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")))
```

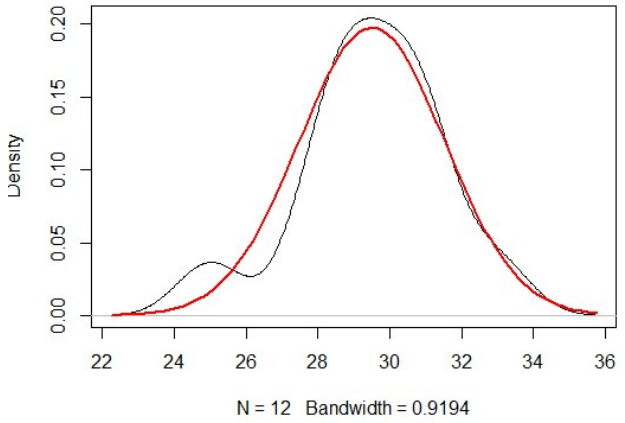
Equivalents

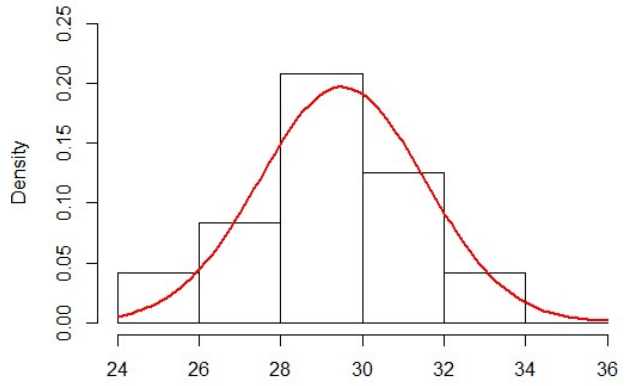
*# Superposer une courbe normale (avec $\bar{x}$ et **S*** estimés de l'échantillon) :*

```
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

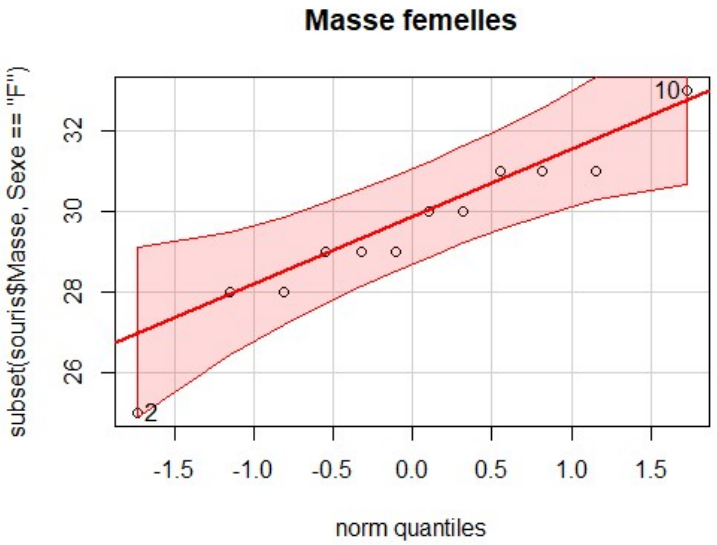
Etape 3 : Vérification de la condition de normalité des distributions

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 :

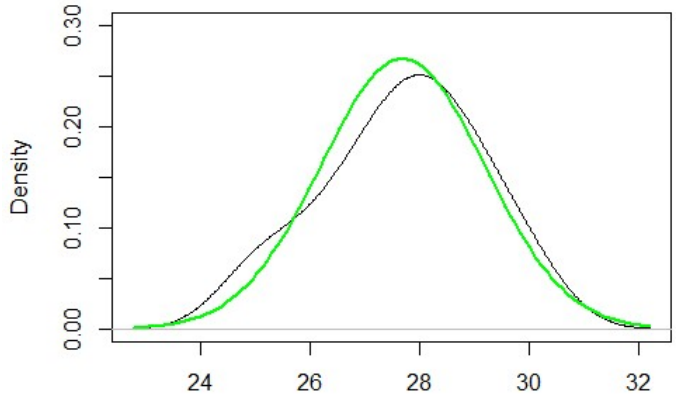
```
plot(density(subset(souris$Masse,Sexe=="M")),
     ylim=c(0,0.3))
```

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

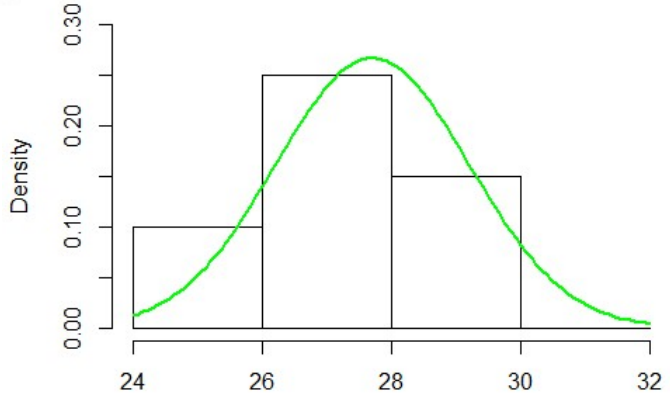
```
curve(dnorm(x,
            mean=mean(subset(souris$Masse,Sexe=="M")),
            sd=sd(subset(souris$Masse,Sexe=="M"))),
            col="green",lwd=2,add=TRUE)
```

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

Sorties :



N = 10 Bandwidth = 0.7416



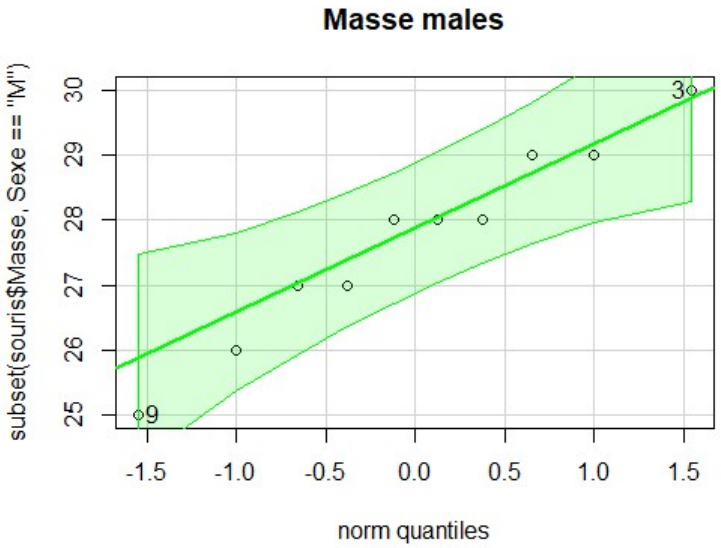
Effectuer le test sous R

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
```

Effectuer le test sous R

Etape 4 : Vérification de la condition d'homogénéité des variances (homoscedaticité)

Code :

```
var.test(souris$Masse~souris$Sexe)
var.test(Masse~Sexe,data=souris)
```

Equivalents

~ *= « **en fonction de** » (étudie les variations de VD ~ VI)*

*S'obtient avec : **alt gr** + touche **2** (PC) ou **alt** ⌥ + touche **N** (Mac)*

Sortie :

```
        F test to compare two variances

data:  souris$Masse by souris$Sexe
F = 1.8318, num df = 11, denom df = 9, p-value = 0.3726
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.4682299 6.5721346
sample estimates:
ratio of variances
           1.83175
```

Effectuer le test sous R

Etape 5 : le test t de Student pour 2 échantillons indépendants

Code : *(Equivalents)*

```
t.test(souris$Masse~souris$Sexe,
       var.equal=TRUE)

t.test(Masse~Sexe,data=souris,
       var.equal=T)
```

Précise que l'égalité des variances est respectée (test F Fisher)

Sorties :

```
        Two Sample t-test

data:  Masse by Sexe
t = 2.3301, df = 20, p-value = 0.03039
alternative hypothesis: true difference
 in means is not equal to 0
95 percent confidence interval:
 0.1886017 3.4113983
sample estimates:
mean of x mean of y
     29.5      27.7
```

Pour chez vous

Pour calculer la variance commune et t_{obs} avec R :

```
var_com1<-((n_Fem-1)*var_Fem+(n_Mal-1)*var_Mal)/(n_Fem+n_Mal-2)

t_obs<-(xbarre_Fem-xbarre_Mal)/sqrt(var_com1*(1/n_Fem+1/n_Mal))
```

```
> var_com1
[1] 3.255

> t_obs
[1] 2.330109
```

Effectuer le test sous R

Etape 5 : le test t de Student pour 2 échantillons indépendants

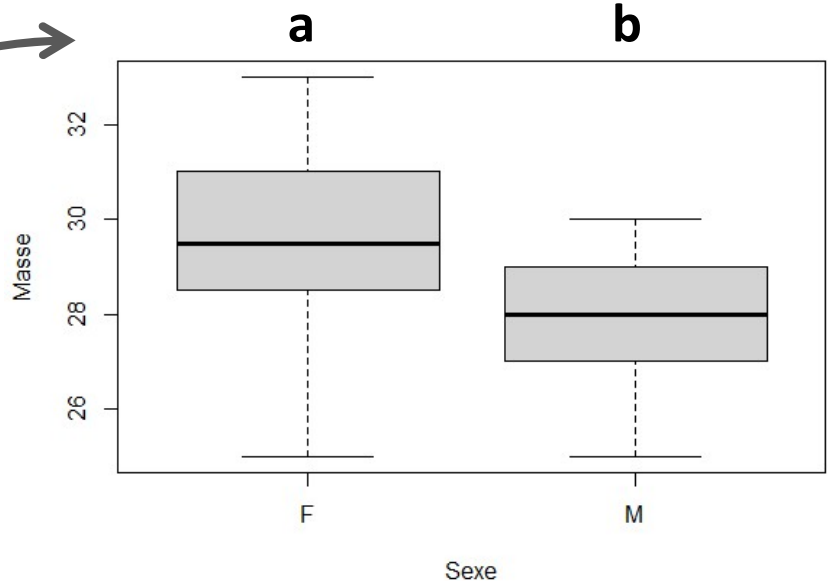
Boxplot = graphique adapté à la représentation visuelle de résultats de tests de comparaison moyenne (ou médiane) entre plusieurs échantillons.

Code :

```
boxplot(Masse~Sexe,data=souris)
```

La VI doit être un Facteur

Sorties :



Lettre en commun = pas de différence significative

CCL : La masse des souris des cactus diffère significativement selon le sexe (p-value < 0. 05).

Effectuer le test sous R

Etape 5 : le test t de Student pour 2 échantillons indépendants (unilatéral)

Code :

 # Masse F > M ?

```
t.test(Masse~Sexe,data=souris,
       var.equal=T, alternative = "greater")
```

 # Masse F < M ?

```
t.test(Masse~Sexe,data=souris,
       var.equal=T, alternative = "less")
```

Rappel : (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 ...
```

Sorties :

```
        Two Sample t-test

data:  Masse by Sexe
t = 2.3301, df = 20, p-value = 0.01519
alternative hypothesis: true difference in means
 between group F and group M is greater than 0
95 percent confidence interval:
 0.4676621       Inf
sample estimates:
mean in group F mean in group M
           29.5            27.7
```

```
        Two Sample t-test

data:  Masse by Sexe
t = 2.3301, df = 20, p-value = 0.9848
alternative hypothesis: true difference in means
 between group F and group M is less than 0
95 percent confidence interval:
     -Inf 3.132338
sample estimates:
mean in group F mean in group M
           29.5            27.7
```

Effectuer le test sous R

Cas 2 : homoscédasticité rejetée → **Test de Welch**

→ Version corrigée du test t : sans variance commune, mais avec un nombre de ddl inférieur.

Code :

```
t.test(Masse~Sexe,data=souris,
       var.equal=F)
```

NB : **par défaut** c'est le test de **Welch** qui est réalisé
→ R part du principe que l'**homoscédasticité n'est pas respectée**

Sorties :

```
        Welch Two Sample t-test

data:  Masse by Sexe
t = 2.3963, df = 19.765, p-value = 0.02658
alternative hypothesis: true difference in
 means between group F and group M is not
 equal to 0
95 percent confidence interval:
 0.2319136 3.3680864
sample estimates:
mean in group F mean in group M
          29.5            27.7
```

Une autre solution est d'appliquer un test non paramétrique : Wilcoxon-Mann-Whitney

Plan du cours

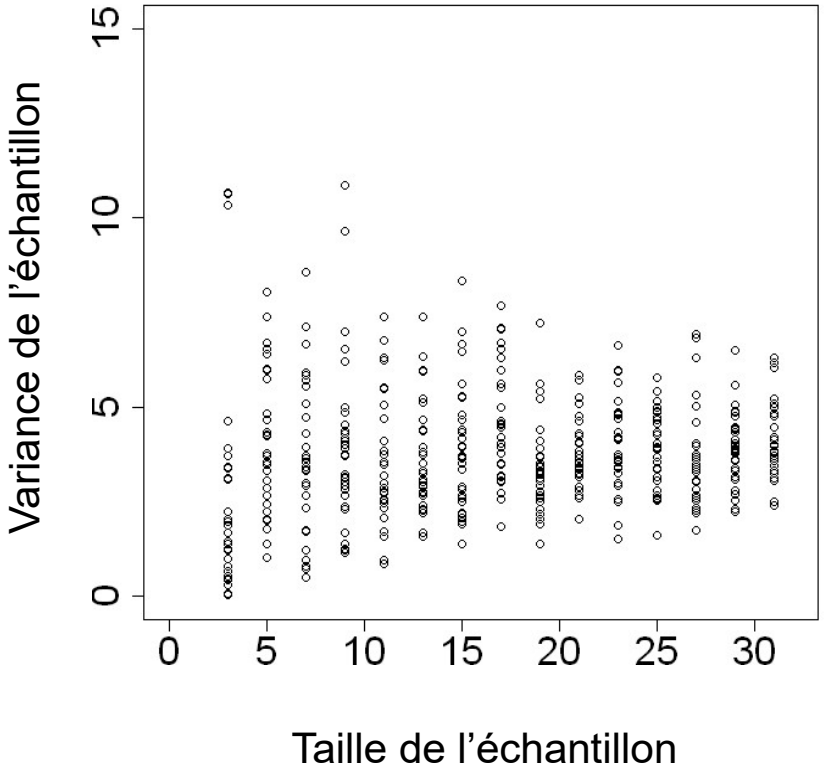
Partie 4. Les tests sur les différences de moyennes

Rappel sur les tests non-paramétriques

S'utilisent quand les **conditions tes tests paramétriques ne sont pas remplies** (distribution variable inconnue ou non Normale, hétéroscédasticité) et sont **adaptés aux très petits échantillons** (ex: n<6).

Ce type de test est fortement recommandé quand les échantillons sont de petites tailles car la variance est de plus en plus variable, même pour des échantillons issus de la même population N(μ, σ).

→ Augmente le risque de ne pas satisfaire la condition d'homogénéité des variances en paramétrique.



Rappel sur les tests non-paramétriques

- Permettent de tirer des conclusions intéressantes à partir d'échantillons sans faire **aucune hypothèse sur la distribution**, et donc loi de probabilité, de la VD.
  → pas besoin de paramètres exactes comme pour tests paramétriques (ex: Student).

- On **remplace les valeurs réelles** prises par la VD, **par le rang** de ces valeurs
  → passage donc à une échelle ordinale (doit pouvoir ordonner les valeurs).

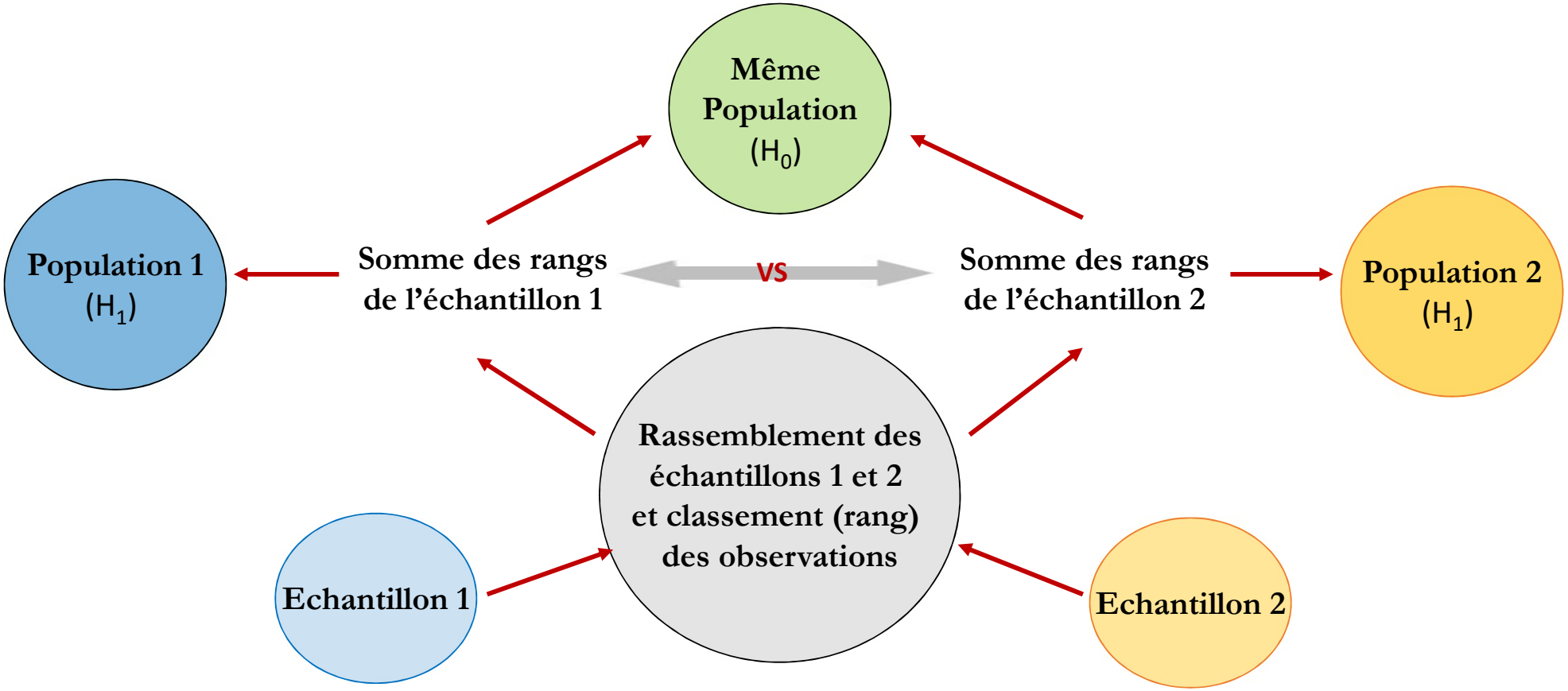
- **Tests universels :** VA quantitatives continues (non normales, loi de proba. inconnue, petits échantillons), VA quantitatives discrètes, VA qualitatives ordinales (nominales = χ^2).

- Légère perte de puissance par rapport tests paramétriques pour déceler un effet faible sur la VA étudiée, mais **méthodes plus robustes** aux valeurs extrêmes (outliers).

Comparaison deux échantillons indépendants – Non paramétrique

Test Wilcoxon-Mann-Whitney : 2 tests concurrents, mêmes résultats, cross déductibles.

ECQ, au risque α choisi, deux échantillons indépendants sont issus d'une **même population (H_0)** ou bien de **2 populations différentes (H_1)** vis-à-vis de la variable étudiée ?



Comparaison deux échantillons indépendants – Non paramétrique

Principe de la méthode des rangs :

ECQ les éléments de deux échantillons indépendants, **classés par ordre croissant** sur une même échelle de valeurs, **occupent des** positions (= « **rangs** ») **équivalents ou pas** ?
→ Similitude des distributions (ex : proximité de leurs médianes)… ou non….

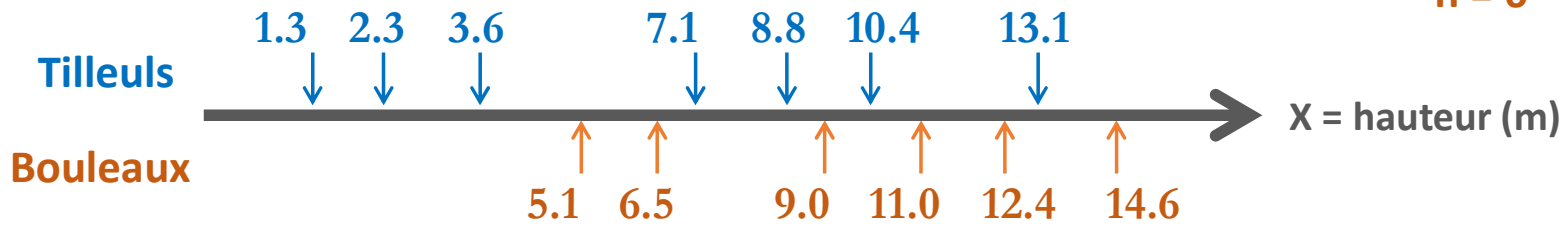
n = 7

Hauteur des tilleuls (m) :
1.3, 7.1, 8.8, 2.3, 13.1, 3.6, 10.4

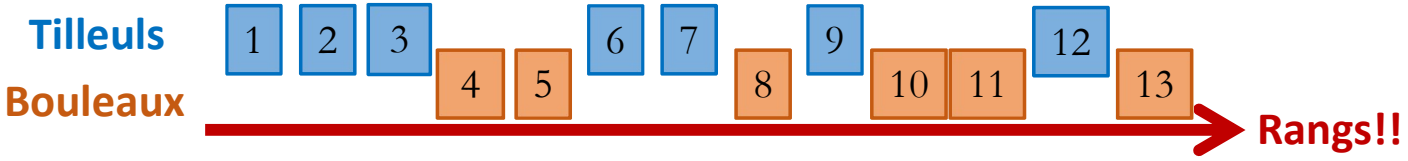
Hauteur des bouleaux (m) :
9.0, 14.6, 5.1, 6.5, 12.4, 11.0

n = 6

1) On classe les valeurs de X :



2) On les range :

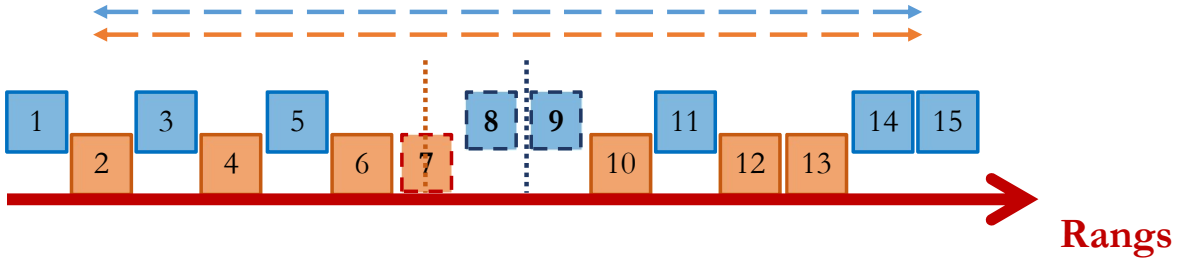


Comparaison deux échantillons indépendants – Non paramétrique

On s'intéresse à l'interclassement des 2 échantillons.

H_0 : La somme des rangs des éléments de l'éch.1 est égale à celle de l'éch.2

Ech.1 et Ech. 2 «interclassés/bien mélangés»

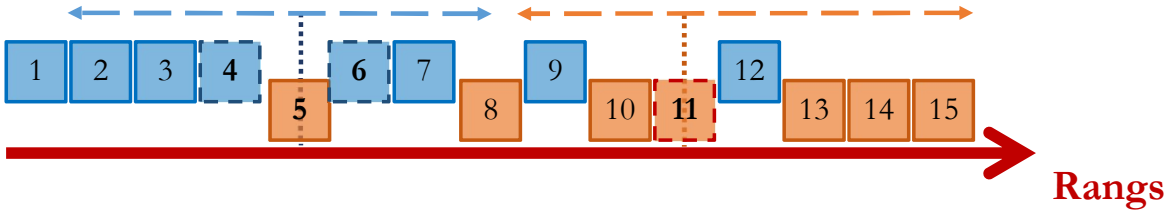


P(X1>X2) = P(X1<X2) = 0.5

Distribution homogène des rangs
(Les médianes sont proches)
→ même population

H_1 : La somme des rangs des éléments de l'éch.1 est différente de celle de l'éch.2

Ech.1 et Ech. 2 «mal interclassés/séparés»



P(X1>X2) ≠ P(X1<X2) ≠ 0.5

Distribution hétérogène des rangs
(Les médianes sont éloignées)
→ populations différentes

Comparaison deux échantillons indépendants – Non paramétrique

2.1.2.1. Test de Wicoxon-Mann-Whitney : Cas des petits échantillons ($9 \leq n_1$ et/ou $n_2 \leq 20$)

On va calculer un **score (U)** basé sur la ∑ des rangs des observations de chaque échantillon.

Les hypothèses :

$H_0 : U_1 = U_2$

$H_1 : U_1 \neq U_2$

La statistique de test :

$U_{obs} = \min (U_1 , U_2)$

La méthode :

1. Regrouper et **classer** par ordre croissant l'ensemble des observations des 2 échantillons (en repérant l'origine de chaque valeur -échantillon 1 ou 2-, et les ex-aequo)
2. Affecter un **rang** à chaque observation (rang médian en cas d'ex-aequo)
3. Calculer la **somme des rangs** de l'éch.1 (R_1) et celle de l'éch.2 (R_2)

Comparaison deux échantillons indépendants – Non paramétrique

4. **Calculer le score U_i** de chaque échantillon (***U_1*** et ***U_2***) avec : $$U_i = R_i - \frac{n_i(ni+1)}{2}$$

   NB : $U_2 + U_1 = n_2 * n_1$

5. **U_{obs}** correspond à **la plus petite valeur** entre ***U_1*** et ***U_2***.

6. On compare **U_{obs}** à la valeur **$U_{théo}$** → Table de Mann-Whitney bilatérale (seuil **α**, **n_1** et **n_2**).

Règle de décision :

Si **$U_{obs} \leq U_{théo}$** alors **on rejette H_0** au risque α

C'est l'inverse des test t et Z !!!

TABLE DE MANN-WHITNEY

Valeurs critiques (U_{crit}) à comparer avec la valeur observée (U_{obs}) à partir de vos 2 écha
un test bilatéral au seuil α = 0.05 ou 0.01.

NB : n_1 et n_2 représentent le nombre d'observations dans chaque échantillon.

n_2	α	n_1 3	4	5	6	7	8	9	10	11	12	13	14
3	.05	--	0	0	1	1	2	2	3	3	4	4	5
	.01	--	0	0	0	0	0	0	0	0	1	1	1
4	.05	--	0	1	2	3	4	4	5	6	7	8	9
	.01	--	--	0	0	0	1	1	2	2	3	3	4
5	.05	0	1	2	3	5	6	7	8	9	11	12	13
	.01	--	--	0	1	1	2	3	4	5	6	7	7
6	.05	1	2	3	5	6	8	10	11	13	14	16	17
	.01	--	0	1	2	3	4	5	6	7	9	10	11
7	.05	1	3	5	6	8	10	12	14	16	18	20	22
	.01	--	0	1	3	4	6	7	9	10	12	13	15

Comparaison deux échantillons indépendants – Non paramétrique

Et pour un test unilatéral ?

Pour chez vous
+ voir vidéo

- ***Si on veut tester H_1 : Ech.1 > ech.2***

  Cela implique que $P(X_1>X_2) > 0.5 \rightarrow U_1 > U_2$

  On s'attend à ce que les rangs de l'éch.1 soient décalés vers les grandes valeurs de X

  On prendra $U_{obs} = U_2 \rightarrow$ ***U** de l'éch. dont on fait l'hypothèse qu'il est le plus petit*

- ***Si on veut tester H_1 : Ech.1 < ech.2***

  Cela implique que $P(X_1>X_2) < 0.5 \rightarrow U_1 < U_2$

  On s'attend à ce que les rangs de l'éch.1 soient décalés vers les petites valeurs de X

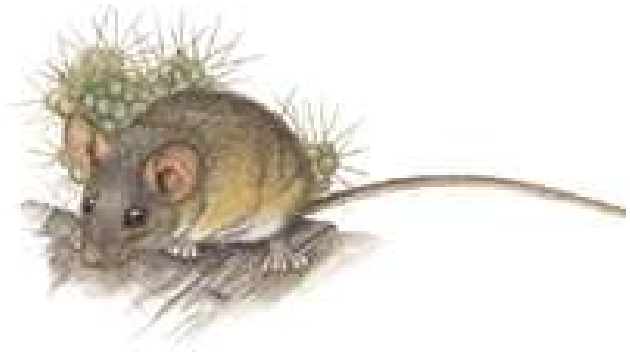
  On prend $U_{obs} = U_1 \rightarrow$ ***U** de l'éch. dont on fait l'hypothèse qu'il est le plus petit*

➔ Dans tous les cas, si $U_{obs} \leq U_{théo,\ \alpha}$: rejet de H_0 au profit de H_1(idem test bilatéral).

Comparaison deux échantillons indépendants – Non paramétrique

Application :

On s'intéresse à nouveau à la différence de masse (g) entre mâles et femelles de souris des cactus adultes (*Peromyscus eremicus*).



On a prélevé cette fois des échantillons plus petits:

Femelles (n = 7) : 24, 30, 30, 38, 40, 32, 28

Mâles (n = 5) : 18, 20, 28, 24, 26

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

Importer un jeu de données sous R

1. **Télécharger** et **ouvrir** le data frame «Souris2_excel » (Moodle, section Jeux de données)

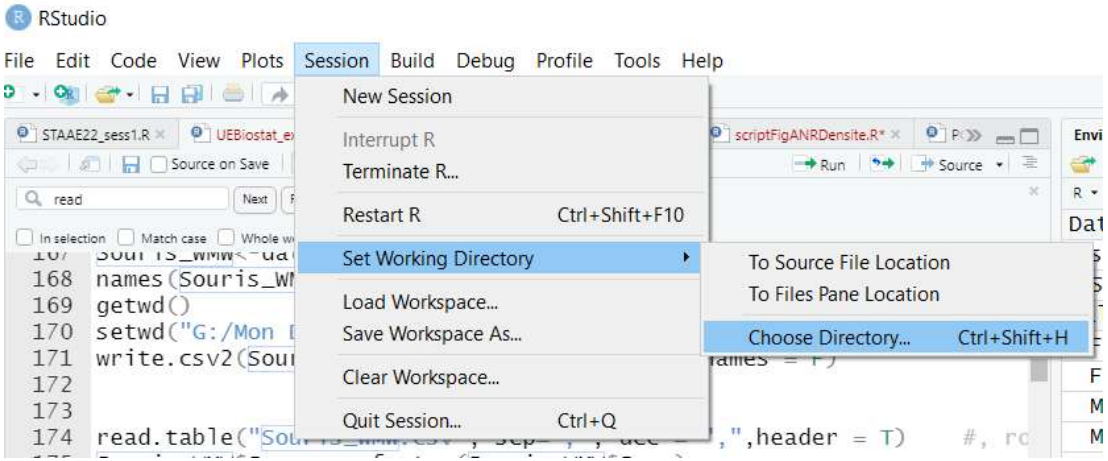
2. Via votre tableur, **enregistrer** le data frame dans un **format** facilement lisible par R :

   Enregistrer sous → votre répertoire de travail R (ex: TP_R/data)

   *Type de fichier → **CSV** (séparateur: point virgule)*

   Nom → «Souris_WMW»

3. Rappeler à R le chemin d'accès à votre **répertoire de travail** :



→ **Copier/coller** dans votre script la ligne de code apparaissant dans la console (**fonction setwd()**, voir TP).

Ex : setwd("G:/Mon Drive/Biostat/DataTP/JDD_CMTD")

Importer un jeu de données sous R

VD quantitative

VI = Facteur (k = 2 modalités)

4. Ouvrir le fichier .csv dans R :

```
Souris_WMW<-read.table("Souris_WMW.csv", sep=";", dec = ",",header = T)
```

Permet d'ouvrir une grande variété de type de fichiers (csv, csv2, txt…)

Souris_WMW = format csv2 (français) : sep= ";", dec=","

Pour préciser que la 1ère ligne = en tête de colonnes (noms variables)

	Masse	Sexe
1	24	Femelle
2	30	Femelle
3	30	Femelle
4	38	Femelle
5	40	Femelle
6	32	Femelle
7	28	Femelle
8	18	Male
9	20	Male
10	28	Male
11	24	Male
12	26	Male

5. Contrôler la qualité de l'importation du jeu de données sous R:

```
View(Souris_WMW)
```

Remarque :

***Format LONG** = 1 colonne par variable*

```
str(Souris_WMW)
```

```
'data.frame':   12 obs. of  2 variables:
 $ Masse: num  24 30 30 38 40 32 28 18 20 28 ...
 $ Sexe : chr  "Femelle" "Femelle" "Femelle" "Femelle" ...
```

```
Souris_WMW$Sexe<-as.factor(Souris_WMW$Sexe)
```

```
str(Souris_WMW)
```

```
'data.frame':   12 obs. of  2 variables:
 $ Masse: num  24 30 30 38 40 32 28 18 20 28 ...
 $ Sexe : Factor w/ 2 levels "Femelle","Male": 1 1 1 1 1 1 1 2
```

Facteur à 2 modalités