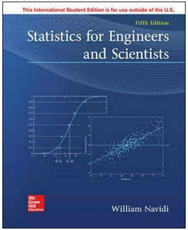
Introduction aux statistiques pour l'Ingénieur
Test Statistique

Stéphane Canu
asi.insa-rouen.fr/enseignants/~scanu
scanu@insa-rouen.fr

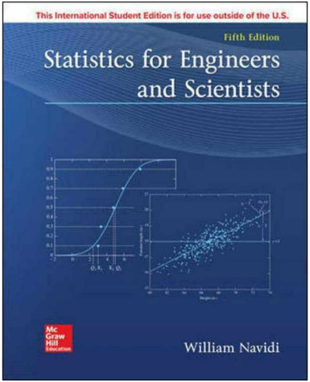


ITI 3, INSA Rouen Normandie, mai 2022

Lecture road map

1. La question du test

2. Formalisation des test d'hypothèse



https://moodle.insa-rouen.fr/course/view.php?id=93

La question du test

- Laquelle de ces deux variétés s'agit-il Varonne ou Touselles de Mayan ?
  -> seule la variété Varonne permet de faire du bon whisky

- Le taux de naissance des deux genres est égal (Laplace 1778)
  réponses : oui ou non

- La durée de vie de ce téléphone portable est inférieure à 1000 jours
  réponses : oui ou non

- Ce nouveau vaccin est efficace (plus efficace que l'ancien) ?
  actions : si oui on va le distribuer

- Ce candidat va gagner les élections
  réponses : oui, non, on ne peu pas dire

Attention aux raisonnements faux

- Laquelle de ces deux variétés s'agit-il Varonne ou Touselles de Mayan ?
  -> seule la variété Varonne permet de faire du bon whisky

- Le taux de naissance des deux genres est égal (Laplace 1778)
  non parce qu'on à observé 502 naissance de garçon et 498 de filles (et $502 \neq 498$)

- La durée de vie de ce téléphone portable est inférieure à 1000 jours
  non parce que le mien est tombé en pane au bout de 978 jours

- Ce nouveau vaccin est efficace (plus efficace que l'ancien) ?
  non parce qu'on a observé des personnes malades qui ont été vaccinées

- Ce candidat va gagner les élections
  oui parce que les sondage lui prédisent un score de 51 %

La question du test : deux hypothèses s'affrontent

La durée de vie de ce téléphone portable est inférieure à 1000 jours -> réponses : oui ou non
du point de vue du client : H0 : non H1 : oui On n'y croit pas
Ce nouveau vaccin est efficace (plus efficace que l'ancien) ? -> actions : si oui on va le distribuer H0 : non H1 : oui On ne va pas se lancer dans le fabrication d'une nouveau vaccin pour rien.
Ce candidat va gagner les élections -> réponses : oui, non, on ne peu pas dire
il faut faire deux tests
L'hypothèse nulle notée H0
L'hypothèse alternative notée H1 est l'hypothèse complémentaire à l'hypothèse nulle

La question, hypothèse et modèle statistique

si H_0 est vraie que peut-on s'attendre à observer

hypothèses, modèle statistique et p-valeur

p-valeur, est la probabilité pour un modèle statistique donné sous l'hypothèse nulle d'obtenir la même valeur ou une valeur encore plus extrême que celle observée.
La valeur p n'est pas la probabilité que l'hypothèse de test soit vraie. La valeur p indique dans quelle mesure les données sont conformes à l'hypothèse de test et à ses hypothèses (i.e. le modèle statistique sous-jacent)
erreur à priori : α
erreur a posteriori : p-valeur

Test = hypothèses + décision

Soient

- X une variable aléatoire.
- On considère des hypothèses qui portent sur la nature d'une variable aléatoire
- Deux hypothèses non symétriques : H_0 et H_1
- On cherche à tester H_0 à partir d'un Échantillon $(X_1, \ldots, X_n)$

Definition (Test statistique)

Tester deux hyothèses H_0 et H_1 c'est définir une fonction de décision D

$$D : \begin{array}{ccc} \mathbb{R}^n & \longmapsto & \{0, 1\} \\ (x_1, \ldots, x_n) & \longrightarrow & D(x_1, \ldots, x_n) \end{array}$$

si $D(x_1, \ldots, x_n) = 1$ on rejette H_0 (on décide H_1)

Cette fonction D réalise une partition de l'espace des échantillons

Hypothèses et décisions

Cette fonction de décision D réalise une partition de l'ensemble des échantillons

Definition (Régions de décision)

Région d'acceptation : l'ensemble $\overline{W}$ des réalisations $(x_1, \ldots, x_n)$ pour lesquelles on garde H_0 ($D(x_1, \ldots, x_n) = 0$)

Région critique : l'ensemble W des réalisations $(x_1, \ldots, x_n)$ pour lesquelles on rejette H_0 ($D(x_1, \ldots, x_n) = 1$)

On a deux hypothèses et deux décisions possibles

Mesures d'erreur et d'efficacité d'un test

4 situations :

	H_0	H_1
$D(x_1, \ldots, x_n) = 0$	$1 - \alpha$	β
$D(x_1, \ldots, x_n) = 1$	α	$1 - \beta$

2 erreurs

α probablité de rejeter H_0 ($D(x_1, \ldots, x_n) = 0$) alors que H_0 est vraie

$$\alpha = \mathbb{P}(D = 1|H_0) = \int_{(x_1,\ldots,x_n)\in W} L(x_1, \ldots, x_n, H_0)\, dx_1, \ldots, dx_n$$

β probabilité de garder H_0 ($D(x_1, \ldots, x_n) = 0$) alors que H_1 est vraie

$$\beta = \mathbb{P}(D = 0|H_1) = \int_{(x_1,\ldots,x_n)\in \overline{W}} L(x_1, \ldots, x_n, H_1)\, dx_1, \ldots, dx_n$$

Definition (puissance d'un test (d'une règle de décision))

$1 - \beta$ la probabilité de rejeter hypothèses nulle avec raison

Méthodes de construction des tests

Comment construire D ?

Cas des tests à hypothèses simples (HS = une hypothèse paramétrique qui caractérise entièrement la loi de probabilité)

Definition (Tests du rapport de vraisemblance)

$$W = \left\{ x_1, \dots, x_n \mid \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k \right\}$$

$$D(x_1, \dots, x_n) = \begin{cases} 0 & \text{si } \frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} \leq k \\ 1 & \text{sinon} \end{cases}$$

Exemple : tester deux valeurs pour l'espérance d'une loi normale

Comment trouver k ?

Pour un test de risque α on a

$$\alpha = \mathbb{P}(D = 1 | H_0) = \mathbb{P}\left(\frac{L(x_1, \dots, x_n, H_1)}{L(x_1, \dots, x_n, H_0)} > k \,\middle|\, H_0 \right)$$

La méthode de Neyman–Pearson

Definition (Le principe de Neyman–Pearson)

pour un α fixé, trouver la fonction de décision qui minimise β
(ou qui maximise $1-\beta$ la puissance du test)

$$
\begin{array}{ll}
\min\limits_{D} & \mathbb{P}(D=0|H_1) = \int_{(x_1,\ldots,x_n)\in\overline{W}} L(x_1,\ldots,x_n,H_1)\; dx_1,\ldots,dx_n \\
\text{avec} & \alpha = \mathbb{P}(D=1|H_0) = \int_{(x_1,\ldots,x_n)\in W} L(x_1,\ldots,x_n,H_0)\; dx_1,\ldots,dx_n
\end{array}
$$

La méthode de Neyman–Pearson

Definition (Le théorème de Neyman–Pearson)

Le test du rapport de vraisemblance

$$W = \left\{ x_1, \ldots, x_n \mid \frac{L(x_1, \ldots, x_n, H_1)}{L(x_1, \ldots, x_n, H_0)} > k \right\}$$

tel que k soit fixé de sorte que

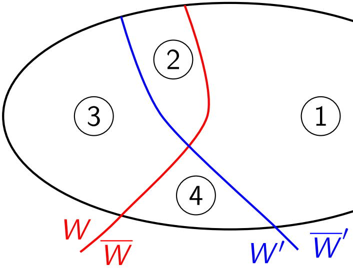
$$\alpha = \int_{(x_1,\ldots,x_n)\in W} L(x_1, \ldots, x_n, H_0)\, dx_1, \ldots, dx_n$$

est optimal au sens du principe de Neyman–Pearson (parmi tous les tests de risque α, c'est celui de puissance maximal).

Preuve : soit W' un autre test de risque α on a donc

$$\alpha = \int_W L(x_1, \ldots, x_n, H_0)\, dx_1, \ldots, dx_n = \int_{W'} L(x_1, \ldots, x_n, H_0)\, dx_1, \ldots, dx_n$$

Comparaison de deux fonctions de décision



Zone 1 $\overline{W}$ et $\overline{W}' = \dfrac{L(x_1, \ldots, x_n, H_1)}{L(x_1, \ldots, x_n, H_0)} \leq k$ et $\overline{W}'$

Zone 2 W et $\overline{W}' = \dfrac{L(x_1, \ldots, x_n, H_1)}{L(x_1, \ldots, x_n, H_0)} > k$ et $\overline{W}'$

Zone 3 W et $W' = \dfrac{L(x_1, \ldots, x_n, H_1)}{L(x_1, \ldots, x_n, H_0)} > k$ et W'

Zone 4 $\overline{W}$ et $W' = \dfrac{L(x_1, \ldots, x_n, H_1)}{L(x_1, \ldots, x_n, H_0)} \leq k$ et W'

Comparaison de deux fonctions de décision

W et W' sont tout les deux de même risque α :

$$\begin{aligned}
\alpha &= \int_W L(x_1,\ldots,x_n,H_0)\, dx_1,\ldots,dx_n &&= \int_{W'} L(x_1,\ldots,x_n,H_0)\, dx_1,\ldots,dx_n \\
&= \int_{2} L(0)\, dx + \int_{3} L(0)\, dx &&= \int_{3} L(0)\, dx + \int_{4} L(0)\, dx \\
&= \int_{2 = \frac{L(\mathbf{1})}{L(\mathbf{0})} > k \cap \overline{W}'} L(0)\, dx &&= \int_{4 = \frac{L(\mathbf{1})}{L(\mathbf{0})} \leq k \cap W'} L(0)\, dx
\end{aligned}$$

et

$$\begin{aligned}
\beta_W - \beta_{W'} &= \int_{\overline{W}} L(x_1,\ldots,x_n,H_1)\, dx_1,\ldots,dx_n - \int_{\overline{W}'} L(x_1,\ldots,x_n,H_1)\, dx_1,\ldots,d \\
&= \int_{1} L(1)\, dx + \int_{4} L(1)\, dx - \int_{1} L(1)\, dx - \int_{2} L(1)\, dx \\
&= \int_{4 = \frac{L(\mathbf{1})}{L(\mathbf{0})} \leq k \cap W'} \frac{L(1)}{L(0)} L(0)\, dx - \int_{2 = \frac{L(\mathbf{1})}{L(\mathbf{0})} > k \cap \overline{W}'} \frac{L(1)}{L(0)} L(0)\, dx \leq 0
\end{aligned}$$