# *Effectuer le test sous R*

***2. Le test statistique :***

***Code :***

```
shapiro.test(musaraignes)
```

***Sortie :***

```
        Shapiro-Wilk normality test

data:  musaraignes
W = 0.92652, p-value = 0.2146
```

**Mais que test ce test ??**

H0 : la distribution de l'échantillon est compatible avec la loi normale

H1 : la distribution de l'échantillon n'est pas compatible avec la loi normale




**Que conclure??**

**A. H0 : ~ Normale** ✓ **A** **B. H1 : pas Normale**

# *Effectuer le test sous R*

***Code :***

```
t.test(musaraignes, mu=23.3, alternative = "two.sided")
```

Pour un test bilatéral (par défaut)

***Sortie :***

```
        One Sample t-test

data:  musaraignes
t = 2.7846, df = 15, p-value = 0.01389
alternative hypothesis: true mean is not equal to 23.3
95 percent confidence interval:
 24.28511 30.71489
sample estimates:
mean of x
     27.5
```

$t_{obs}$

p-value

Moyenne échantillon

Rappel $H_1$ testée

Moyenne de référence (23.3) pas incluse dans $IC_{\mu}95\%$ de moyenne μ de la population dont est issue l'échantillon.

# *Effectuer le test sous R*

***Code :***

```
t.test(musaraignes, mu=23.3, alternative = "greater")
```

Pour un test unilatéral droit

***Sortie :***

```
        One Sample t-test

data:  musaraignes
t = 2.7846, df = 15, p-value = 0.006943
alternative hypothesis: true mean is greater than 23.3
95 percent confidence interval:
 24.85586      Inf
sample estimates:
mean of x
     27.5
```

$t_{obs}$

p-value

Rappel $H_1$ testée

Moyenne échantillon

Moyenne de référence (23.3) pas incluse dans $Ic_{\mu}95\%$

# *Effectuer le test sous R*

***Code :***

```
t.test(musaraignes, mu=23.3, alternative = "less")
```

Pour un test unilatéral gauche

***Sortie :***

```
        One Sample t-test

data:  musaraignes
t = 2.7846, df = 15, p-value = 0.9931
alternative hypothesis: true mean is less than 23.3
95 percent confidence interval:
     -Inf 30.14414
sample estimates:
mean of x
     27.5
```

Moyenne de référence (23.3) **incluse** dans $Ic_{\mu}95\%$

# Plan du cours

## Partie 4. Les tests sur les différences de moyennes


1. Test de conformité : moyenne observée vs théorique

**2. Test d'homogénéité : Comparer les moyennes de plusieurs échantillons**

**2.1. Comparaison de deux échantillons indépendants**

**2.1.1. Test paramétrique : t de Student (Welch)**
***(Comparaison de variances : test F de Fisher)***

2.1.2. Test non paramétrique : Wilcoxon -Mann-Whitney

2.3. Comparaison de deux échantillons appariés

2.3.1. Test paramétrique : t de Student apparié

2.3.2. Test non paramétrique : Wilcoxon apparié

2.4. Comparaison de trois échantillons ou plus

2.4.1. Test paramétrique : ANOVA 1 facteur

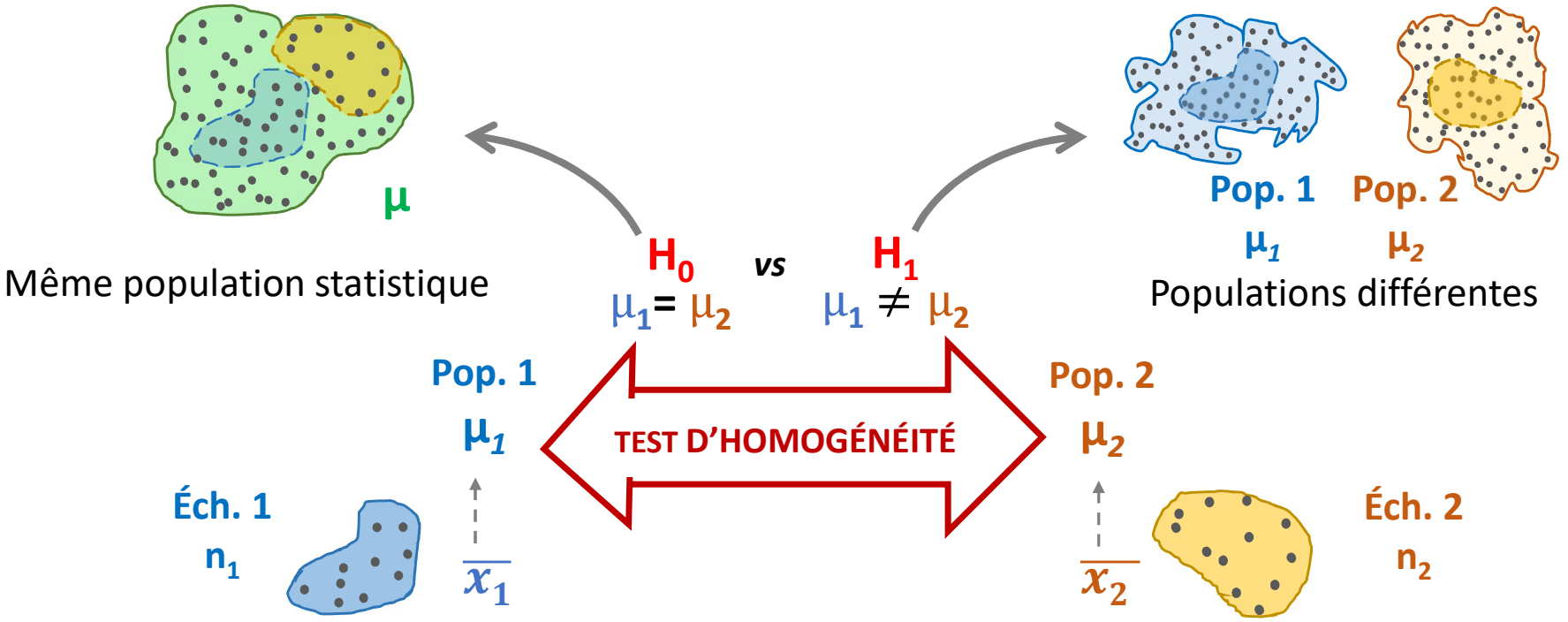
2.4.2. Test non paramétrique : Kruskal-Wallis

# Comparaison deux échantillons indépendants - Paramétrique

**Les données :**

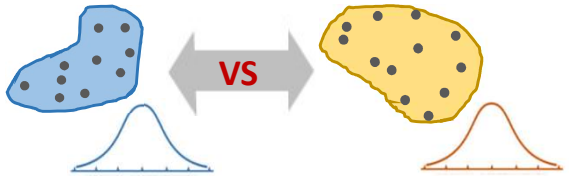
1 VA quantitative X mesurée sur 2 échantillons indépendants (VI qualitative à 2 modalités).
Ils sont issus de 2 populations dont on estime les moyennes ($\mu_1$, $\mu_2$) et variances ($\sigma^2_1$ $\sigma^2_2$).

**Principe :** On calcule une différence entre les moyennes pour trancher entre 2 hypothèses.



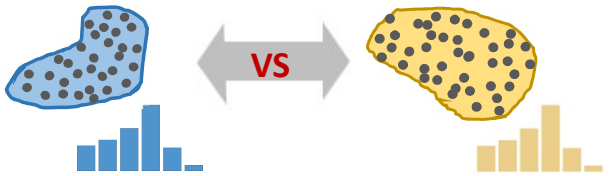
# Comparaison deux échantillons indépendants - Paramétrique

Comme pour le calcul de l'intervalle de confiance, on distingue globalement **deux cas** :




**Petits ($n_1 \leq 30$ et/ou $n_2 \leq 30$), Distri. ~ loi Normale**

→ ***Cas 1 :*** Approximation par les lois de Student : **Test t de Student**




**Grands ($n_1 > 30$ et $n_2 > 30$), Distri. quelconque**

→ ***Cas 2 :*** Approximation par la loi Normale CR (TLC) : **Test Z**

On ne **détaillera que** le cas du test de **Student** car :

- La statistique t tend vers Z pour les grands échantillons (TLC)
- R fait automatiquement la correction via la fonction t.test()

# Comparaison deux échantillons indépendants - Paramétrique

**Conditions d'application :** (vérifier avant de réaliser le test !)

- **Indépendance** des observations au sein des échantillons (par opposition à appariés)

(mesure de X sur des individus ≠ et sans liens, effectifs éch1 et éch2 pas forcement égaux)

- **Normalité** distribution de $X_1$ et $X_2$ dans les 2 échantillons (graphique + test Shapiro-Wilk)

→ Peut s'en affranchir si les échantillons sont grands ($n_1 > 30$ **et** $n_2 > 30$) (TLC)

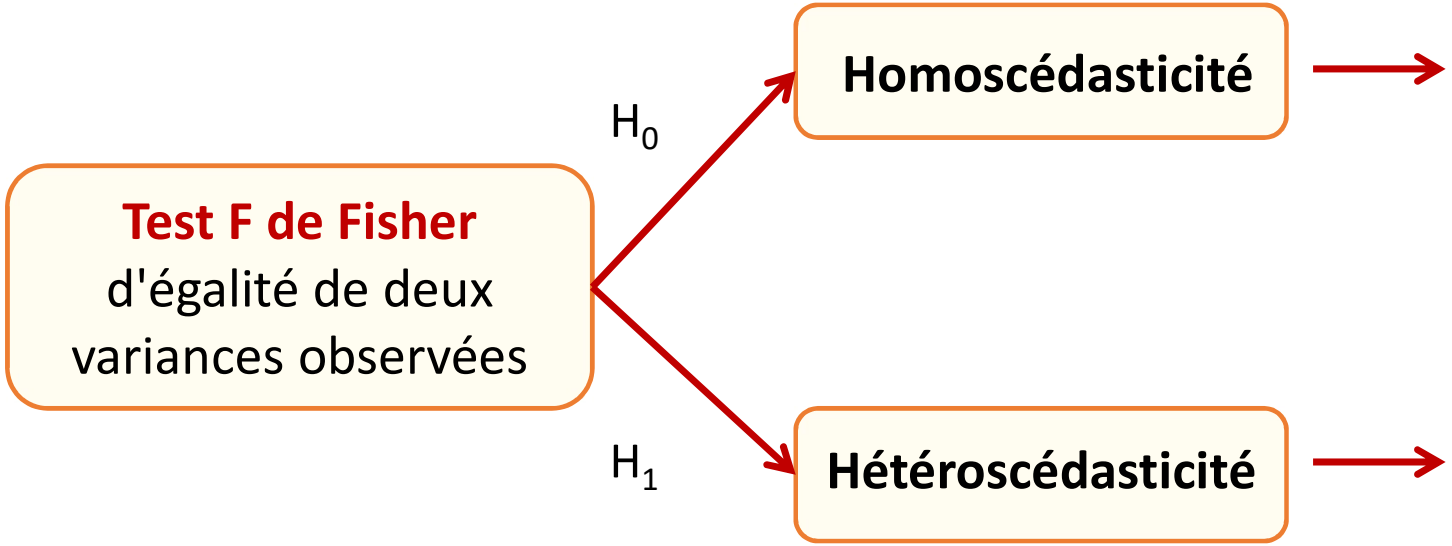
- **Homoscédasticité** = égalité des variances des 2 échantillons (test Fisher)

→ Peut s'en affranchir si les tailles des échantillons sont voisines ($n_1 \approx n_2$)

Si au moins 1 des 2 échantillons est trop petit ($n_1 \leq 30$ ou $n_2 \leq 30$) **et** que la variable étudiée ne suit pas la loi Normale → approche non-paramétrique.

# Comparaison deux échantillons indépendants - Paramétrique

**Avant tout, il faut vérifier la condition d'Homoscédasticité** = « égalité » des variances**:**




***Cas 1 :***

Approximation par les lois de Student :

**Test t de Student**

***Cas 2 :***

**Test t de Welch**

(« Student corrigé »)

***Ou***

**Test de Wilcoxon-Mann-Whitney**

# (Comparaison de variances)

***Pourquoi comparer des variances ?***

**Homoscédasticité** : condition nécessaire à la réalisation d'un bon nombre de tests…

→ La différence entre $S_1^{2*}$ et $S_2^{2*}$ est elle significative ou bien s'explique-t-elle par les fluctuations d'échantillonnage?

**Test F de Fisher** — Pour **2 échantillons** seulement

**Test Bartlett (ou Levene)** — Pour **> 2 échantillons**

Ces tests sont très **sensibles à l'hypothèse de Normalité**, valide que si :

- Les échantillons sont grands ( ≥ 30)
- Au moins 1 des échantillons est petit (<30) mais de distribution Normale

**Test Fligner-Killeen** — Option **non paramétrique pour ≥ 2 échantillons**

# (Comparaison de variances – Test F de Fisher)

**Objectif : comparer le rapport des 2 variances estimées** ($S_2^{2*}$ et $S_1^{2*}$) avec la loi de Fisher.

Ce rapport est supposé égal à 1 sous $H_0$ (cad si les variances sont égales).

**Hypothèses :**

$\mathbf{H_0}$ : $\sigma_1^2 = \sigma_2^2 \rightarrow \sigma_1^2 / \sigma_2^2 = 1$

$\mathbf{H_1}$ : $\sigma_1^2 \neq \sigma_2^2 \rightarrow \sigma_1^2 / \sigma_2^2 \neq 1$

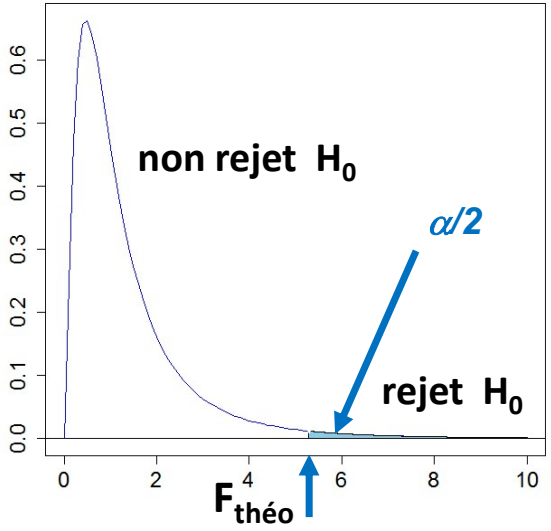
Généralement en bilatéral.

**Statistique de test :**

$$F = \frac{\max(S_1^{2*}, S_2^{2*})}{\min(S_1^{2*}, S_2^{2*})}$$

*Variance max en numérateur car loi de Fisher pas symétrique!*

- Valeur critique $F_{théo}$ → **Table Fisher**, ddl = ($n_{S^2max}$-1, $n_{S^2min}$-1)
- Comparaison de $F_{obs}$ à la distribution de F sous $H_0$.




**Règle de décision :**

Si $F_{obs} \geq F_{théo}$ ou p-value < seuil α fixé

➔ Rejet de $H_0$ (Homoscédasticité) et on accepte $H_1$ (Hétéroscédasticité)

# Comparaison deux échantillons indépendants - Paramétrique

***Cas 1 : homoscédasticité non rejetée*** → Approximation par les **lois de Student (Test t)**

**Hypothèses :**

$H_0 : \overline{x_1} = \overline{x_2}$
$H_1 : \overline{x_1} \neq \overline{x_2}$

**Statistique de test :**

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Calcul de la **variance commune $\widehat{\sigma^2}$ :**

$$\widehat{\sigma^2} = \frac{(n_1 - 1)s_1^{2*} + (n_2 - 1)s_2^{2*}}{n_1 + n_2 - 2}$$

- Valeur critique $t_{théo}$ → **table de Student**, ddl = $n_1+n_2-2$ → borne(s) de (non)rejet de $H_0$.
- Calculer $t_{obs}$ à partir des observations et comparaison à la distribution de **t** sous $H_0$. .

**Règle de décision :** Si | $t_{obs}$ | ≥ | $t_{théo}$ | ou p-value < α ➔ Rejet de $H_0$ et on accepte $H_1$.

# Comparaison deux échantillons indépendants - Paramétrique

***Application :***



On dispose de deux échantillons (mâles vs. femelles) de souris des cactus (*Peromyscus eremicus*) dont on a mesuré la masse (g) chez des adultes.

**Femelles :** 31, 25, 29, 30, 31, 28, 31, 29, 29, 33, 30, 28

**Mâles :** 28, 29, 30, 29, 27, 26, 27, 28, 25, 28

**Peut-on déceler, avec un risque d'erreur $\alpha$ de 5%, une différence de la moyenne du poids (g) chez l'individu adulte entre mâles et femelles ?**

→ *Voir la réalisation du test « à la main » en vidéo sur MesCoursJV* ←

# *Effectuer le test sous R*

## *Etape 1 : Créer le jeu de données (data frame)*

### *Code :*

```
Fem<-c(31, 25, 29, 30, 31, 28, 31, 29, 29, 33, 30, 28)
Mal<-c(28, 29, 30, 29, 27, 26, 27, 28, 25, 28)

n_Fem<-length(Fem)
n_Mal<-length(Mal)

Sexe<-c(rep("F",12),rep("M",10))

souris<-data.frame(c(Fem,Mal),Sexe)
souris$Sexe<-as.factor(souris$Sexe)
names(souris)<-c("Masse","Sexe")

str(souris)
View(souris)
```

### *Sorties :*

```
> n_Fem
[1] 12
> n_Mal
[1] 10
```

| | Masse | Sexe |
|---|---|---|
| 1 | 31 | F |
| 2 | 25 | F |
| 3 | 29 | F |
| 4 | 30 | F |
| 5 | 31 | F |
| 6 | 28 | F |
| 7 | 31 | F |
| 8 | 29 | F |
| 9 | 29 | F |
| 10 | 33 | F |
| 11 | 30 | F |
| 12 | 28 | F |
| 13 | 28 | M |

*(…)*

**Facteur à « k » modalités :**

échantillons à comparer (F=1, M=2) →

```
'data.frame':   22 obs. of  2 variables:
 $ Masse: num  31 25 29 30 31 28 31 29 29 33 ...
 $ Sexe : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
```

# *Effectuer le test sous R*

## *Etape 2 : Quelques stat descriptives*

### *Code :*

```
xbarre_Fem<-mean(souris$Masse[souris$Sexe=="F"])
xbarre_Fem<-mean(subset(souris$Masse,Sexe=="F"))

xbarre_Mal<-mean(subset(souris$Masse,Sexe=="M"))
```

*Equivalents (voir TP)*

```
var_Fem<-var(subset(souris$Masse,Sexe=="F"))
var_Mal<-var(subset(souris$Masse,Sexe=="M"))
```

```
sd_Fem<-sd(subset(souris$Masse,Sexe=="F"))
sd_Mal<-sd(subset(souris$Masse,Sexe=="M"))
```

### *Sorties :*

```
> xbarre_Fem
[1] 29.5
> xbarre_Mal
[1] 27.7
```

```
> var_Fem
[1] 4.090909
> var_Mal
[1] 2.233333
```

```
> sd_Fem
[1] 2.0226
> sd_Mal
[1] 1.494434
```

# *Effectuer le test sous R*

## *Etape 3 : Vérification de la condition de normalité des distributions*

### *Code :*

***# Pour les femelles :***

```
plot(density(souris$Masse[souris$Sexe=="F"]))

plot(density(subset(souris$Masse,Sexe=="F")))
```

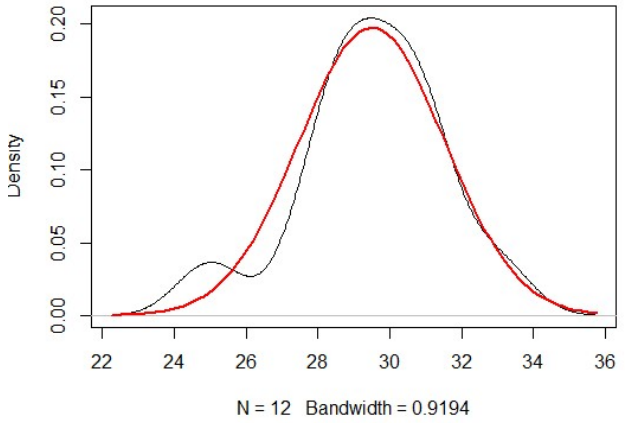
*Equivalents*

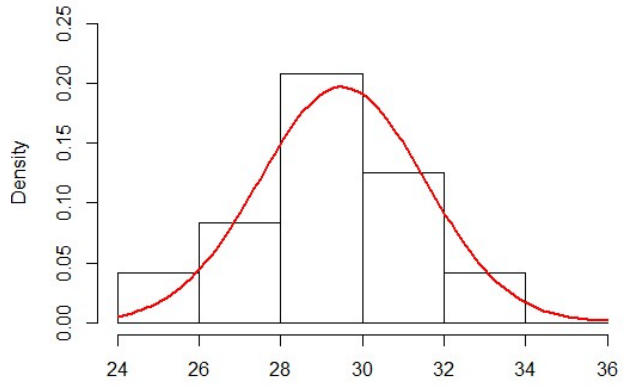
*# Superposer une courbe normale (avec $\bar{x}$ et **S*** estimés de l'échantillon) :*

```
curve(dnorm(x,
            mean=mean(subset(souris$Masse,Sexe=="F")),
            sd=sd(subset(souris$Masse,Sexe=="F"))),
      col="red",lwd=2,add=TRUE)


hist(subset(souris$Masse,Sexe=="F"),
     breaks=seq(24,36,2),
     prob=T, ylim=c(0,0.25))
```

### *Sorties :*



# *Effectuer le test sous R*

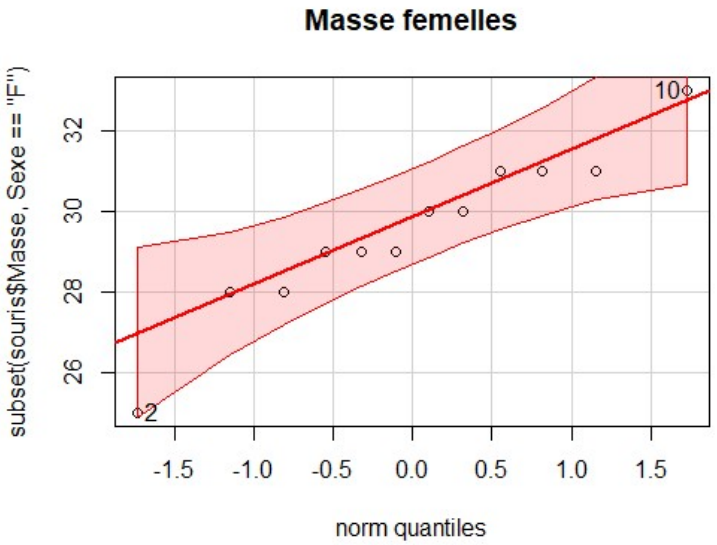
## *Etape 3 : Vérification de la condition de normalité des distributions*

### *Code :*

***# Pour les femelles :***

```
qqPlot(subset(souris$Masse,Sexe=="F"),
       main="Masse femelles", col.lines = "red")
```

### *Sorties :*




```
shapiro.test(subset(souris$Masse,Sexe=="F"))
```

```
        Shapiro-Wilk normality test

data:  subset(souris$Masse, Sexe == "F")
W = 0.94506, p-value = 0.5662
```

# *Effectuer le test sous R*

## *Etape 3 : Vérification de la condition de normalité des distributions*

### *Code :*

***# Pour les mâles :***

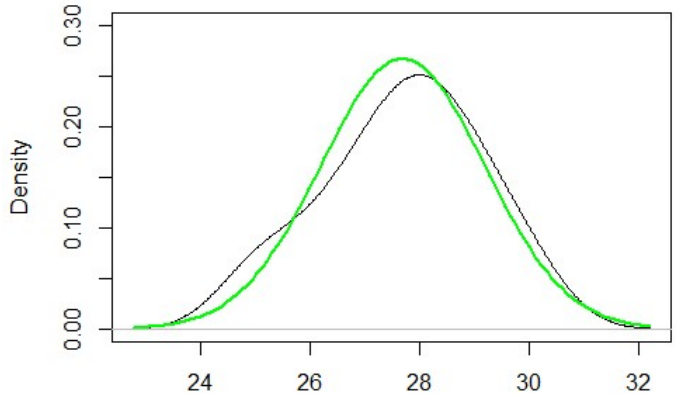
```
plot(density(subset(souris$Masse,Sexe=="M")),
     ylim=c(0,0.3))
```

*# Superposer une courbe normale (avec $\bar{x}$ et **S** estimés de l'échantillon) :*

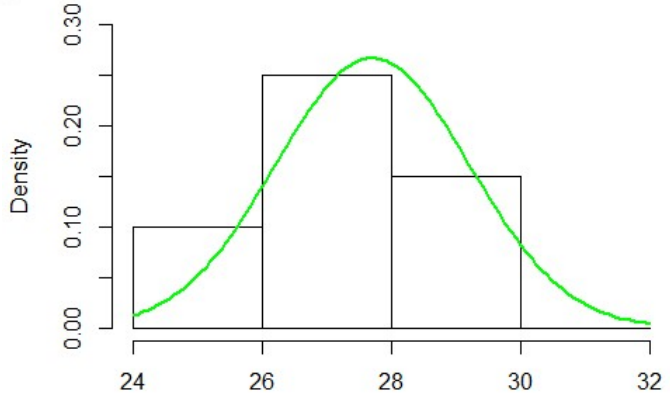
```
curve(dnorm(x,
            mean=mean(subset(souris$Masse,Sexe=="M")),
            sd=sd(subset(souris$Masse,Sexe=="M"))),
            col="green",lwd=2,add=TRUE)
```

```
hist(subset(souris$Masse,Sexe=="M"),
     breaks=seq(24,32,2),
     prob=T, ylim=c(0,0.3))
```

### *Sorties :*




N = 10 Bandwidth = 0.7416



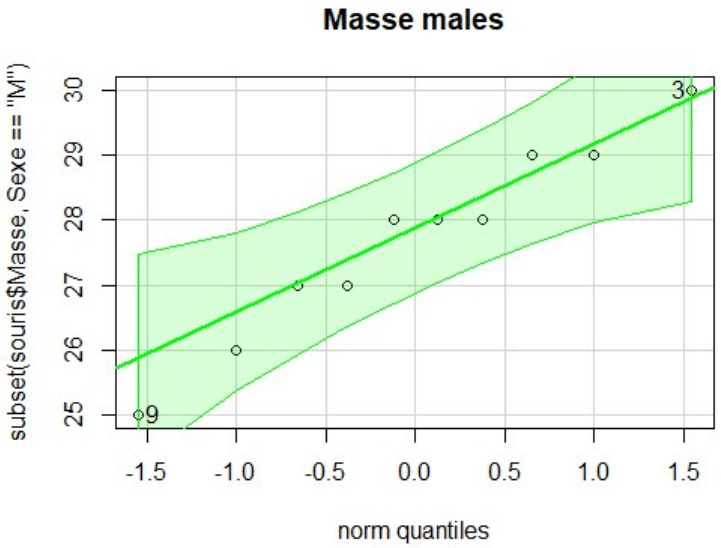
# *Effectuer le test sous R*

## *Etape 3 : Vérification de la condition de normalité des distributions*

### *Code :*

***# Pour les mâles :***

```
qqPlot(subset(souris$Masse,Sexe=="M"),
       main="Masse males", col.lines = "green")
```




```
shapiro.test(subset(souris$Masse,Sexe=="M"))
```

```
        Shapiro-Wilk normality test

data:  subset(souris$Masse, Sexe == "M")
W = 0.96624, p-value = 0.854
```