

SPLEX

Statistiques pour la classification et fouille de données en génomique

Tests d'hypothèse

Pierre-Henri WUILLEMIN

DEcision, Système Intelligent et Recherche opérationnelle
LIP6

pierre-henri.wuillemin@lip6.fr
<http://webia.lip6.fr/~phw/splex>

Les tests : introduction

► Définition

Un **test d'hypothèse** est une **règle de décision** permettant de déterminer laquelle parmi deux hypothèses concernant la valeur d'un paramètre (p, μ, σ^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_θ 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_θ .

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

- Θ = ensemble des valeurs du paramètre θ
- Θ partitionné en Θ_0 et Θ_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 Θ_i est un singleton ; sinon elle est *multiple*
- test unilatéral = valeurs dans Θ_1 toutes soit plus grandes, soit plus petites, que celles dans Θ_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



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.



Régions critiques

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 μ_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 μ_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 μ_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

α = risque de première espèce

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

β = 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 α (1/2)

exemple

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

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

Sous H_0 : peu probable que \bar{X} éloignée de plus de 2 écart-types de μ (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 α (2/2)

Exemple

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

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



en principe α est fixé et on cherche la région critique



Puissance du test

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

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

α et β varient en sens inverse l'un de l'autre

\Rightarrow test = compromis entre les deux risques

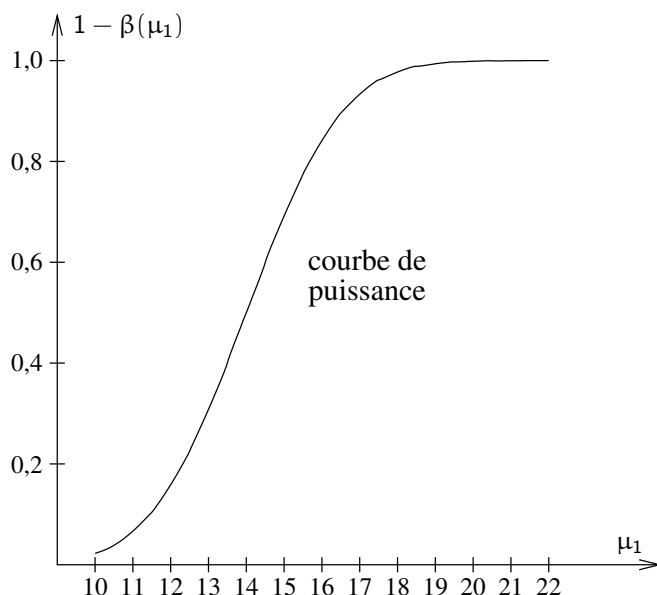
H_0 = hypothèse privilégiée, vérifiée jusqu'à présent et que l'on n'aimerait pas abandonner à tort

\Rightarrow on fixe un *seuil* α_0 :

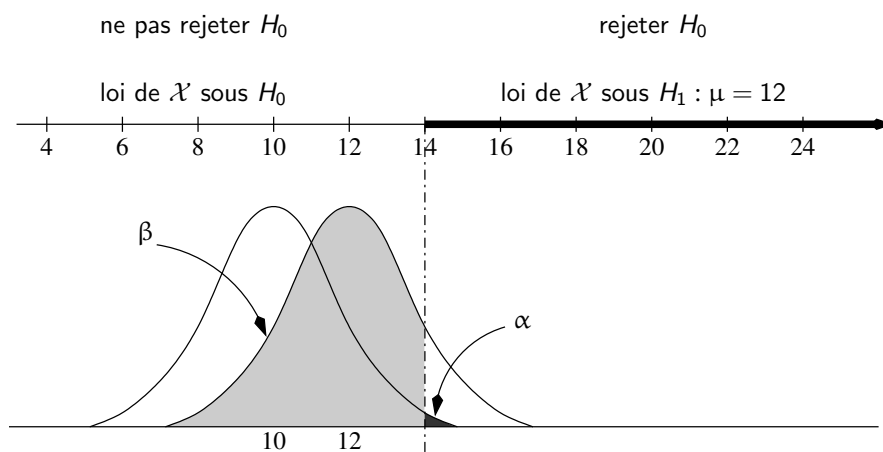
- α doit être $\leq \alpha_0$
- test minimisant β sous cette contrainte
- $\min \beta = \max 1 - \beta$
- $1 - \beta = \text{puissance du test}$



Courbe de puissance du test



Interprétation de α et β



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



Maximisation de la vraisemblance (MLE)

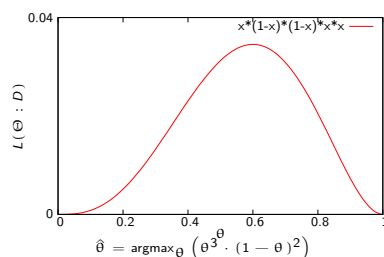
Soit une variable binaire X . Avec $\theta = P(X = 1)$:

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

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

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

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 :

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



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 α_0
- c'est un test du rapport de

$$\frac{L(x, \theta_0)}{L(x, \theta_1)} > k \Rightarrow x \in A \text{ (accepter } H_0)$$

$$\text{vraisemblance : } \frac{L(x, \theta_0)}{L(x, \theta_1)} < k \Rightarrow x \in W \text{ (rejeter } H_0)$$

$$\frac{L(x, \theta_0)}{L(x, \theta_1)} = k \Rightarrow \delta(x) = \rho \text{ (accepter } H_0 \text{ avec proba } 1 - \rho \\ H_1 \text{ avec proba } \rho)$$

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



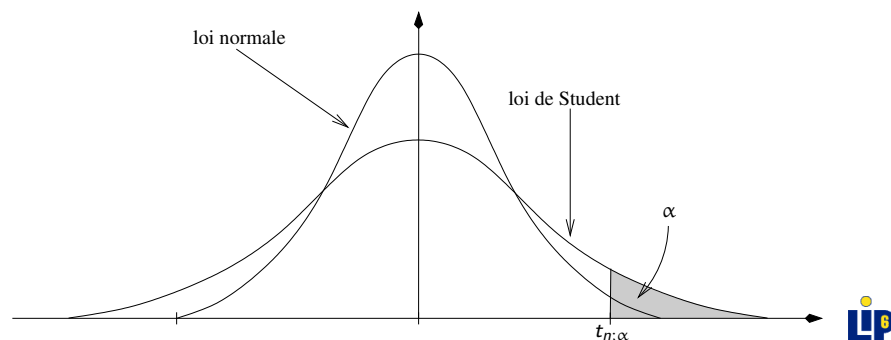
Loi de Student

Loi de Student

La loi de Student ne possède qu'un seul paramètre n (degré de liberté). La loi à n degrés est notée T_n .

- L'espérance d'une variable obéissant à une loi T_n est 0,
- Sa variance est $\frac{n}{n-2}$ pour $n > 2$.

. Plus n est grand, plus T_n se rapproche de $\mathcal{N}(0; 1)$.



Utilisation de la loi de Student : Intervalles de confiance pour μ

Soit un échantillon $(x_i)_{i \leq n}$ et \bar{x} sa moyenne et s^2 sa variance.

Intervalles de confiance pour μ de niveau de confiance $1 - \alpha$

Situation	Loi utilisée	Bornes de l'intervalle
σ^2 connue $X \sim$ loi normale ou n grand (> 75)	$\frac{\mathcal{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0; 1)$	$\bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$
σ^2 inconnue n très grand (> 75)	$\frac{\mathcal{X} - \mu}{S/\sqrt{n}} \sim \mathcal{N}(0; 1)$	$\bar{x} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}$
σ^2 inconnue $X \sim$ loi normale	$\frac{\mathcal{X} - \mu}{S/\sqrt{n}} \sim T_{n-1}$	$\bar{x} \pm t_{n-1; \alpha/2} \times \frac{s}{\sqrt{n}}$



Utilisation de la loi de Student : comparaison d'échantillons

Soit deux échantillons :

- n_1, \bar{X}_1, s_1^2 issu d'une population μ_1, σ_1^2 ,
- n_2, \bar{X}_2, s_2^2 issu d'une population μ_2, σ_2^2 ,

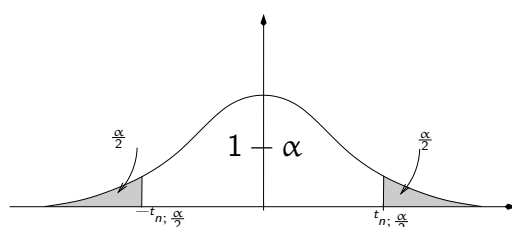
Comparaison de μ_1 et de μ_2

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Règle de décision

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim T_{n_1 + n_2 - 2} \quad \text{avec } s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$



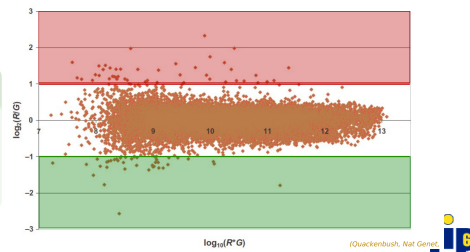
Application : Analyse d'expression différentielle

Analyse différentielle

- L'objectif est de trouver une liste de gènes différentiellement exprimés sur des résultats de puces avec plus de 22000 gènes.
- L'expression d'un gène est résumée par la **moyenne des niveaux d'expression dans le groupe**. Il s'agit donc bien de comparaison de moyenne pour l'analyse différentielle.
- Pour chaque gène i , $H_0^i : \mu_1^i = \mu_2^i$
- Classiquement : niveau de risque 5%

fold-change

Règle arbitraire : si $|\log_2(\frac{\bar{x}_1}{\bar{x}_2})| > 1$ alors le gène i est exprimée différentiellement.



SAM : Significance Analysis of MicroArrays

- Utilisation d'un test de Student pour l'analyse différentielle,
- Problème : analyse multiples (22000 gènes)
- Solution : *SAM calculates a test statistic for relative difference in gene expression based on permutation analysis of expression data and calculates a false discovery rate (FDR).*

SAM is distributed by Stanford University in an R-package.