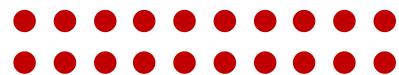


Notes de cours de Probabilités et statistiques



GI(S3)-2021/2022

1

Chapitre 2



Statistique inférentielle & Théorie des Tests

II- Echantillonnage

Soit (Ω, Θ, P) un espace de probabilité et \mathbb{R} un corps des nombres réels.

Définition 1: On dit qu'une application

$$\begin{aligned} \mathbf{X} : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow \mathbf{X}(\omega) \end{aligned}$$

telle que à chaque événement élémentaire (ω) fait correspondre un nombre réel, est *une variable aléatoire* si, pour tout nombre réel x ,

$$\mathbf{A} = (\mathbf{X} \leq x) = \{\omega \mid \mathbf{X}(\omega) \leq x\} \in \Theta \quad (= \text{ensemble de tous les évènements possibles})$$

c'est à dire \mathbf{A} est un événement.

II- Echantillonnage (suite)

Définition 2:

Soit \mathbf{X} une variable aléatoire sur un référentiel Ω . Un échantillon de \mathbf{X} de taille n est un n -uplets $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ de variables aléatoires indépendantes de même loi que \mathbf{X} .

Une réalisation de cet échantillon est un n -uplet de réels

(x_1, \dots, x_n) où $\mathbf{X}_i(\omega) = x_i$.

III- Estimateurs et estimations

- On s'intéresse à la caractéristique X d'une population (éventuellement à un vecteur de caractéristiques), dont la loi dépend d'un paramètre inconnu θ .
- On note $f_\theta(x)$ la densité de la loi de X au point x (resp. la loi $P_\theta(X=x)$ de X au point x) si X est continue (resp. si X est discrète).
- On souhaite estimer θ d'une population (cela peut être sa moyenne μ , son écart-type σ , une proportion p) .

III. Estimateurs et estimations (suite)

- **Un estimateur** de θ est une statistique $\hat{\theta}_n(X_1, \dots, X_n)$ dont la réalisation est envisagée comme une “*bonne valeur*” du paramètre θ .
- On parle d'**estimation** de θ associée à cet estimateur la valeur observée lors de l’expérience, c'est-à-dire la valeur prise par la fonction $\hat{\theta}_n(X_1, \dots, X_n)$ au point observé (x_1, \dots, x_n) (i.e.: $\hat{\theta}_n(x_1, \dots, x_n)$).

Exemples d'estimateurs

1) Estimateur de la moyenne empirique (\bar{X}_n):

Définition 3: On appelle moyenne de l'échantillon ou *moyenne empirique*, la statistique notée \bar{X}_n définie par:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Définition 4: On appelle *statistique* sur un n -échantillon *une fonction* de (X_1, \dots, X_n) .

Exemples d'estimateurs (suite)

2) Estimateur de la variance empirique ($\tilde{s}^2(X)$):

Définition 5: On appelle *Variance empirique*, la statistique notée $\tilde{s}^2(X)$ définie par: $\tilde{s}^2(X) :=$

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

Définition 6: Pour une réalisation donnée de l'échantillon aléatoire, $\tilde{s}^2(x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$ est l'estimation associée.

IV. Convergence des variables aléatoires

Par nature, les probabilités s'intéressent aux phénomènes limites : que se passe-t-il lorsque l'on réalise à la suite un très grand nombre d'expériences aléatoires?

Définition 7:

- ❖ La suite $(\mathbf{X}_n)_n$ converge en probabilité vers X *si*:

$$\forall \varepsilon > 0, \lim_{n \rightarrow +\infty} \mathbb{P}(\{\omega | \mathbf{X}_n(\omega) - \mathbf{X}(\omega)| > \varepsilon\}) = 0$$

- ❖ La suite $(\mathbf{X}_n)_n$ converge presque sûrement vers X *si*, pour presque tout $\omega \in \Omega$: $\lim_{n \rightarrow +\infty} \mathbf{X}_n(\omega) = \mathbf{X}(\omega)$.

- ❖ La suite $(\mathbf{X}_n)_n$ converge en loi vers X *si*, $\lim_{n \rightarrow +\infty} \mathbf{F}_n(x) = \mathbf{F}(x)$ avec \mathbf{F}_n est la fonction de répartition de \mathbf{X}_n et \mathbf{F} est celle de X.

Le théorème central – limite

Théorème 1:

Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires *i.i.d* et notons \bar{X}_n sa moyenne empirique. Si $\text{Var}(X_1) < +\infty$.

Alors:
$$\frac{\left(\bar{X}_n - E(X_1)\right)}{\sqrt{\text{Var}(X_1) / n}} \xrightarrow{L} N(0, 1)$$

Remarque:

C'est ce théorème qui affirme que la loi normale est la loi des phénomènes naturels.

Propriétés des estimateurs

Proposition 1:

Soit X une variable aléatoire de moyenne μ et d'écart-type σ . On a :

$$E(\bar{X}_n) = \mu, \quad \text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

De plus, par *le théorème central limite*, \bar{X}_n converge en loi vers $N(\mu, \frac{\sigma}{\sqrt{n}})$ lorsque n tend vers l'infini.

Propriétés des estimateurs

Proposition 2:

Soit X une variable aléatoire d'écart-type σ et de moment centré d'ordre 4, μ_4 . On a:

$$E(\tilde{S}^2) = \frac{n-1}{n} \sigma^2,$$

$$\text{Var}(\tilde{S}^2) = \frac{n-1}{n^3} ((n-1)\mu_4 - (n-3)\sigma^4)$$

Précision d'un estimateur

1) CONVERGENCE:

- Un estimateur $\hat{\theta}_n$ d'une grandeur θ est une fonction qui dépend uniquement du n -échantillon (X_1, \dots, X_n) . Il est dit convergent s'il est “proche” de θ au sens de la convergence en probabilité.
- Formalisation :

Pour tout: $\varepsilon > 0$, $\lim_{n \rightarrow +\infty} \mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) = 0$.

2) Biais d'estimateur:

Définition 8: Soit $\hat{\theta}_n$ un estimateur d'un paramètre θ .

On appelle biais la quantité $E(\hat{\theta}_n) - \theta$.

L'estimateur $\hat{\theta}_n$ est dit sans biais si $E(\hat{\theta}_n) - \theta = 0$. Il est biaisé si et seulement si $E(\hat{\theta}_n) \neq \theta$.

Exemple: Moyenne empirique: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ estimateur de l'espérance mathématique μ et sans biais (**Proposition 1**).

3) Erreur quadratique moyenne

La qualité d'un estimateur se mesure également par l'erreur *quadratique moyenne* (**EQM**) définie par **EQM**= $E\{(\hat{\theta}_n - \theta)^2\}$:

$$\square \quad E\{(\hat{\theta}_n - \theta)^2\} = E\{(\hat{\theta}_n - E(\hat{\theta}_n))^2\} + (E(\hat{\theta}_n) - \theta)^2$$

$$= \text{var}(\hat{\theta}_n) + (\text{biais})^2$$

$$\square \quad \text{Si } \hat{\theta}_n \text{ est sans biais, alors } E\{(\hat{\theta}_n - \theta)^2\} = \text{var}(\hat{\theta}_n).$$

D'où:

- Un "meilleur" estimateur est **sans biais** et de **variance minimum** (efficace).

❖ Résumé: Estimateur d'une moyenne μ

Soit un n -échantillon (X_1, \dots, X_n) issu d'une loi de moyenne μ et de variance σ^2 , toutes deux inconnues, alors:

- D'après la LGN, la moyenne empirique \bar{X}_n est un estimateur *convergent* de μ .
- l'estimateur \bar{X}_n est *sans biais*.
- par *indépendance* : $\text{Var}(\bar{X}_n) = \sigma^2/n$.
- **loi de \bar{X}_n :** si $X \sim N(\mu, \sigma^2)$ alors $\bar{X}_n \sim N(\mu, \sigma^2/n)$.
- lorsque n est grand, d'après le **TCL**, la loi de \bar{X}_n est approchée par une loi normale $N(\mu, \sigma^2/n)$.

❖ Résumé: Estimateur d'une variance

La variance empirique associée à un n -échantillon

(X_1, \dots, X_n) est définie par $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$

□ S_n^2 est un estimateur *convergent* de la variance σ^2

□ S_n^2 est *sans biais.*

□ loi de S_n^2 :

Si $X \sim N(\mu, \sigma^2)$, alors $\frac{(n-1)S_n^2}{\sigma^2}$ suit une loi du chi-deux à $(n-1)$ degrés de libertés $\chi^2_{(n-1)}$

V- Maximum de vraisemblance

Estimation par la méthode du maximum de vraisemblance

Soit X une variable aléatoire réelle de loi paramétrique (discrète ou continue), dont on veut estimer le paramètre θ . Alors on *définit* une fonction f telle que:

$$f(x; \theta) =:$$

- $f_\theta(x)$ si X est une v.a. continue de densité f
- $P_\theta(X=x)$ si X est une v.a. discrète de probabilité ponctuelle P .

V-Maximum de vraisemblance

Définition 9:

La méthode consistant à estimer θ par la valeur qui maximise L (vraisemblance) s'appelle méthode du maximum de vraisemblance:

$$\hat{\theta}_n = \{\theta / L(\hat{\theta}_n) = \sup_{\Theta} L(\theta)\}$$

- ❖ On appelle vraisemblance de θ au vu des observations d'un n -échantillon indépendamment et identiquement distribué, le nombre:

$$\begin{aligned}
 L(x_1, \dots, x_i \dots, x_n; \theta) &= \\
 &= f(x_1; \theta) \times f(x_2; \theta) \times \dots \times f(x_n; \theta) \\
 &= \prod_{i=1}^n f(x_i; \theta)
 \end{aligned}$$

- ❖ Le maximum s'obtient sous la condition nécessaire suivante:

$$\frac{\partial L(x_1, \dots, x_i \dots, x_n; \theta)}{\partial \theta} = 0 \text{ ou}$$

$$\frac{\partial \ln(L(x_1, \dots, x_i \dots, x_n; \theta))}{\partial \theta} = 0$$

- Cette condition permet de trouver la valeur $\theta = \hat{\theta}_n$.
- $\theta = \hat{\theta}_n$ est le maximum local si la condition suffisante est remplie au point critique $\theta = \hat{\theta}_n$:

$$\frac{\partial \mathbf{L}(x_1, \dots, x_i, \dots, x_n; \theta)}{\partial \theta^2} < 0 \text{ ou}$$

$$\frac{\partial \ln (\mathbf{L}(x_1, \dots, x_i, \dots, x_n; \theta))}{\partial \theta^2} < 0$$

Exemple1:

- ❖ Soient X_1, \dots, X_n suivent une loi géométrique de paramètre $\theta = p$, alors la fonction de vraisemblance est:

$$\begin{aligned} L(x_1, \dots, x_i \dots, x_n; p) &= \prod_{i=1}^n p(1-p)^{x_i} \\ &= p^n(1-p)^{\sum x_i - n} \end{aligned}$$

- ❖ La log-vraisemblance vaut:

$$\begin{aligned} \log(L(x_1, \dots, x_i \dots, x_n; \theta)) \\ = n \log(p) + \left(\sum x_i - n \right) \log(1-p) \end{aligned}$$

Exemple1 (suite):

- ❖ L'équation du premier ordre s'écrit:

$$\frac{\partial L(x_1, \dots, x_i \dots, x_n; p)}{\partial p} = \frac{1}{p} - \frac{\sum x_i - n}{1-p} = 0,$$

Ou encore $\frac{1}{p} = \frac{\bar{x}_n - 1}{(1-p)}$ dont la solution est égale:

$p = \frac{1}{\bar{x}_n}$. L'estimateur du maximum de

vraisemblance de p est donc: $\widehat{\theta}_n = \widehat{p}_n = \frac{1}{\bar{X}_n}$.

Exemple2:

- ❖ Soient X_1, \dots, X_n une suite des variables aléatoires suivant la loi normale $N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$, alors la fonction de vraisemblance est:

$$L(x_1, \dots, x_i \dots, x_n; \mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}}$$
$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{\sum(x_i - \mu)^2}{2\sigma^2}\right)$$

- ❖ La log-vraisemblance vaut:

$$\begin{aligned} & \log(L(x_1, \dots, x_i \dots, x_n; \theta)) \\ &= \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \left(\sum (x_i - \mu)^2 \right) \end{aligned}$$

❖ Les équations du premier ordre s'écrivent:

$$\frac{\partial L(x_1, \dots, x_i \dots, x_n; \mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} (\sum (x_i - \mu)^2) = 0$$

$$\frac{\partial L(x_1, \dots, x_i \dots, x_n; \mu, \sigma^2)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4} (\sum (x_i - \mu)^2) = 0$$

La solution de ce système est $\mu = \bar{x}_n$ et $\sigma^2 = \frac{1}{n} \sum (x_i - \mu)^2$. Les estimateurs du maximum de vraisemblance sont donc:

$$\mu = \bar{X}_n \text{ et } \widehat{\sigma^2} = \frac{1}{n} \sum (X_i - \mu)^2.$$

Intervalles de confiance

Intervalles de confiance (1)

- ❖ On dispose d'un n -échantillon (X_1, \dots, X_n) de variables aléatoires suivant une loi dépendant d'un certain paramètre θ inconnu que l'on cherche à **estimer**.
- ❖ Un intervalle de confiance aléatoire de niveau (de confiance) $1-\alpha$ pour θ est la donnée de deux variables aléatoires $A := A(X; \alpha)$ et $B := B(X; \alpha)$ telles que $P(A \leq \theta \leq B) = 1-\alpha$.

Intervalles de confiance (2)

- ❖ Les variables aléatoires A et B constituent les bornes de cet intervalle de confiance aléatoire, que l'on note en général

$$\mathbf{IC}_{1-\alpha}(\theta) = [A, B]$$

- ❖ Lorsque l'échantillon est effectivement observé, et que l'on dispose des données (x_1, \dots, x_n) , on notera

$$\mathbf{ic}_{1-\alpha}(\theta) = [a, b]$$

l'intervalle de confiance réalisé résultant, ou a
 $a = a(x_1, \dots, x_n; \alpha)$ et $b = b(x_1, \dots, x_n; \alpha)$.

Intervalles de confiance (3)

- α est la probabilité que le paramètre θ n'appartienne pas à l'intervalle $\mathbf{ic}_{1-\alpha}$, c'est à dire la probabilité que l'on se trompe en affirmant que $\theta \in \mathbf{ic}_{1-\alpha}$.
- C'est donc une probabilité d'erreur, qui doit être assez petite.
- Les valeurs usuelles de α sont 10%, 5%, 1%, etc...

Intervalles de confiance (4)

- Il semble logique de chercher un intervalle de confiance pour θ de la forme $[\hat{\theta}_n - \varepsilon, \hat{\theta}_n + \varepsilon]$, où $\hat{\theta}_n$ est un estimateur de θ .
- Il reste alors à déterminer ε de sorte que:
 - $P(\hat{\theta}_n - \varepsilon \leq \theta \leq \hat{\theta}_n + \varepsilon) = P(|\hat{\theta}_n - \theta| \leq \varepsilon) = 1 - \alpha$
 - et que ε ne dépendant que des observations et pas de θ .
- **Problème:** Ce n'est pas toujours possible!.

Exemple1: IC_{1-α}(μ), σ² connue

Cas des grands échantillons ($n > 30$) ou des petits échantillons avec hypothèse de normalité

Soit (X_1, \dots, X_n) un n -échantillon de v.a. de loi $N(\mu, \sigma^2)$. La moyenne empirique, \bar{X}_n a pour loi $N(\mu, \sigma^2/n)$.

- On cherche un intervalle de confiance pour μ de la forme donc $[\bar{X}_n - \varepsilon, \bar{X}_n + \varepsilon]$.
- Conformément à ce qui précède, le problème revient, pour α fixé, à chercher ε tel que: $P(|\bar{X}_n - \mu| \leq \varepsilon) = 1 - \alpha$
- Les propriétés élémentaires de la loi normale permettent

d'établir que $Z = \left(\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \right) \sim N(0,1)$. Alors, $P(|\bar{X}_n - \mu| \leq \varepsilon) = P\left(\left|Z\right| \leq \frac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - P\left(\left|Z\right| > \frac{\varepsilon\sqrt{n}}{\sigma}\right) = 1 - \alpha$

- Or la table de la loi normale (*ou le logiciel R*) donne la valeur $z_\alpha = F_Z^{-1}(1 - \frac{\alpha}{2})$ telle que $P(|Z| > z_\alpha) = \alpha$. La fonction quantile F_Z^{-1} , fonction réciproque de la fonction de répartition,

- Par conséquent, $\frac{\varepsilon\sqrt{n}}{\sigma} = z_\alpha$. D'où le résultat:

Un intervalle de confiance de seuil α pour le paramètre μ de la loi $N(\mu, \sigma^2)$ est:

$$[\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}, \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}]$$

- Le problème est que cet intervalle n'est utilisable que si on connaît la valeur de σ .

Sous **R**: z_α est obtenu par la commande `qnorm(1-alpha/2)`.

Exemple 2: $\text{IC}_{1-\alpha}(\mu)$, σ^2 inconnue

Cas des grands échantillons ($n > 30$) ou des petits échantillons
avec hypothèse de normalité

- Une idée naturelle est alors de remplacer σ par un estimateur $\hat{\sigma}$. Mais si on fait cela,

$$P\left(\bar{X}_n - \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2} \leq \mu \leq \bar{X}_n + \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}\right) = P\left(\left|\frac{\bar{X}_n - \mu}{\hat{\sigma}} \sqrt{n}\right| \leq \varepsilon\right)$$

n'est pas égale à $1 - \alpha$, car $\frac{\bar{X}_n - \mu}{\hat{\sigma}} \sqrt{n}$ n'est pas $N(0,1)$.

- Donc $[\bar{X}_n - \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}, \bar{X}_n + \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}]$ n'est pas I.C de seuil α .

```
if(!require("BSDA")){install.packages("BSDA") }

library(BSDA)

z.test(NUTRIAGE$poids ,mu=60 ,sigma.x=12) )
```

Exemple 2: $\text{IC}_{1-\alpha}(\mu)$, σ^2 inconnue (suite) $(n > 30)$ ou des petits échantillons avec hypothèse de normalité

□ On peut cependant résoudre le problème en utilisant le **Théorème de Fisher** : Si X_1, X_2, \dots, X_n sont n variables aléatoires indépendantes et de même loi normale $N(\mu, \sigma^2)$, alors:

- \bar{X}_n a pour loi $N(\mu, \sigma^2/n)$.
- $\frac{n\hat{\sigma}^2}{\sigma^2}$ est de loi de khi deux à $n-1$ degrés de libertés χ_{n-1}^2
- \bar{X}_n et $\hat{\sigma}^2$ sont indépendants
- $\frac{\bar{X}_n - \mu}{\hat{\sigma}} \sqrt{n} = \sqrt{n} \left(\frac{\bar{X}_n - \mu}{S_n} \right)$ est de loi de Student \mathcal{T}_{n-1}
- $S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ variance empirique (sans biais).

Exemple 2: $\text{IC}_{1-\alpha}(\mu)$, σ^2 inconnue

Cas des grands échantillons ($n > 30$) ou des petits échantillons avec hypothèse de normalité

□ D'où les résultats:

Un intervalle de confiance de niveau $(1-\alpha)$ pour la moyenne μ est:

$$\text{ic}_{1-\alpha}(\mu) = [\bar{x} - t_{1-\alpha/2}^{n-1} \frac{s_n}{\sqrt{n}}, \bar{x} + t_{1-\alpha/2}^{n-1} \frac{s_n}{\sqrt{n}}]$$

$$s_n^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Sous R: $t_{1-\alpha/2}^{n-1}$ est obtenu par la commande `qt(1-alpha/2, n-1)`.

IC_{1-α}(μ) pour une moyenne

- ❖ Pour $\alpha = 5\%$, ce résultat signifie que "la vraie moyenne, μ ", de la population a une probabilité de 95% d'être dans cet intervalle. On notera par commodité cet intervalle de confiance IC₉₅.
- ❖ Cas des petits échantillons ($n < 30$) + non normale
 - ▶ **Définition:** *Dans le cas où aucune hypothèse n'est faite sur les données, nous conseillons d'utiliser une approche par bootstrap.*

Applications numériques: TP

Importation des données (R/Rcmdr)

Lecture des données

De nombreux types de jeu de données peuvent être importés :

- ❖ feuilles Excel
- ❖ bases de données
- ❖ fichiers SAS, SPSS, Stata. . .
- ❖ Presse-papier, . . .

Exemple de lecture d'un fichier Excel : **nutriage.xls**

- Un échantillon de personnes âgées résidant en France a été interrogé en 2000 dans le cadre d'une enquête nutritionnelle.
- Données concernent 226 personnes.
- Source : <http://www.biostatisticien.eu/springeR/nutriage.xls>

Importation des données

Exemple: On prend les données du fichier "nutriage.xls"

Variables et codage :

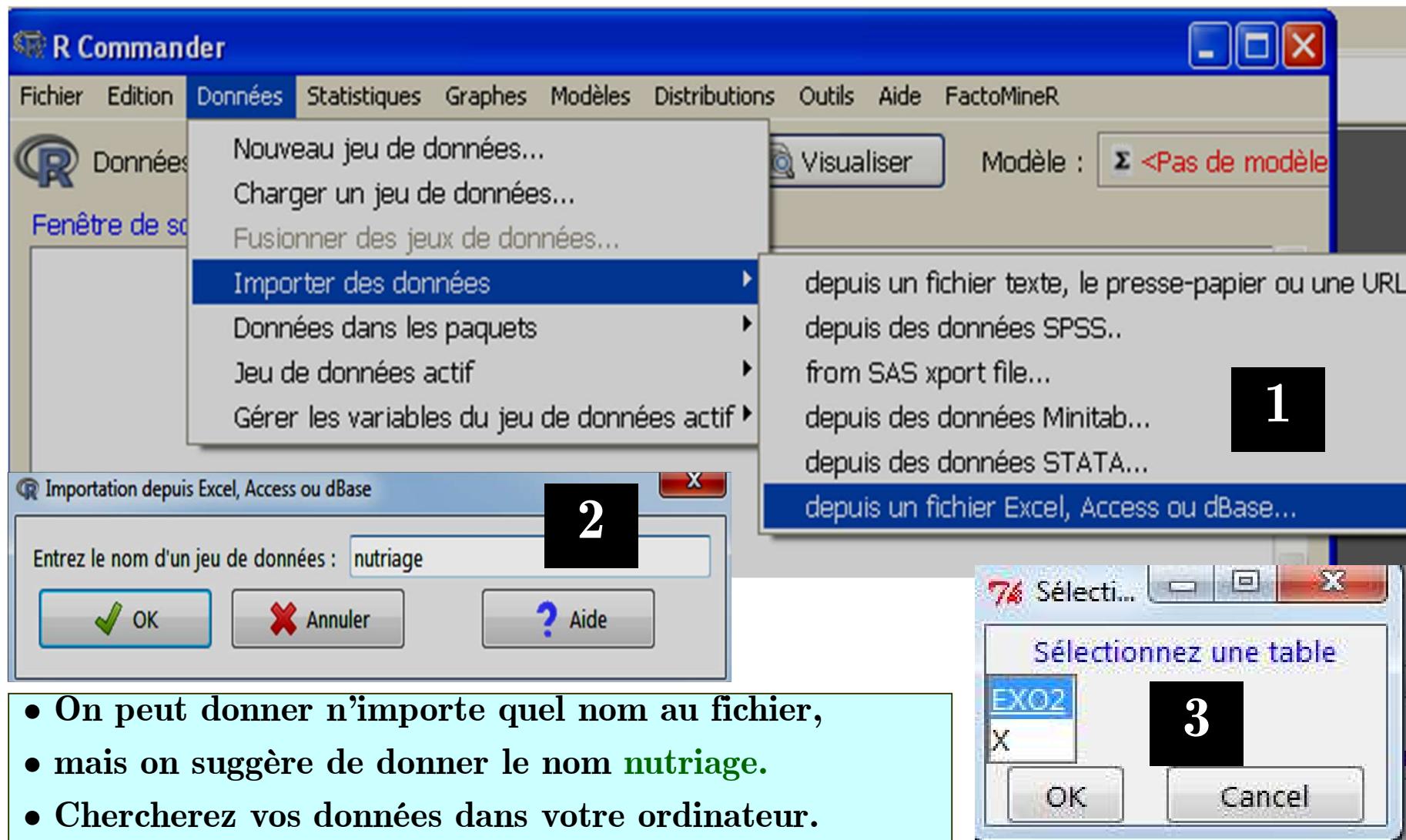
Description	Unité ou Codage	Variable
Sexe	2=Femme ; 1=Homme	sexe
Situation familiale	1=Vit seul 2=Vit en couple 3=Vit dans sa famille 4=Autre type de cohabitation	situation
Consommation journalière de thé	Nombre de tasses	the
Consommation journalière de café	Nombre de tasses	cafe
Taille	Cm	taille
Poids	Kg	poids
Âge le jour de l'entretien	Années	age

Importation des données

Consommation de viande	0=Jamais 1=Moins d'une fois par semaine 2=Une fois par semaine 3=2/3 fois par semaine 4=4/6 fois par semaine 5=Tous les jours	viande
Consommation de poisson	Idem	poisson
Consommation de fruits crus	Idem	fruit_crus
Consommation de fruits et légumes cuits	Idem	fruit_legume_cuits
Consommation de chocolat	Idem	chocol
Matière grasse préférentiellement utilisée pour la cuisson	1=Beurre 2=Margarine 3=Huile d'arachide 4=Huile de tournesol 5=Huile d'olive 6=Mélange d'huile (type Isio4) 7=Huile de colza 8=Graisse de canard ou d'oie	matgras

De Excel à R Commander

Importation de données à partir d'un fichier .xls



Visualiser et éditer une base de données

Le jeu de données est saisi, on peut le constater avec l'option **Visualiser** et **éditer**.

The screenshot shows the R Commander interface. At the top, there is a menu bar with options: Fichier, Edition, Données, Statistiques, Graphes, Modèles, Distributions, Outils, Aide, and FactoMineR. Below the menu bar, there is a toolbar with buttons for 'Données' (set to 'nutriage'), 'Editer', 'Visualiser' (which is highlighted with a red arrow), and 'Modèle' (set to '<Pas de modèle>').

On the left, there is an 'Editeur de données' window titled 'nutriage'. It contains a table with columns: sexe, situation, the, cafe. The rows are numbered from 1 to 16. The first 16 rows of data are identical, showing values for sexe (2), situation (1), the (0), and cafe (0). The last row (17) shows different values: sexe (1), situation (1), the (0), and cafe (3).

On the right, there is a larger table titled 'nutriage' with columns: taille, poids, age, viande, poisson, fruit_crus, f. The rows are numbered from 1 to 27. The data starts with 16 rows of identical values (taille 151-162, poids 58-68, age 72-78, etc.) and ends with 11 rows of different values.

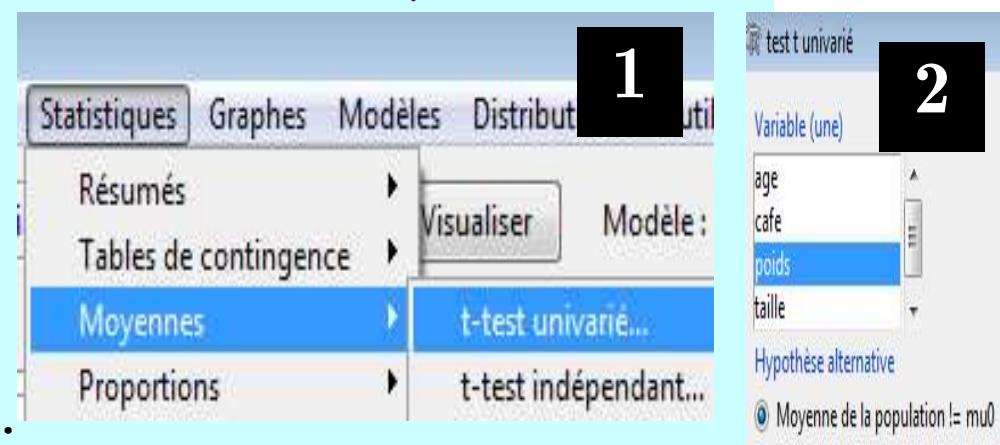
	taille	poids	age	viande	poisson	fruit_crus	f
1	151	58	72	4	3	1	1
2	162	60	68	5	2	5	5
3	162	75	78	3	1	5	5
4	154	45	91	0	4	4	4
5	154	50	65	5	3	5	5
6	159	66	82	4	2	5	5
7	160	66	74	3	3	5	5
8	163	66	73	4	2	5	5
9	154	60	89	4	3	5	5
10	160	77	87	2	3	5	5
11	175	68	91	5	2	5	5
12	165	75	81	5	2	2	2
13	158	53	89	4	2	5	5
14	155	63	79	3	1	5	5
15	154	80	83	3	3	5	5
16	166	80	78	5	0	5	5
17	159	57	74	3	2	5	5
18	157	55	74	3	2	5	5
19	165	57	78	5	2	5	5
20	156	90	78	5	1	5	5
21	175	90	73	5	1	5	5
22	161	68	76	4	2	5	5
23	168	83	85	4	2	5	5
24	168	90	73	3	2	4	4
25	156	56	77	3	2	5	5
26	170	70	74	3	2	5	5
27	162	58	67	4	3	5	5

I.C. pour une moyenne (Application)

Exemple d'application: A partir de l'étude alimentaire, on s'intéresse à l'estimation par intervalle de confiance de la moyenne du poids des personnes âgées vivant en France.

Instruction sous R: *nutriage.txt*

```
```
>t.test(nutriage$poids,conf.level=.95)$conf.int
[1] 64.90497 68.05963
attr("conf.level")
[1] 0.95
Nous obtenons l'intervalle
de confiance [65.16,67.80]
de niveau de confiance 0.95.
```



# I.C. pour une moyenne (Application)

## Cas des petits échantillons ( $n < 30$ )

**Instruction sous R:** Il est possible d'utiliser les fonctions `bootstrap()` disponibles dans le package [RVAideMemoire](#).

```
library(RVAideMemoire)
I.C pour la moyenne
samp <- sample(nutriage$poids, 10, replace=TRUE)
bootstrap(samp, function(x, i) mean(x[i]))
Bootstrap
data: samp
1000 replicates

95 percent confidence interval:
 22.4 39.2
sample estimates:
original value
 30.9
```

# Intervalle de confiance pour la variance

Cas des échantillons avec une hypothèse de normalité

## Construction de I. C:

- On recherche une fonction pivotale, c'est à dire une fonction de  $X_1, X_2, \dots, X_n$  et de  $\sigma^2$ , dont la loi de probabilité ne dépend ni de  $\mu$  ni de  $\sigma^2$ .
- Une telle fonction est donnée par le **théorème de Fisher** :  $\frac{nS_n^2}{\sigma^2}$  est de loi  $\chi_{n-1}^2$ .
- On a donc quels que soient les réels  $a$  et  $b$ ,  $0 < a < b$ :  
$$P(a \leq \frac{nS_n^2}{\sigma^2} \leq b) = P(\frac{nS_n^2}{b} \leq \sigma^2 \leq \frac{nS_n^2}{a}) = F_{\chi_{n-1}^2}(b) - F_{\chi_{n-1}^2}(a)$$

# Intervalle de confiance pour la variance

Cas des échantillons avec une hypothèse de normalité

## Construction de I. C:

- Il y a une infinité de façons de choisir  $a$  et  $b$  de sorte que cette probabilité soit égale à  $1-\alpha$ . On montre que les valeurs pour lesquelles  $b-a$  est minimum (on cherche à obtenir l'intervalle de confiance le plus étroit possible)

sont telles que:  $F_{\chi^2_{n-1}}(b)=1-\frac{\alpha}{2}$  et  $F_{\chi^2_{n-1}}(a) = \frac{\alpha}{2}$ .

- La table de  $\chi^2_{n-1}$  donne la valeur  $z_{n,\alpha}$  telle que  $Z$  est une v.a. de loi  $\chi^2_n$ , alors  $P(Z > z_{n,\alpha})=1 - F_{\chi^2_n}(z_{n,\alpha})=\alpha$ .

Alors,  $b=z_{n-1,\alpha/2}$  et  $a=z_{n-1,1-\alpha/2}$ , et  $P\left(\frac{nS_n^2}{b} \leq \sigma^2 \leq \frac{nS_n^2}{a}\right)=1-\alpha$ .

# Intervalle de confiance pour la variance

- ❖ Cas des échantillons avec une hypothèse de normalité

**Définition:** Un intervalle de confiance de niveau  $(1-\alpha)$  pour la variance  $\sigma^2$  est:

$$ic_{1-\alpha}(\sigma^2) = \left[ \frac{(n-1)S_{n-1}^2}{\chi_{n-1,1-\alpha/2}^2}, \frac{(n-1)S_{n-1}^2}{\chi_{n-1,\alpha/2}^2} \right]$$
$$S_{n-1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$$

# I.C. variance (Application)

Voyons comment le faire sous R. La fonction qui nous donne les quantiles du Ki-deux est `qchisq`. Prenons les données. *nutriage.txt*

```
alpha <- 0.05; n<-226

sn2 <- (n*var(nutriage$poids)) / (n-1)

born.inferior <- (n-1)*sn2/qchisq(1-alpha/2,df=n-1)

born.superior <- (n - 1)*sn2/qchisq(alpha/2,df = n - 1)

c(born.inferior, born.superior)
```

Nous obtenons l'intervalle de confiance [121.37, 175.78] de niveau de confiance 0.95.

# I.C. : erreur standard pour la moyenne

## ❖ Cas des échantillons sans hypothèse de normalité

Dans le cas où aucune hypothèse n'est faite sur les données, nous conseillons d'utiliser une approche par *bootstrap* comme pour la moyenne.

```
I.C. de l'erreur standard pour la moyenne
> bootstrap(samp,function(x,i) sd(x[i])/sqrt(length(x[i])))

 Bootstrap

data: samp
1000 replicates

95 percent confidence interval:
 3.121431 5.420537
sample estimates:
original value
 4.739433
```

# I. C. pour une proportion p

Si ( $n > 30$ ), ( $np \geq 5$ ) et  $n(1-p) \geq 5$ )

Si une population contient une proportion  $p$  d'individus possédant un caractère donné, l'estimateur de ce paramètre est la fréquence du caractère dans l'échantillon, noté  $\hat{p}$ .

**Définition:** Un intervalle de confiance de niveau( $1-\alpha$ ) pour la proportion inconnue  $p$  est :

$$ic_{1-\alpha}(p) = \left[ \hat{p} - z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

où  $z_{1-\alpha/2}$  représente le quantile de la loi normale centrée réduite. Pour  $\alpha=5\%$ ,  $z_{1-\alpha/2}=1,96$ .

# I.C. pour une proportion p

## Application

```
library(epitools)
table(nutriage$sexe) # Répartition de la
variable sexe.

sexes
1 2
85 141

binom.approx(141,226) [c("lower","upper")] #
Calcul de l'ic avec n=226.

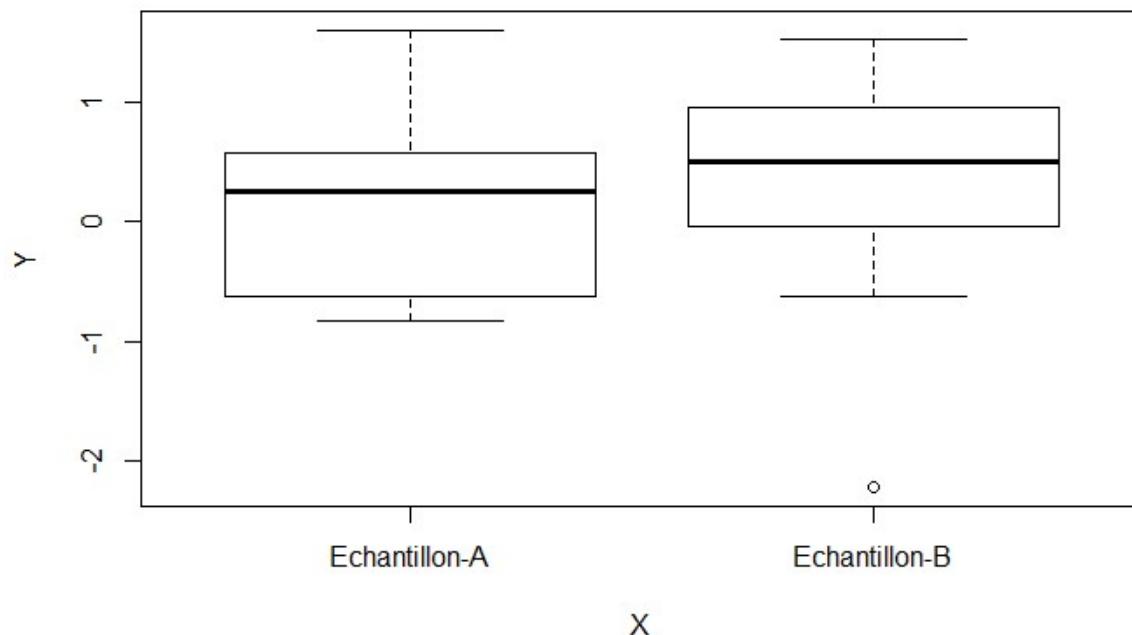
 lower upper
1 0.5607393 0.6870483
```

# TESTS statistiques, ...

# Problématique:

Exemple de la comparaison de 2 échantillons:

```
> set.seed(1)
> Echantillons=rep(c("Echantillon-A","Echantillon-B"),each=10)
> Y=rnorm(20)
> don=data.frame(Y,X=Echantillons)
> attach(don)
> boxplot(Y~X)
```



- ❖ On observe une différence entre les deux groupes.
- ❖ Est elle significative?
- ❖ Pour cela on effectue un test statistique.

# Introduction

## La théorie des tests :

- Il s'agit à partir de l'étude d'un ou plusieurs échantillons de prendre des décisions concernant l'ensemble de la population.
- On est alors amené à émettre des hypothèses concernant la population qui peuvent s'avérer *vraies* ou *fausses*.
- Ces hypothèses sont faites pour être soit confirmées soit rejetées à l'aide d'un test d'hypothèse.

# Introduction (suite)

Un test d'hypothèse est une méthode basée sur **la probabilité** pour prendre une **décision**, *sur la base de résultats d'échantillon*, concernant *la valeur* du paramètre  $\theta$  d'une population (par exemple, *la moyenne* ( $\mu$ ) ou *l'écart type* ( $\sigma$ ) dans le cas d'un problème d'un seul échantillon), ...

# Introduction (suite)

ou *les valeurs* relatives aux paramètres  $\theta_1$  ( $\mu_1$ ) et  $\theta_2$  ( $\mu_2$ ) pour deux populations (par exemple, *la différence entre les moyennes de population  $\mu_1 - \mu_2$  dans le cas d'un problème à deux échantillons*).

Sa construction se base sur les notions de l'échantillonnage et les estimations.

# Introduction

- L'une des fonctions des statistiques est de proposer, à partir d'observations d'un phénomène aléatoire, une estimation d'un des paramètres du phénomène.
- Les statistiques servent aussi à prendre **des décisions**, comme par exemple:
  - Peut on considérer qu'un médicament est plus efficace qu'un placebo ?
  - Le nombre de consultations de Google par seconde suit il une loi de Poisson?

# Tests d'hypothèses

## Notions générales:

Un test d'hypothèse (ou test statistique) est un processus composé de plusieurs étapes très concrètes, qui a pour but de fournir une règle de **décision** permettant, sur la base de résultats d'échantillon, d'évaluer (**accepter** ou **rejeter**) une information hypothétique (notée  $H_0$ ).

**Hypothèse nulle  $H_0$**  : C'est l'hypothèse principale, que l'on va supposer vraie pour faire le test. C'est toujours une hypothèse d'égalité.

## Exemple ( $H_0$ ):

Deux populations d'étudiants (de même niveau) ayant suivi des méthodes pédagogiques différentes ont les mêmes notes moyennes aux examens.

# Tests d'hypothèses

L'hypothèse alternative notée  $H_1$  est l'hypothèse contraire que l'on souhaite prouver en rejetant l'hypothèse nulle ( $H_0$ ). Elle traduit une différence ou un effet statistiquement significatifs.

## Exemple ( $H_1$ ):

Deux populations d'étudiants ayant suivi des méthodes pédagogiques différentes ont des notes moyennes significativement différentes aux examens.

# Tests d'hypothèses (suite)

L'idée de base des tests, est de trouver une statistique (*une fonction des observations*) dont on connaît la loi (ou qui s'approxime par une loi connue) si  $H_0$  est vraie, et qui ne se comporte pas de la même manière selon que  $H_0$  ou  $H_1$  est vraie.

# Exemple pratique

## Durée de vie de moteurs électriques:

Les moteurs des appareils électroménagers d'une marque M ont une durée de vie moyenne de  $\mu_r=3000$  heures avec un écart-type  $\sigma_r= 150$  heures. A la suite d'une modification dans la fabrication des moteurs, le fabricant affirme que les nouveaux moteurs ont une durée de vie moyenne supérieure à celle des anciens.

- On teste un échantillon de  $N = 50$  nouveaux moteurs. On note  $X_i$  les durées de vie observées et on calcule la durée de vie moyenne (empirique) des nouveaux moteurs :  $\bar{X}_{50} = (\sum_{i=1}^{50} X_i)/50 = \textcolor{blue}{3040,3}$  heures.

# Exemple pratique(suite)

Question :

Les nouveaux moteurs apportent-ils une amélioration statistiquement significative dans la durée de vie des appareils électroménagers ?

# Exemple pratique(suite)

## « réponse ! »

➤ **Référence** : durée de vie moyenne des anciens moteurs :  $\mu_r = 3000$  h,  $\sigma_r = 150$  h.

➤ **Données** : L'échantillon de nouveaux moteurs ( $N = 50$ ):  $\bar{X}_{50} = 3040,3$  h.

**Hypothèse  $H_0$**  : le nouveau procédé ne change pas la durée de vie moyenne

**Hypothèse  $H_1$**  : le nouveau procédé augmente la durée de vie moyenne

**Comment décider ?**  $\Rightarrow$  on utilise l'écart entre  $\bar{X}_{50}$  et la référence  $\mu_r$  :

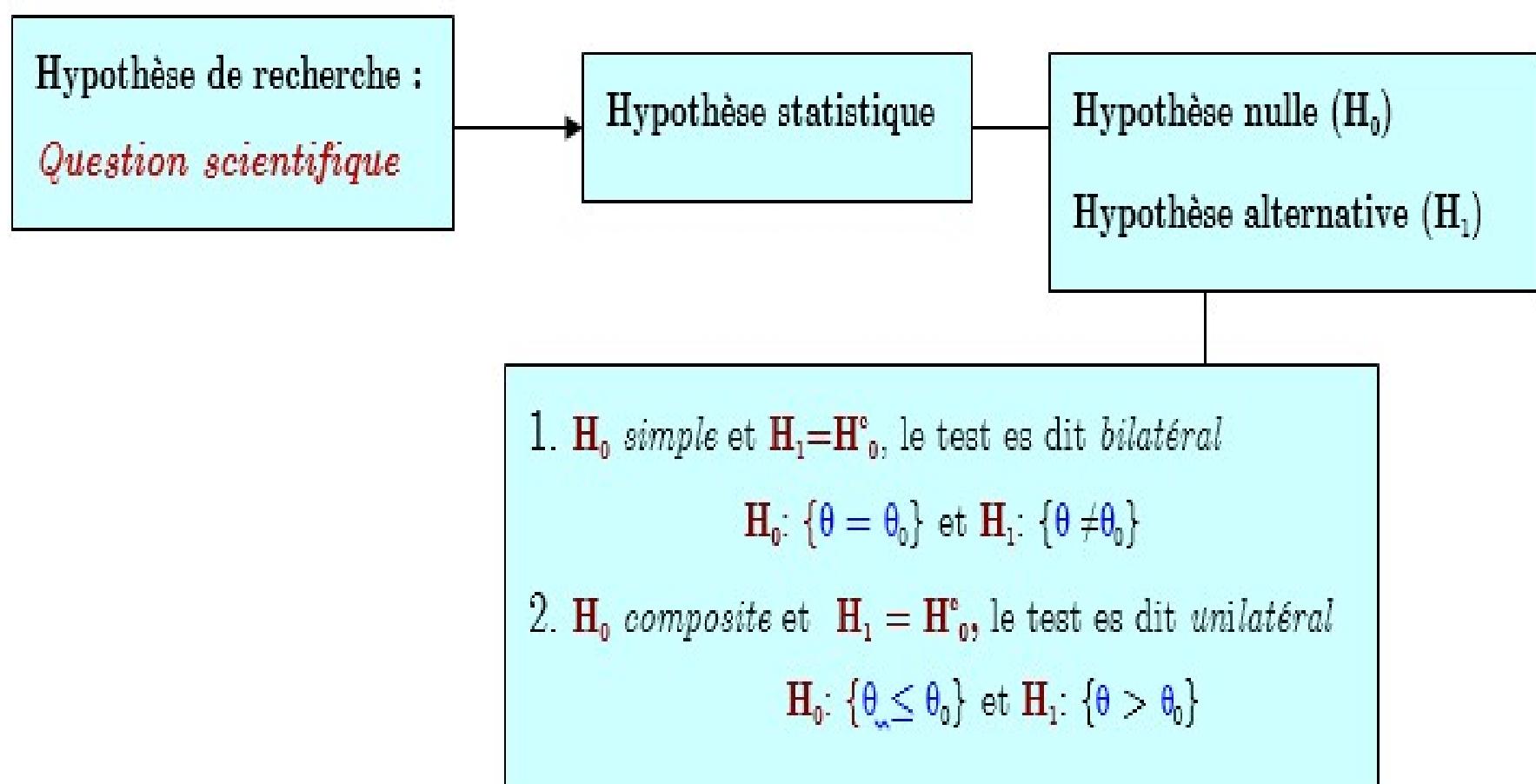
- Si l'écart  $(\bar{X}_{50} - \mu_r)$  est faible, il est possible d'accepter  $H_0$ .
- Si l'écart  $(\bar{X}_{50} - \mu_r) >> 0$  et grand, il est possible d'accepter  $H_1$ .

**Question** : l'écart observé  $\bar{X}_{50} - \mu_r = 40,3$  heures est-il «suffisamment grand» pour rejeter  $H_0$ ?

# Différentes étapes

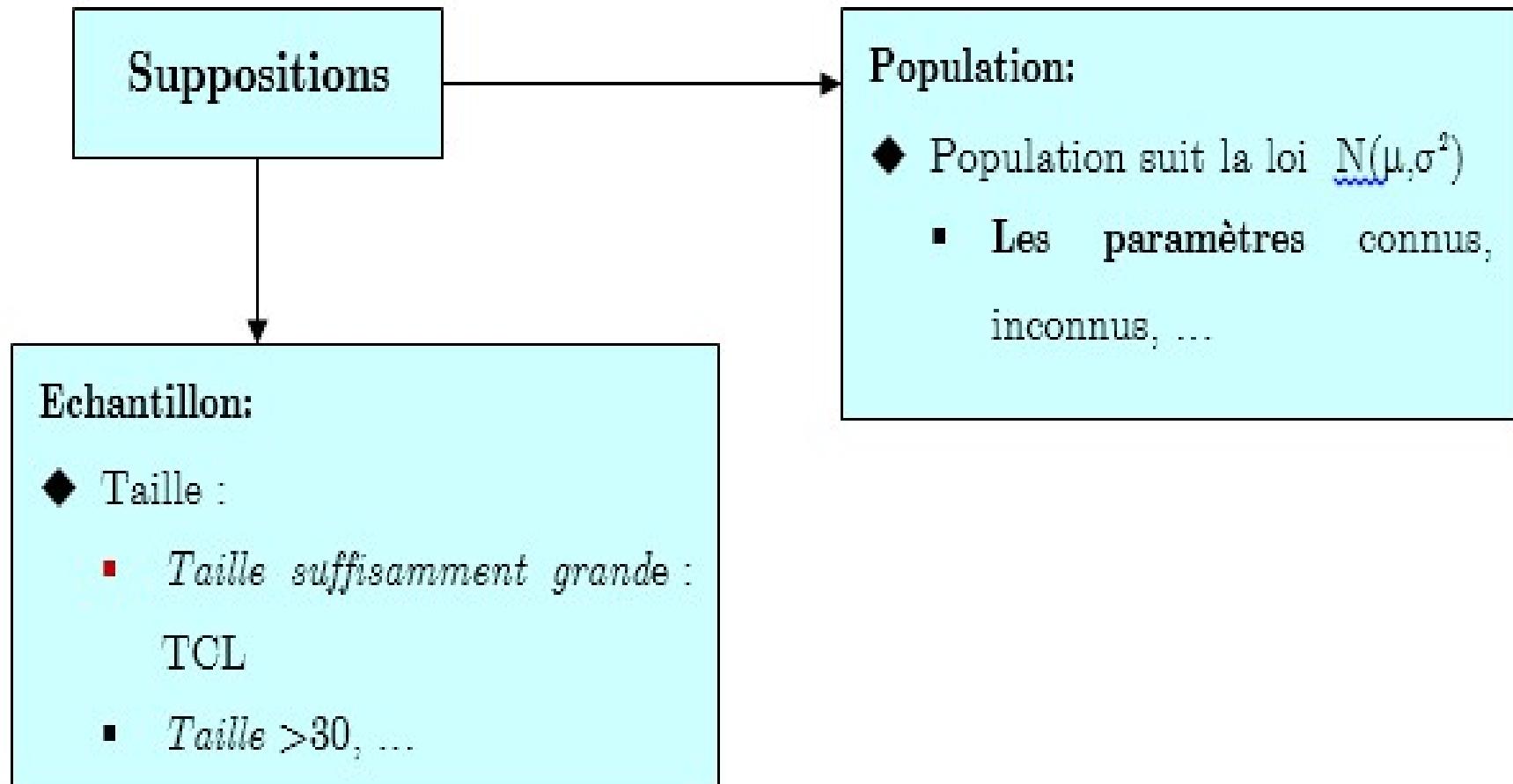
En voici les étapes à suivre pour la réalisation du test

## 1) Poser les hypothèses statistiques:



# Etapes en la réalisation du test (suite)

## 2) Spécifier les suppositions qu'on va assumer



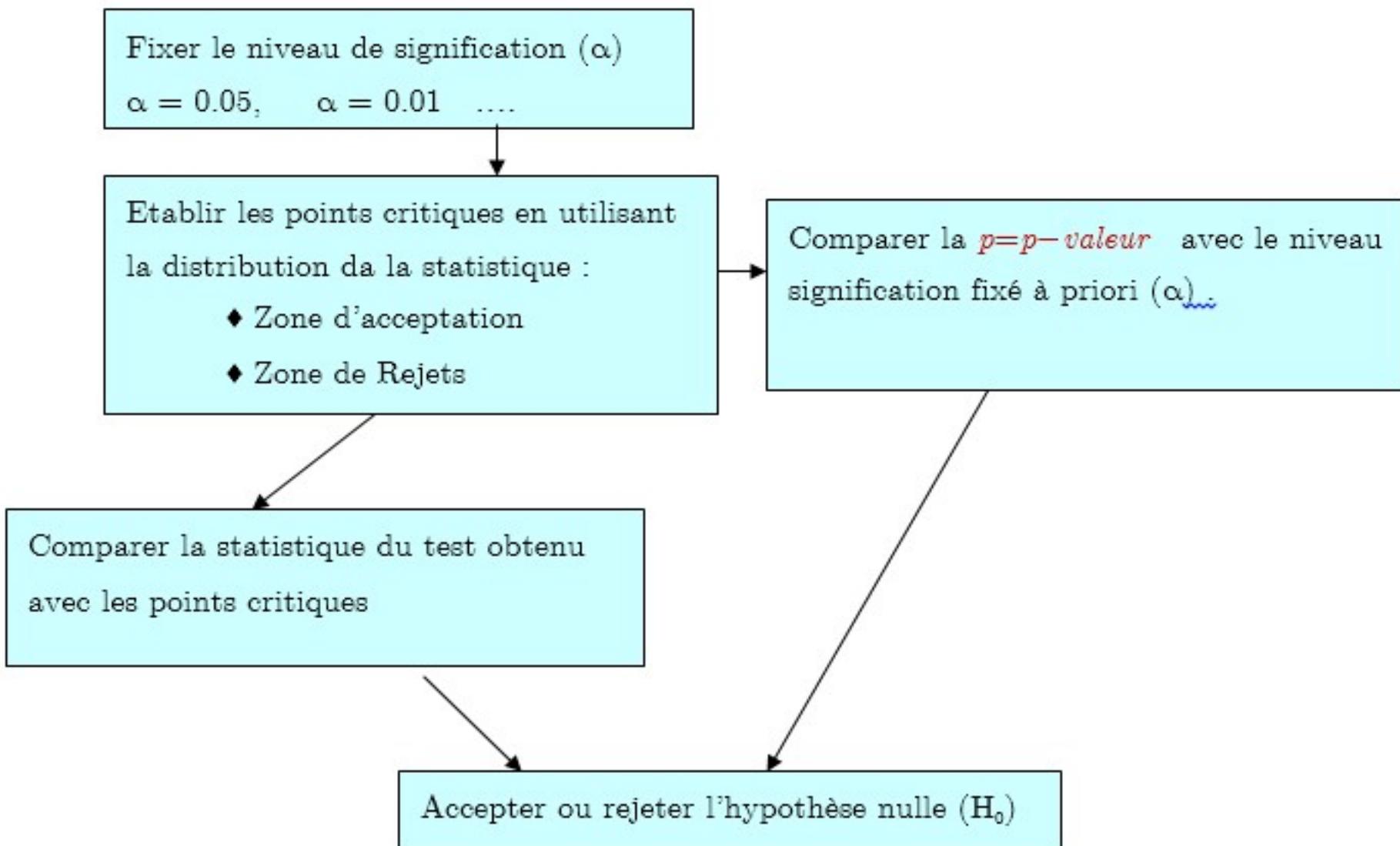
# Etapes en la réalisation du test (suite)

## 3) Calculer la Statistique de Test:



# Etapes en la réalisation du test (suite)

## 4) Prendre une décision:



# Région critique et règle de décision

Définition: Le niveau critique observé (ou  $P$ -value) est la valeur minimale de  $\alpha$  telle que  $H_0$  est toujours rejetée.

Avantage :

1. Une fois que la p-value est connue, le décideur peut déterminer la décision du rejet ou non-rejet en utilisant n'importe quel seuil  $\alpha$
2. Comparer la p-value à  $\alpha$  au lieu de comparer la valeur de la statistique de test à une valeur critique (table).

$P$ -value  $< \alpha \Rightarrow$  rejet de  $H_0$  (*Généralement on considère  $\alpha=0.05$* )

$P$ -value  $\geq \alpha \Rightarrow$  non rejet de  $H_0$

# Exemple pratique (suite)

## Durée de vie de moteurs électriques:

- Sous l'hypothèse  $H_0$ , la durée de vie d'un moteur testé est une v.a.  $X$  de moyenne  $\mu_r=3000$  et d'écart-type  $\sigma_r=150$  (NB. sa distribution n'est pas connue).
- D'après le TCL, sous  $H_0$ , la moyenne  $\bar{X}_N$  de l'échantillon (avec  $N=50>30$ ) :  $\bar{X}_N = (\sum_{i=1}^N X_i)/N \sim N(\mu_r, \sigma_r^2/N) \Rightarrow Z = \frac{(\bar{X}_N - \mu_r)}{\sigma_r/\sqrt{N}} \sim N(0,1)$ . Donc, si  $H_0$  est vraie, on calcul  $z_\alpha$  tel que  $p(|Z| > z_\alpha) = 0,05 = 5\%$  (faible).
- Et la valeur observé de  $Z$ :

$$z_{\text{obs}} = \frac{40,3}{150/\sqrt{50}} = 1,90 < z_{\alpha=0,05} = 1,96$$

- **Décision:** on accepte l'hypothèse  $H_0$  (avec « 5 chances sur 100 » de se tromper)

# Tests d'hypothèses (suite)

## Erreurs associées au contraste :

Dans un test d'hypothèse statistique, il y a deux manières de se tromper :

- **L'erreur de type I** : la possibilité de rejeter une hypothèse nulle,  $H_0$ , alors qu'elle est vraie.

$\alpha = P[\text{erreur de type I}] = P(\text{rejeter } H_0 | H_0 \text{ est vraie}) =$  Risque d'affirmer qu'il y a une différence significative alors qu'elle n'existe pas réellement.

- **L'erreur de type II** : La possibilité de ne pas rejeter une hypothèse nulle,  $H_0$ , alors qu'elle est fausse.

$\beta = P[\text{erreur de type II}] = P[\text{ne pas rejeter } H_0 | H_0 \text{ est fausse}] =$  Risque d'affirmer qu'il n'y a pas de différence significative alors qu'elle existe réellement.

# Tests d'hypothèses (suite)

**Exemple:** On juge une personne et on formule les hypothèses suivantes :  $H_0$  : personne innocente  
 $H_1$  : personne coupable

|                    |                               | Vérité                        |
|--------------------|-------------------------------|-------------------------------|
| $H_0$              | personne coupable             | personne innocente            |
| personne innocente | Erreur de type I ( $\alpha$ ) | Correcte                      |
| personne coupable  | Correcte                      | Erreur de type II ( $\beta$ ) |

- $P(\text{condamner un innocent}) = \alpha$ .
- $P(\text{libérer un coupable}) = \beta$ .
- L'erreur de première espèce a alors pour conséquence de *condamner un innocent*, alors que l'erreur de deuxième espèce conduit à *libérer un coupable*...
- **Remarque :** Ne pas condamner un innocent **est prioritaire** par rapport à ne pas libérer un coupable.

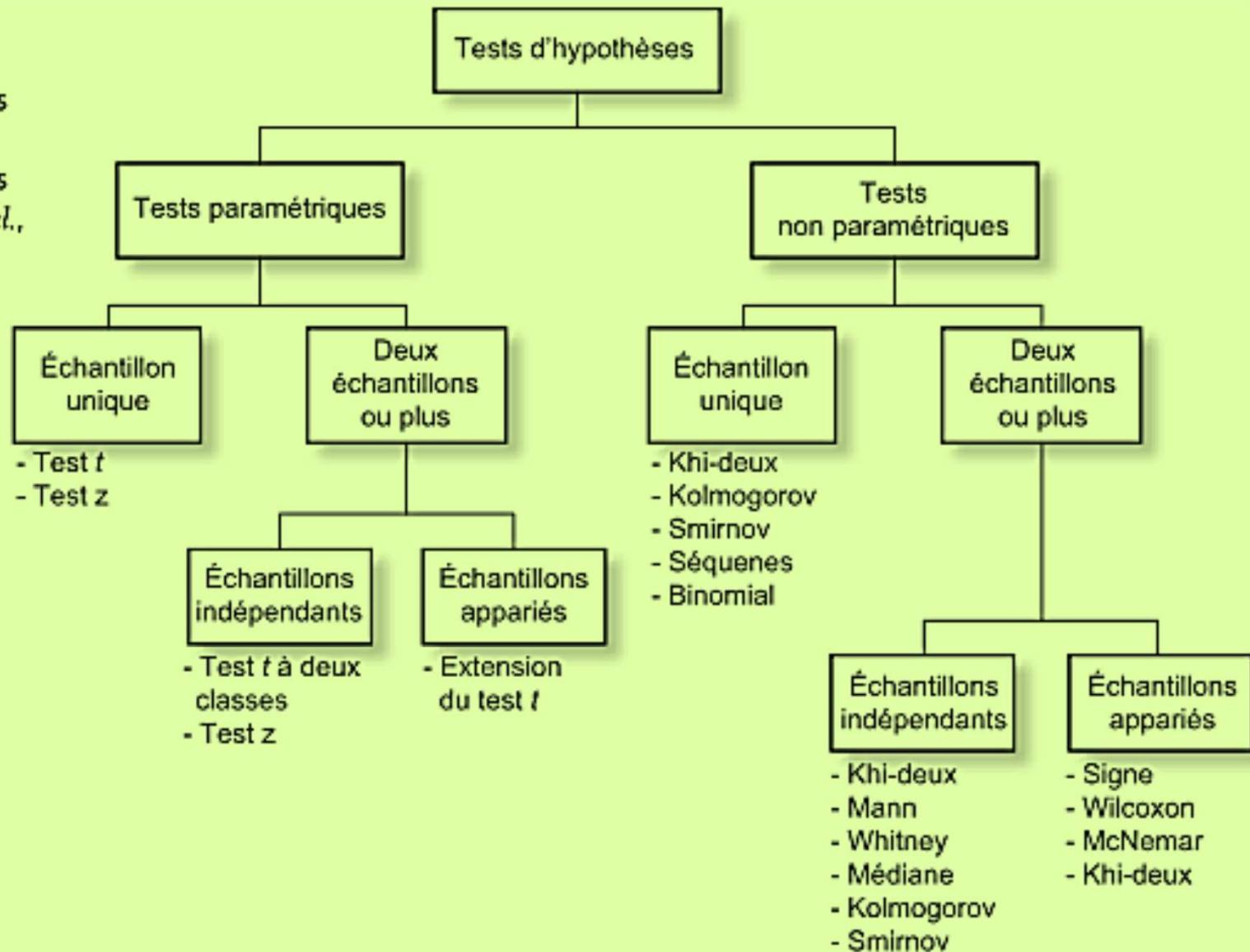
# Tests d'hypothèses (suite)

## Puissance du test :

- La puissance de test : La puissance du test est la probabilité de rejeter  $H_0$  si  $H_0$  est effectivement fausse, c'est-à-dire  $1-\beta$ .
- $1 - \beta = P[\text{rejeter } H_0 | H_0 \text{ est fausse}] =$  Probabilité de détecter une différence si elle existe réellement.
- Les conséquences de ces deux erreurs peuvent être d'importances diverses. En général, une des erreurs est plus grave que l'autre.
- Un bon test est un test qui, pour  $\alpha$  donné, maximise la puissance  $(1 - \beta) \geq 0,80$  .

# Tests paramétriques ou non ?

Tests paramétriques et tests non paramétriques (Malhotra *et al.*, 2007).



# Tests paramétriques de comparaison de deux moyennes

## **t-test pour deux échantillons indépendants**

**Objectif :** Etude de l'effet d'un facteur à deux modalités sur une variable dépendante

**Principe :** On s'intéresse à la différence entre les moyennes  $\mu_1$  et  $\mu_2$  au sein de deux populations au travers de deux échantillons indépendants.

**Données:** On doit disposer d'une variable **qualitative** qui prend **2 modalités** et on désire tester si la moyenne d'une autre variable **quantitative reste similaire selon ces 2 modalités**.

**Hypothèse testée:**

## Hypothèse testée:

Nous voulons savoir si il s'agit de la même population en ce qui concerne les moyennes, c'est-à-dire si  $\mu_1 = \mu_2$ .

On va donc tester:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \begin{matrix} > \\ \neq \\ < \end{matrix} \mu_2 \end{cases}$$

## Conditions de validité :

1. Indépendance des observations (*Cela signifie que la valeur d'une observation ne doit aucunement influencer la valeur d'une autre observation*).
2. La variable suit une loi normale dans chaque population ou  $n_1$  et  $n_2 > 30$ .
3. La variable a la même variance dans les deux populations.

## Statistique du Test:

*Variances inconnues, mais égales ( $\sigma_1 = \sigma_2$  : Homocédasticité)*

| Hypothèse                                                                 | Statistique de test                                                                                                                                    | On rejette $H_0$ si :                                                                                                                         |
|---------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$ | $t = \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{S}^2 \left( \frac{1}{n_x} + \frac{1}{n_y} \right)}} \sim t_{n_x + n_y - 2}$                                   | $ t_{\text{exp}}  > t_{n_x + n_y - 2, \alpha/2}$ ou<br>$p\text{-value} = 2 \times p(t_{n_x + n_y - 2, \alpha/2} >  t_{\text{exp}} ) < \alpha$ |
| $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$    | $\rightarrow$ Calcul de t sous $H_0$ :<br>$t_{\text{exp}} = \frac{\bar{x} - \bar{y}}{\hat{s} \sqrt{\left( \frac{1}{n_x} + \frac{1}{n_y} \right)}}$     | $t_{\text{exp}} > t_{n_x + n_y - 2, \alpha}$ ou<br>$p\text{-value} = p(t_{n_x + n_y - 2} > t_{\text{exp}}) < \alpha$                          |
| $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$    | $\rightarrow t_{n_x + n_y - 2, \alpha/2}$ : <u>lue dans la table de Student</u> pour un risque d'erreur $\alpha$ fixé et $(n_x + n_y - 2)$ d. liberté. | $-t_{\text{exp}} > t_{n_x + n_y - 2, \alpha}$ ou<br>$p\text{-value} = p(t_{n_x + n_y - 2, \alpha} > -t_{\text{exp}}) < \alpha$                |

## Variances inconnues, mais différentes ( $\sigma_1 \neq \sigma_2$ : Hétérocédasticité)

| Hypothèse                                                                 | Statistique de test                                                                                                                                                                                                       | On rejette $H_0$ si                                                                                                   |
|---------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$ | $t = \frac{\bar{X} - \bar{Y}}{\sqrt{\left( \frac{\hat{S}_x^2}{n_x} + \frac{\hat{S}_y^2}{n_y} \right)}} \sim t_{f, \alpha/2}$                                                                                              | $ t_{\text{exp}}  > t_{f, \alpha/2}$ ou<br>$p\text{-value} = 2 \times p(t_{f, \alpha/2} >  t_{\text{exp}} ) < \alpha$ |
| $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$    | <p>telle que :</p> $f = \frac{\left( \frac{\hat{S}_x^2}{n_x} + \frac{\hat{S}_y^2}{n_y} \right)}{\frac{1}{n_x+1} \left( \frac{\hat{S}_x^2}{n_x} \right)^2 + \frac{1}{n_y+1} \left( \frac{\hat{S}_y^2}{n_y} \right)^2} - 2$ | $t_{\text{exp}} > t_{f, \alpha}$ ou<br>$p\text{-value} = p(t_{f, \alpha} > t_{\text{exp}}) < \alpha$                  |
| $\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases}$    | <p>Si la valeur de <math>f</math> n'est pas un entier, on la redonde au entier antérieur.</p>                                                                                                                             | $-t_{\text{exp}} > t_{f, \alpha}$ ou<br>$p\text{-value} = p(t_{f, \alpha} > -t_{\text{exp}}) < \alpha$                |

$$\hat{S}_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

# Applications numériques

## (Sous R et /ou Rcommander)

# Comparaison de deux moyennes

## Exemple:

- ❖ On reprend les données du fichier "**nutriage.xls**".
- ❖ On souhaite étudier si **le poids** est significativement différent pour les **Hommes** et les **Femmes** dans la population étudiée.

**Validité:** Avant de comparer **ces deux moyennes**, on doit vérifier si:

- 1) **Les deux variances sont égales ou non.**
  - ⇒ **Test de Fisher** pour deux variances
- 2) **La distribution normale dans deux groupes.**
  - ⇒ **Hypothèse la moins importante pour la qualité du test**

# Validité: Comparaison de 2 variances

The screenshot shows the R Commander interface with the 'Statistiques' menu open, specifically the 'Variances' submenu which is highlighted. A callout arrow points from the text 'Sélectionner la variable qui code les deux groupes' to the 'Groupes (un)' field in the dialog, where 'sexe' is selected. Another callout arrow points from the text 'Sélectionner la variable d'intérêt' to the 'Variable réponse' field, where 'poids' is selected.

R Commander window:

- Fichier Edition Données Statistiques Graphes Modèles Distributions Outils Aide Facto
- Données : nutriage
- Fenêtre de script:

```
tapply(nutriage$poids, sexe)
var.test(poids ~ sexe, alternative='two.sided', conf.level=.95,
+ data=nutriage)
```
- Fenêtre de sortie:

```
F test to compare two variances

data: poids by sexe
F = 0.9636, num df = 84, denom df = 140, p-value = 0.8625
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.6620041 1.4289895
sample estimates:
ratio of variances
0.9635637
```

Test F de deux variances dialog:

- Groupes (un): sexe
- Variable réponse: poids
- Définition: <Pas de groupes sélectionnés>
- Hypothèse alternative:
  - Bilatéral (selected)
  - Définition < 0
  - Définition > 0
- Niveau de confiance : .95
- OK (button)
- Annuler (button)

Sélectionner la variable qui code les deux groupes

Sélectionner la variable d'intérêt

# 1. Validité: Lecture du résultat

## Explication des lignes de résultats

**F test to compare two variances :** C'est le nom anglais de la procédure, littéralement « F test à deux échantillons».

**data: poids by sexe:**

Rappel précisant qu'on a testé la variable poids, en construisant des groupes basés sur les deux modalités de la variable sexe.

**F =F<sub>obs</sub> = 0.9574, num df = 84, denom df = 139, p-value =0.8368**

**F:** C'est la valeur observée de la statistique  $F=S^2_1 /S^2_2= 0.9574$ .

**num df= (n<sub>1</sub>-1)=(85-1) ; denom df=(n<sub>2</sub>-1)=(140-1) .**

**P-value =0.83668.** On conclut que les variances ne diffèrent pas de façon significative au seuil de 5% ( P-value>  $\alpha=0.05$ ).

## 1. Validité: Lecture du résultat(suite)

95 percent confidence interval:

0.6573299 1.4203820

- C'est l'intervalle de confiance à 95% pour  $ratio = \sigma^2_1 / \sigma^2_2$ .

sample estimates:

ratio of variances

0.9573563

- L'estimation de  $ratio$ , montre que le rapport est presque 1.

## 2. Validité: Vérification de la normalité

### Étape 1.

#### Séparation des deux échantillons:

- Le poids des Hommes (échantillon 1)
- Le poids des Femmes (échantillon 2)

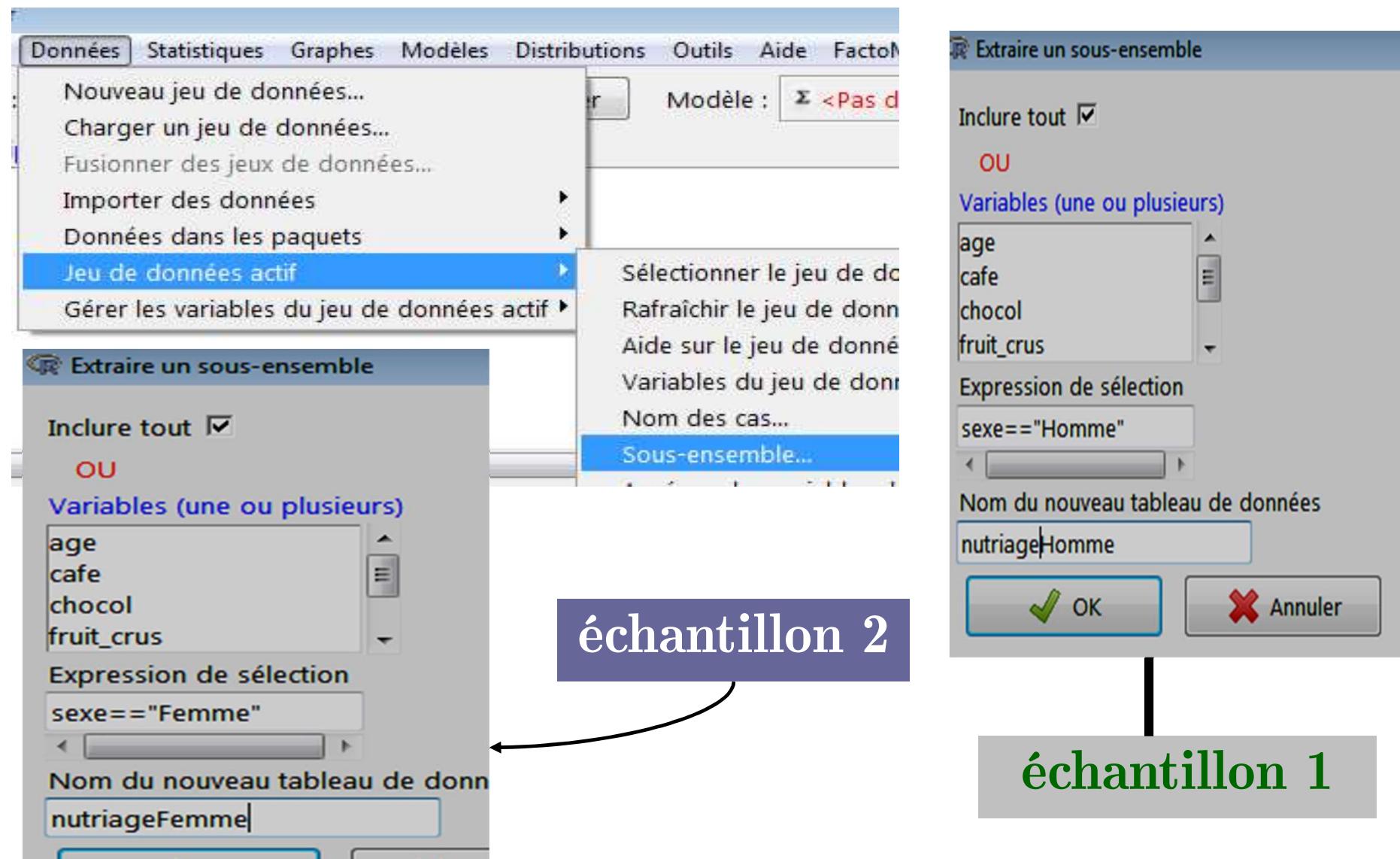
### Étape 2.

#### Tests de Shapiro et Wilk

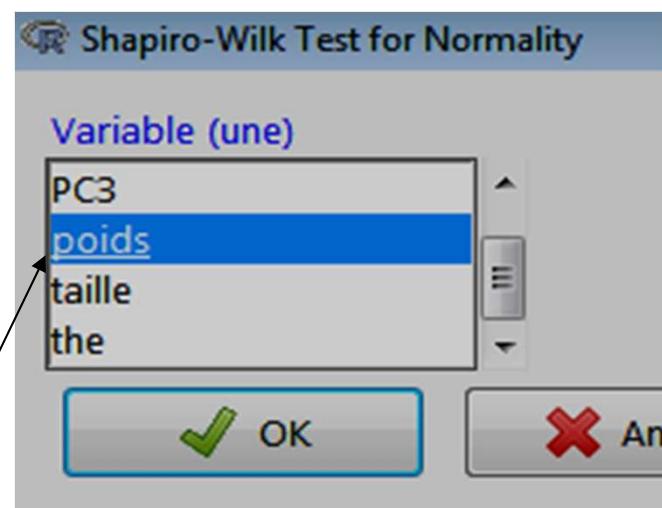
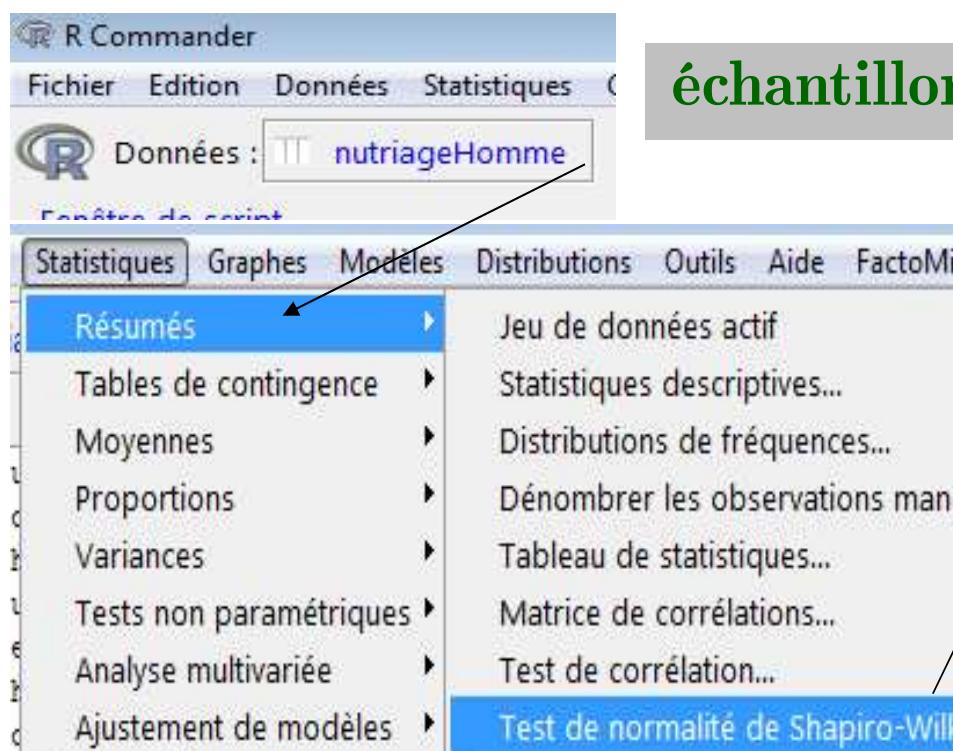
### Étape 3.

#### Histogrammes et/ou graphiques quantile-quantile

## 2. Vérification de la normalité (Étape 1)



## 2. Vérification de la normalité (Étape 2)



The screenshot shows the R console output for the Shapiro-Wilk test on the 'poids' variable of the 'nutriageHomme' dataset. The command entered was `> shapiro.test(nutriageHomme$poids)`. The output shows the test results: `Shapiro-Wilk normality test`, `data: nutriageHomme$poids`, and `W = 0.9743, p-value = 0.087`. A tooltip labeled 'échantillon 2' is positioned above the console window.

```
> shapiro.test(nutriageHomme$poids)

Shapiro-Wilk normality test

data: nutriageHomme$poids
W = 0.9743, p-value = 0.087
```

The screenshot shows the R console output for the Shapiro-Wilk test on the 'poids' variable of the 'nutriageFemme' dataset. The command entered was `> shapiro.test(nutriageFemme$poids)`. The output shows the test results: `Shapiro-Wilk normality test`, `data: nutriageFemme$poids`, and `W = 0.9905, p-value = 0.4618`.

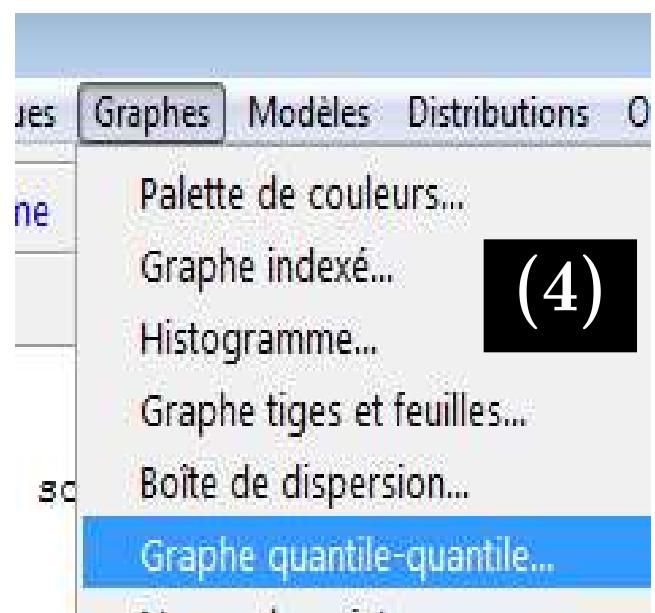
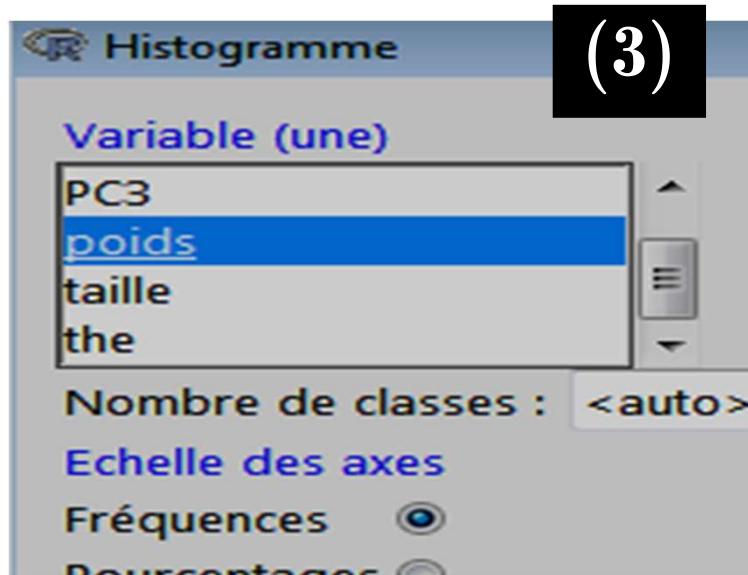
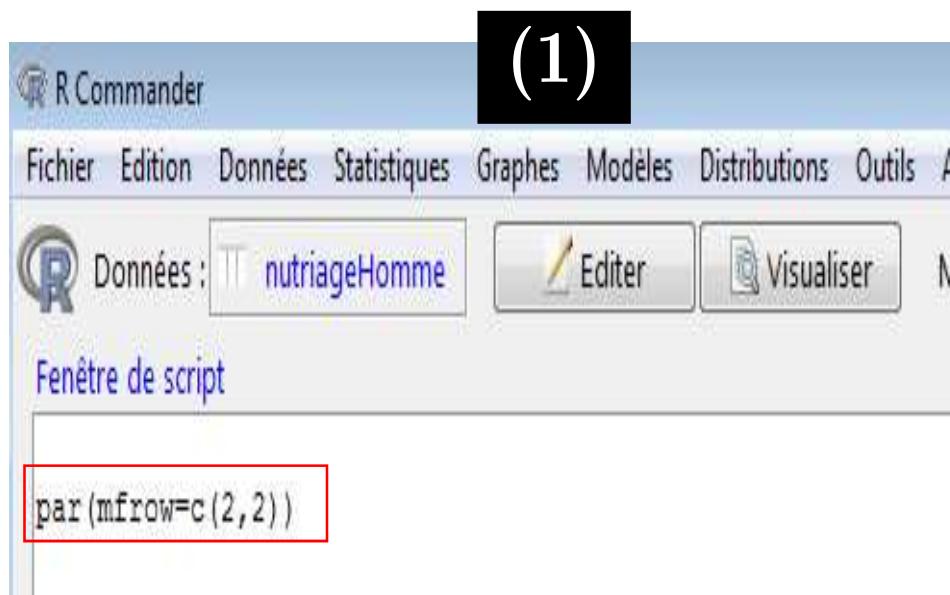
```
> shapiro.test(nutriageFemme$poids)

Shapiro-Wilk normality test

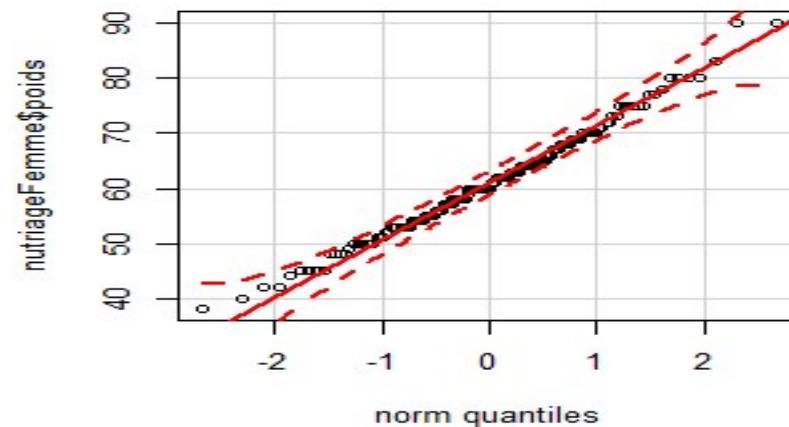
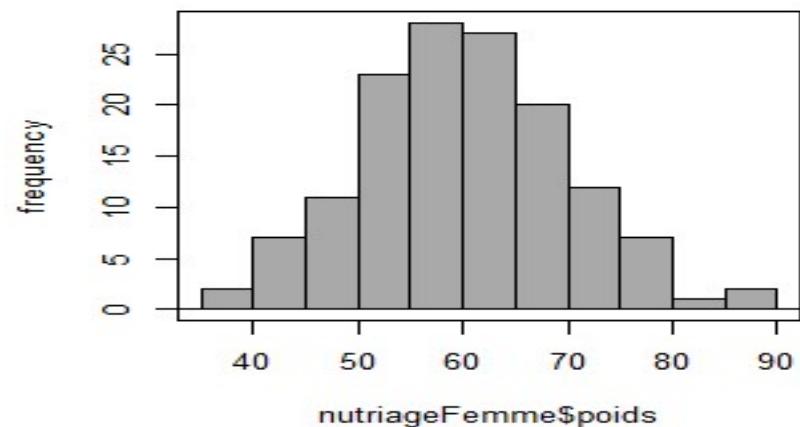
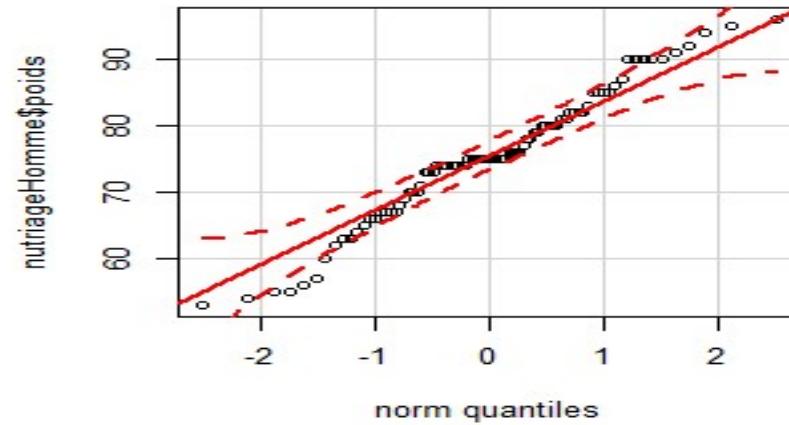
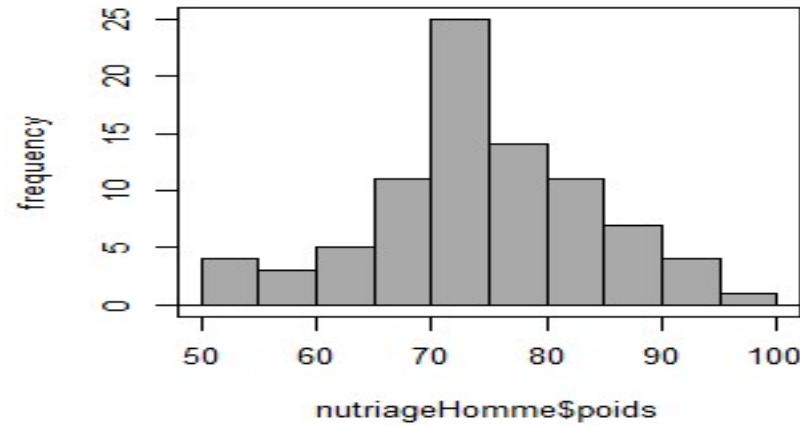
data: nutriageFemme$poids
W = 0.9905, p-value = 0.4618
```

On accepte  $H_0$ :  
échantillon normale

## 2. Vérification de la normalité (Étape 3)

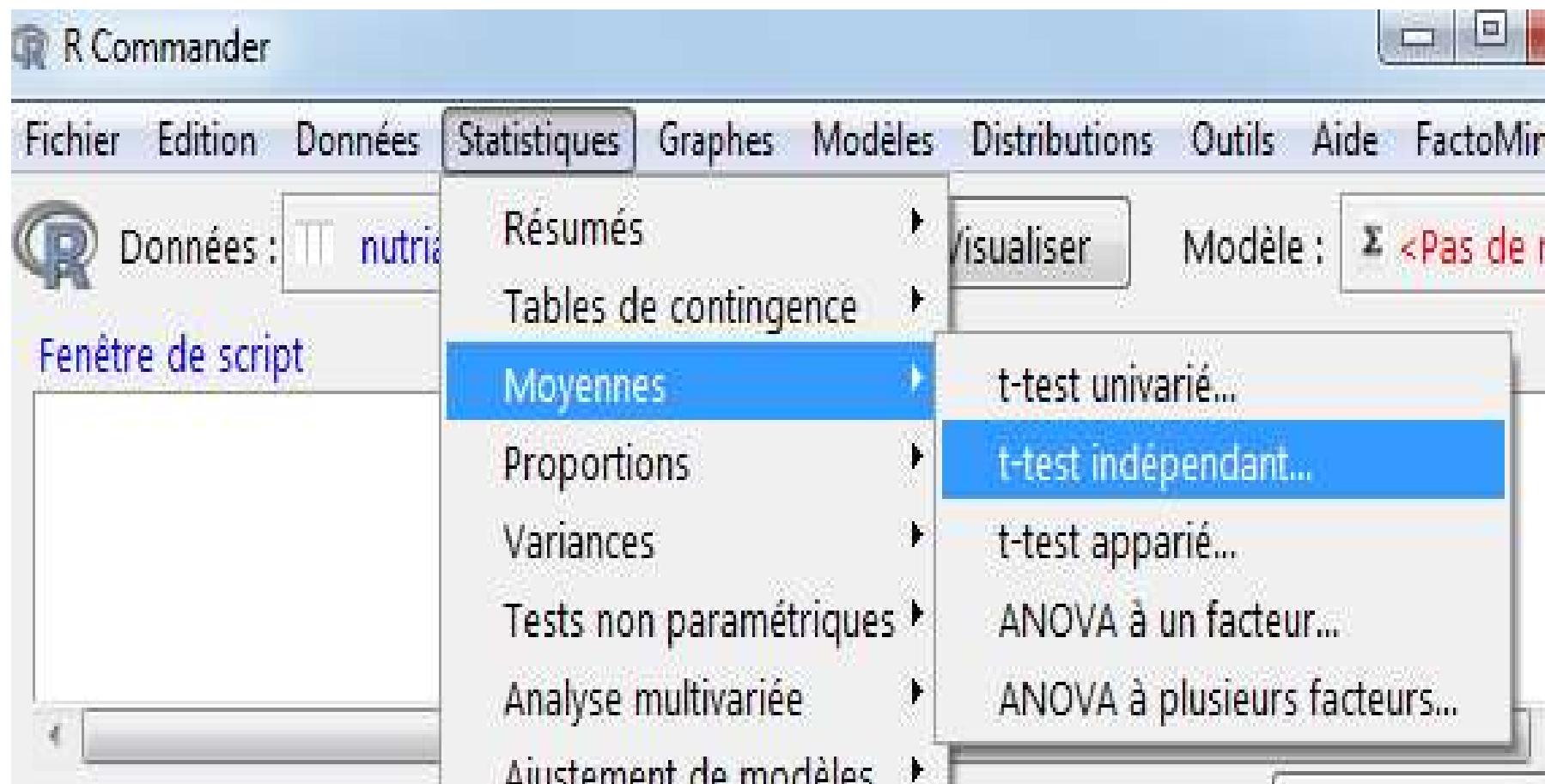


## 2. Vérification de la normalité (Étape 3)



# Exécution de t-test

On lance t-test à partir du package **Rcmdr**, menu:  
"Statistiques/Moyennes/ t-test indépendant"



# t-test indépendant: Résultat

**test t indépendant**

**Groupes (un)**  
sexe

**Variable réponse (une)**  
age  
cafe  
**poids**  
taille

**Différence :** <Pas de groupes sélectionnés>

**Hypothèse alternative**  
Bilatéral   
Différence < 0   
Différence > 0

**Niveau de confiance** .95

**Variances égales ?**  
Oui   
Non

**OK** **Annuler** **Réinitialiser**

**Fenêtre de sortie**

```
> t.test(poids~sexe, alternative='two.sided', conf.level=.95,
+ var.equal=TRUE, data=nutriage)

Two Sample t-test

data: poids by sexe
t = 10.5644, df = 224, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
11.62737 16.95986
sample estimates:
mean in group Homme mean in group Femme
75.40000 61.10638
```

**Soumettre**

# t-test: Lecture des résultats

## Explication des lignes de résultats

**Two Sample t-test:** C'est le nom anglais de la procédure, littéralement "test t à deux échantillons".

**data: poids by sexe:**

Rappel précisant qu'on a testé la variable poids, en construisant des groupes basés sur les deux modalités de la variable sexe.

**t = 10.5644, df = 224, p-value < 2.2e-16=(2.2\*10<sup>-16</sup>)**

**t:** C'est tout simplement la valeur du t de student, **10.5644** arrondie à la quatrième décimale.

$$t = T_{\text{obs}} = \frac{\bar{x}_H - \bar{x}_F}{S \sqrt{\frac{1}{n_H} + \frac{1}{n_F}}} = 10.5644$$

# t-test: Lecture des résultats

## Explication des lignes de résultats

**df:** abréviation anglaise de "degrees of freedom". C'est le nombre de degrés de libertés de la comparaison, soit, pour une comparaison à échantillons indépendants, le nombre de sujets moins 2, soit  $(226-2)=224$ .

**p-value =**  $= 2.2 \cdot 10^{-16}$  ( $<< 0.05$ ).

Donc  $H_0 : \mu_H = \mu_F$  est rejetée au seuil  $\alpha = 0.05$ . On peut conclure que le poids **discrimine les deux sexes**.

**data: poids by sexe:**

Rappel précisant qu'on a testé la variable poids, en construisant des groupes basés sur les deux modalités de la variable sexe.

# t-test: Lecture des résultats (suite)

## Explication des lignes de résultats

**alternative hypothesis: true difference in means is not equal to 0**

- L'hypothèse nulle ( $H_0$ ) d'une comparaison de moyennes représente l'égalité. Ici on compare les moyennes de la variable **Poids** obtenues pour les hommes et les femmes.
- L'hypothèse alternative est que  $H_0$  est fausse et donc que la différence "réelle" (*true difference en anglais*) dans la population d'où est tiré l'échantillon **n'est pas égale à 0**.

# t-test: Lecture des résultats (suite)

## Explication des lignes de résultats

5 percent confidence interval:

11.62737 16.95986

- Il s'agit de l'intervalle de confiance à 95% de la différence des moyennes, les bornes inférieure et supérieure de l'intervalle calculé de manière à ce que la moyenne "réelle" de la population ait 95% de chance d'être contenue de dans.

# t-test: Lecture des résultats (suite)

## Explication des lignes de résultats

**sample estimates:**

| <b>mean in group Homme</b> | <b>mean in group Femme</b> |
|----------------------------|----------------------------|
| <b>75.40000</b>            | <b>61.10638</b>            |

- Il s'agit des estimations des moyennes réalisées à partir de l'échantillon arrondies à la cinquième décimale, pour les hommes et pour les femmes.

## **Test Z pour l'égalité de deux proportions (distribution binomiale)**

Ce test paramétrique est utilisé pour étudier la supposition selon laquelle les proportions  $p_1$  (*inconnue d'individus présentant un certain caractère dans une population  $\mathcal{P}_1$* ) et  $p_2$  (*population  $\mathcal{P}_2$* ) des éléments de deux populations sont égales, sur la base de deux échantillons, l'un de chaque population.

### **Formulation des deux hypothèses:**

Les deux échantillons proviennent de deux populations de proportions  $p_1$  et  $p_2$ . Nous voulons savoir si  $p_1 = p_2$ . Pour cela, on désire tester l'hypothèse  $H_0$  contre l'hypothèse  $H_1$ :

$$\begin{cases} H_0 : p_1 = p_2 \\ H_1 : p_1 \begin{matrix} > \\ \neq \\ < \end{matrix} p_2 \end{cases}$$

♦ Conditions d'applications :

- 1) Le test est approximatif et suppose que le nombre d'observations dans les deux échantillons est suffisamment grand (c'est-à-dire  $n_1, n_2 \geq 30$ ) pour justifier l'approximation de la loi binomiale à la loi normale.
- 2) Les deux échantillons doivent vérifiées si  $n_1 \hat{p} \geq 5, n_1(1 - \hat{p}) \geq 5, n_2 \hat{p} \geq 5$  et  $n_2(1 - \hat{p}) \geq 5$  (

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}.$$

✓ Statistique de test sous  $H_0$ :

On détermine la variable aléatoire qui convient pour ce test. Ici, l'estimateur usuel de la proportion  $p$ , noté, par exemple,  $\hat{P}$ . La statistique de test

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

avec  $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ .

Peut être comparé avec le quantile de la distribution normale,  $N(0,1)$ , en utilisant un test uni ou bilatérale.

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

## Objectif:

On veut tester à partir d'un tableau de contingence s'il y a une relation entre deux caractères X et Y.

## Principe du test :

On calcule alors la statistique  $\chi^2$  :

$$\chi^2 = \sum_{ij} \frac{(O_{ij} - T_{ij})^2}{T_{ij}} \stackrel{H_0}{\sim} \chi^2(l-1)(c-1)$$

avec,

$O_{ij}$  = effectifs observes;

$T_{ij}$  = effectifs théoriques sous l'hypothèse  $H_0$ ;

$l$  = nombre de lignes;  $c$  = nombre de colonnes;

$(l-1)(c-1) = \text{ddl}$

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

Effectifs théoriques ( $T_{ij}$ ):

$$T_{ij} = \frac{O_{i+} \times O_{j+}}{N} \quad \text{avec}$$

$$O_{i+} = \sum_{j=1}^c O_{ij} = \text{total a la marge en ligne}$$

$$O_{+j} = \sum_{i=1}^l O_{ij} = \text{total a la marge en colonne}$$

$N$  = effectif total.

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

## En pratique:

- 2) Le test de khi-deux est une méthode pour comparer le tableau  $T_{\text{theo}}$  et le tableau  $T_{\text{obs}}$ . Faire l'opération termes à termes  $T_{\text{obs}} - T_{\text{theo}} = R$  (**Tableau des écarts à l'indépendance**).
  - 3) Élever chaque terme au carré pour obtenir le tableau  $R^2$ .
  - 4) Enfin on divise termes à termes le tableau  $R^2$  par le tableau des effectifs théoriques  $T_{\text{theo}}$ . On aura ensuite
- $$\chi^2 = \sum \frac{(T_{\text{obs}} - T_{\text{theo}})^2}{T_{\text{theo}}}$$

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

Exemple :

- Le tableau suivant résume la *présence* ou l'*absence* d'une **Infection** en fonction de l'utilisation ou non d'une *antibiothérapie* ou d'un *placebo* du ce **Traitement**.

| Effectifs observés |          | Y= Traitement |         |       |
|--------------------|----------|---------------|---------|-------|
| X= Infection       | Absence  | Antibio       | Placebo | Total |
|                    |          | 75            | 27      | 102   |
|                    | Présence | 10            | 29      | 39    |
| Total              |          | 85            | 56      | 141   |

- Y- a t' il une relations entre ces deux variables ?

⇒ Test d'indépendance ( $\chi^2$ )

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

Tester l'indépendance entre deux variables revient à mesurer l'écart entre ce qu'on observe et ce que l'on s'attend à observer dans une situation théorique d'indépendance.

| Effectifs observés |          | Y= Traitement |         |       |
|--------------------|----------|---------------|---------|-------|
| Infection          | Absence  | Antibio       | Placebo | Total |
|                    |          | 75            | 27      | 102   |
|                    | Présence | 10            | 29      | 39    |
| Total              |          | 85            | 56      | 141   |

| Effectifs attendus: $(n_i \times n_j) / N$ |          | Traitement |         |       |
|--------------------------------------------|----------|------------|---------|-------|
| X= Infection                               | Absence  | Antibio    | Placebo | Total |
|                                            |          | 61.49      | 40.51   | 102   |
|                                            | Présence | 23.51      | 15.49   | 39    |
| Total                                      |          | 85         | 56      | 141   |

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

- **Test :** On teste l'hypothèse  $H_0$  : "la variable **Infection** est indépendante de la variable **Traitement**" contre  $H_1$ : "la dépendance entre les deux variables".

## ↳ Utilisation de test du $\chi^2$

- ❖ si  $\chi^2_{\text{obs}} > \chi^2_{1-\alpha} (l-1)(c-1)$ , on **rejette**  $H_0$ 
  - ↳  $\alpha$  est le seuil pour la décision ( souvent fixé à 5%).
  - ↳  $l = \text{nombre de lignes}; c = \text{nombre de colonnes}$
- ❖ Le logiciel fournit en réponse une **p-value** :
  - ↳ **p-value** = Niveau de signification = **Probabilité de se tromper en rejetant l'hypothèse  $H_0$** .
  - ↳ **p-value**  $< \alpha (=5\%) \Rightarrow \text{rejeter } H_0$
  - ↳ **p-value**  $\geq \alpha (=5\%) \Rightarrow \text{ne pas rejeter } H_0$

# LE TEST DU $\chi^2$ D'INDÉPENDANCE

## □ Conditions d'application:

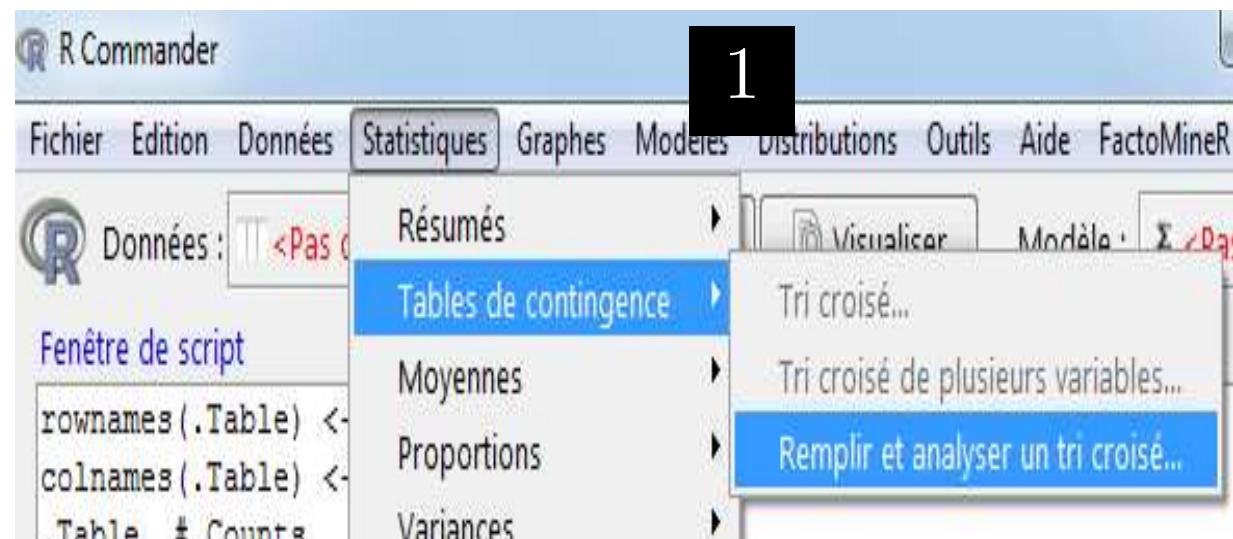
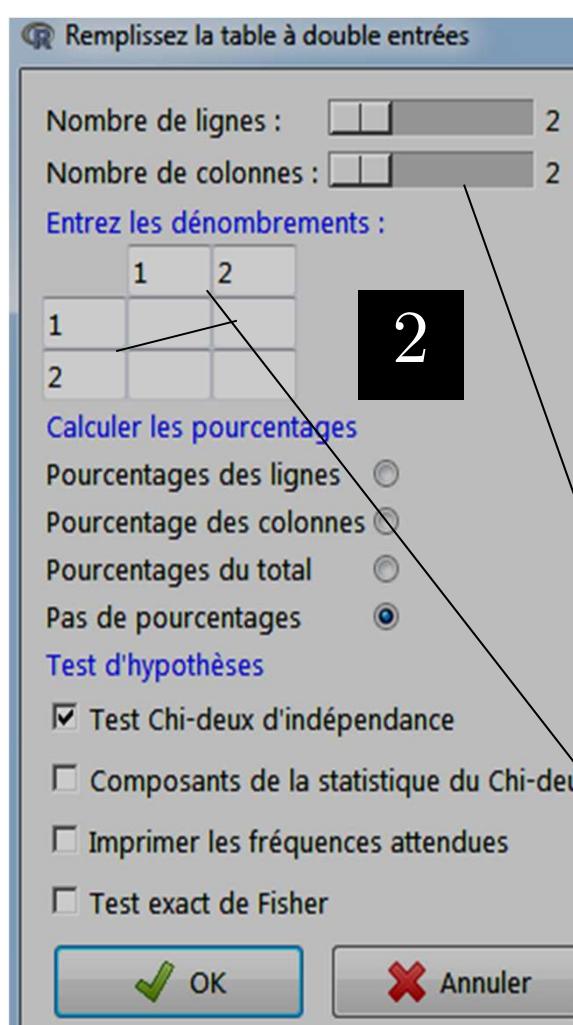
L'effectif théorique calculé sous l'hypothèse  $H_0$  doit être supérieur à 5.

## □ Remarques :

1. Pour des effectifs faibles (un au moins eff.  $\leq 5$ ), il existe des corrections a ce test de Yates ou d'autres tests (test exact de Fisher).
2. Si l'hypothèse d'indépendance est rejetée, il est intéressant d'observer la contribution de chaque modalité à ce rejet → Analyse factorielle des correspondances simples.

# $\chi^2$ avec RCommander

## Création des tableaux croisés à la main



On peut changer le nombre de lignes et de colonnes en glissant les buttons.

On peut mettre des noms de variables dans les case de 1, 2 et des chiffres dans les cases vides.

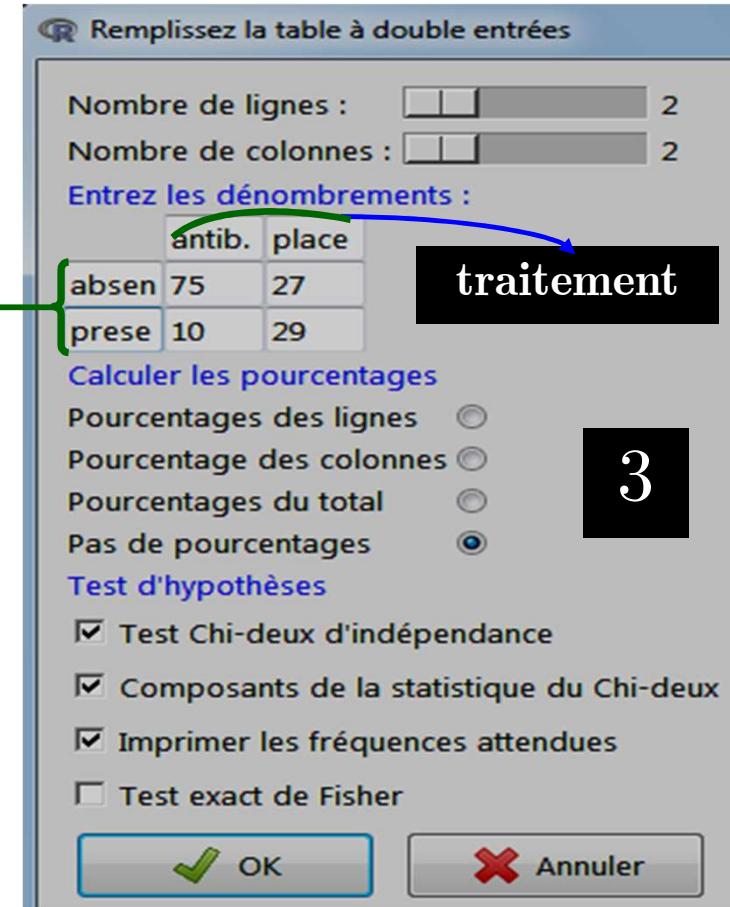
# Exécution du $\chi^2$

## Lire les résultats:

- Affichage de notre tableau d'origine

```
> .Table # Counts
 antib placebo
absence 75 27
presence 10 29
```

Infection



- Examinons le résultat du test :

```
> .Test <- chisq.test(.Table, correct=FALSE)
> .Test
Pearson's Chi-squared test
data: .Table
X-squared = 27.0232, df = 1, p-value = 2.01e-07
```

- R indique d'abord qu'il applique le test de khi deux de Karl Pearson qui est son "inventeur" : Pearson's Chi-squared test.
- Ensuite le test est appliqué sur l'objet ".Test" : data: .Table

## Exécution du $\chi^2$ ( fenêtre de sortie)

- ◆ La valeur de l'indicateur de  $\chi^2$  est  
 $27.02 = \text{xsquared}$ , doit être positionnée par rapport à la loi de Khi\_deux à degré de liberté ( $df = 1$ ).  
↳  $p\text{-value} = 2.01 \times 10^{-7}$
- ◆ Dans ce cas: **p-value** est inférieure à 5%  
( $2.0 \times 10^{-7} << 0.05$ ), donc on **rejette** largement l'hypothèse d'indépendance.
- ◆ Affirmation avec seulement 5 chances sur 100 de se tromper.

# Exécution du $\chi^2$ ( fenêtre de sortie) (3)

## □ Analysons les effectifs des modalités:

### ❖ Tableau des résidus

```
> round(.Test$residuals,2)
 antib. placebo
absence 1.72 -2.12
presence -2.79 3.43
```

|          | antib. | placebo |
|----------|--------|---------|
| absence  | +      | -       |
| présence | -      | +       |

## Interpretation:

- ↳ On voit grâce au tableau des résidus que c'est la "présence" qui est le principale responsable du calcul de l'indicateur de  $\chi^2$ .
- ↳ En effet, en terme de présence, les infections au placebo sont significativement plus nombreux que le antibiothérapie. Inversement l'absence du maladie est significativement plus fréquent que placebo.

# $\chi^2$ avec RCommander

## Exercice:

|  |  | matgras |        |           |          |           |       |       |       |        |
|--|--|---------|--------|-----------|----------|-----------|-------|-------|-------|--------|
|  |  | sexes   | beurre | margarine | arachide | tournesol | olive | Isio4 | colza | canard |
|  |  | Homme   | 10     | 10        | 16       | 21        | 20    | 5     | 0     | 3      |
|  |  | Femme   | 5      | 17        | 32       | 47        | 20    | 18    | 1     | 1      |

# Le $\chi^2$ à partir d'un tableau de données

**Exemple 2:** On reprend les données du fichier "nutriage.xls".

## Deux variables qualitatives:

- sexe: 2=Femme; 1=Homme
- matgras (Matière grasse préférentiellement utilisée pour la cuisson): ↗

"beurre" =Beurre,

"margarine" =Margarine,

"arachide" =Huile d'arachide,

"tournesol"=Huile de tournesol

"olive" =Huile d'olive,

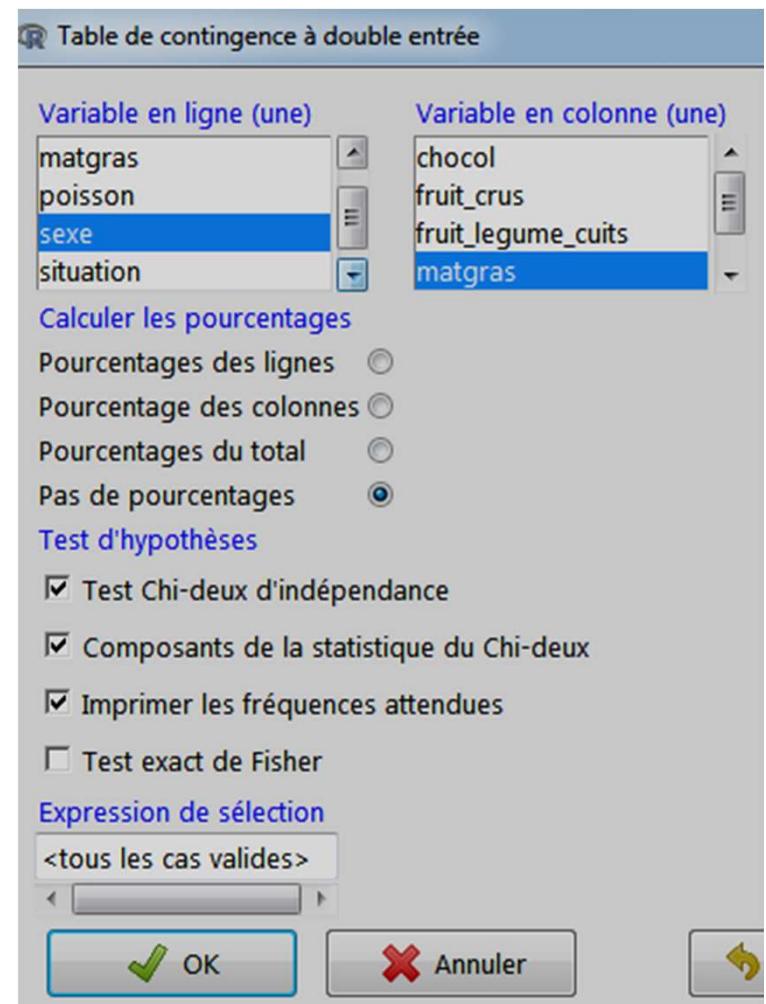
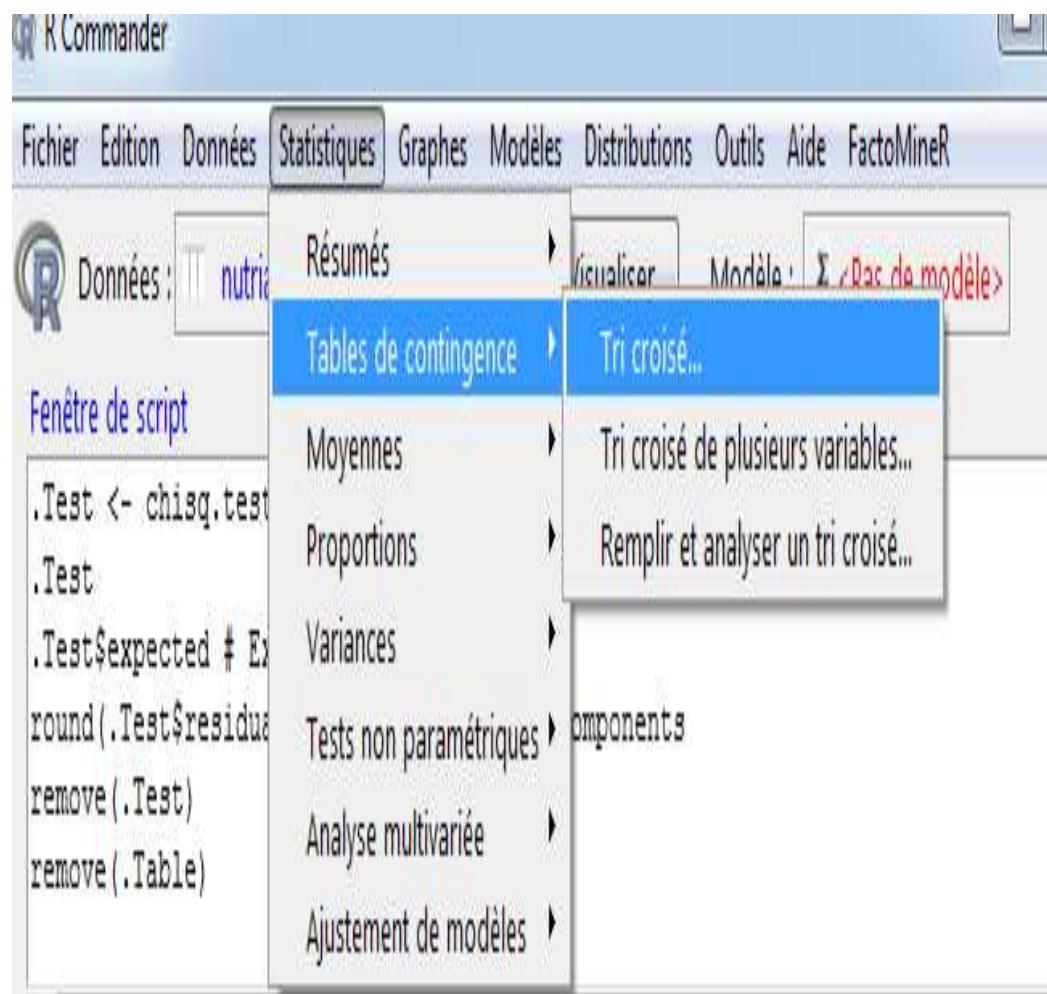
"Isio4"=Mélange d'huile (type Isio4),

"colza"=Huile de colza,

"canard"=Graisse de canard ou d'oie

# Le $\chi^2$ à partir d'un tableau de données

Chargez R, puis R Commander et importez le classeur nutriage.xls comme jeu de données.



# Exécution du test de khi deux

Fenêtre de sortie Soumettre

```
> .Table <- xtabs(~sexe+matgras, data=nutriage)

> .Table
 matgras
sexe beurre margarine arachide tournesol olive Isio4 colza canard
Homme 10 10 16 21 20 5 0 3
Femme 5 17 32 47 20 18 1 1

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

 Pearson's Chi-squared test
data: .Table
X-squared = 15.1584, df = 7, p-value = 0.03402

> .Test$expected # Expected Counts
 matgras
sexe beurre margarine arachide tournesol olive Isio4
Homme 5.641593 10.15487 18.0531 25.57522 15.04425 8.650442
Femme 9.358407 16.84513 29.9469 42.42478 24.95575 14.349558
 matgras
sexe colza canard
Homme 0.3761062 1.504425
Femme 0.6238938 2.495575
```