E(\hat{\theta}) = \theta$).

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

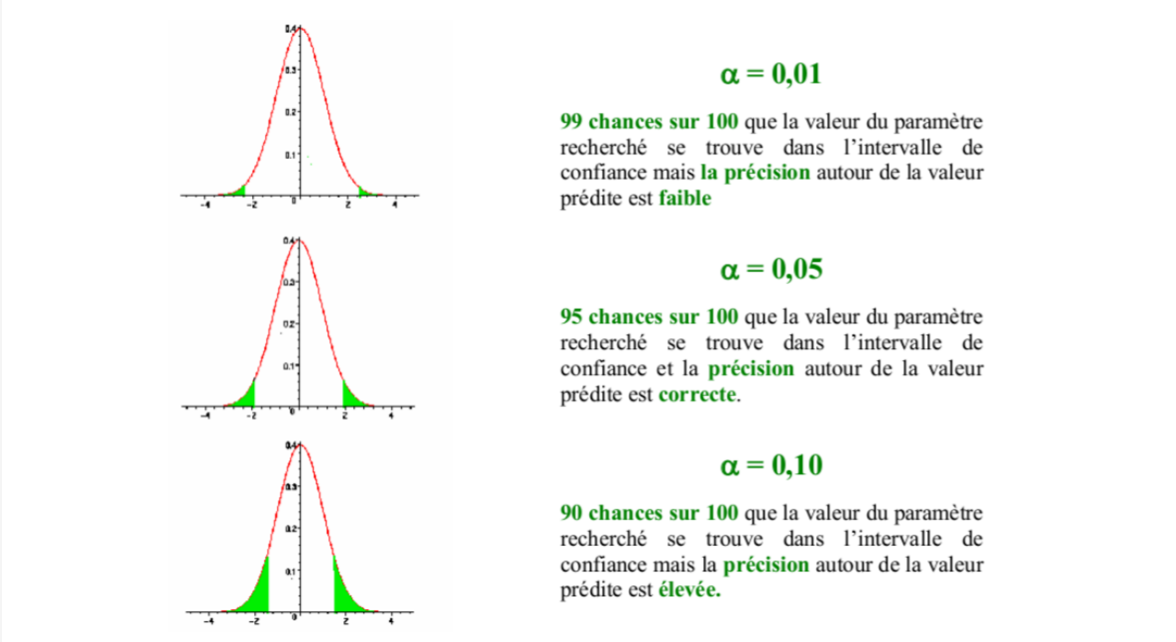
Estimation par intervalle de confiance

Introduction

L'estimation par intervalle de confiance d'un paramètre est un procédé qui consiste à déterminer un intervalle dans lequel le paramètre à une certaine probabilité de se trouver. En général, on fixe l'intervalle de confiance à 95% soit un niveau de confiance de $0,95$. En opposition au niveau de confiance se situe le niveau d'incertitude α qui vaut ici $(1-0,95)$ soit $\alpha = 0,05$.

Niveau d'incertitude

Le niveau d'incertitude est fonction de la précision souhaitée lors de l'étude. On aura donc tendance à diminuer celui-ci dans le cadre d'une étude nécessitant une haute précision.



La zone colorée représente la **marge d'erreur tolérée**.

Procédé

On cherches donc à estimer un paramètre. On décide d'un niveau d'incertitude (généralement 0,05). On peut ensuite calculer sur cet échantillon : une moyenne sur l'échantillon. Le but est donc d'obtenir 95% des résultats entre 2 bornes. On aura donc les 2 bornes suivantes : (Moyenne $\pm$ 2.σ^2).

L_i et L_s vont donc représenter respectivement la limite inférieur et la limite supérieur.

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

La variable aléatoire recherchée suivra généralement une loi binomiale. Cette loi peut être approximée par une gaussienne si elle respecte les critères suivants :

- Le nombre de personnes interrogées est supérieur à 30. ($n > 30$)
- Au moins 5 personnes dans les personnes interrogées valident le préposât ($n.p > 5$)

> Si tel est le cas, on peut approximer cette variable aléatoire par une gaussienne d'espérance n*p

Intervalle de confiance d'une proportion

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 np et de variance $np.(1-p)$.on s'intéresse 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- $\frac{1-0.95}{2}$

> $X = 1 - \frac{\alpha}{2}$

à la lecture de la table, 0.975 nous donne un a équivalent 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)

Que faire si l'on traite le cas d'un échantillon exhaustif ?

Si l'échantillon avait été **exhaustif**, l'intervalle recherché aurait été :
$\left[f - 1.96.\sqrt{\frac{N-n}{N-1}}.\sqrt{\frac{p.(1-p)}{n}}; f + 1.96.\sqrt{\frac{N-n}{N-1}}.\sqrt{\frac{p.(1-p)}{n}} \right]$

Que faire si Le niveau de confiance n'existe pas dans la table ?

Dans le cas d'un niveau de confiance $(1 - \alpha)$ où a qui vaut $(1 - \frac{\alpha}{2})$ n'existerait pas dans la table directement.

Nous allons donc chercher les deux valeurs qui l'encadrent soit a^- qui est équivalent à la valeur précédente dans la table (avec une valeur associée de g^-) et une valeur a^+ qui est la valeur suivante dans la table (avec une valeur associée de g^+).

par interpolation linéaire, $a \approx g^- + \frac{a-a^-}{a^+-a^-}.(g^+ - g^-) = g$

> Si on prends l'exemple d'un $a = 0,99$. Le g vaut donc 2.575714 par interpolation linéaire.

Intervalle de confiance d'une moyenne

Les échantillons de grande taille $(n \geq 30)$

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 :

$\overline{x} \approx X_G$ et donc par conséquent $\frac{\overline{x}-E(\overline{x})}{\sqrt{var(\overline{x})}} \approx X^*_G$

On peut donc en déduire les formules suivantes :

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

et pour un échantillon exhaustif :

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

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). Son graph est semblable à celui de gauss plus aplatit. 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.

Comme pour la gaussienne, il existe une table à laquelle on peut se référencer pour trouver quelle est la probabilité.

$t_v \approx \frac{\overline{x}-E(\overline{x})}{\sqrt{var(\overline{x})}}$

> exemple : Sur un échantillon de 20 patients, on constate une durée moyenne de 125 min et d'écart type 30. Quel est l'intervalle qui reprends 95% des moyennes de temps des interventions.
>
> On à un v = 19 =(20-1).
>
> On recherche donc Pr(-a<t_{19}<a)=0,95
>
> on à donc les 2 paramètres α = 0.05 et v = 19. la simple lecture de la table de student donne donc 2.093.

En sachant que $E(\overline{x}) = m$ et $var(\overline{x}) = \frac{\sigma^2}{n}$

On obtiens la fonction suivante

$$Pr(\overline{x})\left\{\overline{x} - SudentVal * \sqrt{\frac{\sigma^2}{n}} \leq E(\overline{x}) \leq \overline{x} + SudentVal * \sqrt{\frac{\sigma^2}{n}}\right\}$$

Echantillon exhaustif

$$Pr(\overline{x})\left\{\overline{x} - SudentVal * \sqrt{\frac{N-n}{N-1}} * \sqrt{\frac{\sigma^2}{n}} \leq E(\overline{x}) \leq \overline{x} + SudentVal * \sqrt{\frac{N-n}{N-1}} * \sqrt{\frac{\sigma^2}{n}}\right\}$$

Intervalle de confiance d'une variance

SKIP

Marge d'erreur associée à un intervalle de confiance

La marge d'erreur associée à un intervalle correspond à la **demi amplitude de cet intervalle**.

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**.

Marge d'erreur Maximale

La marge d'erreur maximale est obtenue lorsque $p = 0.5$