ANOVA/t-test

3/12/2021

Questions sceances passees

- Comment creer un scatterplot avec R?
- Si les noms des colonnes sont modifies a cause de l'encoding ou autre, comment regler ce probleme?
- Comment selectionner une colonne a laquelle on s'interesse?
- Comment selectionner une rangee a partir de la valeur d'une colonne?

Rappel semaine derniere

1. On a vu l'importance de faire des plots pour explorer ses donnees et identifier certaines valeurs extremes
2. Les tests d'hypotheses / methode scientifique:
    a. **Question de recherche**: les lettres sont elles plus facile a identifier dans les mots que seules?
    b. **Hypothese de recherche** (i.e., l**'hypothese alternative**): Les lecteurs seront plus rapide a trouver que la lettre "L" apparait a l'ecran lorsque qu'elle est presentee dans un mot (e.g., "elan") versus pas (juste l).
    c. **Hypothese nulle**?
    d. Choisir un test statistique pour obtenir la **p-value.**
    e. **P-value =** la probabilite que, *sachant l'hypothese nulle*, on obtienne les donnees que l'on a.

Deux types d’erreurs

Faux positif: dire qu’il existe une vraie difference entre nos groupes alors qu’en realite il y en a une / dire qu’il y a un effet quand il n’y en a pas

Faux negatif: dire qu’il n’y a pas de difference entre nos groupes alors qu’en realite il y en a une / dire qu’il n’y a pas d’effet quand il y en a un.

Choisir le bon test statistique pour vos donnees

1. **Etape de la methode scientifique**
    a. **Question de recherche**: les lettres sont elles plus facile a identifier dans les mots que seules?
    b. **Hypothese de recherche** (i.e., l'**hypothese alternative**): Les lecteurs seront plus rapide a trouver que la lettre "L" apparait a l'ecran lorsque qu'elle est presentee dans un mot (e.g., "elan") versus pas (juste l).
    c. **Hypothese nulle**?
    d. Choisir un test statistique pour obtenir la **p-value.**
    e. **P-value =** la probabilite que, *sachant l'hypothese nulle*, on obtienne les donnees que l'on a.

Type de variables: rappel

Variable = un indicateur d'un phenomene (e.g., la fumee est un indicateur de feu; l'argent un indicateur de la richesse etc…). Cet indicateur peut changer de valeur (donc il est *variable*). Une variable est mesurable et quantifiable.

Variable quantitative (continuous): contiennent des valeurs numériques faisant référence à une unité de mesure reconnue / peut prendre un grand nombre de valeurs en général. Exemple: l'age, la taille, etc…

Variable qualitative: contiennent des valeurs qui expriment une qualité, un etat, une condition, un statut unique et exclusif. Exemple: le sexe, la religion, etc...

Types de variables

Variable independante = la variable manipulee par l'experimentateur.

Variable dependante = la variable qui peut changer en fonction de la variable independante (i.e., qui subit ses effets).

Exemple:

- Question: Est ce que l'age du premier mot est different pour les monolingues vs. les bilingues?
- Hypothese: Les bilingues mettent plus longtemps a produire leur premier mot
- Variable independante?
- Variable dependante?

Variables

Probleme: Je veux comparer les capacites de lecture des enfants dyslexiques aux enfants normolecteurs.

- Variable independante? De quel type est elle (qualitative ou quantitative)?
- Variable dependante? Type?

Probleme: Quel est le lien entre vitesse d’apprentissage de la lecture et la transparence d’une langue?

- Variable independante? Type?
- Variable dependante? Type?

Variables

Probleme: Quel est le lien entre nombre d’heures d’exposition a une langue et capacite orale dans cette langue?

- Variable independante? Type?
- Variable dependante? Type?

Type de variable et test stats

En fonction des types de variables dependantes et independantes que l'on a, on ne va pas avoir recours aux memes tests statistiques.

Par exemple:

- Deux variables (IV et DV) quantitatives → probablement une correlation/regression
  - On s'interesse a la relation entre les deux variables: est ce qu'elles covarient?
- Une variable independante categorique + une variable dependante continue → un t-test
  - On s'interesse a "est ce que deux groupes/choses sont differentes"

Correlation

On sait comment:

1. Representer graphiquement un nuage de points avec une droite de regression/correlation
2. Obtenir le coefficient de correlation

On ne sait pas encore tester si la correlation entre les deux variables est *significative* ou non. C'est a dire: est ce que la relation entre les deux variables que l'on observe est du a la chance ou bien est sans doute du a une vraie relation?

Correlation

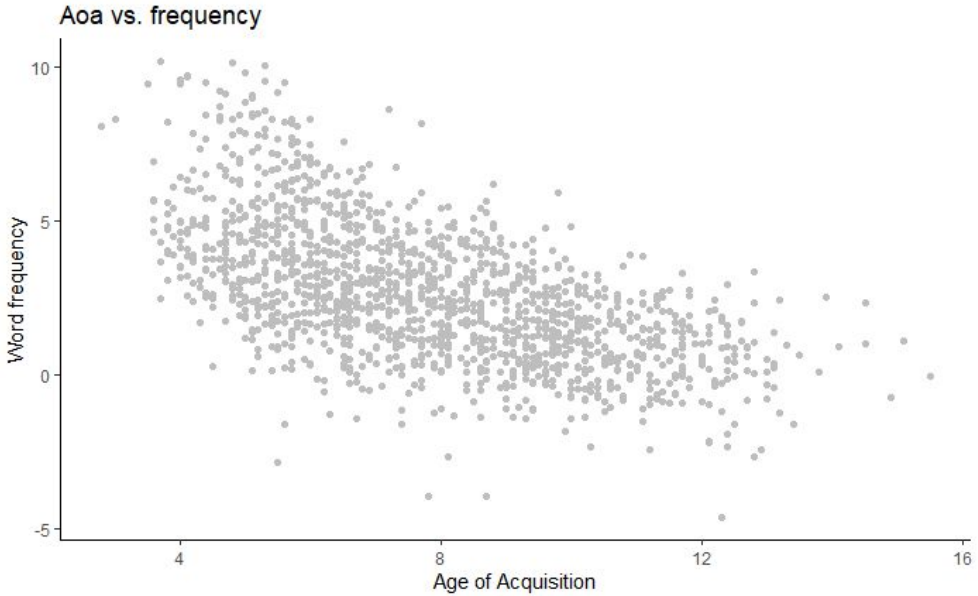
1. Ouvrir RStudio
2. Ouvrir votre Project TD, ou simplement un nouveau script
3. Importer chronolex

On va s’interesser a Age of Acquisition et la frequence des mots (freqfilms)

Correlation

Faites un plot de AoA vs freqfilms

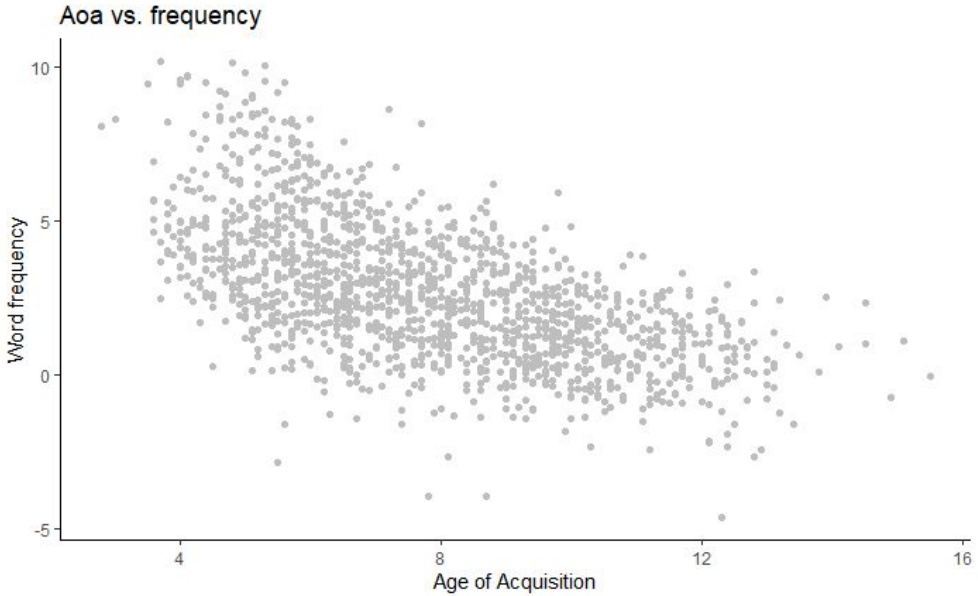
Est ce qu'on voit des outliers evident juste avec une inspection visuelle?



Correlation

Quel est le coefficient de correlation entre ces deux variables?

Comment interprete-t-on ce resultat?



Tester la significativite d’une correlation

On peut tester:

- Est ce que r est significativement different de zero?
  - C’est a dire: la relation entre les deux variables est significativement plus large que 0 de facon a ce qu’on puisse ecarter l’idee que c’est du a des variations aleatoires
- Est ce que deux valeurs de r sont significativement differentes?
- Est ce que r differe significativement d’un point autre que 0?

Note que r est une estimation du parametre de population ρ (rho)

Tester la significativite d'une correlation

On peut tester:

- **Est ce que r est significativement different de zero?**
  - **C'est a dire: la relation entre les deux variables est significativement plus large que 0 de facon a ce qu'on puisse ecarter l'idee que c'est du a des variations aleatoires**
- Est ce que deux valeurs de r sont significativement differentes?
- Est ce que r differe significativement d'un point autre que 0?

Note que r est une estimation du parametre de population ρ (rho)

Correlation

Est ce que r differe significativement de zero?

Pour repondre a cette question on fait deux hypotheses sur la relation entre nos deux variables au niveau de la population:

1) Hypothese nulle = ?
2) Hypothese alternative = ?

Correlation

Est ce que r differe significativement de zero?

Pour repondre a cette question on fait deux hypotheses sur la relation entre nos deux variables au niveau de la **population, c'est a dire pour** ρ. ρ represente le "vrai" coefficient.

1) Hypothese nulle: $\rho = 0$
2) Hypothese alternative: ρ != 0

L'hypothese nulle ici voudrait dire que les deux variables sont lineairement independantes l'une de l'autre (i.e., elles n'ont aucune relation entre elle)

Parenthese: crud factor

Le Crud Facteur decrit par Meehl en 1990 est le phénomène qui veut que tout correle un peu avec tout.

Starbuck a trouve que “choosing two variables at random, a researcher has a 2-to-1 odds of finding a significant correlation on the first try, and 24-to-1 odds of finding a significant correlation within three tries”. Sa conclusion: “social sciences are drowning in statistically significant but meaningless noise”.

Attention du coup en big data! C’est pourquoi il faut aussi différencier significative d’importance de l’effet.

Correlation

Est ce que r differe significativement de zero?

Pour repondre a cette question on fait deux hypotheses sur la relation entre nos deux variables au niveau de la population:

1) Hypothese nulle: $\rho = 0$
2) Hypothese alternative: ρ != 0

L'hypothese nulle ici voudrait dire que les deux variables sont lineairement independantes l'une de l'autre (i.e., elles n'ont aucune relation entre elle)

Est ce que r differe significativement de 0?

Le niveau de significativite de pour ce test est determine par la significativite du **t-test de Student**

Le t-test de Student est un test statistique qui permet de comparer deux moyennes.

La formule pour calculer la significativite du coefficient de correlation avec un t-test est:

$$T = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

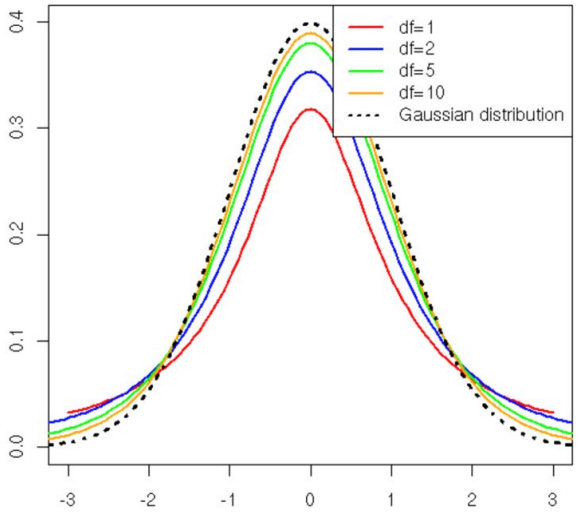
Comment obtenir la p-value?

P-value = la response a la question”quelle est la chance d’obtenir les donnees que l’on a (i.e., la valeur de T que l’on a) si l’hypothese nulle est vraie?”

Pour connaitre la p-value associer aux t-test, il faut se referrer a une distribution de t avec n-2 degrees de liberte.

Bien sur on va pas faire ca a la main

→ R le fait tout seul!



Correlation test avec R

On utilise la fonction cor.test() avec la method “pearson”

Note: “pearson” est la methode par defaut donc il n’y a pas besoin de la specifier mais c’est bien pour s’assurer qu’on se souvient de ce qu’on fait.

cor.test(x, y, method = “pearson”)

Est ce que la correlation entre AoA et freqfilms est significative?

Correlation test avec R

T = t-test statistique

Df = degrees of freedom

Doit etre egale a N-2. Est ce le cas?

P-value: que nous dit elle?

```
        Pearson's product-moment correlation

data:  chronolex$AoA and log(chronolex$freqfilms)
t = -28.93, df = 1480, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6325849 -0.5674717
sample estimates:
       cor
-0.6010247
```

Correlation test avec R

Confidence interval (intervalle de confiance): un set de valeurs dont on est plutot sure contient la vraie valeur (le parametre de population)

Un intervalle de confiance a 95% nous dit que 95% des experiences comme la notre contiendront la vraie valeur de la moyenne (mais 5% non! Donc on a une chance sur 20 (5/100) de se tromper).

```
        Pearson's product-moment correlation

data:  chronolex$AoA and log(chronolex$freqfilms)
t = -28.93, df = 1480, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6325849 -0.5674717
sample estimates:
       cor
-0.6010247
```

Correlation test avec R

ATTENTION: l'intervalle de confidence ne veut PAS forcement dire qu'on est 95% sure du resultat.

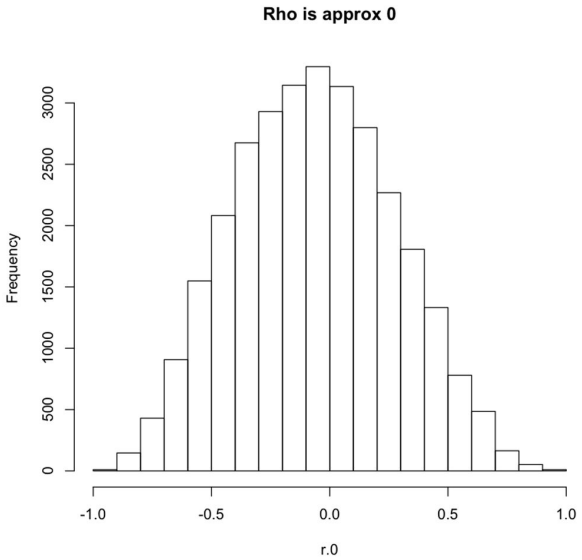
```
        Pearson's product-moment correlation

data:  chronolex$AoA and log(chronolex$freqfilms)
t = -28.93, df = 1480, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6325849 -0.5674717
sample estimates:
       cor
-0.6010247
```

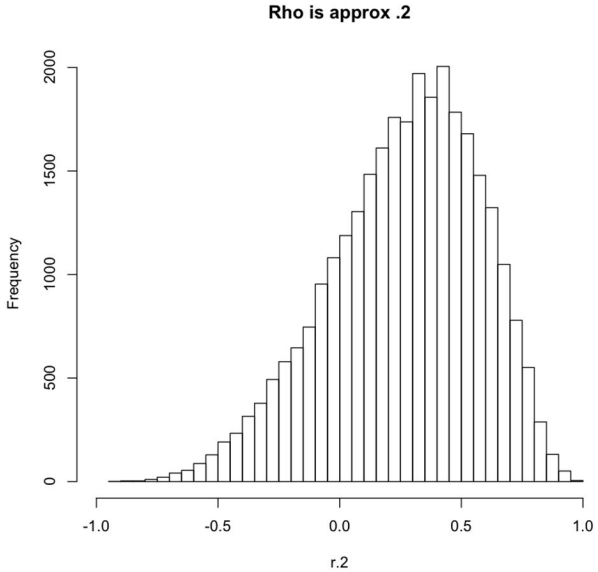
Autre tests

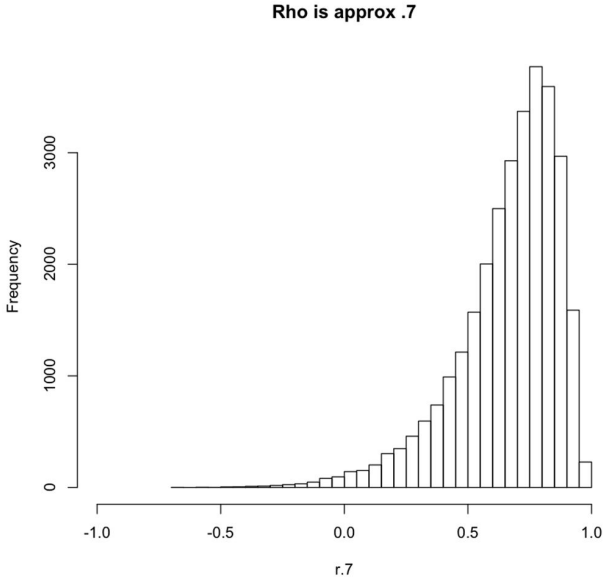
1) Est ce que deux valeurs de r different?
2) Est ce que r differe d'une valeur fixe?

Ici on a un probleme: si on part de l'hypothese que la vraie valeur de ρ n'est pas 0 alors la distribution de r ne sera plus *normalle*. Est ca pose des problemes pour calculer les ecarts types necessaire au t-test (pas besoin de comprendre les details).

Petite demo du probleme







Autre test: Est ce que deux valeurs de r different?

Pour echapper a ce probleme, on peut utiliser la statistique z de Fisher (Fisher's z), aussi appele r'.

C'est un poil plus complique, et plus facile a illustrer quand on s'interesse a des vairables categoriques, donc pour l'instant on va passer.

J'y reviendrai (peut etre) en TD 5 ou 6.

Test d’hypothese et type de variable

On a vu comment analyser la relation entre deux variables continues.

Mais souvent en recherche on a une variable independante categorique et une variable dependante continue. Exemple:

- Monolingue vs. bilingue et temps de reaction en DL
- Treatment vs. Placebo et evolution des symptomes
- ⇒ Groupe A vs Groupe B et une mesure

Student’s t-test

Utile pour comparer deux moyennes. Donc quand on a:

- Une variable independante categorique a deux niveaux:
  - Language background (lvl 1 = monolingue; lvl 2 = bilingue)
  - Political affiliation (lvl 1 = republican; lvl 2 = democrat)
- Une variable dependante continue:
  - Temps de reaction en DL (compris entre 400ms et 1000ms)
  - Average yearly income of their neighborhood (compris entre 0$ et 1M$)

Student’s t-test

Dans chronolex:

1) Creez une variable freqbooksCAT qui divise en deux la variable frequence:
   a) High vs. low
2) Calculer la moyenne des temps de réaction en progressive demasking (PDMrt) pour chaque catégorie de frequence (aggregate, mean)
3) Faire un bar graph de ces moyennes.

chronolex$freqbooksCAT<- ifelse(chronolex$freqbooks<median(chronolex$freqbooks), “low”, “high”)

chronolex$freqbooksCAT[chronolex$freqbooks < median(chronolex$freqbooks)] <- “low”

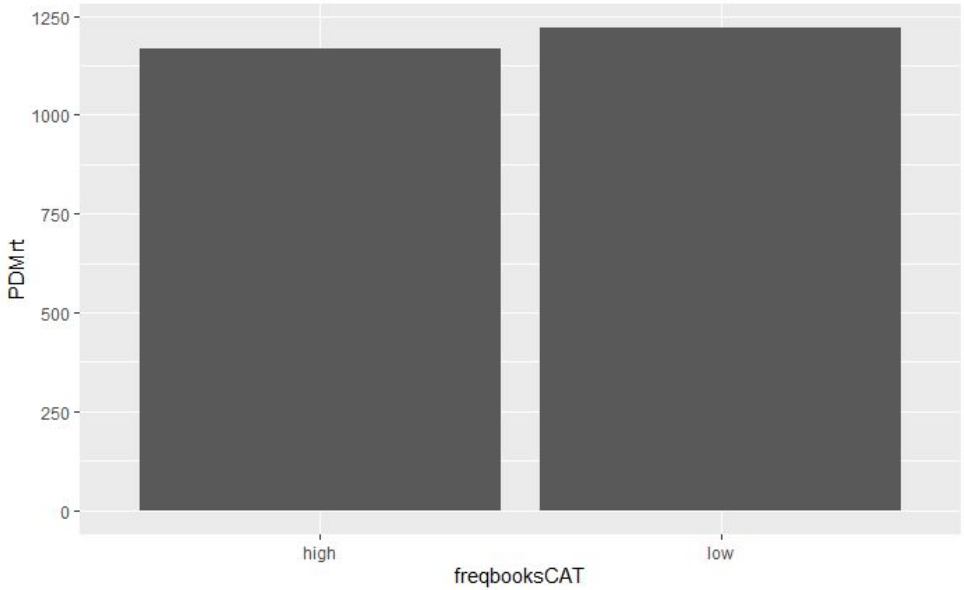
Student’s t-test

On a bien deux moyennes qu’on veut comparer.

ggplot(means, aes(freqbooksCAT, PDMrt))+

geom_bar(stat=”identity”, position=”dodge”)

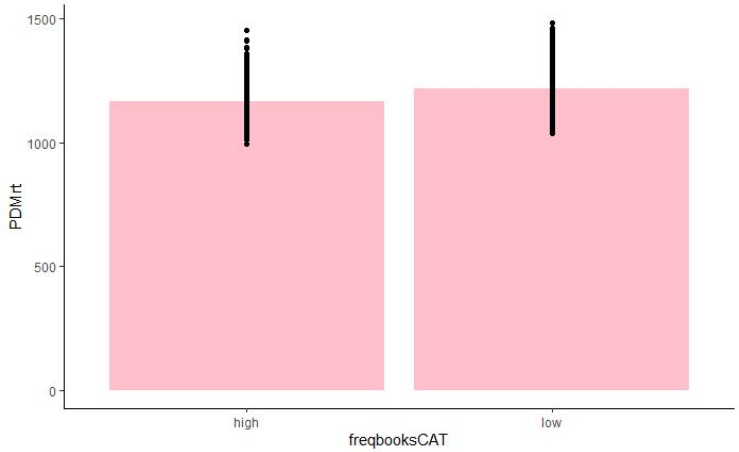
ATTENTION: les bar plots tout seul c’est pas tres utile et surtout ca peut cacher des histoires de distributions des donnees.!



Student’s t-test

Un (un peu) meilleur plot

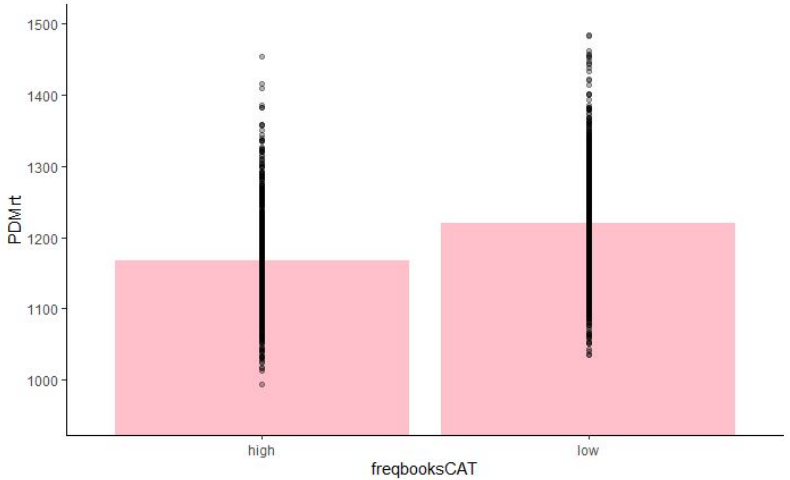
```r
ggplot(meansFreq, aes(freqbooksCAT, PDMrt))+
  geom_bar(stat="identity", position = "dodge", fill = "pink")+
  geom_point(data=chronolex, aes(freqbooksCAT, PDMrt, group=freqbooksCAT))+
  theme_classic()
```



Student t-test

Un encore un peu meilleur plot:

```
```{r}
ggplot(meansFreq, aes(freqbooksCAT, PDMrt))+
  geom_bar(stat="identity", position = "dodge", fill = "pink")+
  geom_point(data=chronolex, aes(freqbooksCAT, PDMrt, group=freqbooksCAT), alpha=0.3)+
  theme_classic()+
  coord_cartesian(ylim=c(950, 1500))
```
```



Y a t-il des outliers repérables graphiquement?

Student t-test

Ici on a un two sample t-test parce que nos deux groupes sont indépendants (un mot dans le groupe low frequence n'est pas dans le groupe high frequence)

Fonction R: t.test(continuous variable ~ categorical variable, dataset)

Note: l'ecriture ~ est tres commune en R et peut se lire "par" (pense a aggregate(x~y) par exemple)

Qu'obtient on pour freqbooksCAT et PDMrt?

Student t-test

T = la statistique t

Df = degrees of freedom (ici calcule avec l'approximation de Satterthwaite)

P-value = la probabilite d'observer nos donnees sachant l'hypothese nulle (pas de difference entre les groupes) est <0.001

DONC on peut dire que les donnees observees ne collent pas avec l'hypothese nulle

```
        Welch Two Sample t-test

data:  PDMrt by freqbooksCAT
t = -13.163, df = 1446.4, p-value < 2.2e-16
alternative hypothesis: true difference in means between group
high and group low is not equal to 0
95 percent confidence interval:
 -61.03780 -45.20511
sample estimates:
mean in group high  mean in group low
          1166.984           1220.105
```

Student t-test

C’est possible de modifier certains des arguments de la fonction t.test.

```
        Two Sample t-test

data:  PDMrt by freqbooksCAT
t = -13.163, df = 1480, p-value < 2.2e-16
alternative hypothesis: true difference in means between group high and
group low is not equal to 0
95 percent confidence interval:
 -61.03765 -45.20526
sample estimates:
mean in group high  mean in group low
          1166.984           1220.105
```

Par exemple vous pouvez faire l’hypothese que les deux groupes ont une variance egale:

t.test(x ~y, data, var.equal = TRUE)

Remarquez la diff avec les DFs

Student t-test

C’est possible de modifier certains des arguments de la fonction t.test.

On peut aussi specifier qu’on veut un paired t.test pour comparer les memes personnes sur deux variables

Exemple avec chronolex?

Formule R: t.test(x~y,dataset ,paired = TRUE)

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

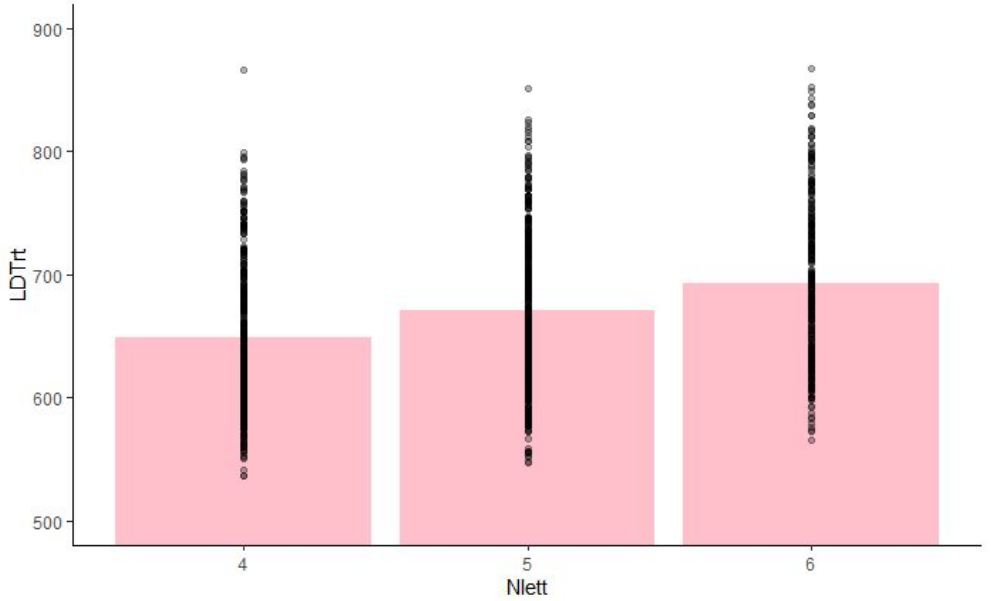
Permet de comparer trois groupes ou plus sur une seule variable dependante continue.

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

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



One-way ANOVA

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

- lm(variable dependante ~ variable independante, dataset)
- anova()

→ anova(lm(x~y, data))

Faites l’anova de 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

Que nous dit la p-value??

```
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
```

Semaine pro

TD 3 due lundi soir max

TD 4 sera poste ajd ou demain et due une semaine a partir du moment ou il est poste (aussi sur 3 points)

Cours:

- Un peu plus de detail sur les one-way ANOVA
- Two-way ANOVA
- Regression (ANOVA est en fait un cas particulier de la regression)