Statistique (MATH-F-315, Cours #1)

Thomas Verdebout

Université Libre de Bruxelles

Plan de la partie Statistique du cours

1. Introduction.
2. Théorie de l'estimation.
3. Tests d'hypothèses et intervalles de confiance.
4. Régression.
5. ANOVA.

Plan de la partie Statistique du cours

1. Introduction.
2. Théorie de l'estimation.
3. Tests d'hypothèses et intervalles de confiance.
4. Régression.
5. ANOVA.

Motivation

La Statistique

Ensemble de méthodes et outils mathématiques visant à *collecter, décrire et analyser* des données afin d'obtenir de l'information permettant de prendre des *décisions* malgré la présence d'incertitude

La statistique joue un rôle essentiel dans de nombreuses disciplines: En Biologie, Géographie, Géologie, Médecine, Chimie, Physique, etc...

$\rightarrow$ La statistique permet de confronter une théorie scientifique à l'observation!

Motivation

Exemple 1

Imaginons que nous sommes intéressés par l'âge moyen d'un Belge. En particulier, on se demande si cet âge moyen est < 40 ans.

Supposons que nous n'ayons pas les "moyens" d'effectuer un recensement de sorte que l'âge moyen en Belgique n'est tout simplement pas connu. Que pouvons-nous faire? C'est un problème statistique.

Nous décidons de façon naturelle de prendre un échantillon de belges. L'âge du belge est donc modélisé par une variable aléatoire X.

La question considérée devient donc est-ce que $\mathrm{E}[X] < 40$?

Motivation

L'idée naturelle consiste à considérer un échantillon $(X_1, \ldots, X_n)$, associé à n belges (on prend leur âge). On dira qu'il s'agit d'un échantillon aléatoire simple si ces variables aléatoires sont inépendantes et identiquement distribuées ("i.i.d.").

Ceci signifie que ces v.a. (i) sont mutuellement indépendantes, et (ii) possèdent toutes la même distribution. Les valeurs prises par l'échantillon $(X_1, \ldots, X_n)$ par $(x_1, \ldots, x_n)$ (les minuscules sont dans tout le cours réservées aux valeurs numériques observées (non aléatoires), tandis que les majuscules désignent les v.a. dont ces valeurs observées sont des réalisations).

Pour estimer $\mathrm{E}[X]$, l'idée la plus naturelle consiste à calculer la moyenne empirique $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (ou $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$) et de fonder ses conclusions sur la valeur de cette moyenne. Est-elle "réellement" plus petite que 40? Comment faire exactement pour tenir compte de la valeur de $\bar{X}$ afin de résoudre notre problème?

Puisque $\bar{X}$ est une fonction des v.a. $X_1, \ldots, X_n$, elle est elle-même une v.a., avec sa propre distribution: on parlera de distribution échantillonnée.

Exemple 2

Election présidentielle française 2012. François Hollande et Nicolas Sarkozy s'affrontent au second tour de l'élection. Un sondage d'avril 2012 effectué auprès de $n = 1000$ personnes donne comme vainqueur François Hollande avec 54% des suffrages contre 46% pour Nicolas Sarkozy.

Si on note p "la vraie" proportion de Français qui vote pour François Hollande, nous n'avons dès lors qu'une estimation $\hat{p} = 54/100$ de cette proportion. Quelles conclusions peut-on vraiment tirer sur ce qui nous intéresse, c'est-à-dire p? Que signifie cette marge d'erreur dont on entend souvent parler dans les médias?

Exemple 3

Un jeu de données assez célèbre en Biostatistique est le jeu de données relatif aux Iris de Fisher. Prenons une seule variable dans ce jeu de données qui est la longueur d'un pétale.

Setosa	Versicolor	Virginica
0.2	1.4	2.5
0.4	1.5	2.3
0.5	1.8	1.9
0.2	1.3	2.3
0.5	1.6	2.4

Peut-on conclure que les longueurs des pétales de ces trois espèces sont différentes sur base de cet échantillon?

Plan de la partie Statistique du cours

1. Introduction.
2. Théorie de l'estimation.
3. Tests d'hypothèses et intervalles de confiance.
4. Régression.
5. ANOVA.

Introduction au problème

Soit le modèle statistique $\mathcal{P} = \{\mathrm{P}_{\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \Theta\},\ \Theta \in \mathbb{R}^k$ engendrant le vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ qui reflète l'objet d'intérêt:

1. âge de n personnes sélectionnées en Belgique;
2. présence ou nom d'un virus chez n personnes;
3. etc

Nous supposerons dans ce cours que les les X_i sont i.i.d et nous noterons $\mathrm{P}_{\boldsymbol{\theta}}$ leur loi commune (obtenue comme produit des loi marginales par l'indépendance). Le plus souvent, nous supposerons que θ est un paramètre unidimensionnel qui peut représenter une caractéristique de la distribution sous-jacente. Par exemple, l'espérance de la distribution. Dans cette première partie du cours, nous nous intéressons à l'estimation de $\boldsymbol{\theta}$.

Introduction au problème

On appelle *statistique* toute fonction *mesurable* des observations.

Ainsi, par exemple, $T(X_1, \ldots, X_n) := \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ est une statistique, alors que $S(X_1, \ldots, X_n) := \frac{1}{n}\sum_{i=1}^{n}(X_i - \theta)$ où θ est le paramètre inconnu du modèle (par exemple $\mathrm{E}[X_1]$), n'en est pas une. En effet, pour chaque valeur de θ, S est différente.

Definition

On appelle *estimateur* de $\boldsymbol{\theta}$ toute *statistique* à valeurs dans Θ.

Remarque: on appelle *estimateur* de $g(\boldsymbol{\theta})$ toute *statistique* à valeur dans $g(\Theta)$.

Introduction au problème

Plusieurs notations peuvent être utilisées pour l'estimateur de $\boldsymbol{\theta}$: $\mathbf{T}(\mathbf{X})$, $\hat{\boldsymbol{\theta}}(\mathbf{X})$, $\hat{\boldsymbol{\theta}}$, etc.

Exemple Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$. Les statistiques suivantes sont toutes à valeurs dans $\mathbb{R}$, et constituent donc des estimateurs de μ:

1. $\bar{X} := \frac{1}{n} \sum_{i=1}^{n} X_i$ (moyenne arithmétique)
2. $X_{1/2}^{(n)}$ (médiane empirique)
3. $\frac{1}{2}(X_{1/4}^{(n)} + X_{3/4}^{(n)})$ (milieu de l'intervalle interquartile)
4. $\frac{1}{80} \sum_{i=11}^{90} X_{(i)}$ (moyenne tronquée ("trimmed"))
5. X_1 (première observation)
6. 0 (l'origine)
7. ...

Comment choisir un bon estimateur dans cette situation?

Les propriétés d'un estimateur sont, en fait, les propriétés de sa *distribution échantillonnée*. Dans la suite de cette première partie, nous nous intéressons à certaines caractéristiques de cette loi échantillonnée.

Estimateur sans biais

Notons $\mathrm{E}_{\boldsymbol{\theta}}[\dots]$ une espérance calculée sous $\mathrm{P}_{\boldsymbol{\theta}}$.

Definition

Un estimateur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ est dit *sans biais* si

$$\mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\boldsymbol{\theta}}\right] = \boldsymbol{\theta}, \qquad \forall \boldsymbol{\theta} \in \Theta$$

(ce qui implicitement requiert que $\mathrm{E}_{\boldsymbol{\theta}}\left[\hat{\boldsymbol{\theta}}\right]$ existe et soit finie pour tout $\boldsymbol{\theta}$).

La différence $\mathrm{E}_{\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}] - \boldsymbol{\theta}$ est appelée *biais* de l'estimateur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$.

Estimateur sans biais

Exemple 1 Soient $X_1, \ldots, X_n$ i.i.d. $\mathrm{Bin}(1, p)$, $p \in (0, 1)$. La proportion empirique

$$\hat{p} := \frac{1}{n} \sum_{i=1}^{n} X_i$$

est un estimateur sans biais de la probabilité correspondante p, puisque

$$\begin{aligned} \mathrm{E}_p \left[\frac{1}{n} \sum_{i=1}^{n} X_i \right] &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_p[X_i] \\ &= \frac{1}{n} np = p \qquad \forall p \in [0, 1]. \end{aligned}$$

Exemple 2 Soient $X_1, \ldots, X_n$ i.i.d. , $\mathrm{E}[X_i] = \mu < \infty$. La moyenne empirique

$$\bar{X} := \frac{1}{n} \sum X_i$$

est un estimateur sans biais pour μ (propriété qu'elle partage toutefois avec X_1)

Estimateur sans biais

Exemple 3 Soient $X_1, \ldots, X_n$ i.i.d. , $\mathrm{Var}(X_i) = \sigma^2 < \infty$, $\mathrm{E}[X_i] = \mu$. La variance empirique

$$s^2 := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

est un estimateur biaisé de σ^2. En effet,

$$\begin{aligned}
\mathrm{E}[s^2] &= \mathrm{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i^2\right] - \mathrm{E}[\bar{X}^2] \\
&= \frac{n}{n} \mathrm{E}[X_1^2] - (\mathrm{Var}(\bar{X}) + \mathrm{E}^2[\bar{X}]) \\
&= \mathrm{Var}(X_1) + \mathrm{E}^2[X_1] - (\mathrm{Var}(\bar{X}) + \mathrm{E}^2[\bar{X}]) \\
&= \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n} \sigma^2 < \sigma^2
\end{aligned}$$

Le biais de s^2 se corrige facilement, et

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

est un estimateur sans biais de σ^2, puisque $\mathrm{E}[S^2] = \mathrm{E}\left[\frac{n}{n-1} s^2\right] = \frac{n}{n-1} \mathrm{E}[s^2] = \sigma^2$.