Chapitre 6

Les tests d'hypothèse

1 - Comparer des moyennes ou des proportions

Une première expérience simple

Ce qu'il faut retenir

Dès qu'on compare deux moyennes empiriques, on en trouve une plus grande que l'autre

$\Rightarrow$ Pas forcément un phénomène biologique

$\Rightarrow$ Peut venir d'un effet d'échantillonnage

ATTENTION:
L'effet d'échantillonnage **ne traduit aucun effet biologique**: ne pas l'interpréter !

Un exemple: effet du sexe sur l'espérance de vie chez une espèce sauvage

Le principe du test d'hypothèse

<u>Exemple de l'effet du sexe sur l'espérance de vie:</u>
On observe une différence de moyenne entre mâles et femelles

<u>Deux hypothèses:</u>
1) Effet d'échantillonnage
2) Effet biologique (meilleure survie d'un sexe)

<u>Principe de parcimonie</u>: de deux explications on choisit toujours la plus simple (l'effet d'échantillonnage car ne fait pas appel à un phénomène biologique incertain)

<u>Autre formulation</u>:
L'effet du sexe sur l'espérance de vie doit être démontré: différence de survie observée trop importante pour être un effet d'échantillonnage

Sinon: prudence… (pas d'effet du sexe)

Le principe du test d'hypothèse

Test d'hypothèse:
L'effet d'échantillonnage seul peut-il expliquer la différence de moyenne (ou de proportion, cf plus loin) observée ?

A) Comparaison à une moyenne théorique

Exemple:
Effet d'une maladie sur X = « concentration d'un certain type de cellule »

Individus non malades: moyenne connue (μ_0= 2.1 g/L)

Individu malades: moyenne plus haute ? Plus basse ?

Expérience: concentration mesurée chez n=40 individus

$$\Rightarrow \boxed{\bar{x} = 2.7}$$

Questions: différence de concentration = effet d'échantillonnage ? AU CONTRAIRE: Effet de la maladie sur X ?

Comparaison à une moyenne théorique

On suppose l'écart type connu (σ=1 g/L).
On note μ=moyenne de X chez les individus malades.
On suppose la normalité des mesures: $X_i \sim N(\mu, \sigma)$

On parle **d'hypothèse nulle H0**:

H0: $\mu = \mu_0$ (=2.1 g/L)

H0 est vraie veut dire que la moyenne observée (2.7 g/L) diffère de la moyenne théorique (2.1 g/L) uniquement à cause de l'effet (ou l'erreur) d'échantillonnage

On parle **d'hypothèse alternative H1:**

H1: $\mu \neq \mu_0$

Si H1 est vrai il y a un effet **REEL** de la maladie sur X

Comparaison à une moyenne théorique

Pour évaluer H0 : écart réduit Z

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Si H0 vrai alors Z ne doit pas être trop grand car Z~N(0,1)

Avec un risque α de se tromper, on a $|Z|<\varepsilon_\alpha$

Décision (z = valeur observée de l'écart réduit Z):

- Si $|z|<\varepsilon_\alpha$: moyenne observée = pas trop loin de moyenne théorique: différence explicable par effet d'échantillonnage
$\Rightarrow$ On accepte H0 (aucun effet de la maladie sur X)

- Si $|z|>\varepsilon_\alpha$: moyenne observée = trop loin de moyenne théorique: différence non explicable par effet d'échantillonnage
$\Rightarrow$ On rejette H0 (effet de la maladie sur X)

Comparaison à une moyenne théorique

Application numérique:
Effet d'une maladie sur X = « concentration d'un certain type de cellule »

H0: $\mu = \mu_0$ (=2.1 g/L)
H1: $\mu \neq \mu_0$

μ_0= 2.1 g/L
$\bar{x} = 2.7 \quad g/L$
σ=1 g/L
n=40
On prend α=0.05, test bilatéral: ε_α=1.96

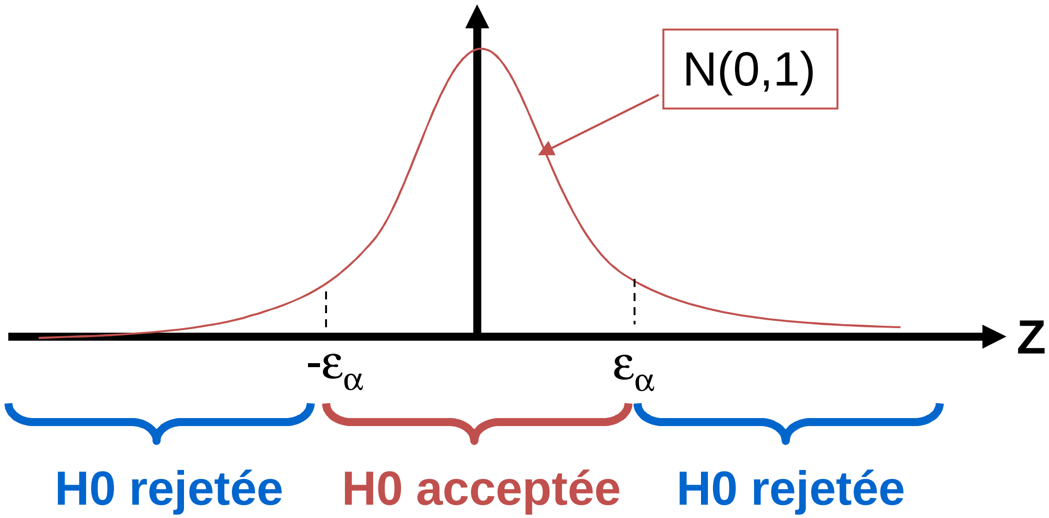
$$\boxed{z = \frac{2.7 - 2.1}{1/\sqrt{40}} = 3.79 > 1.96}$$

$\Rightarrow$ On rejette H0 (risque α=0.05)

$\Rightarrow$ **Conclusion**: effet (positif) de la maladie sur X

Comparaison à une moyenne théorique

Graphiquement



On parle de rejet **bilatéral**

Comparaison à une moyenne théorique

Rejet unilatéral

Exemple d'application:
Mise sur le marché d'un médicament: on souhaite mettre en évidence un effet positif

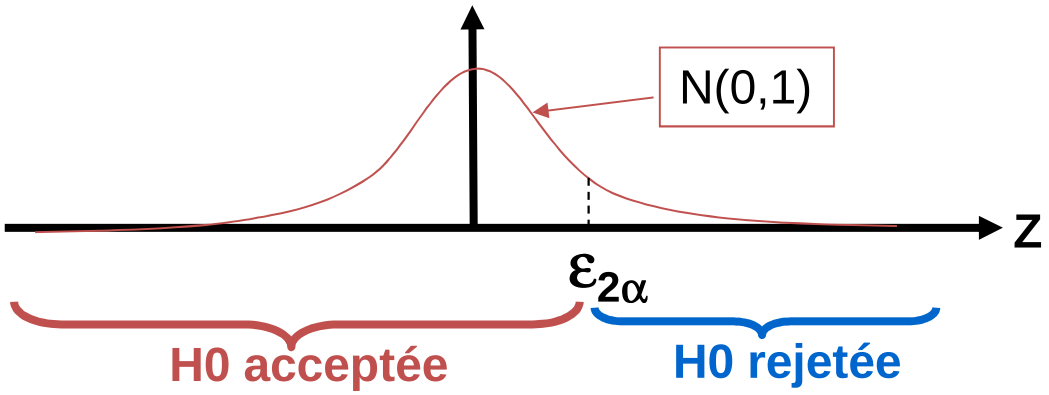
On veut juste savoir si l'effet est positif (effet négatif $\Leftrightarrow$ pas d'effet)

$\Rightarrow$ Test unilatéral droit

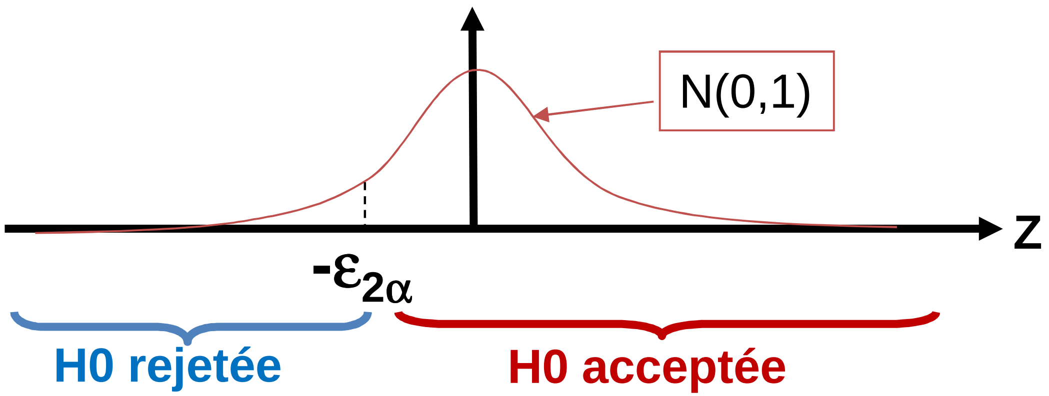
On écrit: H0: $\mu = \mu_0$
H1: $\mu > \mu_0$

Comparaison à une moyenne théorique

Rejet unilatéral



DROIT



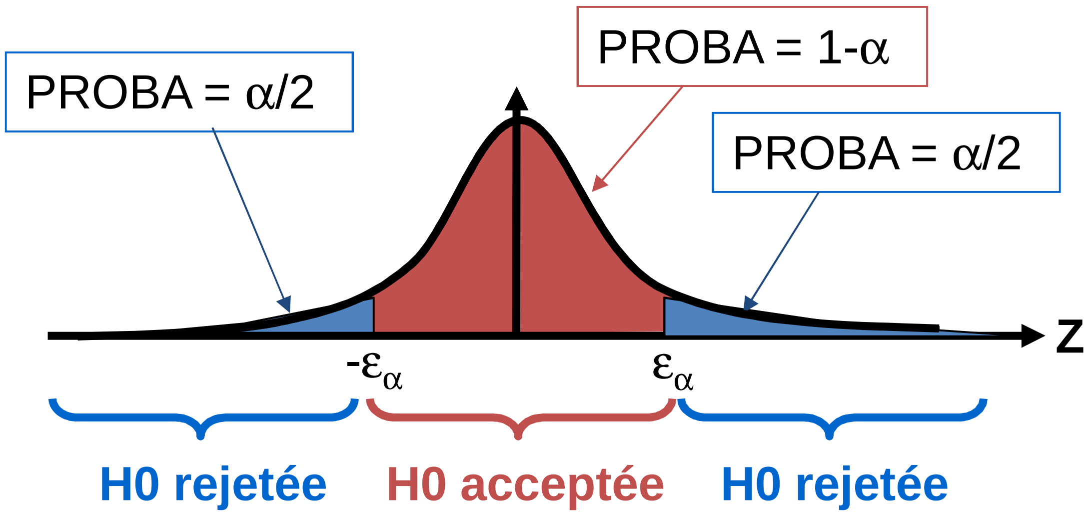
GAUCHE

Comparaison à une moyenne théorique

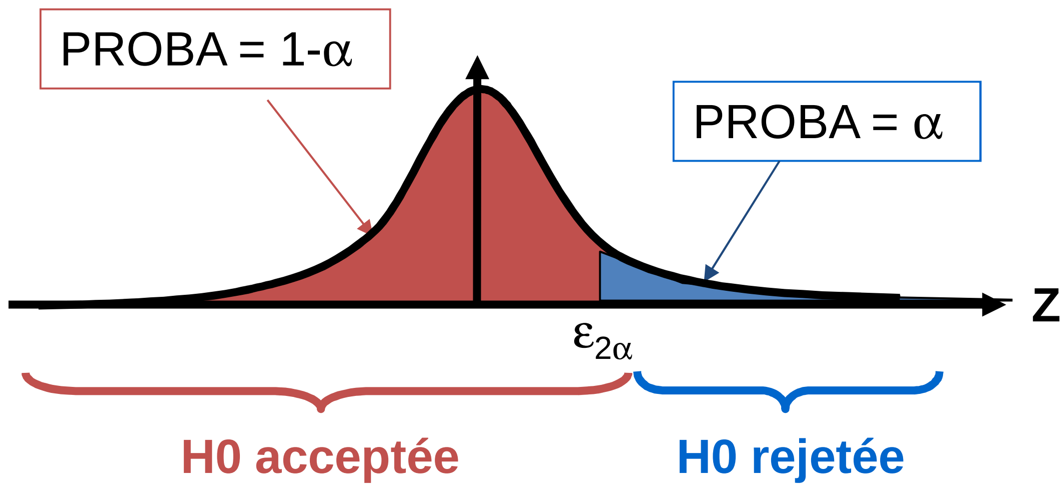
Pourquoi $\varepsilon_{\underline{2\alpha}}$ dans un test unilatéral ???

$\Rightarrow$ Pour conserver un risque d'erreur de α

Comparaison à une moyenne théorique



BILATERAL



UNILATERAL

Généralités sur les tests d'hypothèse

Ils sont basés sur un **raisonnement par l'absurde**

Exemple:

Si Socrate était un chat ← Hypothèse à tester

Alors Socrate pourrait se gratter derrière les oreilles avec ses pattes arrières ← Conséquence mesurable de cette hypothèse

Or, Socrate ne sait pas faire ça ← Incompatibilité avec les observations

Donc Socrate n'est pas un chat ← Conclusion:l'hypothèse est fausse

Généralités sur les tests d'hypothèse

Application à la comparaison à une moyenne théorique

Hypothèse à tester	Si H0 ($\mu=\mu_0$) est vraie
Conséquence mesurable de cette hypothèse	Alors l'écart réduit (z) ne doit pas être trop grand ($\lvert z\rvert<\varepsilon_\alpha$)
Incompatibilité avec les observations	Or $\lvert z\rvert>\varepsilon_\alpha$
Conclusion:l'hypothèse est fausse	Donc H0 est fausse

Généralités sur les tests d'hypothèse

Dans l'autre sens ça ne marche pas

<u>Exemple:</u>

Si Socrate était un chat	Si H0 ($\mu=\mu_0$) est vraie
Alors Socrate aimerait le lait	Alors l'écart réduit (z) ne doit pas être trop grand ($\lvert z\rvert<\varepsilon_\alpha$)
Or, Socrate bien du lait	Or on a bien $\lvert z\rvert<\varepsilon_\alpha$
~~Donc Socrate est un chat~~	~~Donc H0 est vrai~~

<u>Moralité (IMPORTANT):</u> Le fait d'accepter H0 ne veut pas dire qu'elle est vrai. Il est plus correct de dire **« on ne peut pas rejeter H0 »**

Exemple:
Effet d'une (autre) maladie sur X = « concentration d'un certain (autre) type de cellule »

H0: $\mu = \mu_0$ (=2.1 g/L)
H1: $\mu \neq \mu_0$

μ_0= 2.6 g/L
$\bar{x} = 2.8 \;\; g / L$
σ=1.3 g/L
n=35
On prend α=0.05, test bilatéral: ε_α=1.96

$$\boxed{z = \frac{2.8 - 2.6}{1.3 / \sqrt{35}} = 0.91 < 1.96}$$

$\Rightarrow$ On ne peut pas rejeter H0

Conclusions: Il est effectivement possible que μ vaille 2.6 g/L

Si on calcule $IC_\alpha(\mu)$=[2.37; 3.23]
$\Rightarrow$ Toutes les valeurs comprises entre 2.37 et 3.23 sont également possibles

Par exemple il est possible que μ vaille 3.0 g/L (et donc H0 serait fausse)

Parfois la différence de moyennes existe mais est trop faible pour être détectée

Exemple: si μ_0 = 2.4 et μ= 2.45

Dans ce cas H0 est fausse mais le test conduira probablement à ne pas la rejeter

$\Rightarrow$ On dit que le test manque de puissance

Risque de deuxième espèce β=P(accepter H0 / H0 fausse)
(très difficile à calculer en pratique)

Bilan: les risques en statistiques

	H0 vraie	H0 fausse
H0 acceptée	Test OK (proba = 1 - α)	Test pas OK (proba = β)
H0 rejetée	Test pas OK (proba = α)	Test OK (proba = 1 - β)

Comparaison à une moyenne théorique: différents cas de figure

(même problème que plus tôt quand on ne connaît pas la variance: on l'estime)

Est-ce que n≥30 ?

- OUI
  - **Connaît-on la variance ?**
    - NON
      - On estime: $\hat{\sigma}^2 = \frac{n}{n-1} s^2$
      - On considère $z = \frac{\bar{x} - \mu_0}{\hat{\sigma} / \sqrt{n}}$
      - Comparé à ε_α (Table loi normale)
    - OUI
      - On considère $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
      - Comparé à ε_α (Table loi normale)
- NON
  - On doit supposer la normalité de la variable mesurée (Xi~N(μ,σ))
  - **Connaît-on la variance ?**
    - NON
      - On estime: $\hat{\sigma}^2 = \frac{n}{n-1} s^2$
      - On considère $t = \frac{\bar{x} - \mu_0}{\hat{\sigma} / \sqrt{n}}$
      - Comparé à t_α^{n-1} (Table loi student n-1 ddl)
    - OUI
      - On considère $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
      - Comparé à ε_α (Table loi normale)

B) Comparaison à une proportion théorique

Exemple du début du cours:

Maladie qui tue p_0 = 50% des souris sans traitement

Traitement testé sur 40 souris
$\Rightarrow$ 23 survivantes

Effet du traitement ou simple effet d'échantillonnage ?

On va tester l'hypothèse

H0: $p=p_0$ (aucun effet du traitement)
H1: $p \neq p_0$ (effet du traitement, sens non précisé: test bilatéral)

Comparaison à une proportion théorique

On appelle X = nombre de souris survivantes

Sous H0: $X \sim B(n,p_0)$, avec $n = 40$ et $p_0 = 0.5$

Approximation: Si $n \geq 30$, $np_0 \geq 5$ et $n(1-p_0) \geq 5$

Donc $X \sim N(np_0, \sqrt{np_0(1-p_0)})$

Et donc $\boxed{\hat{p} = \frac{X}{n} \sim N(p_0, \sqrt{\frac{p_0(1-p_0)}{n}})}$

On centre et on réduit pour obtenir l'écart réduit: $\boxed{Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}}$

Si H0 est vraie, Z est distribué comme une N(0,1) et l'écart réduit observé z ne doit pas dépasser le seuil d'une N(0,1)

$\Rightarrow$ On compare z à ε_α

Comparaison à une proportion théorique

Application numérique:

$$\left.\begin{aligned} p_0 &= 0.5 \\ n &= 40 \\ \hat{p} &= 23/40 = 0.575 \end{aligned}\right\} \Rightarrow \boxed{z = \frac{0.575 - 0.5}{\sqrt{\dfrac{0.5 \times 0.5}{40}}} = 0.95 < 1.96}$$

Si on prend un risque α=0.05 $\Rightarrow$ H0 non rejetée: on ne peut pas conclure à un effet positif du traitement (données insuffisantes ou aucun effet ??? On ne peut pas savoir)

Conditions d'application: on a bien $n=40 \geq 30$, $np_0 = 20 \geq 5$ et $n(1-p_0)=20 \geq 5$

Combien de survivants aurait-il fallu observer pour conclure à un effet du traitement ?

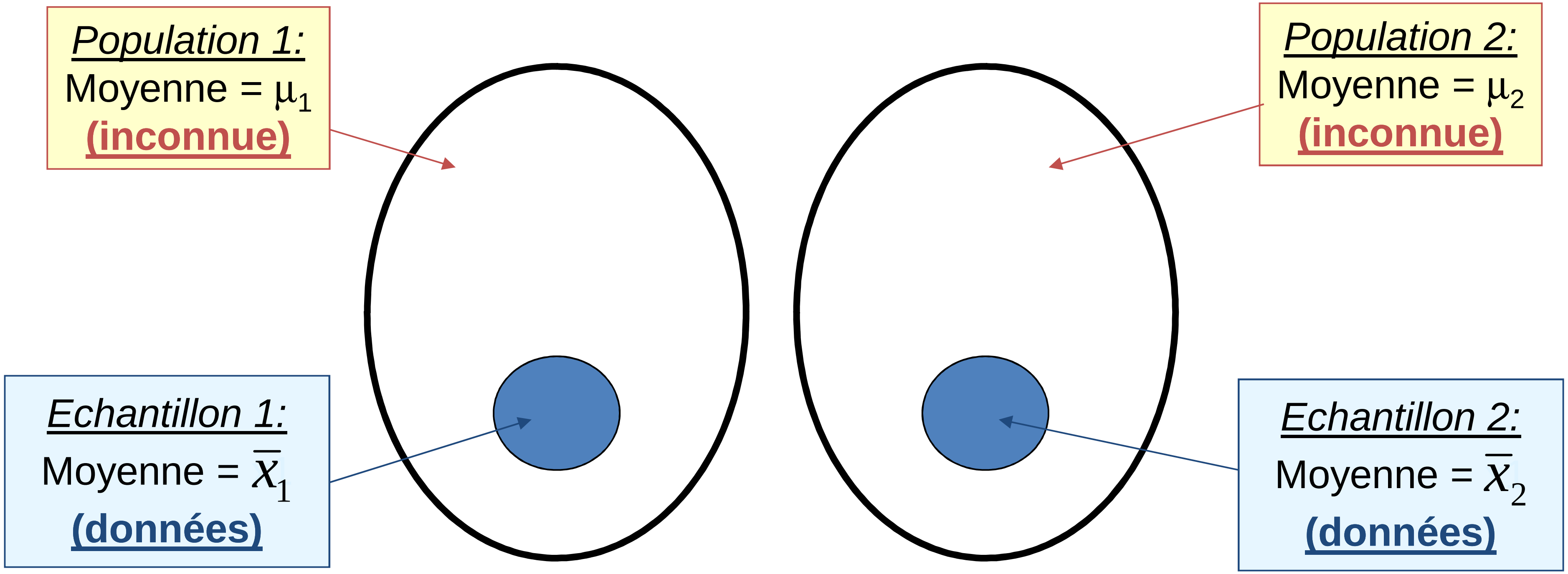
$$24 \;\; \textit{survivants} \quad \Rightarrow \quad z = 1.26 < 1.96$$

$$25 \;\; \textit{survivants} \quad \Rightarrow \quad z = 1.58 < 1.96$$

$$26 \;\; \textit{survivants} \quad \Rightarrow \quad z = 1.89 < 1.96$$

$$\boxed{27 \;\; \textit{survivants} \quad \Rightarrow \quad z = 2.21 > 1.96} \Leftarrow$$

C) Comparaison de deux moyennes observées



Question: la différence entre $\overline{X}_1$ et $\overline{X}_2$ traduit-elle un effet biologique ($\mu_1 \neq \mu_2$) ?
Un effet d'échantillonnage (avec $\mu 1 = \mu 2$) ?

$\Rightarrow$ Différence de moyennes significative entre les deux populations ?

Exemple: différence d'espérance de vie entre mâles et femelles d'une espèce
Population 1 = les mâles de l'espèce, population 2 = les femelles de l'espèce

Comparaison de deux moyennes observées

Test d'hypothèse: quelle différence maximum peut-on attendre entre $\bar{x}_1$ et $\bar{x}_2$ si les deux moyennes sont bien égales (simple effet d'échantillonnage)

On note X1 = « durée de vie des mâles », X2 = « durée de vie d'une femelle »
($\mu_1 = E(X1)$, $\mu_2 = E(X2)$)

H0: $\mu_1 = \mu_2$
H1: $\mu_1 \neq \mu_2$

On mesure la différence de moyennes observée en supposant:

1) Que les variables mesurées (X1 et X2) suivent des lois normales
2) Que la variance de ces lois normales est connue

$$\boxed{X1 \sim N(\mu_1, \sigma_1) \qquad X2 \sim N(\mu_2, \sigma_2)}$$

Comparaison de deux moyennes observées

On tire n_1 individu dans la population 1 (mâles) et n_2 dans la population 2 (femelles)

On détermine la loi de probabilité des moyennes observées

$$\overline{X}_1 \sim N(\mu_1, \sigma_1 / \sqrt{n_1}) \qquad \overline{X}2 \sim N(\mu_2, \sigma_2 / \sqrt{n_2})$$

Ecart entre les deux moyennes observées

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

Si H0 est vraie, alors $\mu_1 = \mu_2$ et donc

$$\overline{X}_1 - \overline{X}_2 \sim N(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

Comparaison de deux moyennes observées

On défini Z qui mesure l'écart réduit entre les deux moyennes observées

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Si H0 est bien vraie, alors Z à une probabilité 1-α d'être supérieure à la valeur seuil ε_α d'une N(0,1)

$\Rightarrow$ On compare z (écart réduit observé) à ε_α

Remarque: on ne doit faire l'hypothèse de normalité que si n < 30

Un exemple…

Comparaison de deux moyennes observées

Dans la pratique on connaît rarement la variance

$\Rightarrow$ Estimation

Pour des raisons mathématiques, on doit distinguer:

1) $n \geq 30$, on estime alors les deux variances séparément

2) $n < 30$, on doit supposer une variance commune à X1 et X2 (que l'on estime)

Comparaison de deux moyennes observées

Cas variance inconnue, n ≥ 30

On estime:

$$\hat{\sigma}_1^{\ 2} = \frac{n_1}{n_1 - 1} s_1^{\ 2}$$

(s_1^2 = variance observée dans l'échantillon 1)

$$\hat{\sigma}_2^{\ 2} = \frac{n_2}{n_2 - 1} s_2^{\ 2}$$

(s_2^2 = variance observée dans l'échantillon 2)

On a alors

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\hat{\sigma}_1^2}{n_1} + \dfrac{\hat{\sigma}_2^2}{n_2}}} \sim N(0,1)$$

On compare z à la valeur seuil d'une N(0,1) = ε_α

Comparaison de deux moyennes observées

Cas variance inconnue, n < 30

On estime une variance commune:

$$\hat{\sigma}^2 = \frac{n_1 s_1{}^2 + n_2 s_2{}^2}{n_1 + n_2 - 2} = \frac{(n_1 - 1)\hat{\sigma}_1{}^2 + (n_2 - 1)\hat{\sigma}_2{}^2}{n_1 + n_2 - 2}$$

On a alors

$$T = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T\ (n_1 + n_2 - 2\ \ ddl)$$

(c'est des maths)

On compare t à la valeur seuil d'une T(n_1+n_2-2 ddl) = $t_\alpha^{n1+n2-2}$

Comparaison de deux moyennes observées

Exemple:

Y a-t-il une différence significative dans le jour de naissance moyen entre les étudiants de la gauche de l'amphi et ceux de droite ?

Récapitulatif

Est-ce que n≥30 ?

OUI

Connaît-on la variance ?

NON

On estime:

$$\hat{\sigma}_1^{\,2} = \frac{n_1}{n_1 - 1} s_1^{\,2}$$

$$\hat{\sigma}_2^{\,2} = \frac{n_2}{n_2 - 1} s_2^{\,2}$$

On considère

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}}$$

Comparé à ε_α
(Table loi normale)

OUI

On considère

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Comparé à ε_α
(Table loi normale)

NON

On doit supposer la normalité des variables mesurées

Connaît-on la variance ?

NON

On suppose $\sigma_1 = \sigma_2 = \sigma$ et on estime σ par :

$$\hat{\sigma}^2 = \frac{n_1 s_1^{\,2} + n_2 s_2^{\,2}}{n_1 + n_2 - 2}$$

On considère

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Comparé à $t_\alpha^{n1+n2-2}$
(Table loi student n_1+n_2-2 ddl)

OUI

On considère

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Comparé à ε_α
(Table loi normale)

D) Comparaison de deux fréquences observées

Exemple:

On souhaite savoir si le fait de fumer augmente le risque de développer une certaine maladie

Echantillonnage
n_1=197 non fumeurs, 12 malades
n_2=178 fumeurs, 23 malades

On calcule la fréquence de la maladie pour estimer la proportion de malades dans chaque groupe:

$$\hat{p}_1 = f_1 = \frac{12}{197} = 0.061$$

$$\hat{p}_2 = f_2 = \frac{23}{178} = 0.129$$

A priori: fumeurs plus malades, mais il faut se méfier des effets d'échantillonnage

Comparaison de deux fréquences observées

On note:

p_1 = proportion réelle d'individus malades chez les non fumeurs (population 1)
p_2 = proportion réelle d'individus malades chez les fumeurs (population 2)

Les fréquences observées sont des estimations des proportions réelles

Si $n_1 \geq 30$, $n_2 \geq 30$, $n_1p_1 \geq 5$, $n_1(1-p_1) \geq 5$, $n_2p_2 \geq 5$, $n_2(1-p_2) \geq 5$ on a:

$$\hat{p}_1 \sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \qquad \hat{p}_2 \sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right)$$

On s'intéresse à la différence des deux fréquences observées:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right)$$

Comparaison de deux fréquences observées

On va tester l'hypothèse:

H0: $p_1=p_2$ (aucun effet du fait de fumer sur la maladie)
H1: $p_1 \neq p_2$ (effet du fait de fumer sur la maladie)

L'écart entre les deux fréquences observées devrait alors être:

$$\boxed{\hat{p}_1 - \hat{p}_2 \sim N\left(0, \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right)}$$

On passe à l'écart réduit:

$$\boxed{Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)}$$

Comparaison de deux fréquences observées

Problème: on ne connaît pas la proportion commune p de malades dans les populations

$\Rightarrow$ Estimation

$$\boxed{\hat{p} = f = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2} = \frac{nombre\ de\ cas\ \ positif}{nombre\ de\ cas\ total}}$$

On remplace p par son estimation, on montre que (maths):

$$\boxed{Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)}$$

On compare z à ε_α (table loi normale)

Comparaison de deux fréquences observées

Application numérique:

$$\hat{p}_1 = 0.061$$
$$\hat{p}_2 = 0.129$$
$$\hat{p} = \frac{12+23}{197+178} = 0.093$$
$$n_1 = 197$$
$$n_2 = 178$$

$$z = \frac{0.061-0.129}{\sqrt{0.093*(1-0.093)*\left(\frac{1}{197}+\frac{1}{178}\right)}} = -2.17$$

|z|=2.17 > 1.96 donc on rejette H0 (risque α=0.05)

$\Rightarrow$ fumer augmente la probabilité de développer la maladie

conditions d'application:

$$\mathrm{n}_1 = 197 \geq 30, \;\; \mathrm{n}_2 = 178 \geq 30$$
$$\mathrm{n}_1 \hat{p} = 18.4 \geq 5, \;\; \mathrm{n}_1(1-\hat{p}) = 178.6 \geq 5$$
$$\mathrm{n}_2 \hat{p} = 16.6 \geq 5, \;\; \mathrm{n}_2(1-\hat{p}) = 161.4 \geq 5$$