

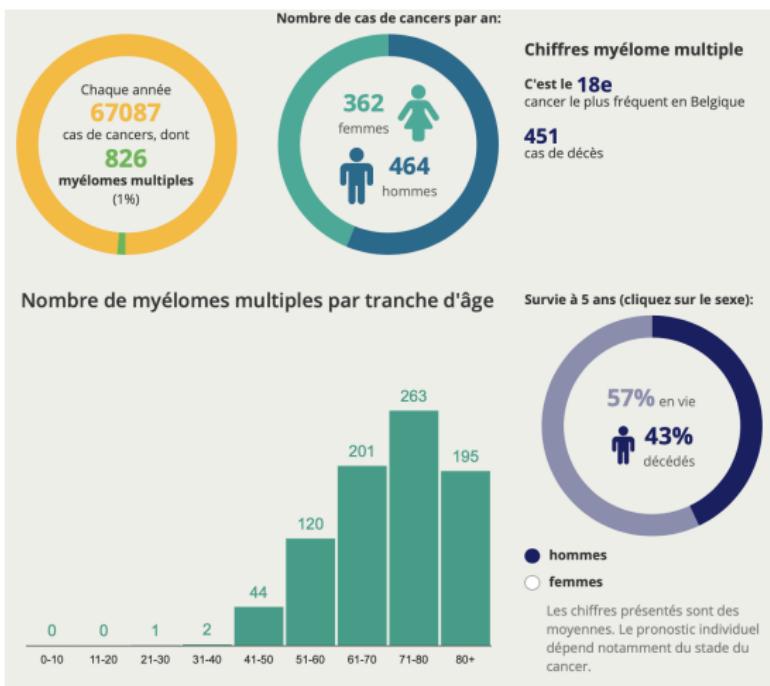
Module 3 - Analyse statistique avec R - Séance 2 DUBii 2019

Leslie REGAD

2019-02-05

Etude de cas : le myélome multiple (MM)

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



Question statistique

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

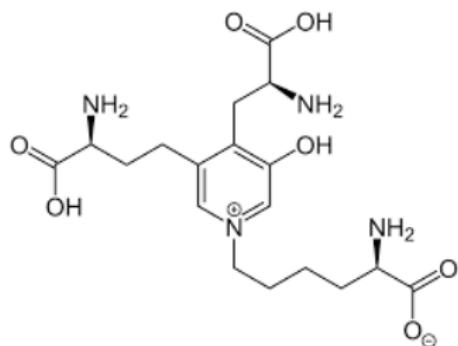
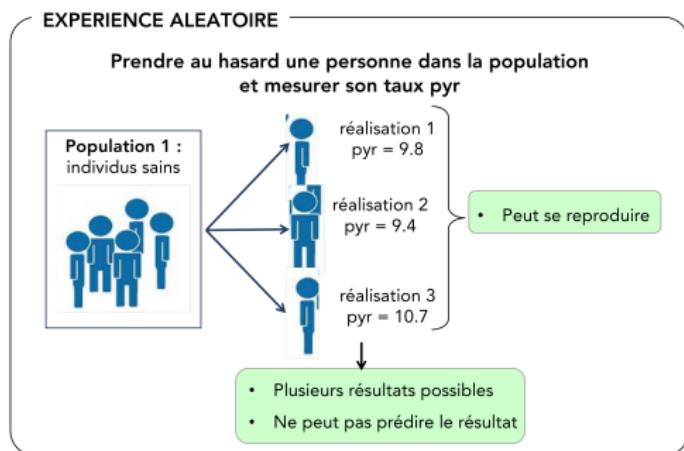


Figure 1:

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

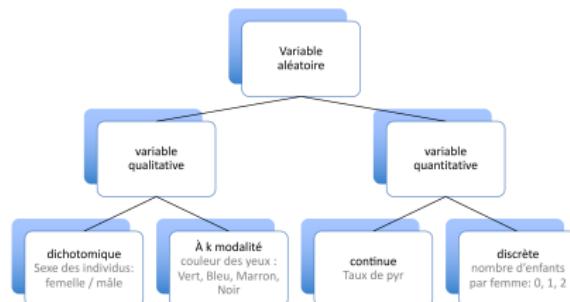
Expérience aléatoire / variable aléatoire



- ▶ Que mesure-t-on lors de l'expérience : $X = \text{'le taux de pyr'}$ → **Variable aléatoire**
- ▶ Sur qui/quoi la variable aléatoire est mesurée : '1 patient' → **Unité statistique**

Variables aléatoires (VA)

- ▶ Notations :
 - ▶ Variable aléatoire : X
 - ▶ Ses valeurs : x_i
- ▶ Caractérisée par son **espérance** et sa **variance**
 - ▶ **Espérance ($E(X)$)** : caractérise la tendance centrale, la valeur moyenne, prise par la VA
 - ▶ **Variance ($V(X)$)** : caractérise la dispersion de la variable autour de son espérance

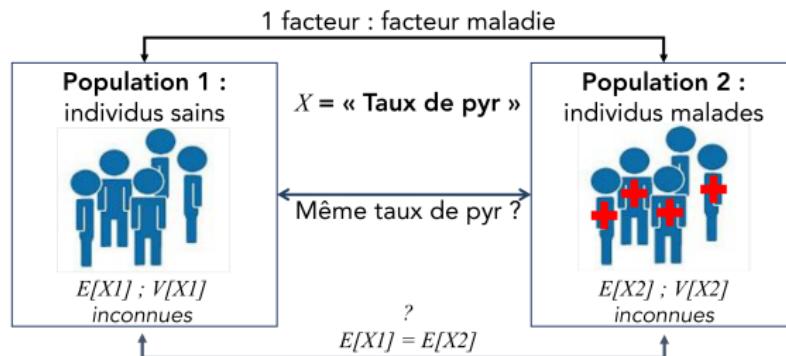


Cas d'étude

Question biologique

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

- ▶ X = “Taux de pyr”
- ▶ US = un individu
- ▶ type = quantitative continue



Cas d'étude : présentation des données

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

```
[1] 218 38
```

```
colnames(dataMyelom)
```

```
[1] "numero"      "diagn"        "stade"        "sstade"       "statut"
[27] "crp"          "ploid"         "tt1"          "etatdn"       "agedg"
```

```
str(dataMyelom[,c("diagn", "pyr")])
```

```
'data.frame': 218 obs. of 2 variables:
$ diagn: int 1 1 1 1 1 1 1 1 1 ...
$ pyr  : num 5.48 8.93 6.1 13.29 5.35 ...
```

Description des deux échantillons

- ▶ taille des échantillons

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

0	1
40	178

- ▶ échantillon 1 : 40 individus sains

```
ind.sain <- which(dataMyelom[, "diagn"]==0)
pyr0 <- dataMyelom[ind.sain, "pyr"]
```

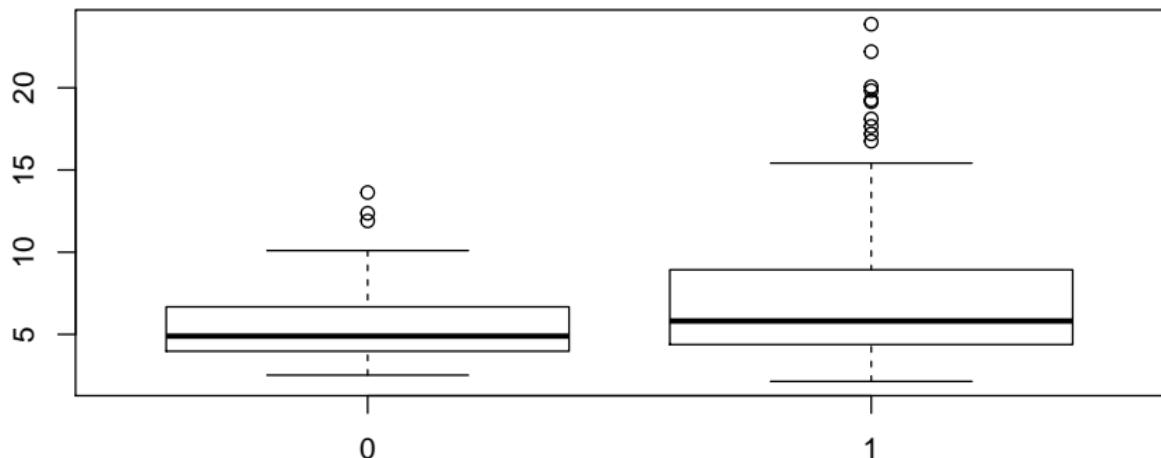
- ▶ échantillon 2 : 178 patients malades

```
ind.malade <- which(dataMyelom[, "diagn"]==1)
num1 <- dataMyelom[ind.malade, "num"]
```

Description des deux échantillons

- ▶ distribution de pyr dans les deux échantillons

```
par(mar=c(3,3,1,1))  
boxplot(pyr~diagn, data=dataMyelom)
```



Estimation de paramètres

- ▶ estimateur de l'espérance :

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

- ▶ estimateur de la variance :

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

- ▶ calcul de la moyenne de pyr dans les deux échantillons

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

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

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

Estimation de paramètres

- ▶ estimateur de l'espérance :

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

- ▶ estimateur de la variance :

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

- ▶ calcul de la variance de pyr dans les deux échantillons

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

```
diagn: 0
[1] 2.527974
```

```
diagn: 1
[1] 4.089944
```

Quelle confiance donner aux résultats ?

- ▶ 1 population : $X \sim \mathcal{N}(\mu = 14.3; \sigma^2 = 24.6)$
- ▶ Tire 1000 échantillons dans la population et calcule la moyenne de X dans les échantillons

```
nb.simul <- 1000 ; val.mean <- rep(NA, length = nb.simul)
for(i in 1: nb.simul){
  ech <- rnorm(n = 100, mean=14.3, sd = sqrt(24.6)) ;  val
}
hist(val.mean,xlab="estimateur de mu", main=paste(nb.simul,
```

1000 simulations



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
- ▶ Prend en compte
 - ▶ la **variabilité** des données
 - ▶ la **taille de l'échantillon**
- ▶ **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

- ▶ si $n > 30$:

$$IC_{1-\alpha} = M \pm z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{s^2}{n}}$$

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

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

$$IC_{1-\alpha} = M \pm t_{\alpha; (n-1)\text{ddl}} \times \sqrt{\frac{s^2}{n}}$$

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

IC du taux de pyr

- ▶ échantillon 1 : 15 patients sains

```
alpha <- 0.5
borneInf.0 <- mean(pyr0) - qt(1-alpha/2, df=(length(pyr0)-1))
borneSup.0 <- mean(pyr0) + qt(1-alpha/2, df=(length(pyr0)-1))
round(c(borneInf.0, borneSup.0),2)
```

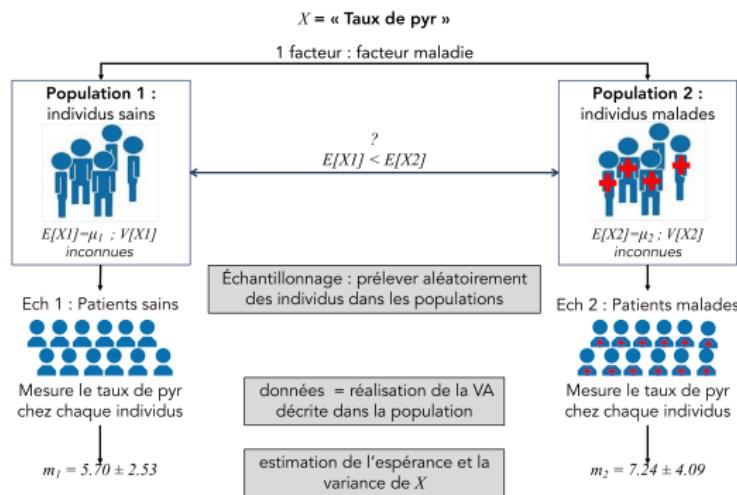
[1] 5.43 5.97

- ▶ échantillon 2 : 14 patients atteints de MM

```
borneInf.1 <- mean(pyr1) - qt(1-alpha/2, df=(length(pyr1)-1))
borneSup.1 <- mean(pyr1) + qt(1-alpha/2, df=(length(pyr1)-1))
round(c(borneInf.1, borneSup.1),2)
```

[1] 7.03 7.45

Cas d'étude : Introductions 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) : H0 (hypothèse nulle)
 - ▶ elles s'expliquent par la maladie : H1

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)** : égalité des paramètres :
 $\mu_1 < \mu_2$
Les différences observées sont expliquées uniquement par le hasard et par le facteur

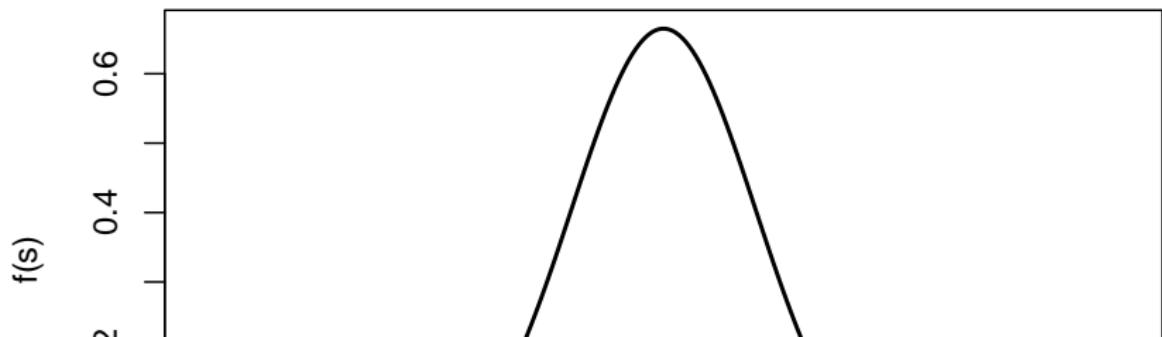
Réaliser le test va consister à choisir une des deux hypothèses (H_0 ou H_1) en se basant sur les données des échantillons

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

- ▶ si oui → non rejet de H_0

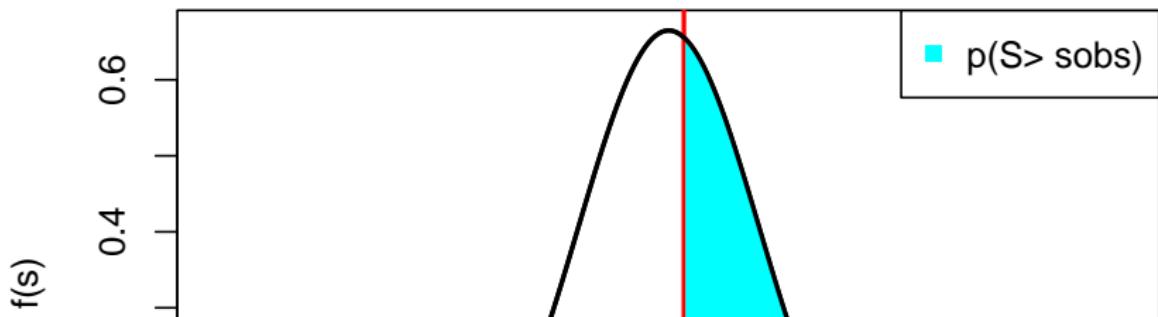
Est-ce que les données sont compatibles avec H0 ?

- ▶ **H0 : Hypothèse de référence** : $\mu_1 = \mu_2$
→ calculer la probabilité d'obtenir sous H0 le même résultat que celui observé sur les échantillons = **p-value**
- 1. Définir un **critère statistique** S dont la loi sous H0 est connue
 $S = M_1 - M_2$ avec M =moyenne de X dans 1 échantillon
- 2. Calculer $s_{obs} =$ valeur de S calculée sur les échantillons :
 $s_{obs} = m_1 - m_2$



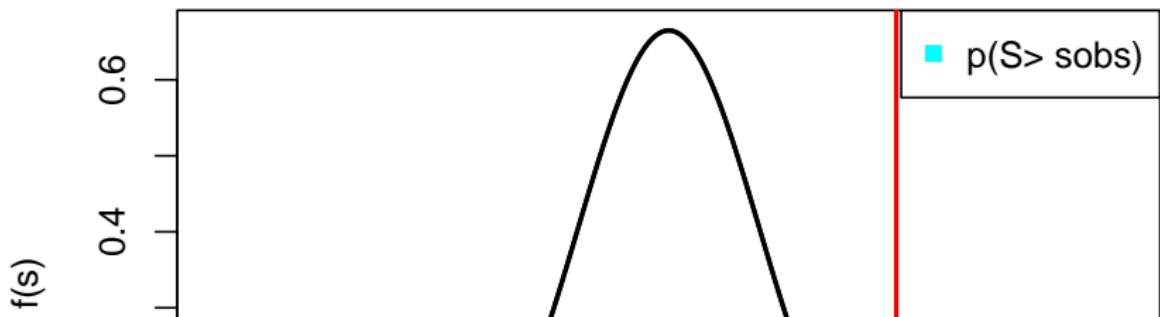
Est-ce que les données sont compatibles avec H0 ?

- ▶ **H0 : Hypothèse de référence** : $\mu_1 = \mu_2$
 → calculer la probabilité d'obtenir sous H0 le même résultat que celui observé sur les échantillons = **p-value**
1. Définir un **critère statistique** S dont la loi sous H0 est connue
 $S = M_1 - M_2$ avec M =moyenne de X dans 1 échantillon
 2. Calculer s_{obs} = valeur de S calculée sur les échantillons :
 $s_{obs} = m_1 - m_2$



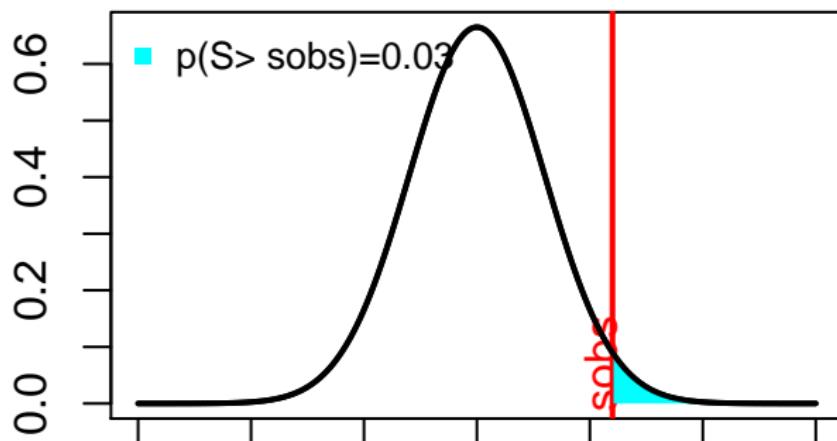
Est-ce que les données sont compatibles avec H0 ?

- ▶ **H0 : Hypothèse de référence** : $\mu_1 = \mu_2$
→ calculer la probabilité d'obtenir sous H0 le même résultat que celui observé sur les échantillons = **p-value**
- 1. Définir un **critère statistique** S dont la loi sous H0 est connue
 $S = M_1 - M_2$ avec M =moyenne de X dans 1 échantillon
- 2. Calculer $s_{obs} =$ valeur de S calculée sur les échantillons :
 $s_{obs} = m_1 - m_2$



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 une différence plus grande) que celle observée



les erreurs

Deux erreurs possibles quand on conclut au test :

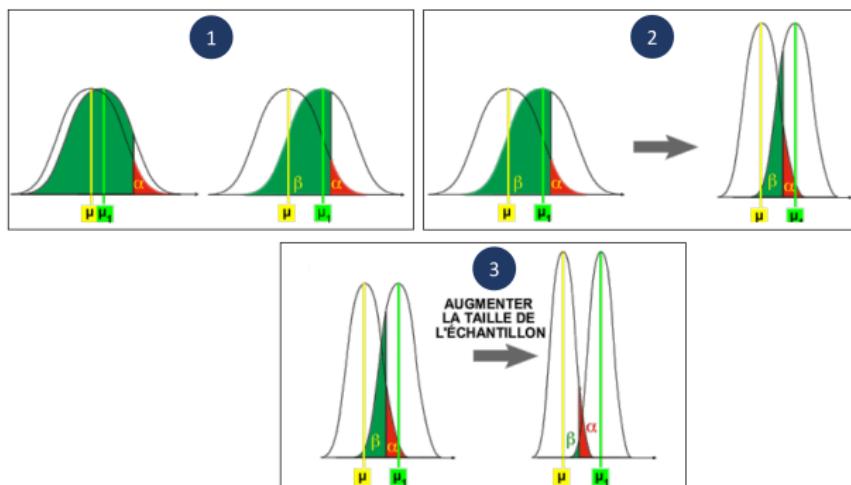
- ▶ **erreur de type I** : je conclus à 1 effet de la maladie alors qu'en réalité il n'y en a pas.
- ▶ **erreur de type II** : je conclus à \emptyset effet de la maladie alors qu'en réalité il existe.

		Décision du test	
		Rejet de H0	Non Rejet de H0
Réalité	H0 vraie	α	$1-\alpha$
	H0 fausse	<i>Puissance</i>	β



Diminuer l'erreur de type II / Augmenter la puissance du test

1. Augmenter la taille d'effet
2. Diminuer la variabilité des données (variance des données)
3. Augmenter la taille de l'échantillon



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

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 < \mu_2$$

```
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 zero
95 percent confidence interval:
```

Les différents types de tests : suivant H1

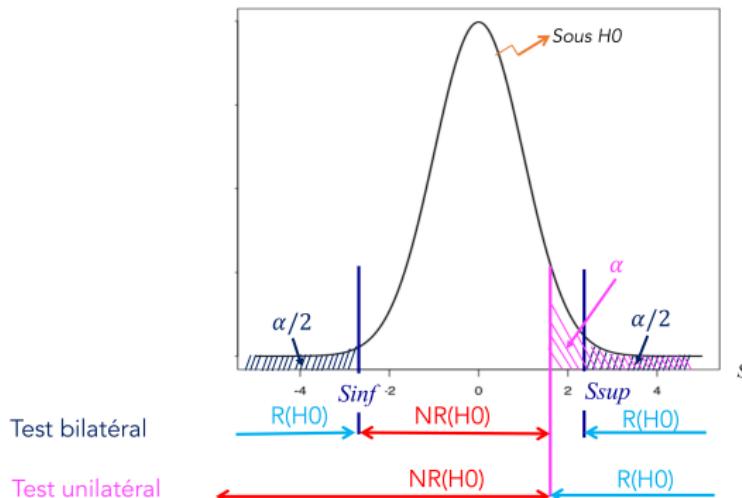
$$H_0 : \mu_1 = \mu_2$$

Test unilatéral : un sens dans l'effet

$$H_1 : \mu_1 \geq \mu_2$$

Test bilatéral : aucun sens dans l'effet

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



Les différents types de tests : suivant la VA

