Statistiques II

STATISTIQUES MULTIVARIÉES

Rémi Frossard
PSYCHOLOGIE | SSP

Table des matières

Mise en marche

Démarrer le programme

- Lancer le programme R
- Dans la console R, écrire **library(Rcmdr)** pour ouvrir la console Commander

Entrer des données

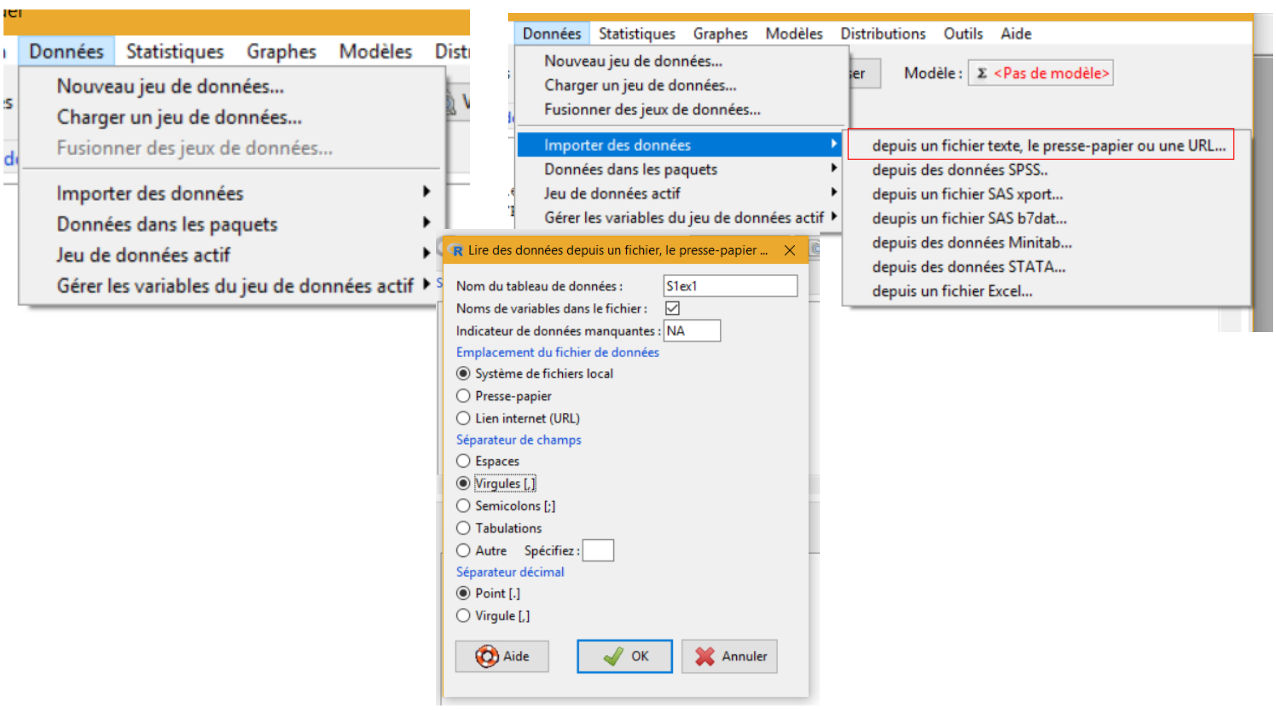
1. Entrer les données sur **Calque**.
   ⚠ *La disposition des données varie selon le type de teste que l'on souhaite effectuer !*

2. Pour entrer correctement les données sur calques, il faut déterminer les différents **groupes** et les différentes **variables**
   ⚠ *Ne pas mettre d'accents sur les mots lorsque l'on entre les données sur Calque, cela empêche le programme de fonctionner correctement !*

	A	B
1	**Var 1**	
2	Scores	
3	49	
4	46	
5	43	
6	40	
7	38	
8	37	
9	48	
10	45	
11	43	
12	39	
13	38	
14	37	
15	48	
16	44	
17	42	
18	38	
19	38	
20	36	

	A	B	C
1	**Var 1**	**Var 2**	
2	Avant	Apres	
3	25	21	
4	28	23	
5	29	22	
6	26	21	
7	28	26	
8	27	29	
9	22	21	
10	25	22	
11	27	23	
12	28	22	
13	29	25	
14			

3. Enregistrer les données au format **CSV**, choisir **Virgule** comme séparateur de champs.
4. Importer les données sur le programme R, en respectant les paramètres suivants :



5. Ensuite, il faut cliquer sur le bouton **Visualiser** pour faire apparaître les données dans un tableau annexe
6. En cliquant sur Editer, il est possible de changer les noms des colonnes

⚠ **Dans la plupart des tests, nous entrons les données sous la forme suivante : 1 ligne par individu.**

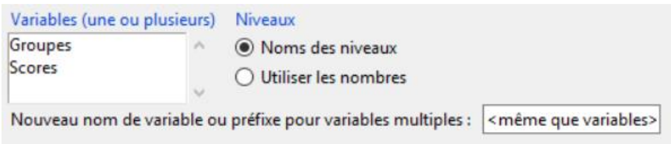
Transformer des variables numériques en variables nominales

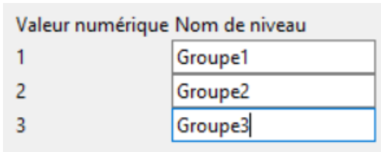
Chemin d'accès : Données / Gérer les variables du jeu de données actif / Convertir des variables numériques en facteur

Scores	Groupes
-9	1
-4	1
-2	1
-2	1
-1	1
0	1
0	1
1	1
-1	2
-1	2
0	2
1	2
1	2
3	2
5	2
6	2
-9	3
-5	3
-5	3
-3	3
-2	3
-1	3
0	3
0	3

Lorsque nous récupérons des données, il est possible que des variables se trouvent sous forme numérique alors que nous souhaiterions des **variables nominales**. C'est le cas par exemple lorsque nous numérotons différents groupes (1, 2 et 3 pour groupe 1, groupe 2 et groupe 3). Si nous laissons ces variables au format numérique, cela faussera les résultats.

Dans la fenêtre, sélectionner la **variable à convertir**, puis cocher **Noms des niveaux**. Ne pas renommer la variable. Ensuite, cliquer sur OK. Une boîte de dialogue s'ouvre, il faut cliquer sur OK. Enfin, une fenêtre s'ouvre avec les facteurs à renommer





Les données numériques seront ainsi **automatiquement remplacées par les noms attribués**.

Faire des opérations mathématiques sur R

Il est possible d'effectuer des opérations mathématiques sur le programme R. Pour cela, il suffit **d'entrer l'opération dans la console Rcmdr**. Une fois l'opération entrée, il faut la **surligner** et cliquer sur le bouton **Soumettre** de la console.

Notation en R	**Valeur mathématique**
a + b	Somme de a et b
a – b	b soustrait de a
a * b	Produit de a par b
a / b	Division réelle de a par b
a %/% b	Division euclidienne (entière) de a par b
a %% b	a mod b
a^b	a à la puissance b
sqrt (a)	Racine carrée de a
abs (a)	Valeur absolue de a
log (a)	Logarithme naturel de a
exp (a)	Exponentielle de a
sin (a)	Sinus de a (en radians)
cos (a)	Cosinus de a (en radians)
tan (a)	Tangente de a (en radians)
mean (a)	Moyenne de a

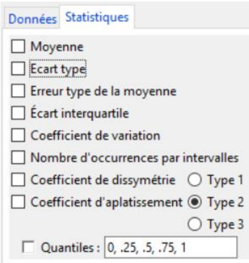
Résumer numériquement les scores

Chemin d'accès : Statistiques/Résumés/Jeu de données actif

```
> summary(Dataset)
  i..Score...
 Min.   :36.00
 1st Qu.:38.00
 Median :41.00
 Mean   :41.61
 3rd Qu.:44.75
 Max.   :49.00
```

Obtenir des données descriptives

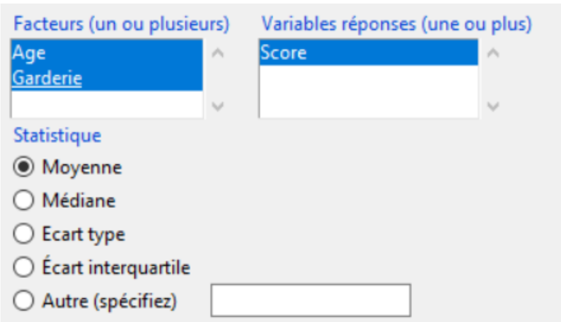
Chemin d'accès : Statistiques / Résumés / Statistiques descriptives



Dans l'onglet **Données**, sélectionner la variable dont on veut extraire des données descriptives et dans l'onglet **Statistiques**, cocher les valeurs dont on a besoin.

Obtenir des données descriptives à partir d'un tableau de données 2x2

Chemin d'accès : Statistiques / Résumés / Tableau de statistiques



Dans la fenêtre, sélectionner les **Facteurs** et les **variables réponses** dont on veut obtenir des données, puis sélectionner le **type de données voulues**.

Obtenir la valeur d'un p unilatéral à partir d'un p bilatéral

Pour nombre de test sur R, ceux-ci se font automatiquement sous forme **bilatérale**. Dans certains cas, il va falloir les **transformer en p-valeur d'un test unilatéral**. Cela se fait de la manière suivante :

Test bilatéral : $H_0: \mu = 0$ $H_1: \mu \neq 0$

$t(42) = 2.694$, $p = 0.010$

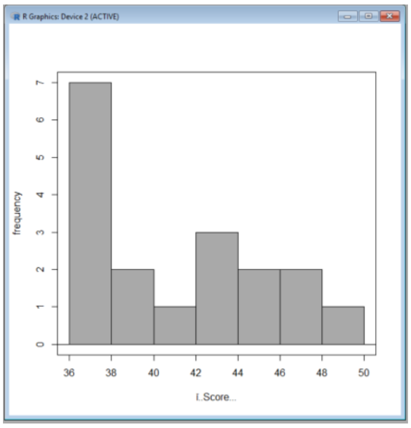
Test unilatéral : $H_0: \mu = 0$ $H_1: \mu > 0$ ➡ $t(42) = 2.694$, $p = {}^{p}/_{2} = {}^{0.010}/_{2} = 0.005$

$H_0: \mu = 0$ $H_1: \mu < 0$ ➡ $t(42) = 2.694$, $p = 1 - {}^{p}/_{2} = 1 - {}^{0.010}/_{2} = 0.995$

Représenter graphiquement les données

Histogramme

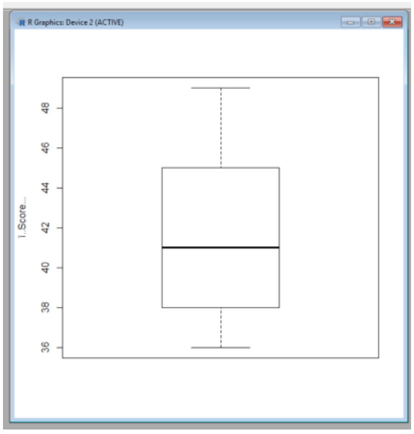
Chemin d'accès : Graphes / Histogramme



Ne pas oublier de nommer les axes.

Boîte de dispersion (boîte à moustache)

Chemin d'accès : Graphes / Boîte de dispersion

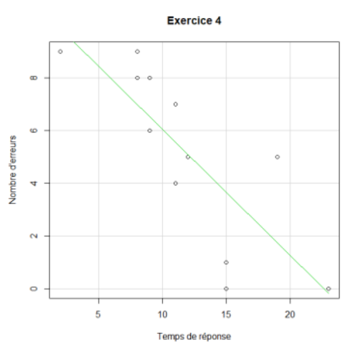


Ne pas oublier de nommer les axes.

Pour obtenir **plusieurs boîtes à moustache**, dans la fenêtre du graphique, cliquer sur la variable puis sur le bouton **Graphe par :** et sélectionner la variable de regroupement.

Nuage de points (diagramme de dispersion)

Chemin d'accès : Graphes / Nuage de points



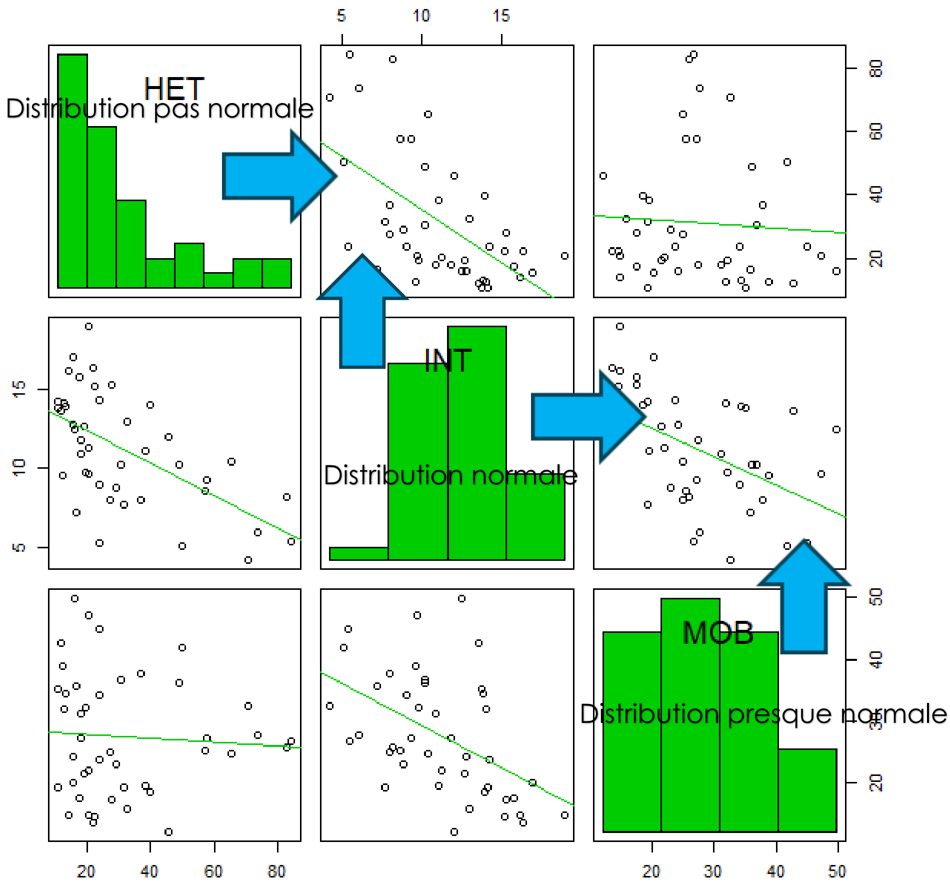
Dans l'onglet **Données**, sélectionner la **variable prédictive (X)** et la **variable réponse (Y)**. Dans l'onglet **Options**, nommer **l'axe des X, l'axe des Y, donner un titre au graphique** et cliquer sur **Ligne des moindres carrés** pour faire apparaître la droite de régression.

La ligne des moindres carrés indique le **sens de la corrélation**.

Matrice de nuage de points

Chemin d'accès : Graphes / Matrice de nuage de points

Dans l'onglet **Données**, sélectionner les variables que l'on veut observer. Dans l'onglet **Options**, cocher **Histogrammes** et **Lignes des moindres carrés** pour faire apparaître la droite de régression.



Interprétation

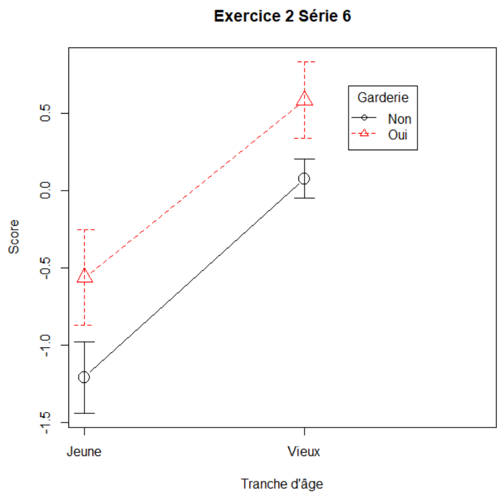
1) **Corrélation négative entre HET et INT** : plus l'hétérogénéité est élevée, moins il y a d'intégration
2) **Corrélation nulle entre HET et MOB** : l'hétérogénéité n'a pas d'effet sur la mobilité, et vice versa
3) **Corrélation négative entre INT et MOB** : plus il y a de mobilité, moins il y a d'intégration.

Diagramme d'interaction

Chemin d'accès : Graphes / Graphe des moyennes

Dans la fenêtre, sélectionner les **Facteurs** et la **variable réponse**.

Sous l'onglet **Options**, sélectionner les données que l'on veut voir apparaître, et **nommer** les différents axes.



Interprétation du graphique :

- **Effet d'âge** : dans les deux conditions, les vieux ont toujours des scores plus élevés que les jeunes
- **Effet de garderie** : les enfants ayant été en garderie, qu'importe qu'ils soient jeunes ou vieux, ont de meilleurs résultats que les enfants n'étant pas allés en garderie
- **Pas d'effet d'interaction** : comme les lignes sont parallèles, il n'y a pas d'effet d'interaction.

Série 1 : Tests non-paramétriques

La p-valeur est la valeur utilisée dans les tests d'hypothèses, tests qui permettent ou non de rejeter l'hypothèse nulle (H_0). Elle représente la probabilité de faire une erreur de premier-type, ou de rejeter l'hypothèse nulle si elle est vraie.

Plus la valeur de p est petite, plus la probabilité de faire une erreur en rejetant l'hypothèse nulle est faible. Une valeur limite de 0.05 est souvent utilisée (seuil α).

Autrement dit : **on rejette l'hypothèse nulle si la p-valeur est inférieure à 0.05**

⚠ *Si la case du test statistiques que l'on doit effectuer est grisée, cela signifie qu'il y a un problème avec la façon dont les données ont été entrées sur Calques.*

Test du Khi-Carré (Série 1, exercice 2)

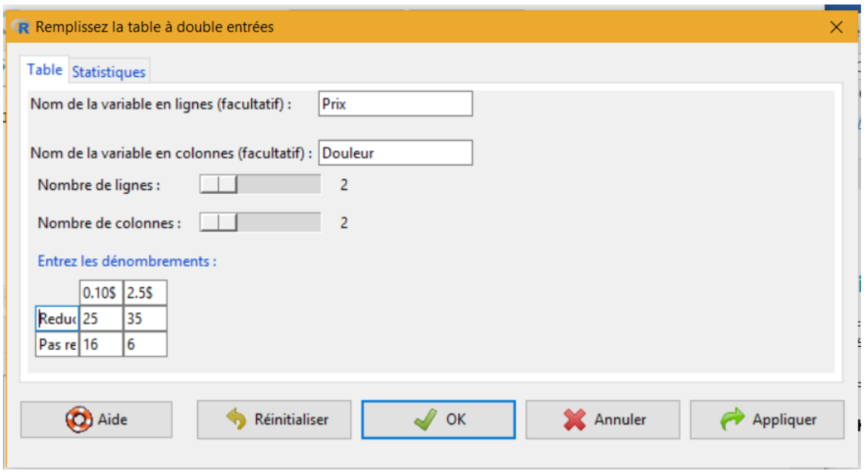
Chemin d'accès : Statistiques / Tables de contingences / Remplir et analyser un tir croisé

Avant de commencer : Poser les hypothèses nulles et les hypothèses alternatives et définir α

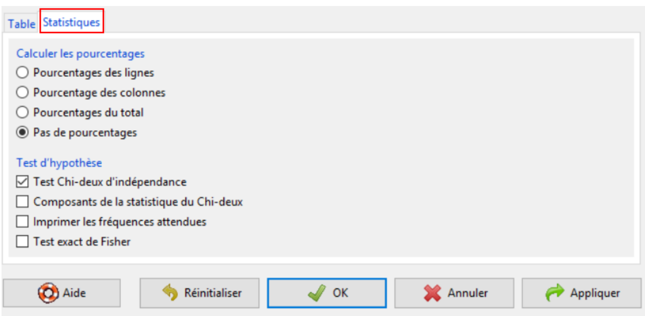
H_0 : $\pi_1 = \pi_2$ α = 5% (soit 0.05)

H_1 : $\pi_1 \neq \pi_2$

1. Remplir le tableau avec les données.



2. Pour afficher les pourcentages, il faut choisir l'onglet **Statistique** dans la fenêtre de la table à double entrée, puis cocher le pourcentage souhaité, ainsi que le type de test souhaité (ici, un test **Chi-deux d'indépendance**)



3. On obtient les résultats sous la forme suivante :

```
         Douleur
Prix       0.10$ 2.5$
  Reduc       25   35
  Pas reduc   16    6

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

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 6.2121, df = 1, p-value = 0.01269
```

```
              Prix
Douleur     0.10$  2.5$
  Reduc        61  85.4
  Pas reduc    39  14.6
  Total       100 100.0
  Count        41  41.0

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

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 6.2121, df = 1, p-value = 0.01269
```

Avec pourcentage

4. La notation en **norme APA** se fait ainsi :

a. χ^2 (N = total des observations**,** nbre colonne **–** 1) = X-squared ;
p-valeur = .xxx
Exemple : χ^2 (82, 1) = 6.212, p = .013

5. **Conclusion** : on rejette l'hypothèse nulle **si la p-valeur est plus petite que le seuil α.**

Test de Wilcoxon apparié *(Exercice 3)*

Chemin d'accès : Statistiques / Tests non paramétriques / Test de Wilcoxon apparié

1 échantillon ; 2 mesures

Avant de commencer : Poser les hypothèses nulles et les hypothèses alternatives et définir α
H$_0$: $\mu_1 = \mu_2$
H$_1$: $\mu_1 < \mu_2$ (unilatéral gauche) ; $\mu_1 > \mu_2$ (unilatéral droite) ; $\mu_1 \neq \mu_2$ (bilatéral)
α = 5%

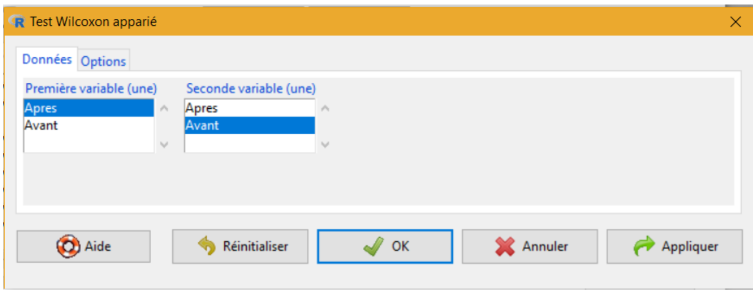
⚠ Si on fait un test $\mu_1 > \mu_2$, dans la fenêtre du test, on doit comparer Var1 à Var2. Dans ce cas, on sélectionne $\mu < 0$

1. Entrer les données sur **Calque**, de la manière suivante :

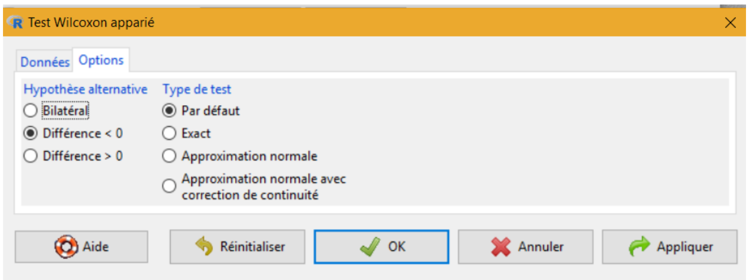
	A	B	C
1	**Var 1**	**Var 2**	
2	Avant	Apres	
3	25	21	
4	28	23	
5	29	22	
6	26	21	
7	28	26	
8	27	29	
9	22	21	
10	25	22	
11	27	23	
12	28	22	
13	29	25	
14			

2. Importer les données dans le programme R

3. Sélectionner la **première variable** et la **deuxième variable**, puis sous l'onglet **Options**,



4. Sélectionner le **type d'hypothèse alternative** (unilatéral gauche/droite ou bilatéral)



5. On obtient les résultats sous la forme suivante :

```
        Wilcoxon signed rank test with continuity correction

data:  Apres and Avant
V = 2.5, p-value = 0.003733
alternative hypothesis: true location shift is less than 0
```

6. La notation aux **normes APA** se fait ainsi :
   a. V = 2.5 ; P-valeur = .004
7. **Conclusion** : on rejette l'hypothèse nulle **si la p-valeur est plus petite que le seuil α.**

Test de Wilcoxon bivarié ou Test de Mann-Witney *(Exercice 4)*

Chemin d'accès : Statistiques / Tests non paramétriques / Test de Wilcoxon

2 échantillons ; 1 mesure

Avant de commencer : Poser les hypothèses nulles et les hypothèses alternatives et définir α
H_0 : $\mu_1 = \mu_2$
H_1 : $\mu_1 > \mu_2$ (unilatéral gauche) ; $\mu_1 < \mu_2$ (unilatéral droite) ; $\mu_1 \neq \mu_2$ (bilatéral)
α = 5%

⚠ Si on fait un test $\mu_1 > \mu_2$, dans la fenêtre du test, on doit comparer Var1 à Var2. Dans ce cas, on sélectionne $\mu < 0$

1. Entrer les données sur **Calques**, de la manière suivante :

	A	B
1	**Notes**	**Groupe**
2	73	Matin
3	87	Matin
4	79	Matin
5	75	Matin
6	82	Matin
7	66	Matin
8	95	Matin
9	75	Matin
10	70	Matin
11	86	Apresmidi

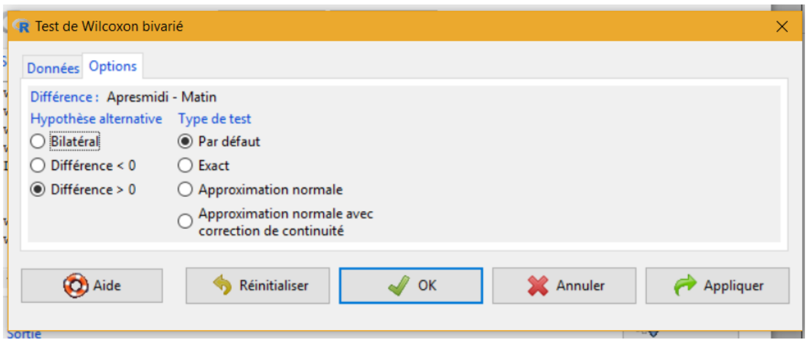
⚠ *Le test se fait toujours ainsi : Lettre plus proche de Y – Lettre plus proche de A*

Il faut donc faire attention au libellé des groupes

Si on obtient la bonne p-valeur, mais que le W n'est pas le même, c'est dû à un « problème » de libellé des groupes.

2. Importer les données dans R

3. Dans la fenêtre du test, sous l'onglet **Options**, sélectionner le type d'hypothèse alternative



4. Les données apparaissent sous la forme suivante :

```
> with(Dataset, tapply(Notes, Groupe, median, na.rm=TRUE))
Apresmidi     Matin
     84.5      75.0

> wilcox.test(Notes ~ Groupe, alternative="greater", data=Dataset)

        Wilcoxon rank sum test with continuity correction

data:  Notes by Groupe
W = 80, p-value = 0.03475
alternative hypothesis: true location shift is greater than 0
```

5. La notation aux **normes APA** se fait ainsi :
   a. W = 80, p = .035

6. **Conclusion** : on rejette l'hypothèse nulle **si la p-valeur est plus petite que le seuil α.**

Série 2 : Tests paramétriques

Le test de normalité permet de montrer sur la **distribution des données** suit une loi normale ou non. Si l'on souhaite utiliser un test paramétrique, **il est nécessaire que les distributions des données suivent une loi normale**, sans quoi il faudra recourir à un test non-paramétrique.

Tester la normalité d'une distribution

Chemin d'accès : Statistiques / Résumés / Test de normalité

On utilise le test de **Shapiro-Wilk**

Hypothèses
$H_0: F(x) = F_0(x)$
$H_0: F(x) \neq F_0(x)$

On rejette l'hypothèse nulle si p < 0.05. **La distribution suit une loi normale si on obtient une p-valeur supérieure à 0.05**

Les résultats se notent au **format APA** de la façon suivante :
W = 0.986 , p = 0.989

Test de Student sur 1 échantillon (t-test univarié)

Chemin d'accès : Statistiques / Moyenne / T-test univarié

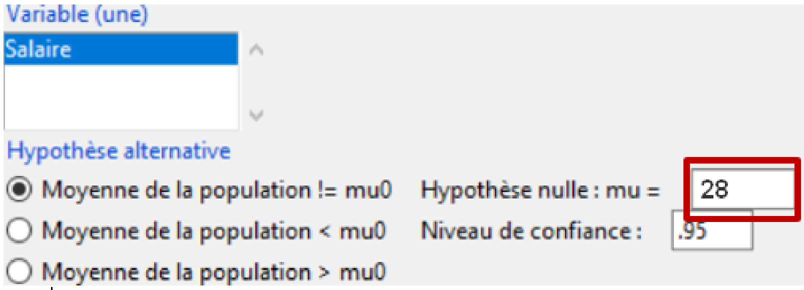
1 échantillon ; 1 mesure

Avant de commencer : poser les hypothèses et déterminer le seuil
H0 = $\tilde{\mu}_1 = \tilde{\mu}_2$
H1 = $\tilde{\mu}_1 \neq \tilde{\mu}_2$ (bilatérale) $\tilde{\mu}_1 > \tilde{\mu}_2$ (unilatéral gauche) $\tilde{\mu}_1 < \tilde{\mu}_2$ (unilatéral droite)
α = 5%

1. Entrer les données sur **Calque** de la manière suivante :

Salaire
24
27
31
21
19
26
28
22
15
25

2. Importer les données dans R
3. Tester la **normalité** de la distribution **(p.14)**
4. Dans la fenêtre du T-test, entrer la **moyenne** de l'hypothèse nulle



5. Les données apparaissent de la façon suivante :

```
        One Sample t-test

data:  Salaire
t = -2.8345, df = 9, p-value = 0.01958
alternative hypothesis: true mean is not equal to 28
95 percent confidence interval:
 20.44807 27.15193
sample estimates:
mean of x
     23.8
```

6. La notation en **norme APA** se fait ainsi :
t(9) = **-2.835** , p = **0.02**

7. **Conclusion :** on rejette l'hypothèse nulle **si la p-valeur est plus petite que le seuil** α**.**

Test de Student à mesures répétées

Chemin d'accès : Statistiques / Moyenne / T-Test apparié

1 échantillon, 2 mesures

Avant de commencer : poser les hypothèses et déterminer le seuil
H₀ = $\tilde{\mu}_1 = \tilde{\mu}_2$
H₁ = $\tilde{\mu}_1 \neq \tilde{\mu}_2$ (bilatérale) $\tilde{\mu}_1 > \tilde{\mu}_2$ (unilatéral gauche) $\tilde{\mu}_1 < \tilde{\mu}_2$ (unilatéral droite)
α = 5%

1. Entrer les données sur **Calque** de la manière suivante :

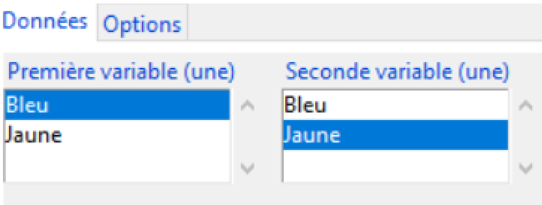
Bleu	Jaune
17	24
23	18
18	15
21	25
25	21
18	23
14	19
22	20

2. Importer les données dans R

3. Tester la **normalité des deux distributions (p.13)**

4. Effectuer le **T-Test apparié**

5. Dans la fenêtre du test, sélectionner la **variable 1 et la variable 2** ; sous **Options** choisir le type de test (unilatéral, bilatéral)



6. Les résultats apparaissent de la façon suivante :

```
        Paired t-test

data:  Bleu and Jaune
t = 0.73988, df = 15, p-value = 0.4708
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.998368  4.123368
sample estimates:
mean of the differences
                 1.0625
```

7. La notation en norme APA se fait ainsi :
t(9) = 0.74, p = 0.471

8. **Conclusion** : on rejette l'hypothèse nulle **si la p-valeur est plus petite que le seuil α.**

Test de Student/Fisher sur 2 groupes indépendants

Chemin d'accès : Statistiques / Moyennes / T-test indépendant

2 échantillons ; 1 mesure

1. Entrer les données sur **Calques** de la manière suivante :

Score	Age
24	Jeune
25	Jeune
17	Jeune
13	Jeune
24	Jeune
0	Jeune
12	Vieux
...	...
12	Vieux
15	Vieux
6	Vieux
3	Vieux
3	Vieux
16	Vieux

2. Importer les données dans R

3. **Tester la normalité des distributions**. Pour pouvoir le faire, il va falloir créer des **sous-ensembles**. On teste la normalité des deux groupes séparément. Il faut créer un premier sous-ensemble, faire le test de normalité, puis créer un deuxième sous-ensemble et tester à nouveau la normalité

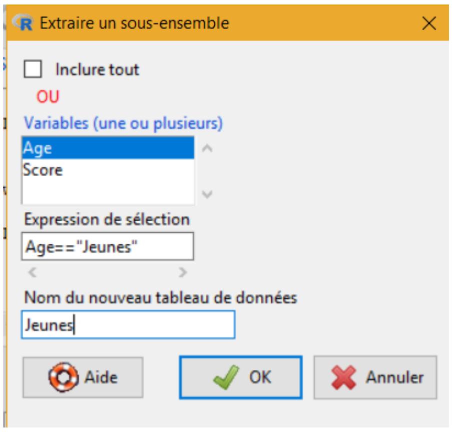
Créer des sous-ensembles

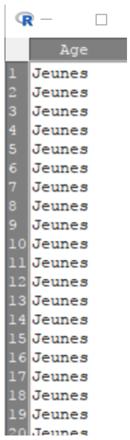
Chemin d'accès : Données / Jeu de données actif / Sous-ensemble

a. Dans la fenêtre des sous-ensembles, sélectionner la **variable** et indiquer **l'expression de la sélection**, c'est-à-dire la partie de l'échantillon que l'on va extraire dans notre sous-ensemble.

⚠ *Il faut absolument respecter les libellés du document des données, sinon ça ne marche pas. Il ne faut pas oublier les " "*

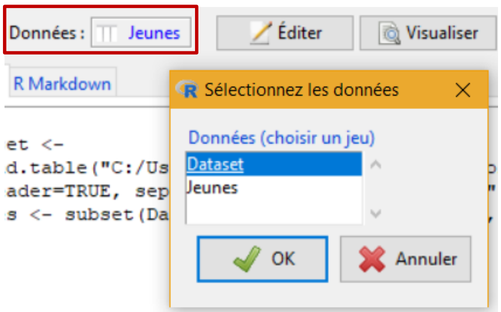
Il est préférable de **renommer le tableau** des données pour ne pas remplacer le tableau précédent.





Voilà notre sous-ensemble

b. Pour retourner au tableau de donnée de départ, il faut cliquer sur **Données** :, puis sélectionner le tableau désiré.

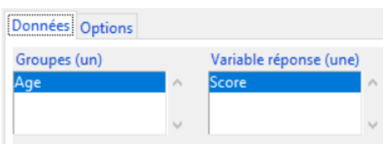


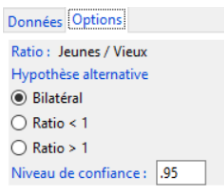
4. Avant d'utiliser un test paramétrique, il faut **déterminer si les variances sont homogènes ou non**. Pour cela on utilise un test F de deux variances. Cela se fait de la manière suivante.

Test F de deux variances

Chemin d'accès : Statistiques / Variances / Test F de deux variances

a. Dans la fenêtre du test, sélectionner les variables. Sous l'onglet **Options**, sélectionner le type de test (unilatéral, bilatéral)





b. Les résultats apparaissent de la manière suivante :

```
        F test to compare two variances

data:  Score by Age
F = 3.1156, num df = 19, denom df = 15, p-value = 0.02987
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.123522 8.153807
sample estimates:
ratio of variances
          3.115568
```

c. La notation en **normes APA** se fait ainsi :
$F(19,15) = 3.116$, $p = 0.030$

d. On rejette notre hypothèse nulle si la p-valeur est plus petite que 0.05. Lorsque l'on **rejette l'hypothèse nulle, cela signifie que les variances sont différentes**.

5. Une fois le test de variance effectué, on peut effectuer notre **t-test indépendant**.

Définir les hypothèses et le seuil
H_0 : $\tilde{\mu}_1 = \tilde{\mu}_2$
H_1 : $\tilde{\mu}_1 \neq \tilde{\mu}_2$(Bilatérale) H_1 : $\tilde{\mu}_1 > \tilde{\mu}_2$(unilatéral gauche) H_1 : $\tilde{\mu}_1 < \tilde{\mu}_2$(unilatéral droite)
α = 5%

6. Dans l'onglet **Options**, sélectionner le **type d'hypothèses** et indiquer si les **variances sont égales ou non**
7. Les résultats apparaissent sous la forme suivante :

```
        Welch Two Sample t-test

data:  Score by Age
t = 3.5032, df = 30.986, p-value = 0.001421
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  3.217059 12.182941
sample estimates:
mean in group Jeunes  mean in group Vieux
                18.7                 11.0
```

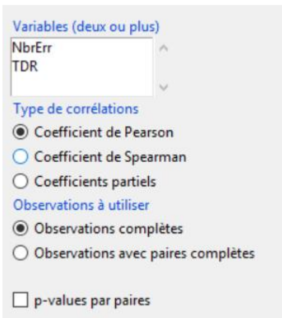
8. La notation en **normes APA** se fait ainsi :
$t(31) = 3.503$, $p = 0.001$

9. Conclusion : on rejette l'hypothèse nulle si la **p-valeur est plus petite que le seuil α**

Coefficient de corrélation de Bravais-Pearson *(Coefficient de corrélation linéaire)*

Chemin d'accès : Statistiques / Résumés / Matrice de corrélations

1. Dans la fenêtre du test, sélectionner les **variables** et le **type de coefficient** souhaité. Ici, on choisira **Coefficient de Pearson**



2. Les données apparaissent de la manière suivante :

```
          NbrErr        TDR
NbrErr  1.0000000 -0.8001774
TDR    -0.8001774  1.0000000
```

3. Le coefficient de corrélation se note ainsi :
$r_{xy} = -0.8$

Le coefficient de corrélation permet de montrer le **lien entre deux variables**. Le coefficient de corrélation peut prendre une valeur comprise entre –1 et 1. **Plus la valeur est proche de 0, plus la corrélation est faible**.
Un coefficient de corrélation **négatif** indique une relation **inversément proportionnelle** : plus la variable X est élevée, plus la variable Y est basse.
Un coefficient de corrélation **positif** indique une **relation proportionnelle** : plus la variable X est élevée, plus la variable Y est élevée également.
Un coefficient de **0** indique **une absence de corrélation**. Les deux variables sont **indépendantes.**

Test de corrélation

Chemin d'accès : Statistiques / Résumés / Test de corrélation

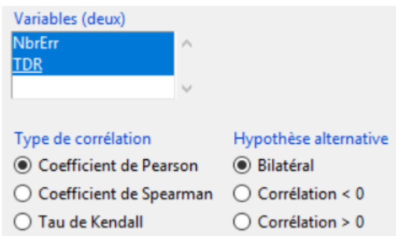
Avant de commencer : poser les hypothèses et déterminer le seuil
H₀ : $\rho_{xy} = 0$
H₁ : $\rho_{xy} \neq 0$ (bilatérale) $\rho_{xy} > 0$(unilatéral gauche) $\rho_{xy} < 0$ (unilatéral droite)
α = 5%

1. Entrer les données dans **Calques** de la manière suivante :

TDR	NbrErr
23	0
8	8
15	1
9	8
9	6
11	7
11	4
19	5
12	5
8	9
15	0
2	9

2. Importer les données dans R

3. Sélectionner les **variables**, le **type de coefficient de corrélation** et le **type d'hypothèse**.



4. Les données apparaissent de la manière suivante :

```
                Pearson's product-moment correlation

data:  NbrErr and TDR
t = -4.219, df = 10, p-value = 0.001775
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9416511 -0.4184273
sample estimates:
       cor
-0.8001774
```

5. La notation en **normes APA** se fait ainsi :
   t(10) = -4.219, p = 0.002

6. **Conclusion** : on rejette l'hypothèse nulle **si la p-valeur est plus petite que le seuil α.**

Droite de régression

Chemin d'accès : Statistiques / Ajustement des modèles / Régression linéaire

Formules

La **pente de la droite des moindres carrés** (droite de régression) vaut :

$$\hat{\beta}_1 = \frac{sY}{sX}$$

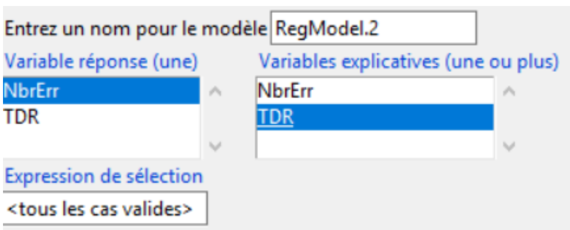
L'ordonnée à l'origine de cette droite vaut :

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X}$$

1. Entrer les données dans **Calques** de la manière suivante :

TDR	NbrErr
23	0
8	8
15	1
9	8
9	6
11	7
11	4
19	5
12	5
8	9
15	0
2	9

2. Dans la fenêtre du test, choisir la **variable réponse (Y)** et **la/les variable(s) explicative(s) (X)**



3. Les résultats apparaissent de la manière suivante :

```
Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  10.8337     1.4727   7.356 0.0000244 ***
TDR          -0.4789     0.1135  -4.219   0.00177 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. L'équation de la droite de régression s'exprime de la manière suivante
$\hat{Y}$= 10.834 – 0.479 X

Avec X = temps de réponse et Y = Nombre d'erreurs

Valeurs prédites et résidus

On peut prédire les valeurs de Y en connaissant les valeurs de X. Pour cela, on peut utiliser la formule suivante :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

On peut aussi faire calculer les valeurs prédites sur R, de la manière suivante :

Chemin d'accès : Modèles / Ajouter les statistiques des observations aux données

Dans la fenêtre d'options, il faut sélectionner **Valeurs ajustées** et **Résidus**.

- [x] Valeurs ajustées
- [x] Résidus
- [ ] Résidus studentisés
- [ ] Valeurs estimées (hat-values)
- [ ] Distances de Cook
- [ ] Indices des observations

Les valeurs apparaissent dans deux nouvelles colonnes sur le tableau de données.

residuals.RegModel.1	fitted.RegModel.1
-24.0233786	126.0234
-0.5935143	138.5935
20.7510235	115.2490
14.7581340	118.2419
-16.0589313	111.0589
10.9979530	135.0020
-15.8120017	130.8120
-6.2703081	106.2703
18.3709330	123.6291
-12.6432881	117.6433
2.7794656	127.2205
7.7439129	112.2561

La colonne **residuals.RegModel1.1** correspond aux **résidus**

La colonne **fitted.RegModel1.1** correspond aux **valeurs prédites**

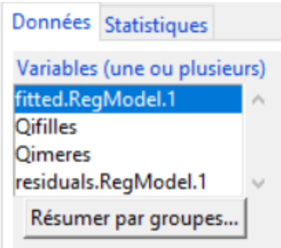
Variance expliquée – Variance Totale – Part de variance expliquée

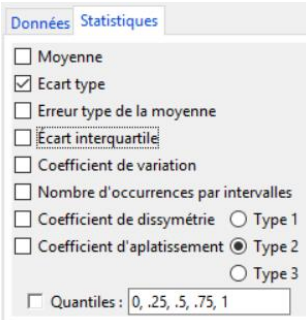
Chemin d'accès : Statistiques / Résumés / Statistiques descriptives

Variance expliquée

La **variance expliquée** fait référence à la **variance des valeurs ajustées**.

Dans l'onglet **Données**, on sélectionne la Variable **Fitted.Regmodel** et dans l'onglet **Statistiques** on coche la case **Ecart-Type**.





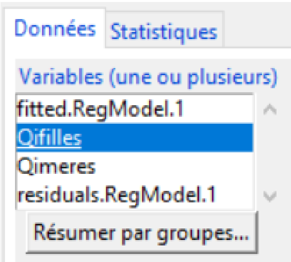
L'écart-type se nomme **sd**.

```
      sd
10.03886
```

La **variance expliquée** n'est rien d'autre que **l'écart-type obtenu élevé au carré (sd²)** donc :
$Var(\hat{Y}) = sd^2$ donc $Var(\hat{Y}) = 10.039^2 = 100.779$

Variance totale

Pour la variance totale, la démarche est la même, sauf que l'on sélectionne notre **variable réponse** (Y) au lieu de sélectionner les valeurs ajustées.



On obtient ainsi un écart-type sd

La variance totale correspond à cet écart-type élevé au carré
$Var(Y) = sd^2$ donc $Var(Y) = 18.009^2 = 324.333$

Part de variance expliquée de variables distinctes

La **part de variance expliquée** se calcule de la manière suivante :
$\frac{Var(\hat{Y})}{Var(Y)}$ donc $\frac{Var(\hat{Y})}{Var(Y)} = \frac{100.779}{324.333} = 0.311$

Part de variance expliquée de variables combinées (examen blanc 225)

Lorsque l'on veut obtenir la part de variance expliquée de de variables combinées, il nous faut créer des modèles linéaires *(ex. modèle 1 et modèle 2)*

Ensuite, on se sert du coefficient de détermination R^2 des deux modèles comparés.

La part de variance expliquée se calcule donc de la manière suivante : on soustrait au R^2 du modèle le plus complet le R^2 du modèle plus parcimonieux.

$$Part\ de\ variance\ expliquée : R_b^2 - R_a^b$$

Exemple de l'examen
Modèle 1 est notre modèle parcimonieux

```
Multiple R-squared:  0.2302
```

Modèle 2 est notre modèle « complet »

```
Multiple R-squared:  0.2354
```

La part de variance expliquée par les valeurs lorsqu'on contrôle le statut socio-économique et les aspirations vaut :

$$0.235 - 0.230 = 0.005 = 0.5\%$$

Série 3

Régression linéaire

Chemin d'accès : Statistiques / Ajustement des modèles / Régression linéaire

Les données apparaissent de la manière suivante :

```
Call:
lm(formula = Sattrav ~ Resp + PerSup + Env + AnnServ, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1441 -1.0830  0.4068  1.1422  2.3796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.66926    2.03155   0.822    0.430
Resp         0.60516    0.42784   1.414    0.188
PerSup      -0.33399    0.53673  -0.622    0.548
Env          0.48552    0.27610   1.758    0.109
AnnServ      0.07023    0.26223   0.268    0.794

Residual standard error: 2.057 on 10 degrees of freedom
Multiple R-squared:  0.4864, Adjusted R-squared:  0.2809
F-statistic: 2.367 on 4 and 10 DF,  p-value: 0.1227
```

Estimations des paramètres du modèle

Formule du modèle linéaire :
$\hat{Y} = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_k \cdot X_k$ où k correspond au **nombre de variables prédictives**

L'estimation des paramètres du modèle se fait à partir des données dans la colonne **Estimate**. Ainsi, notre formue devient :
$Satisfaction = 1.669 + 0.605 \cdot Resp - 0.334 \cdot PerSup + 0.486 \cdot Env + 0.070 \cdot AnnServ$

Interpréter les paramètres du modèle

Ordonnée à l'origine
L'ordonnée à l'origine correspond à β_0. C'est la **valeur théorique** que prend la **variable Y** lorsque les variables β_1 β_2 et β_k prennent la valeur de 0

Premier coefficient de régression partiel
$\beta_1$est le coefficient de régression partiel correspondant à la variable X_1. Lorsque la variable X_1 **augmente d'une unité** et que les variables β_2 et restent **constantes**, la variable Y **augmente ou diminue** (selon si β_1 est positif ou négatif)

Deuxième coefficient de régression partiel
$\beta_2$est le coefficient de régression partiel correspondant à la variable X_2. Lorsque la variable X_2 **augmente d'une unité** et que les variables β_1 et β_k restent **constantes**, la variable Y **augmente ou diminue** (selon si β_2 est positif ou négatif)

Kième coefficient de régression partiel *(k étant le nombre de variables prédictives)*
$\beta_k$est le coefficient de régression partiel correspondant à la variable X_k. Lorsque la variable X_k **augmente d'une unité** et que les variables $\boldsymbol{\beta_1}$ et β_2 restent **constantes**, la variable Y **augmente ou diminue** (selon si β_k est positif ou négatif)

Erreur standard de régression et indice de séparation

L'erreur standard de régression s'écrit $\hat{\sigma}$. On trouve cette valeur dans le tableau de régression linéaire à la ligne suivante :

```
Residual standard error: 2.057 on 10 degrees of freedom
```

Ce qui signifie que si les erreurs se distribuaient normalement, 68% seraient distribuées sur une bande de +/- 2.057

L'indice de séparation équivaut à l'écart-type de Y divisée par l'erreur standard de régression.

$\frac{S_y}{\hat{\sigma}}$ donc $\frac{S_y}{\hat{\sigma}} = \frac{2.426}{2.057} = 1.179$ ce qui équivaut à un nuage relativement diffus.

Coefficient de détermination, coefficient de détermination ajusté et coefficient de corrélation multiple

Le **coefficient de détermination** s'écrit R^2. Le **coefficient de détermination ajusté** s'écrit $R_{2ajust\acute{e}}$
Tous deux se trouvent à la ligne suivante :

```
Multiple R-squared:  0.4864, Adjusted R-squared:  0.2809
```

La valeur du coefficient de détermination correspond au pourcentage de la variance expliquée par les variables prédictives.

Le **coefficient de corrélation multiple** vaut $R = \sqrt{R^2}$
Exemple : $R = \sqrt{R^2} = \sqrt{0.486} = 0.697$

Test global (régression multiple)

Chemin d'accès : Statistiques / Ajustement des modèles / Régression linéaire

Avant de commencer : poser les hypothèses et déterminer le seuil
H_0 = $(\beta_1 = 0) \wedge (\beta_2 = 0) \wedge (\beta_3 = 0) \wedge \ldots \wedge (\beta_k = 0)$
H_1 = $(\beta_1 \neq 0) \vee (\beta_2 \neq 0) \vee (\beta_3 \neq 0) \vee \ldots \vee (\beta_k \neq 0)$
α = 5%

1. Charger les données dans R

2. Sélectionner la **variable réponse (Y)** et **les variables prédictives (X)**

3. Les résultats apparaissent de la manière suivante :

```
Residuals:
   Min     1Q Median     3Q    Max
-5.071 -1.194 -0.206  1.738  4.195

Coefficients:
            Estimate Std. Error t value    Pr(>|t|)
(Intercept) 19.94076    1.19265  16.720     < 2e-16 ***
HET         -0.10856    0.01699  -6.389 0.000000134 ***
MOB         -0.19331    0.03543  -5.456 0.000002739 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.243 on 40 degrees of freedom
Multiple R-squared:  0.6244, Adjusted R-squared:  0.6056
F-statistic: 33.25 on 2 and 40 DF,  p-value: 0.000000003126
```

4. La notation des résultats en **normes APA** se fait de la manière suivante :
F(k, n–k–1) = ..., p = … → ou k = nombre de variables prédictives
et n = le nombre d'observations

Donc F(2,40) = 33.25, p <.001

5. **Conclusion** : on rejette l'hypothèse nulle **si la p-valeur est plus petite que le seuil α.**

Significativité des paramètres (test de régression partiel – test marginal)

A partir du tableau de régression linéaire, on prend les valeurs indiquées dans les **deux dernières colonnes**.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.66926    2.03155   0.822    0.430
Resp         0.60516    0.42784   1.414    0.188
PerSup      -0.33399    0.53673  -0.622    0.548
Env          0.48552    0.27610   1.758    0.109
AnnServ      0.07023    0.26223   0.268    0.794
```

Coefficient de régression partiel associé à RESP : t(40) = 1.414, p = 0.188
Coefficient de régression partiel associé à PerSup : t(40) = -0.622 = 0.548

On considère que les paramètres sont significatifs **si leur p-valeur est plus petite que 0.05**. Généralement, on peut observer une ou plusieurs * à côté des paramètres significatifs

⚠ Le test marginal s'utilise en présence de :

- Variables numériques si elles ne sont pas combinées entre elles
- Variables non-numériques / facteurs si ceux-ci n'ont pas plus de 2 modalités !

  Si les facteurs ont plus de 2 modalité, ou que les variables numériques sont combinées (régression hiérarchique) on utilisera la comparaison de modèles ! *(ex. 1 série 5, ou examen blanc 225)*

Série 4 et 5

Codage des variables

Le codage permet de créer une variable muette à X modalités à partir d'une variable nominale.

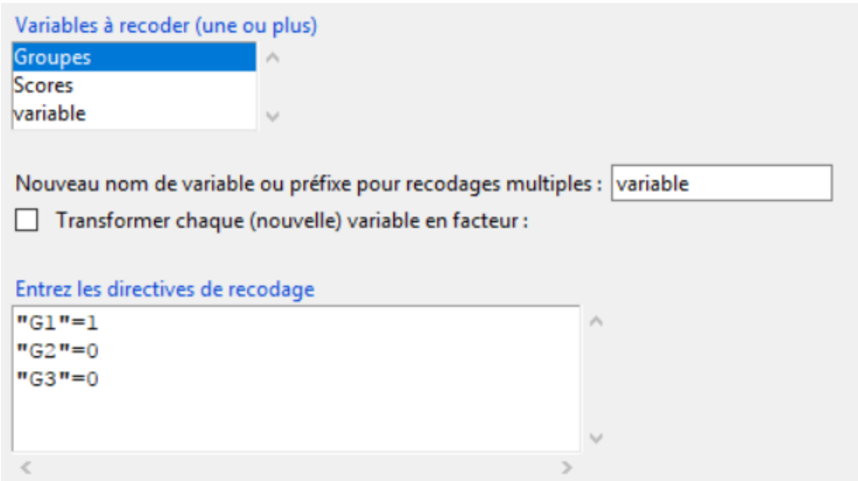
Exemple : on peut recoder une variable nominale sexe en une variable muette avec les modalités Garçon – Filles.

Méthode 1 : codage Manuel

Chemin d'accès : Données / Gérer les variables du jeu de données actif / Recoder les variables

1. Sélectionner la **variable à recoder**

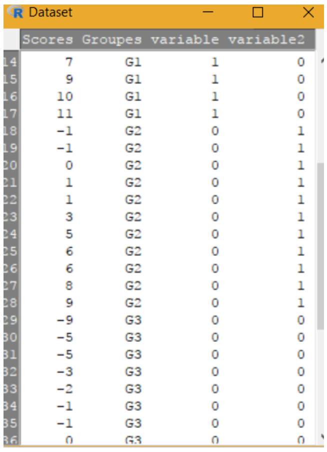
2. Entrer les **directives de recodage**
   a. On donne la valeur de **0 à la modalité de contrôle**
   b. On attribue la valeur de **1 à une des modalités de comparaison**



3. **Décocher** la case « Transformer chaque (nouvelle) variable en facteur » !!!

4. Répéter les 3 premiers points et attribuer la valeur de 1 à chacune des modalités de comparaison l'une après l'autre.

5. En cliquant sur visualiser, on devrait voir les 0 et les 1 apparaître dans le tableau.


Dataset

	Scores	Groupes	variable	variable2
14	7	G1	1	0
15	9	G1	1	0
16	10	G1	1	0
17	11	G1	1	0
18	-1	G2	0	1
19	-1	G2	0	1
20	0	G2	0	1
21	1	G2	0	1
22	1	G2	0	1
23	3	G2	0	1
24	5	G2	0	1
25	6	G2	0	1
26	6	G2	0	1
27	8	G2	0	1
28	9	G2	0	1
29	-9	G3	0	0
30	-5	G3	0	0
31	-5	G3	0	0
32	-3	G3	0	0
33	-2	G3	0	0
34	-1	G3	0	0
35	-1	G3	0	0
36	0	G3	0	0

Méthode 2 : codage Traitement

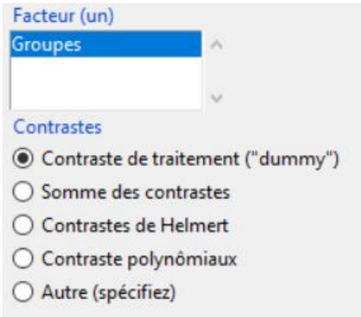
Chemin d'accès : Données / Gérer les variables du jeu de données actif / Réordonner les niveaux d'un facteur

1. Dans la fenêtre, **sélectionner le facteur à réorganiser**. Puis on donne la valeur de **1 à la modalité de contrôle**.

2. Ensuite, aller sous :

Données / Gérer les variables du jeu de données actif / Définir les contrastes d'un facteur

3. Sélectionner le facteur, et cliquer sur **Contraste de traitement**



4. Dans la console, **surligner contrasts(Dataset$Groupes)**, puis cliquer sur **soumettre**

5. Un tableau apparaît. Si notre manipulation est exacte, la **variable contrôle devrait avoir deux fois la valeur de 0**, et les autres variables devraient avoir chacune à leur tour la valeur 1.

```
   [T.G2] [T.G1]
G3      0      0
G2      1      0
G1      0      1
```

Méthode 3 : codage Somme

⚠ Le codage somme est une parenthèse dans l'exercice. Une fois les questions relatives au codage somme faites, il faut revenir à un codage traitement pour répondre au reste !

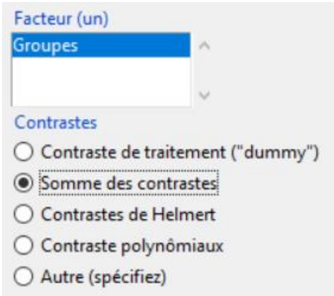
Chemin d'accès : Données / Gérer les variables du jeu de données actif / Réordonner les niveaux d'un facteur

1. Dans la fenêtre, sélectionner le **facteur à réorganiser**. Puis on donne la **valeur de 3** à la **modalité de contrôle**.

2. Ensuite, aller sous :

Données / Gérer les variables du jeu de données actif / Définir les contrastes d'un facteur

3. Sélectionner **le facteur**, et cliquer sur **Somme des contrastes**



4. Dans la console, **surligner contrasts(Dataset$Groupes)**, puis cliquer sur **soumettre**

5. Un tableau apparaît. Si notre manipulation est exacte, la **variable contrôle devrait avoir deux fois la valeur de -1**, et les autres variables devraient avoir chacune à leur tour la valeur 1.

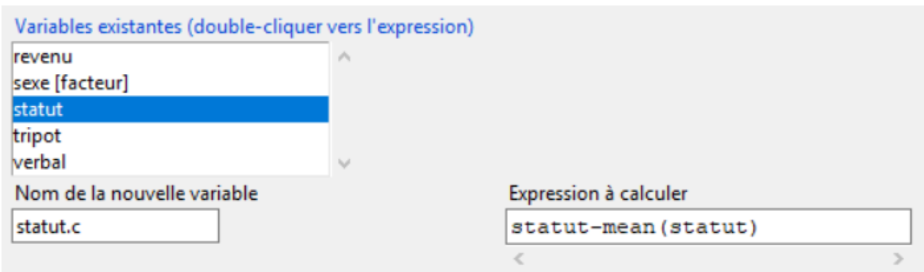
```
   [S.G1] [S.G2]
G1      1      0
G2      0      1
G3     -1     -1
```

Calculer de nouvelles variables

Chemin d'accès : Données / Gérer les variables du jeu de données actif / Calculer une nouvelle variable

1. Sélectionner la **variable à calculer**

2. Entrer **l'expression à calculer**
   a. Pour centrer une variable : entrer la formule $variable - mean(variable)$



3. **Renommer** la variable et valider.

4. En cliquant sur visualiser, les nouvelles variables calculées devraient apparaître.

Créer un modèle linéaire

Chemin d'accès : Statistiques / Ajustement des modèles / Modèle linéaire

Modèle purement additif

Formule du modèle purement additif

Avec une **variable nominale dichotomique** *(ex. sexe = fille/garçon)*

$Y_i = \alpha + \beta_{Xi} + \gamma_{Di} + \varepsilon$ où γ_{Di} correspond à la variable muette

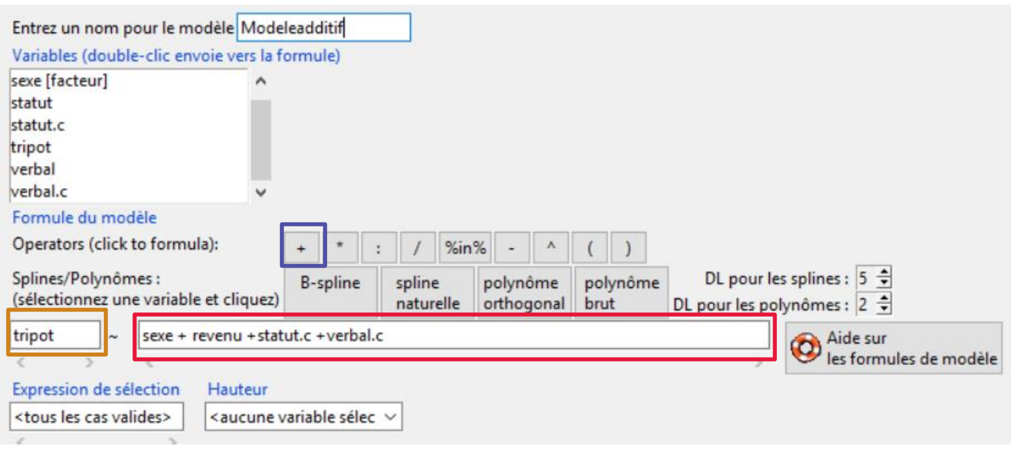
Avec une **variable nominale polytomique**

$Y_i = \alpha + \beta_{Xi} + \gamma_{D1} + \gamma_{D2} + \varepsilon$ qui permet ainsi de générer 3 droites :

D_{1i}	D_{2i}	Equation de la droite
0	0	$Y_i = \alpha + \beta_{Xi} + \gamma_1 \cdot 0 + \gamma_2 \cdot 0 + \varepsilon_i$ Donc $\mathbf{Y_i = \alpha + \beta_{Xi} + \varepsilon_i}$
1	0	$Y_i = \alpha + \beta_{Xi} + \gamma_1 \cdot 1 + \gamma_2 \cdot 0 + \varepsilon_i$ Donc $\mathbf{Y_i = (\alpha + \gamma_1) + \beta_{Xi} + \varepsilon_i}$
0	1	$Y_i = \alpha + \beta_{Xi} + \gamma_1 \cdot 0 + \gamma_2 \cdot 1 + \varepsilon_i$ Donc $\mathbf{Y_i = (\alpha + \gamma_2) + \beta_{Xi} + \varepsilon_i}$

Il s'agit d'un modèle qui comprend **uniquement les variables prédictives**. Il n'y a pas l'intégration des interactions.

1. Dans la fenêtre de **Modèle linéaire**, sélectionner la **variable indépendante (Y)** et les **différentes variables explicatives (X)**. Comme il n'y a pas d'interactions, on n'utilise que le symbole **+**



2. Les données apparaissent ainsi :

```
Residuals:
    Min      1Q  Median      3Q     Max
-51.082 -11.320  -1.451   9.452  94.252

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -16.90890    6.62606  -2.552    0.0144 *
sexe[T.G]    22.11833    8.21111   2.694    0.0101 *
revenu        4.96198    1.02539   4.839 0.0000179 ***
statut.c      0.05223    0.28111   0.186    0.8535
verbal.c     -2.95949    2.17215  -1.362    0.1803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
F-statistic: 11.69 on 4 and 42 DF,  p-value: 0.000001815
```

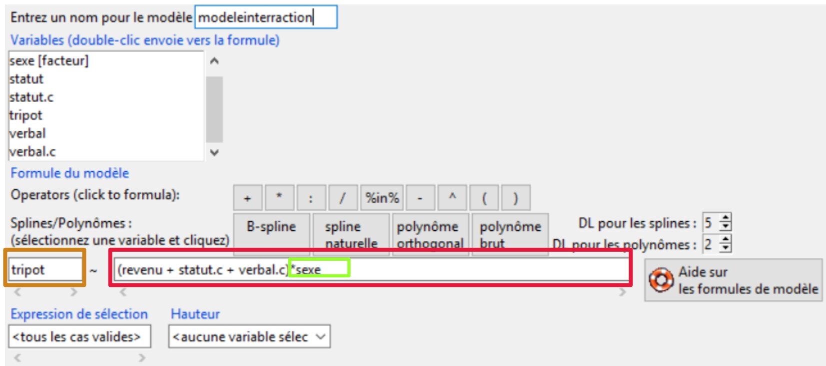
⚠ **On ne peut pas utiliser la régression linéaire car on ne peut pas intégrer de variable muette.**

Modèle avec interactions (Série 5 exercice 1)

Formule du modèle avec interactions
$Y_i = \alpha + \beta_{Xi} + \gamma_{Di} + \delta(D_i \cdot X_i) + \varepsilon_i$

Ce modèle intègre les **interactions** entre les variables et la variable muette. Pour insérer des interactions, on utilise le symbole *

1. Dans la fenêtre, sélectionner la variable indépendante **(Y)**, et les variables explicatives **(X)**. On met les variables explicatives entre parenthèses, puis on ajoute ***variable muette**.



2. Les données apparaissent ainsi :

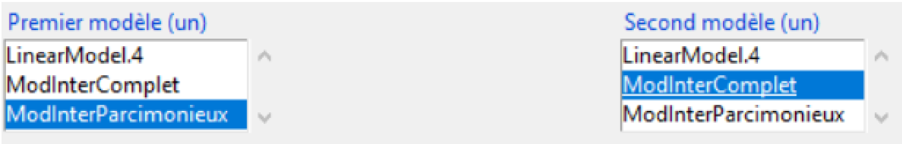
```
Residuals:
    Min      1Q  Median     3Q    Max
-56.654  -7.589  -1.016  3.323 83.903

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.0727     9.2858   0.331   0.7425
statut.c              0.2073     0.4378   0.473   0.6385
revenu                0.6813     2.1833   0.312   0.7567
verbal.c             -0.1392     3.9048  -0.036   0.9717
sexe[T.G]            -1.8324    11.8505  -0.155   0.8779
statut.c:sexe[T.G]   -0.3529     0.5492  -0.643   0.5243
revenu:sexe[T.G]      5.3478     2.4244   2.206   0.0334 *
verbal.c:sexe[T.G]   -2.8355     4.5973  -0.617   0.5410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.98 on 39 degrees of freedom
Multiple R-squared:  0.6243, Adjusted R-squared:  0.5569
F-statistic:  9.26 on 7 and 39 DF,  p-value: 0.00000106
```

Comparaison d'un modèle parcimonieux avec un modèle complet (Série 5 exercice 1)

Chemin d'accès : Modèles / Test d'hypothèse / Comparaison de modèles

1. Il faut avoir créer **deux modèles** **(cf Modèle linéaire p.29)** :
    a. Un modèle **complet** qui contient toutes les variables/interactions
    b. Un modèle parcimonieux, qui ne contient que les variables/interactions statistiquement significatives.

2. Dans la fenêtre, sélectionner le **modèle parcimonieux comme premier modèle** et le **modèle complet comme deuxième modèle**



3. Les données apparaissent ainsi :

```
Analysis of Variance Table

Model 1: tripot ~ revenu * sexe
Model 2: tripot ~ (statut.c + revenu + verbal.c) * sexe
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     43 18930
2     39 17164  4    1766.4 1.0034 0.4175
```

4. Poser les **hypothèses et le seuil**.
   $H_0 = (\beta_1 = 0) \wedge (\beta_3 = 0) \wedge (\delta_1 = 0) \wedge (\delta_3 = 0)$
   $H_1 = (\beta_1 \neq 0) \vee (\beta_3 \neq 0) \vee (\delta_1 \neq 0) \vee (\delta_3 \neq 0)$
   ↳ Notre hypothèse nulle est que le modèle parcimonieux est préférable au modèle complet
   $\alpha = 5\%$

5. Les résultats en **norme APA** se notent de la manière suivante :
   F(**4**,**39**)= **1.003**, p = 0.418

6. Conclusion. Ici, comme p > 0.05, nous ne rejetons pas H0. Le modèle

Test de la significativité des paramètres

Ce test fonctionne selon le même principe qu'une comparaison de modèle. On crée un modèle complet, puis on en crée un autre sans la variable ou l'interaction dont on veut tester la significativité.

Analyse de variance à 1 facteur de classification (Série 5 exercice 2)

Chemin d'accès : Statistiques / Moyennes / ANOVA à un facteur

1. Entrer les données sur Calque et les importer sur le programme R

2. Dans la fenêtre, sélectionner le groupe et la variable réponse, et renommer le modèle

3. Les données apparaissent ainsi :

```
            Df Sum Sq Mean Sq F value Pr(>F)
Groupes      2    4.2   2.077   0.067  0.936
Residuals   46 1434.7  31.188
```

Ce qui correspond à la table de l'ANOVA :

Table de l'ANOVA				
Source de variation	SC	dl	CM	F_{emp}
expliquée	4.153	2	2.077	0.067
résiduelle	1434.663	46	31.188	
totale	1438.816	48		

4. Poser les hypothèses et déterminer le seuil

Hypothèses

$H_0: \mu_1 = \mu_2 = \mu_3$

$H_1: (\exists\ i)(\exists\ j)\ \mu_i \neq \mu_j$

$\alpha = 0.05$

Résultats

P = 0.0936

5. Conclusion. Comme p > 0.05, on ne rejette pas H0 au seuil de 5%. On peut donc conclure qu'il n'y a pas de différence significative entre les trois méthodes d'entraînement.

Calcul des paramètres du modèle à la main (Série 5 exercice 2)

Lorsque l'on fait un test de l'ANOVA à 1 facteur de classification, en plus d'obtenir les résultats de la table de l'ANOVA, nous obtenons d'autres données sous la forme suivante :

```
       mean
G1 2.588235
G2 3.363636
G3 3.000000
```

Où :

$\text{Mean G1} = \overline{Y}_1 = \mu_1$

$\text{Mean G2} = \overline{Y}_2 = \mu_2$

$\text{Mean G3} = \overline{Y}_3 = \mu_3$

C'est à partir de ces données que nous pouvons calculer les paramètres du modèle à la main.

⚠ Respecter l'ordre donné par R pour l'attribution des μ

Calcul des paramètres $\boldsymbol{\alpha}, \boldsymbol{\gamma_1}$ **et** $\boldsymbol{\gamma_2}$

$$\begin{cases} \hat{\alpha} = \overline{Y}_3 = 3.000 \\ \hat{\gamma}_1 = \overline{Y}_1 - \overline{Y}_3 = 2.588 - 3.000 = -0.412 \\ \hat{\gamma}_2 = \overline{Y}_2 - \overline{Y}_3 = 3.364 - 3.000 = 0.364 \end{cases}$$

Calcul des paramètres d'une estimation des scores moyens (codage Traitement)

Pour calculer les paramètres de prédiction, nous avons besoin des paramètres calculé à la page 32. Ces données peuvent aussi être obtenues en estimant les paramètres du modèle (c.f page 23)

```
Call:
lm(formula = Scores ~ Groupes, data = Dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-12.0000  -4.0000  -0.3636   4.0000  13.0000

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.0000     1.2187   2.462   0.0176 *
Groupes[T.G1]  -0.4118     1.8220  -0.226   0.8222
Groupes[T.G2]   0.3636     2.0786   0.175   0.8619
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.585 on 46 degrees of freedom
Multiple R-squared:  0.002887, Adjusted R-squared:  -0.04047
F-statistic: 0.06658 on 2 and 46 DF,  p-value: 0.9357
```

Estimation des scores moyens

$$\begin{cases} \hat{\mu}_1 = \hat{\alpha} + \hat{\gamma}_1 = 3.000 + (-0.412) = 2.588 \\ \hat{\mu}_2 = \hat{\alpha} + \hat{\gamma}_2 = 3.000 + 0.364 = 3.364 \\ \hat{\mu}_3 = \hat{\alpha} = 3.000 \end{cases}$$

Calcul des paramètres d'une estimation des scores moyens (codage Somme)

Il faut tout d'abord calculer les paramètres $\mathbf{\mu}, \mathbf{\alpha_1}$ **et** $\mathbf{\alpha_2}$. Pour ce faire, il nous fait d'abord estimer les paramètres du modèle (cf page 23)

Calcul des paramètres $\mathbf{\mu}, \mathbf{\alpha_1}$ **et** $\mathbf{\alpha_2}$

$$\begin{cases} \hat{\mu} = \bar{Y} = \dfrac{\bar{Y}_1 + \bar{Y}_2 + \bar{Y}_3}{3} = \dfrac{2.588 + 3.364 + 3.000}{3} = 2.984 \\ \hat{\alpha}_1 = \bar{Y}_1 - \bar{Y} = 2.588 - 2.984 = -0.396 \\ \hat{\alpha}_2 = \bar{Y}_2 - \bar{Y} = 3.364 - 2.984 = 0.380 \end{cases}$$

A partir de ces paramètres, il nous est possible d'estimer les scores moyens

Estimation des scores moyens

$$\begin{cases} \hat{\mu}_1 = \hat{\mu} + \hat{\alpha}_1 = 2.984 + (-0.396) = 2.588 \\ \hat{\mu}_2 = \hat{\mu} + \hat{\alpha}_2 = 2.984 + 0.380 = 3.364 \\ \hat{\mu}_3 = \hat{\mu} - \hat{\alpha}_1 - \hat{\alpha}_2 = 2.984 - (-0.396) - 0.380 = 3.000 \end{cases}$$

Ces estimations sont les mêmes que celles obtenues en faisant une estimation à partir d'un codage traitement.

Lorsque l'on fait une estimation à partir d'un codage Somme ou Traitement, nous pouvons constater que :

$$\begin{cases} \hat{\mu}_1 = \overline{Y}_1 \\ \hat{\mu}_2 = \overline{Y}_2 \\ \hat{\mu}_3 = \overline{Y}_3 \end{cases}$$

Série 6

Analyse de variance de type II (série 6 exercice 2)

1. Entrer les données sur Calque de la manière suivante :

Age	Garderie	Score
Jeune	Non	-0.139
Jeune	Non	-2.002
Jeune	Non	-1.631
…	…	…
Jeune	Oui	-1.412
Jeune	Oui	-0.681
Jeune	Oui	0.638
…	…	…
Vieux	Non	-0.167
Vieux	Non	-0.285
Vieux	Non	0.851
…	…	…
Vieux	Oui	0.859
Vieux	Oui	0.782
Vieux	Oui	0.851

2. Il faut commencer par **créer plusieurs modèles** :

Chemin d'accès : Statistiques / Ajustement de modèle / Modèle linéaire

a. Modèle 1 : $Y \sim \alpha_j$ → *y expliquée par variable explicative 1*
b. Modèle 2 : $Y \sim \beta_k$ → *y expliquée par variable explicative 2*
c. Modèle 3 : $Y \sim \alpha_j + \beta_k$ → *y expliquée par interaction de v1 et v2*
d. Modèle 4 : $Y \sim \alpha_j + \beta_k + \gamma_{jk}$ → *y expliquée par v1, v2 et leur interaction*

⚠ **RENOMMER LES MODELES POUR MIEUX S'Y RETROUVER**

3. Remplir la **table d'ANOVA**

ANOVA de type II à pas de fourmi

Table de l'ANOVA

Sources	**SC** ***(Somme des carrés)***	**Ddl (Degré de liberté)**	**CM (Carré Moyen)**	**F_{emp}**
Age	M2 vs M3	p – 1	SC_{age} / (p – 1)	CM_{age} / CM_R
Garderie	M3 vs M1	q – 1	$SC_{garderie}$ / (q–1)	$CM_{garderie}$ / CM_R
Age:Garderie	M3 vs M4	pq	$SC_{a:g}$ / pq	$CM_{a:g}$ / CM_R
Résidus (R)	RSS ligne 2 comparaison M3 vS M4	N – pq	SC_R / (N – pq)	

Où :

- **p** = le nombre de modalités de la variable âge
  - Ex. Age = jeune / vieux = 2
- **q** = le nombre de modalités de la variable garderie
  - Ex. Garderie = oui / non = 2
- **N** = le nombre total d'observations

a. La colonne de **Sommes des carrés** se calcule par **comparaison de deux modèles** (voir comparaison de modèle en page 31)

a. Le SC_R se trouve lors de la comparaison entre le modèle 3 et 4

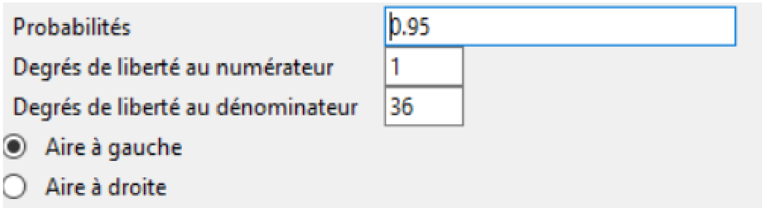
On obtient ainsi le tableau suivant :

Sources	**SC**	**Ddl**	**CM**	$\mathbf{F_{emp}}$	**P**
Age	14.371	1	14.371	24.566	0.0000
Garderie	3.171	1	3.171	5.421	0.0256
Age:Garderie	0.037	1	0.037	0.063	0.8019
R	21.050	36	0.585	$F_{0.95}(1,36)$	

b. Calculer le F_{crit} de la manière suivante :

Chemin d'accès : Distributions / Distributions continues / Distribution F / Quantiles F

Dans la fenêtre, remplir la case **probabilité** (correspond à $1-\alpha$) et les cases de **degré de liberté**, puis sélectionner l'aire voulue.


Probabilités 0.95
Degrés de liberté au numérateur 1
Degrés de liberté au dénominateur 36
Aire à gauche
Aire à droite

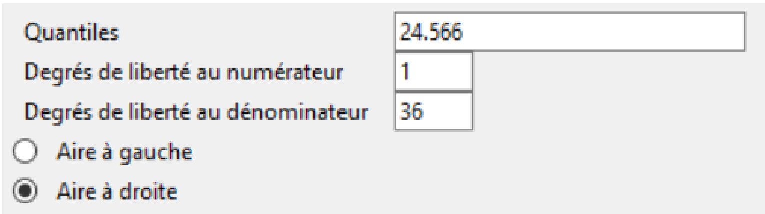
On obtient ainsi un F_{crit} de 4.113

c. Pour remplir la colonne P, on procède de la manière suivante :

Chemin d'accès : distributions / distributions continues / Distributions F / Probabilités F

a. Remplir la case **quantiles**. Celle-ci corresond à la valeur du **Femp** de chaque ligne, puis remplir les **degrés de liberté**.

b. Sélectionner l'**aire**. Ici, on choisira **Aire à droite**.


Quantiles 24.566
Degrés de liberté au numérateur 1
Degrés de liberté au dénominateur 36
Aire à gauche
Aire à droite

c. Répéter la procédure pour chaque ligne du tableau, sauf la ligne R

ANOVA de type II à pas de géant

Chemin d'accès : Modèles / Tests d'hypothèse / Table d'ANOVA

1. Etablir la table **à partir du modèle complet**

2. Dans la fenêtre du test, sélectionner **Moyennes marginales partielles (Type II)**. Il n'est pas nécessaire de toucher au reste.

3. On obtient ainsi notre table de l'ANOVA sous la forme suivante :

```
> Anova(ModeleComplet, type="II")
Anova Table (Type II tests)

Response: Score
              Sum Sq Df F value     Pr(>F)
Age          14.3706  1 24.5771 0.00001711 ***
Garderie      3.1714  1  5.4239    0.02559 *
Age:Garderie  0.0374  1  0.0639    0.80187
Residuals    21.0497 36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interprétation de la table de l'ANOVA

1. On commence par **tester la significativité de l'interaction** des deux variables explicatives.
   a. **Poser les hypothèses et le seuil**
   <u>Hypothèses</u>
   $\mathbf{H_0} = (\forall_j)(\forall_k)\gamma_{jk} = 0$
   $\mathbf{H_0} = (\exists_j)(\exists_k)\gamma_{jk} \neq 0$ } *Interaction*
   α : 5%

   b. **Relever les résultats aux normes APA**
   $F(1,36) = 0.064$, $p = 0.802$,

   c. **Conclusion**. Comme $p > 0.05$, on ne rejette pas H_0. L'effet de l'interaction entre l'âge et l'expérience en garderie sur la prise de rôles est égal à zéro.

2. Etant donné que **l'effet d'interaction est nul**, il nous faut **tester la significativité des effets principaux**. Nous commençons par l'effet de la **première variable explicative**

   a. **Poser les hypothèses et le seuil**
   <u>Hypothèses</u>
   $\mathbf{H_0} = (\forall_j)\alpha_j = 0$
   $\mathbf{H_0} = (\exists_j)\alpha_j \neq 0$ } *Effet principal de la variable 1*
   α : 5%

   b. **Relever les résultats aux normes APA**
   $F(1,36) = 24.578$, **$p < 0.001$**,

   c. **Conclusion**. Comme $p < 0.05$, on ne rejette pas H_0. L'âge a un effet statistiquement significatif sur la prise de rôles. Les enfants « vieux » sont plus aptes à la prise de rôles que les enfants « jeunes »

3. Répéter la procédure avec la **deuxième variable explicative**.
   a. **Poser les hypothèses et le seuil**
   <u>Hypothèses</u>
   $\mathbf{H_0} = (\forall_k)\beta_k = 0$
   $\mathbf{H_0} = (\exists_k)\beta_k \neq 0$ } *Effet principal de la variable 2*
   α : 5%

   b. **Relever les résultats aux normes APA**
   $F(1,36) = \mathbf{5.424}$, $p = 0.026$,

   c. **Conclusion**. Comme $p < 0.05$, on ne rejette pas H_0. L'expérience en garderie a un effet statistiquement significatif sur la prise de rôles. Les enfants ayant fréquenté une garderie sont plus aptes à la prise de rôles que ceux n'y étant pas allés.

Généralement, on peut corroborer ces résultats en représentant les données sous la forme d'un graphique des moyennes.

Calculer la somme des carrés

Pour calculer la somme des carrés, il faut tout d'abord créer les différents modèles.

Ensuite on procède par comparaison de modèles. Nous obtenons, par exemple, les données suivantes :

```
Anova Table (Type II tests)

Response: Score
              Sum Sq Df F value     Pr(>F)
Age          14.3706  1 24.5771 0.00001711 ***
Garderie      3.1714  1  5.4239    0.02559 *
Age:Garderie  0.0374  1  0.0639    0.80187
Residuals    21.0497 36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La somme des carrés se situe dans la colonne Sum Sq

Série 7

Analyse en composantes principales

Pour faire une analyse en composantes principales, il y a un certain nombre d'étapes à faire.

1. Entrer les données sur **Calques**, puis les importer sur R

2. Construire le **diagramme des valeurs propres** :

Chemin d'accès : Statistiques / Analyse multivariée / Analyse en composantes principales

a. Dans la fenêtre, sélectionner les **variables** qui nous intéressent. Sous **options**, cocher **Analyser la matrice de corrélation**, **Graphes des éboulis** et **Ajouter les composantes principales au jeu de données** **(à faire après avoir déterminé quelles composantes nous devons garder)**

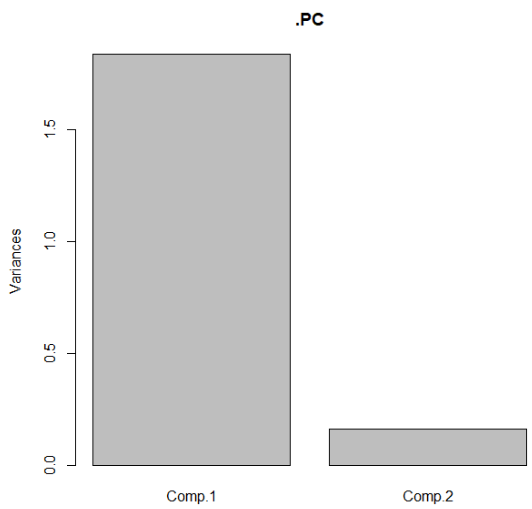
b. Les données nous apparaissent comme suit :

```
Component loadings:
         Comp.1     Comp.2
Test1 0.7071068 -0.7071068
Test2 0.7071068  0.7071068

Component variances:
   Comp.1    Comp.2
1.8364284 0.1635716

Importance of components:
                          Comp.1     Comp.2
Standard deviation     1.3551488 0.40443991
Proportion of Variance 0.9182142 0.08178582
Cumulative Proportion  0.9182142 1.00000000
```

c. Les **valeurs propres** des composantes principales se trouve sous **Component variances**

Comp.1 = **1.837** Comp.2 = **0.164**

3. Le **graphe des éboulis** apparaît sous la forme d'un histogramme

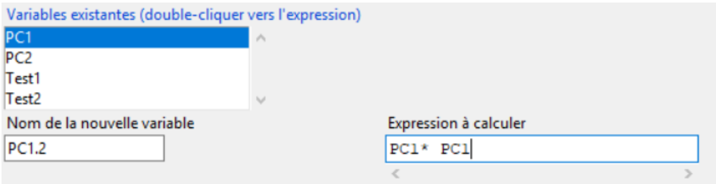


4. Calculer la **moyenne** et la **variance** des composantes principales.

   a. Pour calculer la **moyenne**, il faut aller sous Statistiques/Résumés/Statistiques descriptives, sélectionner les **variables PC1, PC2, PCn**, puis sélectionner **Moyenne**. Normalement, nous devrions avoir une moyenne de **0**.

   Lors d'une ACP, la **variance** correspond à la **moyenne des carrés** des composantes, il nous faut donc **Calculer de nouvelles variables**.

   b. Dans la fenêtre, renommer la nouvelle variable, puis inscrire la formule. On **multiplie PC1 par PC1**



   c. Reproduire la **même chose avec PC2, PCn**, …

   d. Les variances doivent normalement avoir la **même valeur que les valeurs propres** des composantes.

5. Vérifier l'**orthogonalité** des composantes. Pour ce faire, il faut faire une **matrice de corrélation** entre les composantes.

	PC1	PC2
PC1	1	0
PC2	0	1

Les composantes sont orthogonales les unes par rapport aux autres si leur **coefficient de corrélation vaut 0**.

6. Dresser la **matrice de saturation**. Pour ce faire, il faut faire **une matrice de corrélation** entre les composantes et les variables x, y, … Les données apparaissent ensuite sous la forme suivante :

	PC1	PC2	Test1	Test2
PC1	1.000	0.000	0.958	0.958
PC2	0.000	1.000	-0.286	0.286
Test1	0.958	-0.286	1.000	0.836
Test2	0.958	0.286	0.836	1.000

La matrice de saturation correspond aux **coordonnées des variables** dans le plan factoriel. Les coordonnées se notes ainsi :

	$G_1 = PC_1$	$G_2 = PC_2$
x	0.958	-0.286
y	0.958	0.286

7. **Représenter les individus** dans le premier plan factoriel. Il faut faire un nuage de points. Dans la fenêtre, sélectionner PC1 pour l'axe des X et PC2 pour l'axe des Y. Pour faire apparaître les individus en fonctions des groupes, cliquer sur Graphe par groupe.

8. **Représenter les variables** dans le premier plan factoriel. Cette étape se fait à la main. Avec un compas, tracer un **rond de diamètre 1**. Tracer **l'axe des X et l'axe des Y**. Le **point d'intersection** a les coordonnées **(0 ; 0)**. Ensuite, placer les composantes selon leurs coordonnées. Puis tracer une **flèche (vecteur)** depuis le point (0 ; 0) jusqu'à chaque point.

    a. **Interpréter la représentation**. Pour le faire, on observe **l'angle formé par les vecteurs** reliant les points. Plus l'angle est **petit**, plus la **corrélation** entre les composantes sont **fortes**.

Série 8

Critères de sélection des composantes d'une ACP

Critère de Kaiser

Selon ce critère, nous sélectionnons toutes les composantes dont la **valeur propre** est **supérieure ou égale à 1**.

On peut aussi regarder sur le graphe des éboulis et sélectionner toutes les composantes dont la barre dépasse le 1.

Critère de Jolifé

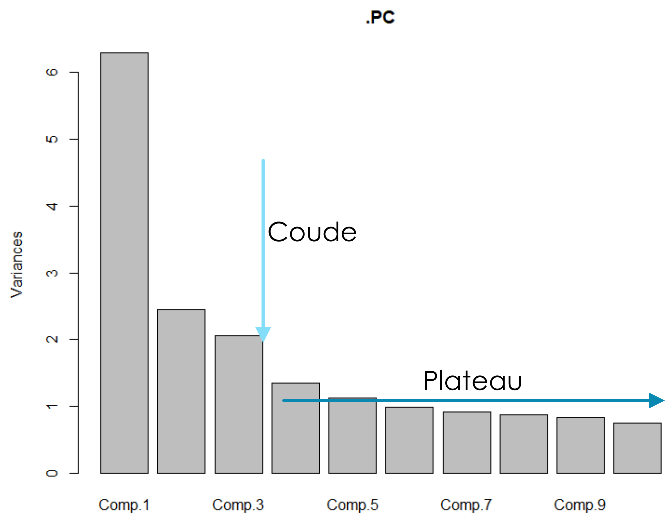
Selon ce critère, nous conservons toutes les composantes dont la **variance cumulée** correspond **au minimum à 80%** (donc 0.80) de la variance totale.

Ce critère est très peu sélectif puisqu'il faut souvent garder un très grand nombre de composantes.

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     2.3094489 1.0366659 0.99874324 0.68828793 0.68472508 0.65805379
Proportion of Variance 0.5333554 0.1074676 0.09974881 0.04737403 0.04688484 0.04330348
Cumulative Proportion  0.5333554 0.6408231 0.74057186 0.78794588 0.83483073 0.87813421
                           Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     0.60742533 0.55845837 0.52701296 0.5099745
Proportion of Variance 0.03689655 0.03118758 0.02777427 0.0260074
Cumulative Proportion  0.91503076 0.94621833 0.97399260 1.0000000
```

Critère de Cattel

Selon le critère de Cattell, lorsque l'on observe le graphe des éboulis, on peut observer un pic, avec une pente, suivie d'un plateau. Selon le critère de Cattell, nous conservons toutes les composantes se situant **au « coude » du graphique**, c'est-à-dire avant le plateau.

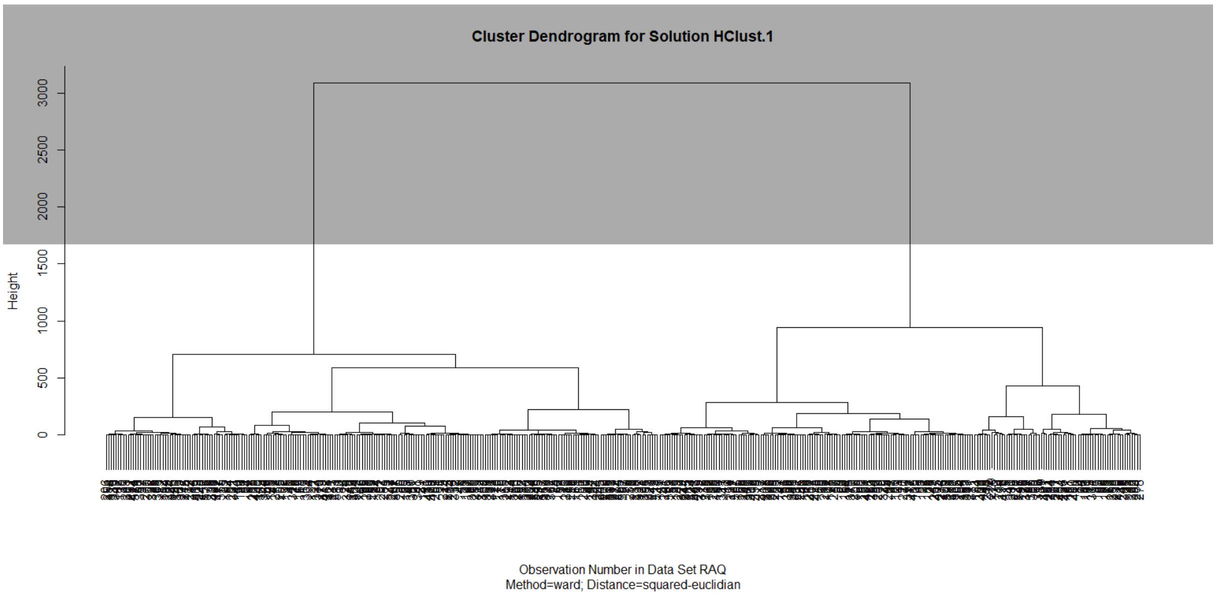


Classification hiérarchique

Chemin d'accès : Statistiques / Analyse multivariée / Classification / Classification hiérarchique

Dans la fenêtre **Données**, sélectionner les **composantes principales**. Sous l'onglet **Options**, sélectionner **Méthode de Ward** pour la méthode de classification et **Euclidienne au carré** pour la mesure de distance.

Ensuite, cocher le **dendrogramme.** On devrait obtenir un graphique ressemblant à ceci :



Classification selon la méthode du saut maximal

1. Recopier le tableau de la consigne

d	a	b	c	d	e
a	0	5	10	9	$\boxed{1}$
b		0	3	99	70
c			0	12	25
d				0	50
e					0

2. Il faut toujours entourer la **plus petite distance**, ici, il s'agit du **1** entre a et e (on ne tient pas compte des 0)

3. On reconstruit le tableau. A la place de a et de e, on crée une ligne et une colonne {a, e}. On ajoute ensuite les 0 entre les mêmes lettres et entre la ligne et la colonne {a, e}

4. Pour la ligne {a, e}, on regarde la distance **entre a et e** et chaque autre lettre, et on **garde la valeur la plus grande**, par rapport à notre tableau complet. Ce qui donne ceci :

    a. <u>Entre {a, e} et b</u> : la distance entre a et b est de 5, la distance entre e et b est de 70. A l'intersection entre {a, e} et b, on va donc mettre 70
    b. <u>Entre {a, e} et c</u> : la distance entre a et c est de 10, la distance entre e et c est de 25. A l'intersection entre {a, e} et c on va donc mettre 25

c. Entre {a, e} et d : la distance entre a et d est de 9, la distance entre e et d est de 50. A l'intersection entre {a, e} et d on va donc mettre 50

	$\{a,e\}$	b	c	d
$\{a,e\}$	0	70	25	50
b		0	$\underline{3}$	99
c			0	12
d				0

5. On crée un troisième tableau en suivant la même méthode.

	$\{a,e\}$	$\{b,c\}$	d
$\{a,e\}$	0	70	$\underline{50}$
$\{b,c\}$		0	99
d			0

6. On crée encore un nouveau tableau, encore plus petit. Cette fois-ci, on ne peut plus aller plus loin. C'est la fin de la méthode. On se retrouve ainsi avec deux groupes distincts.

	$\{a,e,d\}$	$\{b,c\}$
$\{a,e,d\}$	0	$\underline{99}$
$\{b,c\}$		0

Série 9

Analyse parallèle

L'analyse parallèle est un moyen qui permet de **déterminer le nombre de composantes à conserver** lors d'une analyse en composantes principales ou une analyse factorielle.

1. Au lancement de R, dans la **console**, entrer **library(paran)**

2. Charger le jeu de données

3. Dans la console de R commander, entrer la formule suivante :

   Pour une analyse en composantes principales :
   paran(NOM DU FICHIER, graph=TRUE, cfa=FALSE, centile=95)

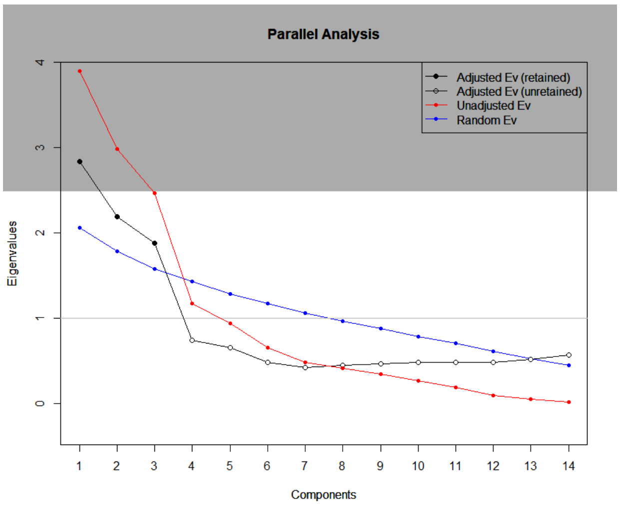
   Pour une analyse factorielle :
   paran(NOM DU FICHIER, graph=TRUE, cfa=TRUE, centile=95)

4. On obtient les résultats sous la forme suivante :

```
Using eigendecomposition of correlation matrix.
Computing: 10%  20%  30%  40%  50%  60%  70%  80%  90%  100%


Results of Horn's Parallel Analysis for component retention
420 iterations, using the 95 centile estimate

--------------------------------------------------
Component   Adjusted    Unadjusted    Estimated
            Eigenvalue  Eigenvalue    Bias
--------------------------------------------------
1           2.826482    3.898824      1.072342
2           2.205700    2.982781      0.777081
3           1.894906    2.467714      0.572808
--------------------------------------------------

Adjusted eigenvalues > 1 indicate dimensions to retain.
(3 components retained)
```

5. Dans ce cas, nous retiendront 3 composantes. Ainsi, lors d'une analyse en composantes principales ou une analyse factorielle, nous conserverons les 3 premières composantes.

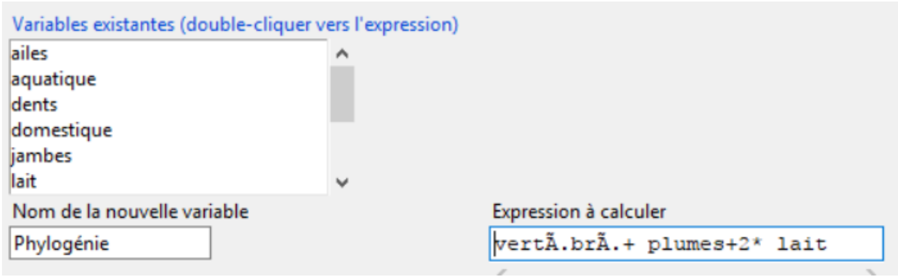
6. On peut aussi se servir du graphique de l'analyse parallèle. On retient un nombre de composantes égal au **nombre de points remplis en noirs** sur la ligne noire.



Calculer une nouvelle variable pour classifier les individus

Chemin d'accès : Données / Gérer les variables du jeu de données actif / Calculer une nouvelle variable

1. Dans la fenêtre, entrer **l'expression à calculer** et **renommer** la nouvelle variable
    a. *Normalement, les modalités d'expressions de la variable sont indiquées dans l'exercice.*

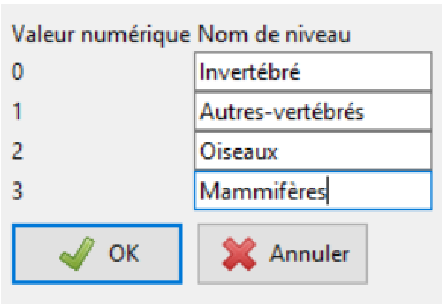


2. Une fois la nouvelle variable calculée, il faut la **transformer en facteur** :

Chemin d'accès : Données / Gérer les variables du jeu de données actif / Convertir des variables numériques en facteurs

3. Sélectionner la variable à transformer en facteur et cocher **Nom des niveaux**. Cliquer sur Ok

4. Une fenêtre va apparaître, dans laquelle il faudra **entrer le nom des niveaux du facteur**. Le 0 correspond à la classe la plus basse et le 3 à la plus haute



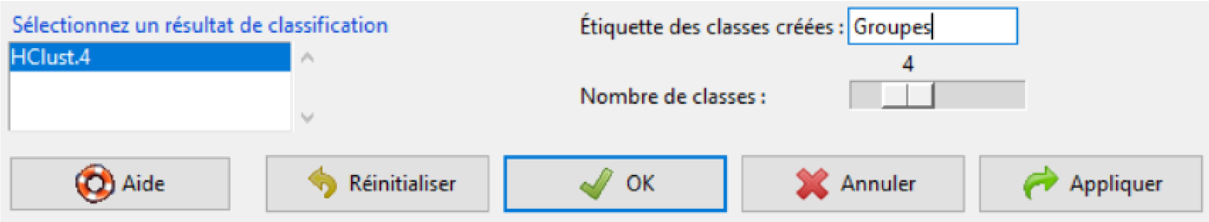
Dresser une table de contingence croisée entre une classification et une variable de classification

1. Il faut tout d'abord effectuer une **analyse en composante principale** (ou analyse factorielle), une **classification hiérarchique de Ward** (voir dossier) ainsi que calculer une variable de classification (vu précédemment)

2. Une fois la classification hiérarchique (et le dendrogramme) faits, il faut **ajouter la classification au jeu de donnée** :

Chemin d'accès : Statistique / Analyse multivariée / Classification / Ajouter les classes au jeu de données

3. **Renommer** la classification, sélectionner la classification à ajouter ainsi que **le nombre de classes**. Le nombre de classe se détermine à partir de l'analyse du dendrogramme



4. Une fois la classification ajoutée au jeu de données, il faut faire une **table de contingence**. Pour ce faire :

Chemin d'accès : Statistique / Tables de contingence / Tir croisé

5. Sous variable en **ligne**, il faut sélectionner la **variable de classification** que l'on a créée, et sous variable en **colonne** sélectionner la **classification hiérarchique**. Sous l'onglet **Statistique**, il faut cocher **Test Chi-deux d'indépendance**.

6. Les données apparaissent comme ceci :

```
Frequency table:
                    Groupes
Phylogénie            1  2  3  4
  Invertébrés         0  0 13  0
  Autres-Vertébrés    0  0  3  6
  Oiseaux             0 14  0  0
  Mammifères         27  0  1  3
```

7. Une fois la table de contingence remplie, on peut effectuer un **test du Khi-carré** pour déterminer si les variables et groupes créés sont dépendants ou non.

8. Les résultats apparaissent en dessous de la table de contingence, de la manière suivante :

```
        Pearson's Chi-squared test

data:  .Table
X-squared = 145.6, df = 9, p-value < 2.2e-16
```

9. Les résultats en **normes APA** s'écrivent de la manière suivante :

$$\chi^2(9, N^* = 67) = \mathbf{145.6}\,; p < .001$$

ou N = le nombre d'individus

10. Si $p < \alpha$, on **rejette l'hypothèse nulle** voulant que les variables et les groupes soient indépendants. Les variables et les groupes sont donc dépendants.

V de Cramer

Le **V de Cramer** est une valeur numérique qui permet de déterminer la **force de la dépendance des variables avec les groupes**. Il se calcule de la manière suivante :

$$V = \sqrt{\frac{\chi^2}{\chi^2_{max}}} = \sqrt{\frac{\chi^2}{N \cdot [min(I^*, J^*) - 1]}}$$

** ou I correspond au nombre de lignes de la table de contingence*
** ou J correspond au nombre de colonnes de la table de contingence*

Plus le V de Cramer est **proche de 1**, plus la dépendance entre les variables et les groupes est **fonctionnelle**.

Classification par k-means *(méthode des nuées dynamiques)*

Chemin d'accès : Statistique / Analyse multivariée / Classification / Classification par k-means

Indice de Calinski et Harabasz

L'indice de **Calinski et Harabasz** permet de déterminer quel modèle est le plus efficace lors d'une classification par k-means.

Plus cet indice est **élevé**, **plus le modèle est adéquat**. Cet indice se calcule de la manière suivante :

$$\mathrm{CH} = \frac{\mathrm{B_q^*} / (\mathrm{q}^* - 1))}{\mathrm{W_q^*} / (\mathrm{n}^* - \mathrm{q})}$$

**Bq correspond à between cluster sum of squares*
**Wq correspond à total within sum of squares*
**q = nombre de classes*
**n = nombre d'individus*

Plus la valeur de ce rapport est **élevée**, plus **les cluster sont cohérent** (donc faible variance à l'intérieur du cluster) et **plus les cluster seront distincts entre eux** (forte variance entre les clusters)

Transformer des items

Lors d'une analyse factorielle, il se peut que des items aillent dans un sens opposé à ce que l'on souhaiterait.
Il nous faut donc effectuer une petite opération pour en changer la direction. Pour ce faire, il nous faut **calculer une nouvelle variable** pour chaque item devant subir une rotation

Chemin d'accès : Données / Gérer les variables du jeu de données actif / Calculer une nouvelle variable

Dans **expression à calculer**, nous entrons la **formule** de l'inversion :

- Dans le cas d'un test de score entre 1 et 5, nous entrons la formule suivante : **6 – item de la rotation**

Renommer la variable, comme ça nous ne perdons pas les données d'origine

Analyse factorielle

Chemin d'accès : Statistique / Analyse multivariée / Analyse factorielle

1. Dans la fenêtre du test, **sélectionner les variables**. Sous **options**, dans **Rotation des facteurs**, cliquer sur **Sans**, car on n'effectue pas de rotation.

2. Une fenêtre va s'ouvrir en demandant **le nombre de facteur à retenir**. Cela se décide selon les critères de Kaiser, Jolifé, Cattell ou suite à une analyse parallèle.

3. Les données apparaissent de la manière suivante :

```
Uniquenesses:
sppa01.r sppa02.r   sppa03 sppa04.r   sppa05
   0.508    0.441    0.398    0.442    0.336

Loadings:
         Factor1
sppa01.r 0.701
sppa02.r 0.748
sppa03   0.776
sppa04.r 0.747
sppa05   0.815

               Factor1
SS loadings      2.875
Proportion Var   0.575

Test of the hypothesis that 1 factor is sufficient.
The chi square statistic is 158.14 on 5 degrees of freedom.
The p-value is 2.47e-32
```

4. Pour déterminer si le modèle est satisfaisant, on effectue un **test du χ^2**. Les informations du test se trouvent dans la console.
   Les résultats se notent de la manière suivante aux **normes APA** :

$$\chi^2(5) = 158.14, p < .001$$

   Si p < α, on rejette l'hypothèse nulle voulant que le modèle soit satisfaisant.

Unicité et communalité

5. L'**unicité** correspond à la **part propre de chaque item**, c'est-à-dire la part du modèle qui n'est pas expliquée par le facteur commun. L'unicité se note σ^2

6. La **communalité** des items correspond à la **part du modèle expliquée par les facteurs communs**.
   Elle se calcule de la manière suivante : β^{2*}
   **se trouve dans la/les colonne(s) Factor, sous Loadings*

   Pour calculer la communalité lorsque nous avons plusieurs facteurs, il suffit **d'additionner les carrés de chaque valeur des facteurs à chaque ligne**.

```
  Factor1 Factor2 Factor3
A  0.683  -0.017   0.022
B -0.030   0.591   0.166
C  0.792  -0.012  -0.143
D  0.162  -0.045  -0.190
E  0.639   0.029  -0.140
F -0.072   0.845  -0.027
G  0.768  -0.052  -0.073
H  0.018   0.825   0.081
I -0.079   0.592   0.701
J  0.645  -0.043  -0.062
K -0.019   0.515   0.103
```

Ex : communalité de A : $0.683^2 + -0.017^2 + 0.022^2 = 0.467$

Matrice de corrélation théorique et matrice de corrélation observée

La **matrice de corrélation théorique se calcule à la main**. Elle prend la forme d'un tableau de X lignes sur X colonnes (X étant le nombre d'items)

	Item 1	**Item 2**	**Item 3**	**Item 4**	**Item 5**	**Item X**
Item 1	$\beta_1^2 + \sigma_1^2 = 1$	$\beta_1^2\beta_2^2$	$\beta_1^2\beta_3^2$	$\beta_1^2\beta_4^2$	$\beta_1^2\beta_5^2$	$\beta_1^2\beta_x^2$
Item 2	$\beta_2^2\beta_1^2$	$\beta_2^2 + \sigma_2^2 = 1$	$\beta_2^2\beta_3^2$	$\beta_2^2\beta_4^2$	$\beta_2^2\beta_5^2$	$\beta_2^2\beta_x^2$
Item 3	$\beta_3^2\beta_1^2$	$\beta_3^2\beta_2^2$	$\beta_3^2 + \sigma_3^2 = 1$	$\beta_3^2\beta_4^2$	$\beta_3^2\beta_5^2$	$\beta_3^2\beta_x^2$
Item 4	$\beta_4^2\beta_1^2$	$\beta_4^2\beta_2^2$	$\beta_4^2\beta_3^2$	$\beta_4^2 + \sigma_4^2 = 1$	$\beta_4^2\beta_5^2$	$\beta_4^2\beta_x^2$
Item 5	$\beta_5^2\beta_1^2$	$\beta_5^2\beta_2^2$	$\beta_5^2\beta_3^2$	$\beta_5^2\beta_4^2$	$\beta_5^2 + \sigma_5^2 = 1$	$\beta_5^2\beta_x^2$
Item X	$\beta_x^2\beta_1^2$	$\beta_x^2\beta_2^2$	$\beta_x^2\beta_3^2$	$\beta_x^2\beta_4^2$	$\beta_x^2\beta_5^2$	$\beta_x^2 + \sigma_x^2 = 1$

On peut ensuite comparer cette matrice de corrélations théorique avec la matrice de corrélation observée. Pour ce faire il faut :

1. **Créer une matrice de corrélation sur R**
2. **Soustraire** la matrice de corrélation théoriques à la matrice de corrélation observée

Ainsi, cette procédure nous permet de montrer quelle(s) corrélation(s) est(sont) moins bien décrite(s) par le modèle.

Série 10

Analyse factorielle

Varimax – Promax

Chemin d'accès : Statistique / Analyse multivariée / Analyse factorielle

1. Dans la fenêtre du test, **sélectionner les variables**. Sous **options**, dans Rotation des facteurs, cliquer sur :

   a. **Varimax** si on effectue une rotation **orthogonale**
   b. **Promax** si on effectue une rotation **oblique**

2. Une fenêtre va s'ouvrir en demandant le **nombre de facteur à retenir**. Cela se décide selon les critères de Kaiser, Jolifé, Cattell ou suite à une analyse parallèle.

3. Les données apparaissent de la manière suivante :

```
Uniquenesses:
    A     B     C     D     E     F     G     H     I     J     K
0.532 0.622 0.352 0.936 0.571 0.281 0.402 0.313 0.151 0.579 0.723

Loadings:
  Factor1 Factor2 Factor3
A  0.683
B          0.591   0.166
C  0.792          -0.143
D  0.162          -0.190
E  0.639          -0.140
F          0.845
G  0.768
H          0.825
I          0.592   0.701
J  0.645
K          0.515   0.103
```

   a. La console ne fait apparaître que les valeurs étant **strictement supérieures à 0.1**. Pour les faire apparaître il faut rajouter : **,cutoff=0.0 dans après .FA**

4. Pour déterminer si le modèle est satisfaisant, on effectue un **test du** χ^2. Les informations du test se trouvent dans la console.
   Les résultats se notent de la manière suivante aux normes APA :

$$\chi^2(25) = 50.04, p < .001$$

5. **Si p < α, on rejette l'hypothèse nulle** voulant que le modèle soit satisfaisant.

Analyse factorielle selon la méthode de régression : Résultat des facteurs

Dans la fenêtre de l'analyse factorielle, **sélectionner les variables**. Sous **Options**, sélectionner le **type de rotation** et sous **Scores des facteurs**, sélectionner **Méthode de régression**.

Les **résultats des facteurs** vont apparaître dans le jeu de donner. Il faut cliquer sur visualiser pour les voir dans les deux dernières colonnes, sous la forme suivante :

F1	F2
1.011655220	-0.152736335
1.465399325	1.320162243
0.349885631	-1.279522305
-1.546517680	-0.013652655
-0.794468863	1.447512646
0.090722360	-0.077124757
-1.561812671	-0.486273404
-0.356389702	-1.079432641
0.577217028	0.940080230
-1.097564167	-0.081145110
0.601456117	0.656219999
1.748483782	0.491894584
1.365610564	1.305106471
0.657827941	-0.119462635
-0.889005698	0.411939207
1.329014094	1.044693802

Construire une échelle *(exercice 2 série 10)*

Pour **construire une échelle**, il suffit de **calculer une nouvelle variable**. Sous expression de la nouvelle variable, on additionne **tous les items qui participent à l'échelle**.

Pour l'échelle 1 : on additionne toutes les variables qui ont une **corrélation positive** avec le facteur 1 ainsi que celles qui ont une **corrélation négative** avec le facteur 2

Pour l'échelle 2 : on additionne toutes les variables qui ont une corrélation positive avec le facteur 2, et toutes celles qui ont une corrélation négative avec le facteur 1

Série 11

Analyse factorielle confirmatoire

Chemin d'accès : Statistique / Analyse multivariée / Analyse factorielle confirmatoire

1. Effectuer avant une analyse factorielle exploratoire pour déterminer les variables correspondant à quels facteurs.

2. Dans la fenêtre de l'analyse factorielle confirmatoire, **sélectionner les variables qui composent le premier facteur**, **renommer** ce facteur, puis appuyer sur **définir un facteur**

3. Refaire l'étape 2 **autant de fois qu'il y a de facteurs à construire**.

4. Dans l'onglet **Options** :

    a. Sous Matrice à analyser : sélectionner **covariance**
    b. Sous Corrélation des facteurs sélectionner
        i. **Corrélés** si les variables sont corrélées entre elles (**Obliques**)
        ii. **Orthogonaux** si les variables sont **orthogonales entre elles**
    c. Sous contrainte, sélectionner **Variance du facteur égale à 1**
    d. Sélectionner les indices d'ajustement : **AIC, BIC, RMSEA, SRMR, NFI et CFI**

5. Les données apparaîtront sous la forme suivante :

$$\chi^2(51) = 199.429 \text{ p} < .001$$
$$\chi^2/_{ddl} = 3.910$$
$$\text{RMSEA} = 0.076$$
$$\text{SRMR} = 0.034$$
$$\text{NFI} = 0.963$$
$$\text{CFI} = 0.972$$
$$\text{AIC} = 253.429$$
$$\text{BIC} = -117.516$$

Indices de validation

Indices incrémentaux : NFI et CFI
Indices absolus : RRSEA et SRMR
Indices de parcimonie : AIC et BIC

Critères de validation des indices

Indices	Critère de validation
$\chi^2(dl) = \cdots\ p = \cdots$	p > 0.05
$\chi^2/_{ddl}$	> 3
RMSEA	< 0.08
SRMR	< 0.05
NFI	< 0.9
CFI	< 0.9
AIC	Plus il est petit, mieux c'est
BIC	Plus il est petit, mieux c'est

Test statistique de l'analyse factorielle confirmatoire

Ce test statistique s'effectue lorsque l'on est face à deux modèles issus d'une analyse factorielle confirmatoire.

On nommera $\mathcal{M}_0$ notre modèle parcimonieux
On nommera $\mathcal{M}_1$ notre modèle complet
Modèles orthogonaux ou oblique, cela dépend de la consigne

1. Calculer les **degrés de liberté**. Pour cela, on **soustrait** les degrés de liberté du $\mathcal{M}_1$ à ceux du $\mathcal{M}_0$: $Ddl_0 - Ddl_1$

2. Calculer la **valeur du χ^2**. Pour cela, on **soustrait** la valeur du χ^2 du $\mathcal{M}_1$ à la valeur du χ^2 du $\mathcal{M}_0$

3. Calculer la **p-valeur** :

Chemin d'accès : Distribution / Distribution continue / Distribution du chi-deux / Probabilité du chi-deux

   a. Dans la fenêtre, **entrer la valeur du χ^2** ainsi que les **degrés de liberté**, puis sélectionner **aire à droite**. Puis valider. On obtient ainsi la p-valeur de notre test statistique.

4. Les résultats sous **forme APA** se notent de la manière suivante :
   $\chi^2(ddl) = \cdots,\ p =$

5. **Si p < α, nous rejetons l'hypothèse nulle** qui veut que le modèle parcimonieux s'accorde bien avec les variables

Contrôle Continu Automne 2017

Exercice 3 – Analyse de covariance

a) Estimer les paramètres

Chemin d'accès : Statistique / Résumé / Statistiques descriptives

1. Dans la fenêtre, cliquer sur Résumer par groupe et sélectionner Prénom. Sous variables, sélectionner Longévité. Dans l'onglet Statistique, cocher moyenne. Les données apparaissent de la manière suivante :

```
        mean
A   73.00637
B   72.07674
C   70.87692
D   71.27222
E-Z 72.17247
```

b) Estimer à nouveau les paramètres, avec D comme groupe de référence

1. Tout d'abord, il faut créer les variables centrées :

Chemin d'accès : Données / Gérer les variables du jeu de données actif / Calculer une nouvelle variable

   a. Sous Variables existantes, sélectionner Naissance, puis sous Expression à calculer, entrer la formule Naissance – 1900 et renommer la variable. Faire la même chose pour la variable Carrière en entrant la formule Carrière – 10 et renommer la variable

2. Coder les variables afin que la variable D soit notre groupe de référence

Chemin d'accès : Données / Gérer les variables du jeu de données actif / Réordonner les niveaux d'un facteur

   a. Sous Facteur, sélectionner Prénom et valider
   b. Une nouvelle fenêtre va s'ouvrir, avec les anciens niveaux et les nouveaux ordres à définir.
   c. Donner la valeur de 1 à D, puis valider
   d. Définir les contrastes d'un facteur :

Chemin d'accès : Données / Gérer les données du jeu de données actif / Définir les contrastes d'un facteur

   e. Sélectionner Contraste de Traitement puis valider.
   f. Pour vérifier que le contraste est correcte, on peut sélectionner dans la console la formule contrasts(LONGEVITE$Prenom) et cliquer sur Soumettre. On obtient un tableau :

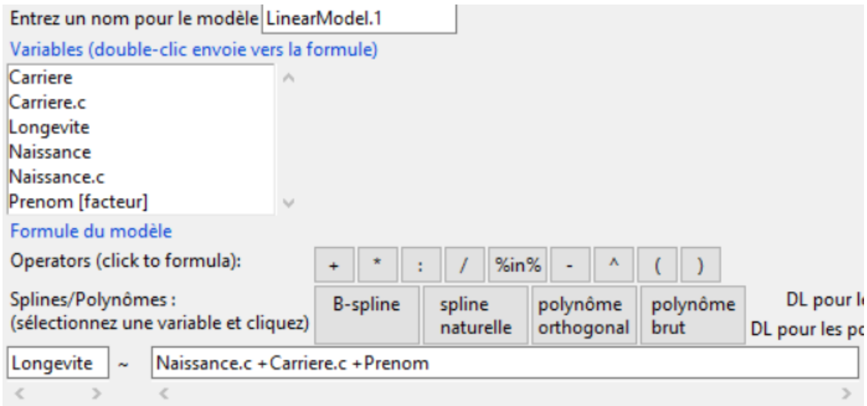
```
      [T.A] [T.B] [T.C] [T.E-Z]
D         0     0     0       0
A         1     0     0       0
B         0     1     0       0
C         0     0     1       0
E-Z       0     0     0       1
```

g. Si D a la valeur de 0 sur toute la ligne, notre contraste est correctement établi.

3. Créer un modèle linéaire purement additif :

Chemin d'accès : Statistiques / Ajustement de modèle / Modèle linéaire

a. Dans la fenêtre, sélectionner pour la première case la variable longévité, puis ajouter les variables naissance centrée, carrière centrée et Prénom :



b. Nous avons donc le modèle suivant : $Longevit\acute{e} \sim Naissance.c + Carriere.c + Prenom$

c. Les données apparaissent sous cette forme :

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.078388   0.851436  83.481   <2e-16 ***
Naissance.c  -0.026704   0.012032  -2.219   0.0265 *
Carriere.c    0.027330   0.045002   0.607   0.5437
Prenom[T.A]   2.200077   1.252047   1.757   0.0790 .
Prenom[T.B]   1.244620   1.026363   1.213   0.2253
Prenom[T.C]  -0.003714   1.181654  -0.003   0.9975
Prenom[T.E-Z] 1.191541   0.882859   1.350   0.1772
```

4. Estimer les paramètres du modèle :

a. L'intercept correspond au groupe de référence, ici D. La valeur sous D est donc de 71.078

b. Pour estimer les autres paramètres nous devons : prendre la valeur de D et lui ajouter la valeur des valeurs Estimate correspondant à chaque lettre donc :

A	$\mathbf{71.078 + 2.2000 = 73.278}$
B	$71.078 + 1.245 = 72.323$
C	$71.078 + (-0.004) = 71.004$
D	71.078
E-Z	$71.078 + 1.192 = 72.270$

c) Test de différences significatives

1. Pour effectuer ce test, il va nous falloir deux modèles. Le premier modèle est celui que nous avons créé au point B, c'est notre modèle complet.

2. Nous allons créer un deuxième modèle qui ne comprendra que les covariables Naissance centrée et Carrière centrée. Nous obtenons le modèle suivant :

$$Longevite \sim Naissance.c + Carriere.c$$

   C'est notre modèle parcimonieux ! Il apparaît ainsi :

```
Call:
lm(formula = Longevite ~ Carriere.c + Naissance.c, data = LONGEVITE)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.693  -7.767   0.223   8.193  27.286 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 72.17061    0.24061 299.955   <2e-16 ***
Carriere.c   0.02700    0.04499   0.600   0.5485    
Naissance.c -0.02371    0.01176  -2.017   0.0438 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.3 on 3358 degrees of freedom
Multiple R-squared:  0.001317,	Adjusted R-squared:  0.0007217 
F-statistic: 2.213 on 2 and 3358 DF,  p-value: 0.1095
```

3. Une fois nos deux modèles créés, nous allons en faire la comparaison :

Chemin d'accès : Modèles / Tests d'hypothèses / Comparer deux modèles

   a. Dans la fenêtre, sélectionner le modèle parcimonieux sous Premier modèle et le modèle complet sous Second modèle, puis valider

   b. Les résultats apparaissent ainsi :

```
Analysis of Variance Table

Model 1: Longevite ~ Carriere.c + Naissance.c
Model 2: Longevite ~ Carriere.c + Naissance.c + Prenom
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   3358 428725                           
2   3354 428060  4    665.47 1.3036 0.2663
```

4. Notre hypothèse nulle est que notre modèle parcimonieux plus adéquat que notre modèle complet, donc que la lettre débutant le prénom n'a aucun effet sur la longévité.

5. Les résultats du test se notent ainsi : $F(4, 3354) = 1.304, p = 0.266$

6. Conclusion : comme p > 0.05, nous ne pouvons pas rejeter l'hypothèse nulle. Ainsi, la lettre débutant le prénom n'a aucun effet sur la longévité !

Deuxième méthode !

On peut aussi obtenir les mêmes résultats en faisant une table d'ANOVA :

Chemin d'accès : Modèles / Tests d'hypothèses / Table d'ANOVA

Dans la fenêtre, on ne change absolument rien et on valide. On obtient un tableau sous la forme suivante :

```
Anova Table (Type II tests)

Response: Longevite
             Sum Sq   Df F value   Pr(>F)
Prenom          665    4  1.3036  0.26625
carriere.c       47    1  0.3688  0.54369
naissance.c     629    1  4.9257  0.02653 *
Residuals    428060 3354
```

Les résultats du test sont : $F(4, 3354) = 1.304, p = 0.266$

Constat : que l'on fasse une comparaison entre un modèle complet et un modèle parcimonieux, ou que l'on utilise une table d'ANOVA, le résultat est le même !

⚠ La table d'ANOVA est néanmoins plus rapide à mettre en place !

d) Tests marginaux

1. Pour déterminer si les baseballeurs dont le prénom commencent par la lettre D vivent significativement plus longtemps que ceux dont la lettre commence par A, B, C, E-Z, nous devons effectuer des tests marginaux

2. Nous avons donc besoin de notre table du modèle complet que nous avons créé au point b) :

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.078388   0.851436  83.481   <2e-16 ***
Carriere.c    0.027330   0.045002   0.607   0.5437
Naissance.c  -0.026704   0.012032  -2.219   0.0265 *
Prenom[T.A]   2.200077   1.252047   1.757   0.0790 .
Prenom[T.B]   1.244620   1.026363   1.213   0.2253
Prenom[T.C]  -0.003714   1.181654  -0.003   0.9975
Prenom[T.E-Z] 1.191541   0.882859   1.350   0.1772
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.3 on 3354 degrees of freedom
Multiple R-squared:  0.002867, Adjusted R-squared:  0.001083
F-statistic: 1.607 on 6 and 3354 DF,  p-value: 0.1409
```

3. Pour effectuer les tests marginaux, nous devons relever les valeurs sous les colonnes t value et Pr(>|t|) (que nous appellons p) pour chaque lettre.

⚠ Il s'agit d'un test unilatéral ! Or, la table nous donne les p-valeurs d'un test bilatéral. Nous devons donc calculer les p-valeurs de la manière suivante :

a. Si le coefficient est positif, la p-valeur se calcule ainsi : $p/2$
b. Si le coefficient est négatif, la p-valeur se calcule ainsi : $1 - p/2$

4. Notre hypothèse nulle est que les personnes dont le prénom commence par D meurent plus tôt. Ainsi les tests marginaux sont les suivants :

Tests	**Valeur de p**	**Résultat**	**Conclusion**
D vs A	$0.079/_2 = 0.0395 = 0.040$	$t(3354) = 1.757, p = 0.040$	p < 0.05 = pas de différence
D vs B	$0.225/_2 = 0.1125 = 0.113$	$t(3354) = 1.757, p = 0.113$	p > 0.05 = différence !
D vs C	$1 - 0.998/_2 = 0.501$	$t(3354) = 1.757, p = 0.501$	p > 0.05 = différence !
D vs E-Z	$0.177/_2 = 0.0885 = 0.089$	$t(3354) = 1.757, p = 0.089$	p < 0.05 = pas de différence