MSV601 - Rappels sur les tests

Marie-Luce Taupin
marie-luce.taupin@univ-evry.fr

Laboratoire LaMME, Université d'Evry val d'Essonne
http://www.math-evry.cnrs.fr/members/mtaupin/welcome

2018-2019

Probabilité

Définition axiomatique (Kolmogorov-1933)

Une **probabilité** est une application $\mathbb{P} : \Omega \to [0, 1]$ telle que :

- pour tout $A \in \Omega$, on a $\mathbb{P}(A) \geq 0$,
- $\mathbb{P}(\Omega) = 1$,
- Si $A \cap B = \emptyset$, alors $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Remarque

Une probabilité est une mesure.

Définition

On appele **espace probabilisé** le triplet $(\Omega, \mathcal{P}(\Omega), \mathbb{P})$

Loi de probabilité

Définitions

- Soit X une variable aléatoire **discrète** telle que $\Omega_X = x_1, ..., x_N$. La loi de probabilité de X est définie/caractérisée par sa **fonction de probabilité** qui donne, pour tout $i \in 1, ..., N$

$$p_i = \mathbb{P}(X = x_i)$$

- Soit X une variable aléatoire **continue**. On appelle **densité** de probabilité la fonction $f(x)$ définie par :

$$f(x) = \lim_{\delta \to 0} \frac{\mathbb{P}(X \in [x; x + \delta])}{\delta}$$

Remarque : Pour δ proche de 0, $f(x)dx \approx \mathbb{P}(X \in [x; x + \delta])$

Fonction de répartition

Définition

On appelle **fonction de répartition** la fonction F définie par :

$$\begin{array}{rccl} F & : & D_X & \longrightarrow [0,1] \\ & & x_i & \longmapsto \mathbb{P}(X \leq x_i) \end{array}$$

Espérance

Définition

L'**espérance** $E(X)$ d'une variable aléatoire X est définie par :

$$E(X) = \sum_{i=1}^{N} x_i p_i \qquad \text{(cas discret)}$$

$$E(X) = \int_{-\infty}^{+\infty} x f(x) dx \qquad \text{(cas continu)}$$

Remarques

- L'espérance ne fait pas nécessairement partie de D_x
- L'espérance n'est pas toujours définie

Variance et écart-type

Définitions

- La **variance** $Var(X)$ d'une variable aléatoire X est définie par :

$$Var(X) = E\left[(X - E(X))^2\right]$$

- L'**écart-type** est défini par

$$\sigma = \sqrt{Var(X)}$$

Covariance

Définition

Pour un couple de variables aléatoires (X, Y), la **covariance** est définie par :

$$\begin{aligned} Cov(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

Remarques

- $X \perp\!\!\!\perp Y \Rightarrow Cov(X, Y) = 0$ mais $Cov(X, Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$
- $Var(X) = Cov(X, X)$

Coefficient de corrélation

Définition

Pour un couple de variables aléatoires (X, Y), le **coefficient de corrélation** est défini par :

$$Cor(X, Y) = \rho_{X,Y} = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Propriété

$$\rho_{X,Y} \in [-1; 1]$$

Loi normale

Définition

- Une variable aléatoire X suit une loi normale (ou loi de Gauss-Laplace) de paramètres μ et σ^2 si sa densité f s'écrit :

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- On note : $X \sim N(\mu, \sigma^2)$
- $E(X) = \mu$ et $Var(X) = \sigma^2$

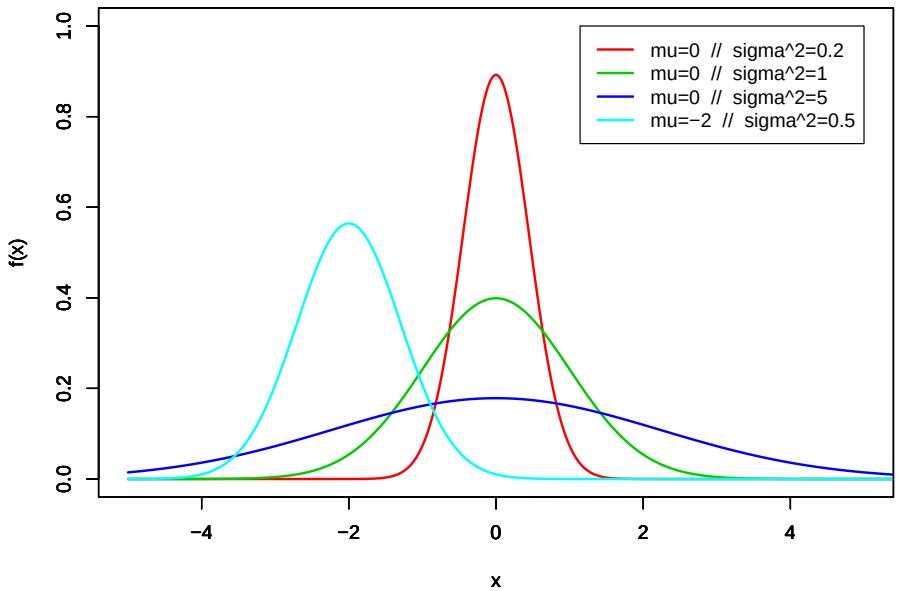
Remarque

Il n'existe pas de forme analytique de la fonction de répartition F.

Loi normale

Densité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



Loi normale centrée réduite

Propriété

Si $X \sim N(\mu, \sigma^2)$, alors $Y = \dfrac{X - \mu}{\sqrt{\sigma^2}} \sim N(0, 1)$

Notations

Par convention, on note ϕ la densité d'une $N(0, 1)$ et Φ sa fonction de répartition.

Loi normale centrée réduite

Propriétés

- ϕ est une fonction paire
- Le graphe de ϕ est symétrique par rapport à l'axe des ordonnées
- $\Phi(x) = 1 - \Phi(-x)$
- $\mathbb{P}(|X| \leq x) = \mathbb{P}(-x \leq X \leq x) = 2(\Phi(x) - 1/2)$
- $\mathbb{P}(|X| \geq x) = \mathbb{P}((X \leq -x) \cap (X \geq x)) = 2(1 - \Phi(x))$

Théorème de la limite centrale

Théorème

Soit $X_1, ..., X_n$ n v.a. indépendantes et identiquement distribuées d'éspérance μ :

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu) \xrightarrow[n\to\infty]{\mathcal{L}} N\left(0, \sigma^2\right)$$

ou aussi

$$Y = \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow[n\to\infty]{\mathcal{L}} N(0,1)$$

Remarque

Ce théorème est appelé théorème de la limite centrale (TLC) ou théorème de la limite centrée (TLC) ou théorème central limite (TCL)

Loi du χ^2

Définition

Soient $X_1, ..., X_n$ n variables aléatoires indépendantes et identiquement distribuées de loi normale centrée réduite.
La variable aléatoire $Y = X_1^2 + ... + X_n^2$ suit une loi continue appelée loi du χ^2 à n degrés de liberté :

$$Y = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$$

.

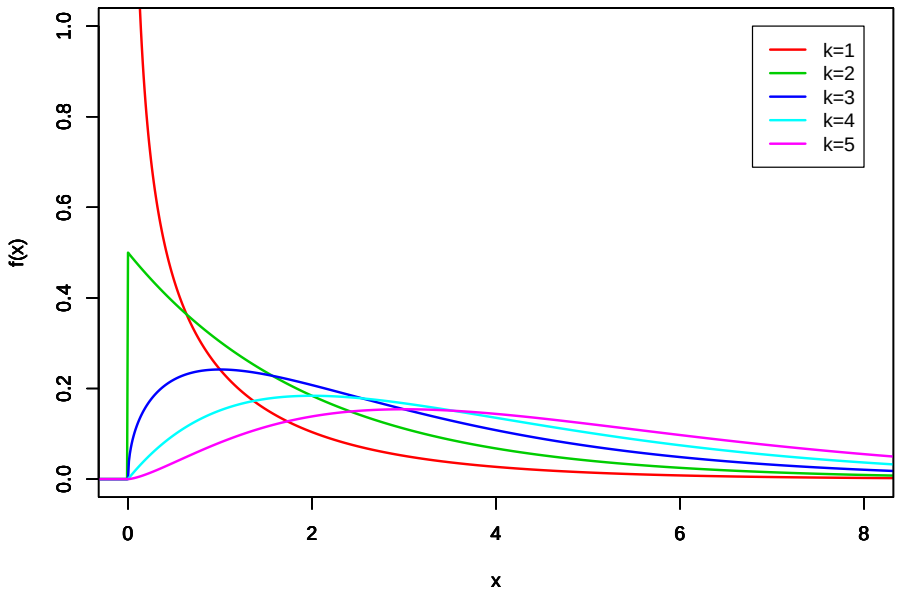
Propriétés

- Si $Y_1 \sim \chi_{n_1}^2$ et $Y_2 \sim \chi_{n_2}^2$ avec $Y_1 \perp\!\!\!\perp Y_2$, alors $Y = Y_1 + Y_2 \sim \chi_{n_1+n_2}^2$
- Si $Y \sim \chi_n^2$, alors $E(Y) = n$ et $Var(Y) = 2n$

Loi du χ^2

Densité

$$f(y) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n-2)/2} e^{-x/2}$$



Loi de Student

Définition

Soient X et Y deux variables aléatoires indépendantes telles que $X \sim N(0,1)$ et $Y \sim \chi^2_n$. La variable aléatoire $T = X/\sqrt{Y/n}$ suit une loi continue appelée loi de Student à n degrés de liberté :

$$T = \frac{X}{\sqrt{Y/n}} \sim t_n$$

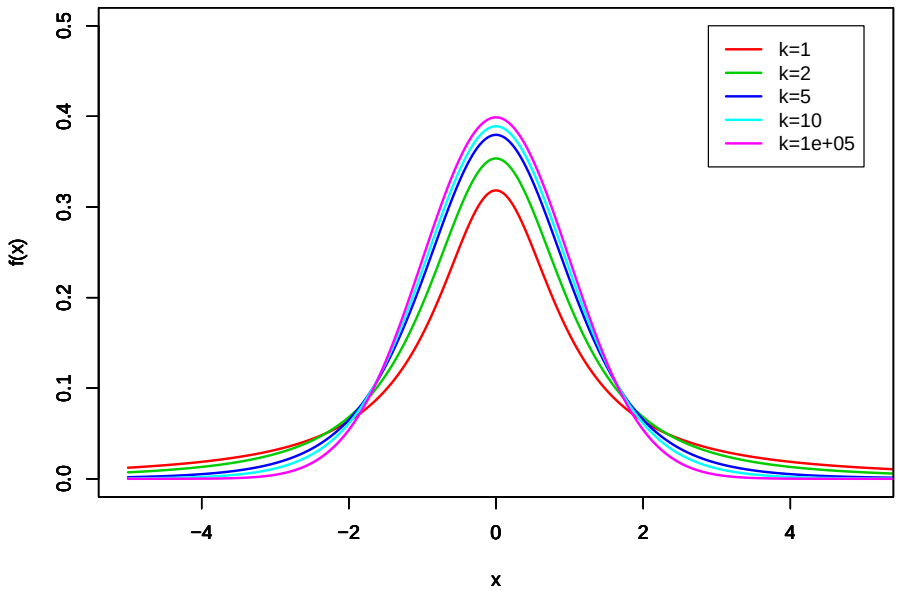
Propriétés

- $E(T) = 0$
- $Var(T) = \frac{n}{n-2}$ si $n > 2$

Loi de Student

Densité

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})(1+\frac{x^2}{n})^{\frac{n+1}{2}}}$$



Loi de Fisher

Définition

Soient Y_1 et Y_2 deux variables aléatoires indépendantes telles que $Y_1 \sim \chi^2_{n_1}$ et $Y_2 \sim \chi^2_{n_2}$. La variable aléatoire $Z = (Y_1/n_1)/(Y_2/n_2)$ suit une loi continue appelée loi de Fisher à n_1 et n_2 degrés de liberté :

$$Z = \frac{Y_1/n_1}{Y_2/n_2} \sim \mathcal{F}(n_1; n_2)$$

Remarques

- $Z_1 \sim F(n_2; n_1) \Rightarrow Z_2 = 1/Z_1 \sim \mathcal{F}(n_1; n_2)$
- $T \sim t_n \Rightarrow Z = T^2 \sim \mathcal{F}(1; n)$

Statistique

On considère $X_1, \cdots, X_n$ n variables indépendantes de te même loi P_θ où θ est inconnu, à estimer.

Définition

On appelle **statistique** toute fonction du n-échantillon $X_1, ..., X_n$:

$$\begin{array}{rcl} T & : & \mathbb{R}^n \longrightarrow \mathbb{R} \\ & & (X_1, ..., X_n) \longmapsto T(X_1, ..., X_n)) \end{array}$$

Remarque

T est aussi une variable aléatoire

Statistique

Exemples d'estimateurs

- Moyenne empirique (ou moyenne observée) : $\overline{X}_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Variance empirique (ou variance observée) :
  $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2$
- Variance empirique corrigée : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{n}{n-1}\hat{\sigma}^2$

Formule de Huygens

Théorème (de König-Huygens)

- $Var(X) = E\left((X - E(X))^2\right) = E(X^2) - E(X)^2$
- $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2$
- $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}x_i^2 - n\overline{x}^2$

Estimation

On distingue...

- Estimation ponctuelle
- Estimation par intervalles de confiance

Estimation ponctuelle

Définition

Un **estimateur ponctuel** est une statistique dont la réalisation (pour un échantillon donnée) constitue une estimation de l'un des paramètres θ de la distribution ou de l'une des fonctions permettant de la caractériser.

Notation

On note généralement $\hat{\theta}$ l'estimateur du paramètre θ.

Qualités d'un estimateur

Définitions

- On appelle **biais** d'un estimateur la quantité :

$$b(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- Un estimateur est dit **sans biais** si

$$b(\hat{\theta}) = 0, \text{ soit aussi si } \mathbb{E}(\widehat{\theta}) = \theta$$

- Un estimateur est dit **asymptotiquement sans biais** si

$$\lim_{n \to +\infty} b(\hat{\theta}) = 0$$

Qualités d'un estimateur

Définitions (suite)

- On appelle **erreur quadratique moyenne** la quantité :

$$EQM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [b(\hat{\theta})]^2$$

- Un estimateur est dit **consistant** (ou **convergent**) si :

$$\lim_{n \to \infty} EQM(\hat{\theta}) = 0$$

Remarque

Pour montrer qu'un estimateur sans biais est consistant, il suffit de montrer que $\lim_{n \to \infty} Var(\hat{\theta}) = 0$

Théorème de Cochran

Théorème

Soit $X_1, ..., X_n$ un n-échantillon de la variable aléatoire X où $X \sim N(\mu, \sigma^2)$.

- μ est estimé sans biais par la moyenne $\overline{X}$; $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- σ^2 est estimé sans biais par la variance empirique corrigée : $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$; $Q^2 = \sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \sim \sigma^2\chi^2_{n-1}$
- Q^2 et $\overline{X}$ sont indépendants.

Estimation par intervalles

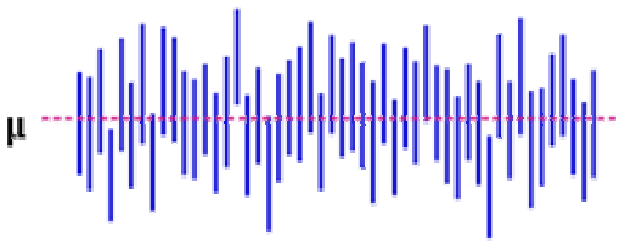
Définition

Un **intervalle de confiance** de niveau $1-\alpha$ du paramètre θ est un intervalle aléatoire $[A_n, B_n]$ tel que :

$$\mathbb{P}(\theta \in [A_n, B_n]) = 1 - \alpha$$

Remarque

Les extrémités l'intervalle sont aléatoires, dépendent de n et sont reliées à un estimateur de θ.



Estimation par intervalle

Exercice

Pour déterminer la teneur en potassium d'une solution, on effectue des dosages à l'aide d'une technique expérimentale donnée. On admet que le résultat d'un dosage est une variable aléatoire suivant une distribution normale $N(\mu, \sigma^2)$ dont l'espérance μ est la valeur que l'on cherche à déterminer, et dont l'ecart-type σ est de 1 mg/litre si l'on suppose que le protocole expérimental a été suivi scrupuleusement. Les résultats pour cinq dosages indépendants sont les suivants (en mg/litre) :

$$74.0,\ 71.6,\ 73.4,\ 74.3,\ 72.2$$

1. Déterminer à partir de ces mesures un intervalle de confiance pour μ de niveau 95%.
2. Recalculez l'intervalle de confiance en supposant que la variance est inconnue.

Estimation par intervalle

Famille gaussienne, variance connue

Soit $(X_1, ..., X_n)$ un n-échantillon de loi $N(\mu, \sigma^2)$. on suppose σ^2 connu. L'intervalle pour μ, de niveau de confiance $(1-\alpha)$:

$$IC_{(1-\alpha)} = \left[\overline{X_n} - q_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}; \overline{X_n} + q_{1-\alpha/2}\sqrt{\frac{\sigma^2}{n}}\right],$$

où

$$\mathbb{P}(\mathcal{N}(0,1) \leq q_{1-\alpha/2} \leq 1-\alpha/2.$$

Estimation par intervalle

Famille gaussienne, variance inconnue

Soit $(X_1, ..., X_n)$ un n-échantillon de loi $N(\mu, \sigma^2)$. on suppose σ^2 inconnu. L'intervalle pour μ, de niveau de confiance $(1-\alpha)$:

$$IC_{(1-\alpha)} = \left[\overline{X_n} - t_{1-\alpha/2}\sqrt{\frac{S_n^{*2}}{n}}; \overline{X_n} + t_{1-\alpha/2}\sqrt{\frac{S_n^{*2}}{n}}\right]$$

où

$$\mathbb{P}(\mathcal{S}_{n-1} \leq t_{1-\alpha/2} \leq 1-\alpha/2.$$

Tests d'hypothèses

Définitions

- Un **test statistique** est une procedure permettant de décider entre deux hypothèses au vu des observations.
- Une **hypothèse statistique** est un énoncé portant sur les caractéristiques d'une population (paramètre ou forme d'une distribution)

Démarche

1. Choisir les hypothèses à tester (H_0) contre (H_1)
2. Choisir une statistique de test, en fonction d'un estimateur du paramètre sur lequel porte le test
3. Déterminer la règle de décision (région de rejet Γ)
4. Calculer la statistique (et la p-valeur)
5. Conclure

Erreurs et risques associés

Résultats possibles

Décision / Réalité	Ne pas rejeter H_0 (conclure H_0)	Rejeter H_0 (conclure H_1)
H_0 vraie	OK	Erreur de type I
H_1 vraie	Erreur de type II	OK

Définitions

- **Risque de première espèce** : $\alpha = \mathbb{P}(\text{rejeter } H_0 | H_0 \text{ vraie})$
  (probabilité de commettre une erreur de type I)
- **Risque de seconde espece** : $\beta = \mathbb{P}(\text{ne pas rejeter } H_0 | H_1 \text{ vraie})$
  (probabilité de commettre une erreur de type II)
- **Puissance** : $\Pi = 1 - \beta = \mathbb{P}(\text{rejeter } H_0 | H_1 \text{ vraie})$
  (probabilité de prendre la bonne décision en rejetant H_0)

Erreurs et risques associés

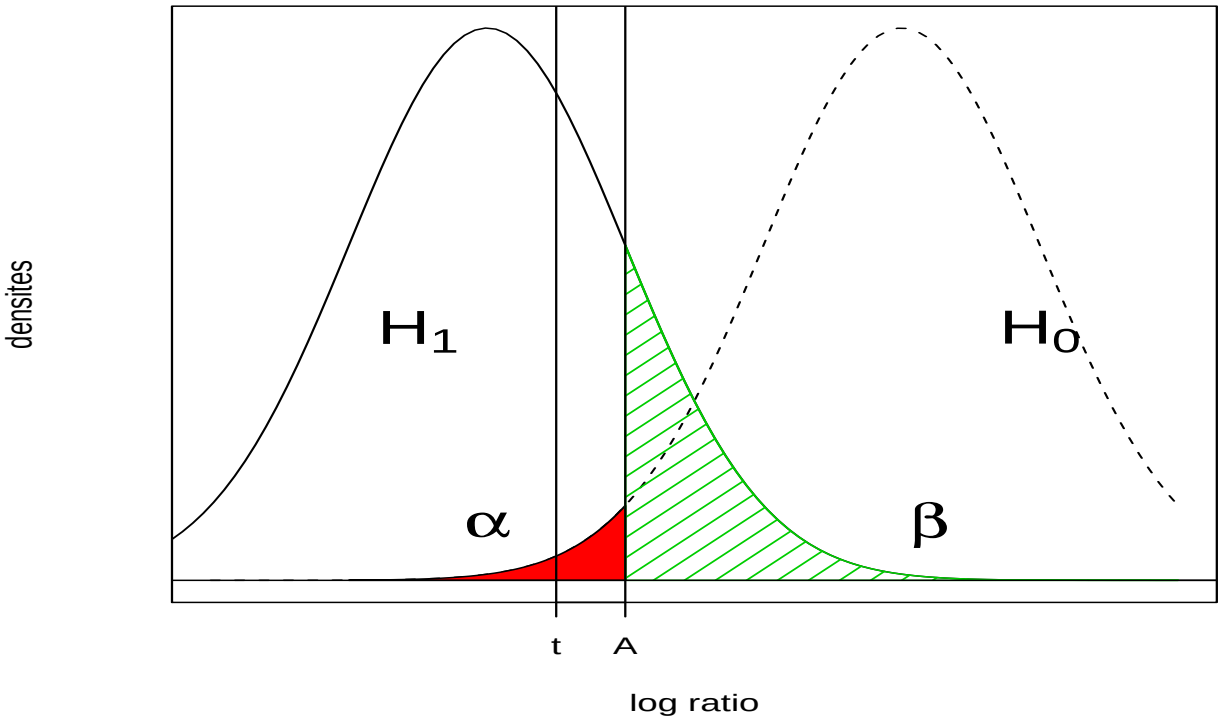


FIGURE – Erreur de type I et II et région critique de la forme $\Gamma =]-\infty, A]$

Asymétrie

Hypothèse nulle et hypothèse alternative

- L'**hypothèse nulle** (notée H_0) est l'hypothèse privilégiée. C'est celle qui est supposée vraie par défaut (vérité établie) et qui sera conservée en cas de doutes (trop importants).
- L'**hypothèse alternative** (notée H_1) contredit l'hypothèse nulle. C'est l'hypothèse que l'on cherche à montrer.

Statistique de test

Définition

Une **statistique de test** est une statistique (dont la loi est connue sous H_0) qui permet de mesurer l'écart à l'hypothèse nulle.

Règle de décision

Choix du niveau du test

- Le **niveau de signification** (ou **seuil de signification**) du test est le risque de première espèce α consenti.
- Le niveau de signification du test est souvent fixé à 0.05 ou 0.01, mais ce seuil est arbitraire est toute autre valeur peut être choisie.

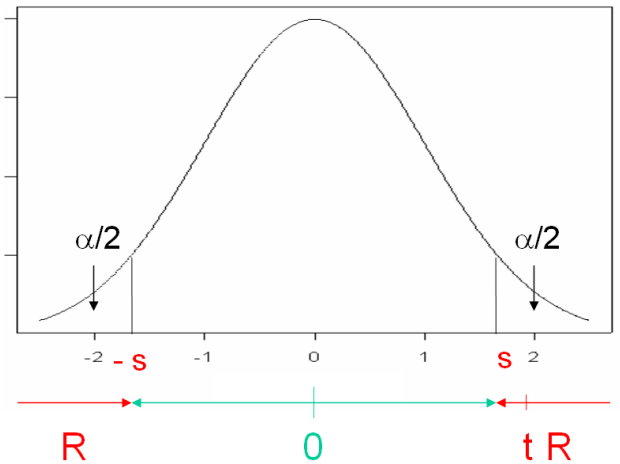
Région de rejet

Définition

La **Région de rejet** est l'ensemble R des valeurs (de la statistique de test) pour lesquelles l'hypothèse nulle est rejetée.

Démarche de Neyman Pearson

Maximiser la puissance tout en contr ôlant α.



Degré de signification (p-valeur)

Définition

Le **degré de signification** (ou **p-valeur**) est défini par :

$$p = min\{\alpha | T \in \Gamma_\alpha\}$$

Test unilatéral à droite	Test unilatéral à gauche	Test bilatéral
$p = \mathbb{P}(T > t \| H_0)$	$p = \mathbb{P}(T < t \| H_0)$	$p = \mathbb{P}(\|T\| > \|t\| \| H_0)$

Remarques

- La p-valeur est la probabilité d'obtenir une valeur de la statistique de test au moins aussi extr ême que celle observée lorsque H_0 est vraie
- En pratique, on rejette H_0 lorsque $p < \alpha$

Degré de signification

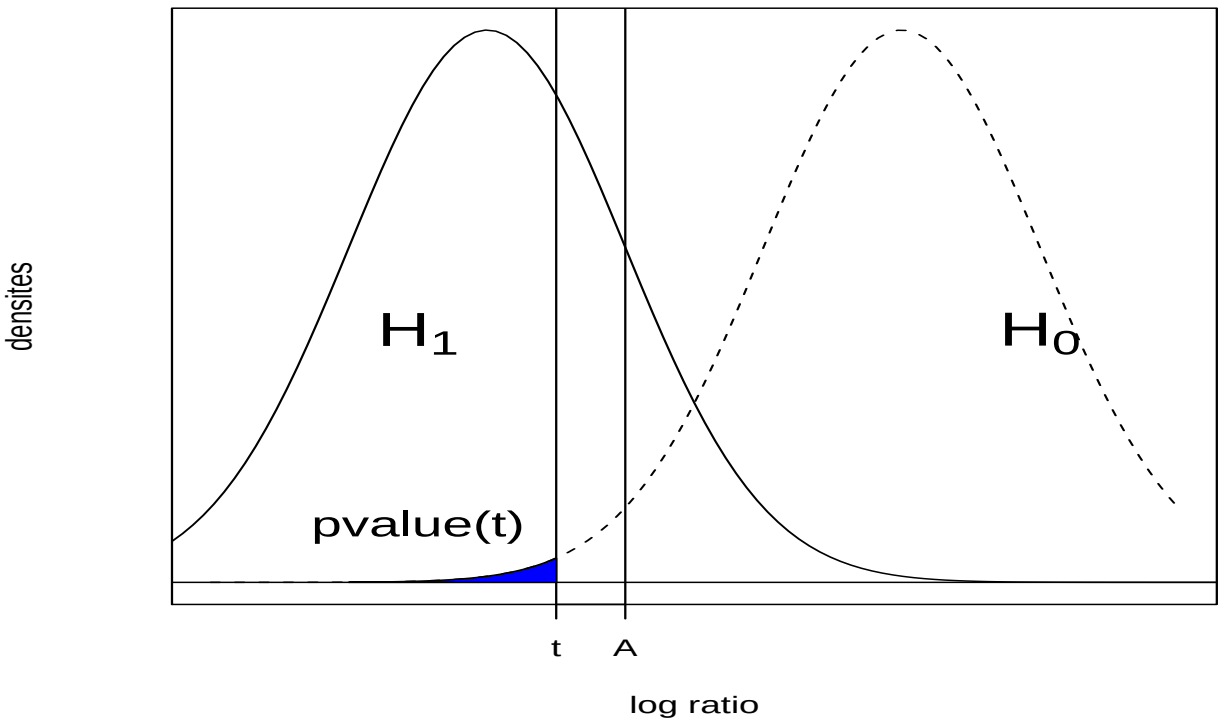


FIGURE – p-valeur associée à la réalisation t de la statistique de décision pour un test unilatéral à gauche.

Test sur l'espérance d'un échantillon gaussien

Exemple

On dispose d'un lot de 500 souriceaux et on se demande si ce lot est bien standard du point de vue de la taille. Dans des conditions normales, la taille adulte de ce type de souris suit une loi normale d'espérance 10 *cm* et de variance inconnue. Un échantillon de 5 sujets, tirés au hasard dans ce lot atteint à l'âge adulte la taille suivante :

$$12{,}4 \quad 13{,}0 \quad 9{,}8 \quad 10{,}5 \quad 14{,}2$$

Au niveau $\alpha = 5\%$, peut-on considérer que ce lot est bien standard ?

Test sur l'espérance d'un échantillon gaussien

Cas 1 : variance connue (test z)

- Présupposés : $X_1, ..., X_n$ iid avec $X_i \sim N(\mu, \sigma_0^2)$, σ_0^2 connu.
- Hypothèse nulle : $H_0 : \mu = \mu_0$
- Statistique de test :

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \underset{H_0}{\sim} N(0,1)$$

Test sur l'espérance d'un échantillon gaussien

Cas 2 : variance inconnue (test de Student ou test t)

- Présupposés : $X_1, ..., X_n$ iid avec $X_i \sim N(\mu, \sigma^2)$, σ^2 inconnu.
- Hypothèse nulle : $H_0 : \mu = \mu_0$
- Statistique de test :

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{S^{\star 2}}{n}}} \underset{H_0}{\sim} t_{n-1}$$

Fonction R

`t.test`

Comparaison des espérances de deux échantillons gaussiens

Test de Student

- Présupposés :
  $X_{11}, ..., X_{1n_1}$ iid avec $X_{1i} \sim N(\mu_1, \sigma_1^2)$
  $X_{21}, ..., X_{2n_2}$ iid avec $X_{2i} \sim N(\mu_2, \sigma_2^2)$
  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ inconnu.
- Hypothèse nulle : $H_0 : \mu_1 = \mu_2$
- Statistique de test :

$$\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{(n_1-1)S_1^{\star 2} + (n_2-1)S_2^{\star 2}}{n_1+n_2-2}}} \underset{H_0}{\sim} t_{n_1+n_2-2}$$

Fonction R

t.test

Test sur la variance d'un échantillon gaussien

- Présupposés : $X_1, ..., X_n$ iid avec $X_i \sim N(\mu, \sigma^2)$.
- Hypothèse nulle : $H_0 : \sigma^2 = \sigma_0^2$
- Statistique de test :

$$\frac{n-1}{\sigma_0^2} S^{\star 2} \underset{H_0}{\sim} \chi^2_{n-1}$$

Comparaison des variances de deux échantillons gaussiens

- Présupposés :
  $X_{11}, ..., X_{1n_1}$ iid avec $X_{1i} \sim N(\mu_1, \sigma_1^2)$
  $X_{21}, ..., X_{2n_2}$ iid avec $X_{2i} \sim N(\mu_2, \sigma_2^2)$
- Hypothèse nulle : $H_0 : \sigma_1^2 = \sigma_2^2$
- Statistique de test :

$$\frac{S_1^{\star 2}}{S_2^{\star 2}} \underset{H_0}{\sim} \mathcal{F}(n_1 - 1, n_2 - 1)$$

Fonction R

var.test