One-way ANOVA (suite)
Two-way ANOVA & interactions

10/12/2021

Recap semaine derniere

Variable independante qualitative + Variable dependante quantitative

Test d'hypothese: y a t'il une difference significative entre les donnees observees?

Correlation: est ce que r est different de 0? cor.test(x, y, method= "pearson")

Student t-test: est ce que deux moyennes sont differentes?

- t.test(x~y, dataset, paired=FALSE) si les deux moyennes sont basees sur des donnees independante (i.e., between subject design)
- t.test(x~y, dataset, paired=TRUE) si les deux moyennes sont basees sur des donnees dependantes (i.e., within subject design).

Ouvrir R

Ouvrez RStudio

Placez vous dans le Project du TD si vous en aviez cree un

Ouvrez un nouveau script.

Pro tip: n'oubliez pas de passer a la ligne entre les instructions sur R, ca rend le fichier plus lisible + utilisez les commentaires pour structurer

```
data<-read.csv(data, header=TRUE)
mean(data$column)
ggplot(data, (x,y))+
  geom_point(aes(color=z))
aggregate(x~y, data, FUN="mean")
```

```
#loading data
data<-read.csv(data, header=TRUE)

#calculating some means
mean(data$column)
aggregate(x~y, data, FUN="mean")

#plotting
ggplot(data, (x,y))+
  geom_point(aes(color=z))
```

Idee de paired vs independent t.test

Imaginez qu'on veuille connaître l'effet d'une nouvelle methode d'apprentissage de la lecture. Deux designs possibles:

1) Un groupe avec la methode d'apprentissage A1 et l'autre groupe avec la methode A2 et de comparer les deux (between subject)

2) Un echantilon avec mesure repetees.

Test d’hypothese et type de variable

On a vu comment analyser la relation entre deux variables continues (correlation) ET une variable categorique a deux niveaux + une variable continue (t-test)

Et si on a plus de deux niveaux a notre variable categorique?? Exemple:

- Monolingue vs. bilingue **vs. trilingue** et temps de reaction en DL
- Treatment 1 vs. Treatment 1 + Treament 2 vs. Treament 2 vs. Placebo et evolution des symptomes
- ⇒ Groupe A vs Groupe C vs Groupe B vs …. et une mesure

One-way ANOVA

Permet de comparer trois groupes ou plus sur une seule variable dependante continue.

Dans chronolex:

1) On va traiter Nlett (i.e., nombre de lettre dans les mots) comme un facteur.
    a) Comment faire pour transformer Nlett en facteur?
2) Utiliser summary() pour voir la distribution du nombre de mots dans chaque categories.
    a) Est ce que les samples size sont bien reparti?
3) Cree un subset de chronolex qui ne contient que les mots de 4, 5 et 6 lettres.
    a) Comment faire?? (pas vraiment vu en cours, des idees?)

One-way ANOVA

Principe de l’anova: comparer la variations des moyennes par groupes a la variation dans chaque groupe.

ATTENTION: l’anova nous permet seulement de detecter si toutes les moyennes sont les memes ou si au moins une est differente *mais sans nous dire laquelle.*

Comment savoir alors? Avec des post-hoc tests, ou “tests de comparaison multiples”

ANOVA illustration (de Wikipedia)

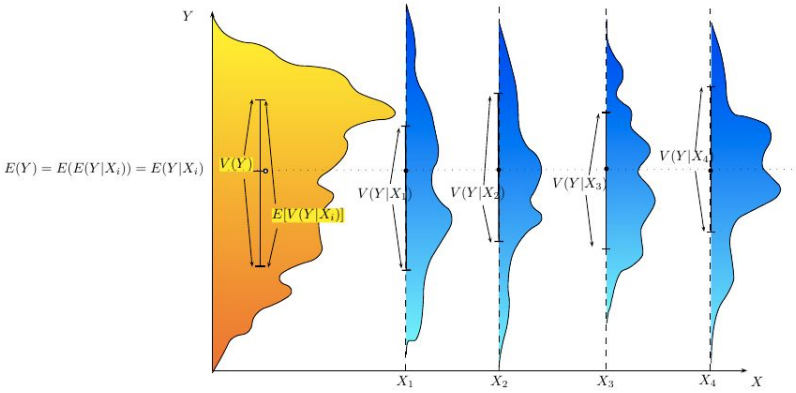


Figure 2: ANOVA : No fit

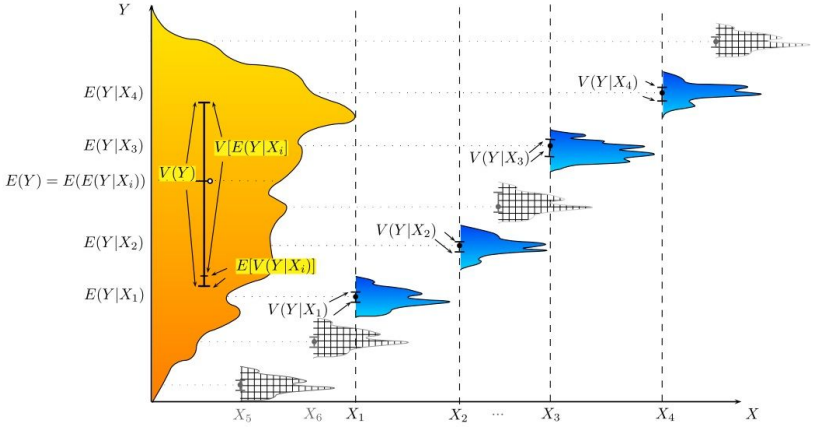


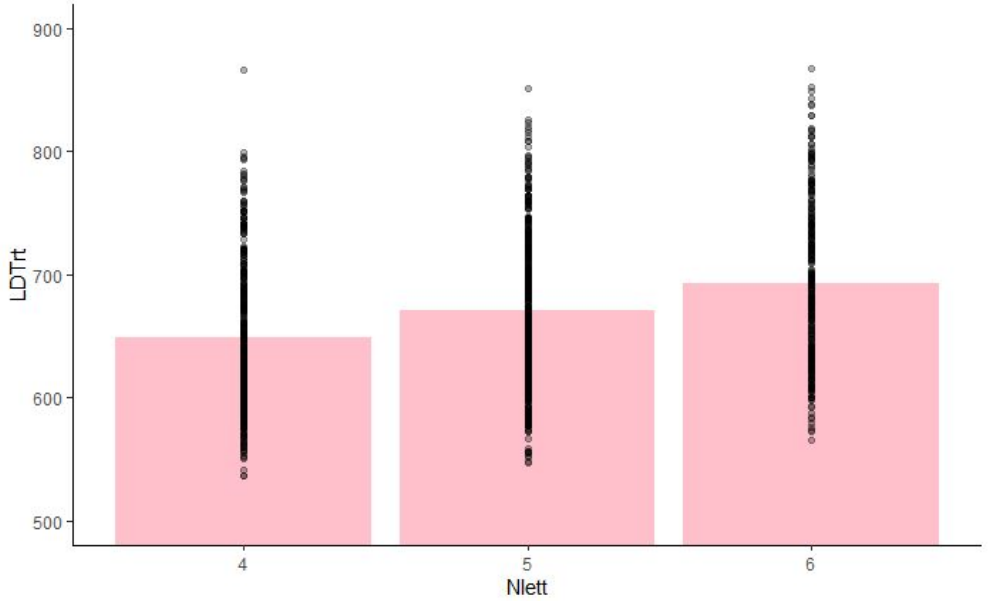
Figure 3: ANOVA : very good fit

One-way ANOVA

Fais un bar plot des moyennes de LDTrt en fonction de Nlett, avec les points de chaques observations.

Outliers?

Si oui comment retirer?



One-way ANOVA

En R c’est un peu complique bizarrement. Il faut deux fonctions:

- lm(variable dependante ~ variable independante, dataset):
  - Fit un modele lineaire sur les donnees
- anova()
  - Fais une anova sur un modele: test si les facteurs du modele influent significativement sur la variable dependante

→ anova(lm(x~y, data))

Faites une ANOVA avec Nlett et LDTrt

One-way ANOVA

La ligne Nlett = la variabilite entre les deux groupes

Df = N - 1 ou N est le nombre de groupe

La ligne Residuals = la variabilite a l'interieur des groupes

```
Analysis of Variance Table

Response: LDTrt
            Df  Sum Sq Mean Sq F value    Pr(>F)
Nlett        2  334995  167498  44.635 < 2.2e-16 ***
Residuals 1212 4548133    3753
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Que nous dit la p-value?

Post-hoc tests / test de comparaison mutliples

On sait qu'il existe une difference de moyenne entre nos trois groupes, mais on ne sait pas lesquels sont responsables de cette difference.

Il existe plusieurs tests qui nous permettent de faire ce:

- Bonferroni
- Tukey
- Dunnett
- Sidak
- …

Pairwise comparisons

Ici on voit que Nlett a un effet significatif mais on ne sait pas ou sont les differences:

Est-ce entre 4 et 5, 4 et 6 ou 5 et 6?

→ il y a trois comparaisons a faire

Avec r on peut utiliser la fonction:

pairwise.t.test(x, y, p.adj = “none”)

```
Analysis of Variance Table

Response: LDTrt
            Df  Sum Sq Mean Sq F value    Pr(>F)
Nlett        2  334995  167498  44.635 < 2.2e-16 ***
Residuals 1212 4548133    3753
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pairwise comparisons

pairwise.t.test(chronolexNlett$LDTrt, chronolexNlett$Nlett, p.adj = “none”)

```
        Pairwise comparisons using t tests with pooled SD

data:  chronolexNlett$LDTrt and chronolexNlett$Nlett

  4       5
5 1.4e-07 -
6 < 2e-16 7.3e-07

P value adjustment method: none
```

Ajuster la p-value

Attention: quand on fait plein de t.test pour comparer nos groupes, forcement on augmente notre risque de faire une erreur. Meme si le risque d'erreur pour chaque comparaison est inférieure a 5%, le risque de faire une erreur "en tout" est plus grand

Heureusement il y a des methodes pour ajuster la p-value de maniere a garder l'erreur générale en dessous de 5%

Par exemple la methode de Bonferroni (une des plus 'conservatrice')

```
pairwise.t.test(chronolexNlett$LDTrt, chronolexNlett$Nlett, p.adj = "bonferroni")
```

Bonferroni method

```
        Pairwise comparisons using t tests with pooled SD

data:  chronolexNlett$LDTrt and chronolexNlett$Nlett

  4       5
5 1.4e-07 -
6 < 2e-16 7.3e-07

P value adjustment method: none
```

```
        Pairwise comparisons using t tests with pooled SD

data:  chronolexNlett$LDTrt and chronolexNlett$Nlett

  4       5
5 4.2e-07 -
6 < 2e-16 2.2e-06

P value adjustment method: bonferroni
```

Recap

Variable independante	**Variable dependante**	**Test stats**	**Note**
1 quantitative	1 quantitative	correlation	On ne parle pas vraiment de IV et DV dans ce cas, juste de variable 1 et 2
1 qualitative avec deux niveaux (e.g., Taille: grand vs petit)	1 quantitative	Student’s t-test	
1 qualitative avec 3+ niveaux (e.g., Taille: grand, moyen, petit)	1 quantitative	One-way ANOVA	Nécessite souvent des test de comparaisons multiples apres, avec methode Bonferroni, Tukey etc…

Plus d’une variable independante

Et si on a plus d’une variable independante (facteur) qui nous interesse?

Par exemple, on peut s'intéresser au effet du nombre de lettre et de la frequence.

Commencons par:

1) Creer la variable freqCat qui vaut “high” si la frequence est superieur a la mediane de freqfilms et “low” sinon (deja fait plusieurs fois)
2) Calculer la moyenne de LDTrt en fonction de Nlett ET freqCAT (pas fait encore: une idee?)

Representation graphique

Ca devient plus compliqué quand on a plus que deux variables a representer, mais a notre niveau (deux IVs, une DV) c'est encore jouable avec des barplots.

Comment creer ce barplot?

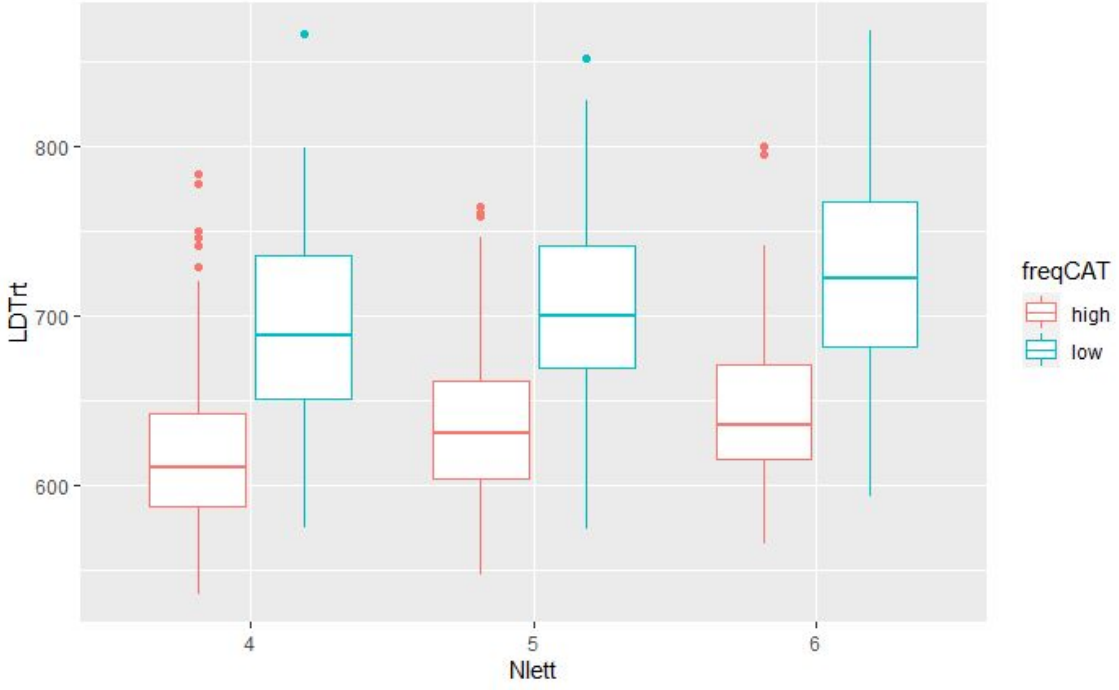
Mais pour changer on va faire un boxplot (et aussi pour gagner du temps sur le calcul des SD etc…)

Representation graphique

Comment creer ce graph a votre avis?

ggplot(chronolexNlett, aes(Nlett, LDTrt))+

….?



Two-way ANOVA

Comme pour le One-Way ANOVA on doit d'abord creer un modele lineaire avec lm() et ensuite utiliser anova() pour obtenir la F-statistics, les DFs et la p-value.

```
modelLettFreq <- lm(LDTrt ~ Nlett + freqCAT, chronolexNlett)
```

```
anova(modelLettFreq)
```

Two-way ANOVA

Que voit on?

```
Analysis of Variance Table

Response: LDTrt
             Df  Sum Sq Mean Sq F value    Pr(>F)
Nlett         2  334995  167498   67.97 < 2.2e-16 ***
freqCAT       1 1563894 1563894  634.63 < 2.2e-16 ***
Residuals  1211 2984239    2464
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

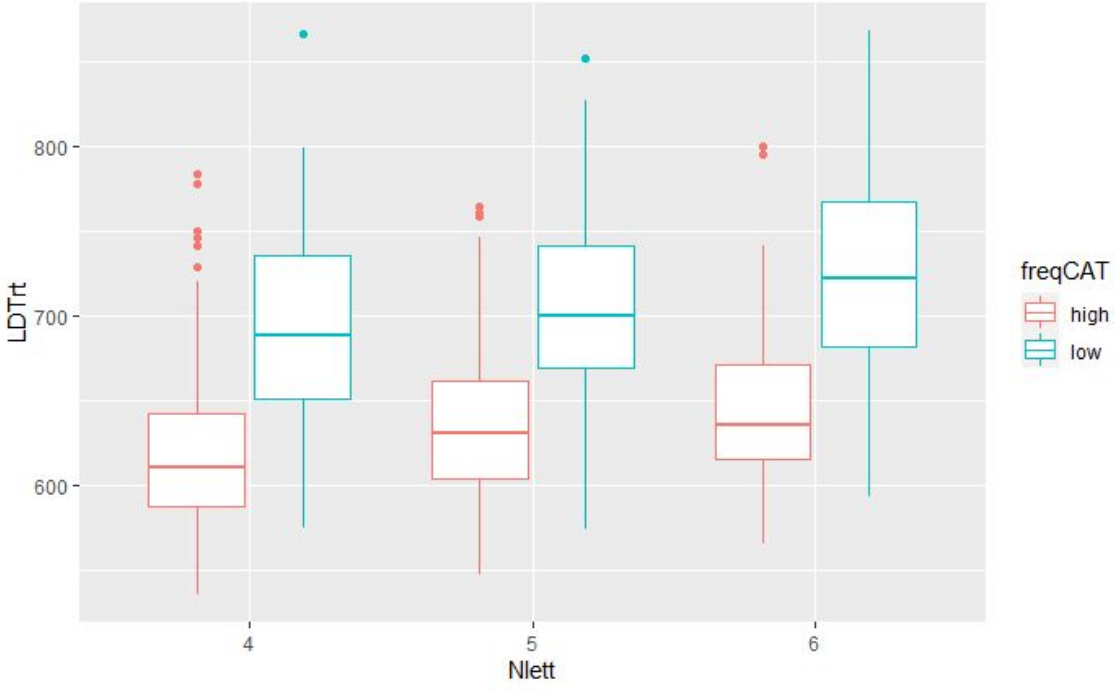
Interaction

Quand on a plus d'une variable indépendante on peut s'interesser a:

- L'effet de X
- L'effet de Y
- Mais aussi, l'effet combine de X et Y

Une interaction existe si l'effet de X et Y ensemble est plus grand que l'effet de X + l'effet de Y

Est ce qu'on peut suspecter une interaction ici?

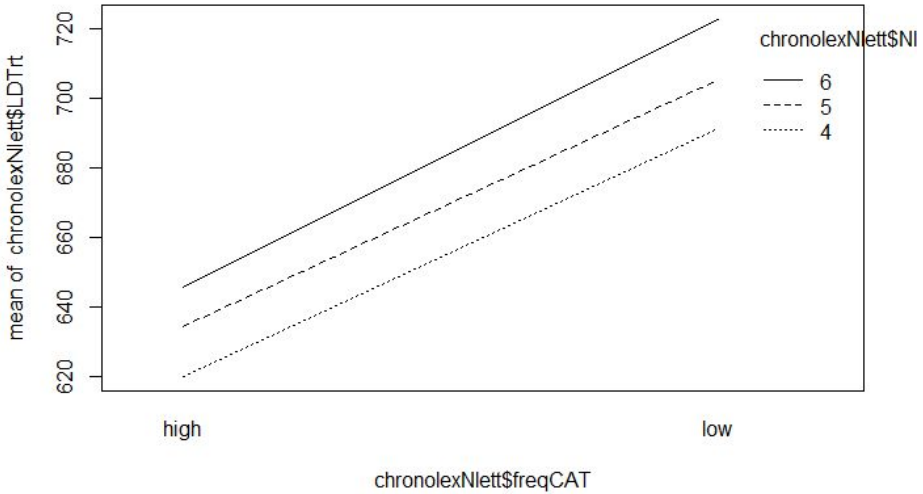


Meilleur représentation graphique

La fonction interaction.plot permet une bonne représentation des interactions:

interaction.plot(chronolexNlett$freqCAT, chronolexNlett$Nlett, chronolexNlett$LDTrt, fun = mean)

Droites paralleles = l'effet de la fréquence est comparable pour chaque longueur de mot = probablement pas d'interaction



Testons cela formellement

modelLettFreq_interaction<- lm(LDTrt ~ Nlett+ freqCAT + Nlett:freqCAT, chronolexNlett)

anova(modelLettFreq)

NOTE: modelLettFreq_interaction<- lm(LDTrt ~ Nlett * freqCAT, chronolexNlett) fait la meme chose.

Interaction: R output

Que voit-on?

```
Analysis of Variance Table

Response: LDTrt
               Df  Sum Sq Mean Sq F value Pr(>F)
Nlett           2  334995  167498  67.902 <2e-16 ***
freqCAT         1 1563894 1563894 633.985 <2e-16 ***
Nlett:freqCAT   2    1914     957   0.388 0.6785
Residuals    1209 2982324    2467
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

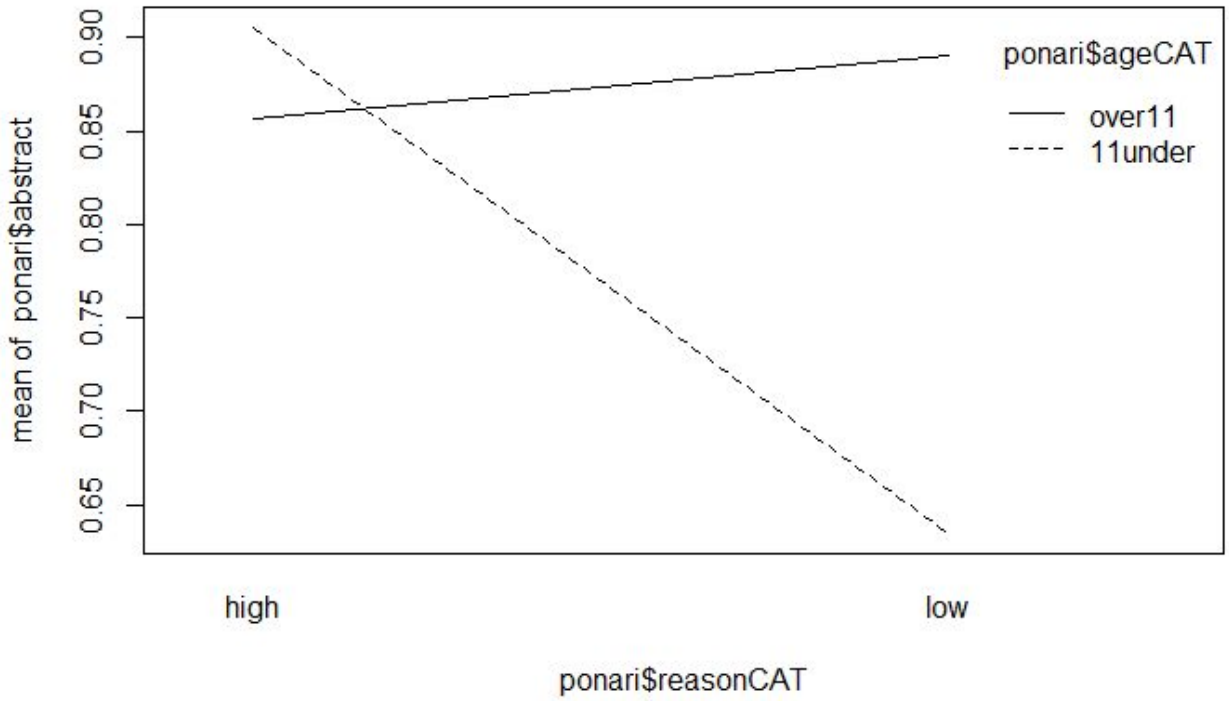
Interaction: autre exemple avec Ponari

1) Lire le dataset Ponari
   a) On s'interesse a la relation entre le groupe des participants (TD ou DLD) et leur reasoning Skills, et les effets sur leur reconnaissance des mots abstraits.
2) Creez une variable **reasonCAT** qui prend pour valeur "**low**" quand reasoningSkills est en dessous de la **mediane** pour cette variable, et "**high**" sinon
3) Creez une variable **ageCAT** qui prend pour valeur "**over11**" quand l'age est **plus que 11** et "**11under**" quand l'age est **11 ou moins**
4) De quel type sont ces deux variables transformees?
5) Creez un interaction plot avec les deux variables independantes et la variable dependante

Interaction plot

Que voit on?

Comment interpreter ce graph?



Analyse formelle

Quelle formule dois je utiliser pour faire ma two-way anova?

Step 1: _____?

Step 2: _____?

Output R

Ca nous dit que les reasoning skills influent sur les capacites a reconnaitre les mots abstraits differement en fonction de l'age

```
Analysis of Variance Table

Response: abstract
                Df  Sum Sq  Mean Sq F value    Pr(>F)
ageCAT           1 0.16074 0.160743  12.726 0.0011959 **
reasonCAT        1 0.19960 0.199602  15.802 0.0003914 ***
ageCAT:reasonCAT 1 0.14507 0.145067  11.485 0.0019272 **
Residuals       31 0.39158 0.012631
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction recap:

Il y a interactions lorsque l'impact d'un facteur dépend des valeurs d'un autre facteur.

Pour analyser plus loin un interaction, il faut parfois la decomposer. Par exemple ici on pourrait faire une one-way anova pour ageCAT = over11 et une opur ageCAT = 11under et voir quel sont les effets de reasonCAT dans les deux cas.

Souvent pour une interaction avec deux facteurs seulement, on a tendance a se baser sur le graph tout simplement. Mais pour des three-way interactions ou plus, il faut avoir recours a ce genre de decomposition.

Conditions pour ANOVA

Attention, pour utiliser une ANOVA (ou un t-test) il faut respecter certaines conditions, en autre:

- Les donnees sont distribuees de facon “normale” (i.e., follow the normal distribution)
- Homogeneite de la variance
- La taille des echantillons des differents groupes est equivalente

Normalite

Pour s’assurer qu’on a respecte cette condition, on peut comparer la distribution de nos donnees par rapport a la distribution normale.

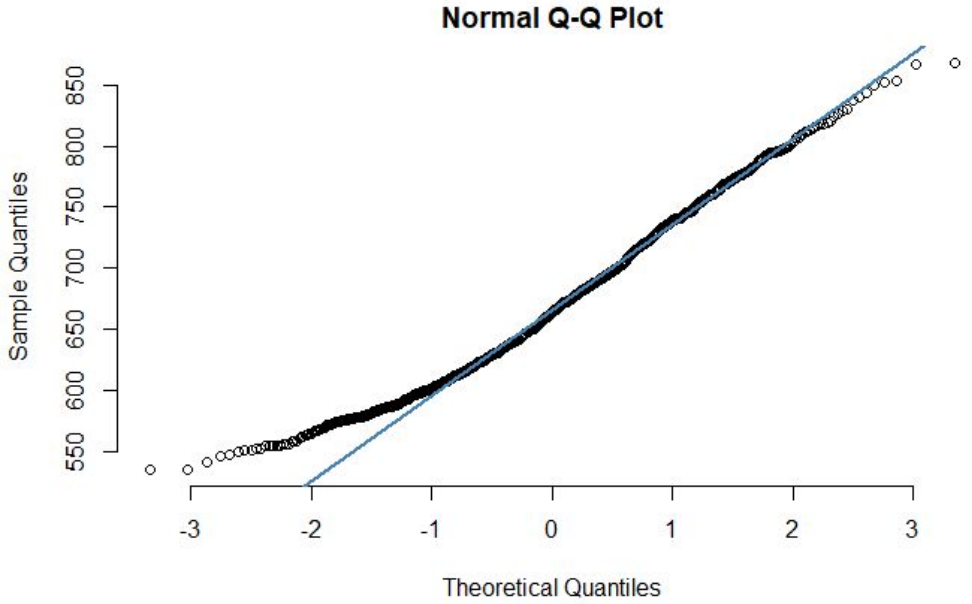
On va revenir sur chronolex et LDTrt

La fonction qqnorm permet de creer un qqplot (= un plot qui represente les quantiles de deux distributions l’un par rapport a l’autre - en particulier ici par rapport aux quantiles de la distribution normale)

Normalite

```
qqnorm(chronolexNlett$LDTrt,
        pch = 1,
        frame = FALSE)

qqline(chronolexNlett$LDTrt,
    col = "steelblue",
    lwd = 2)
```

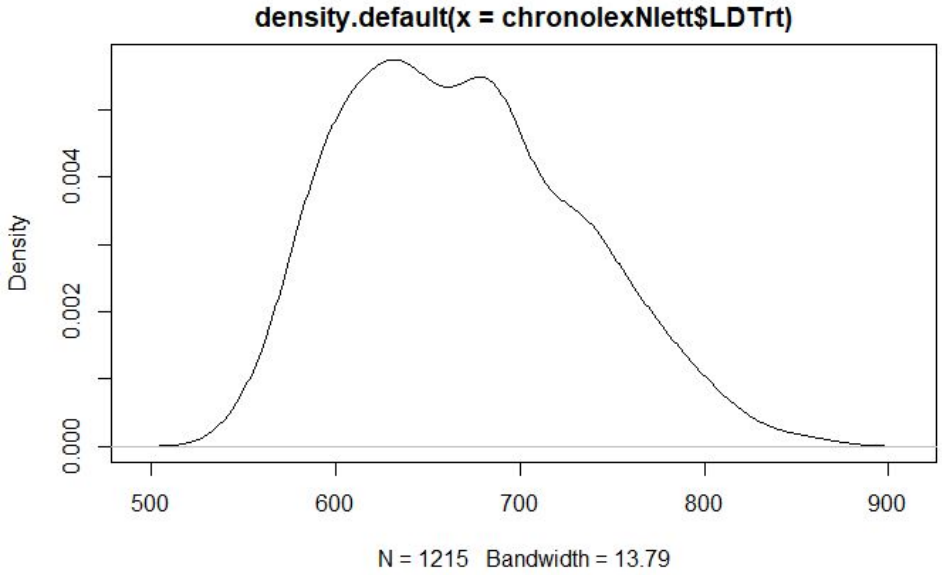


Normalite

On peut regarder la distribution de nos reaction times:

plot(density(chronolexNlett$LDTrt))

Effectivement avec right skew



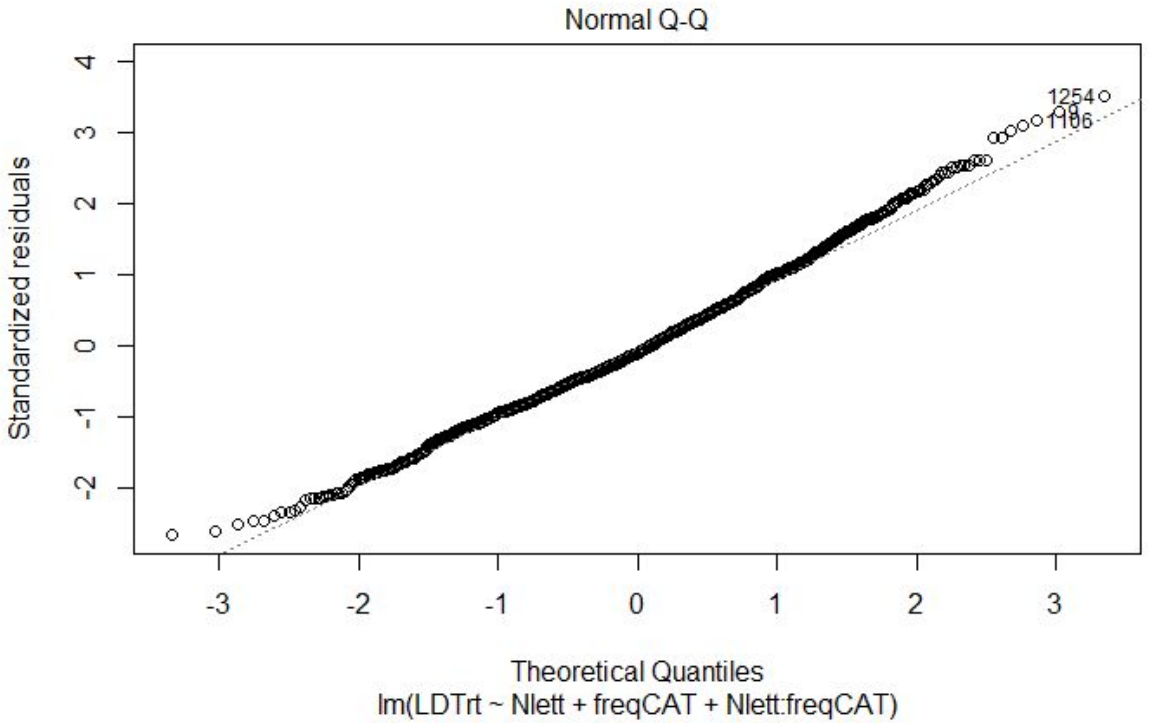
Normalite: autre facon de diagnostiquer

1) La fonction qqPlot du package “car”
   a) Comment installer le package car?
   b) qqPlot(chronolexNlett$LDTrt)

2) On peut aussi regarder la distribution des *residuals* du modele (i.e., la variabilite qu’il reste une fois qu’on prend en compte l’effet de nos facteurs. Si il n’y a pas de truc bizarre, elle doit etre normalement distribue).
   a) plot(model, 2)

Normalite: autre

plot(model, 2)



Homoscedasticity

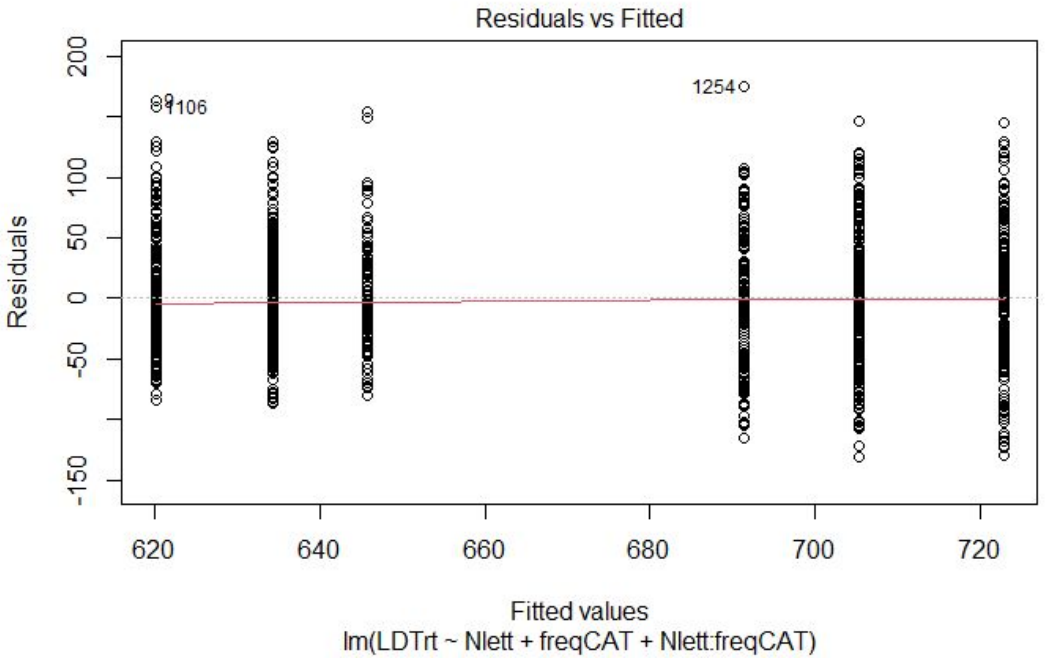
Pour s’assurer que la variance est plutot egale pour tous les niveaux de nos variables, on peut aussi regarder les residuals:

plot(model, 1)

Homoscedasticity

On ne voit pas de pattern bizarre: i.e. pas de relation entre les residuals et les valeurs de notre modele.

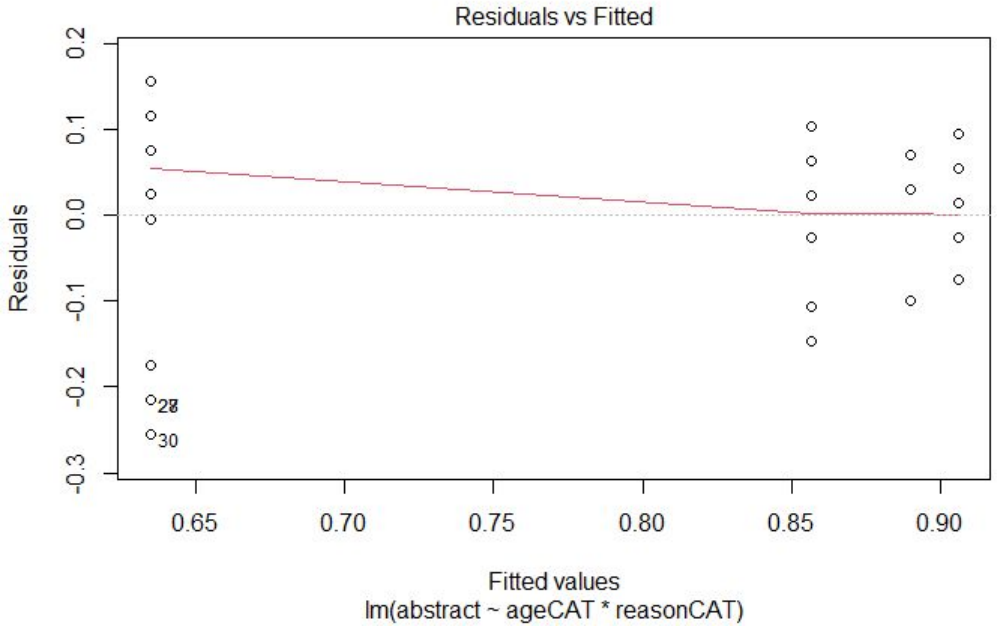
Note que les point 1254 et 1106 sont identifes comme outliers et peuvent etre interessants a retirer



Exemple d’heteroscedasticity

plot(modelponari,1)

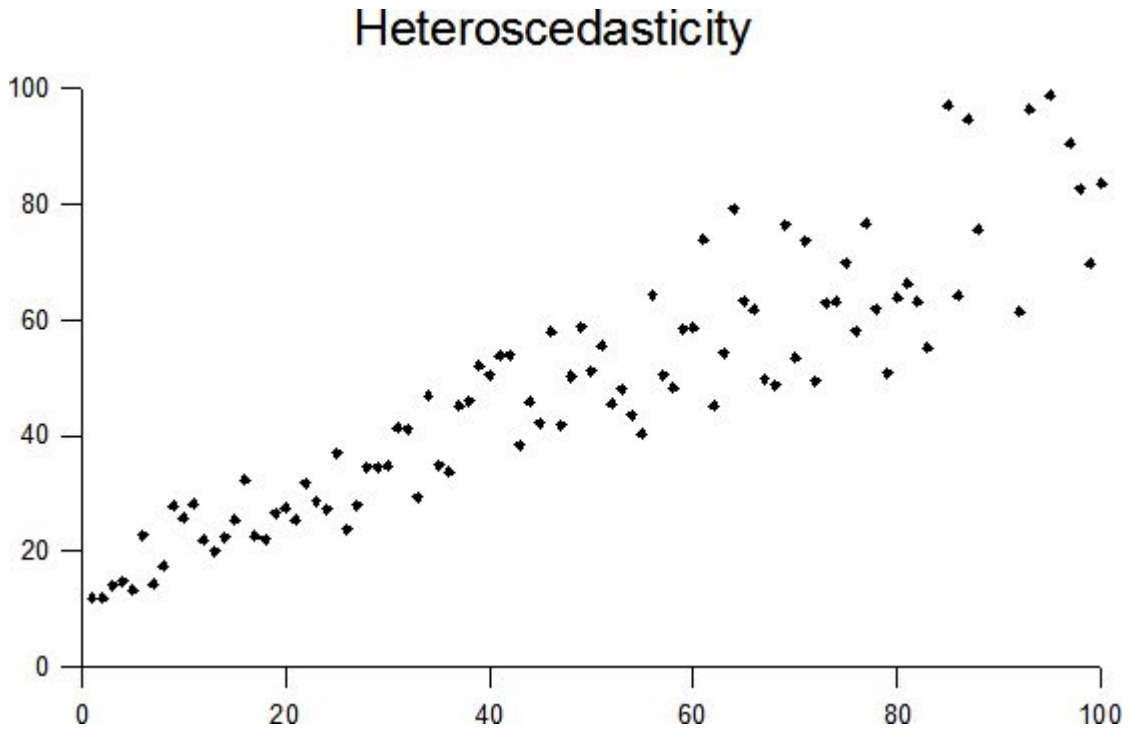
La variance est plus grande pour les valeurs faibles de “abstract” et moins grande pour les valeurs elevees.



Exemple d’heteroscedasticity

Exemple Wikipedia:

Ici la variance augmente pour les grandes valeurs de X par rapport au petites valeurs



Semaine pro

Derniere sceance! Yay!

On parlera enfin de regression + recap du cours

TD 4, 5 et 6 sont a rendre avant le 10 Janvier