Statistique (MA101) Cours 5

ENSTA 1ère année

Christine Keribin

christine.keribin@math.u-psud.fr

Laboratoire de Mathématiques
Université Paris-Sud

2017-2018

Sommaire

Do you remember ?

Un test est une procédure de décision qui permet de trancher, au vu des résultats d'un échantillon X, entre deux hypothèses l'hypothèse nulle (H_0) et une hypothèse alternative (H_1), dont une seule est vraie.

- Définir le modèle
- Définir les hypothèses nulle (H_0) et alternative (H_1)
- Choisir une statistique de test $T(X)$, calculer sa loi sous (H_0)
- Définir la règle de décision en calibrant la région de rejet $\mathcal{R}$ suivant le risque de première espèce α :

$$\alpha = \mathbb{P}_{H_0}(T(X) \in \mathcal{R})$$

- Calcul de la puissance $\pi = \mathbb{P}_{H_1}(T(X) \in \mathcal{R})$
- Calcul de la statistique observée et décision : rejet ou acceptation de (H_0).

Do you remember ?

Dans le test de l'espérance d'une loi gaussienne $\mu = \mu_0$ contre $\mu > \mu_0$, la valeur observée de la statistique de Student sur un échantillon de 12 individus est $t_{obs} = 1.5$.

- Quelle est la décision au niveau 5% ? Quel est le risque de cette décision ?
- Quelle est la décision au niveau 10% ? Quel est le risque de cette décision ?

α	0.9	0.95	0.975	0.99
$q_{\mathcal{T}(11)}(\alpha)$	1.36	1.80	2.20	2.72

Probabilité critique ou p-value

- C'est le niveau obtenu si le seuil est remplacé par la statistique observée, "plus petit niveau qui fait rejeter (H_0)" au vu des données
- Exemple : test de Student de $\mu = \mu_0$ contre $\mu > \mu_0$,
  - $\hookrightarrow$ rejet : $\mathcal{R} = \{t; t = T(x) > qt(1-\alpha, n-1)\}$
  - $\hookrightarrow$ niveau : $\mathbb{P}_{H_0}(T(X) > qt(1-\alpha, n-1)) = \alpha$
  - $\hookrightarrow$ valeur observée de la stat de test : $t_{obs} = T(x_{obs})$
  - $\hookrightarrow$ p-value : $P_c(t_{obs}) = \mathbb{P}_{H_0}(T(X) > t_{obs})$
- Donc,
  - $\hookrightarrow$ si $P_c(t_{obs}) \leq \alpha$, c'est que t_{obs} est dans la région de rejet de (H_0)
  - $\hookrightarrow$ si $P_c(t_{obs}) > \alpha$, c'est que t_{obs} est dans la région d'acceptation de (H_0)

$\hookrightarrow$ dessin !

p-value

Définition

Soit la fonction test $\varphi(x;\alpha)$ associée à la région de rejet $\mathcal{R}_\alpha$ de niveau α. La p-value est définie par

$$P_c(t_{obs}) = \inf\{\alpha \in [0,1]; \varphi(x_{obs};\alpha) = 1\}$$

C'est une variable aléatoire qui qualifie "non-adéquation" des données avec (H_0)

Dans un test de niveau α, (H_0) est rejetée si $\alpha >$ p-value, conservée si $\alpha <$ p-value :

- si $0.05 >$ p-value > 0.01, le test est significatif,
- si $0.01 >$ p-value > 0.001, le test est très significatif,
- si $0.001 >$ p-value, le test est hautement significatif.

p-value : cas d'une région de rejet bilatère

Exemple : test de Student de $\mu = \mu_0$ contre $\mu \neq \mu_0$,

- rejet : $\mathcal{R} = \{x; |T(x)| > qt(1-\alpha/2, n-1)\}$
- niveau : $\mathbb{P}_{H_0}(T(X) \in \mathcal{R}) = \alpha$
- valeur observée de la stat de test : $t_{obs} = T(x_{obs})$
- p-value : $P_c(t_{obs}) = \mathbb{P}_{H_0}(|T(X)| > t_{obs})$

  $\hookrightarrow$ si t_{obs} est supérieur à la médiane de T :

$$P_c(t_{obs}) = 2\,\mathbb{P}_{H_0}(T(X) > t_{obs})$$

  $\hookrightarrow$ si t_{obs} est inférieur à la médiane de T :

$$P_c(t_{obs}) = 2\,\mathbb{P}_{H_0}(T(X) < t_{obs})$$

- dans le cas de l'exemple introductif ($t_{obs} = 2$, $n = 12$)

$$P_c(t_{obs}) = 2(1 - \mathrm{F}^{\mathcal{T}(11)}(t_{obs} = 2) = 0.07$$

p-value

Attention :
p-value= proba d'observer sous (H_0) des valeurs au moins aussi "extrêmes" que celle qui a été observée sur l'échantillon x

$$\neq$$

proba que H_0 soit vraie sachant qu'on a observé x

Il n'est pas possible de calculer cette dernière à partir de la seule connaissance de la p-value.

Utiliser un test existant

La p-value permet de prendre une décision à partir du résultat d'un test sans utiliser les tables

Exemple : Test de Shapiro-Wilks à partir d'un n échantillon

$$(H_0) : X_i \sim \mathcal{N} \text{ contre } (H_1) : X_i \not\sim \mathcal{N}$$

- La statistique de test est $W = \frac{\left(\sum_i a_i X_{[i]}\right)^2}{\sum_i (X_i - \bar{X})^2}$ où $X_{[i]}$ est la i-ème statistique d'ordre, a_i les coordonnées du vecteur $a = \frac{m' V^{-1}}{\sqrt{m' V^{-1} V^{-1} m}}$ et $m = (m_1, \ldots, m_n)$ est le vecteur des espérances des statistiques d'ordre d'un échantillon i.i.d. $\mathcal{N}(0,1)$
- La région de rejet est unilatérale à gauche
- Les logiciels calculent la pvalue

```
> shapiro.test(x)
Shapiro-Wilk normality test
data:  x
W = 0.9284, p-value = 0.4323
```

Propriété des tests

- Le test est sans biais si $1 - \beta(\theta) = \pi(\theta) > \alpha$ pour tout $\theta \in \Theta_1$
- Le test est consistant (ou convergent) si la suite des puissance $\pi_n(\theta)$ tend vers 1 quand n tend vers l'infini
- A niveau fixé a priori, on choisit un test de puissance maximale

Définition

Un test est uniformément plus puissant (UPP) si, quelle que soit la valeur de θ, sa puissance $\pi(\theta)$ est supérieure à la puissance de tout autre test de niveau α.

Optimalité

Le cadre de la théorie de Neyman-Pearson permet de construire des tests les plus puissants parmi les tests de niveau fixé

Théorème (Neyman-Pearson)

Soit $L(\theta; x)$ la vraisemblance des observations. La région critique optimale (de NP) du test de $\theta = \theta_0$ vs $\theta = \theta_1$ au niveau α est définie par

$$\mathcal{R}_\alpha^{opt} = \left\{ x \in \mathbb{R}^n ; \frac{L(\theta_1; x)}{L(\theta_0; x)} > k_\alpha \right\}$$

Traduction Tout test de niveau inférieur ou égal à α est de puissance inférieure à celle du test de NP

Rem : $L(\theta_1; X)/L(\theta_0; X)$ est bien une statistique de test car θ_1 et θ_0 sont donnés.

Preuve (cas continu)

Soit $\mathcal{R}_\alpha$ une région de rejet qcq de niveau α. Supposons qu'il existe k_α définissant la région de rejet $\mathcal{R}_\alpha^{opt}$ de NP de risque α.

- Puissance associée au test de région de rejet $\mathcal{R}_\alpha$

$$\pi(\mathcal{R}_\alpha) = \underbrace{\int_{\mathcal{R}_\alpha \setminus \mathcal{R}_\alpha^{opt}} L(\theta_1; x)dx}_{\leq k_\alpha \mathbb{P}_{\theta_0}(\mathcal{R}_\alpha \setminus \mathcal{R}_\alpha^{opt})} + \int_{\mathcal{R}_\alpha^{opt} \cap \mathcal{R}_\alpha} L(\theta_1; x)dx$$

- Puissance associée au test de région de rejet $\mathcal{R}_\alpha^{opt}$

$$\pi(\mathcal{R}_\alpha^{opt}) = \underbrace{\int_{\mathcal{R}_\alpha^{opt} \setminus \mathcal{R}_\alpha} L(\theta_1; x)dx}_{k_\alpha \mathbb{P}_{\theta_0}(\mathcal{R}_\alpha^{opt} \setminus \mathcal{R}_\alpha) <} + \int_{\mathcal{R}_\alpha^{opt} \cap \mathcal{R}_\alpha} L(\theta_0; x)dx$$

- niveau α :

$$\mathbb{P}_{\theta_0}(\mathcal{R}_\alpha \setminus \mathcal{R}_\alpha^{opt}) + \mathbb{P}_{\theta_0}(\mathcal{R}_\alpha \cap \mathcal{R}_{opt}) \leq \alpha = \mathbb{P}_{\theta_0}(\mathcal{R}_\alpha^{opt} \setminus \mathcal{R}_\alpha) + \mathbb{P}_{\theta_0}(\mathcal{R}_\alpha \cap \mathcal{R}_\alpha^{opt})$$

- D'où $\pi(\mathcal{R}_\alpha) < \pi(\mathcal{R}_\alpha^{opt})$
- Existence de k_α : TVI

Tests hyp. simple contre hyp. composite

La puissance est une fonction de θ

- $\theta = \theta_0$ contre $\theta > \theta_0$ est UPP pour tester l'espérance dans un modèle gaussien
- $\theta = \theta_0$ contre $\theta \neq \theta_0$ n'est pas UPP