

Introduction aux Statistiques et à l'utilisation du logiciel R

Christophe Lalanne* Christophe Pallier†

1 Introduction

2 Comparaisons de deux moyennes

2.1 Objet de l'étude

On a mesuré le temps de sommeil (en heure) de 10 sujets après administration d'un certain médicament (numéroté 1 et 2). Chaque sujet teste les 2 médicaments l'un après l'autre. On s'intéresse à l'effet du type de médicament sur la durée de sommeil; en particulier, on cherche à savoir si les sujets dorment plus après avoir pris le médicament 2 que lorsqu'ils prennent le médicament 1.

Les données sont contenues dans le fichier `student.dat`, et sont organisées sous forme tabulaire, avec un en-tête et 3 colonnes, correspondant respectivement au n° de sujet, à la durée de sommeil avec le médicament 1, et à la durée de sommeil avec le médicament 2.

2.2 Analyse

On se situe ici dans le cadre d'échantillons appariés : les observations ne sont pas indépendantes entre les deux conditions. Cela a pour principale conséquence que l'on doit tenir compte de cette covariance dans l'estimation de la variance de population (celle relative à l'ensemble formé de la différence observée entre les deux conditions), contrairement au cas précédent où les observations étaient indépendantes (entre et à l'intérieur des groupes).

Avant de procéder à l'analyse inférentielle, il convient de regarder un peu ce que nous disent les données brutes : c'est l'objet de la statistique descriptive. On peut par exemple produire un résumé numérique, à l'aide des commandes suivantes :

*christophe.lalanne@gmx.net

†www.pallier.org

```

> data<-read.table(data2.dat, header=T)
> data
  SUJET  M1  M2
1    S1 5.7 6.9
2    S2 3.4 5.8
3    S3 4.8 6.1
4    S4 3.8 5.1
5    S5 4.9 4.9
6    S6 8.4 9.4
7    S7 8.7 10.5
8    S8 5.8 6.6
9    S9 5.0 9.6
10   S10 7.0 8.4
> summary(data$M1)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
3.400  4.825  5.350  5.750  6.700  8.700
> summary(data$M2)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.900  5.875  6.750  7.330  9.150 10.500

```

La commande `summary()` affiche les principaux indicateurs de position et d'étendue des valeurs observées. Il y manque cependant l'écart-type que l'on peut évaluer comme suit :

```

> sd(data$M1); sd(data$M2)
[1] 1.789010
[1] 2.002249

```

On voit que les sujets dorment en moyenne un peu moins avec le premier médicament (5.75 h) en comparaison du deuxième médicament (7.33 h), ce que l'on observe également sur le graphique en boîte à moustaches suivant :

```

> boxplot(data$M1,data$M2,col="gray",main="n=10",ylab="Temps de sommeil
(h)",xlab="Médicament",names=c("1","2"))

```

La variabilité est comparable entre les deux ensembles de réponse, et il n'y a pas de valeurs atypiques. La distribution des réponses mesurées semblent relativement symétrique. Comme les données sont appariées, il peut être intéressant de représenter les données sous forme d'un nuage bi-varié, dans lequel on fait apparaître la droite d'équation $y = x$:

```

> plot(data$M1,data$M2,pch=16,main="n=10",xlim=c(2,12),ylim=c
(2,12),xlab="Médicament 1",ylab="Médicament 2")
> abline(0,1)

```

Pour tester l'existence d'une différence entre les deux conditions au niveau de la population parente, on utilise le test t de Student (pour échantillons appariés) :

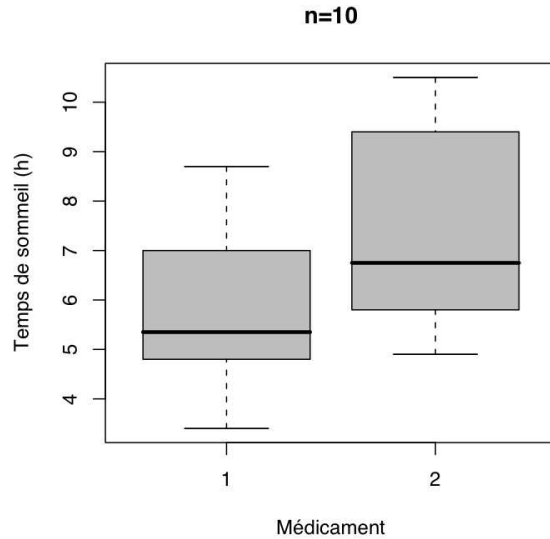


FIG. 1: Effet du type de médicament sur le temps de sommeil chez les 10 sujets.

```
> (out <- t.test(data$M1, data$M2, alternative="less",
  var.equal=T, paired=T))
```

Paired t-test

```
data: data$M1 and data$M2
t = -4.0621, df = 9, p-value = 0.001416
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
 -Inf -0.8669947
sample estimates:
mean of the differences
 -1.58
```

On pourrait vérifier la normalité des distributions, mais comme l'examen des données ne révèle pas d'asymétries flagrantes, ce n'est pas utile. On pourrait de même tester l'homogénéité des variances à l'aide d'un test de xx ? Celle-ci étant supposée vérifiée, on spécifie l'option `var.equal=T`. Le cas échéant, R calculerait un test t corrigé par la méthode de Welch. Ici, comme on cherche à savoir si le médicament 2 entraîne des durées de sommeil plus importantes que le médicament 1, on a spécifié une hypothèse alternative orientée ($\mu_1 - \mu_2 < 0$), d'où l'option `alternative="less"`. Les résultats indiquent que le test est significatif : la statistique de test vaut -4.06 et la probabilité d'observer un échantillon au moins aussi extrême sous H_0 est de $.0014$ (largement inférieure au seuil conventionnel de 5%).

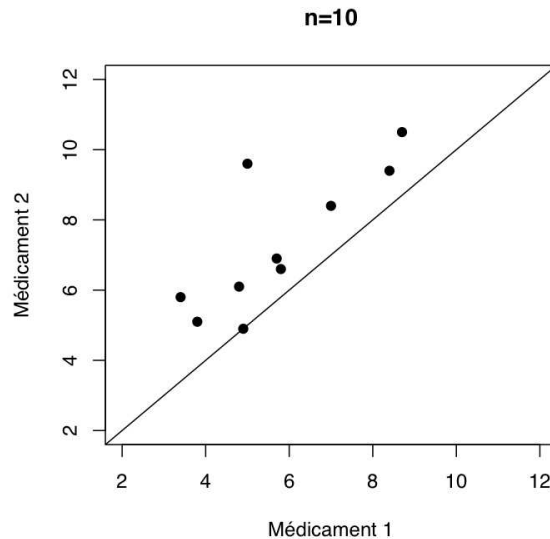


FIG. 2: Relation entre les temps de sommeil en fonction du type de médicament chez les 10 sujets.

L'intervalle de confiance à 95% pour la moyenne estimée est indiqué et correspond à l'intervalle $[-2.46; -0.70]$.

On peut enfin évaluer la puissance de ce test pour les paramètres considérés. Pour cela, il nous faut calculer l'écart-type estimé sd . Comme celui-ci sert à calculer la statistique de test utilisée, on peut l'obtenir directement puisque l'on connaît la valeur de t (`out$statistic`), la différence de moyennes (`out$estimate`) et la taille de l'échantillon ($n=10$) :

```
> cat((unlist(out$estimate)/unlist(out$statistic))*sqrt(10))
1.229995
```

On peut également le calculer tout simplement comme l'écart-type associé à la différence des moyennes :

```
> diff_mean <- data$M1-data$M2
> sd(diff_mean)
[1] 1.229995
```

A présent, on peut appeler la fonction `power.t.test()` avec tous les arguments requis, sauf `power` puisque c'est la valeur que l'on veut évaluer :

```
> power.t.test(n=10,sd=1.23,delta=1.58,type="paired",
  alt="one.sided",sig.level=0.05)
```

```
Paired t test power calculation
```

```

n = 10
delta = 1.58
sd = 1.23
sig.level = 0.05
power = 0.9816248
alternative = one.sided

```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

On peut illustrer l'extrémalité de la statistique t calculée à l'aide du schéma suivant, où l'on a tracé la densité de la loi de t à 9 degrés de liberté, la zone hachurée correspondant à la zone de rejet de H_0 (la zone couvre 2.5% de la surface totale sous la courbe) :

```

> x<-seq(-4.5,4.5,by=0.05)
> plot(x,dt(x,df=9),type="l")
> seuil <- qt(0.05,df=9)
> xx <- c(seq(-4.5,seuil,by=.05),rev(seq(-4.5,seuil,by=.05)))
> yy <- c(rep(0,54),dt(xx[55:108],df=9))
> polygon(xx, yy, col="gray")
> abline(h=0)
> abline(v=out$statistic, col="red")
# ou abline(v=qt(1-unlist(out$p.value),df=9))
> title(expression(p(t[obs]>t[alpha] / H[0])==0.05))

```

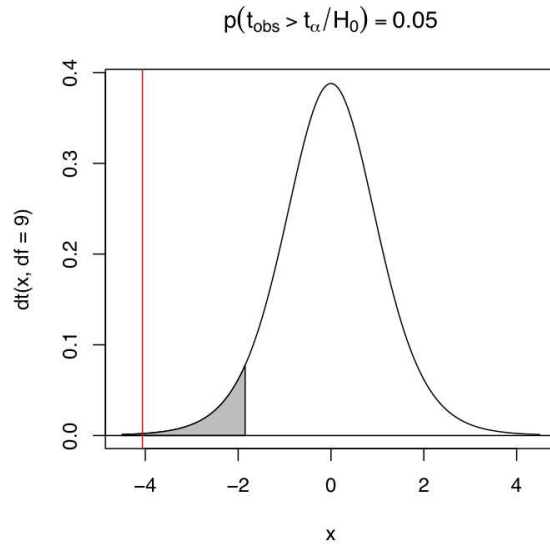


FIG. 3: Distribution de la statistique de test calculée sous H_0 .

2.3 Remarques

- Si on avait fait comme si les échantillons étaient indépendants, les valeurs de t et de p s’en trouveraient sensiblement modifiées ($t = -1.8606$ et $p = 0.03959$). Cela s’explique par le fait que dans ce cas, en ignorant la covariance due à l’appariement, la valeur de la variance estimée augmente (beaucoup plus que la racine carrée de la taille de l’effectif, ici passé à 20) et par conséquent la valeur de t diminue. L’avantage des échantillons appariés est que l’on contrôle une source de variabilité (la variance intra-sujet dans ce cas), et que l’on peut en tenir compte dans l’estimation de la variance de population.
- D’autre part, si on n’avait pas eu d’hypothèse a priori sur le sens de la différence entre les 2 moyennes (hypothèse alternative stochastique), le test aurait été bilatéral et dans ce cas, il aurait toujours été significatif pour des échantillons appariés ($p = 0.002833$), mais il ne l’aurait pas été dans le cas d’échantillons indépendants ($p = 0.07919$). En effet, avec un test bilatéral, les seuils critiques se réfèrent à des probabilités associées de 0.025, et non pas 0.05 comme c’est le cas dans un test unilatéral (cas d’une hypothèse alternative orientée), et sont par conséquent plus importants (± 2.262157).

3 ANOVA

L’ANOVA (“ANalysis Of VAriance”) constitue une extension du test t lorsqu’il n’y a qu’une seule variable qualitative, ou facteur dans la terminologie de R. Si c’est un facteur de groupe (ou de classification, ou « emboîtant »), alors il s’agit d’une extension du test t pour échantillon indépendant ; si le facteur comporte des niveaux répétés sur les sujets (facteur « croisé »), alors c’est une extension du test t pour échantillons appariés. Dans le cas où il y a plus d’un facteur, plusieurs cas de figures sont envisageables, selon le type (nominale ou ordinale) et le rôle au sein de la structure du plan expérimental (facteur d’emboîtement ou de croisement) des variables impliquées (Tab. 1).

Les principales commandes R utiles dans les analyses de variance sont les suivantes :

```
aov() effects() tapply() pairwise.t.test() conf.int() contr()
plot(aov()) boxplot() interaction.plot()
```

3.1 ANOVA à un facteur

3.1.1 Objet de l’étude

Pour ce premier exemple d’ANOVA, nous allons reprendre les données de Howell (chap. 11, pp. 340-341). Il s’agit d’une étude visant à étudier les

plan	formule R
$S < G_k >$	<code>aov(y~g)</code>
$S * A_n$	<code>aov(y~a+Error(subj/a))</code>
$S < G_k > * A_n$	<code>aov(y~g*a+Error(subj/a))</code>
$S * A_n * B_m * E_k$	<code>aov(y~a*b+Error(subj/(a*b)))</code>

TAB. 1: Exemples de plans expérimentaux traités par ANOVA, avec les formules correspondantes sous R. Les variables de groupe sont notées G , les facteurs croisés A et/ou B , et le facteur sujet est noté S . La variable dépendante, ou mesurée, est supposée être y sous R, et `subj` désigne le vecteur sujet.

capacités de mémorisation d'un matériel verbal de sujets âgés de 55 à 65 ans en fonction du niveau de traitement imposé préalablement sur ce matériel.

Les données sont contenues dans le fichier `eysenck1974.dat`; celui-ci comprend l'ensemble des scores (nombre de mots rappelés en fonction du niveau de traitement), et ils sont délimités par une virgule.

3.1.2 Analyse

Les principales étapes de l'analyse sont résumées dans les points suivants :

1. résumé numérique et graphique
2. vérification (graphique ou à l'aide des tests appropriés) des conditions d'application de l'ANOVA : normalité des résidus et homoscedasticité
3. test de l'ANOVA et calcul des $IC_{95\%}$
4. comparaisons multiples
5. conclusions (descriptive et inférentielle)

Il faut dans un premier temps importer le fichier de données, créer un `data.frame` croisant les scores avec les 5 modalités du groupe d'apprentissage.

```
> a <- read.table('eysenck1974.dat', sep=",")
> g <- gl(5,10,50, label=c("Addition", "Rimes", "Adjectifs", "Images",
  "Intentionnel"))
> data <- data.frame(t(a), g)
> colnames(data) <- c("score", "group")
> attach(data)
```

On procède ensuite à une simple étude descriptive des données : il faut étudier la distribution des données par groupe, pour vérifier l'éventuelle présence de données atypiques, comparer les moyennes de 5 groupes et l'homogénéité des variances intra-groupes (homoscedasticité).

```

> table(data)
> summary(data)
      score      group
Min.   : 3.00  Addition   :10
1st Qu.: 7.00  Rimes       :10
Median :10.00  Adjectifs   :10
Mean   :10.06  Images      :10
3rd Qu.:11.75  Intentionnel:10
Max.   :23.00

> stripchart(score~group,vertical=T,pch='x',method="jitter",
             xlab="Niveau de traitement",ylab="Nombre de mots rappelés")
> abline(h=mean(score),lty=2)
> bp <- boxplot(score~group,col="lightgrey")
> bp$stats
> gmean <- tapply(score,group,mean)
> gvar <- tapply(score,group,var)
> gsd <- tapply(score,group,sd)
> matrix(round(c(gmean,gsd,gvar),digits=2),nrow=3,byrow=T,
         dimnames=list(c("mean","sd","var"),levels(group)))
      Addition Rimes Adjectifs Images Intentionnel
mean    7.00  6.90    11.00  13.40     12.00
sd     1.83  2.13     2.49   4.50     3.74
var    3.33  4.54     6.22  20.27    14.00

```

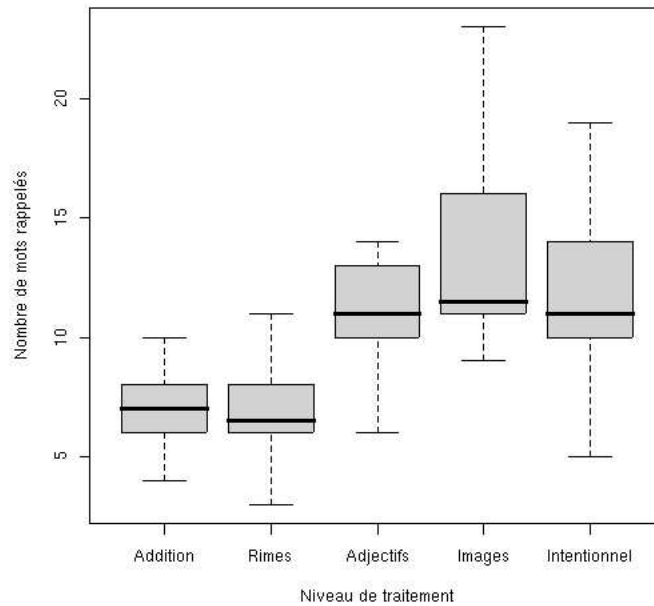


FIG. 4: Distribution des scores par niveau de traitement.

Après vérification des conditions d'application de l'ANOVA, il ne nous reste plus qu'à produire le tableau d'ANOVA et calculer les intervalles de confiance associés à chaque moyenne de groupe.

```
> model <- aov(score~group)
> plot(model)
> anova(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	4	351.52	87.88	9.08	0.0000
Residuals	45	435.30	9.67		

TAB. 2: Tableau d'ANOVA pour l'expérience de Eysenck (1974).

3.2 ANOVA à deux facteurs

3.2.1 Objet de l'étude

Il s'agit des mêmes données de Howell (chap. 13, pp. 454-456), avec une distinction entre sujets jeunes et sujets âgés (facteur âge supplémentaire).

Les données sont contenues dans le fichier `eysenck1974-2.dat` : celui-ci comprend l'ensemble des scores (nombre de mots rappelés en fonction du niveau de traitement), en fonction de l'âge et de la condition expérimentale. Les données sont cette fois-ci organisées par colonnes, les codes des facteurs utilisés sont les suivants :

- 1ère colonne : âge, 1=sujets âgés, 2=sujets jeunes
- 2ème colonne : condition, 1=Addition, 2=Rimes, etc.
- 3ème colonne : score

3.2.2 Analyse

Les principales étapes de l'analyse sont résumées dans les points suivants :

1. résumé numérique et graphique
2. vérification (graphique ou à l'aide des tests appropriés) des conditions d'application de l'ANOVA : normalité des résidus et homoscédasticité
3. test de l'ANOVA : étude de l'effet d'interaction et des effets simples
4. intervalles de confiance pour les moyennes
5. comparaisons multiples
6. conclusions (descriptive et inférentielle)

Commençons par importer les données et créer une structure de données pour l'analyse.

```

> a <- read.table('eysenck1974-2.dat')
> data <- data.frame(factor(a$V1),factor(a$V2),a$V3)
> colnames(data) <- c("age","cond","score")
> attach(data)

```

On procède ensuite, comme dans le cas de l'ANOVA à un facteur, à la visualisation graphique des données et à l'examen des données numériques, résumées par condition. Cette fois-ci, on est en présence de deux facteurs de groupe (ou de classification), donc il sera intéressant de regarder à la fois les données par groupe, mais également les données résultant du croisement des modalités de chacun des facteurs.

```

> summary(data)
> tapply(score,age,mean)
> tapply(score,cond,mean)
> tapply(score,list(age,cond),mean)
  1  2  3  4  5
1 7.0 6.9 11.0 13.4 12.0
2 6.5 7.6 14.8 17.6 19.3
> boxplot(score~age*cond,border=c("red","blue"),
           xlab="Condition",ylab="Score")
> library(lattice)
> bwplot(score~age|cond)

```

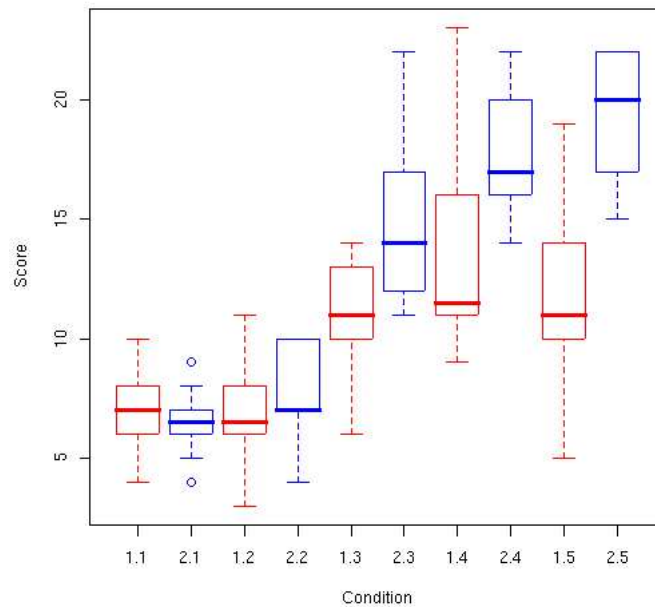


FIG. 5: Distribution des scores par niveau de traitement.

Passons à l'analyse de variance proprement dit, ainsi qu'aux comparaisons multiples :

```

> model1 <- aov(score~age*cond)
> summary(model1)
> effects <- model.tables(model1,se=T)

```

Les résultats de l'ANOVA (Tab. 3) indiquent que l'interaction $\text{age} \times \text{cond}$ est significative, ce qui signifie que l'on ne peut pas interpréter les effets principaux de l'âge et du niveau de traitement indépendamment l'un de l'autre. En d'autres termes, l'effet de l'âge n'est pas le même selon le niveau de traitement considéré. L'interprétation seule des effets principaux n'a aucun sens. Pour mieux visualiser cet effet croisant les deux facteurs, on peut produire un graphique d'interaction (Fig. 6) avec, par exemple, les commandes suivantes :

```

> interaction.plot(age,cond,score,xlab="Age",ylab="Score")
> xx <- 1.5
> yy <- mean(score)
> lwb <- yy-effects$se[3]
> upb <- yy+effects$se[3]
> points(xx,yy,pch=16)
> arrows(xx,lwb,xx,upb,length=.05,angle=90,code=3)

```

Les comparaisons multiples (Fig. 7) peuvent être visualisées comme suit :

```

> hsd1 <- TukeyHSD(model1,which="cond")
> hsd1
> plot(hsd1)
> hsd2 <- TukeyHSD(model1)
> plot(hsd2)

```

Une solution alternative pour l'estimation des IC associés aux comparaisons multiples est d'utiliser la fonction `simint` (`multcomp`)¹ :

```

> library(multcomp)
> mc <- simint(score~cond,type="Tukey")
> plot(mc)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	240.25	240.25	29.94	0.0000
cond	4	1514.94	378.74	47.19	0.0000
age :cond	4	190.30	47.58	5.93	0.0003
Residuals	90	722.30	8.03		

TAB. 3: Tableau d'ANOVA pour l'expérience complète de Eysenck (1974).

¹Le temps de calcul est un peu plus long.

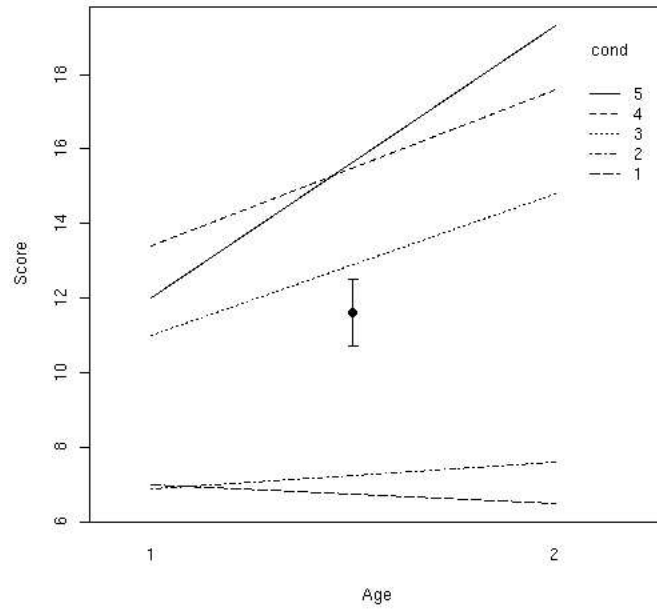


FIG. 6: Graphique d'interaction.

3.3 ANOVA sur mesures répétées

4 Mesures d'association et de dépendance

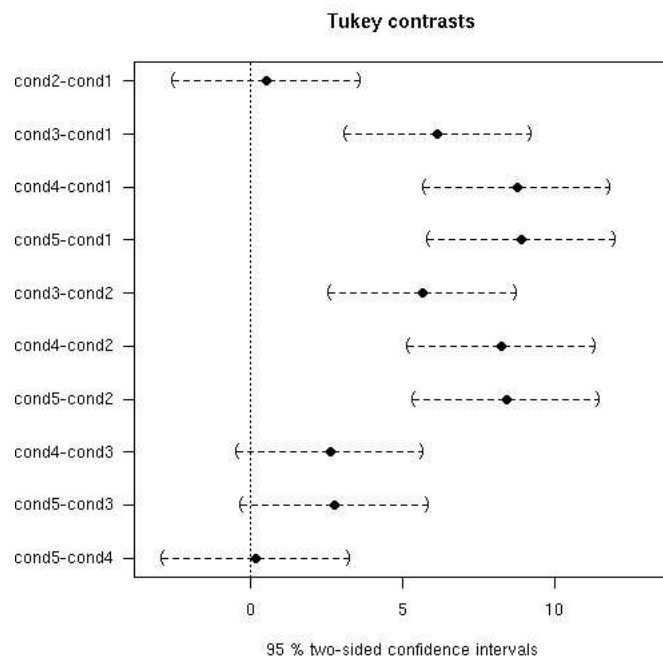


FIG. 7: Comparaisons multiples par la méthode HSD de Tukey.

Références

1. Laboratoire de Biostatistiques de Lyon, pbil.univ-lyon1.fr/

Table des matières

1	Introduction	1
2	Comparaisons de deux moyennes	1
2.1	Objet de l'étude	1
2.2	Analyse	1
2.3	Remarques	5
3	ANOVA	6
3.1	ANOVA à un facteur	7
3.1.1	Objet de l'étude	7
3.1.2	Analyse	7
3.2	ANOVA à deux facteurs	9
3.2.1	Objet de l'étude	9
3.2.2	Analyse	9
3.3	ANOVA sur mesures répétées	11
4	Mesures d'association et de dépendance	11

Table des figures

1	Effet du type de médicament sur le temps de sommeil chez les 10 sujets.	3
2	Relation entre les temps de sommeil en fonction du type de médicament chez les 10 sujets.	4
3	Distribution de la statistique de test calculée sous H_0	5
4	Distribution des scores par niveau de traitement.	8
5	Distribution des scores par niveau de traitement.	10
6	Graphique d'interaction.	12
7	Comparaisons multiples par la méthode HSD de Tukey.	12