

Statistique II



Enseignant : Jean-Philippe Antonietti

Assistants : Eric Eboulet

Semestre automne 2016

Table des matières

I. Bonus :	3
1. Analyse de covariance (0.5 points)	3
2. Taille de l'échantillon (0.25 points)	4
a. Tests de Student.....	4
b. Analyse de variance à un facteur	4
c. corrélation	5
d. régression multiple	5
e. régression logistique avec une variable explicative binaire :.....	6
II.Examen	7
1. Analyse de variance à deux facteurs	7
Codage pour vérifier les égalités :	8
2. Plan Mixte	13
3. régression logistique	20

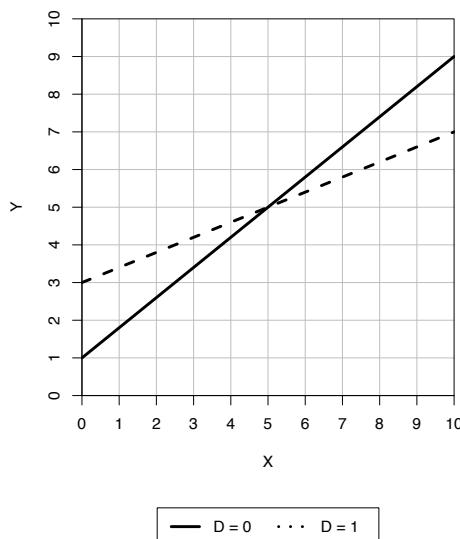
I. Bonus :

(Prends peu de temps)

1. Analyse de covariance (0.5 points)

3) La partie déterministe du modèle :

$$Y = \alpha + \beta \cdot X + \gamma \cdot D + \delta \cdot (D \times X) + \varepsilon$$

est représentée dans le graphique ci-dessous. Que valent les paramètres α , β , γ et δ ? α : la valeur de α est l'intersection de la droite D=0 avec l'axe Y. ici $\alpha = 1$ γ : La valeur de γ est l'intersection de la droite D = 1 avec l'axe Y moins la valeur de α . Ainsi

$$\gamma = 3 - 1 = 2$$

 β : la valeur de β c'est la pente de la droite D=0. Ainsi $\beta = \frac{h}{l} = 8/10$ δ : pour trouver la valeur de δ il faut faire de substitution du modèle

$$\begin{aligned} 1. \quad \mathbf{D=0} : Y &= \alpha_0 + \beta_0 x &= 1 + 0.8x \\ 2. \quad \mathbf{D = 1} : Y &= \alpha_1 + \beta_1 x + \gamma + \delta x &= 2 + 0.4 + 2 + \delta x \end{aligned}$$

On sait que $(\beta_0 + \delta) = \beta_1$ Ainsi $0.8 + \delta = 0.4$ et donc $\delta = -0.4$

2. Taille de l'échantillon (0.25 points)

La taille de l'échantillon permet de savoir quel est le nombre de participant minimum (=N) pour avoir une bon Test. Les démarches varie en fonction du teste.

a. Tests de Student

Définition

$$d = \frac{|\mu_2 - \mu_1|}{\sigma}$$

Guide d'interprétation

Petite	Moyenne	Grande
0.20	0.50	0.80

Sortie

```
Two-sample t test power calculation

n = 393.4057
d = 0.2
sig.level = 0.05
power = 0.8
alternative = two.sided

NOTE: n is number in *each* group
```

```
▶ library(pwr)
▶ pwr.t.test(n,
d,
sig.level = 0.05,
power,
type = "two.sample",
alternative = "two.sided")
```

ici il faut faire $n \times d = N$

b. Analyse de variance à un facteur

Définition

$$f = \sqrt{\frac{\frac{1}{N} \sum n_j (\mu_j - \mu)^2}{\sigma^2}}$$

Guide d'interprétation

Petite	Moyenne	Grande
0.10	0.25	0.40

```
▶ library(pwr)
▶ pwr.anova.test(k,
n,
f,
sig.level = 0.05,
power)
```

Sortie

```
Balanced ANOVA power calculation

k = 4          nombre de facteur
n = 44.5993
f = 0.25        taille d'effet
sig.level = 0.05 seuil alpha
power = 0.8      puissance donnée en %

NOTE: n is number in each group
```

Pour trouver N il faut faire $K \times n = N$

c. corrélation

Définition

La taille d'effet est le coefficient de corrélation lui-même :

$$ES = r$$

Guide d'interprétation

Petite	Moyenne	Grande
0.10	0.30	0.50

Sortie

Approximate correlation power calculation
(arctanh transformation)

```
n = 66.55463
r = 0.3
sig.level = 0.05
power = 0.8
alternative = greater
```

```
► library(pwr)
► pwr.r.test(n,
r,
sig.level = 0.05,
power,
alternative = "two.sided")
```

ici pour trouver N il faut faire :

$$n \times r = N$$

d. régression multiple

Définition

$$f^2 = \frac{R^2}{1 - R^2}$$

Guide d'interprétation

Petite	Moyenne	Grande
0.02	0.15	0.35

```
► library(pwr)
► pwr.f2.test(u,
v,
f2,
sig.level = 0.05,
power)
```

Avec :

- $u = k$;
- $v = N - k - 1$.

$$R^2 = 0.2 \Rightarrow f^2 = \frac{0.2}{1 - 0.2} = 0.25$$

Entrée

```
► pwr.f2.test(u = 3,
f2 = 0.25,
sig.level = 0.05,
power = 0.80)
```

Sortie

Multiple regression power calculation

```
u = 3
v = 43.70447
f2 = 0.25
sig.level = 0.05
power = 0.8
```

Pour trouver N il faut faire :

$$N = v + k + 1$$

e. régression logistique avec une variable explicative binaire :

Définition

La taille d'effet est le rapport de chances lui-même :

$$ES = OR$$

Guide d'interprétation

Petite	Moyenne	Grande
1.5	3.5	9.0

- ▶ library(powerMediation)
- ▶ SSizeLogisticBin(p1,
p2,
B,
alpha = 0.05,
power = 0.8)

Avec :

- ▶ p1 = $P(Y = 1|X = 0)$;
- ▶ p2 = $P(Y = 1|X = 1)$;
- ▶ B = $P(X = 1)$

Entrée

- ▶ SSizeLogisticBin(p1 = 0.256,
p2 = 0.075,
B = 0.576,
alpha = 0.05,
power = 0.8)

Sortie

```
[1] 130
```

II. Examen

1. Analyse de variance à deux facteurs

Statistique > Moyenne > ANOVA à plusieurs facteurs

$$y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}$$

```
> AnovaModel.agexformation <- lm(Durée ~ âge*Formation, data=TP7Ex01,
+   contrasts=list(âge ="contr.Sum", Formation ="contr.Sum"))

> Anova(AnovaModel.agexformation)
Anova Table (Type II tests)

Response: Durée
            Sum Sq Df F value    Pr(>F)
âge          122.722  1 16.7348 0.001497 ***
Formation     100.778  2  6.8712 0.010261 *
âge:Formation 52.111  2  3.5530 0.061385 .
Residuals    88.000 12

Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(TP7Ex01, (tapply(Durée, list(âge, Formation), mean, na.rm=TRUE))) # means
      primaire secondaire supérieure
moinsde40       3  2.000000  1.333333
plusde40        13  5.333333  3.666667

> xtabs(~ âge + Formation, data=TP7Ex01) # counts
           Formation
âge      primaire secondaire supérieure
moinsde40      3        3        3
plusde40       3        3        3
```

Pour répondre à la question il faut regarder le premier tableau bleu

Effet d'interaction :

- Test de l'effet de l'interaction entre **formation** et **âge** au seuil de $\alpha = 5\%$:

$$\begin{aligned} H_0 & : (\forall j)(\forall k) \gamma_{jk} = 0 \\ H_1 & : (\exists j)(\exists k) \gamma_{jk} \neq 0 \end{aligned}$$

$F(2, 12) = 3.553$, $p = 0.061$. Comme $p > \alpha$, nous ne rejetons pas H_0 . L'interaction n'a pas d'effet statistiquement significatif. Vu la petite taille de nos échantillons, nous manquons de puissance et commettons peut-être une erreur de seconde espèce en affirmant l'absence d'interaction.

L'interaction étant égale à zéro, nous pouvons analyser les effets principaux.

Effet principal :

- Test de l'effet principal de la variable **âge** au seuil de $\alpha = 5\%$:

$$\begin{aligned} H_0 & : (\forall j) \alpha_j = 0 \\ H_1 & : (\exists j) \alpha_j \neq 0 \end{aligned}$$

$F(1, 12) = 16.735$, $p = 0.002$. Comme $p < \alpha$, nous rejetons H_0 . L'âge a un effet statistiquement significatif sur la durée du chômage.

- Test de l'effet principal de la variable **formation** au seuil de $\alpha = 5\%$:

$$\begin{aligned} H_0 & : (\forall k) \beta_k = 0 \\ H_1 & : (\exists k) \beta_k \neq 0 \end{aligned}$$

$F(2, 12) = 6.871$, $p = 0.010$. Comme $p < \alpha$, nous rejetons H_0 . Le niveau de formation a un effet significatif sur la durée du chômage.

Tant l'âge que la formation influencent la durée du chômage. Ces derniers résultats sont en contradiction avec les résultats des analyses de variance simples. Ceci s'explique par le fait qu'une analyse de variance à deux facteurs est plus puissante que deux analyses de variance simples. En effectuant une analyse de variance à deux facteurs, si H_1 est vraie, il y a plus de chance de rejeter H_0 .

Codage pour vérifier les égalités :

variable	composition du modèle	
1	Modèle avec constance	Durée = 1
2 alpha	Modèle âge	Duré = jeune
3 beta	Modèle formation	durée = primaire + secondaire
4 gamma	modèle interaction âge x formation	durée = Jeunes:Primaire +Jeunes:Secondaire
5 alpha + beta	modèle âge + formation (sans interaction)	duré = Jeunes +Primaire +Secondaire
6 alpha + gamma	modèle âge + âge x formation (l'interaction)	Jeunes +(Jeunes:Primaire +Jeunes:Secondaire)
7 bêta + gamma	modèle formation + âge x formation	Primaire + Secondaire +Jeunes:Primaire +Jeunes:Secondaire
8 alpha + bêta + gama	Modèle formation + âge + âge x formation	Jeunes + Primaire + Secondeaire + Jeunes:Primaire + Jeunes:Secondaire

Codage somme

Codage Somme

- 1) Définir la variable de référence
 - a. Âge : >40ans
 - b. Formation : supérieur
- 2) Réordonné les niveau
 - a. Âge : <40ans = 1 → 1 >40an = 0 → 2
 - b. Formation : Supérieur = 0 → 3 secondaire = 1 → 1 Primaire = 1 → 2
- 3) Définir un contraste de traitement
- 4) Construction de la variable d'interaction :

- Codage somme

$$\begin{aligned}
 SC(\alpha) &= 122.72 \\
 SC(\beta) &= 100.78 \\
 SC(\gamma) &= 52.11 \\
 SC(\alpha, \beta) &= 223.50 \\
 SC(\alpha, \gamma) &= 174.83 \\
 SC(\beta, \gamma) &= 152.89 \\
 SC(\alpha, \beta, \gamma) &= 275.61 \\
 SC(\alpha|\beta, \gamma) &= SC(\alpha, \beta, \gamma) - SC(\beta, \gamma) = 275.61 - 152.89 = 122.72 \\
 SC(\alpha|\beta) &= SC(\alpha, \beta) - SC(\beta) = 223.50 - 100.78 = 122.72 \\
 SC(\alpha) &= 122.72 \\
 \\
 SC(\beta|\alpha, \gamma) &= SC(\alpha, \beta, \gamma) - SC(\alpha, \gamma) = 275.61 - 174.83 = 100.78 \\
 SC(\beta|\alpha) &= SC(\alpha, \beta) - SC(\alpha) = 223.50 - 122.72 = 100.78 \\
 SC(\beta) &= 100.78 \\
 \\
 SC(\gamma|\alpha, \beta) &= SC(\alpha, \beta, \gamma) - SC(\alpha, \beta) = 275.61 - 223.5 = 52.11 \\
 SC(\gamma) &= 52.11
 \end{aligned}$$

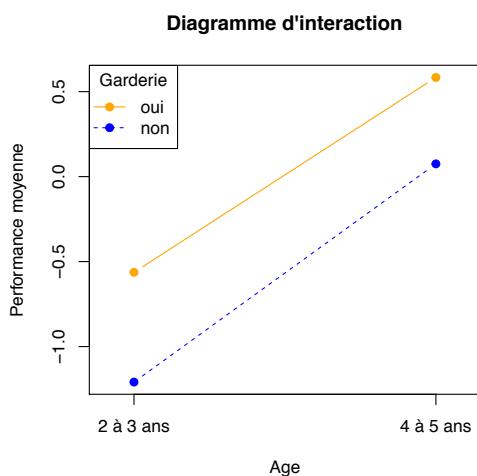
	Codage somme					
	S_1	S_2	S_3	...	S_{m-1}	
a_1	1	0	0	...	0	
a_2	0	1	0	...	0	
a_3	0	0	1	...	0	
:	:	:	:	..	:	
a_{m-1}	0	0	0	...	1	
a_m	-1	-1	-1	...	-1	

Rappel : la variable de référence est la dernière.

Analyse de variance à deux facteur de type II**Résumé numérique****Statistique**

Moyennes		
Âge	Garderie	
	non	oui
2-3 ans	-1.209	-0.563
4-5 ans	0.075	0.584

Écarts-types		
Âge	Garderie	
	non	oui
2-3 ans	0.861	0.977
4-5 ans	0.436	0.496

Diagramme d'interaction = Motif des moyennes

Le graphe ci-contre, semble démontrer une absence d'interaction.

Mais une présence d'effet simple de l'expérience et de l'âge.

Il faut vérifier avec une analyse de variance afin de s'en assurer.

ANOVA de type II

$$\mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}.$$

Avec un codage somme ; Codage traitement

Codage somme

- 1) Réorganiser :
 - a. Données > Géré > réorganiser les niveaux
 - b. **Attention la variable qui n'apparaîtras pas est la variable la plus grande. (Ainsi ici la variable de référence sera la valeur la plus grande et non la plus petite).**
- 2) Définir le contraste
 - a. Donnée > Gérer > définir ...
 - b. Choisir contraste sommes
- 3) Calculer le modèle linéaire
 - a. Statistique > ajustement de modèle > modèle linéaire
 - b. La plus grande variable n'apparaîtra pas.
 - c. **Attention de pas prendre la valeur du modèle !!!**
- 4) Analyse de variance de Type 2
 - a. Modèle > Test d'hypothèse > Table d'Anova

Score ~ age*Expérience

```
> Anova(LinearModel.8, type="II")
Anova Table (Type II tests)

Response: Score
            Sum Sq Df F value    Pr(>F)
age          14.3717  1 24.5780 1.71e-05 ***
Expérience   3.1719  1  5.4245  0.02558 *
age:Expérience 0.0374  1  0.0640  0.80178
Residuals    21.0506 36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interprétation des résultats

- Test de l'effet de l'interaction entre **age** et **garderie** au seuil de $\alpha = 5\%$:

$$\begin{aligned} H_0 &: (\forall j)(\forall k) \gamma_{jk} = 0 \\ H_1 &: (\exists j)(\exists k) \gamma_{jk} \neq 0 \end{aligned}$$

$F(1, 36) = 0.064$, $p = 0.802$. Comme $p > \alpha$, nous ne rejetons pas H_0 . L'effet de l'interaction entre âge et expérience en garderie sur la prise de rôles est égal à zéro. Voyons si les effets principaux sont significatifs.

- Test de l'effet principal de la variable **age** au seuil de $\alpha = 5\%$:

$$\begin{aligned} H_0 &: (\forall j) \alpha_j = 0 \\ H_1 &: (\exists j) \alpha_j \neq 0 \end{aligned}$$

$F(1, 36) = 24.577$, $p < 0.001$. Comme $p_{valeur} < \alpha$, nous rejetons H_0 . L'âge a un effet statistiquement significatif sur la prise de rôles. Les enfants qui 4-5 ans sont plus aptes à la prise de rôles que ceux qui ont 2-3 ans.

- Test de l'effet principal de la variable **garderie** au seuil de $\alpha = 5\%$:

$$\begin{aligned} H_0 &: (\forall k) \beta_k = 0 \\ H_1 &: (\exists k) \beta_k \neq 0 \end{aligned}$$

$F(1, 36) = 5.424$, $p = 0.026$. Comme $p < \alpha$, nous rejetons H_0 . Les enfants ayant bénéficié d'expériences en garderie obtiennent de meilleures performances à l'épreuve de prise de rôles.

Les hypothèses des chercheurs sont corroborées par l'expérience : la prise de rôles est une compétence qui croît avec l'âge, d'une part, et avec l'intensité des échanges sociaux, d'autre part.

Codage traitement

Théoriquement les résultats d'une analyse de variance de type II ne devraient pas être influencés par le type de codage. C'est ce que l'on observe : que l'on utilise un codage *somme* ou un codage *traitement* les résultats sont les mêmes.

Reconstruction du tableau de moyenne à partir des coefficients de régression du modèle construit

Codage somme

Le modèle que nous cherchons à estimer est le suivant :

$$\mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}.$$

Comme $\sum_j \alpha_j = 0$, $\sum_k \beta_k = 0$, $(\forall k) \left(\sum_j \gamma_{jk} = 0 \right)$ et $(\forall j) \left(\sum_k \gamma_{jk} = 0 \right)$, nous pouvons poser les égalités suivantes :

$$\begin{aligned}\alpha_2 &= -\alpha_1 \\ \beta_2 &= -\beta_1 \\ \gamma_{12} &= -\gamma_{11} \\ \gamma_{21} &= -\gamma_{11} \\ \gamma_{22} &= +\gamma_{11}\end{aligned}$$

Ainsi :

$$\begin{aligned}\mu_{11} &= \mu + \alpha_1 + \beta_1 + \gamma_{11} = \mu + \alpha_1 + \beta_1 + \gamma_{11} \\ \mu_{12} &= \mu + \alpha_1 + \beta_2 + \gamma_{12} = \mu + \alpha_1 - \beta_1 - \gamma_{11} \\ \mu_{21} &= \mu + \alpha_2 + \beta_1 + \gamma_{21} = \mu - \alpha_1 + \beta_1 - \gamma_{11} \\ \mu_{22} &= \mu + \alpha_2 + \beta_2 + \gamma_{22} = \mu - \alpha_1 - \beta_1 + \gamma_{11}\end{aligned}$$

```
> summary(LinearModel.8)
```

```
Call:
lm(formula = Score ~ age * Expérience, data = TP7Exo2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5329 -0.4353 -0.1474  0.3508  1.3880 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.27840   0.13582 -2.050   0.0477 *  
age[S.A]      0.60765   0.13582  4.474   7.4e-05 *** 
Expérience[S.N] -0.28860   0.13582 -2.125   0.0405 *  
age[S.A]:Expérience[S.N]  0.03435   0.13582  0.253   0.8018 
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7647 on 36 degrees of freedom
Multiple R-squared:  0.4277, Adjusted R-squared:  0.38 
F-statistic: 8.967 on 3 and 36 DF,  p-value: 0.0001435
```

nous pouvons estimer la moyenne de chacune des cellules du plan expérimental par :

$$\begin{aligned}\hat{\mu}_{11} &= -0.278 + (-0.608) + (-0.289) + (-0.034) = -1.209 \\ \hat{\mu}_{12} &= -0.278 + (-0.608) - (-0.289) - (-0.034) = -0.563 \\ \hat{\mu}_{21} &= -0.278 - (-0.608) + (-0.289) - (-0.034) = 0.075 \\ \hat{\mu}_{22} &= -0.278 - (-0.608) - (-0.289) + (-0.034) = 0.585\end{aligned}$$

Ces estimations correspondent exactement aux moyennes empiriques calculées en a).

Codage traitement

Le modèle est, comme précédemment :

$$\mu_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}.$$

Selon le codage *traitement*, les contraintes sont : $\alpha_1 = 0$, $\beta_1 = 0$ et $\gamma_{11} = \gamma_{12} = \gamma_{21} = 0$

Ainsi :

$$\begin{aligned}\mu_{11} &= \mu \\ \mu_{12} &= \mu + \beta_2 \\ \mu_{21} &= \mu + \alpha_2 \\ \mu_{22} &= \mu + \alpha_2 + \beta_2 + \gamma_{22}\end{aligned}$$

```
Call:
lm(formula = Score ~ T.age * T.Exp, data = TP7Exo2)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5329 -0.4353 -0.1474  0.3508  1.3880 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.2090    0.2044 -5.916 9.01e-07 ***
T.age[T.A]   1.2840    0.3008  4.268 0.000137 ***  
T.Exp[T.0]   0.6459    0.3166  2.040 0.048733 *   
T.age[T.A]:T.Exp[T.0] -0.1374    0.5433 -0.253 0.801778  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7647 on 36 degrees of freedom
Multiple R-squared:  0.4277, Adjusted R-squared:  0.38 
F-statistic: 8.967 on 3 and 36 DF,  p-value: 0.0001435
```

nous pouvons estimer la moyenne de chacune des cellules du plan expérimental par :

$$\begin{aligned}\hat{\mu}_{11} &= -1.209 &= -1.209 \\ \hat{\mu}_{12} &= -1.209 + 0.646 &= -0.563 \\ \hat{\mu}_{21} &= -1.209 + 1.284 &= 0.075 \\ \hat{\mu}_{22} &= -1.209 + 1.284 + 0.646 + (-0.137) &= 0.584\end{aligned}$$

Ces estimations correspondent également aux moyennes empiriques calculées en a).

2. Plan Mixte

Chargé le plugins

Outils > Chargé plugins > Rcmd...>

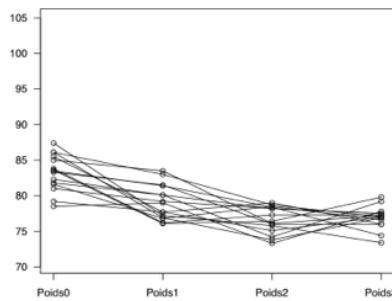
Représentez graphiquement

Évolution du poids

Ici on nous demande de représentez graphiquement l'évolution du poids de chaque sujet.

Graphs ans tables > line graphe (mesure répété)

Choisir toute les variable pour le premier et aucun pour le deuxième.



On peut dire qu'il y a une **tendance** de régression entre 1 et 3. On constate également une tendance de constance entre 2 et 3.

Plan à mesure répétés

Ici on va utiliser le plan à mesure répété, car l'étude mesure le poids des mêmes sujets à 4 moment différentes, le plan d'expérience est un plan à mesure répétées. L'analyse à appliquer est donc une analyse de variance à un facteur intra-sujet.

$$x_{ij} = \mu + \pi_j + \alpha_i + \epsilon_{ij} \text{ avec } \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2).$$

Hypothèse de sphéricité

H_0 : La Sphéricité est satisfaite

H_1 : La Sphéricité n'est pas satisfaite

Statistical Analysis > continuous > Repeated-mesure ANOVA

Nommé la variable > sélectionner les variables > sélectionner Pairwise comparaison (Holm)*

```
> res <- Anova(AnovaModel.2, idata=time, idesign=~Time, type="III")
> summary(res, multivariate=FALSE)

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

          SS num Df Error SS den Df      F      Pr(>F)
(Intercept) 374097     1   51.819     14 101069.491 < 2.2e-16 ***
Time         423       3   209.097     42    28.301 3.582e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Mauchly Tests for Sphericity
Test statistic p-value
Time        0.92142 0.9593
```

43

- Sorties informatiques

```
Test statistic p-value
Time      0.92142  0.9593
```

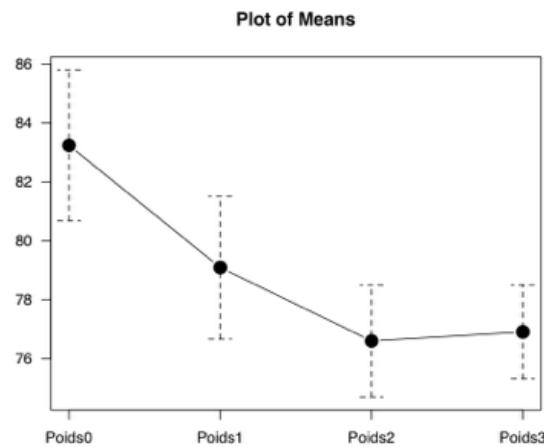
- Interprétation

$W = 0.921$, $p = 0.959$. Comme $p > \alpha$, on ne rejette pas H_0 au seuil $\alpha = 5\%$. L'hypothèse de sphéricité est corroborée. Nous n'aurons pas besoin de corriger les résultats de l'analyse de variance.

Comme la p-valeurs est plus grand que alpha ($p=0.959 > 0.05$) on ne rejette pas H_0 . Ainsi la sphéricité est satisfaite. Ainsi on n'a pas besoin de faire des correction avec des test de correction.

Graph de l'évolution du poids moyen

Le test de sphéricité nous donne également l'illustration graphique de l'évolution du poids moyen des 15 sujets ayant été suivis.



ANOVA sur mesure répétée

```
> res <- Anova(AnovaModel.2, idata=time, idesign=~Time, type="III")
> summary(res, multivariate=FALSE)

Univariate Type III Repeated-Measures ANOVA Assuming Sphericity

SS num Df Error SS den Df          F          Pr(>F)
(Intercept) 374097     1   51.819     14 101069.491 < 2.2e-16 ***
Time        423       3   209.097     42    28.301 3.582e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Mauchly Tests for Sphericity

Test statistic p-value
Time      0.92142  0.9593
```

- Interprétation

$F(3, 42) = 28.301$, $p < 0.001$. On rejette donc H_0 au seuil $\alpha = 5\%$. Le poids moyen des personnes participant à l'expérience n'est pas toujours le même, il évolue.

Il faut donc faire des comparaison 2 par 2 avec le teste d'Holm afin d'avoir les différence entre les moyenne

***Le test de pairwise comparaison holm ou Comparaison multiple**

(Ce test nous permet de comparer les moyenne 2 par 2. Ainsi d'avoir une meilleure analyse)

Les résultats se trouvent vers la fin du tableau du test fait précédemment.

Comparaisons multiples

Comme le facteur intra-sujet possède 4 modalités, nous allons réaliser la comparaison de 6 paires de variables.

```
> pairwise.pairedt.test(with(TempDF, cbind(Poids0, Poids1, Poids2, Poids3)),
+   group=NULL, "Boulimie", p.adjust.method="holm")
Pairwise comparisons using Paired t-test
data: Boulimie
Poids0  Poids1 Poids2
Poids1 0.0013 -
Poids2 1.5e-05 0.0139 -
Poids3 2.3e-05 0.0282 0.6755
P value adjustment method: holm
```

Résultats

- 0 Vs 1:** $p = 0,001$
1 vs 2 : $p = 0,014$
2 vs 3: $p = 0,676$

- Interprétation

Toutes les différences sont statistiquement significatives au seuil de 5% à l'exception d'une seule : celle qui compare le poids mesuré en fin de traitement à celui mesuré six mois plus tard ($p = 0.676$). Du début à la fin de l'intervention, les sujets perdent donc régulièrement du poids : la perte de poids est significative entre le début et le milieu du traitement ($p = 0.001$) ; elle l'est aussi entre le milieu et la fin du traitement ($p = 0.014$). Par contre, une fois le traitement terminé, leur poids ne change plus, il reste stable ($p = 0.676$).

Exercice 3 : Correction Greenhouse-Geisser

En cas non satisfaction de l'hypothèse de la sphéricité, il faut appliquer la correction de Greenhouse-Geisser.

	F_{emp}	df_{num}	df_{den}	$\hat{\epsilon}$
a)	1.113	3	27	0.548
b)	2.515	5	90	0.734
c)	2.515	5	90	0.319
d)	5.719	2	22	0.845

H_0 : La sphéricité est satisfaite

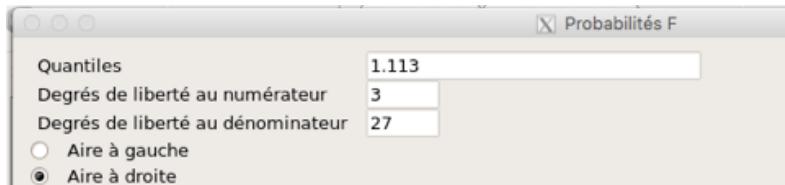
H_1 : La sphéricité n'est pas satisfaite

Lorsque la sphéricité est satisfaite

$$p = P(F > F_{\text{emp}} | F \sim F(df_{\text{num}}, df_{\text{den}}))$$

Avec le tableau ci-dessus on sait que : $F(3,27) = 1.113$ on cherche que vaut la p-valeur :

Menu originale > distribution > Distribution continu > Distribution F > probabilité de F



ATTENTION il ne faut pas mettre une virgule mais un point ! Sans quoi la réponse sera fausse !

45

Exemple de faute si on met une virgule Vs point :

```
> pf(c(1,113), df1=3, df2=27, lower.tail=FALSE)
[1] 4.079149e-01 2.139732e-15
> pf(c(1.113), df1=3, df2=27, lower.tail=FALSE)
[1] 0.3610892
```

Avec une
virgule

Avec un
point

La réponse est donc : $F(3 ; 27) = 1.113, p=0,361$

Lorsque la sphéricité n'est pas satisfaite

$$p = P(F > F_{\text{emp}} | F \sim F(\hat{\varepsilon} \cdot df_{\text{num}}, \hat{\varepsilon} \cdot df_{\text{den}}))$$

Avec le tableau ci dessus, on fait une correction : $F(0,548 \times 3 ; 0,548 \times 27) = 1.113$. il faut chercher la p-valeur

Menu originale > distribution > Distribution continu > Distribution F > probabilité de F

Quantiles	1.113
Degrés de liberté au numérateur	1.644
Degrés de liberté au dénominateur	14.796
<input type="radio"/> Aire à gauche	
<input checked="" type="radio"/> Aire à droite	

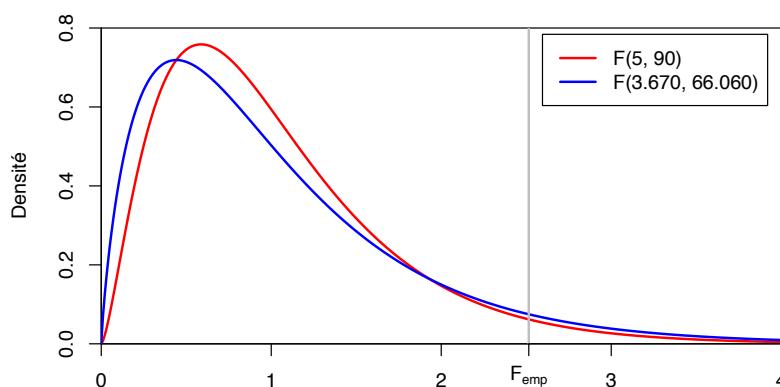
```
> pf(c(1.113), df1=1.644, df2=14.796, lower.tail=FALSE)
[1] 0.3428916
```

Réponse : $F(1.644 ; 14.796) = 1.113 ; p = 0,343$

Les probabilités critiques prennent donc les valeurs suivantes :

	<i>Hypothèse de sphéricité supposée</i>	
	VRAIE	FAUSSE
a)	$F(3, 27) = 1.113, p = 0.361$;	$F(1.644, 14.796) = 1.113, p = 0.343$.
b)	$F(5, 90) = 2.515, p = 0.035$;	$F(3.670, 66.060) = 2.515, p = 0.054$.
c)	$F(5, 90) = 2.515, p = 0.035$;	$F(1.595, 28.710) = 2.515, p = 0.109$.
d)	$F(2, 22) = 5.719, p = 0.010$;	$F(1.690, 18.590) = 5.719, p = 0.015$.

Distributions du point b)



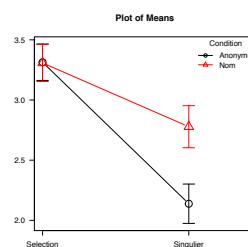
Exercice 2

Les questionnaires auto-reportés sont sujets à un biais bien étudié : la désirabilité sociale. Effectivement, lorsque l'on répond à des questions sur soi, la tentation peut être grande de se présenter sous son plus beau jour. Il est clair, que lorsque le test est passé dans une situation de sélection (par exemple un entretien d'embauche), la désirabilité sociale risque d'être augmentée, mais qu'en est-il lorsque le test est passé dans des conditions singulières ? Est-ce que les individus risquent quand même de biaiser leurs résultats ?

Pour répondre à cette question, certains de nos collègues de l'Université de Lausanne proposeront à 200 étudiants de passer une fois un test en situation "singulière" et une fois en situation de "sélection". De plus, ils demanderont à la moitié des étudiants de noter leurs noms sur les deux tests, et demanderont l'anonymat à l'autre moitié.

Analyse globale

- a) Représentez les résultats de cette expérience par un diagramme d'interaction.



- b) Réalisez une analyse de variance sur plan mixte, avec comme facteur intra-sujet la situation (Sélection vs Singulière) et comme facteur inter-sujets la condition (Nom vs Anonyme).

- Sorties informatiques

	SS	num Df	Error SS	den Df	F	Pr(>F)
(Intercept)	1664.07	1	238.023	98	685.1397	< 2.2e-16 ***
Condition	5.02	1	238.023	98	2.0687	0.1535
Situation	36.38	1	16.558	98	215.3196	< 2.2e-16 ***
Condition:Situation	5.22	1	16.558	98	30.8738	2.37e-07 ***

Signif. codes:	0	'***'	0.001	'**'	0.01	'*'
					0.05	'.'
					0.1	' '
					1	

- Résultats de l'analyse de variance sur plan mixte

L'effet principal du facteur inter-sujets *Condition* (Nom vs Anonyme) n'est pas significatif au seuil de 5% ($F(1, 98) = 2.069, p = 0.154$). L'effet principal du facteur intra-sujet *Situation* (sélection vs singulière) est significatif au seuil de 5% ($F(1, 98) = 215.32, p < 0.001$). L'effet d'interaction entre *Condition* et *Situation* est également significatif ($F(1, 98) = 30.874, p < 0.001$). Pour pouvoir interpréter plus finement les résultats de l'expérience, nous allons maintenant analyser les effets simples.

Statistical analyse > Continous variable > Repeated mesure

- . nom Planiste
- . sélection les deux variables (selection & Singulier)
- . sélection la variable grouping (condition)

Comment interaction significatif ($p < 0,05$); On va donc observer les effet simples afin de mieux comprendre les interactions.

Analyse des effets simples

- c) Après avoir sélectionné uniquement les étudiants qui ont noté leur nom, testez si les étudiants biaisen (positivement) leur désirabilité sociale en situation de sélection en comparaison de la situation singulière.

Il faut dans un premier temps créer un sous-ensemble ne contenant que les individus ayant répondu de manière dévoilée.

Sélectionner sous ensemble

Original menu > donné > jeu de données actif > sous ensemble

Originale

Originale menu > Statistique > moyenne > T-test apparié
Comment on compare X1 < X2 il faut choisir différence > 0

- Sorties informatiques

Paired t-test

```
data: Selection and Singulier
t = 6.1722, df = 49, p-value = 6.348e-08
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.3860364      Inf
sample estimates:
mean of the differences
                  0.53
```

- Interprétation

La valeur empirique de la variable de décision du test de Student sur mesures pairees vaut $t(49) = 6.172$, $p < 0.001$. En condition dévoilée, les étudiants biaisen donc leurs résultats pour paraître plus désirables en situation de sélection.

- d) Faites de même avec les étudiants qui ont répondu anonymement.

Désirabilité en condition d'anonymat

Il faut dans un premier temps créer un sous-ensemble ne contenant que les individus ayant répondu de manière anonyme.

- Sorties informatiques

Paired t-test

```
data: Selection and Singulier
t = 15.004, df = 49, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.044593      Inf
sample estimates:
mean of the differences
                  1.176
```

- Interprétations

La valeur empirique de la variable de décision du test de Student sur mesures pairees vaut $t(49) = 15.004$, $p < 0.001$. En condition d'anonymat, les étudiants biaisen leurs résultats pour paraître plus désirables en situation de sélection.

- e) En situation de sélection, existe-t-il une différence de désirabilité sociale entre les étudiants qui ont indiqué leur nom et ceux dont l'anonymat était garanti ?

Différence entre les conditions “Nom” et “Anonyme” en situation de sélection

Il s'agit ici d'un test de Student sur deux groupes indépendants. Nous ne posons pas d'hypothèse sur le sens de la différence, nous réaliserons donc un test bilatéral.

- Sorties informatiques

Welch Two Sample t-test

```
data: Selection by Condition
t = 0.027781, df = 97.97, p-value = 0.9779
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.42259 0.43459
sample estimates:
mean in group Anonyme      mean in group Nom
            3.314                  3.308
```

- Interprétation

La valeur empirique de la variable de décision du test de Student sur deux groupes indépendants vaut $t(97.97) = 0.028$, $p = 0.978$. Dans une situation de sélection, la désirabilité sociale semble être la même, que les étudiants répondent de manière anonyme ou non.

- f) Et en situation singulière, y a-t-il une différence de désirabilité sociale entre les étudiants qui ont répondu anonymement et ceux qui ont dévoilé leur identité.

Différence entre les conditions “Nom” et “Anonyme” en situation singulière

- Sorties informatiques

Welch Two Sample t-test

```
data: Singulier by Condition
t = -2.6744, df = 97.571, p-value = 0.008778
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1.1149187 -0.1650813

sample estimates:
mean in group Anonyme      mean in group Nom
            2.138                  2.778
```

- Interprétation

La valeur empirique de la variable de décision du test de Student sur deux groupes indépendants vaut $t(97.57) = -2.674$, $p = 0.009$. Dans une situation singulière, le fait de répondre anonymement ou non influence significativement la désirabilité sociale. Les étudiants qui ne répondent pas anonymement se présentent sous un meilleur jour que ceux qui répondent anonymement.

Bilan

- g) Interprétez vos résultats.

Dans l'analyse de variance globale, l'interaction est significative. Cela veut donc dire que l'effet de la condition sur la désirabilité sociale n'est pas le même selon la situation.

En situation de sélection, les étudiants se présentent globalement de manière très positive. Ils mettent volontiers en évidence leurs qualités et estompent peut-être un peu leurs défauts. Mais dans cette situation, ils donnent la même image d'eux-mêmes, qu'ils aient à se décrire voilé ou dévoilé.

En situation singulière, les étudiants se présentent en moyenne de manière moins désirable qu'en situation de sélection. Et ici la condition a un effet sur la désirabilité sociale : les étudiants qui doivent donner leur nom, se présentent sous un jour meilleur que ceux qui peuvent se décrire anonymement.

3. régression logistique

Série 9 : régression logistique

La cote

La cote = Chance ou probabilité. En générale une cote de A est égale à ça :

$$\text{cote}(A) = \frac{p_A}{1 - p_A}.$$

Nous allons voir comment trouvé la cote, lorsqu'on connaît la probabilité (ex : Si l'on attribue au non-tabagisme une probabilité de 0,8, quelle est sa cote ?).

Supposons que la probabilité qu'un événement A survienne soit égale à p . La probabilité que cet événement A ne survienne pas est égale à $1 - p$. La cote en faveur de la survenue de l'événement A est $p/(1 - p)$. La cote contre la survenue de l'événement A est $(1 - p)/p$. La probabilité du non-tabagisme est égale à 0.8. La cote en faveur du non-tabagisme est alors de $0.8/(1 - 0.8) = 0.8/0.2 = 4$. De manière usuelle, nous dirons que la cote du non-tabagisme est de 4 : 1 (*i.e.* 4 contre 1). La cote du tabagisme est de 1 : 4 (*i.e.* 1 contre 4).

Probabilité qu'un événement survient

a) Reconstituez le tableau des effectifs observés

- $f_{ij} = \frac{nij}{n.j}$

		<i>Taux de testostérone (X)</i>	
		<i>Normal</i>	<i>Élevé</i>
<i>Casier (Y)</i>	<i>Vierge</i>	$4016 - 402 =$ 3614	345
	<i>Non vierge</i>	$4016 \times 0.1001 =$ 402	$446 \times 0.2265 =$ 101
		4016	446

- On a 2 distributions conditionnelles :
 - Y quand $X = 0$
 - Y quand $X = 1$

b) Réalisez un test du khi carré au seuil de 5%

- Statistiques → table de contingence → remplir et analyser un tri croisé
 - Pourcentages en colonne
- Pourcentages :

	<i>Normal</i>	<i>Eleve</i>
<i>Vierge</i>	90	77.4
<i>Non_vierge</i>	10	22.6
<i>Total</i>	100	100.0
<i>Count</i>	4016	446.0

Pearson's Chi-squared test

- Résultats au test du χ^2 :


```
data: .Table
X-squared = 64.078, df = 1, p-value = 1.196e-15
```
- $\chi^2(1) = 64.078, p < 0.001 \rightarrow$ on rejette H_0 . Il y a un lien significatif entre le taux de testostérone et la présence ou non d'un casier judiciaire vierge.

c) Quel est le rapport de chances ?

- On calcule le rapport de chance pour l'événement casier non vierge testostérone élevée par rapport au groupe ayant une testostérone normale (% = groupe au dénominateur)
- On peut aussi parler d'odd ration (ou OR) pour désigner le rapport des chances
- $\hat{\theta} = \frac{P_1 / (1 - p_1)}{P_0 / (1 - p_0)} = \frac{0.2265 / 0.7735}{0.1001 / 0.0999} = \frac{101 \times 3614}{345 \times 402} = 2.632$
- Comme l'OR > 1 → il y a plus de chances qu'on ait un casier judiciaire non vierge si on a un haut taux de testostérone. Cette chance est même 2.6x plus importante.

Régression logistique**1. Saisir les données**

		Patient sans chien (X = 0)	Patient avec chien (X = 1)
		29	49
Patient vivant	(Y = 0)	29	49
Patient décédé	(Y = 1)	10	4



ID	X « Chien »	Y « Décès »
1	0	0
2	0	0
...
92	1	1

2. On transforme X et Y en facteurs car ce sont des variables nominales dichotomiques

- Statistiques → gérer les variables → convertir une variable numérique en facteur
 - Niveaux : utiliser les nombres (on garde les codes 0 et 1)

a) Calculez la distribution conditionnelle de Y pour X = 0 et pour X = 1 en pourcentage

- Statistiques → table de contingence → tri croisé
 - Pourcentages en colonne (car on veut les pourcentages de X)

```

Chien
Deces  0   1
      0 29 49
      1 10  4

Column percentages:
Chien
Deces      0      1
      0    74.4  92.5
      1    25.6  7.5
Total 100.0 100.0
Count 39.0  53.0

```

Pearson's Chi-squared test

```

data: .Table
X-squared = 5.7012, df = 1, p-value = 0.01695

```

c) Déterminez la valeur du rapport des chances de l'événement $Y = 0$ pour $X = 1$ par rapport à $X = 0$. Interprétez.

- $OR = \frac{P(Y=0|X=1)}{P(Y=0|X=0)} = \frac{49}{29} = 4.224$
- Il semble donc qu'on ait 4x plus de chances d'être encore en vie une année après une crise cardiaque si on a un chien

d) Estimez les paramètres du modèle logistique qui prédit la probabilité de ne pas survivre plus d'une année après une 1^{ère} crise cardiaque selon que le patient possède ou non un chien

$$P(Y=1) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

- Régression logistique : statistique → ajustement de modèle → modèle linéaire généralisé
 - Sélectionner les variables réponse et explicative
- Résultats :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.0647    0.3667 -2.903  0.00369 **
Chien[T.1]   -1.4408    0.6363 -2.264  0.02355 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 78.469 on 91 degrees of freedom
Residual deviance: 72.765 on 90 degrees of freedom
AIC: 76.765

Number of Fisher Scoring iterations: 5

> exp(coef(GLM.1)) # Exponentiated coe
(Intercept) Chien[T.1]
0.3448276  0.2367347
```

- Lecture du tableau :
 - $\hat{\beta}_0 = -1.0647$
 - $\hat{\beta}_1 = -1.4408$
- $OR = e^{\hat{\beta}_1} = 0.2367 = \frac{1}{4.224}$ (inverse)

e) Le fait de posséder un chien a-t-il un effet significatif au seuil de 5% sur les chances de survie ? Réalisez un test de Wald

→ Le résultat apparaît dans la fenêtre de sortie :

- ```
Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.0647 0.3667 -2.903 0.00369 **
Chien[T.1] -1.4408 0.6363 -2.264 0.02355 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
- $z = -2.264, p = 0.0236 \rightarrow$  rejette  $H_0$  : la présence d'un chien a une influence significative sur le fait de survivre ou non

**Interpréter  $\hat{\beta}_1$** 

- *Signe positif: quand X augmente, la chance de voir apparaître Y (donc que  $Y = 1$ ) augmente aussi*
- *Signe négatif: quand X augmente, la probabilité que  $Y = 1$  diminue*

→ Ici le signe est négatif: quand X augmente (quand on acquiert un chien), la probabilité de  $Y = 1$  diminue (donc la probabilité de décès diminue).

**f) Selon votre modèle, quelle est la probabilité qu'un patient soit toujours en vie après une année s'il ne possède pas de chien ?**

→ On doit calculer  $P(Y = 0 | X = 0)$ :

- $P(Y = 0 | X = 0) = 1 - P(Y = 1 | X = 0) = 1 - \frac{e^{\beta_0}}{1 + e^{\beta_0}} = \frac{1}{1 + e^{\beta_0}}$
- Modèle → ajouter statistiques → valeurs ajustées → nouvelle colonne :  $P(Y = 1 | X = x)$
- $P(Y = 1 | X = 0) = 0.2564 \rightarrow P(Y = 0 | X = 0) = 1 - 0.2564 = 0.7436$

**f) Selon votre modèle, quelle est la probabilité qu'un patient soit toujours en vie après une année s'il possède un chien ?**

→ On doit calculer  $P(Y = 0 | X = 1)$

- $P(Y = 0 | X = 1) = 1 - P(Y = 1 | X = 1) = 1 - \frac{e^{\beta_0 + \beta_1}}{1 + e^{\beta_0 + \beta_1}} = \frac{1}{1 + e^{\beta_0 + \beta_1}}$
- Modèle → ajouter statistiques → valeurs ajustées → nouvelle colonne :  $P(Y = 1 | X = x)$
- $P(Y = 1 | X = 1) = 0.07547 \rightarrow P(Y = 0 | X = 1) = 1 - 0.07547 = 0.92453$

**Série 10 : régression logistique**

- Variable nominale (facteur) à 2 modalités → régression logistique
- Plugin EZR : à utiliser pour les plans complexes (analyse de variance sur mesures répétées ou sur plan mixte)

**a) Peut-on discriminer les hommes ( $Y=0$ ) et les femmes ( $Y=1$ ) en utilisant le rapport entre la longueur de l'index et celle de l'annulaire ?**

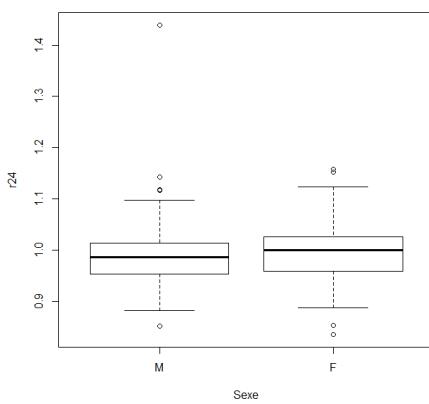
- Fitted : probabilité que ce soit une femme ou un homme

**1. Calcul du rapport en le doigt 2 et le doigt 4**

- Données → gérer → calculer une nouvelle variable
  - Nom de la nouvelle variable : r24
  - Expression à calculer : Index/Annulaire

**2. Représenter la distribution de la variable selon le sexe**

- Graphe → boîte de dispersion
  - Graphe par groupe : sexe

**3. Estimer les paramètres du modèle**

- Statistiques → ajustement de modèle → modèle linéaire généralisé
  - Sexe ~ r24

```
Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.4748 2.4587 0.600 0.549
r24 -0.3673 2.4701 -0.149 0.882
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 262.07 on 233 degrees of freedom
Residual deviance: 262.05 on 232 degrees of freedom
AIC: 266.05

Number of Fisher Scoring iterations: 4
```

- Paramètres :
  - $\hat{\beta}_0 = 1.475$
  - $\hat{\beta}_1 = -0.367$

- Le modèle s'écrit :  $P(\text{sexe} = 1 | r24) = \frac{\exp(1.475 - 0.367 \times r24)}{1 + \exp(1.475 - 0.367 \times r24)}$

→ Plus la variable X augmente, plus la probabilité que  $Y = 1$  (qu'on soit une femme) diminue

#### 4. Test du rapport des vraisemblances

- $GLM.0 : Sexe \sim 1$
- $GLM.1 : Sexe \sim r24$

→ Comparaison des 2 modèles :

Analysis of Deviance Table

| Model 1: Sexe ~ 1   |        |    |        |    |                   |
|---------------------|--------|----|--------|----|-------------------|
| Model 2: Sexe ~ r24 |        |    |        |    |                   |
|                     | Resid. | Df | Resid. | Df | Deviance Pr(>Chi) |
| 1                   | 233    |    | 262.07 |    |                   |
| 2                   | 232    |    | 262.05 | 1  | 0.021965 0.8822   |

- Résultat :  $X^2(1) = 0.022, p = 0.882 \rightarrow$  on accepte  $H_0$ . Cela signifie que le rapport entre les doigts 2 et 4 ne permet pas de discriminer les hommes des femmes

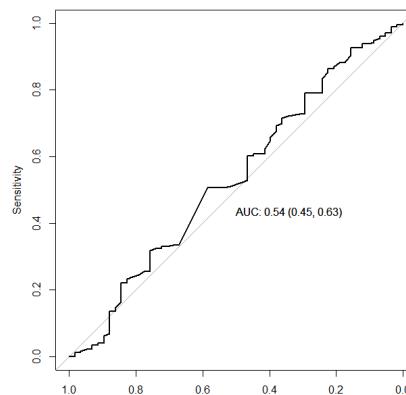
#### 5. Mesure de la force de liaison entre X et Y

- $R^2_L = \frac{\text{Différence de déviance}}{\text{Déviance du modèle le plus mauvais}} = \frac{0.022}{262.05} = 0.000084$

→ La force de liaison entre le sexe et le rapport des longueurs de l'index et de l'annulaire est très faible

#### 6. Calcul du pouvoir discriminant du modèle

- Roc → pRoc → Plot ROC curve...
  - Onglet « plot » : cocher la case AUC (c'est l'aire sous la courbe)



b) Peut-on discriminer les hommes ( $Y=0$ ) et les femmes ( $Y=1$ ) en utilisant simplement la longueur de leur majeur ?

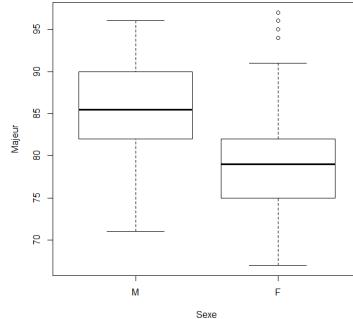
#### 1. Recodage des variables

- Variable sexe :
  - Niveau de référence : Hommes ( $Y=0$ )
- Réordonner les niveaux d'un facteur
  - $M = 1$  (niveau de référence)
  - $F = 2$
- Définir les contrastes d'un facteur → contraste traitement

## 2. Représentation graphique des données

Comment la longueur des majeurs se distribue dans les 2 groupes ?

- Graphes → boites de dispersion
  - Graphe par groupe
- Analyse :
  - Les hommes ont un majeur plus long (médiane = env. 85) que les femmes (médiane = env. 79)
  - La distribution est symétrique dans les 2 groupes



## 3. Estimer les paramètres du modèle

- Régression logistique simple : statistiques → ajustement de modèle → modèle linéaire généralisé
  - « Sexe ~ Majeur »

```
Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) 17.62557 2.73254 6.450 1.12e-10 ***
Majeur -0.20132 0.03281 -6.135 8.52e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 262.07 on 233 degrees of freedom
Residual deviance: 210.62 on 232 degrees of freedom
AIC: 214.62
```

- Paramètres :
  - $\hat{\beta}_0 = 17.626$
  - $\hat{\beta}_1 = -0.201$

→ Plus la variable X augmente, plus la probabilité que  $Y = 1$  (qu'on soit une femme) diminue

## 4. Détermination du rapport des chances

- $OR = e^{\hat{\beta}_1} = 0.818 = \frac{1}{e^{-\hat{\beta}_1}} = \frac{1}{1.223}$

→ Quand la longueur du majeur augmente d'une unité, la chance d'être une femme diminue d'un facteur 1.223.

De manière développée

$$e^{\hat{\beta}_1} = \frac{P(Y=1 | X=x+1)}{P(Y=0 | X=x+1)} \div \frac{P(Y=1 | X=x)}{P(Y=0 | X=x)}$$

## 5. Calcul d'une probabilité

- Que vaut  $P(Y=1 | X=80)$  ?

$$P(Y=1 | X=80) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}}$$

- $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 x = 17.626 - 0.201 \times 80 = 1.546$

$$\rightarrow P(Y=1 | X=80) = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}} = \frac{4.693}{1 + 4.693} = 0.82 = 82\%$$

### 6. Détermination du seuil

Quelle est la valeur à partir de laquelle on peut dire qu'un individu est une femme ou un homme ?

- On doit arriver à  $\frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}} = \frac{1}{2}$
- Donc,  $e^{\hat{\eta}}$  doit être = à 1  $\rightarrow \hat{\eta}$  doit être = à 0

$$\rightarrow \hat{\beta}_0 + \hat{\beta}_1 x_{critique} = 0 \rightarrow x_{critique} = \frac{-\hat{\beta}_0}{\hat{\beta}_1} = \frac{17.626}{0.201} = 97.7$$

$\rightarrow$  On considère que les individus ayant des majeurs < 87.7 sont des femmes

### 7. Test du rapport des vraisemblances

Le test du rapport des vraisemblances est un test global qui permet de déterminer si notre modèle est significatif

- Statistiques  $\rightarrow$  ajustement de modèle  $\rightarrow$  modèle linéaire généralisé
  - GLM.0 : Sexe ~ 1 (hypothèse nulle)
  - GLM.1 : modèle construit précédemment
- On compare les 2 modèles :

```
Analysis of Deviance Table

Model 1: Sexe ~ 1
Model 2: Sexe ~ Majeur
 Resid. Df Resid. Dev Df Deviance Pr(>Chi)
 1 233 262.07
 2 232 210.62 1 51.447 7.354e-13 ***

Signif. codes: 0 '****' 0.001 '***' 0.01 '**' 0.05
```

Analyse du tableau

- On doit calculer la différence de déviance pour savoir si le modèle est significatif  $\rightarrow 51.447$
- $X^2(1) = 51.447, p < 0.001 \rightarrow$  on rejette  $H_0$ , le modèle construit est significatif et meilleur que le modèle vide

### 8. Mesure de la force de liaison entre X et Y

On doit calculer le pseudo  $R^2$  de Mc Fadden

$$\bullet R^2_L = \frac{\text{Différence de déviance}}{\text{Déviance du modèle le plus mauvais}} = \frac{51.447}{262.07} = 0.196$$

### 9. Test de Hosmer-Lemeschow

- Chargement du Plugin ROC
- ROC  $\rightarrow$  pROC  $\rightarrow$  Hosmer-Lemeschow test (sur GLM.1)
  - Number of bins : nombre de classes qu'on va créer  $\rightarrow$  on choisit de créer 8 classes (ce sera précisé à l'examen)

Hosmer and Lemeshow goodness of fit (GOF) test

```
data: .matrix[, 1], .matrix[, 2]
X-squared = 5.0524, df = 6, p-value = 0.5371
```

- **Résultat :**
  - $Df = \text{nombre de classes} - 2$  (ici :  $8-2 = 6$ )
  - $X^2 (6) = 5.052, p = 0.537$

→ On ne rejette pas  $H_0$ , ce qui est positif, car ça signifie qu'il y a un bon ajustement entre le modèle et les données (la distance qui sépare le tableau des effectifs observés du tableau des effectifs théoriques est suffisamment petite) → c'est un bon modèle.

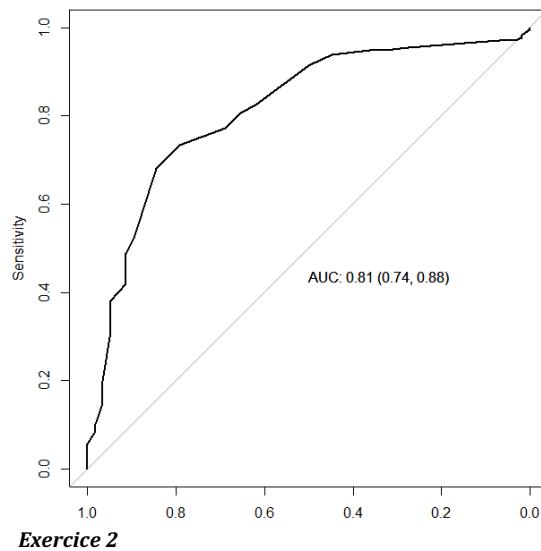
### 10. Calcul du pouvoir discriminant

On peut calculer le pouvoir discriminant du modèle en déterminant l'aire sous la courbe ROC

- On interprète l'aire comme le pouvoir discriminant (c'est l'estimation d'une probabilité)
- On extrait un individu  $j$  du groupe des hommes et un individu  $i$  du groupe des femmes → la paire sera concordante si la probabilité associée à l'homme est plus grande que la probabilité associée à la femme

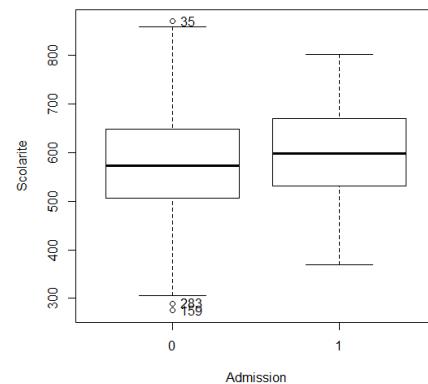
→ Alors quelle est la probabilité que la paire soit concordante ? C'est ce que nous indique l'aire sous la courbe ROC

- Roc → pRoc → Plot ROC curve...
  - Onglet « plot » : cocher la case AUC (c'est l'aire sous la courbe)



a) Calculez la variable Scolarité puis représentez graphiquement cette nouvelle variable selon l'admission à l'université

- Données → gérer → calculer nouvelle variable
  - Nom de la variable : Scolarité
  - Expression :  $(\text{College} + \text{Lycee})/2$
- Représentation graphique :
  - Transformation de la variable admission en un facteur
  - Graphe → boîte à moustaches → graphe par groupe



**b) Transformer la variable « rang » en un facteur en choisissant la modalité 0 comme référence, puis estimez les paramètres du modèle établi par le cabinet**

1. Transformation de la variable « rang »

- Statistiques → gérer → convertir des variables numériques en facteurs
  - Garder les nombres
- Statistiques → gérer → réordonner les niveaux d'un facteur
  - $0 = 1$
  - $1 = 2$
- Contraste traitement

2. Estimation des paramètres du modèle

- Statistiques → ajustement de modèle → modèle linéaire généralisé
  - Admission ~ Scolarité + Rang + Examen

```
Coefficients:
 Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.633993 1.077451 -5.229 1.7e-07 ***
Scolarite 0.002439 0.001106 2.205 0.027458 *
Rang[T.1] 0.968227 0.232421 4.166 3.1e-05 ***
Examen 0.840162 0.222995 3.768 0.000165 ***

Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 499.98 on 399 degrees of freedom
Residual deviance: 462.80 on 396 degrees of freedom
AIC: 470.8

Number of Fisher Scoring iterations: 4
```

• Paramètres :

- $\hat{\beta}_0 = -5.634$
- $\hat{\beta}_1 = 0.002$
- $\hat{\beta}_2 = 0.968$
- $\hat{\beta}_3 = 0.840$

• Le modèle s'écrit

$$P(\text{Admission} = 1 | \text{Scolarité}, \text{Examen}, \text{Rang}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \text{scolarité} + \hat{\beta}_2 \text{rang} + \hat{\beta}_3 \text{examen})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \text{scolarité} + \hat{\beta}_2 \text{rang} + \hat{\beta}_3 \text{examen})}$$

• Analyse :

- Plus la moyenne des scores obtenus au collège et au lycée est grande, plus la probabilité d'admission est grande
- Plus la réputation de l'école est faible, plus la probabilité d'admission est grande
- Plus les scores obtenus aux examens de fin d'année sont grands, plus la probabilité d'admission est grande

c) Réalisez le test de vraisemblance

- GLM.0 : Admission ~ 1
- GLM.1 : Admission ~ Scolarité + Rang + Examen

Comparaison des 2 modèles :

```
Analysis of Deviance Table

Model 1: Admission ~ 1
Model 2: Admission ~ Scolarite + Rang + Examen
 Resid. Df Resid. Dev Df Deviance Pr(>Chi)
 1 399 499.98
 2 396 462.80 3 37.178 4.22e-08 ***

```

- $X^2(3) = 37.178, p < 0.05 \rightarrow$  on rejette  $H_0$ , le modèle construit est significatif

d) Donnez l'équation du modèle logistique ajusté avec les coefficients estimés pour les écoles ayant mauvaise réputation (Rang = 0)

$$\begin{aligned} P(Admission = 1 | Scolarité, Examen, Rang = 0) &= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \text{scolarité} + \hat{\beta}_2 \text{rang} + \hat{\beta}_3 \text{examen})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \text{scolarité} + \hat{\beta}_2 \text{rang} + \hat{\beta}_3 \text{examen})} \\ &= \frac{\exp(-5.634 + 0.002 \text{scolarité} + 0.968 \times 0 + 0.84 \text{examen})}{1 + \exp(-5.634 + 0.002 \text{scolarité} + 0.968 \times 0 + 0.84 \text{examen})} \end{aligned}$$

e) Evaluatez l'adéquation entre le modèle étudié et les données à l'aide du test d'Hosmer-Lemeshow

```
Hosmer and Lemeshow goodness of fit (GOF) test
```

```
data: .matrix[, 1], .matrix[, 2]
X-squared = 8.6725, df = 8, p-value = 0.3707
```

→ On accepte  $H_0$ , l'ajustement est bon !

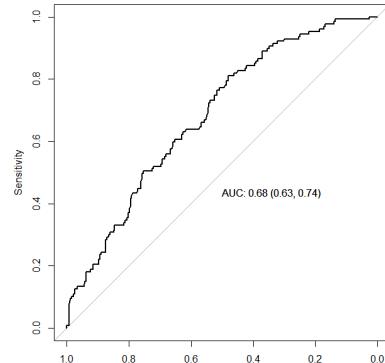
f) Mesurez la force de liaison entre les variables explicatives et le critère en recourant au  $R^2$  de Mc Fadden

- $R^2_L = 37.178 \div 499.98 = 0.074$

→ La part de déviance expliquée par les variables explicatives = 7.4%. Ce n'est pas beaucoup

g) Quel est le pouvoir discriminant du modèle ?

- ROC → pROC → Plot curve
- Comme  $0.6 \leq AUC < 0.7$ , la discrimination est faible



h) Score de 300 au collège, de 200 au lycée, et de 5 à l'examen final, venant d'une école de mauvaise réputation

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$$

- $P(Y=1 | X1 = 250, X2 = 0, X3 = 5) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 2 + \hat{\beta}_3 \times 3}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \times 1 + \hat{\beta}_2 \times 2 + \hat{\beta}_3 \times 3}}$
- $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 = -5.634 + 0.002 \times 250 + 0.968 \times 0 + 0.840 \times 5 = -0.934$
- $P(Y=1 | X1 = 250, X2 = 0, X3 = 5) = \frac{e^{\hat{\eta}}}{1 + e^{\hat{\eta}}} = 0.282$

→ La probabilité d'admission vaut env. 28%. Comme cette probabilité est inférieure à 50%, nous déconseillons au décanat d'admettre ce demandeur à l'université.