

DUBii 2019 - Module 3 : Analyse statistique avec R

Séance 2 : échantillonnage, estimation et tests statistiques

Leslie REGAD

2019-02-10

Contents

Plan	1
Expérience aléatoire	2
Variable aléatoire (VA) / Unité statistique	2
Deux types de VA réelles	2
Variable réelle discrète	2
Variable quantitative continue	3
Etude de cas : le myélome multiple (MM)	4
Cas d'étude	5
Mise en place du protocole	5
Analyse des données avec R	6
Description des deux échantillons	6
Mise en place du protocole	7
Estimation de paramètres	7
Les fluctuations d'échantillonnage	8
Estimation de paramètres	8
Estimation d'un paramètre : Intervalle de confiance (IC)	9
IC de la moyenne : 2 cas	9
IC du taux de pyr à 95%	10
Cas d'étude : Introduction aux tests statistiques	10
Déroulement d'un test statistique	11
Est-ce que les données sont compatibles avec H ₀ ?	11
Est-ce que les données sont compatibles avec H ₀ ?	12
Est-ce que les données sont compatibles avec H ₀ ?	12
Interprétation de la p-value : $p(S > s_{obs})$	13
Les erreurs	13
Interprétation de la p-value : $p(S > s_{obs})$	14
Comparaison du taux de pyr chez les patients malades et contrôles	15
Les différents types de tests	15
Les différents types de tests	16
Merci de votre attention !!!	16

Plan

- Rappels de probabilité
- Introduction du cas d'études
- Estimation d'un paramètre statistique
 - Estimateur ponctuel
 - Intervalle de confiance
- La théorie des tests d'hypothèses
 - Le test de comparaison de deux moyennes
 - Les autres tests d'hypothèses
- A vous de jouer : Etude de cas

Expérience aléatoire

- **Phénomène déterministe** : phénomène dont on peut prévoir le résultat avec certitude lorsqu'on contrôle les causes
exemple : je lâche la craie → elle tombe
- **Expérience aléatoire** : expérience qui, répétée dans des conditions apparemment identiques, peut donner des résultats différents. Elle se définit par trois propositions :
 - l'expérience peut être répétée
 - plusieurs résultats sont possibles
 - le résultat ne peut être prédit avec certitude*exemple* : je lâche la craie → combien de morceaux de craie trouve-t-on par terre ?

Variable aléatoire (VA) / Unité statistique

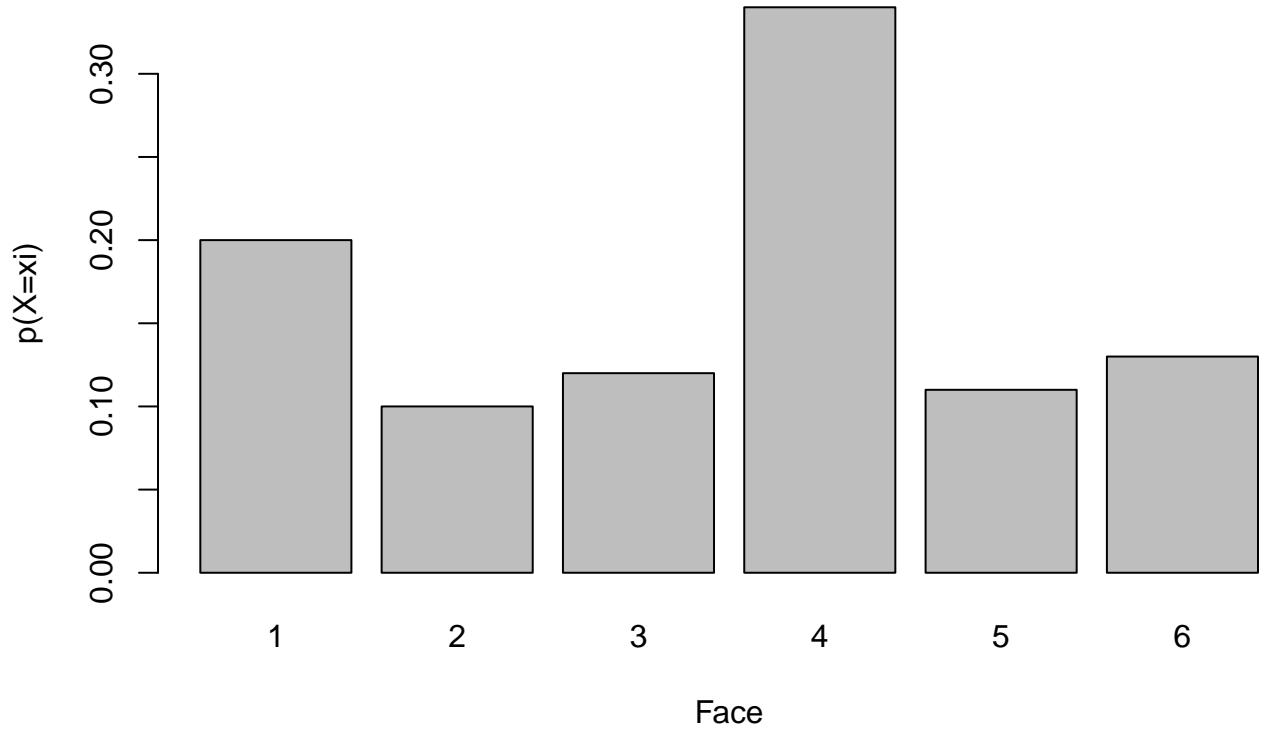
- **Variable aléatoire (VA)** = variable X qui prend des valeurs en fonction du résultat d'une expérience aléatoire
→ elle répond à la question : Que mesure-t-on pendant l'expérience aléatoire ?
- **Unité statistique** = unité sur laquelle est mesurée la VA
- Exemple
 - Expérience aléatoire : je lâche la craie → combien de morceaux trouve-t-on par terre?
 - variable aléatoire : nombre de morceaux obtenus
 - unité statistique (US) : 1 craie
- Notations :
 - X : la variable aléatoire
 - x_i : les réalisations de X
- VA réelles sont caractérisées par 2 paramètres :
 - **Espérance ($E(X)$)** : caractérise la tendance centrale, la valeur moyenne, prise par la VA
 - **Variance ($V(X)$)** : caractérise la dispersion des valeurs de la variable autour de son espérance

Deux types de VA réelles

- L'ensemble des valeurs possibles est défini sur \mathbb{R}
- **discrète** : l'ensemble des valeurs que X peut prendre est fini ou infini dénombrable
 - nombres d'oeufs pondus par une poule : 2, 3, 4, ...
- **continue** : les autres variables
 - poids d'un patient : 75.3kg, 56.4kg, 87kg, ...
 - taux de cholestérol : 2g/l, 1.8g/l, ...

Variable réelle discrète

- se définit par sa **loi de probabilité** : fonction qui associe à chaque valeur de X sa probabilité
$$p_i = p(X = x_i)$$
 - avec $0 < p_i < 1$
 - $\sum_{i=1}^k p_i = 1$, avec k le nombre de modalités de X
- X = “le numéro de la face d'un dé”



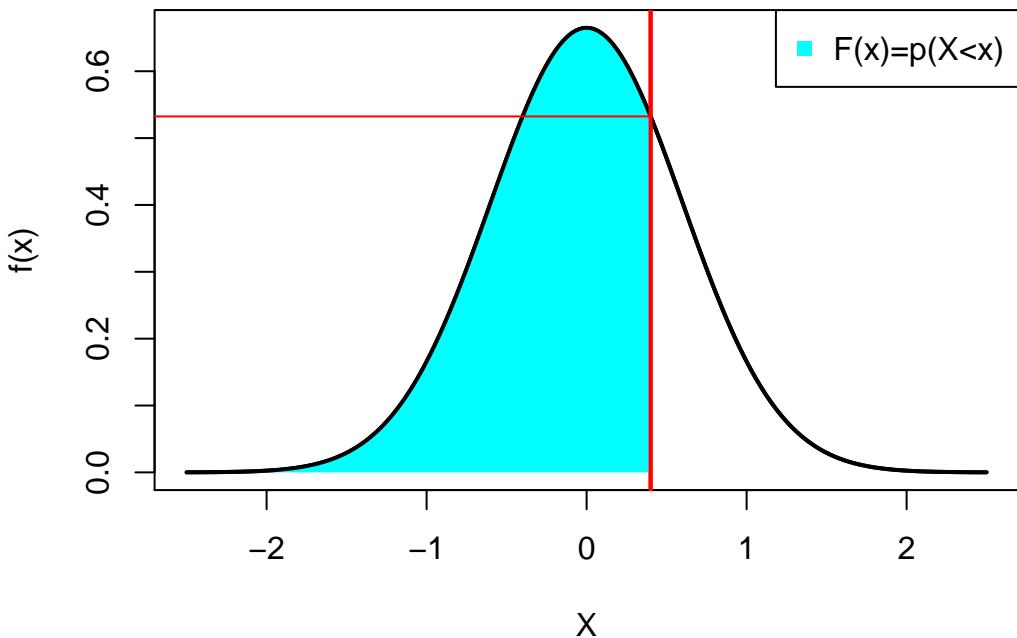
Variable quantitative continue

- l'ensemble des valeurs que X peut prendre est infini
 $\rightarrow p(X = x_i) = 0$
- se caractérise par sa **densité de probabilité** = la fonction $f(x)$ définie par :

$$P[x < X \leq x + dx] = f(x)dx \text{ pour } x \in \mathbb{R}$$

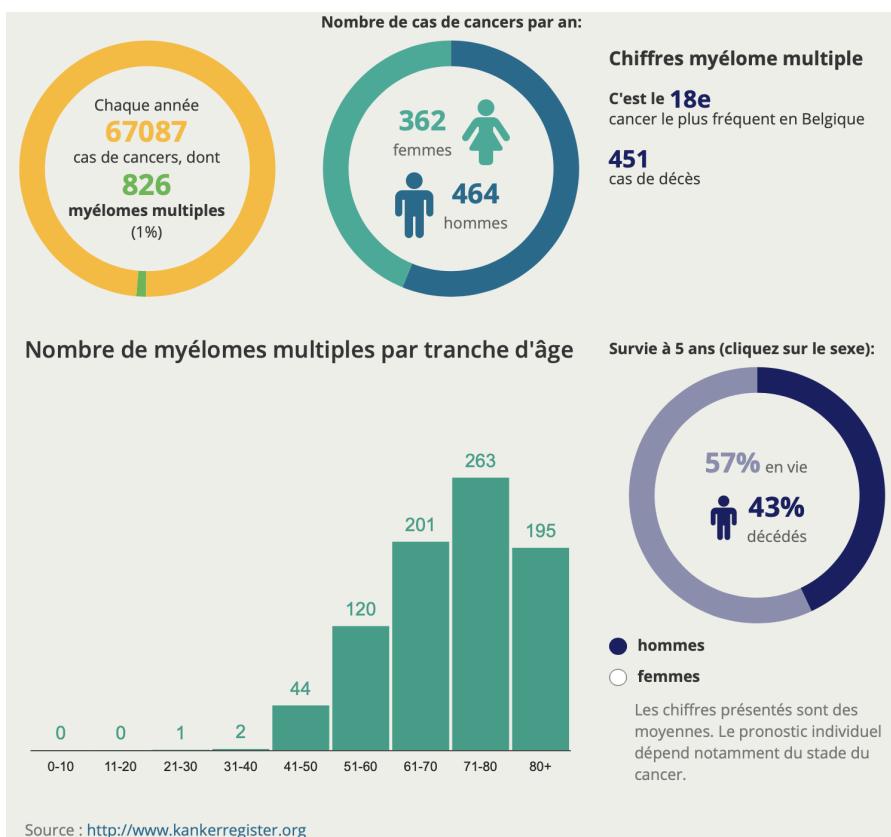
- la **fonction de répartition** est la fonction $F(x)$ telle que :

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(u)du \text{ pour } x \in \mathbb{R}$$



Etude de cas : le myélome multiple (MM)

- Prolifération incontrôlée des plasmocytes → **Affection de la moelle osseuse**



- Cancer relativement **rare**
- Touche proportionnellement davantage les hommes que les femmes
- Atteint principalement les personnes âgées de plus de 60 ans et survient rarement avant 40 ans

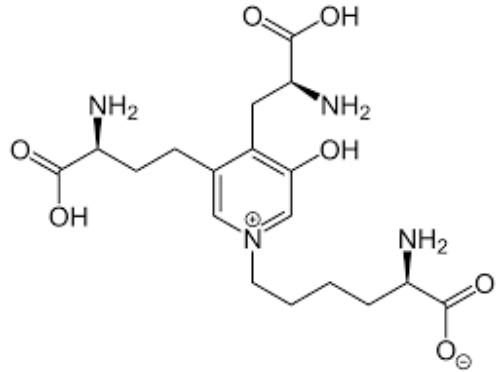


Figure 1:

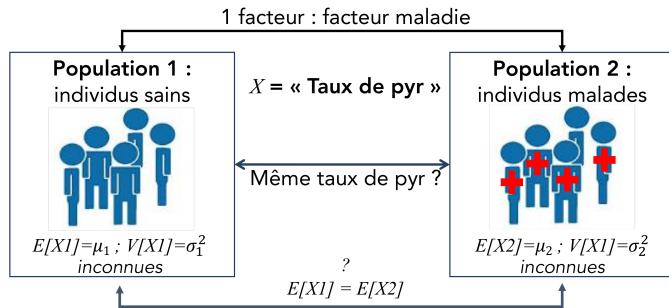
- Symptômes
 - plasmocytes malins prennent peu à peu la place des autres cellules sanguines :
→ **diminution de l'immunité humorale**, anémie, infections, risque d'ecchymoses, ...
 - **Destruction osseuse** qui résulte de la décalcification en certains endroits du squelette
→ **fractures** peuvent aisément se produire à ces endroits

Cas d'étude

Est-ce le dosage urinaire de la molécule déoxypyridinoline (pyr) est un bon marqueur pour détecter le MM ?

→ est-ce que le taux de pyr des patients malades est plus grand que celui des individus sains ?

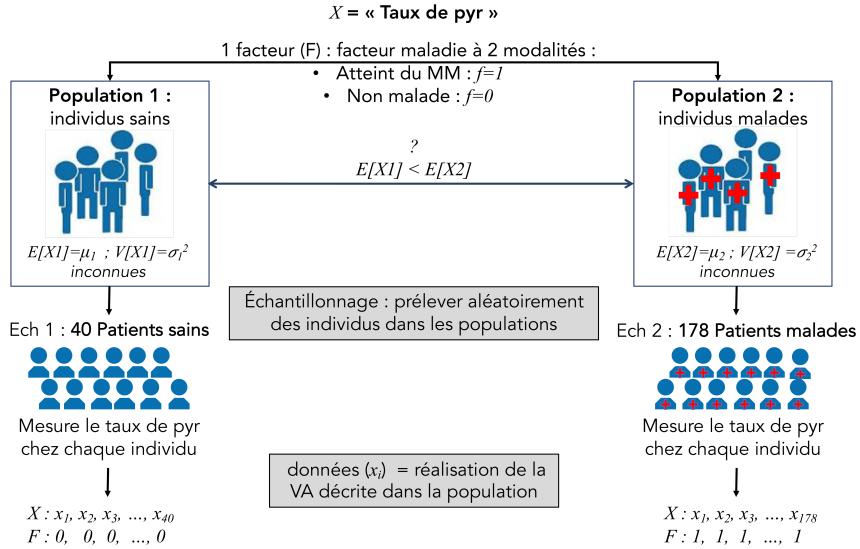
- X = “Taux de pyr” (micmol/mmolcreat)
- US = un individu → type = quantitative continue



Population = ensemble d’individus ayant des caractéristiques qui leur sont propres

Mise en place du protocole

→ Récolte des données



Analyse des données avec R

```
dataMyelom <- read.table("data/myelom.txt", sep="\t", header=T)
dim(dataMyelom)
```

```
[1] 218 38
```

```
table(dataMyelom[, "diagn"])
```

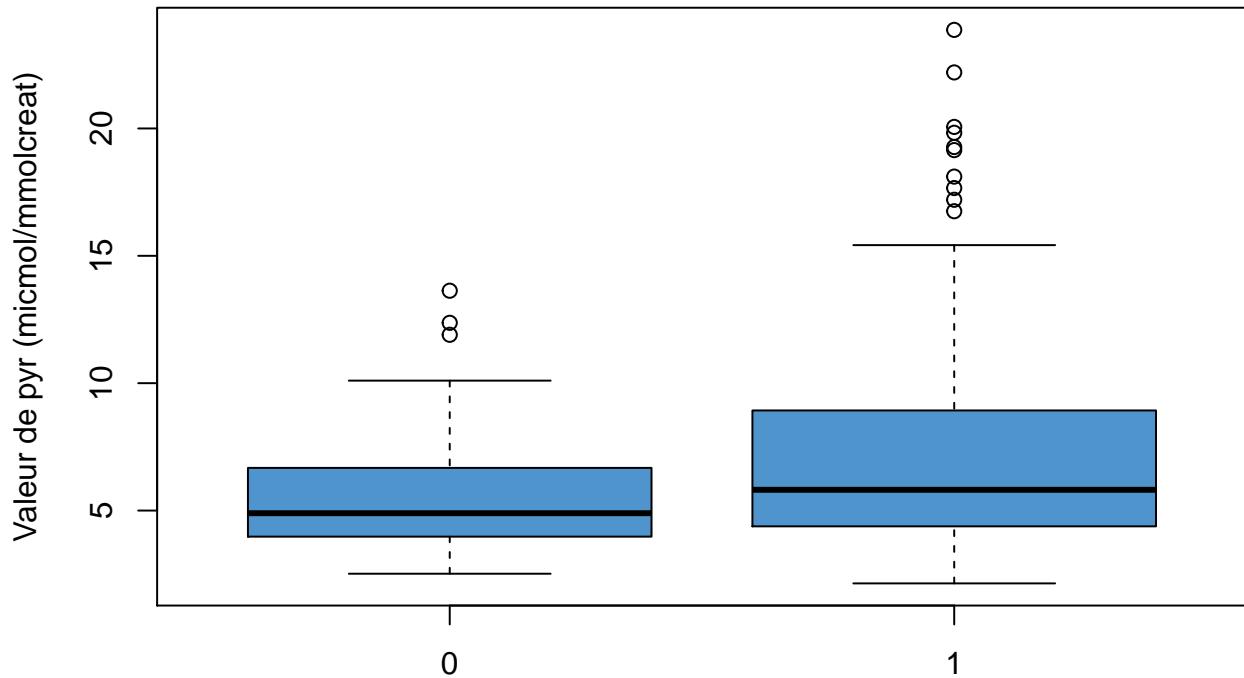
```
0   1
40 178
```

- `diagn = 0` → individus non malades
- `diagn = 1` → patients malades

Description des deux échantillons

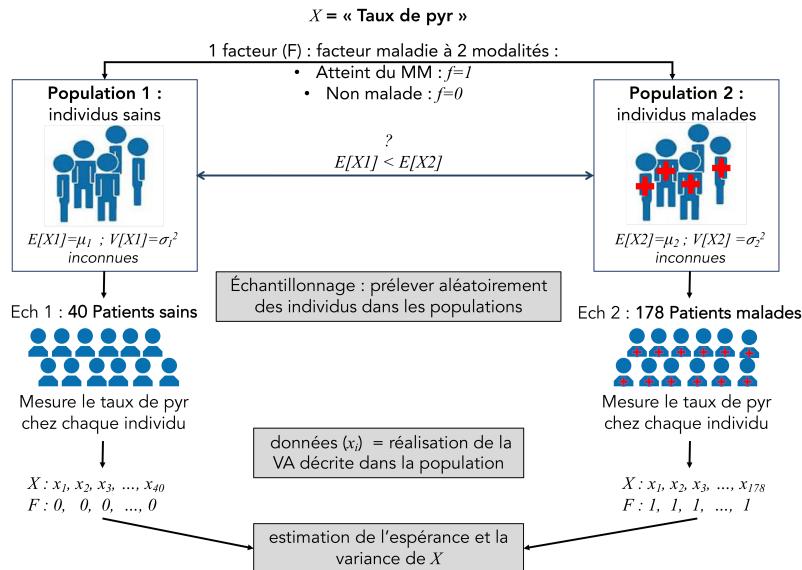
- distribution de pyr dans les deux échantillons

```
boxplot(pyr~diagn, data=dataMyelom, ylab="Valeur de pyr (micmol/mmolcreat)", col = "steelblue3")
```



Mise en place du protocole

→ Récolte des données



Estimation de paramètres

- estimateur de l'espérance de X = moyenne de X dans l'échantillon

$$\hat{\mu} = m = \frac{\sum_{i=1}^n x_i}{n}$$

- estimateur de la variance de X

$$\widehat{\sigma^2} = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2$$

- calcul de l'**estimateur de l'espérance** de X dans les deux échantillons

```
by(dataMyelom[, "pyr"], dataMyelom["diagn"], mean)
```

```
diagn: 0
[1] 5.6985
```

```
diagn: 1
[1] 7.241742
```

- $m_1 \neq m_2$: Ne signifie pas que $\mu_1 \neq \mu_2$

Les fluctuations d'échantillonnage

- 1 population où $X \sim \mathcal{N}(\mu = 4; \sigma^2 = 12)$

1. tire un échantillon de 20 individus et calcule la moyenne de X dans cet échantillon

```
ech1 <- rnorm(n=20, mean=4, sd=sqrt(12))
mean(ech1)
```

```
[1] 5.393009
```

2. tire un deuxième échantillon de 20 individus et calcule la moyenne de X dans cet échantillon

```
ech2 <- rnorm(n=20, mean=4, sd=sqrt(12))
mean(ech2)
```

```
[1] 2.91129
```

→ Les différences entre les deux estimateurs sont dues aux **fluctuations d'échantillonnage**

Estimation de paramètres

- estimateur de l'espérance de X = moyenne de X dans l'échantillon

$$\hat{\mu} = m = \frac{\sum_{i=1}^n x_i}{n}$$

- estimateur de la variance :

$$\widehat{\sigma^2} = s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - m)^2$$

- calcul de l'**estimateur de la variance** de X dans les deux échantillons

```
by(dataMyelom[, "pyr"], dataMyelom["diagn"], var)
```

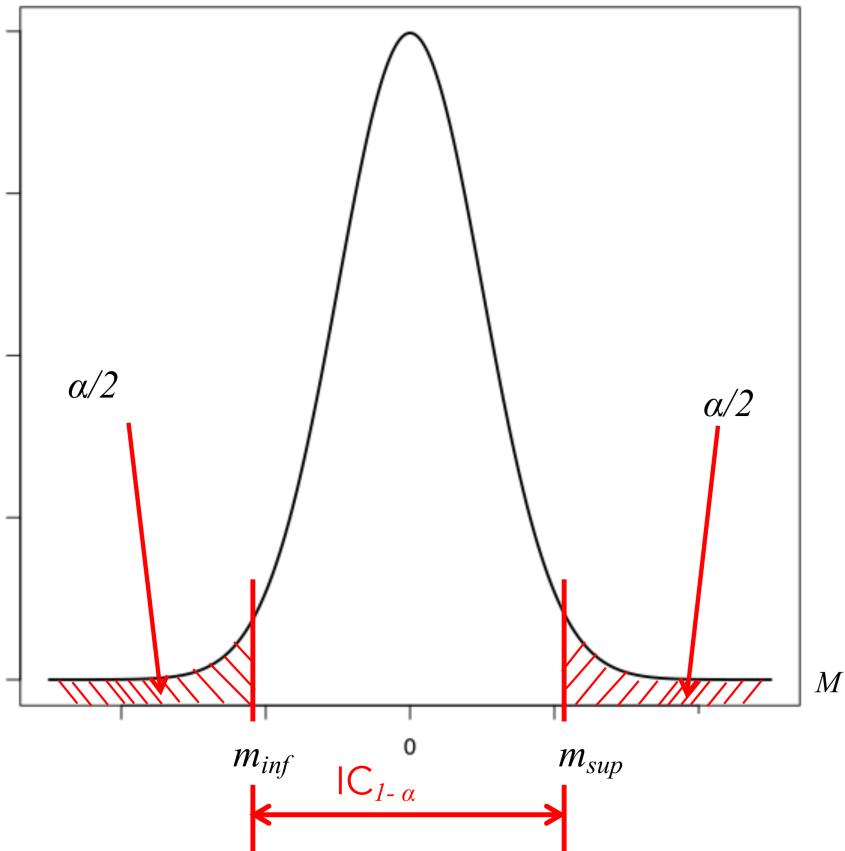
```
diagn: 0
[1] 6.390654
```

```
diagn: 1
[1] 16.72764
```

Estimation d'un paramètre : Intervalle de confiance (IC)

- Apporte deux informations :
 - les valeurs possibles du paramètre θ à estimer
 - le degré de confiance attribué à ces valeurs
- **IC de la moyenne** = intervalle $[m_{inf}; m_{sup}]$ dans lequel on considère que μ a une probabilité $(1 - \alpha)$ de se trouver

$$p(m_{inf} < \mu < m_{sup}) = 1 - \alpha$$



IC de la moyenne : 2 cas

- Prend en compte
 - la **variabilité** des données
 - la **taille de l'échantillon**
- si $n > 30$:

$$IC_{1-\alpha} = \left[m \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{s^2}{n}} \right]$$

– $z_{1-\frac{\alpha}{2}}$: quantile de la loi normale centrée réduite d'ordre α

- si $n < 30$ et $X \sim \mathcal{N}(\mu; \sigma^2)$

$$IC_{1-\alpha} = \left[m \pm t_{\alpha;(n-1)\text{ddl}} \times \sqrt{\frac{s^2}{n}} \right]$$

– $t_{\alpha;(n-1)\text{ddl}}$: quantile de la loi de student d'ordre α à $(n - 1)$ degrés de liberté

IC du taux de pyr à 95%

- échantillon 1 : 40 patients sains

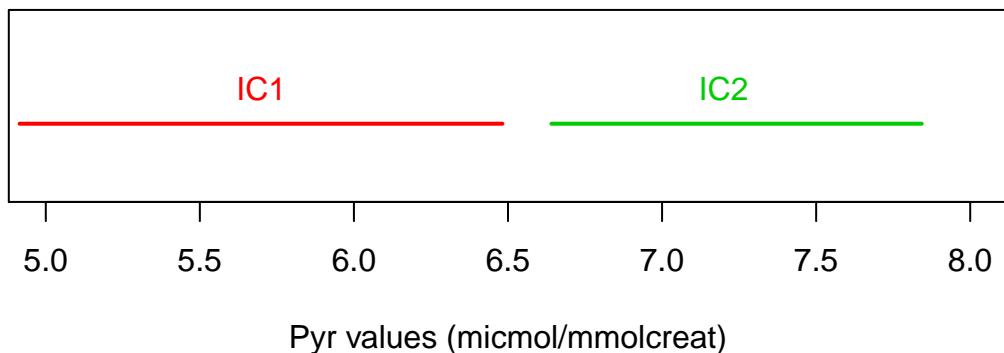
```
alpha <- 0.05
pyr0 <- dataMyelom[which(dataMyelom[, "diagn"]==0), "pyr"]
borneInf.0 <- mean(pyr0) - qnorm(1-alpha/2) * sqrt(var(pyr0)/length(pyr0))
borneSup.0 <- mean(pyr0) + qnorm(1-alpha/2) * sqrt(var(pyr0)/length(pyr0))
round(c(borneInf.0, borneSup.0), 2)
```

[1] 4.92 6.48

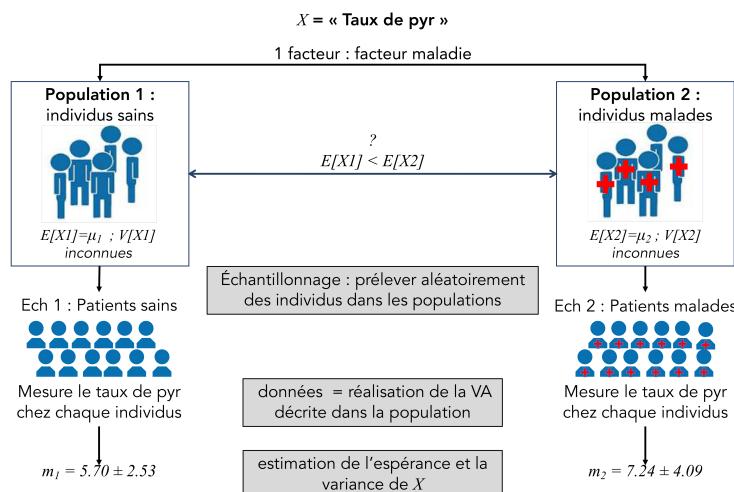
- échantillon 2 : 178 patients atteints de MM

```
pyr1 <- dataMyelom[which(dataMyelom[, "diagn"]==1), "pyr"]
borneInf.1 <- mean(pyr1) - qnorm(1-alpha/2) * sqrt(var(pyr1)/length(pyr1))
borneSup.1 <- mean(pyr1) + qnorm(1-alpha/2) * sqrt(var(pyr1)/length(pyr1))
round(c(borneInf.1, borneSup.1), 2)
```

[1] 6.64 7.84



Cas d'étude : Introduction aux tests statistiques



- 2 possibilités pour expliquer les différences entre m_1 et $m_2 \rightarrow 2$ hypothèses
 - elles s'expliquent seulement par le **hasard** (fluctuations d'échantillonnage) : H_0 (hypothèse nulle)
 - elles s'expliquent par le **hasard** et par le **facteur maladie** : H_1 (hypothèse alternative)

Déroulement d'un test statistique

- Définit deux hypothèses sur un des paramètres de la VA
 - **hypothèse nulle (H_0)** : égalité des paramètres : $\mu_1 = \mu_2 = \mu$
Les différences observées sont expliquées uniquement par le hasard
 - **hypothèse alternative (H_1)** : inégalité des paramètres : $\mu_1 < \mu_2$
Les différences observées sont expliquées à la fois par le hasard et par le facteur

Réaliser le test va consister à choisir une des deux hypothèses (H_0 ou H_1) en se basant sur les données des échantillons

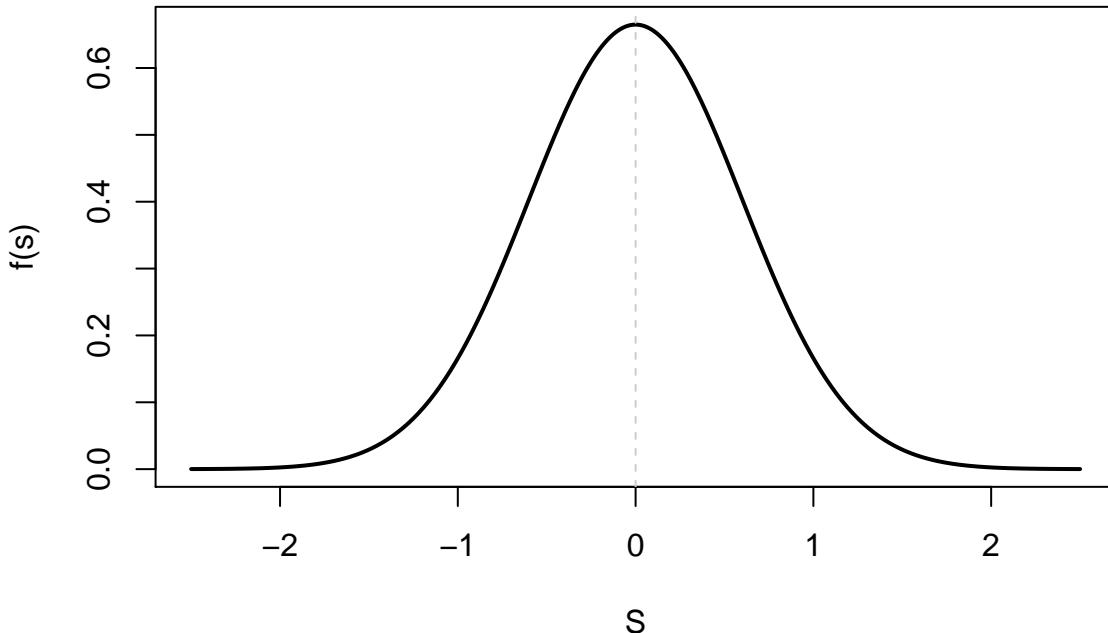
→ Est-ce que les données des échantillons (m_1 et m_2) sont compatibles avec H_0 ?

- si oui → non rejet de H_0
le hasard explique à lui seul les différences entre les deux moyennes observées
- si non → rejet de H_0
le hasard et le facteur expliquent les différences entre les deux moyennes observées

Est-ce que les données sont compatibles avec H_0 ?

1. Définir un **critère statistique** S dont la loi sous H_0 est connue
 $S = M_2 - M_1$ avec M =moyenne de X dans 1 échantillon

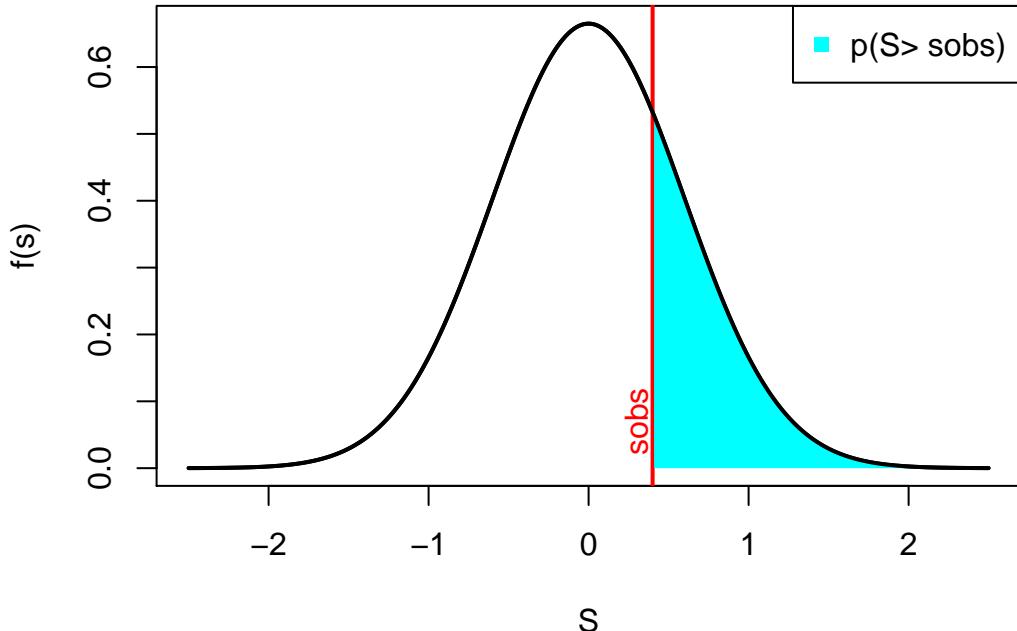
$$\text{Sous } H_0 \text{ (} n_1 \text{ et } n_2 > 30 \text{)} : S \sim \mathcal{N} \left(0; \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$



2. Calculer s_{obs} = valeur de S calculée sur les échantillons : $s_{obs} = m_1 - m_2$
3. Regarde si la réalisation s_{obs} est fortement probable sous H_0 (quand $\mu_1 = \mu_2$)

Est-ce que les données sont compatibles avec H_0 ?

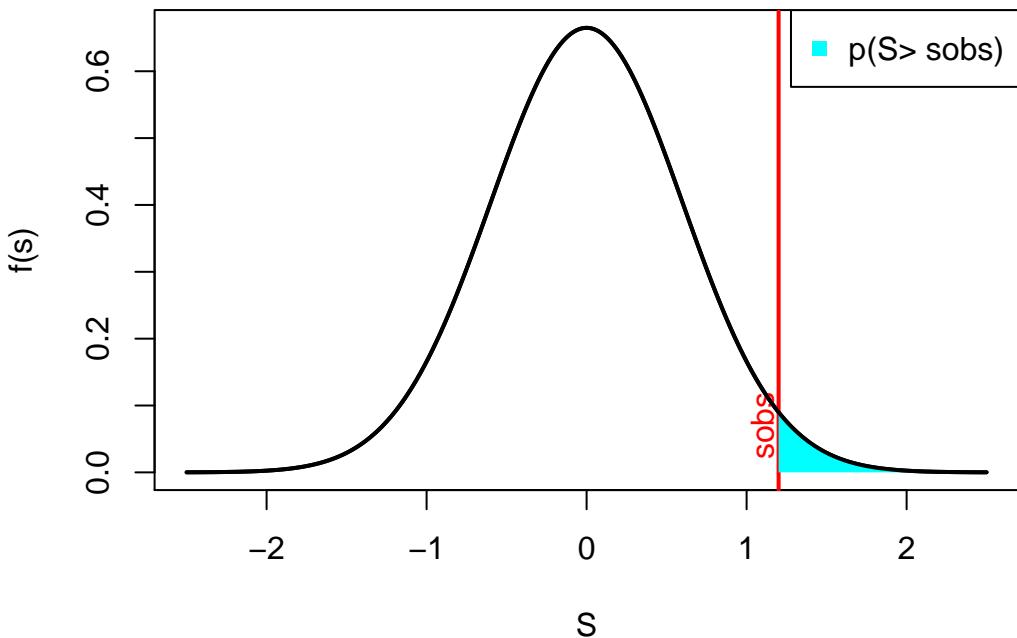
- $S = M_2 - M_1$ avec $S \sim \mathcal{N}\left(0; \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$
- $s_{obs} = m_2 - m_1$
- Calcule la probabilité d'obtenir sous H_0 le même résultat que celui observé (ou plus grand) sur les échantillons = **p-value** = $p(S > s_{obs})$



- forte valeur de p-value \rightarrow : s_{obs} est fortement probable sous H_0

Est-ce que les données sont compatibles avec H_0 ?

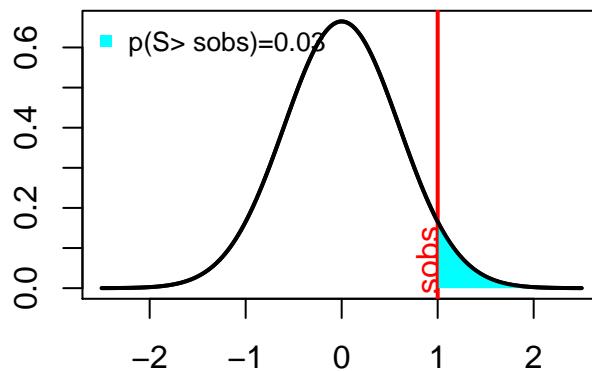
- $S = M_2 - M_1$ avec $S \sim \mathcal{N}\left(0; \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$
- $s_{obs} = m_2 - m_1$
- Calcule la probabilité d'obtenir sous H_0 le même résultat que celui observé (ou plus grand) sur les échantillons = **p-value** = $p(S > s_{obs})$



- forte valeur de p-value \rightarrow : s_{obs} est fortement probable sous H_0
- faible valeur de p-value \rightarrow : s_{obs} est très peu probable sous H_0

Interprétation de la p-value : $p(S > s_{obs})$

- p-value = 0.03
 \rightarrow 3% de chance que la valeur s_{obs} se produise sous H_0
 \rightarrow on a 3% de chance que 2 échantillons pris au hasard aient la même différence (ou plus grande) que celle observée



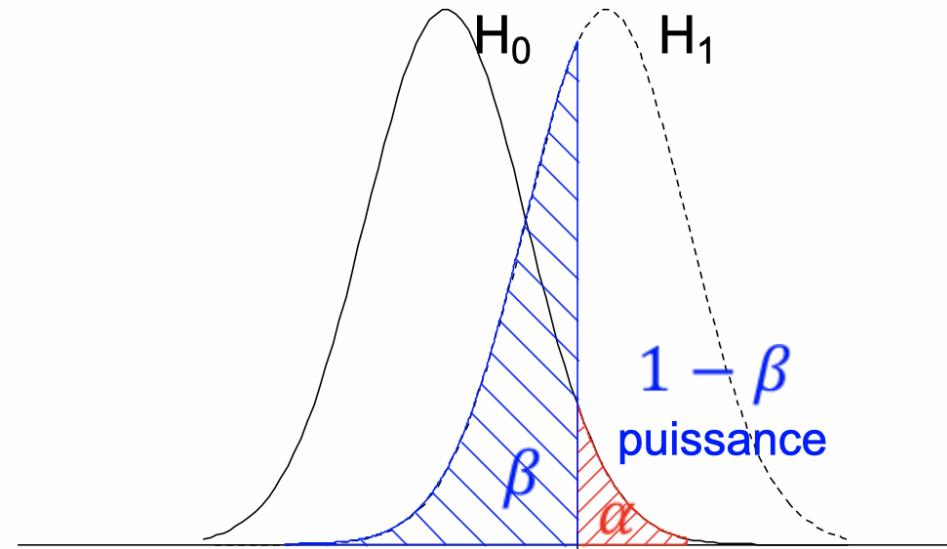
- p-value = 0.67
 \rightarrow 67% de chance que la valeur s_{obs} se produise sous H_0
 \rightarrow on a 67% de chance que 2 échantillons pris au hasard aient la même différence (ou plus grande) que celle observée

Les erreurs

Deux erreurs possibles quand on conclut au test :

- **erreur de type I** : je conclus à 1 effet de la maladie alors qu'en réalité il n'y en a pas.
- **erreur de type II** : je conclus à \emptyset effet de la maladie alors qu'en réalité il existe.

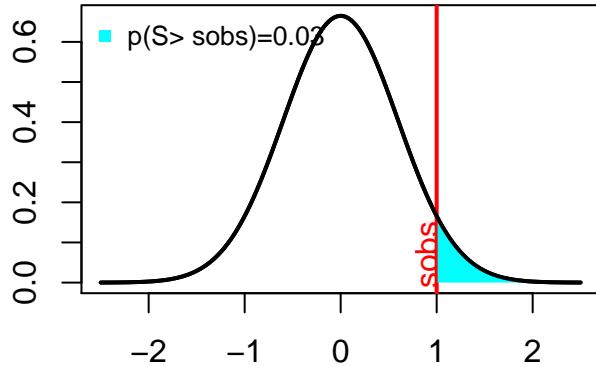
		Décision du test	
		Rejet de H ₀	Non Rejet de H ₀
Réalité	H ₀ vraie	α	$1-\alpha$
	H ₀ fausse	<i>Puissance</i>	β



- **risque de 1ère espèce (α)** : $p(\text{rejet } H_0 \mid H_0 \text{ est vraie})$
- **risque de 2ème espèce (β)** : $p(\text{non rejet } H_0 \mid H_0 \text{ est fausse})$
- **Puissance du test** Puissance = $1 - \beta = p(\text{rejet } H_0 \mid H_0 \text{ est fausse})$
La probabilité de détecter une différence alors qu'il en existe une

Interprétation de la p-value : $p(S > s_{obs})$

- p-value = 0.03
 → 3% de chance que la valeur s_{obs} se produise sous H_0
 → on a 3% de chance que 2 échantillons pris au hasard aient la même différence (ou plus grande) que celle observée



- p-value = 0.67
 - 67% de chance que la valeur s_{obs} se produise sous H_0
 - on a 67% de chance que 2 échantillons pris au hasard aient la même différence (ou plus grande) que celle observée
- on choisit un risque α comme valeur seuil : souvent 5%
 - si p-value > 0.05 → non rejet de H_0
→ les différences observées sur les échantillons sont expliquées uniquement par le hasard
 - si p-value < 0.05 → rejet de H_0
→ Les différences observées sont dues aux fluctuations d'échantillonnage et au facteur

Comparaison du taux de pyr chez les patients malades et contrôles

Est-ce que le taux de pyr des patients malades est plus grand que celui des individus sains ?

- hypothèses

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 < \mu_2$$

```
t.test(pyr~diagn, data=dataMyelom, var.equal=TRUE, alternative="less")
```

Two Sample t-test

```
data: pyr by diagn
t = -2.2878, df = 216, p-value = 0.01156
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
-Inf -0.428922
sample estimates:
mean in group 0 mean in group 1
5.698500    7.241742
```

- p-value < 0.05 → le test est significatif au risque de 5%
- Le taux de pyr des individus contrôles est significativement plus petit de celui des patients malades
- Le dosage de pyr est un bon marqueur de la maladie.

Les différents types de tests

- variable réelle continue : Tests sur les moyennes
 - 1 échantillon et 1 population de référence

- * grand échantillon : test espilon
- * petit échantillon : test de student
 - : test paramétrique : test de student (1 conditions de validité)
 - X doit suivre une loi normale dans la population d'où est tiré l'échantillon
 - tests non paramétriques : test de Wilcoxon ou test de Mann-Whitney
- 2 échantillons
 - * grands échantillons : test espilon
 - * petits échantillons → : test paramétrique : test de student (2 conditions de validité)
 - X doit suivre un loi normale dans les deux populations
 - Egalité des deux variances
 - tests non paramétriques : test de Kruskal-Wallis
- ≥ 3 échantillons
 - * test paramétrique ANOVA (2 conditions de validité) X doit suivre un loi normale dans les différentes populations
 - Egalité des k variances
 - * test non paramétrique : test de Kruskal-Wallis

Les différents types de tests

- test sur les distribution :
 - à une loi normale : test de Shapiro
 - à une distribution théorique : test du Chi² d'homogénéité
 - deux distributions : test du Chi² d'adéquation
- deux variables aléatoires (lien) :
 - réelles continues :
 - * test paramétrique : test du coefficient de corrélation de Pearson
 - binormalité de la variable X
 - * test non paramétrique : test du coefficient de corrélation de Spearman
 - réelles discrètes :
 - * test du chi² d'indépendance

Merci de votre attention !!!

Place au TP : Etude des caractéristiques des patients atteints d'une cirrhose du foie et des individus sains.

