

# Statistiques inférentielles

Pierre-Henri WUILLEMIN

Licence d'Informatique – Université Paris 6

## Les tests : introduction

Grâce aux estimateurs et aux intervalles de confiance, en statistique, on se pose souvent des questions sur la valeur des paramètres  $p, \mu, \sigma^2, \dots$  et il n'est pas rare que l'on ait des décisions à prendre concernant ces valeurs. Les tests d'hypothèses sont des outils pour répondre à ce type de question.

### ➡ Définition

Un **test d'hypothèse** est une **règle de décision** permettant de déterminer laquelle parmi deux hypothèses concernant la valeur d'un paramètre  $(p, \mu, \sigma^2, \dots)$  est la plus plausible.

La première étape dans la construction d'un test d'hypothèse, et peut-être la plus compliquée, consiste à identifier les deux hypothèses et à les formuler dans le langage statistique.

Les deux hypothèses à confronter seront toujours notées :

- $H_0$  : hypothèse nulle et
- $H_1$  : contre-hypothèse

Ces deux hypothèses doivent impérativement être mutuellement exclusives.

En principe,  $H_0$  est l'hypothèse que l'on essaye de vérifier.



## Les tests

### problématique

Soit  $X$  suivant une loi  $P_\theta$  sur  $\mathcal{X}$ , paramétrée par  $\theta \in \Theta$ . On dispose d'un échantillon  $X_1, \dots, X_n$ , toutes i.i.d. de loi  $P_\theta$ .

Soit une partition de  $\Theta = \theta_0 \cup \theta_1$ . Il s'agit de tester, sur l'échantillon, les 2 hypothèses :

$$H_0 : \theta \in \theta_0 \quad H_1 : \theta \in \theta_1$$

### Exemple

Dans une assemblée de 100 personnes, on demande à chacun de donner un chiffre au hasard compris entre 0 et 9. On note  $x_i \in \{0, \dots, 9\}$  le chiffre donné par l'individu  $i$  et  $n_j$  le nombre d'individus ayant donné le chiffre  $j$ . Les résultats (c'est à dire l'ensemble des  $(j, n_j)$  où  $j = 0, \dots, 9$ ) sont les suivants :

$(0, 10), (1, 8), (2, 9), (3, 14), (4, 8), (5, 9), (6, 11), (7, 9), (8, 12), (9, 10)$

Peut-on considérer que ces chiffres ont été effectivement donnés au hasard, au sens où les  $x_i$  sont des réalisations de variables aléatoires i.i.d. distribuées selon une loi uniforme sur  $\{0, \dots, 9\}$  ?

Il s'agit donc de tester :

$$H_0 : X \text{ uniforme sur } \{0, \dots, 9\} \quad H_1 : \text{non}$$



# Tests d'hypothèses en statistique classique

## Hypothèses

- $\Theta$  = ensemble des valeurs du paramètre  $\theta$
- $\Theta$  partitionné en  $\Theta_0$  et  $\Theta_1$
- *hypothèses* = assertions  $H_0 = " \theta \in \Theta_0 "$  et  $H_1 = " \theta \in \Theta_1 "$
- $H_0$  = hypothèse nulle,  $H_1$  = contre-hypothèse
- hypothèse  $H_i$  est simple si  $\Theta_i$  est un singleton ; sinon elle est *multiple*
- test unilatéral = valeurs dans  $\Theta_1$  toutes soit plus grandes, soit plus petites, que celles dans  $\Theta_0$  ; sinon test bilatéral

	hypothèse	test
$H_0 : \mu = 4$ $H_1 : \mu = 6$	simple simple	unilatéral
$H_0 : \mu = 4$ $H_1 : \mu > 4$	simple composée	test unilatéral
$H_0 : \mu = 4$ $H_1 : \mu \neq 4$	simple composée	test bilatéral
$H_0 : \mu = 4$ $H_1 : \mu > 3$	simple composée	formulation incorrecte : les hypothèses ne sont pas mutuellement exclusives



Statistiques inférentielles

4 / 19

## Exemples pratiques d'hypothèses

### Vin

Une association de consommateurs examine un échantillon de 100 bouteilles de Bordeaux afin de déterminer si la quantité de vin est bien égale à 75cl

- paramètre  $\theta$  étudié =  $\mu = E(X)$
- $X$  = quantité de vin dans les bouteilles
- rôle de l'association  $\implies H_0 : \mu = 75\text{cl}$  et  $H_1 : \mu < 75\text{cl}$

### Chômage

Enquête, sur un échantillon de 400 individus de la population active, pour savoir si le taux de chômage, qui était de 10% le mois dernier, s'est modifié

- paramètre étudié =  $p$ , la proportion de chômeurs
- $H_0 : p = 10\%$  et  $H_1 : p \neq 10\%$



Statistiques inférentielles

5 / 19

## Règle de décision

- La règle de décision du test est fondée sur les résultats de l'échantillonnage.
- Les résultats de l'échantillonnage sont examinés **après** la formulation des hypothèses, et non avant.
- Les valeurs du paramètre sous les différentes hypothèses **ne doivent pas** être fixées à partir du résultat observé à partir de l'échantillon.

- Construire la règle de décision, c'est déterminer quelles sont les valeurs qu'il est peu probable que le paramètre étudié (par exemple  $\bar{x}$ ) prenne dans l'échantillon si l'hypothèse  $H_0$  est vraie.
- Il faut examiner la distribution de l'estimateur du paramètre dans l'échantillon lorsque  $H_0$  est vraie et déterminer une **région critique**, ou **région de rejet** de  $H_0$ , telle que si la valeur prise par l'estimateur est dans cette région, il est peu probable que  $H_0$  soit vraie.

- La région critique doit tenir compte de la forme de la contre-hypothèse pour que le rejet de  $H_0$  signifie que  $H_1$  est un choix plausible.



Statistiques inférentielles

6 / 19

## Régions critiques

Hypothèses	Règle de décision
$H_0 : \mu = \mu_0$ $H_1 : \mu > \mu_0$	«rejeter $H_0$ si $\bar{x} > c$ », où $c$ est un nombre plus grand que $\mu_0$
$H_0 : \mu = \mu_0$ $H_1 : \mu < \mu_0$	«rejeter $H_0$ si $\bar{x} < c$ », où $c$ est un nombre plus petit que $\mu_0$
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$	«rejeter $H_0$ si $\bar{x} < c_1$ ou $c_2 < \bar{x}$ », où $c_1$ et $c_2$ sont des nombres respectivement plus petit et plus grand que $\mu_0$ , et également éloignés de celui-ci



## Erreurs dans les décisions

Décision prise \ Réalité	$H_0$ est vraie	$H_1$ est vraie
$H_0$ est rejetée	mauvaise décision : erreur de type I	bonne décision
$H_0$ n'est pas rejetée	bonne décision	mauvaise décision : erreur de type II

$\alpha$  = risque de première espèce

- = probabilité de réaliser une erreur de type I
- = probabilité de rejeter  $H_0$  sachant que  $H_0$  est vraie
- =  $P(\text{rejeter } H_0 | H_0 \text{ est vraie})$ ,

$\beta$  = risque de deuxième espèce

- = probabilité de réaliser une erreur de type II
- = probabilité de rejeter  $H_1$  sachant que  $H_1$  est vraie
- =  $P(\text{rejeter } H_1 | H_1 \text{ est vraie})$ .



## Exemple de calcul de $\alpha$ (1/2)

### exemple

- échantillon de taille 25
- paramètre estimé :  $\mu$  d'une variable  $X \sim \mathcal{N}(\mu; 100)$
- hypothèses :  $H_0 : \mu = 10$      $H_1 : \mu > 10$

Sous  $H_0$  :  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 10}{10/5} = \frac{\bar{X} - 10}{2} \sim \mathcal{N}(0; 1)$

Sous  $H_0$  : peu probable que  $\bar{X}$  éloignée de plus de 2 écart-types de  $\mu$  (4,56% de chance)

$\Rightarrow$  peu probable que  $\bar{X} < 6$  ou  $\bar{X} > 14$

$\Rightarrow$  région critique pourrait être «rejeter  $H_0$  si  $\bar{x} > 14$ »



## Exemple de calcul de $\alpha$ (2/2)

### Exemple

- échantillon de taille 25
- paramètre estimé :  $\mu$  d'une variable  $X \sim \mathcal{N}(\mu; 100)$
- hypothèses :  $H_0 : \mu = 10$      $H_1 : \mu > 10$
- région critique : «rejeter  $H_0$  si  $\bar{x} > 14$ »

$$\begin{aligned}\alpha &= P(\text{rejeter } H_0 | H_0 \text{ est vraie}) \\ &= P(\bar{X} > 14 | \mu = 10) \\ &= P\left(\frac{\bar{X} - 10}{2} > \frac{14 - 10}{2} \middle| \mu = 10\right) \\ &= P\left(\frac{\bar{X} - 10}{2} > 2\right) = 0,0228\end{aligned}$$



en principe  $\alpha$  est fixé et on cherche la région critique



## Puissance du test

$$\alpha = P(\text{rejeter } H_0 | H_0 \text{ est vraie})$$

$$\beta = P(\text{rejeter } H_1 | H_1 \text{ est vraie})$$

$\alpha$  et  $\beta$  varient en sens inverse l'un de l'autre

$\Rightarrow$  test = compromis entre les deux risques

$H_0$  = hypothèse privilégiée, vérifiée jusqu'à présent et que l'on n'aimerait pas abandonner à tort

$\Rightarrow$  on fixe un *seuil*  $\alpha_0$  :

- $\alpha$  doit être  $\leq \alpha_0$
- test minimisant  $\beta$  sous cette contrainte
- $\min \beta = \max 1 - \beta$
- $1 - \beta$  = puissance du test



## Exemple de calcul de $\beta$ (1/2)

### Exemple

- échantillon de taille 25
- paramètre estimé :  $\mu$  d'une variable  $X \sim \mathcal{N}(\mu; 100)$
- hypothèses :  $H_0 : \mu = 10$      $H_1 : \mu > 10$
- région critique : «rejeter  $H_0$  si  $\bar{x} > 14$ »

sous  $H_1$  : plusieurs valeurs de  $\mu$  sont possibles

$\Rightarrow$  courbe de puissance du test en fonction de  $\mu$

Supposons que  $\mu = 11$  :

$$\mu = 11 \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - 11}{2} \sim \mathcal{N}(0; 1)$$



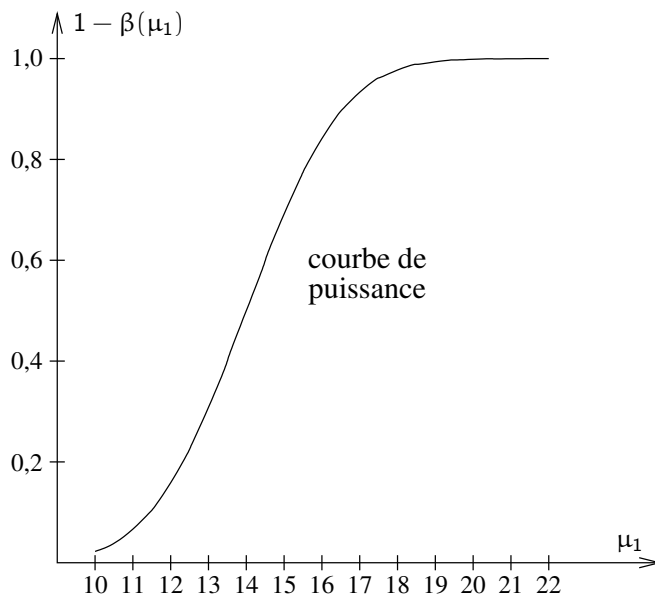
## Exemple de calcul de $\beta$ (2/2)

$$\begin{aligned}
 1 - \beta(11) &= P(\text{rejeter } H_0 | H_1 : \mu = 11 \text{ est vraie}) \\
 &= P(\bar{X} > 14 | \mu = 11) \\
 &= P\left(\frac{\bar{X} - 11}{2} > \frac{14 - 11}{2} | \mu = 11\right) \\
 &= P\left(\frac{\bar{X} - 11}{2} > 1,5\right) = 0,0668
 \end{aligned}$$

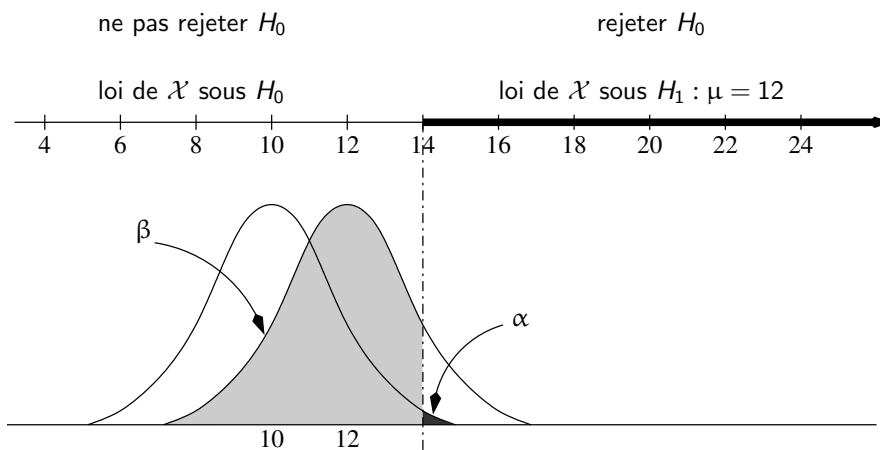
$\mu_1$	$z_1 = \frac{14 - \mu_1}{2}$	$1 - \beta(\mu_1) = P(Z > z_1)$	$\beta(\mu_1)$
10	2,0	0,0228	0,9772
11	1,5	0,0668	0,9332
12	1,0	0,1587	0,8413
13	0,5	0,3085	0,6915
14	0,0	0,5000	0,5000
15	-0,5	0,6915	0,3085
16	-1,0	0,8413	0,1587
17	-1,5	0,9332	0,0668



## Courbe de puissance du test



## Interprétation de $\alpha$ et $\beta$



## Rappel : vraisemblance

On se souvient que :

$$P(X | Y) = \frac{P(Y | X) \cdot P(X)}{P(Y)}$$

Ou encore :

$$P(X | Y) \propto P(Y | X) \cdot P(X)$$

En notant  $\theta$  le paramètre que l'on veut estimer et  $d$  l'observation que l'on fait :

### ➡ Définition (Vraisemblance)

$$P(\theta | d) \propto P(d | \theta) \cdot P(\theta)$$

On nomme :

- $P(\theta)$  la probabilité **a priori** sur  $\theta$ .
- $P(\theta | d)$  la probabilité **a posteriori** sur  $\theta$ .
- $P(d | \theta) = L(d, \theta) = L(\theta : d)$  la **vraisemblance**.



## Maximisation de la vraisemblance (MLE)

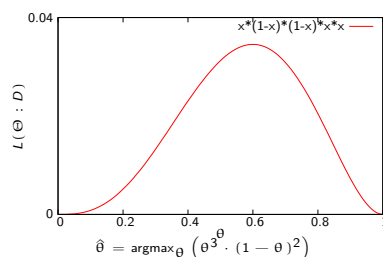
Soit une variable binaire  $X$ . Avec  $\theta = P(X = 1)$  :

$$\Theta = \{\theta, 1 - \theta\}$$

$$D = (1, 0, 0, 1, 1)$$

$$L(\Theta : D) = P(D | \Theta) = \prod_m P(X = d_m | \Theta)$$

Ici :  $L(\Theta : D) = \theta \cdot (1 - \theta) \cdot (1 - \theta) \cdot \theta \cdot \theta$ .



### Estimation de la probabilité par la fréquence

Pour des données qui font apparaître  $p$  fois 1 et  $q = n - p$  fois 0 :

$$L(\Theta : D) = \theta^p \cdot (1 - \theta)^q$$

D'où :

$$\frac{d(\Theta:D)}{d\theta} = p\theta^{p-1}(1-\theta)^q - q(1-\theta)^{q-1}\theta^p$$

$$\frac{d(\Theta:D)}{d\theta} = 0 \iff p(1-\theta) - q\theta = 0$$

finalement :

$$\hat{\theta} = \frac{p}{p+q}$$



## Évaluation des risques pour des tests simples

Cas :  $\Theta_0 = \{\theta_0\}$   $\Theta_1 = \{\theta_1\}$

$$\alpha = P(\text{rejeter } H_0 | H_0 \text{ est vraie})$$

$$= P(X \in W | \theta = \theta_0)$$

$$= \int_W L(x, \theta) dx$$

$$\beta = P(\text{rejeter } H_1 | H_1 \text{ est vraie})$$

$$= P(X \in A | \theta = \theta_1)$$

$$= \int_A L(x, \theta) dx$$



# Lemme de Neyman-Pearson

cas :  $\Theta_0 = \{\theta_0\}$     $\Theta_1 = \{\theta_1\}$

## *Lemme de Neyman-Pearson*

- il existe toujours un test (aléatoire) le plus puissant de seuil donné  $\alpha_0$
- c'est un test du rapport de

$$\begin{aligned} & \frac{L(x, \theta_0)}{L(x, \theta_1)} > k \Rightarrow x \in A \text{ (accepter } H_0) \\ \text{vraisemblance : } & \frac{L(x, \theta_0)}{L(x, \theta_1)} < k \Rightarrow x \in W \text{ (rejeter } H_0) \\ & \frac{L(x, \theta_0)}{L(x, \theta_1)} = k \Rightarrow \delta(x) = \rho \text{ (accepter } H_0 \text{ avec proba } 1 - \rho \\ & \hspace{15em} H_1 \text{ avec proba } \rho) \end{aligned}$$

- $k$  et  $\rho$  déterminés de façon unique par  $\alpha = \alpha_0$

