Statistique inférentielle: tests d'hypothèse simples, régressions linéaires, ANOVA, modèles linéaires généralisés

Lise Vaudor

2016-08-29

Contents

1 Introduction: 2

1.1 Qu'est-ce que l'inférence statistique? . . . 2

1.2 Un premier exemple d'inférence statistique: l'estimation d'un paramètre . . . 2

1.3 Deuxième exemple d'inférence statistique: détection de l'existence d'un lien entre deux variables 3

2 Tests d'hypothèse 7

2.1 Un exemple de test d'hypothèse: le t-test . . . 7

2.2 Le t-test en théorie . . . 8

2.3 Principe général d'un test d'hypothèse *paramétrique* . . . 9

2.4 Quelques simulations pour mieux comprendre . . . 11

3 Effet d'une variable catégorielle sur une variable quantitative: l'ANOVA 13

3.1 Questions associées à une ANOVA . . . 13

3.2 Réalisation de l'ANOVA . . . 14

3.3 Et au fait, elles étaient vérifiées les hypothèses du modèle linéaire? . . . 19

4 Effet d'une variable quantitative sur une autre: régression linéaire 20

4.1 Questions associées à une régression linéaire . . . 20

4.2 Réalisation d'une régression linéaire avec R . . . 21

4.3 Et au fait, elles étaient vérifiées les hypothèses du modèle linéaire? . . . 22

5 Influence de X_1,X_2,X_3... sur Y (quantitative, continue, gaussienne): modèle linéaire classique 24

5.1 X_1,X_2,X_3... sont toutes quantitatives: régression linéaire multiple . . . 24

5.2 X_1,X_2,X_3... sont toutes catégorielles: ANOVA à plusieurs facteurs . . . 26

5.3 X_1,X_2,X_3... sont quantitatives ou catégorielles: ANCOVA . . . 31

1 Introduction:

1.1 Qu'est-ce que l'inférence statistique?

- Considérons une **population** (par exemple, l'ensemble des êtres humains).
- Considérons une **variable d'intérêt** Y (par exemple, la taille des êtres humains)
- Considérons par ailleurs une **variable explicative** X (par exemple, le sexe -homme ou femme- des individus considérés)

Pour étudier et expliquer la taille des êtres humains "en général", on est obligé d'échantillonner la variable Y, puis de tenter d'extrapoler le résultat obtenu sur l'échantillon à l'ensemble de la population. De même, si l'on cherche à savoir si X a un effet "en général" sur Y, il faudra déterminer dans quelle mesure on peut extrapoler l'éventuelle différence observée entre les groupes à la population.

C'est cette généralisation (ou extrapolation) de l'échantillon à la population que l'on appelle **inférence statistique**.

1.2 Un premier exemple d'inférence statistique: l'estimation d'un paramètre

Considérons par exemple, que l'on ait mesuré 5 êtres humains "au hasard" sur terre, et qu'ils mesurent respectivement 159, 179, 183, 166 et 167cm. On s'intéresse à la taille moyenne (μ) de la population humaine.

```
quelques_tailles=c(159,179,183,166,167)
mean(quelques_tailles)
```

```
[1] 170.8
```

Avec une taille d'échantillon aussi petite ($n = 5$), il est évidemment difficile d'extrapoler les résultats pour affirmer "La taille moyenne des êtres humains est 170.8cm"!

En effet, la capacité à réaliser l'inférence statistique dépend de la **taille d'échantillon**.

Considérons maintenant un échantillon plus grand:

```
representations=read.table("../../datasets/representations.csv",sep=";", header=T)
attach(representations)
length(taille) # taille d'echantillon
```

```
[1] 199
```

```
mean(taille)
```

```
[1] 170.7638
```

Avec cet échantillon, on est déjà un peu plus confiant lorsque l'on décrit la taille des êtres humains "en général". Malgré cela, il reste impossible d'affirmer "La taille moyenne des êtres humains est 170.7638". En effet, la moyenne observée sur l'échantillon est une **estimation** $\hat{\mu}$, a priori différente de la valeur réelle de taille moyenne μ du fait du **hasard d'échantillonnage**.

Il convient donc d'être "prudent" dans nos affirmations et d'assortir, par exemple, notre estimation de moyenne d'un **intervalle de confiance**.

L'intervalle de confiance (IC) à 95% est un intervalle de valeurs qui a 95% de chance de contenir la vraie valeur du paramètre μ.

Le calcul de l'intervalle de confiance repose sur un **modèle**.

Ici, supposons que les mesures Y_i soient indépendantes les unes des autres et qu'elles soient toutes distribuées selon la même loi gaussienne (ou normale) autour de la moyenne μ. On peut ajuster ce modèle comme suit:

```
modele=lm(taille~1) # modele lineaire: correspond aux hypotheses sus-mentionnees
modele # renvoie l'estimation du parametre mu du modele
```

```

Call:
lm(formula = taille ~ 1)

Coefficients:
(Intercept)
      170.8
```

Partant de ce modèle, un intervalle de confiance pour μ est:

$$IC = \left[\hat{\mu} - 1.96 * \frac{sd(Y)}{\sqrt{n}}, \hat{\mu} + 1.96 * \frac{sd(Y)}{\sqrt{n}}\right]$$

(Je vous/m'épargne la démonstration mathématique de ce résultat!)

Dans ce cas, au vu des données, l'estimation de μ est $\hat{\mu} = 170.8$, et il y a 95% de chances que la valeur de μ soit comprise entre 169.177cm et 172.3507cm:

```
confint(modele)
```

```
              2.5 %   97.5 %
(Intercept) 169.177 172.3507
```

Notez que l'incertitude sur cette estimation était évidemment beaucoup plus grande lorsque la taille d'échantillon était $n = 5$:

```
modele_bis=lm(quelques_tailles~1)
confint(modele_bis)
```

```
              2.5 %   97.5 %
(Intercept) 158.4956 183.1044
```

1.3 Deuxième exemple d'inférence statistique: détection de l'existence d'un lien entre deux variables

Considérons maintenant le lien entre la variable X et la variable Y.

```
tapply(taille, sexe, "mean")
```

```
   femme    homme
164.7708 176.3495
```

Il semblerait, au vu de l'échantillon, que la taille moyenne diffère selon que l'on s'intéresse aux hommes ou aux femmes...

On peut, comme on l'a fait auparavant, essayer d'estimer la taille moyenne, cette fois-ci en fonction du sexe

```
modele=lm(taille~sexe+0)
confint(modele)
```

```
             2.5 %   97.5 %
sexefemme 162.8019 166.7398
sexehomme 174.4486 178.2504
```

Ici, le fait que les intervalles de confiance soient disjoints tend à prouver que les tailles moyennes sont effectivement différentes en fonction du sexe. Néanmoins, si les intervalles de confiance se recouvraient en partie, ils ne nous permettraient pas d'affirmer si, oui ou non, le facteur de groupement X a *vraiment* un effet, en moyenne, sur la variable Y.

Une autre forme d'inférence statistique vise précisément à extrapoler à l'ensemble de la population le fait qu'il existe un effet de X sur Y. Il s'agit des **tests d'hypothèse**.

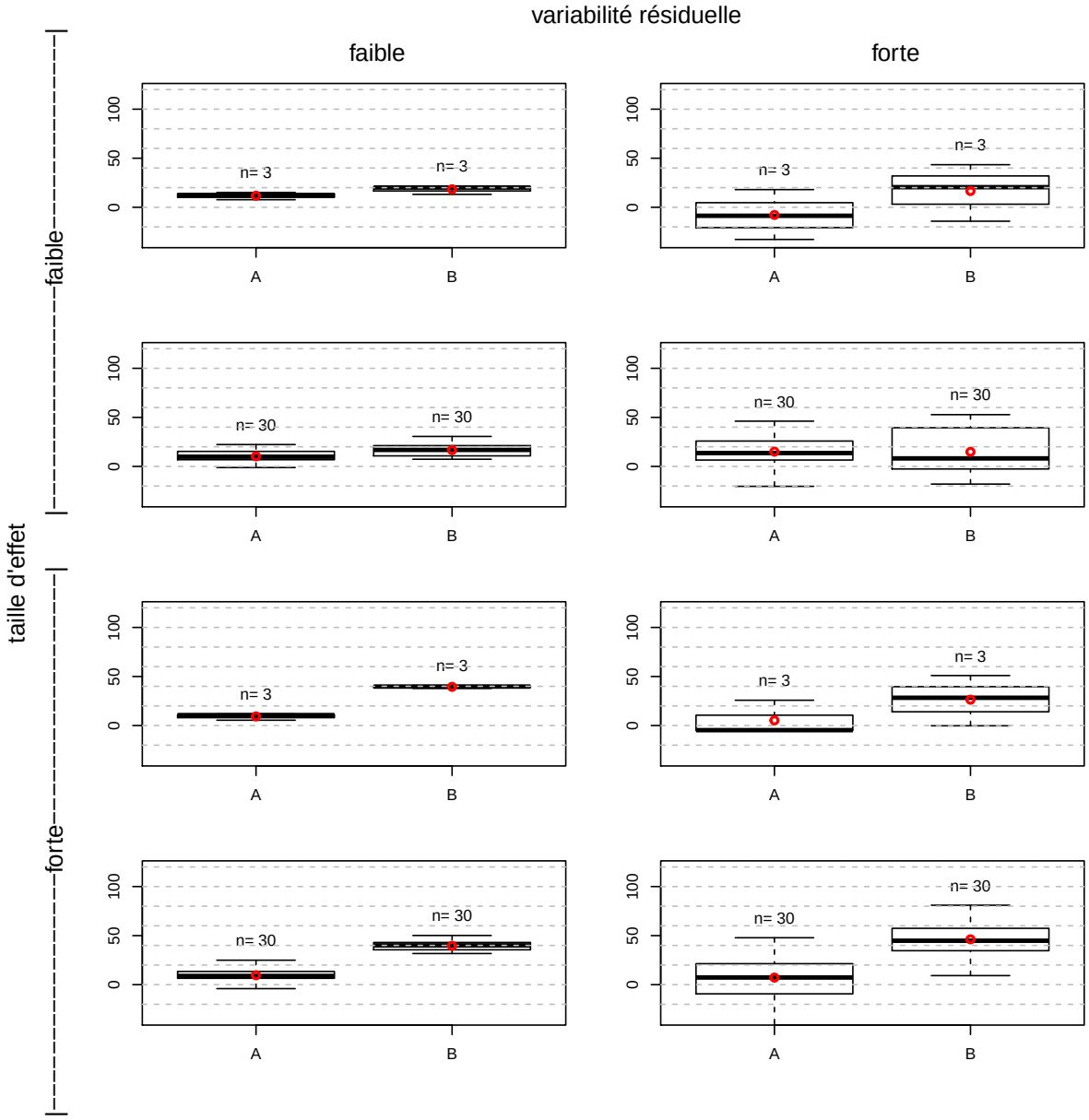
Les tests d'hypothèse s'appuient sur un **modèle statistique** qui décrit Y en fonction de X. Un tel modèle s'accompagne typiquement des questions suivantes:

- En termes d'**effet**: quel est l'effet de X sur Y ?
- En termes de **significativité**: Est-ce que l'effet observé de X sur Y est significatif, ou au contraire pourrait-il simplement s'expliquer par le hasard d'échantillonnage? C'est à cette question que le test d'hypothèse est censé apporter une réponse.
- En termes de **prédiction** et de **qualité d'ajustement**: Serait-on en mesure de prédire la valeur de Y, connaissant celle de X, avec précision? Y est-elle "étroitement" liée à X ou juste "vaguement"?). Autrement dit, la **variabilité résiduelle** des observations par rapport au modèle proposé est-elle forte?

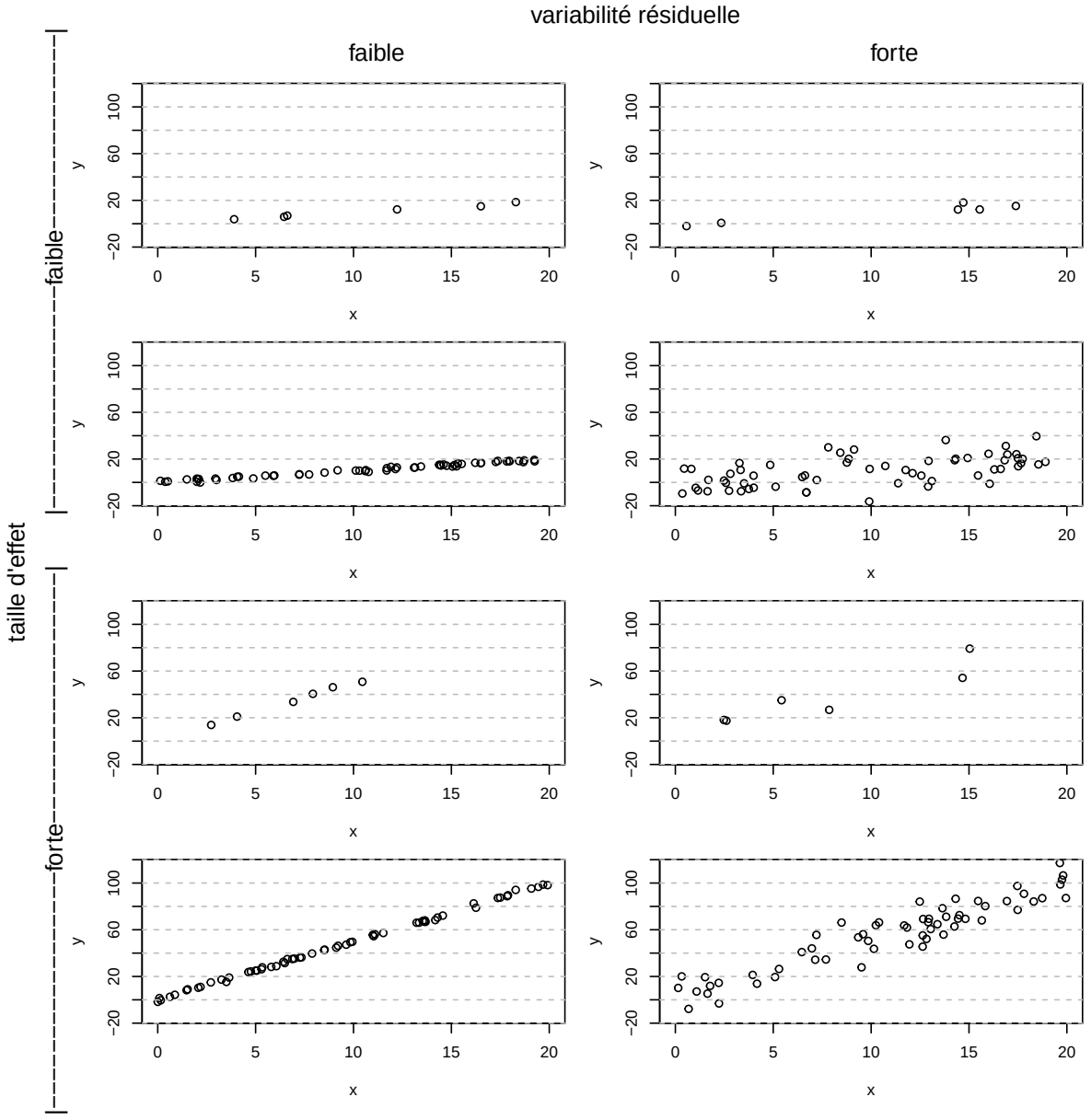
Les figures 1 et 2 vous permettent d'apprécier "intuitivement" l'influence de

- la taille d'échantillon
- la taille d'effet
- la variabilité résiduelle

sur la capacité à extrapoler l'effet observé de X sur Y à la population.



Exemples de relations entre les variables X (qualitative) et Y (quantitative). On a fait varier ici 3 éléments: la taille d'effet, la variabilité résiduelle, et la taille d'échantillon.



Exemples de relations entre les variables X et Y (toutes deux quantitatives. On a fait varier ici 3 éléments: la taille d'effet, la variabilité résiduelle, et la taille d'échantillon.

Exercice 1

Un premier exercice pour se "remettre dans le bain"...

- Tracez les boîtes à moustache représentant les poids des individus en fonction du sexe
- Calculez la moyenne et la variance de la variable poids, en fonction des groupes définis par sexe
- Quel est l'intervalle de confiance pour la moyenne de la variable poids, pour les femmes d'une part, et pour les hommes d'autre part?

2 Tests d'hypothèse

Intéressons-nous désormais plus particulièrement aux tests d'hypothèse. Comme précisé dans l'introduction, l'objectif des tests d'hypothèse est de déterminer si un effet observé à l'échelle de l'échantillon est extrapolable à l'ensemble de la population.

2.1 Un exemple de test d'hypothèse: le t-test

Le **t-test** (ou **test de Student**) est conçu pour tester des différences de moyenne entre deux groupes.

Voici un premier exemple concret de test d'hypothèse: On cherche à savoir si le sexe a bien un effet sur la taille.

Si j'essaie d'expliciter quel est mon modèle pour les données, alors je peux dire la chose suivante: "La taille est une variable qui suit une loi gaussienne avec une certaine moyenne μ_f et un certain écart-type s_f pour les femmes, et avec une certaine moyenne μ_h et un certain écart-type s_h pour les hommes" (et si je veux être tout-à-fait précise, je dois ajouter: "Les observations sont indépendantes les unes des autres").

C'est à l'aide de ce modèle que je vais tester l'hypothèse $H_0 : \{\mu_f = \mu_h\}$

Autrement dit, l'hypothèse que je cherche à tester statistiquement est celle selon laquelle "La taille moyenne ne diffère pas selon les groupes", alors que mon hypothèse au sens "scientifique" est plutôt l'hypothèse inverse!... Il faut donc faire attention à ne pas se mélanger les pinceaux...

2.1.1 Le t-test en pratique

En pratique, voilà comment tester cette hypothèse H_0 avec R:

```
t.test(taille~sexe)
```

```
    Welch Two Sample t-test

data:  taille by sexe
t = -8.3188, df = 192.47, p-value = 1.602e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.323946  -8.833417
sample estimates:
mean in group femme mean in group homme
           164.7708            176.3495
```

Plusieurs informations s'affichent:

- la nature exacte du test (t-test de Welch pour deux échantillons)
- certains éléments (valeur de métrique observée t, nombre de degrés de liberté df) qui interviennent dans le calcul d'une **p-value**
- la p-value elle-même
- l'hypothèse alternative (et non pas H_0) selon laquelle la différence de moyenne entre les groupes **n'est pas** égale à 0
- un intervalle de confiance à 95 % pour cette différence de moyenne
- les moyennes estimées pour chaque groupe

**Plus la p-value est petite, plus on va avoir tendance à rejeter l'hypothèse H_0...

Ici, la p-value étant extrêmement faible, on est très clairement en mesure de rejeter H_0, autrement dit de rejeter l'hypothèse selon laquelle les moyennes sont les mêmes dans les deux groupes... C'est donc que **l'effet observé du sexe sur la taille est significatif**. En pratique, le seuil de significativité, en dessous duquel on rejette l'hypothèse H_0, est très souvent, par convention, de $\alpha = 5\%$ (parfois 1%, parfois 10%).

2.2 Le t-test en théorie

Considérons la métrique suivante:

$$T = \frac{\bar{X_f} - \bar{X_h}}{\sqrt{\frac{s_f^2}{n_f} + \frac{s_h^2}{n_h}}}$$

où

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} X_i^2$$

Cette métrique est donc d'autant plus grande (en valeur absolue) que

- l'écart entre moyennes est important,
- les tailles d'échantillons sont importantes,
- la variance au sein de chaque groupe est petite.

Calculons sa valeur pour les observations:

```
Xbar=as.vector(tapply(taille,sexe,"mean"))
s2=as.vector(tapply(taille,sexe,"var"))
n=as.vector(tapply(taille, sexe,"length"))
Xbar_f=Xbar[1]
Xbar_h=Xbar[2]
s2_f=s2[1]
s2_h=s2[2]
n_f=n[1]
n_h=n[2]
t=(Xbar_f-Xbar_h)/sqrt(s2_f/n_f+s2_h/n_h)
print(t)
```

```
[1] -8.318833
```

On retrouve la valeur t affichée lors de l'exécution du test t par R.

Pour le t-test de Welch, on utilise également la métrique suivante :

$$\nu = \frac{\left(\frac{s_f^2}{n_f} + \frac{s_h^2}{n_h}\right)^2}{\frac{s_4^2}{n_f^2(n_f-1)} + \frac{s_h^4}{n_h^2(n_h-1)}}$$

```
A=(s2_f/n_f+s2_h/n_h)^2
B=s2_f^2/(n_f^2*(n_f-1))+s2_h^2/(n_h^2*(n_h-1))
print(A/B)
```

```
[1] 192.4672
```

On retrouve la valeur *df* ("degrees of freedom" = "degrés de liberté") affichée lors de l'exécution du test t par R.

D'après un théorème démontré par Welch en 1947, si l'hypothèse selon laquelle les moyennes sont identiques est vraie alors la métrique T doit suivre une loi de Student à ν degrés de liberté.

Si l'on compare la valeur observée de t (-8.31) à la loi de Student à ν (192.4673) degrés de liberté alors on s'aperçoit que cette observation est plutôt "extrême". En effet, la probabilité d'observer une valeur de t au moins aussi extrême (ou p-value) est de 1.6e-14 seulement.

```
# pt: loi de répartition de Student
A1=pt(q=-8.318833,df=192.46)
A2=1-pt(q=8.318833,df=192.46)
A=A1+A2
print(A)
```

```
[1] 1.600528e-14
```

Par conséquent on rejette l'hypothèse H_0.

Exercice 2

Est-ce que l'influence du sexe sur le poids des individus est significative? Quelle est la taille de l'effet?

2.3 Principe général d'un test d'hypothèse *paramétrique*

- On considère un modèle statistique décrivant nos données
- On considère une certaine hypothèse H_0 concernant les paramètres du modèle
- On considère une métrique (ou *variable aléatoire*) S, dont la nature dépend du test réalisé.
- On calcule sa **distribution théorique** en se basant sur le modèle assorti de l'hypothèse H_0.
- On calcule la valeur prise par S pour les observations: on obtient la mesure S_{obs}. S_{obs} est une *réalisation* de la variable S.
- On regarde où se place S_{obs} par rapport à distribution théorique de S: on calcule la probabilité que S soit au moins aussi "extrême" que S_{obs}. C'est cette probabilité qui constitue la **p-value**.

Il s'agit en fait d'une logique assez proche d'un raisonnement par l'absurde. Si, partant de l'hypothèse H_0, on calcule que la valeur de métrique observée est "improbable", alors c'est que c'est l'hypothèse H_0 qui est improbable...

Dans tous les cas, lorsque l'on réalise un test statistique, il faut être très attentif aux éléments suivants:

- Les hypothèses du modèle sous-jacent au test (par exemple, pour le t-test, la variable d'intérêt doit être de distribution gaussienne. En revanche, le fait que vous utilisiez le t-test de Welch et non de Student vous permet de supposer que la variance est différente dans les deux groupes)
- La nature de l'hypothèse H_0 (Si vous vous trompez sur la nature de cette hypothèse, alors vous vous tromperez dans la manière dont vous interprétez les résultats du test!)

Maintenant que nous avons abordé la logique "générale" du test d'hypothèse, intéressons-nous plus particulièrement à deux modèles (celui de l'**ANOVA**, et de la **régression linéaire**). Là où le t-test permettait de tester des différences de moyennes entre deux groupes seulement, l'ANOVA permet de considérer un nombre de groupes illimité.

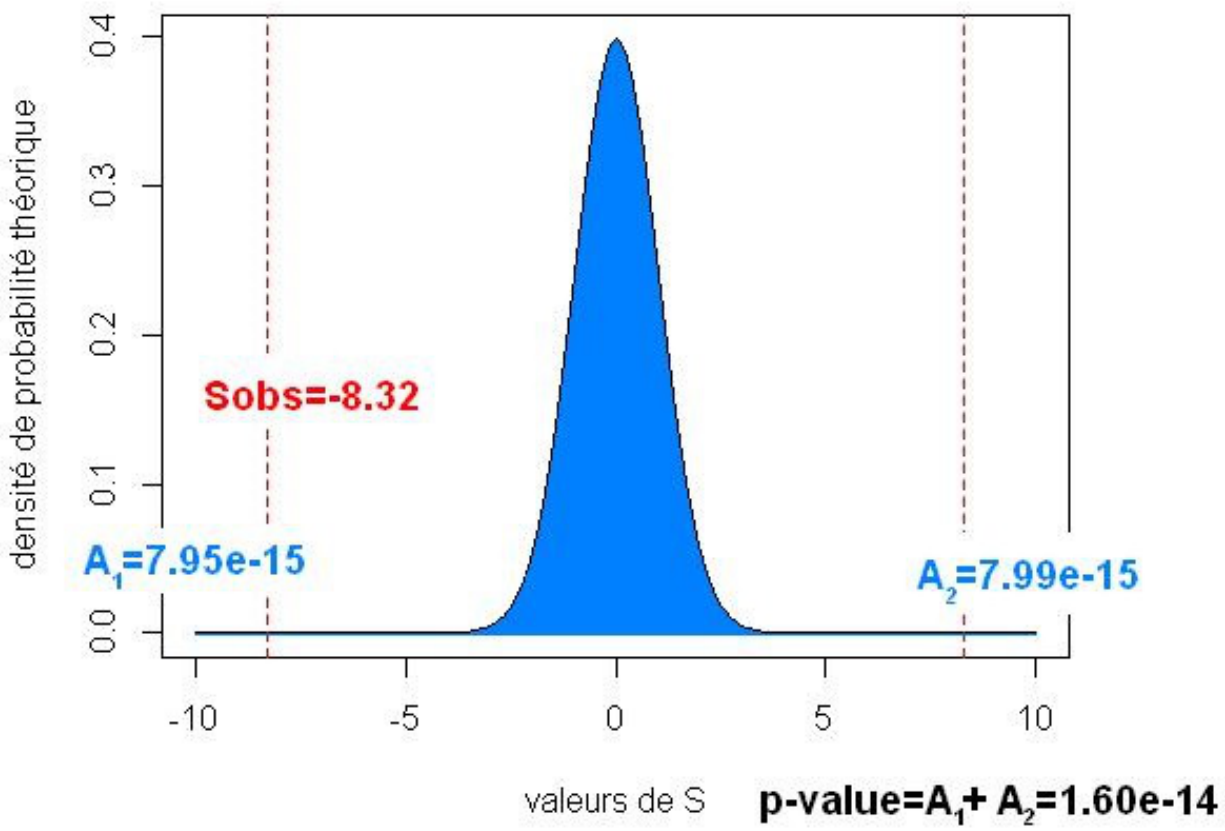


Figure 1: Calcul de p-value dans le cadre d'un test d'hypothèse. Exemple du test de Welch concernant l'effet du sexe sur la taille.

2.4 Quelques simulations pour mieux comprendre

En matière de statistiques, si le détail mathématique sous-jacent nous échappe, il est toujours possible de vérifier et explorer "empiriquement" les méthodes, à travers des simulations...

Ainsi, pour aider à la compréhension de ces tests, j'ai construit une petite application Shiny qui est accessible ici:

https://r-atique.shinyapps.io/Teaching_ttests/

Considérons ainsi

- une variable catégorielle définissant deux groupes, A et B, d'effectifs N_A et N_B respectivement
- une variable quantitative gaussienne Y, de moyenne et écart-type (μ_A,σ_A) pour le groupe A et (μ_B,σ_B) pour le groupe B.

En changeant les différents paramètres (tailles d'échantillons, différence entre les moyennes des groupes, écarts-types) on peut tester la capacité qu'a le t-test de détecter la différence de moyenne (i.e. sa puissance)... ou bien au contraire s'interroger sur le risque qu'il détecte une différence alors qu'il n'y en a pas (son risque d'erreur de type I).

Cas	H_0 rejetée	H_0 acceptée
H_0 vraie	erreur de type I (α)	CORRECT
H_0 fausse	CORRECT	erreur de type II (β)

Lorsque l'on réalise un test, on le fait à un niveau d'erreur de type I donné, i.e. on accepte un certain **risque** α, (par exemple de 5%) de rejeter l'hypothèse H_0 quand bien même elle serait vraie.

La **puissance** est égale à 1-β i.e. elle correspond à la probabilité de détecter un effet quand il existe (H_0 est fausse et on rejette H_0).

Dans la mesure où le résultat du t-test varie d'une simulation à l'autre, il faut évidemment réaliser un grand nombre de simulations pour estimer, globalement, la puissance et le risque d'erreur du test... On peut évidemment réaliser un certain nombre de simulations de manière répétée "à la main" (i.e. en soumettant les mêmes commandes un certain nombre de fois). Dans l'appli, pour plus de commodité, vous pouvez cocher "simulate 1000 such datasets".

Vérifions un certain nombre de choses.

2.4.1 Si H_0 est vraie ($\mu_A = \mu_B$) et les conditions d'applicabilité du test sont respectées

Essayons avec:

- $N_A = N_B = 30$
- $\mu_A = \mu_B = 10$,
- $\sigma_A = \sigma_B = 1$,
- distribution gaussienne

Vous pouvez constater qu'il n'y a effectivement que environ 5% de chances de rejeter H_0 au seuil de 5%...

2.4.2 Si H_0 est fausse ($\mu_A \neq \mu_B$) et les conditions d'applicabilité du test sont respectées

Essayons avec:

- $N_A = N_B = 30$
- $\mu_A = 1$, $\mu_B = 2$,
- $\sigma_A = \sigma_B = 1$,
- distribution gaussienne

Dans ces conditions, l'hypothèse est rejetée (avec raison) dans environ 97% des cas: la puissance du test est très importante. Mais tentez d'augmenter les valeurs de σ_A et σ_B, ou de diminuer les tailles d'échantillon, et la puissance pourrait bien diminuer drastiquement!

L'erreur de type II et la puissance du test dépendent en fait de la taille d'effet (différence des moyennes), de la taille d'échantillon, et de la dispersion des variables autour de leur moyenne... Autant de choses que, justement, on ne connaît pas (sauf, évidemment, lorsque l'on fait des simulations comme ici...). C'est une des raisons pour lesquelles l'usage est de faire un test statistique *en fixant une valeur d'erreur de type I acceptable* (la valeur que l'on prend pour α) plutôt qu'en fixant une valeur de puissance minimale... Avec pour corollaire le fait qu'on limite le risque de détecter un effet qui n'existe pas à défaut de maximiser les chances de détecter un effet qui existe.

2.4.3 Si les conditions d'applicabilité du test ne sont PAS respectées

Par contre, si l'hypothèse de normalité n'est pas vérifiée, alors le risque d'erreur de type I n'est pas forcément celui annoncé...

Vous pouvez tester avec par exemple:

- $N_A = N_B = 30$
- $\mu_A = \mu_B = 1$
- $\sigma_A = \sigma_B = 5$,
- distribution Gamma

Ici, le risque d'erreur de type I tourne plutôt autour de 1%. On pourrait arguer qu'être en dessous du risque de type I, ce n'est pas vraiment un problème. Malheureusement cela veut aussi (vraisemblablement) dire qu'on est dans un cas où la puissance du test est particulièrement faible (vous pouvez tester en augmentant μ_B à 2 par exemple)

2.4.4 Si les conditions d'applicabilité du test ne sont PAS respectées, mais les données sont nombreuses...

Vous pouvez tester avec par exemple:

- $N_A = N_B = 1000$
- $\mu_A = \mu_B = 1$
- $\sigma_A = \sigma_B = 5$,
- distribution Gamma

Ici on constate que quand bien-même l'hypothèse de normalité n'est pas respectée, on a un risque d'erreur de type I très proche de celui recherché. Cela est dû au fait que le t-test est une méthode valide *asymptotiquement* (i.e. quand la taille d'échantillon "tend vers l'infini"), que les hypothèses d'applicabilité soient respectées ou non. Tout le problème est évidemment de décider à partir de quelle taille d'échantillon on est vers l'infini et au-delà (...).

J'ai abordé ce problème philosophico-psychologico-statistique dans le billet suivant:

https://perso.ens-lyon.fr/lise.vaudor/non-respect-des-hypotheses-du-modele-lineaire-anova-regression-cest-grave-docteur/

L'appli vous permet en outre de tester et mesurer la robustesse du t-test en faisant vos propres essais...

Exercice 3

Pour des tailles d'échantillons et variabilités résiduelles du même ordre de grandeur que celles observées pour les variables sexe et taille, quel est le risque réel de commettre une erreur de type I avec un t.test?

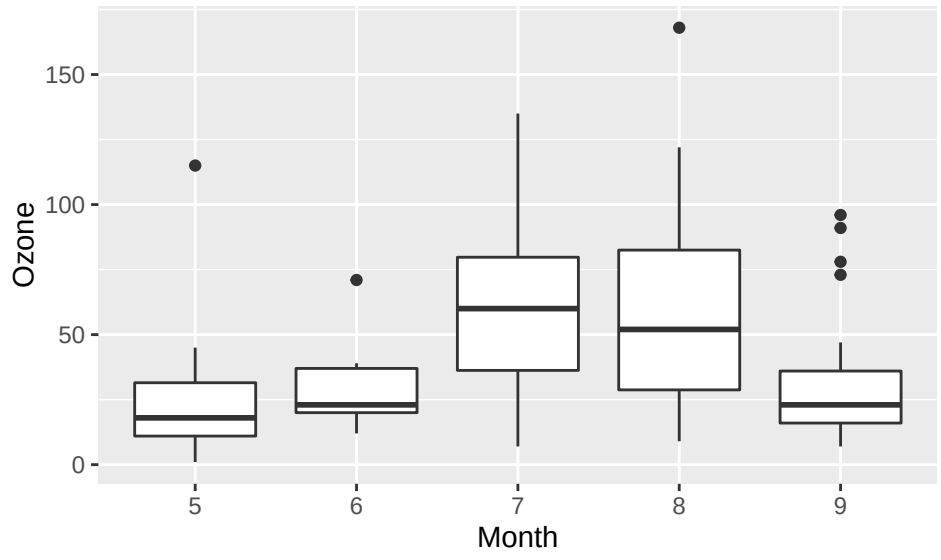
Pour des tailles d'échantillons, variabilités résiduelles, et taille d'effet du même ordre de grandeur que celles observées pour les variables sexe et taille, quel est la puissance du t.test?

3 Effet d'une variable catégorielle sur une variable quantitative: l'ANOVA

Exemple: on analyse l'influence du mois (X) sur la concentration en ozone dans l'air (Y) - données déjà utilisées lors du module d'initiation à R-. Pour cela, on se pose le même type de question que dans le cas d'une régression linéaire (en termes d'effet, de significativité et de qualité d'ajustement).

On cherche à quantifier les différences de moyennes en fonction de groupes (définis par les valeurs de la variable Month):

```
require(ggplot2)
qualitedelair<-read.table("../../datasets/donneesqualitedelair.txt", sep=" ", header=T)
attach(qualitedelair)
Month=as.factor(Month)
qualitedelair$Month=Month
#Month est un facteur (il faut le considérer comme
# variable qualitative et non quantitative)

qplot(data=qualitedelair,x=Month, y=Ozone,geom="boxplot")
```



3.1 Questions associées à une ANOVA

Pour cela, on se pose plusieurs questions:

- En termes d'**effet**: Comment Y évolue-t-elle en fonction de X?
- En termes de **significativité**: Est-ce que l'effet observé est significatif, ou au contraire pourrait-il simplement s'expliquer par le hasard?
- En termes de **prédiction** et de **qualité d'ajustement**: Serait-on en mesure de prédire la valeur de Y, connaissant celle de X, avec précision? Y est-elle étroitement liée à X?)

Réaliser une **ANOVA** va nous permettre de répondre à ces questions, en nous fournissant notamment les éléments de réponse suivants:

- l'**effet** (signe de l'effet et taille d'effet), indicateur de l'ampleur de l'influence de la variable explicative sur la variable réponse.
- la **p-value**, indicatrice de la **significativité} de l'effet
- le $\mathbf{R}^2$, ou coefficient de détermination, indicateur de la **qualité prédictive** et de la **qualité d'ajustement** du modèle

3.2 Réalisation de l'ANOVA

Observez la différence entre lm1 et lm2. L'appel de la fonction lm renvoie les moyennes par groupe (i.e. par niveau de facteur).

```
lm1=lm(Ozone~Month)
summary(lm1)
```

```
Call:
lm(formula = Ozone ~ Month)

Residuals:
    Min      1Q  Median      3Q     Max
-52.115 -16.823  -7.282  13.125 108.038

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.615      5.759   4.101 7.87e-05 ***
Month6         5.829     11.356   0.513    0.609
Month7        35.500      8.144   4.359 2.93e-05 ***
Month8        36.346      8.144   4.463 1.95e-05 ***
Month9         7.833      7.931   0.988    0.325
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.36 on 111 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.2352,    Adjusted R-squared:  0.2077
F-statistic: 8.536 on 4 and 111 DF,  p-value: 4.827e-06
```

En utilisant la première formule les moyennes sont évaluées relativement à un niveau de référence (en l'occurence, le premier): les éléments renvoyés par lm1 répondent donc à la question "les moyennes par groupe sont-elles significativement différentes de celle du premier groupe?"

```r
lm2=lm(Ozone~Month+0)
summary(lm2)
```

```
Call:
lm(formula = Ozone ~ Month + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-52.115 -16.823  -7.282  13.125 108.038

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
Month5   23.615      5.759   4.101 7.87e-05 ***
Month6   29.444      9.788   3.008  0.00325 **
Month7   59.115      5.759  10.266  < 2e-16 ***
Month8   59.962      5.759  10.412  < 2e-16 ***
Month9   31.448      5.453   5.768 7.36e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.36 on 111 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.7109,    Adjusted R-squared:  0.6979
F-statistic: 54.59 on 5 and 111 DF,  p-value: < 2.2e-16
```

En utilisant la formule avec "+0", les moyennes sont évaluées indépendamment les unes des autres : les éléments renvoyés par lm2 répondent donc à la question "les moyennes par groupe sont-elles significativement différentes de zéro?".

Plus classiquement, on va plutôt se poser les questions suivantes:

- "le facteur (dans son ensemble) a-t-il un effet sur la réponse?"
- "les moyennes par groupe sont-elles significativement différentes deux à deux?" (i.e. dans ce cas on ne considère pas qu'il y a un seul groupe de référence)

Pour répondre à la première question, on peut appeler la fonction "anova" (appliquée à un objet de type "modèle linéaire") ou la fonction "aov" (remarquez que les deux lignes de commande suivantes renvoient exactement les mêmes résultats):

```r
anova(lm(Ozone~Month))
summary(aov(Ozone~Month))
```

```
Analysis of Variance Table

Response: Ozone
           Df Sum Sq Mean Sq F value    Pr(>F)
Month       4  29438  7359.5  8.5356 4.827e-06 ***
Residuals 111  95705   862.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

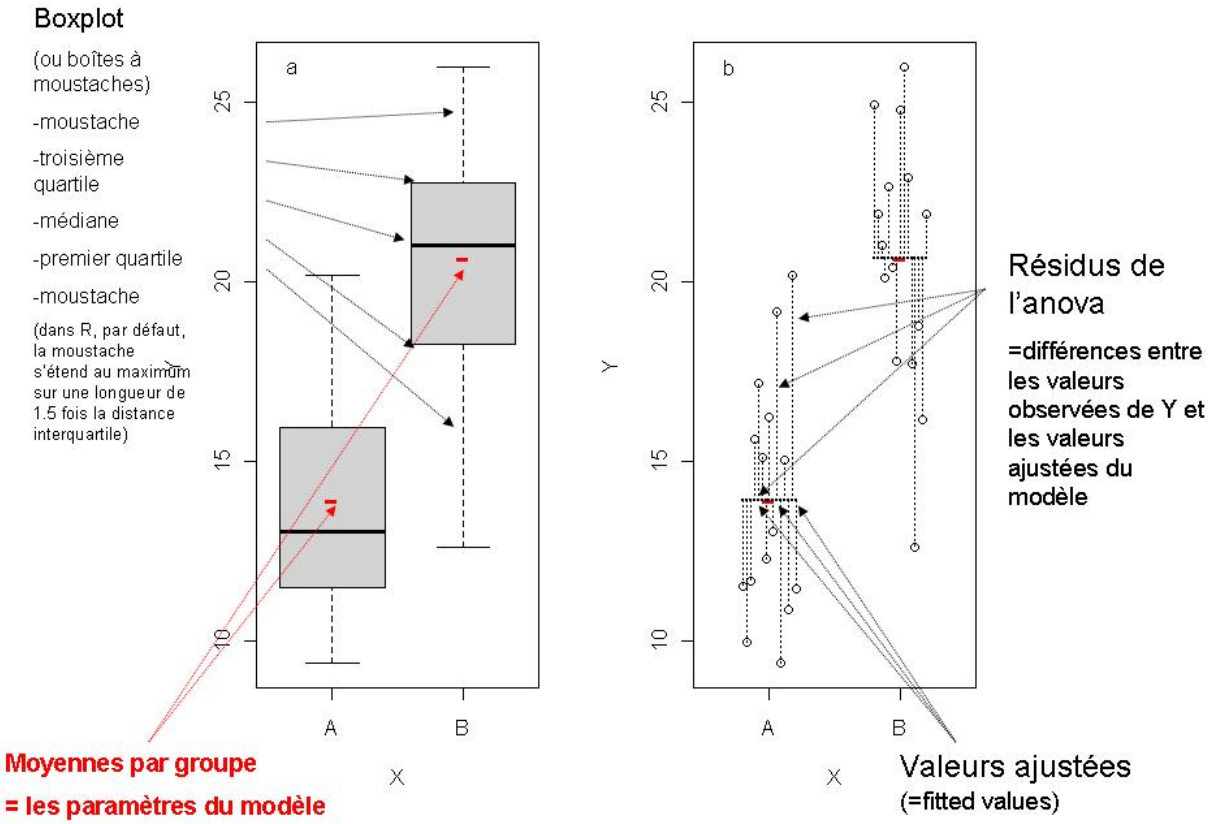
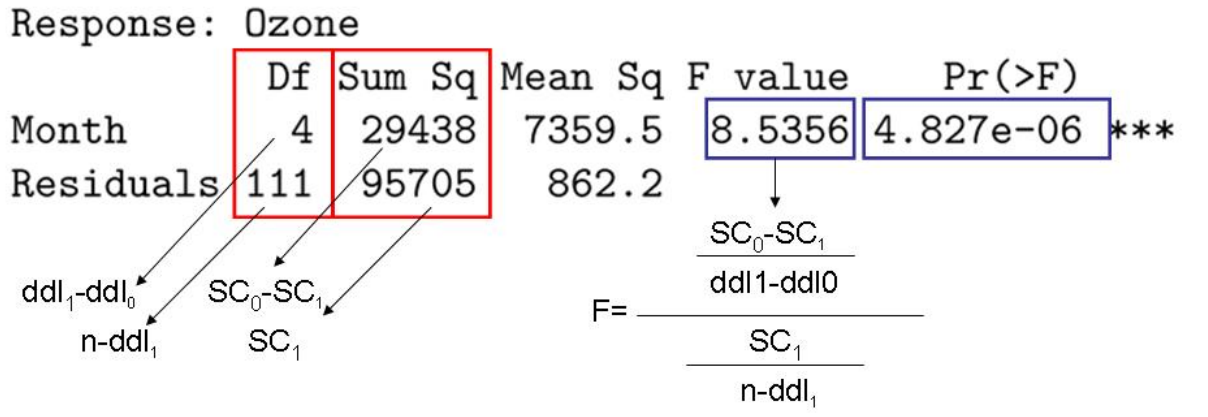


Figure 2: Illustration d'une ANOVA et du vocabulaire associé: a) représentation classique d'une ANOVA. b) représentation "pédagogique" visant à illustrer la notion de résidus dans le cas d'une ANOVA



Degrés de liberté

Quantifie le rapport entre la taille d'échantillon et le nombre de paramètres du modèle

Taille d'échantillon => n=116

Modèle 0: 1 paramètre => ddl_0=n-1=115

Modèle 1: 5 paramètres => ddl_1=n-5=111

Sommes des carrés

Quantifie la qualité explicative du modèle (somme des carrés des résidus du modèle: d'autant plus petite que le modèle est « bon »)

Modèle 0: => SC_0=125143

Modèle 1: => SC_0-SC_1=29438

Statistique F

Quantifie l'apport explicatif du modèle 1 par rapport au modèle 0, en pondérant en fonction du nombre de paramètres utilisés par ces deux modèles

P-value

Permet de répondre à la question « est-ce que le modèle 1 est significativement meilleur que le modèle 0? ». Son calcul découle du calcul de la statistique F

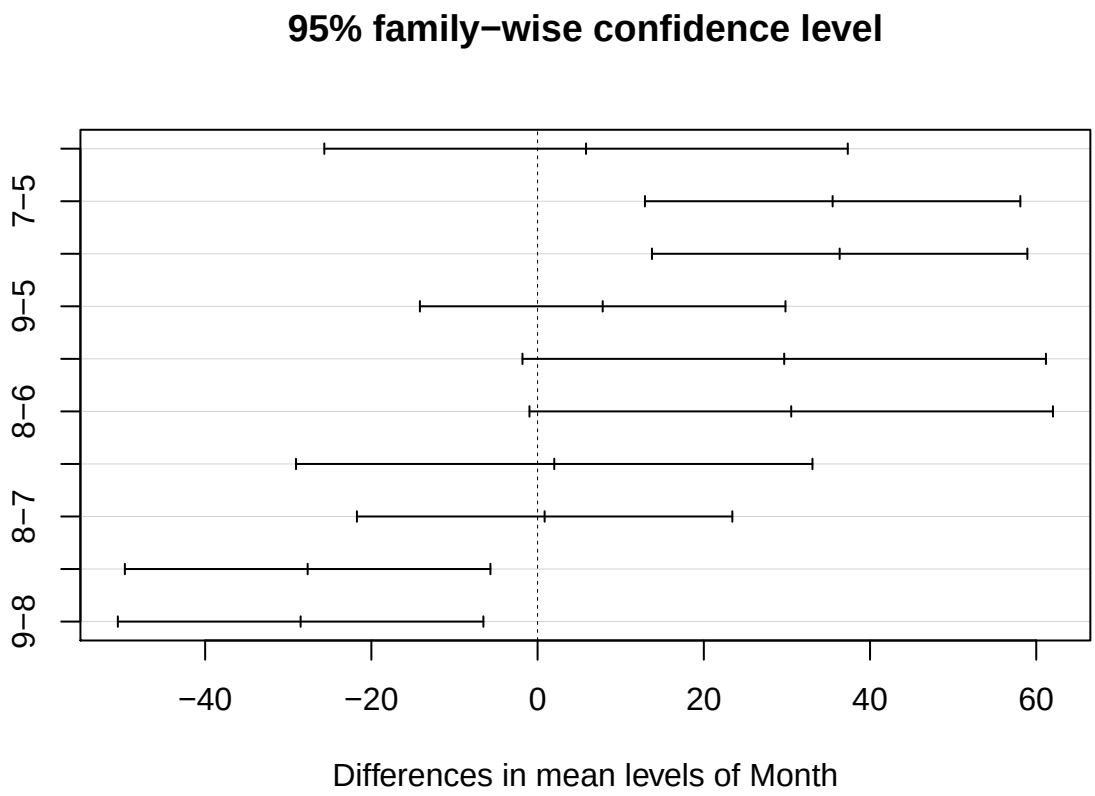
```
1-pf(8.5356, df1=4, df2=111)
```

Figure 3: Explication des différents éléments renvoyés par les fonctions anova(lm(Ozone $\sim$ Month)) et summary(aov(Ozone $\sim$ Month))

L'analyse ci-dessus nous confirme que la variable Month a bien un effet significatif sur la variable Ozone.

Pour répondre à la deuxième question, on peut appeler la fonction "TukeyHSD" (appliquée à un objet de type "aov"):

```
diff_2_a_2=TukeyHSD(aov(Ozone~Month))
plot(diff_2_a_2)
```



La première ligne de commande ci-dessus renvoie la p-value associée au test de Tukey HSD (HSD pour "Honestly Significant Difference"). L'hypothèse testée, pour 2 catégories, est "La différence de moyenne entre les deux catégories est nulle". Ainsi, la commande plot(diff_2_a_2) affiche une estimation de la différence de moyenne entre deux groupes, ainsi qu'une estimation d'un intervalle de confiance à 95% pour cette différence. Comme pour un test F, une p-value proche de zéro nous amène à rejeter l'hypothèse selon laquelle il n'y a pas d'effet (i.e. l'hypothèse selon laquelle il n'y a pas de différence de moyenne significative entre les groupes).

3.2.1 Zoom sur le test F

Le test F permet de répondre à la question "la variable X a-t-elle un effet sur Y?", que la variable explicative soit numérique ou catégorielle.

Soit l'hypothèse H_0, selon laquelle "la variable X n'a pas d'effet sur Y". Le modèle 0 correspond à un modèle dans lequel Y est distribué autour d'une valeur moyenne. Soit l'hypothèse H_1, selon laquelle "la variable X a un effet sur Y". Le modèle 1 correspond à un modèle dans lequel Y est distribué autour de plusieurs moyennes différentes selon les catégories décrites par X.

Soit la statistique F:

$$\frac{\left(\frac{SC_0 - SC_1}{ddl_0 - ddl_1}\right)}{\left(\frac{SC_1}{ddl_1}\right)}$$

où

- SC_0=somme des carrés des écarts au modèle 0
- SC_1=somme des carrés des écarts au modèle 1
- ddl_0=nombre de degrés de liberté du modèle 0
- ddl_1=nombre de degrés de liberté du modèle 1

Supposons que l'hypothèse H_0 est vraie.

Dans ce cas, F suit une distribution F avec (ddl_0-ddl_1, ddl_0) degrés de liberté, et on peut calculer la probabilité d'observer la valeur obtenue pour la statistique F (i.e. la p-value). Si cette p-value est faible, cela signifie que ce qu'on observe est peu vraisemblable sous l'hypothèse que l'on a faite au départ. On rejette donc cette hypothèse. Pour ceux un peu familiers avec la logique mathématique, ce raisonnement est assez semblable à un raisonnement par l'absurde. En faisant une hypothèse, on aboutit à une conclusion absurde ou du moins peu vraisemblable. On en conclut que l'hypothèse de départ est erronéee.

3.2.2 Zoom sur le R^2

Le R^2 permet de caractériser la qualité prédictive et la qualité d'ajustement du modèle. En conservant les notations SC_1 et SC_0 pour la somme des carrés des écarts au modèle 0 et 1 (cf encadré sur le test F), on a:

$$R^2 = \frac{SC_1}{SC_0}$$

Autrement dit, le R^2 est le rapport entre

- la somme des carrés des résidus du modèle 1 (i.e., la variabilité des observations autour d'un modèle qui prend en compte l'effet d'une variable X)
- la somme des carrés des résidus du modèle 0 (i.e., la variabilité "globale" des données, indépendamment de toute variable explicative.)

3.3 Et au fait, elles étaient vérifiées les hypothèses du modèle linéaire?

C'est bien le moment de se poser la question, me direz-vous...

Eh bien, en l'occurence, oui, c'est bien le moment de se poser la question. En effet, ces hypothèses sont que les résidus du modèle sont distribuées de manière gaussienne, et qu'ils sont homoscédastiques. Or, les résidus "n'existent pas" tant qu'on a pas défini et calculé les paramètres du modèle... Il faut donc ajuster le modèle (moyennant quelques lignes de code) avant de se demander s'il est valide...

Examinons donc les résidus du modèle. Globalement, si les résidus sont, grosso modo, gaussiens, l'hypothèse d'homoscédasticité en revanche n'est pas respectée. Néanmoins, compte tenu du nombre de points de mesure, et de la grandeur l'effet du mois sur la concentration en ozone (qui apparaît clairement sur le graphe), un modèle linéaire peut, a priori, être utilisé sans trop d'inquiétudes... L'exercice suivant vise à vérifier cela de manière un peu moins hasardeuse:

Exercice 3

Le code suivant vise à vérifier quel est le taux réel d'erreur de type I dans notre contexte. Décortiquez ce code pour comprendre les calculs réalisés et adaptez-le pour en déduire la puissance du test dans le cas où la différence maximale de température entre deux groupes serait de 5 degrés.

```r
pval=rep(NA,100)
lMonth=levels(Month)
Ozone_sim=rep(NA,length(Ozone))
ma_moy=mean(Ozone,na.rm=T)
for (k in 1:100){
  for (i in 1:5){
    ind=which(Month==lMonth[i])
    mon_sd=sd(Ozone[ind],na.rm=T)
    mon_effectif=length(ind)
    sous_echant=rnorm(mon_effectif,ma_moy,mon_sd)
    Ozone_sim[ind]=sous_echant
  }
  pval[i]=anova(lm(Ozone_sim~Month))$"Pr(>F)"[1]
}
length(which(pval<0.05))/100
```

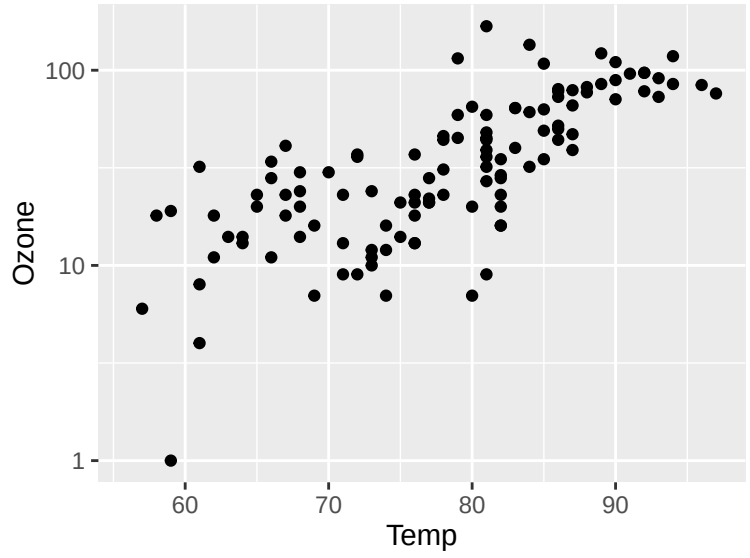
```
[1] 0
```

4 Effet d'une variable quantitative sur une autre: régression linéaire

4.1 Questions associées à une régression linéaire

Exemple: On analyse l'influence de la température (X) sur la concentration en ozone dans l'air (Y).

```r
p=ggplot(data=qualitedelair,aes(x=Temp, y=Ozone))
p=p+scale_y_log10()+geom_point()
plot(p)
```



Pour cela, on se pose plusieurs questions:

- En termes d'**effet**: Comment Y évolue-t-elle en fonction de X?

- En termes de **significativité**: Est-ce que l'effet observé est significatif, ou au contraire pourrait-il simplement s'expliquer par le hasard? (pour prendre un exemple un peu caricatural, on conçoit bien que si l'on ne disposait que de 3 points de mesure, on pourrait "par hasard" observer $y_1 < y_2 < y_3$ pour $x_1 < x_2 < x_3$ sans que cela veuille dire nécessairement que X a un effet positif sur Y).
- En termes de **prédiction** et de **qualité d'ajustement**: Serait-on en mesure de prédire la valeur de Y, connaissant celle de X, avec précision? Y est-elle "étroitement" liée à X ou juste "vaguement"?)

Réaliser une **régression linéaire** va nous permettre de répondre à ces questions, en nous fournissant notamment les éléments de réponse suivants:

- l'**effet** (signe de l'effet et taille d'effet), indicateur du sens et de l'ampleur de l'influence de la variable explicative sur la variable réponse. Dans notre exemple, Y augmente quand X augmente: l'effet de X sur Y est positif
- la **p-value**, indicatrice de la **significativité** de l'effet
- le $\mathbf{R}^2$, ou coefficient de détermination, indicateur de la **qualité prédictive** et de la **qualité d'ajustement** du modèle

4.2 Réalisation d'une régression linéaire avec R

Ici, l'effet de la température sur la concentration en ozone n'est pas vraiment linéaire. Mieux vaut réaliser une transformation de type logarithmique de la variable Ozone pour pouvoir appliquer un modèle de régression linéaire sur les variables étudiées

```
reg=lm(log(Ozone)~Temp)
summary(reg)
```

```

Call:
lm(formula = log(Ozone) ~ Temp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14469 -0.33095  0.02961  0.36507  1.49421

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.83797    0.45100  -4.075 8.53e-05 ***
Temp         0.06750    0.00575  11.741  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5848 on 114 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.5473,    Adjusted R-squared:  0.5434
F-statistic: 137.8 on 1 and 114 DF,  p-value: < 2.2e-16
```

```
names(reg)          # affiche le nom de tous les éléments listés dans l'objet "reg"
```

```
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "na.action"     "xlevels"       "call"          "terms"
[13] "model"
```

```
reg$coefficients   # affiche l'élément "coefficients" de l'objet "reg"
```

```
(Intercept)        Temp
-1.83797055  0.06750275
```

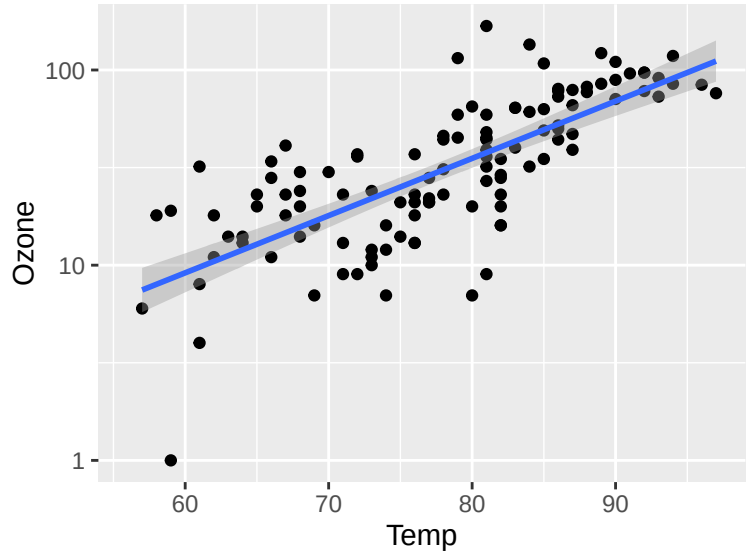
Les éléments renvoyés par summary(reg) correspondent bien à

- l'évaluation de l'effet (estimation des paramètres),
- la significativité de l'effet (p-value) évaluée à travers un test de Student,
- la qualité d'ajustement décrite par le R^2.

Le test de Student correspond à un test d'une **hypothèse nulle** du type H_0 ={le paramètre estimé est nul}. Une **p-value faible** (conventionnellement, inférieure à 5%) entraîne un **rejet de l'hypothèse nulle**. Dans notre cas, on a une valeur d'ordonnée à l'origine (Intercept) significativement non nulle, et une valeur de pente significativement non nulle. Nous reviendrons par ailleurs à la "philosophie" des tests d'hypothèse dans l'encadré sur le test F.

On peut représenter la droite de régression sur le nuage de points (Temp,Ozone) de la manière suivante:

```
p=p+geom_smooth(method="lm")
plot(p)
```



Graphique illustrant l'effet de la température sur la concentration en ozone. En rouge, la droite de régression.

4.3 Et au fait, elles étaient vérifiées les hypothèses du modèle linéaire?

Examinons donc les résidus du modèle. Globalement, si les résidus sont, grosso modo, gaussiens, l'hypothèse d'homoscédasticité n'est pas parfaitement respectée. Néanmoins, compte tenu du nombre de points de mesure, et de la grandeur l'effet de la température sur la concentration en ozone (qui apparaît clairement sur le graphe), un modèle linéaire peut, a priori, être utilisé sans trop d'inquiétudes...

(Ici je vous fais grâce d'une vérification à l'aide de simulations car elle serait légèrement plus difficile à réaliser que dans le cas de l'ANOVA...)

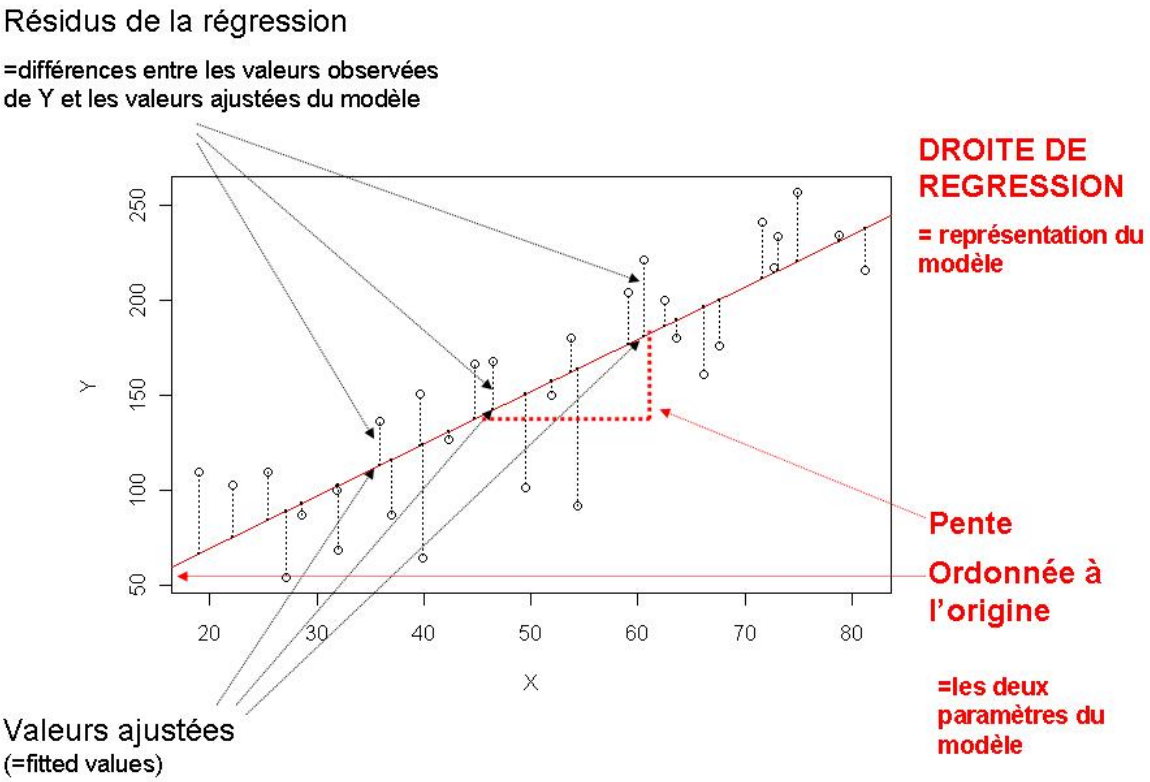


Figure 4: Illustration d'une régression linéaire et du vocabulaire associé

5 Influence de X_1,X_2,X_3... sur Y (quantitative, continue, gaussienne): modèle linéaire classique

Si les modèles multifactoriels reposent en grande partie sur des principes similaires à ceux des modèles "simples", ils sont légèrement plus difficiles à appréhender car il est difficile de produire des représentations graphiques de ces modèles. En effet, s'il est simple de représenter Y en fonction de X, il devient délicat de représenter Y en fonction de X_1 et X_2 (et, a fortiori, en fonction de X_1,X_2,X_3, etc.). Néanmoins, moyennant quelques efforts, et en utilisant astucieusement les possibilités graphiques de R, il est possible de visualiser si, oui ou non, de tels modèles "collent bien" aux observations. Nous allons voir des exemples de tels graphiques dans les paragraphes suivants.

Une autre difficulté inhérente aux modèles multifactoriels est liée aux possibles interactions entre facteurs, et aux possibles déséquilibres dans le plan factoriel. Ce sont deux problèmes que nous aborderons brièvement, en expliquant en quoi ils consistent et en indiquant quelques pistes (ou liens vers d'autres tutoriels) pour y remédier.

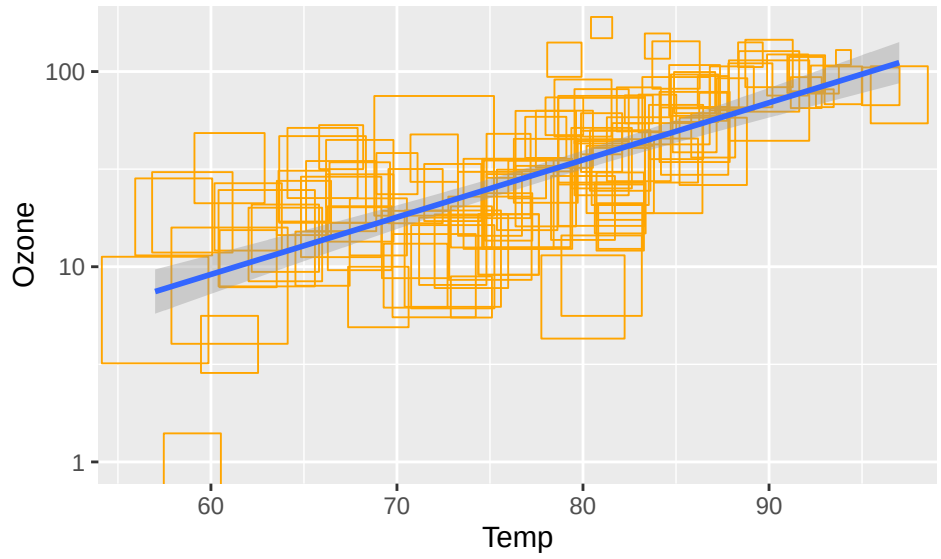
5.1 X_1,X_2,X_3... sont toutes quantitatives: régression linéaire multiple

Si l'on est dans le cas d'une régression multiple:

```
lm(log(Ozone)~Temp+Wind)
```

```
Call:
lm(formula = log(Ozone) ~ Temp + Wind)

Coefficients:
(Intercept)         Temp         Wind
   -0.53193      0.05738     -0.05253
```

On peut par exemple tracer le graphique suivant:



Ce graphe montre deux tendances: la concentration en Ozone dans l'air tend à être plus forte quand la température est plus forte, et la concentration en Ozone dans l'air tend à être plus faible quand la vitesse du vent est plus forte.

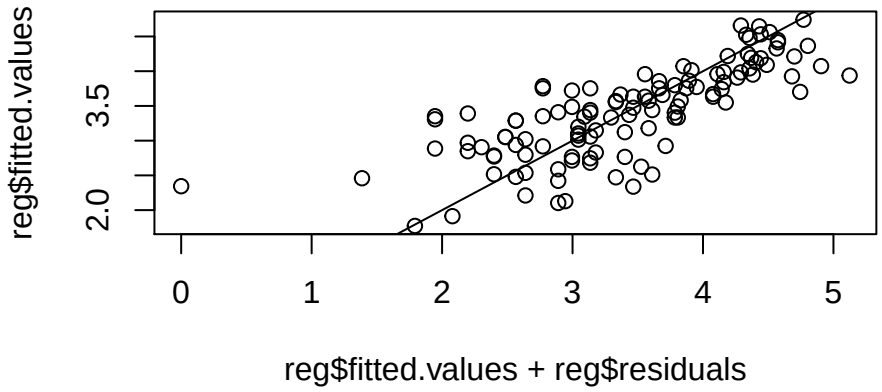
C'est ce que confirme la régression multiple suivante:

```
summary(lm(log(Ozone)~Temp+Wind))
```

```
Call:
lm(formula = log(Ozone) ~ Temp + Wind)

Residuals:
     Min       1Q   Median       3Q      Max
-2.34415 -0.25774  0.03003  0.35048  1.18640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.531932   0.608901  -0.874  0.38419
Temp         0.057384   0.006455   8.889 1.13e-14 ***
Wind        -0.052534   0.017128  -3.067  0.00271 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5644 on 113 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.5821,    Adjusted R-squared:  0.5747
F-statistic: 78.71 on 2 and 113 DF,  p-value: < 2.2e-16
```

(puisque le paramètre de pente associé à Temp est positif et celui associé à Wind est négatif).

Par ailleurs, pour voir si le modèle s'ajuste bien aux observations, on peut représenter les valeurs ajustées du modèle et les valeurs observées de la variable réponse:



Vérification de l'ajustement de la régression multiple

Observez au passage que les deux lignes de commande suivantes

```
summary(lm(log(Ozone) ~ Wind+Temp))
summary(lm(log(Ozone) ~ Month+Temp))
```

renvoient exactement le même résultat (si ce n'est que l'ordre des lignes change dans le tableau renvoyé):

```
Call:
lm(formula = log(Ozone) ~ Wind + Temp)

Residuals:
     Min       1Q   Median       3Q      Max
-2.34415 -0.25774  0.03003  0.35048  1.18640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.531932   0.608901  -0.874  0.38419
Wind        -0.052534   0.017128  -3.067  0.00271 **
Temp         0.057384   0.006455   8.889 1.13e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5644 on 113 degrees of freedom
  (37 observations deleted due to missingness)
Multiple R-squared:  0.5821,    Adjusted R-squared:  0.5747
F-statistic: 78.71 on 2 and 113 DF,  p-value: < 2.2e-16
```

Remarque: dans l'exemple ci-dessus, on a un problème dit de **colinéarité** car nos deux variables explicatives sont corrélées. Cela est gênant car le modèle auquel l'on s'intéresse quantifie l'effet de Temp sur Ozone, et de Wind sur Ozone. En cas de colinéarité, i.e. si Wind a également un effet sur Temp, alors Wind a un effet sur Ozone à la fois directement et indirectement. De même, en cas de colinéarité, Temp a un effet direct et indirect sur Ozone. Il devient alors difficile d'estimer l'effet "individuel" de chacune des variables explicatives. Concrètement, cela se traduit par des incertitudes plus importantes sur les paramètres estimés (incertitudes exprimées par l'erreur standard). Pour plus de détails sur ce problème, voir par exemple ici: http://www.gate.cnrs.fr/perso/fournier/Notes_de_cours/Econometrie/7_Multicolinearite.pdf

[Représentation simplifiée d'un problème de colinéarité]figures/colinearite.jpg}

5.2 X_1,X_2,X_3... sont toutes catégorielles: ANOVA à plusieurs facteurs

Imaginons le plan d'expérience suivant: On s'intéresse à la granulométrie de bancs de galets avant et après une crue importante. Quatre personnes participent aux campagnes terrains avant et après la crue. Chacune de ces quatre personnes va mesurer la taille d'un certain nombre de galets pris "au hasard", sur chacun des bancs considérés. (NB: pour les gens intéressés par l'étude des galets, je précise que les tables de données "galets_1" et "galets_2" sont des tables de données *simulées* et non *réelles*...)

5.2.1 Plan d'expérience en blocs complet et équilibré

Imaginons que chacune des quatre personnes a mesuré la taille de 10 galets, pour chacune des dates (avant/après) et chacun des bancs (A/B) considérés.

Le plan d'expérience est le suivant:

	Date	Banc	Experimentateur	Nb_galets
1	avant	banc A	Fanny	10.00
2	avant	banc A	David	10.00
3	avant	banc A	Guillaume	10.00
4	avant	banc A	Kristell	10.00
5	avant	banc B	Fanny	10.00
6	avant	banc B	David	10.00
7	avant	banc B	Guillaume	10.00
8	avant	banc B	Kristell	10.00
9	après	banc A	Fanny	10.00
10	après	banc A	David	10.00
11	après	banc A	Guillaume	10.00
12	après	banc A	Kristell	10.00
13	après	banc B	Fanny	10.00
14	après	banc B	David	10.00
15	après	banc B	Guillaume	10.00
16	après	banc B	Kristell	10.00

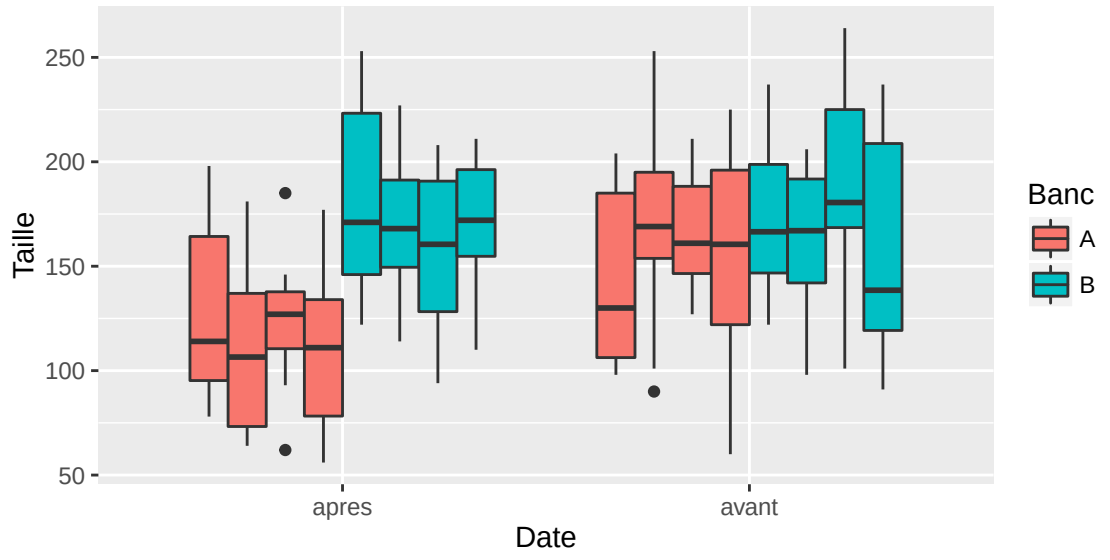
Ce plan est complet car des mesures ont été faites pour tous les cas possibles, en termes de compbinaisons de facteurs. Il est également équilibré car tous les cas correspondent à un même nombre de mesures.}

Supposons maintenant que l'on s'intéresse à l'effet du banc et de la date (avant ou après la crue), sur la taille des galets,

D'après ce graphique, il semblerait que Date et Banc aient tous deux un effet sur Taille. Supposons que l'on s'intéresse également à l'effet de l'expérimentateur. Ce faisant, on veut vérifier que la méthode de mesure est bien "standard" et qu'on n'observe pas de biais en fonction de la personne qui fait la mesure.

Notez, au passage, qu'au delà de 2 facteurs, cela devient difficile de produire des graphiques très clairs : pour vous en convaincre vous pouvez essayer

```
p=ggplot(data=galets_1,aes(x=Date,y=Taille))
p=p+geom_boxplot(aes(fill=Banc, position=Experimentateur))
plot(p)
```



Testons l'effet de ces différentes variables à travers une analyse de la variance (ANOVA). Dans ce cas, où le plan expérimental est équilibré, les commandes:

```
summary(aov(Taille~Date+Banc+Experimentateur,data=galets_1))
summary(aov(Taille~Date+Experimentateur+Banc,data=galets_1))
```

renvoient exactement le même résultat (si ce n'est que l'ordre des lignes dans le tableau change), à savoir que la variable "Experimentateur" n'a pas d'effet significatif sur la taille de galets mesurée (il n'y a donc a priori pas de biais lié à l'expérimentateur), tandis que les variables Date et Banc ont un effet significatif sur la taille de galets:

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
Date              1  15504   15504   8.458  0.00417 **
Experimentateur   3   2353     784   0.428  0.73329
Banc              1  41893   41893  22.855 4.05e-06 ***
Residuals       154 282283    1833
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les tableaux renvoyés sont, grosso modo, de la même forme que ceux renvoyés par l'anova à un facteur, si ce n'est qu'au lieu de comparer seulement 2 modèles, on compare 3 modèles successifs 2 à 2. En effet, avec la commande

```
summary(aov(Taille ~ Date+Banc+Experimentateur, data=galets_1))
```

```
                 Df Sum Sq Mean Sq F value   Pr(>F)
Date              1  15504   15504   8.458  0.00417 **
Banc              1  41893   41893  22.855 4.05e-06 ***
Experimentateur   3   2353     784   0.428  0.73329
Residuals       154 282283    1833
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

les résultats correspondent à:

- la comparaison des modèles 0 et 1 (1ère ligne du tableau),
- la comparaison des modèles 1 et 2 (2ème ligne du tableau),
- la comparaison des modèles 2 et 3 (3ème ligne du tableau),
- la description du modèle complet, i.e. le modèle 3, à travers la somme des résidus (4ème ligne du tableau)

où les modèles en question sont dits "emboîtés 2 à 2":

- {modèle 0}={Taille $\sim$ 1}
- {modèle 1}={Taille $\sim$ Date}
- {modèle 2}={Taille $\sim$ Date+Banc}
- {modèle 3}={Taille $\sim$ Date+Banc+Expérimentateur}

On peut ainsi généraliser la structure d'un tableau d'ANOVA à plusieurs facteurs de la manière suivante (ddl=degrés de liberté, SC=somme des carrés):

- modèle 0=$\{Y \sim 1\}$,
- modèle 1=$\{Y \sim X_1\}$,
- modèle 2=$\{Y \sim X_1 + X_2\}$,
- modèle 3=$\{Y \sim X_1 + X_2 + X_3\}$
- etc.

	Variable	ddl	SCsurddl	Fvalue	pvalue
1	Ajout de la variable X_1	$ddl_0 - ddl_1$	$\frac{SC_0 - SC_1}{ddl_0 - ddl_1}$	F_{0-1}	$pval_{0-1}$
2	Ajout de la variable X_2	$ddl_1 - ddl_2$	$\frac{SC_1 - SC_2}{ddl_1 - ddl_2}$	F_{1-2}	$pval_{1-2}$
3	Ajout de la variable X_3	$ddl_2 - ddl_3$	$\frac{SC_2 - SC_3}{ddl_2 - ddl_3}$	F_{2-3}	$pval_{2-3}$
4	Residus	n-Σddl			

5.2.2 Plan d'expérience en blocs non-complet et/ou non équilibré

Nous avons vu dans la section précédente un cas "idéal", dans lequel le plan d'expérience est complet et équilibré. En pratique, il arrive souvent que les plans d'expériences soient dèséquilibrés (par exemple si il y a davantage de quadrats mesurés après la crue qu'avant) ou même incomplets (par exemple si 2 utilisateurs se sont occupéés exclusivement du banc A, tandis que les deux autres s'occupaient exclusivement du banc B).

Intéressons-nous, notamment, au cas où le plan d'expérience est déséquilibré.

```
galets_2=read.table("../../datasets/galets_2.csv",sep=";", header=T)
summary(aov(Taille~Date+Banc+Experimentateur,data=galets_2))
```

```
                Df Sum Sq Mean Sq F value   Pr(>F)
Date             1  14478   14478   9.410  0.00255 **
Banc             1 110670  110670  71.930 1.71e-14 ***
Experimentateur  3    344     115   0.075  0.97357
Residuals      154 236942    1539
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(Taille~Date+Experimentateur+Banc,data=galets_2))
```

```
                Df Sum Sq Mean Sq F value   Pr(>F)
Date             1  14478   14478   9.410  0.00255 **
Experimentateur  3    174      58   0.038  0.99021
Banc             1 110841  110841  72.041 1.65e-14 ***
Residuals      154 236942    1539
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dans ce cas non-équilibré, les deux commandes ci-dessus ne renverront pas le même résultat. En effet, l'ordre dans lequel on ajoute les variables devient important. Or, comment choisir dans quel ordre on doit les ajouter?

On peut avoir recours à la fonction drop1:

```
drop1(aov(Taille~Date+Banc+Experimentateur,data=galets_2),test="F")
```

```
Single term deletions

Model:
Taille ~ Date + Banc + Experimentateur
               Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                      236942 1180.1
Date            1     13344 250286 1186.8  8.6728  0.003731 **
Banc            1    110841 347783 1239.5 72.0409 1.647e-14 ***
Experimentateur 3       344 237286 1174.3  0.0746  0.973575
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cette fonction renvoie un tableau qui ressemble un peu au tableau d'ANOVA tel que renvoyé par la commande summary(aov(Taille ∼ Date+Banc+Experimentateur,data=galets_2)). Néanmoins, il ne compare pas les mêmes modèles que ce premier tableau, puisqu'il compare au modèle 3 les modèles

- modèle 2a={Taille ∼ Banc+Experimentateur}
- modèle 2b={Taille ∼ Date+Experimentateur}
- modèle 2c={Taille ∼ Date+Banc}

i.e. le modèle complet (modèle 3) auquel on retire à chaque fois une variable explicative différente. Les sommes des carrés renvoyées par cette fonction correspondent aux sommes de carrés dits "de type III", par opposition aux sommes de carrés "de type I" renvoyés par la commande

```r
summary(aov(Taille~Date+Banc+Experimentateur,data=galets_2))
```

```
                Df Sum Sq Mean Sq F value   Pr(>F)
Date             1  14478   14478   9.410  0.00255 **
Banc             1 110670  110670  71.930 1.71e-14 ***
Experimentateur  3    344     115   0.075  0.97357
Residuals      154 236942    1539
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

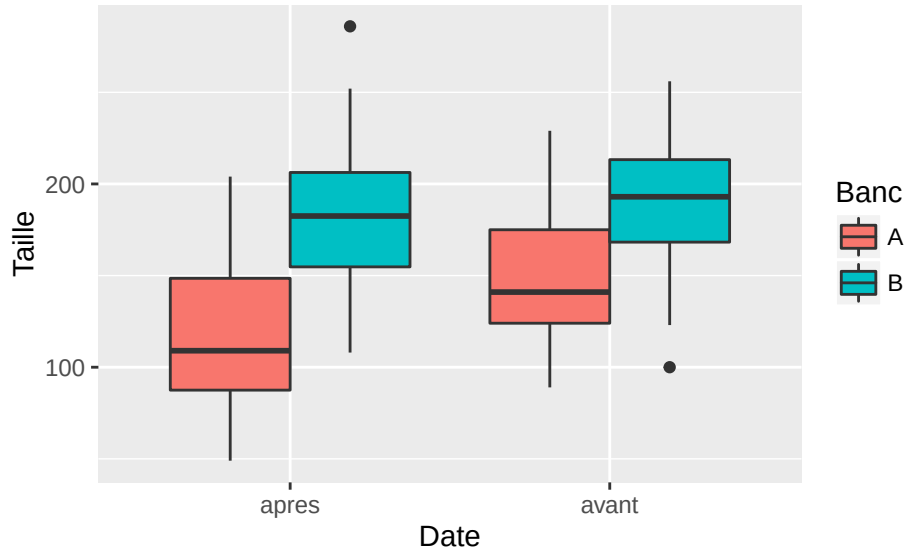
Les sommes de carrés de type III renseignent ainsi sur la part d'explication apportée par une variable en particulier (indépendamment des autres variables explicatives) dans un modèle.

Ainsi, on introduira dans le modèle testé par la fonction aov, et dans cet ordre, les variables Banc, Date, et Expérimentateur.

5.2.3 Interaction entre les variables explicatives

Le graphique renvoyé par

```r
p=ggplot(data=galets_2,aes(x=Date,y=Taille))
p=p+geom_boxplot(aes(fill=Banc))
plot(p)
```



suggère non seulement que Date et Banc ont un effet sur Taille, mais aussi que l'effet de la date est différent selon le banc (et inversement, l'effet du banc est différent selon la date). En d'autre termes, ce graphique suggère une interaction entre les deux facteurs.

On peut tester cette interaction de la manière suivante:

```
my_lm=lm(Taille~Date+Banc+Date*Banc, data=galets_2)
#équivalent à lm(Taille~Date*Banc)
anova(my_lm)
```

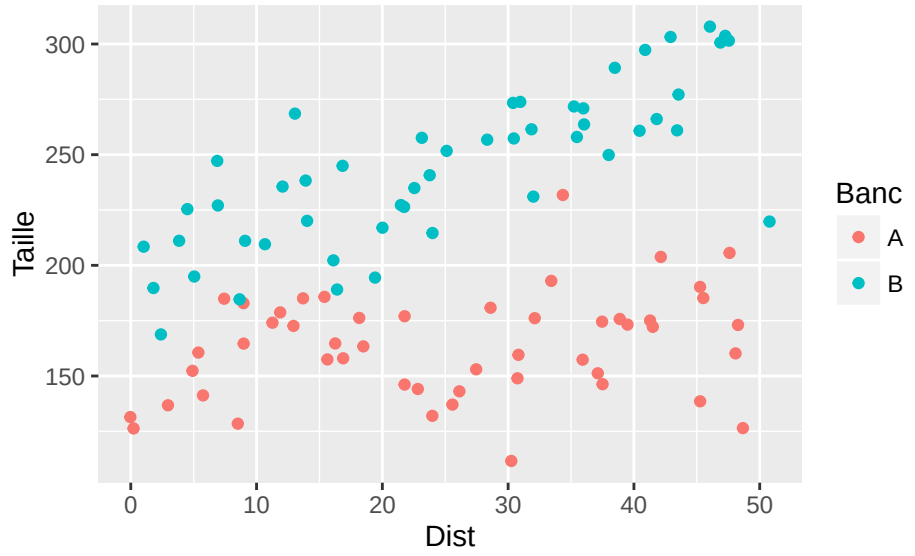
```
Analysis of Variance Table

Response: Taille
           Df Sum Sq Mean Sq F value    Pr(>F)
Date        1  14478   14478  9.8154  0.002067 **
Banc        1 110670  110670 75.0295 5.508e-15 ***
Date:Banc   1   7182    7182  4.8693  0.028799 *
Residuals 156 230104    1475
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Au passage, cela fait un moment que l'on n'a pas vérifié la validité de nos modèles linéaires... Examinons ce que cela donne pour le modèle {Taille ~ Date+Banc+Date*Banc}. Tout semble à peu près OK du côté des hypothèses...

5.3 X_1,X_2,X_3... sont quantitatives ou catégorielles: ANCOVA

Intéressons-nous maintenant à une autre expérience. On s'intéresse (encore!) à des tailles de galets (variable Taille), pour deux bancs différents (variable Banc), mais cette fois on a mesuré où chaque galet était placé au sein du banc (variable Dist: distance en m par rapport à la tête du banc). On cherche à savoir l'influence de la variable Banc (catégorielle) et de la variable Dist (quantitative) sur la variable Taille. On est donc dans le cadre d'une ANCOVA (Analyse de la COVAriance)

```
galets_3=read.table("../../datasets/galets_3.csv",
                    sep=";", header=T)
p=ggplot(data=galets_3,aes(x=Dist,y=Taille))
p=p+geom_point(aes(color=Banc))
plot(p)
```



Le graphique renvoyé par les lignes de commande ci-dessus suggèrent que:

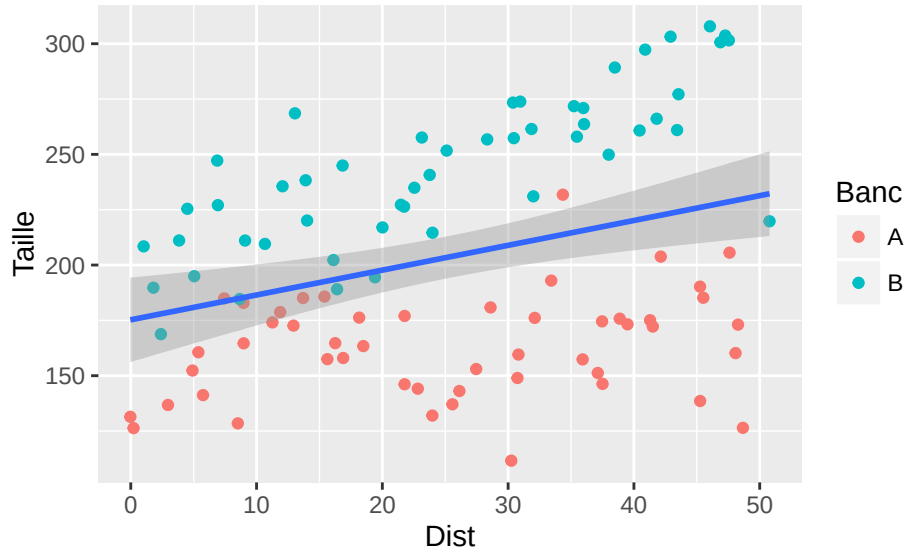
- Dist et Banc influencent Taille
- Dist influence Taille différemment selon la valeur de Banc

Vérifions cette intuition à l'aide d'une ANCOVA:

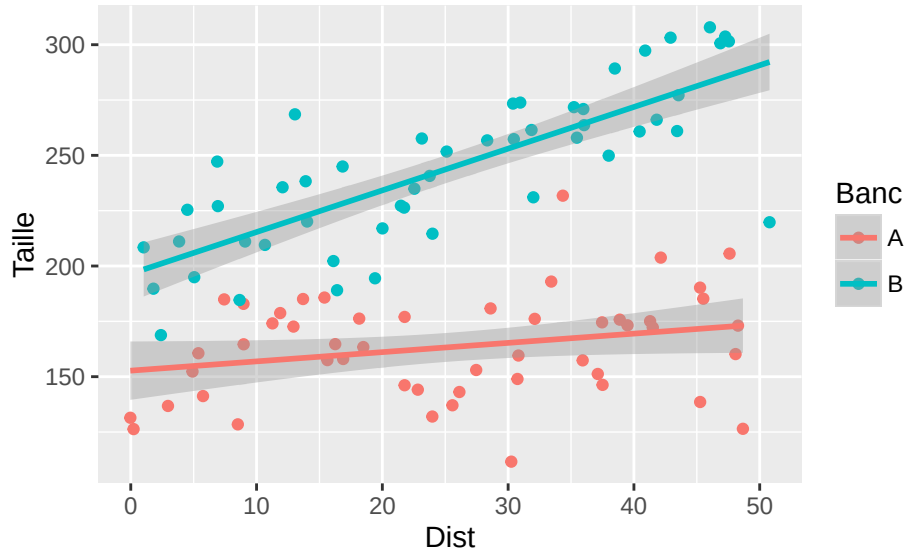
```
my_lm=lm(Taille~Banc+Dist+Banc*Dist,data=galets_3)
anova(my_lm)
```

```
Analysis of Variance Table

Response: Taille
          Df Sum Sq Mean Sq F value    Pr(>F)
Banc       1 162301  162301 319.296 < 2.2e-16 ***
Dist       1  27828   27828  54.746 5.193e-11 ***
Banc:Dist  1  11322   11322  22.274 8.039e-06 ***
Residuals 96  48798     508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
p=ggplot(data=galets_3,aes(x=Dist,y=Taille))
p=p+geom_point(aes(color=Banc))
p=p+geom_smooth(method="lm")
plot(p)
```




```
p=ggplot(data=galets_3,aes(x=Dist,y=Taille,color=Banc))
p=p+geom_point()
p=p+geom_smooth(method="lm")
plot(p)
```



Exercice 4

Expliquez la variable autonotation (données “representations”) par le poids, la taille, et le sexe des individus.