

# Démarche statistique

Rémi Mahmoud

[remi.mahmoud@agrocampus-ouest.fr](mailto:remi.mahmoud@agrocampus-ouest.fr)  
<https://demarche-stat-lesson.netlify.app>



# Introduction

# Objectifs

- Aborder les problèmes courants en analyse de données
- Argumenter le choix de procédures d'analyses
- Mettre en oeuvre une démarche d'analyse de données avec 
- Interpréter & restituer les résultats d'une analyse

On attend:

**Attention**



**Réflexion**



**Participation**



**Evaluation:**

2 CC (50%) + 1 projet (50%)

# Quelques définitions (Wiki)

- **LA statistique:** discipline qui étudie les phénomènes à travers la collecte de **données**, leur traitement, leur analyse, l'interprétation et la présentation des résultats [...]. **Domaine des mathématiques + boîte à outils.** Fait partie de la **science des données**
- **LES statistiques:** type d'information obtenu en soumettant les valeurs à des opérations mathématiques.
- **L'analyse des données:** Famille de méthodes statistiques dont les principales caractéristiques sont d'être multidimensionnelles et descriptives.

▲: en anglais, *data analysis* ⇔ Statistique

## En résumé

La statistique s'intéresse à des jeux de données de taille raisonnable, fréquents dans vos domaines.

# Les statistiques sont (utilisées) partout

 Reporterre, le média de l'écologie

## Entre riches et pauvres, une différence de 5 tonnes de CO<sub>2</sub> chaque année

Les statistiques le prouvent : les riches polluent plus que les populations davantage modestes. Le 28 septembre, le Centre...

2 oct. 2023



 Radio France

## Groupe sanguin et coronavirus, un hasard génétique

D'après une étude chinoise, les personnes de groupe sanguin O sont mieux immunisées contre le coronavirus que les autres groupes.

21 mars 2020



 Le Figaro

## L'inquiétante hausse des homicides et tentatives d'homicide en France

Selon les dernières statistiques du ministère de l'Intérieur, les tentatives ont augmenté de 59% entre 2016 et 2022.

25 janv. 2024



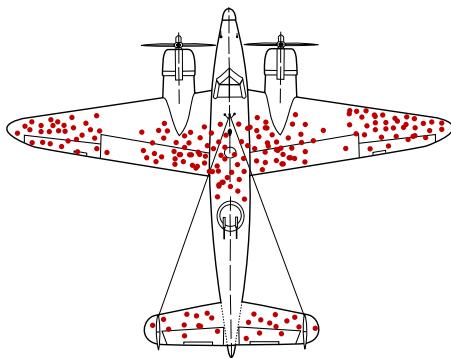
- En cela elles sont un outil *politique*, traduisant une certaine vision du monde et à manipuler avec précaution.

# Cœur de la statistique

- Avoir les *bonnes* données:

- observations d'un phénomène
- sont-elles représentatives ?

Ex. où renforcer l'avion ?



The image shows a screenshot of a video player interface. At the top, there is a thumbnail for a video titled "COMMENT REUSSIR SA VIE ? LE FONDATEUR MILLIARDAIRES DAMAZON 10 CONSEILS DE JEFF BEZOS". Below it is another thumbnail for a video titled "DEVENIR RICHE EST SUPER SIMPLE, IL SUFFIT DE...". To the right of the thumbnails, there is descriptive text and a small video preview window.

JEFF BEZOS : 10 conseils pour réussir de l'homme le plus riche du monde (Motivation français)  
245 k vues • il y a 5 ans  
Miss Imperfekte

Si vous luttez avec votre amour de soi, souhaitez continuer à battre votre confiance en vous et vos rêves, que vous soyiez ...

14 chapitres: Intro Jeff Bezos | Intro Miss Imperfekte | N'ayez aucun regret | Avancez petit à petit | Soyez un expert...

"Je suis devenu Riche quand j'ai compris ça" - Révélations de Jeff Bezos en Français  
38 k vues • il y a 1 an  
Dénichir

Dans cette vidéo, Jeff Bezos partage certains de ses conseils ...

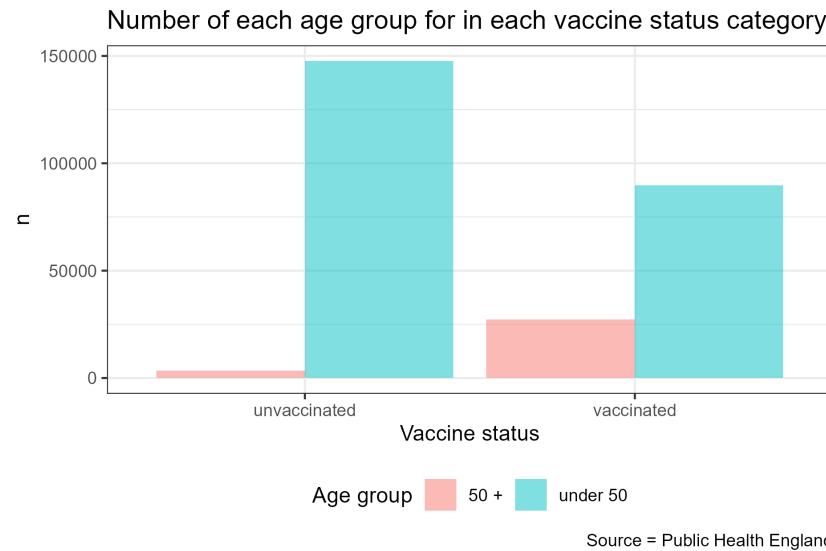
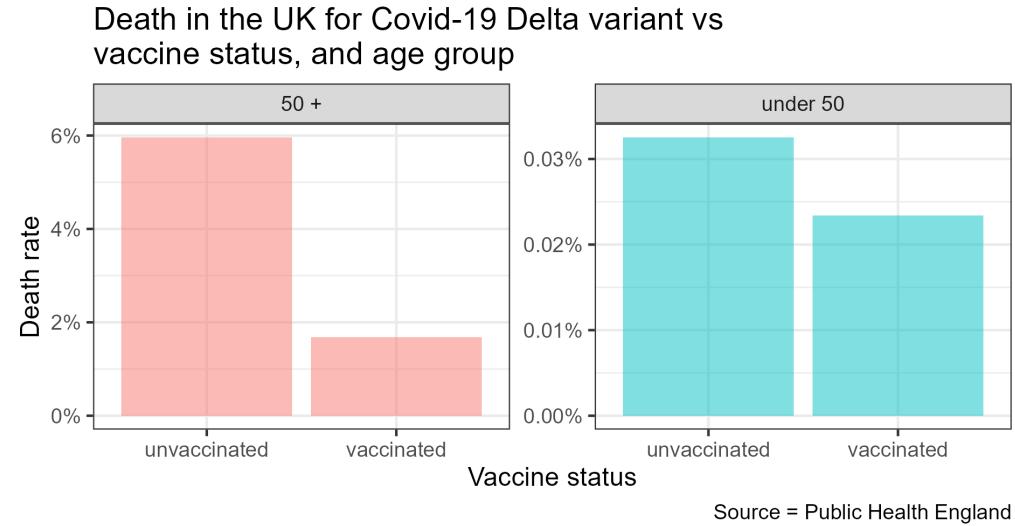
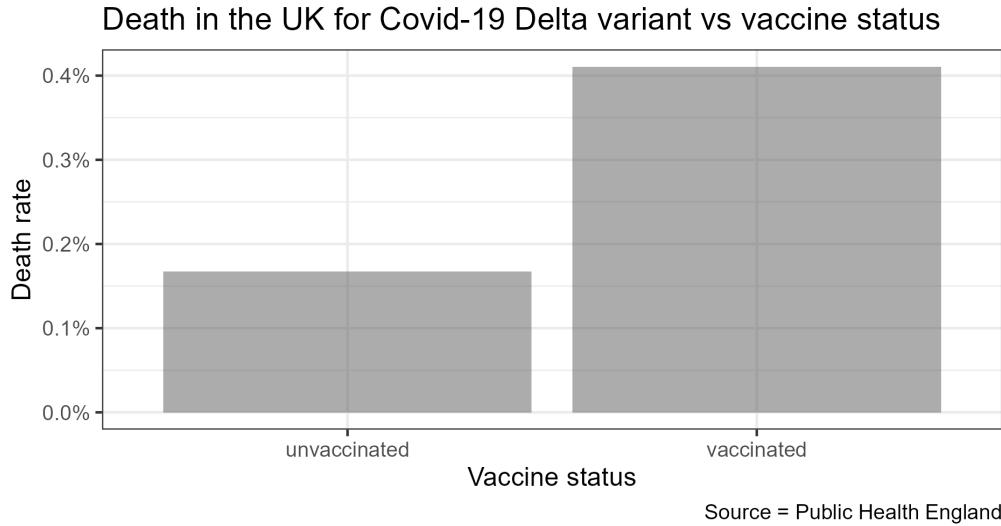
- Les résumer, les visualiser

- Se poser des questions:

- Ex. pollution aux algues vertes: quelle variable utiliser ? Surface ? Masse ? Où/Quand/Comment faire des recueils ?
- Cause / Conséquence (variété de blé / Rendement ; utilisation de phyto / diversité entomologique ; complexité du paysage)
- Variables **explicatives** vs variable(s) **réponse(s)**

- Tous les effets ont-ils été pris en compte ?

Ex. **paradoxe de Simpson** ([Vidéo complète sur le sujet](#) de Science Etonnante)



# Outil: R (**moteur**) / Rstudio (**carrosserie**)

- Gratuit/libre/multi-plateforme
- Nombreux (+++) domaines d'applications (ex. ce cours a été réalisé sous R)
- BEAUCOUP d'aide dispo

## Installation:



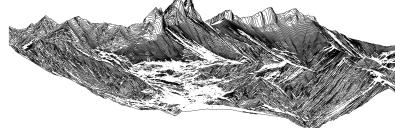
1. Télécharger et installer le **moteur**  
project.org/

→ <https://cran.r-project.org/>

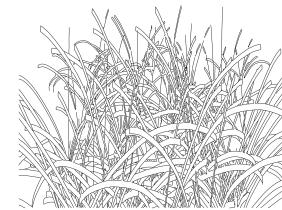


2. Télécharger et installer la **carrosserie**

On peut aussi dessiner avec R (mais c'est une autre histoire)



(a) Ridge

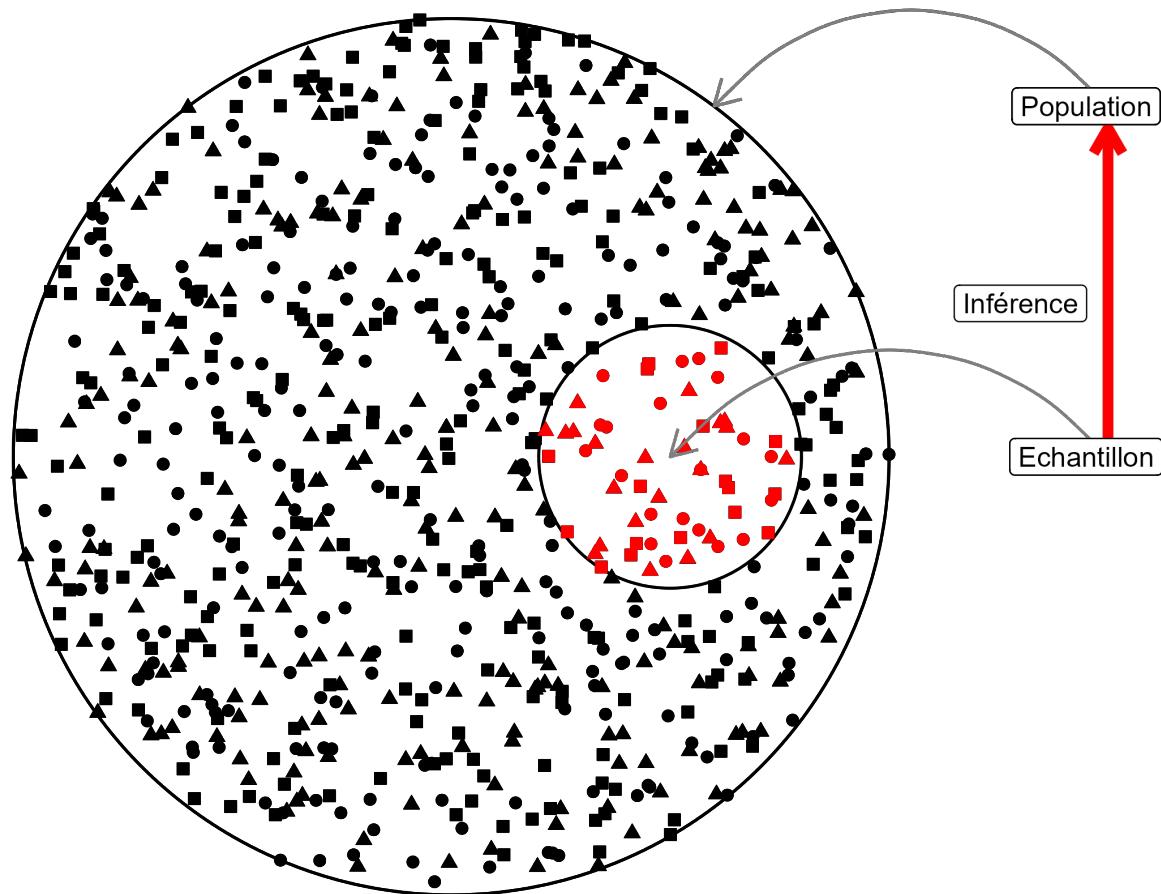


(b) Collatz

Figure 1: Pierre Casadebaig

## Quelques définitions (1/2)

- **Population:** ensemble d'entités objet de l'investigation statistique
- **Individu:** élément de la population d'étude
- **Echantillon:** ensemble des individus pour lesquels des valeurs ont été observées pour les variables de l'étude
- **Variable:** descripteur ou caractère des individus de la population d'étude
- **Inférence:** décider pour une population à partir des données observées de l'échantillon



## Quelques définitions (2/2)

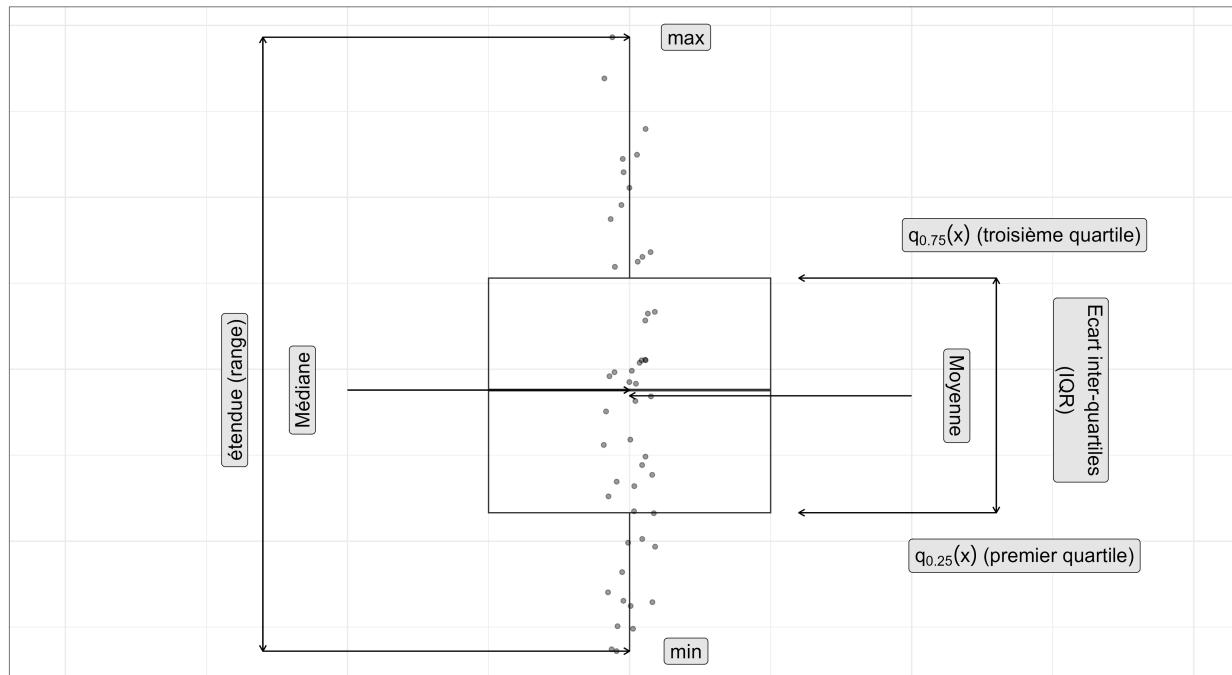
### Nature des variables:

- Qualitatives : les valeurs prises sont des *modalités*
  - nominale : pas de structure d'ordre 
  - ordinale : modalités ordonnées 
- quantitative: les valeurs prises sont *numériques*
  - discrete   
  - continue  

# Décrire les données à l'aide d'indicateurs

Soit  $x_1, x_2, \dots, x_n$  une série de  $n$  valeurs d'une variable X. On peut les décrire en utilisant des indicateurs de **position** et de **dispersion** (que vous connaissez probablement)

- moyenne:  $\bar{x} = \frac{1}{n} \sum x_i = \frac{x_1+x_2+\dots+x_n}{n}$
- médiane:  $q_{0.5}(x) =$  valeur telle que 50% des  $x_i$  ont une valeur inférieure et 50% une valeur supérieure
- 1er quartile:  $q_{0.25}(x) =$  valeur telle que 25% des  $x_i$  ont une valeur inférieure et 75% une valeur supérieure
- 3ème quartile:  $q_{0.75}(x) =$  valeur telle que 75% des  $x_i$  ont une valeur inférieure et 25% une valeur supérieure
- quantile  $\alpha$ :  $q_\alpha(x) =$  valeur telle que  $100 \times \alpha$  % des  $x_i$  ont une valeur inférieure et  $100 - 100 \times \alpha$  % une valeur supérieure
- variance:  $s^2(x) = \hat{\sigma}^2(x) = \frac{1}{n-1} \sum (x_i - \bar{x})^2$
- écart-type:  $s(x) = \sqrt{s^2(x)}$



# Un jeu de données comme fil conducteur

L'association Air Breizh surveille la qualité de l'air et mesure la concentration de polluants comme l'ozone (O<sub>3</sub>) ainsi que les conditions météorologiques comme la température, la nébulosité, le vent, etc. Durant l'été 2001, 112 données ont été relevées à Rennes.

```
1 ozone <- read.table("https://r-stat-sc-donnees.github.io/ozone.txt", header=TRUE, stringsAsFactors = TRUE)
2
3
4 head(ozone)

maxO3    T9    T12    T15 Ne9 Ne12 Ne15      Vx9      Vx12      Vx15 maxO3v
20010601  87 15.6 18.5 18.4   4   4   8  0.6946 -1.7101 -0.6946   84
20010602  82 17.0 18.4 17.7   5   5   7 -4.3301 -4.0000 -3.0000   87
20010603  92 15.3 17.6 19.5   2   5   4  2.9544  1.8794  0.5209   82
20010604 114 16.2 19.7 22.5   1   1   0  0.9848  0.3473 -0.1736   92
20010605  94 17.4 20.5 20.4   8   8   7 -0.5000 -2.9544 -4.3301  114
20010606  80 17.7 19.8 18.3   6   6   7 -5.6382 -5.0000 -6.0000   94

vent pluie
20010601 Nord Sec
20010602 Nord Sec
20010603 Est Sec
20010604 Nord Sec
20010605 Ouest Sec
20010606 Ouest Pluie
```

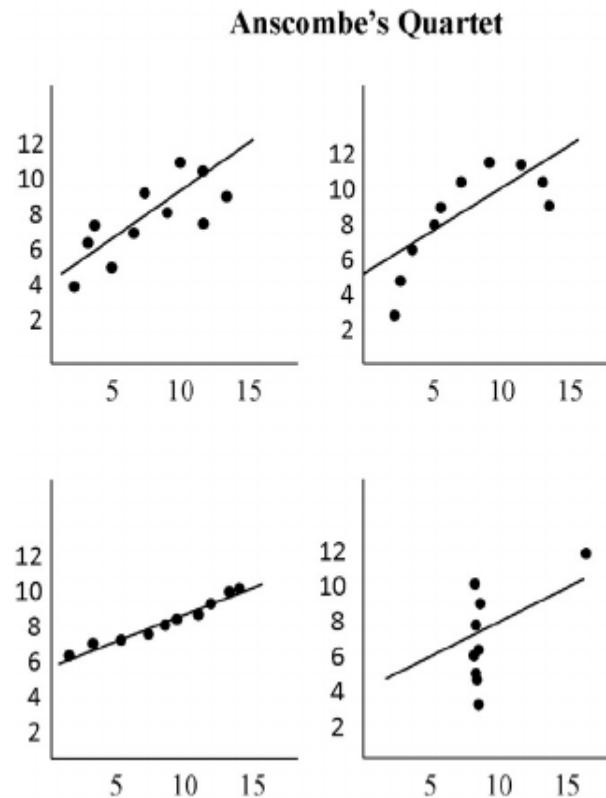
## Une question:

- Peut-on prévoir la concentration en ozone du lendemain pour avertir la population en cas de pic de pollution ?

# Visualisation

# La visualisation: pourquoi ?

- Quels points communs entre ces 4 jeux de données ?



| <u>Property</u>      | <u>Value</u>                             |
|----------------------|--|
| Mean of X (average)  | 9 in all 4 XY plots                      |
| Sample variance of X | 11 in all four XY plots                  |
| Mean of Y            | 7.50 in all 4 XY plots                   |
| Sample variance of Y | 4.122 or 4.127 in all 4 XY plots         |
| Correlation (r)      | 0.816 in all 4 XY plots                  |
| Linear regression    | $y = 3.00 + (0.500 x)$ in all 4 XY plots |

Data sets for the 4 XY plots

| I    | II    | III  | IV    |
|------|-------|------|-------|
| x    | y     | x    | y     |
| 10.0 | 8.04  | 10.0 | 9.14  |
| 8.0  | 6.95  | 8.0  | 8.14  |
| 13.0 | 7.58  | 13.0 | 8.74  |
| 9.0  | 8.81  | 9.0  | 8.77  |
| 11.0 | 8.33  | 11.0 | 9.26  |
| 14.0 | 9.96  | 14.0 | 8.10  |
| 6.0  | 7.24  | 6.0  | 6.13  |
| 4.0  | 4.26  | 4.0  | 3.10  |
| 12.0 | 10.84 | 12.0 | 7.26  |
| 7.0  | 4.82  | 7.0  | 7.26  |
| 5.0  | 5.68  | 5.0  | 4.74  |
|      |       | x    | y     |
|      |       | 8.0  | 6.58  |
|      |       | 8.0  | 5.76  |
|      |       | 8.0  | 5.76  |
|      |       | 8.0  | 8.84  |
|      |       | 8.0  | 8.47  |
|      |       | 8.0  | 7.04  |
|      |       | 8.0  | 5.25  |
|      |       | 19.0 | 12.50 |
|      |       | 8.0  | 5.56  |
|      |       | 8.0  | 7.91  |
|      |       | 8.0  | 6.89  |

# Visualisation: quels choix ?

- La visualisation dépend de:
  - La nature des données (nature des variables, nb d'individus)
  - La problématique: qu'est-ce que je veux montrer ?
  - [le site data-to-viz.com](http://data-to-viz.com) permet de trouver de bonnes idées
- La visualisation permet de:
  - comprendre / explorer / vérifier les données
  - suggérer des analyses
  - *faire passer des idées*

**Les graphiques sont ce qu'on retient de nombreux rapports / analyses / articles scientifiques.**

Sur  , le package [ggplot2](#) permet de faire des visualisations allant du plus simple au plus compliqué.

# GGplot - Quick tuto

## The Basics of ggplot2

There are three main components that get *added* together to build up a plot.

1

`ggplot()`

Required: "Hey, R, I'm building a plot, and I'm using the data in the `data=` argument\* to do it!"

Optional: "Here's the 'default' mappings (`mapping=aes(...)`) of specific data in that dataset and how I want to plot them – as the 'x' values or 'y' values or something else."

+

2

`geom_xxx()`

Required: "Give me a 'layer' of that chart as a column chart or a scatterplot, or a line chart, (or something else.)"

Optional: "If I didn't give it to you above, or if I need to override some part of it, here's the specific mappings (`mapping=aes(...)`) to use for this layer, the specific data (`data=`), and some guidance on what colors to use for what."

+

3

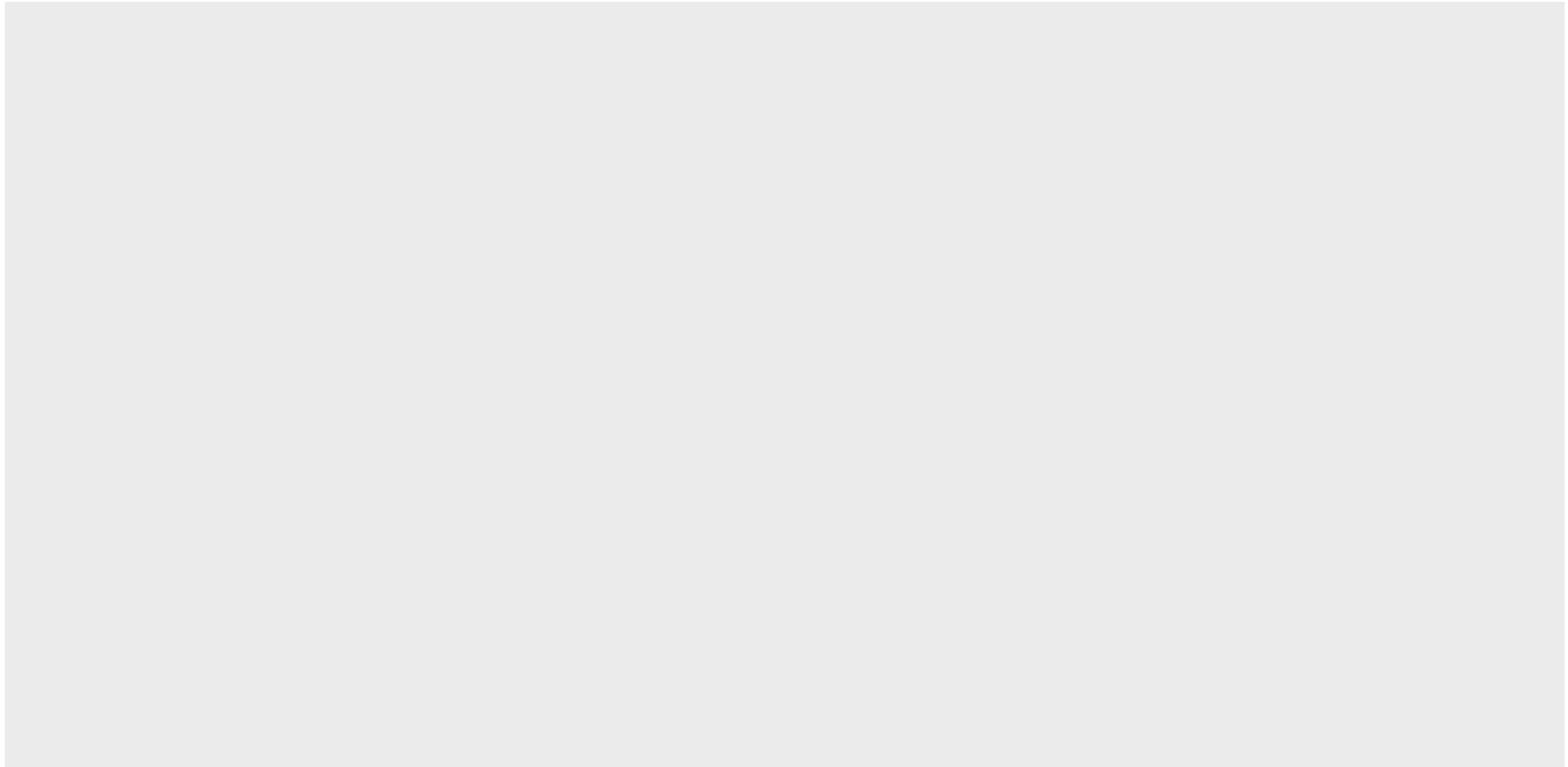
`theme()`

Optional: "Here's how I actually want to *format* the visualization: which tick marks and gridlines to show and what color to make them, which axis labels to show and various font properties, etc."

\* Not always the case, and not strictly required, but ggplot2 works best with data in a long format (or, at least, a tidy format). `gather()` from the `tidyverse` package is your friend here.

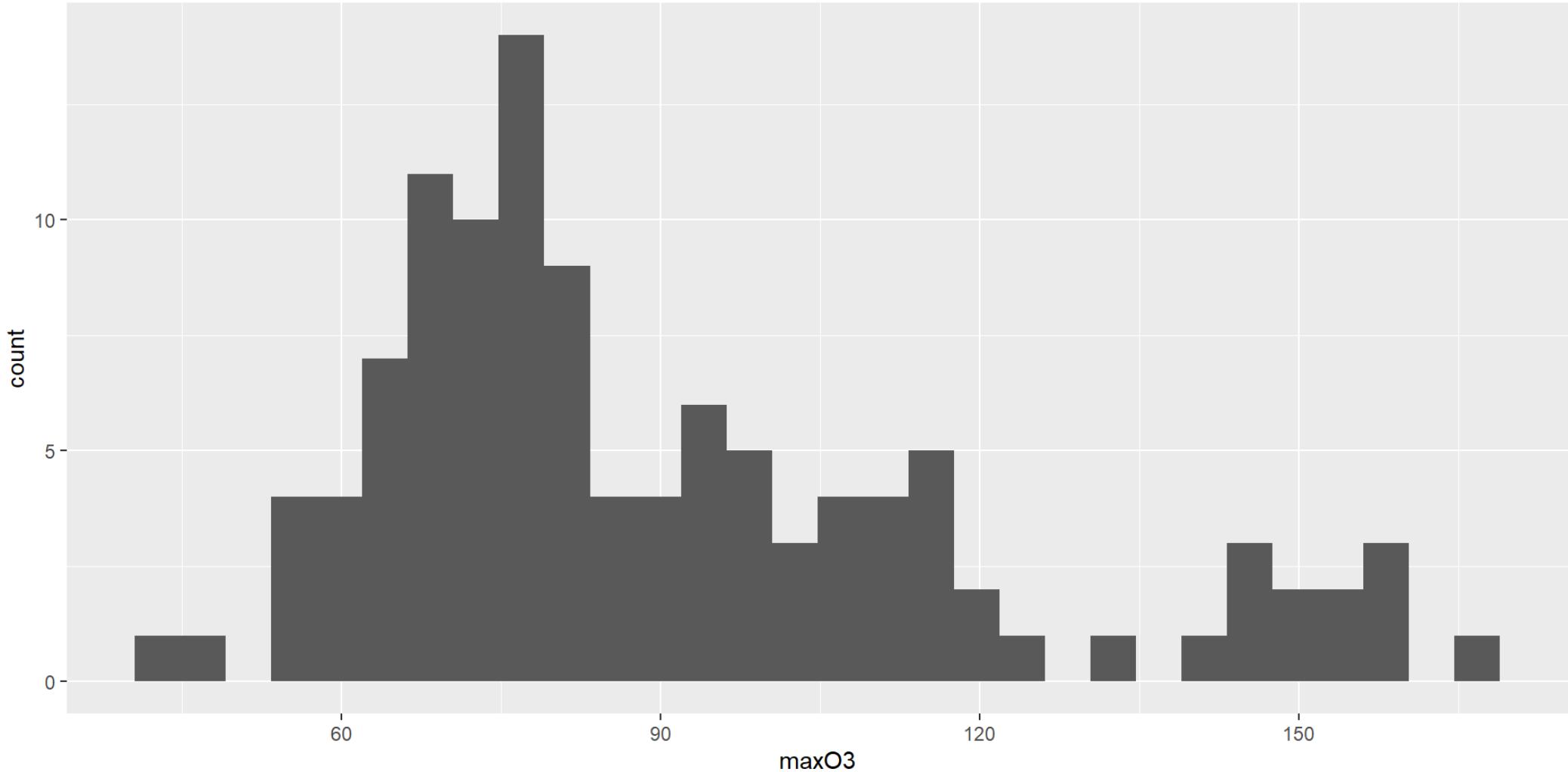
## GGplot - Exemple: distribution d'une variable quantitative

```
1 library(ggplot2)
2
3 ggplot(ozone) # création de l'objet ggplot
```



# GGplot - Exemple: distribution d'une variable quantitative

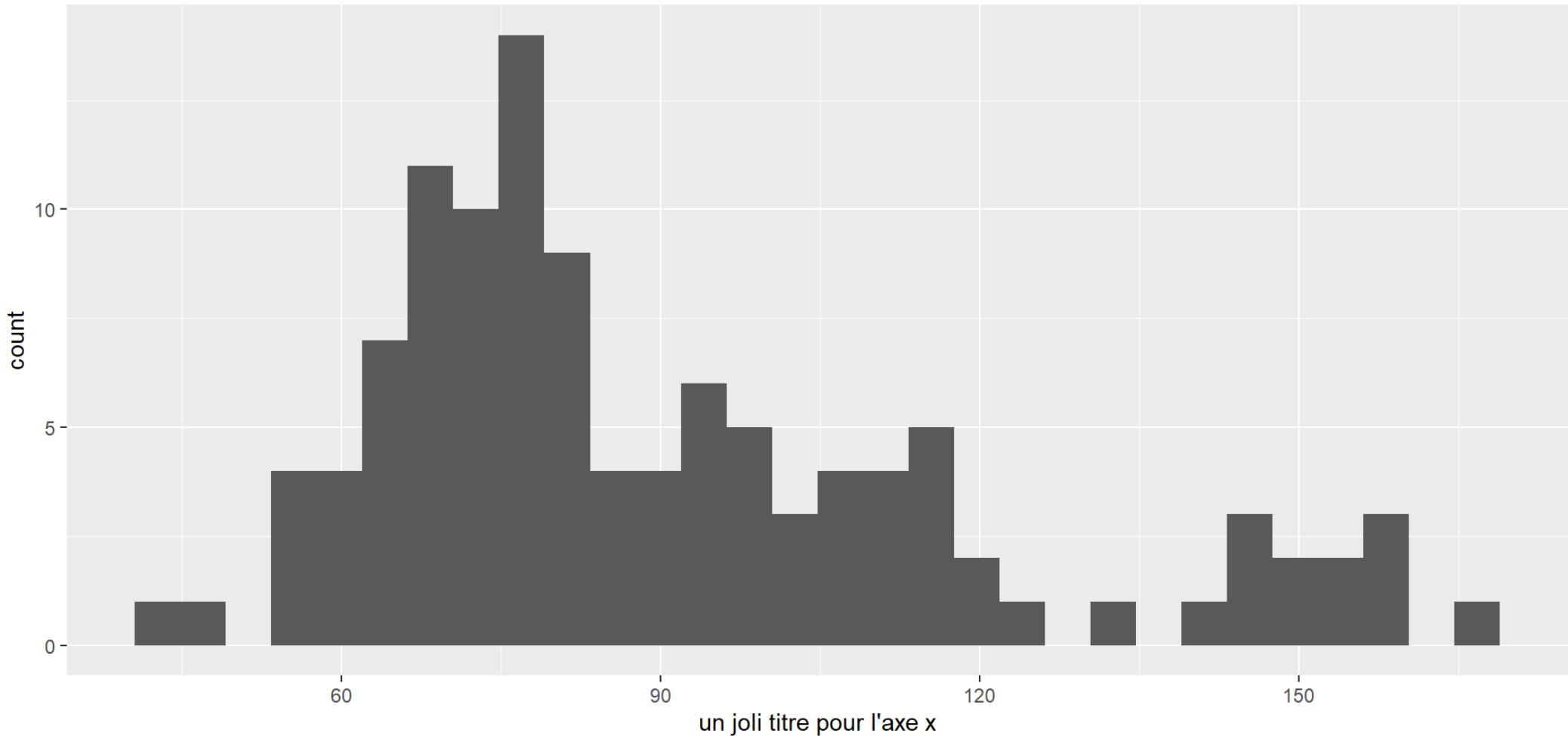
```
1 ggplot(ozone, aes(x = maxO3)) + # création de l'objet ggplot, en spécifiant ce qu'on veut en x  
2   geom_histogram() # ajout d'une couche HISTOGRAMME
```



# GGplot - Exemple: distribution d'une variable quantitative

```
1 ggplot(ozone, aes(x = maxO3)) + # création de l'objet ggplot, en spécifiant ce qu'on veut en x
2   geom_histogram() + # ajout d'une couche HISTOGRAMME +
3   labs(title = "un joli titre", x= "un joli titre pour l'axe x")
```

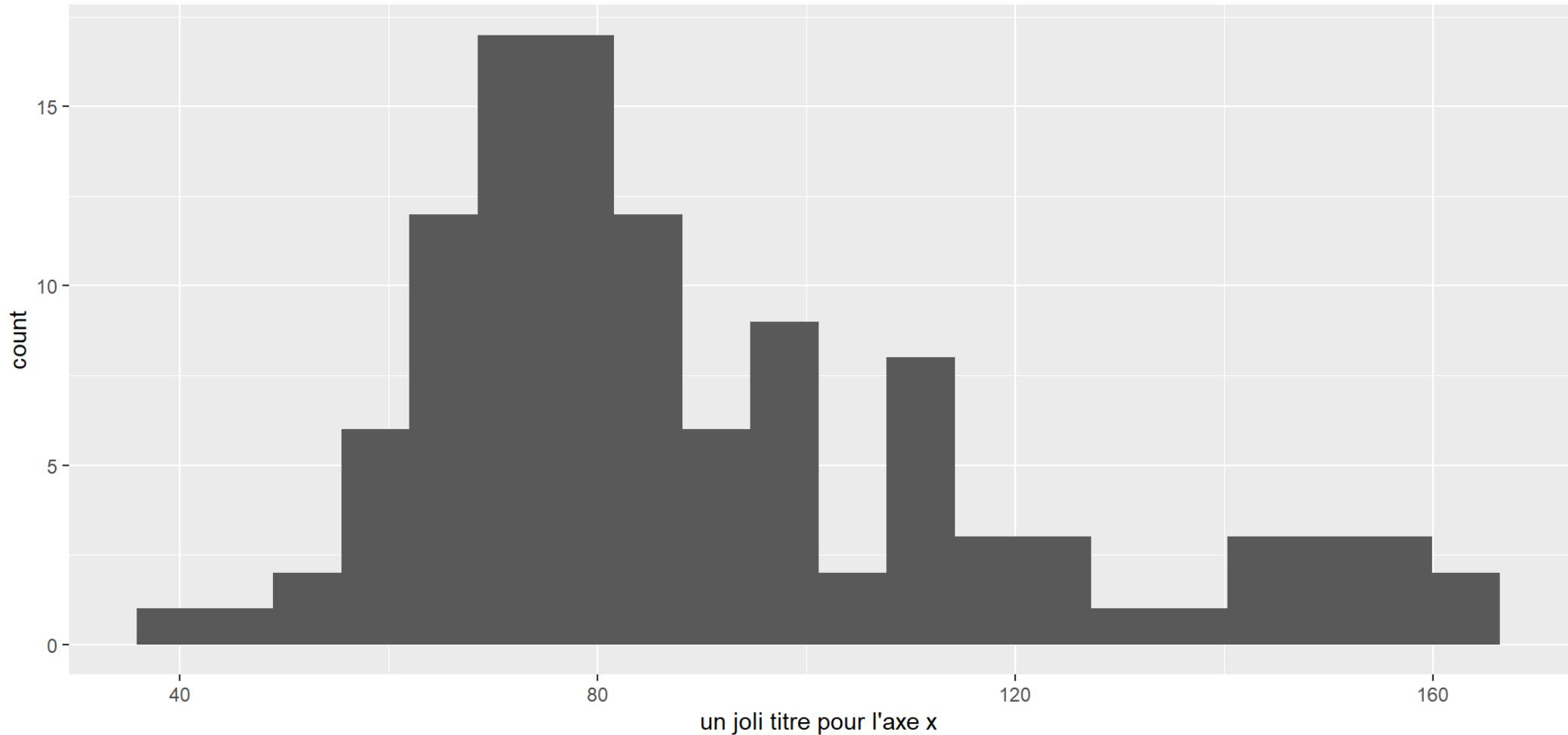
un joli titre



# GGplot - Exemple: distribution d'une variable quantitative

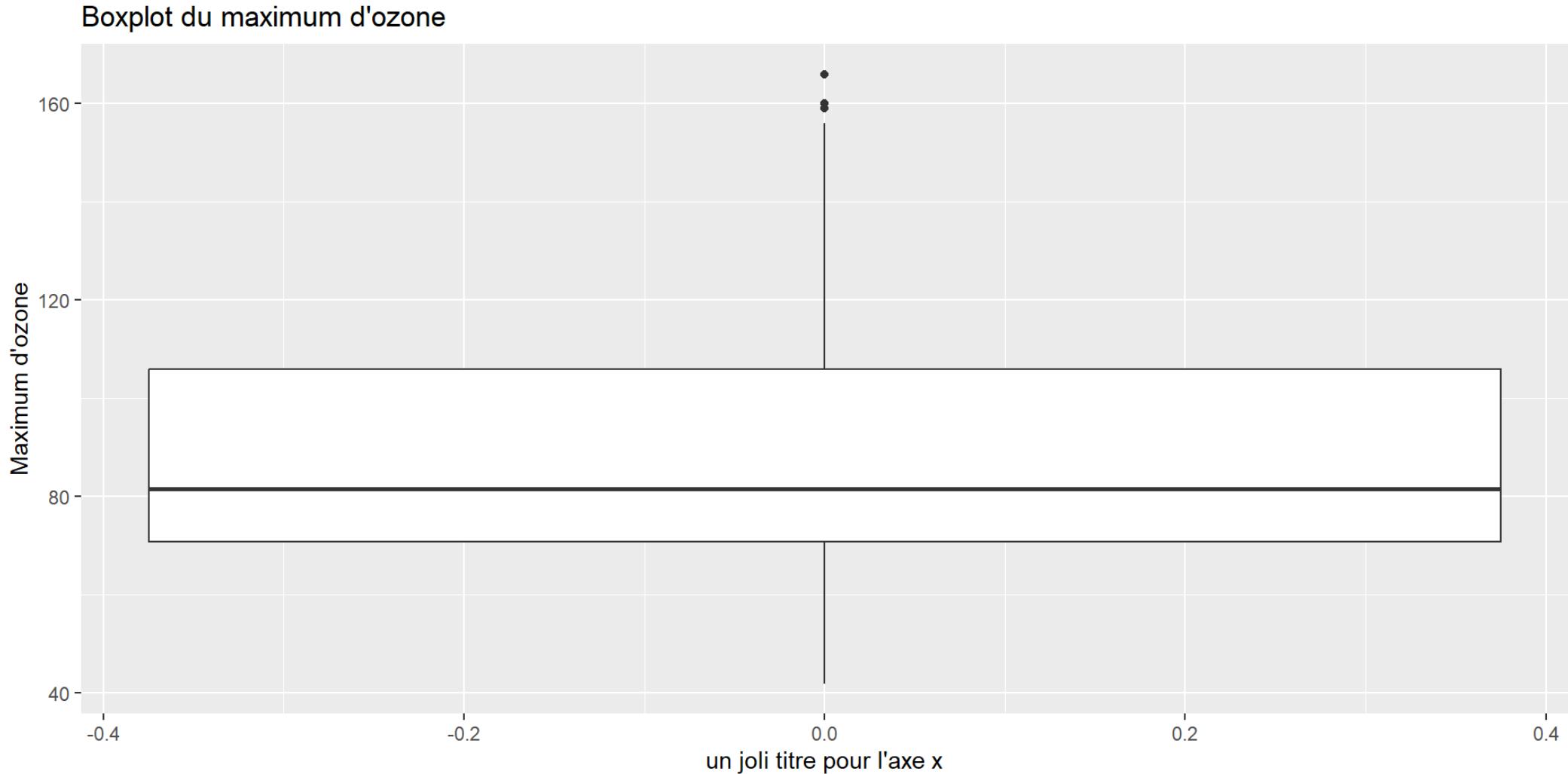
```
1 ggplot(ozone, aes(x = maxO3)) + # création de l'objet ggplot, en spécifiant ce qu'on veut en x  
2   geom_histogram(bins =20) + # ajout d'une couche HISTOGRAMME, classes plus grossières  
3   labs(title = "un joli titre", x= "un joli titre pour l'axe x")
```

un joli titre



# GGplot - Exemple: distribution d'une variable quantitative

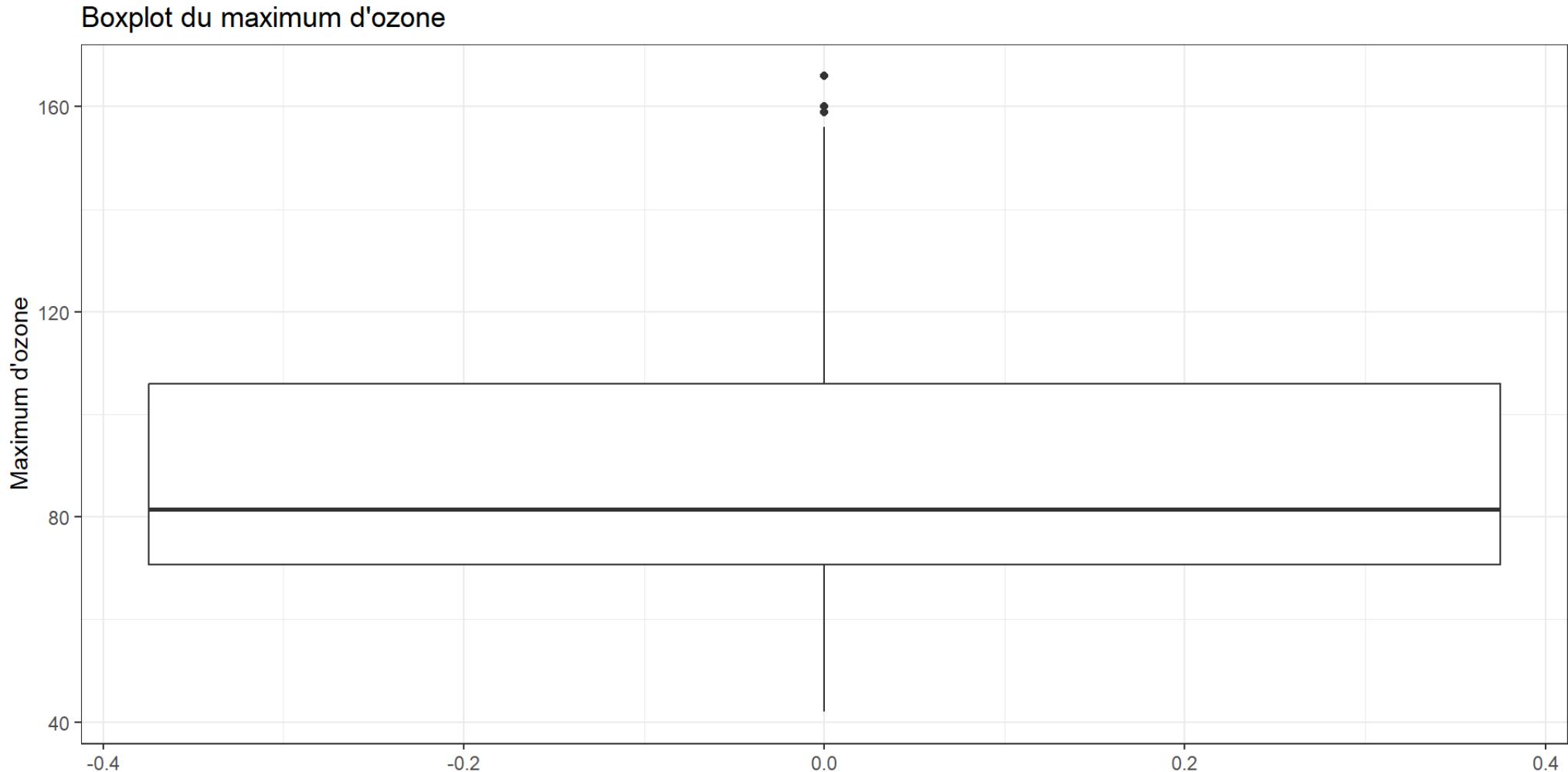
```
1 ggplot(ozone, aes(y = maxO3)) + # création de l'objet ggplot, en spécifiant ce qu'on veut en y
2   geom_boxplot() + # ajout d'une couche boxplot
3   labs(title = "Boxplot du maximum d'ozone", x= "un joli titre pour l'axe x", y = "Maximum d'ozone")
```



# GGplot - Exemple: distribution d'une variable quantitative

Moche ? Pensez à [theme\\_bw\(\)](#) (black & white)

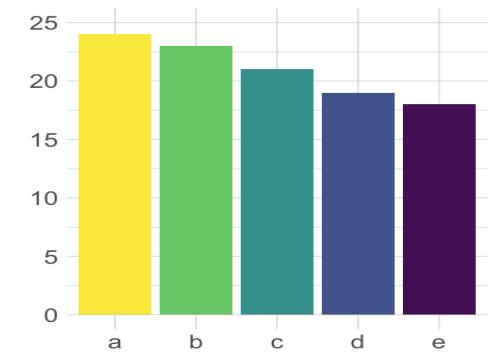
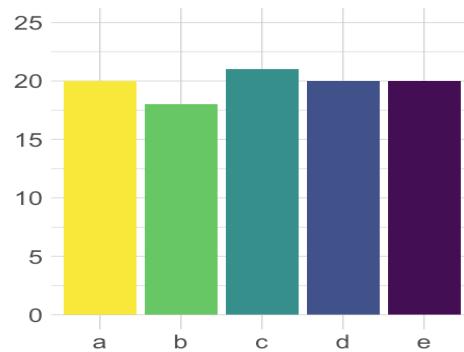
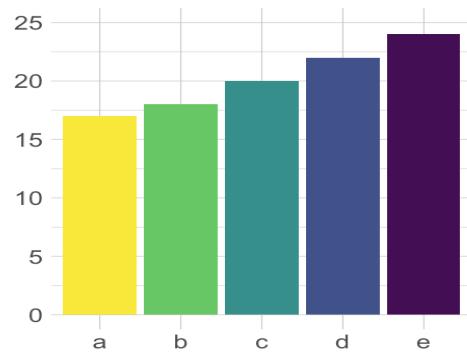
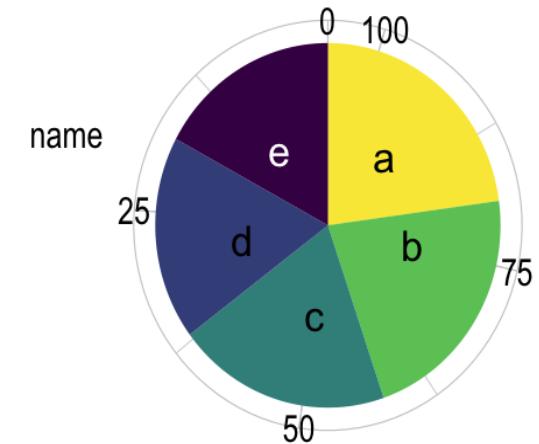
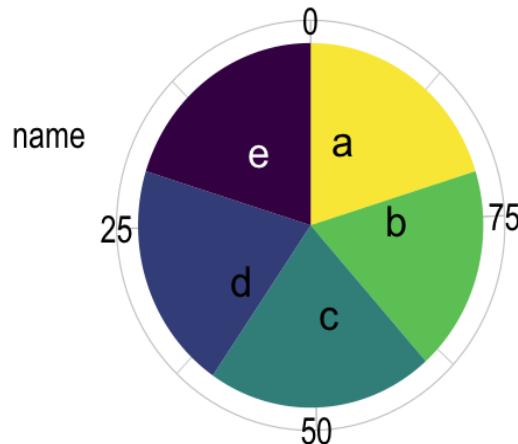
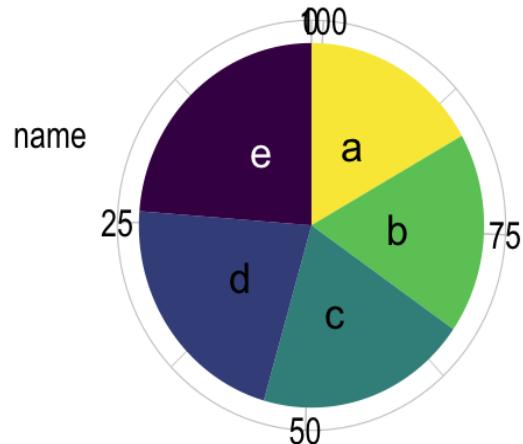
```
1 ggplot(ozone, aes(y = maxO3)) + # création de l'objet ggplot, en spécifiant ce qu'on veut en y
2   geom_boxplot() + # ajout d'une couche boxplot, classes plus grossières
3   labs(title = "Boxplot du maximum d'ozone", y = "Maximum d'ozone") +
4   theme_bw()
```



# Distribution d'une variable qualitative (1/2)

Le camembert c'est bon, mais *seulement en fromage*. SINON CA PUE TROP.

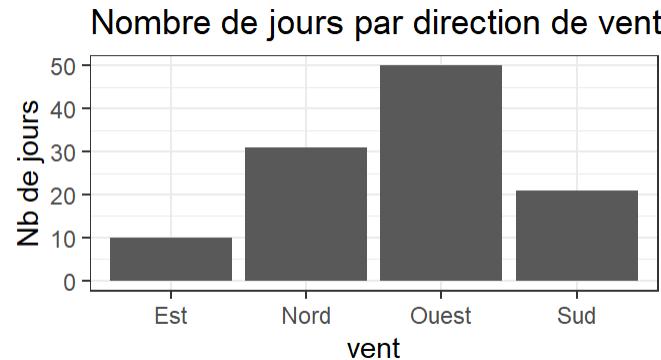
Quelle classe est la plus représentée ?



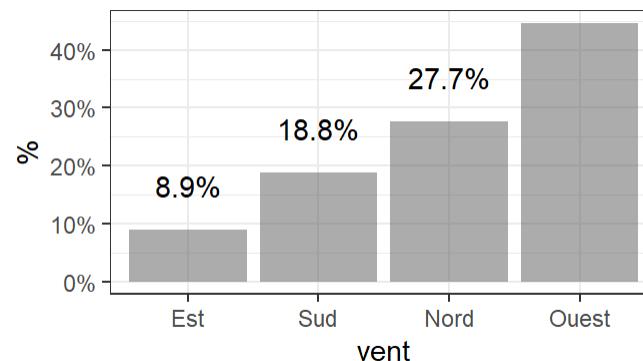
## Distribution d'une variable qualitative (2/2)

Alternative: graphiques en barre

► Code

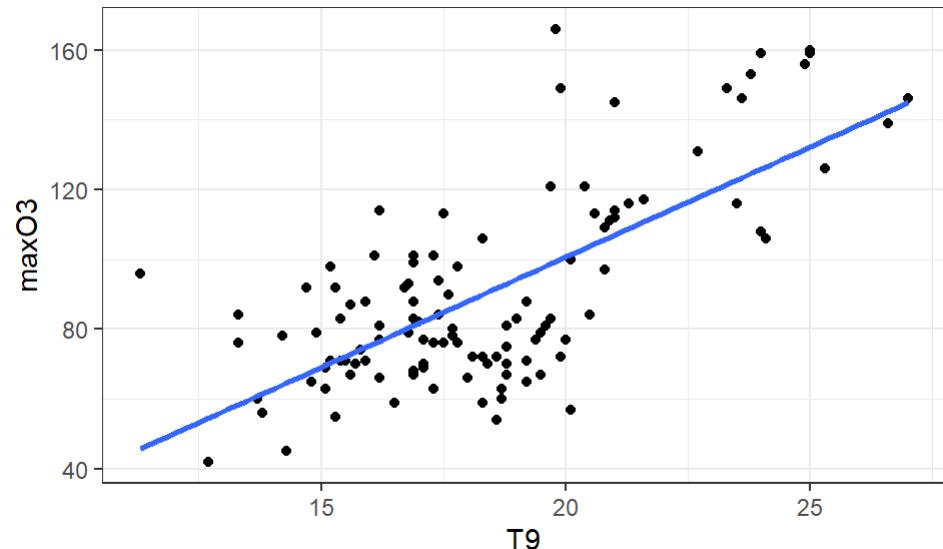


► Code

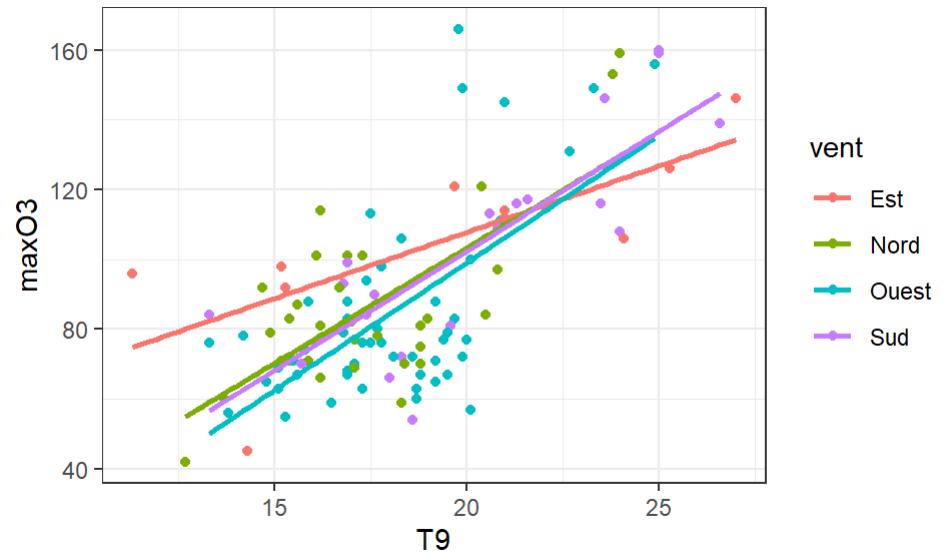


# Effet d'une / deux variable(s) sur une variable quanti

```
1 ggplot(ozone, aes(x = T9, y= maxO3)) +  
2   geom_point() + # ajout points  
3   geom_smooth(method = "lm", se = FALSE) + # Ajout ajustement linéaire  
4   theme_bw()
```

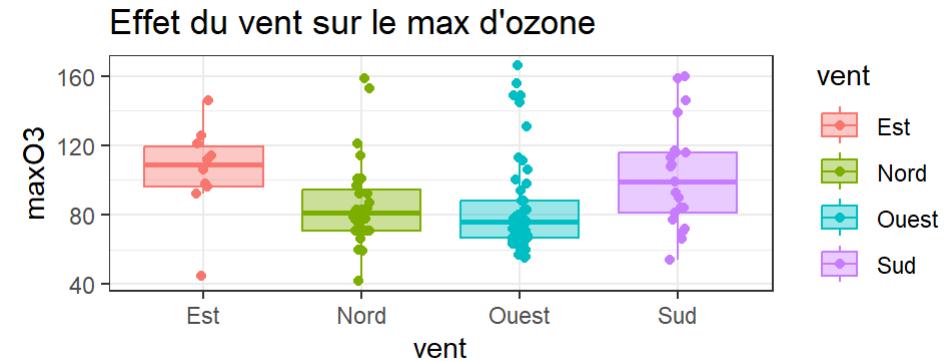


```
1 #!  
2 ggplot(ozone, aes(x = T9, y= maxO3, color = vent)) + # ajout d'une coul  
3   #par direction du vent  
4   geom_point() + # ajout points  
5   geom_smooth(method = "lm", se = FALSE) + # Ajout ajustement linéaire  
6   theme_bw()
```



# Effet d'une / deux variable(s) quali sur une variable quanti (1/2)

```
1 ggplot(ozone, aes(x = vent,
2                   y= maxO3,
3                   fill=vent,
4                   col=vent)) + # dans aes: ce qui DEPEND du jeu de donnees
5 geom_boxplot(outlier.shape=NA,alpha=0.4) + # alpha : transparence
6 geom_point(position = position_jitter(width = 0.05, height= 0)) + #
7 # points, avec perturbation horizontale pour faciliter visu
8 theme_bw() +
9 labs(title ="Effet du vent sur le max d'ozone")
```



## Effet de deux variable(s) quali sur une variable quanti (2/2)

Quid d'une interaction potentielle vent pluie ?

*Point code :* La fonction `%>%` de `dplyr` permet de passer des arguments en chaîne dans des fonctions. Par ex.:

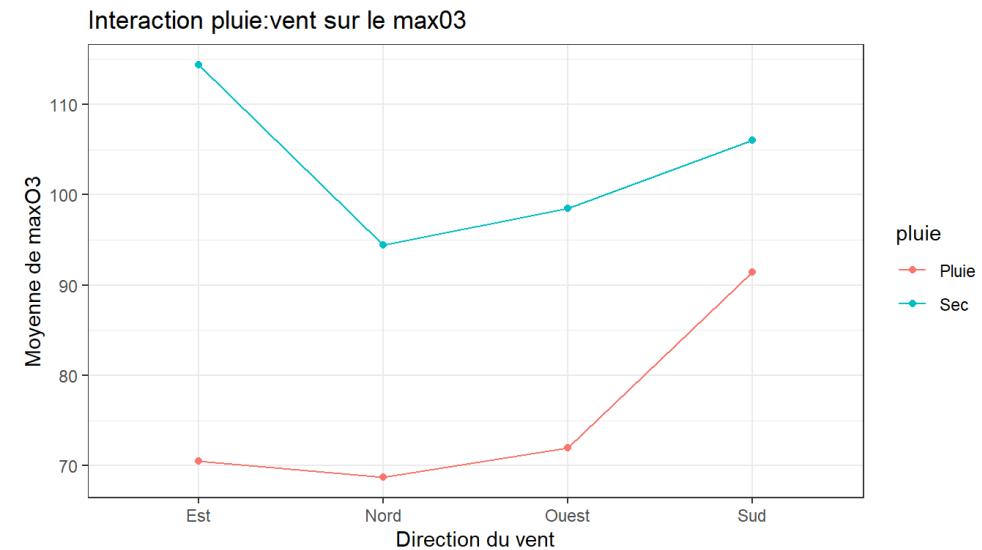
```
1 x <- 1:3  
2 max(x)
```

[1] 3

```
1 library(dplyr) # permet d'utiliser %>%  
2 x %>% max # équivaut à max(x)
```

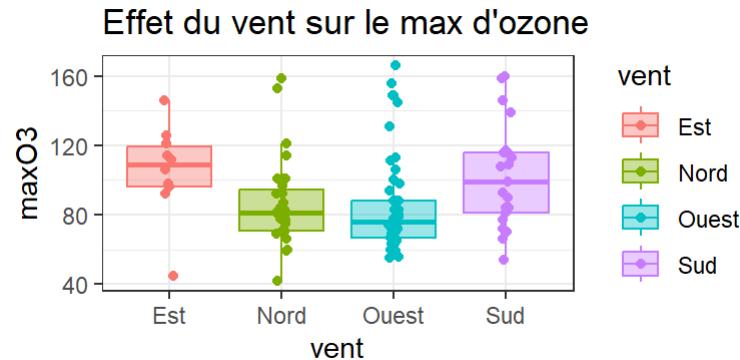
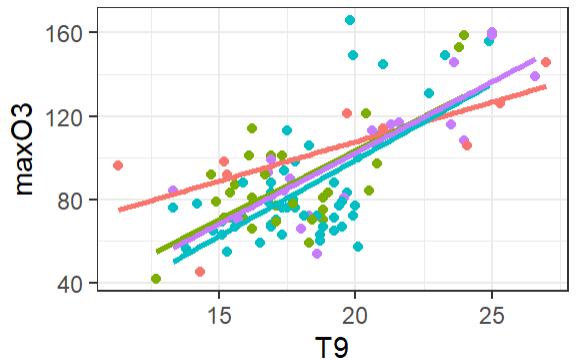
[1] 3

► Code

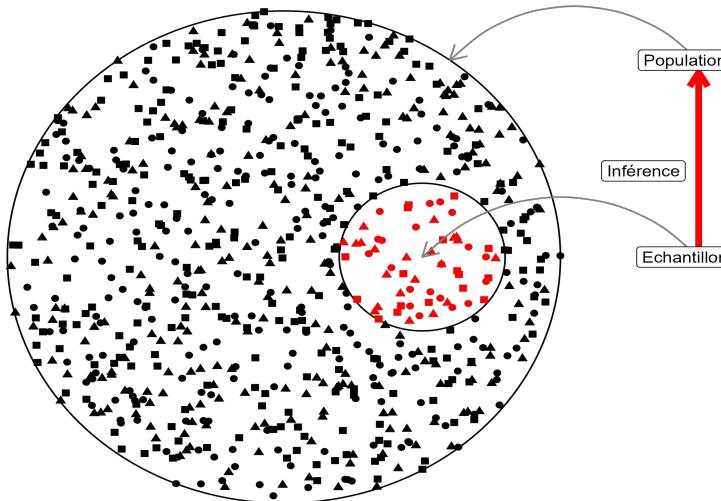


# Des visualisations vers les tests statistiques

Les visus précédentes nous ont permis d'intuiter certaines tendances.



Peut-on les généraliser ? Rappel:



C'est toute la question des *tests statistiques*.

# **Principes et applications des tests statistiques**

# De la question aux tests

- On dispose de multiples variables, toutes présentant potentiellement un intérêt

```
1 colnames(ozone)  
[1] "maxO3"   "T9"      "T12"     "T15"     "Ne9"      "Ne12"    "Ne15"    "Vx9"  
[9] "Vx12"    "Vx15"    "maxO3v"  "vent"    "pluie"
```

- Ici une variable particulière nous intéresse (c'est un *choix* ; guidé par la littérature scientifique / les politiques publiques etc. ): maxO3 (car étant clé pour la pollution de l'air). C'est la **variable réponse**

On peut se poser plusieurs questions:

- Effet du vent sur le max d'ozone ?
- Le temps sec/pluvieux influence-t-il le max d'ozone ? De la même manière selon les différentes directions du vent ?

**Objectif :** Généraliser (inférer) au-delà des données de l'échantillon

**Problème :** Comment gérer l'incertitude liée au fait qu'on a observé qu'une petite partie des données

# Rappel: Théorème central limite

$$\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n) \quad \text{et donc} \quad \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0, 1)$$

- En pratique,  $\sigma^2$  rarement connu, et donc doit être estimé à partir de l'échantillon
- $\Rightarrow$  augmente un peu l'incertitude  $\Rightarrow$  utilisation d'une loi de Student ( $n - 1$ ) degrés de liberté (plutôt qu'une loi normale):  
$$\frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim \mathcal{T}(n - 1)$$

D'où l'intervalle de confiance de  $\mu$  au niveau de confiance 95%:

$$\left[ \bar{x} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1); \bar{x} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1) \right]$$

```
1 t.test(ozone$maxO3)
```

```
One Sample t-test

data: ozone$maxO3
t = 33.905, df = 111, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
85.02578 95.58136
sample estimates:
mean of x
90.30357
```

```
1 x <- ozone$maxO3
2
3 mean(x) - qt(0.975, df = 111)*sd(x)/sqrt(length(x))
```

```
[1] 85.02578
```

```
1 mean(x) + qt(0.975, df = 111)*sd(x)/sqrt(length(x))
```

```
[1] 95.58136
```

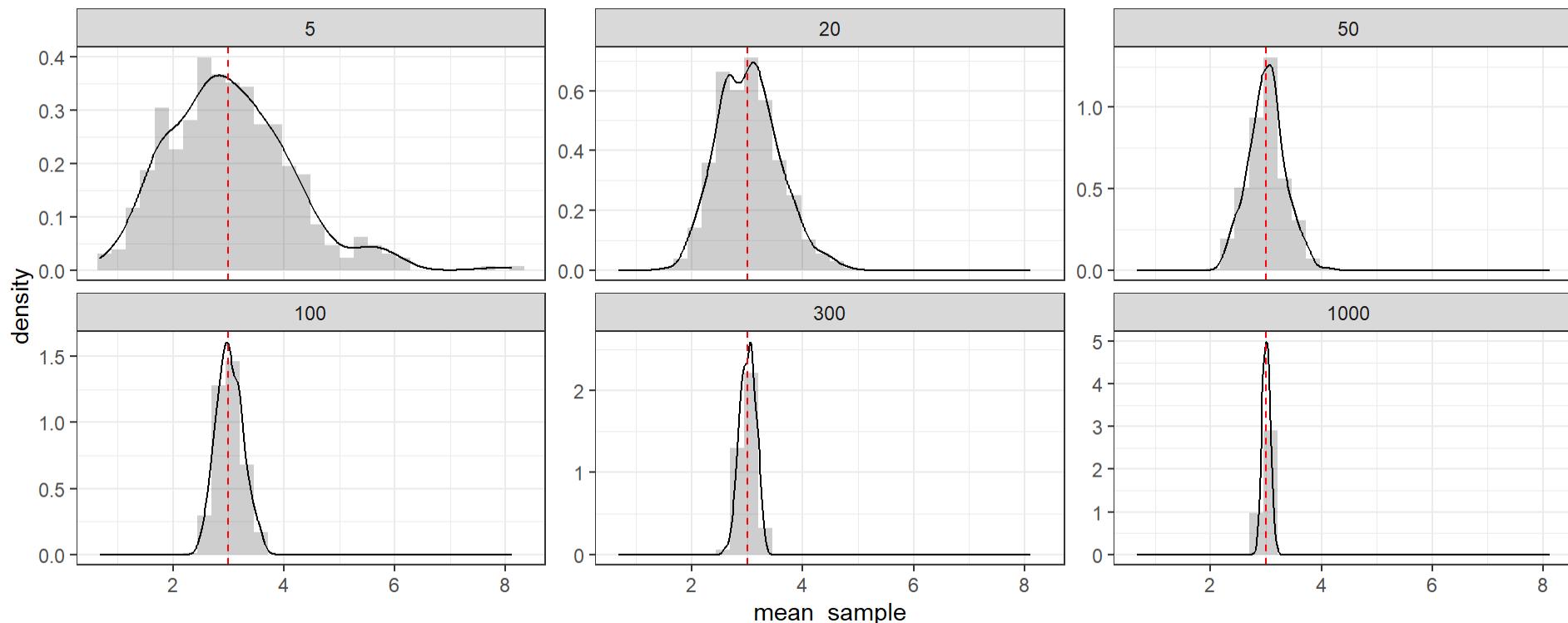
# Le TCL, c'est PUISSANT

- Ce que dit le TCL, c'est que peu importe la distribution de base, les moyennes empiriques d'échantillons issus cette distribution auront une distribution normale centrée sur l'espérance de la distribution !

## Une simulation rapide pour visualiser les choses

- Ex distribution  $\chi_2(3)$
- On connaît la vraie moyenne de cette loi: 3

Mais dans la *vraie vie*, on aurait accès qu'à un échantillon de cette loi, par ex. de taille  $n = 50$ . Regardons comment évolue la distribution des moyennes empiriques en fonction de la taille de l'échantillon.



# Intervalle de confiance vers test de conformité

- On suppose / connaît  $\frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim \mathcal{T}(n - 1)$ .
- On peut maintenant tester si la moyenne est égale à une valeur particulière, par ex est-ce que  $\mu = 100$  ?

A partir d'un échantillon  $x$ , on peut calculer  $\bar{x}$  et  $s^2$  et se demander si la valeur est typique d'une loi de Student à  $n-1$  ddl.

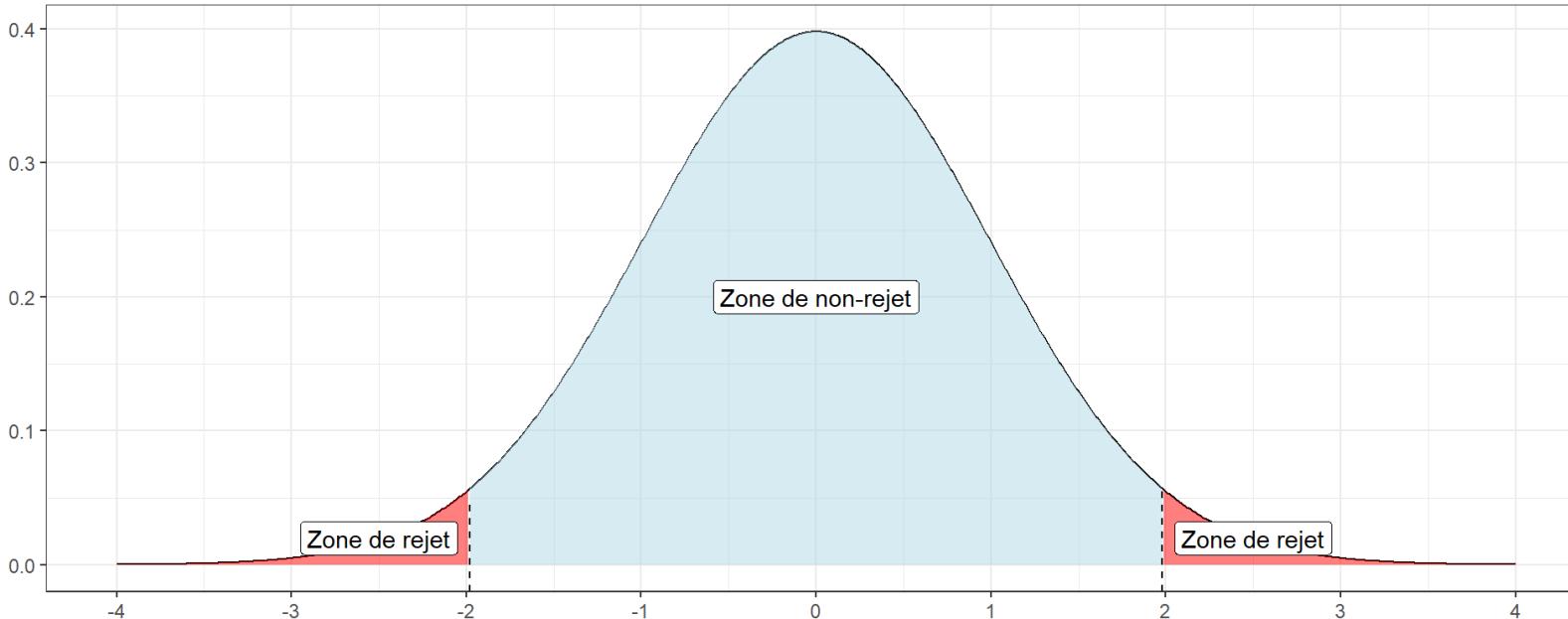
## Quelques définitions

- $H_0$  hypothèse nulle,  $H_0 : \mu = 100$  vs  $H_1$  hypothèse alternative,  $H_1 : \mu \neq 100$
- **Statistique de test:**  $\frac{\bar{X} - \mu}{\sqrt{s^2/n}}$ : il s'agit de la valeur qu'on peut calculer à partir de l'échantillon, et qu'on suppose suivre une certaine loi sous  $H_0$ : ici une loi de Student à  $n - 1$  ddl
- **p-value** du test: probabilité calculée sous  $H_0$ , que la statistique de test soit plus extrême que la valeur observée  $T_{obs}$

p-value: "Dans un monde où  $H_0$  est vraie, la probabilité d'obtenir une valeur au moins aussi extrême pour la statistique de test est de p"

Si  $p < 0.05$ , on rejette l'hypothèse  $H_0$  au seuil de 5%

Dans notre cas,  $n = 112$ , donc  $T$  est censée suivre une loi de Student à 111 ddl, sous  $H_0$ :



```
One Sample t-test
data: ozone$maxO3
t = -3.6406, df = 111, p-value = 0.0004148
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 85.02578 95.58136
sample estimates:
mean of x
 90.30357
[1] 0.0004147533
```

Pour une loi de Student à  $n-1$  ddl,  $-3.64$  est une valeur *peu probable* donc on rejette  $H_0$ , au seuil de 5%.

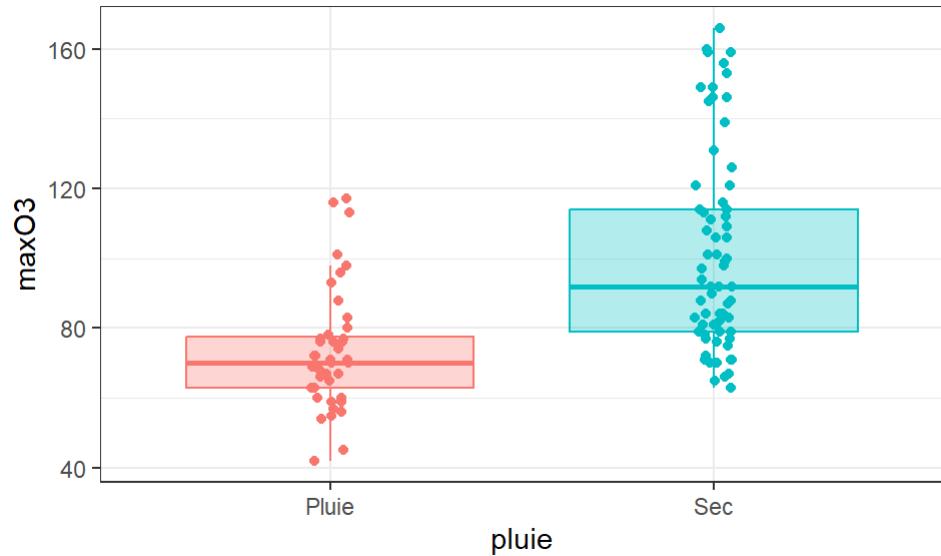
Ici, on a testé  $H_0$  vs  $H_1 : \mu \neq 100$  (alternative = "two.sided") mais on aurait pu tester  $H_1: \mu < 100$  (alternative = "less") ou  $H_1: \mu > 100$  (alternative = "greater")

# Comparaison de 2 moyennes (1/2)

Question: le temps (sec/pluvieux) a-t-il un effet sur le max d'ozone ?

Réflexe: visu !

► Code



Passer de l'échantillon à n'importe quel jour induit de l'incertitude. Mais, il est plus facile de conclure qu'il y a une différence entre les deux moyennes *dans la population* si:

1. Les moyennes sont très différentes ( vs )
2. La variabilité du maximum d'ozone est faible au sein des jours pluvieux et au sein des jours secs
3. Il y a beaucoup de données

## Comparaison de 2 moyennes (2/2)

On considère que les données de la sous-population 1 sont telles que  $(X_{i1})_{1 \leq i \leq n_1} \sim \mathcal{N}(\mu_1, \sigma^2)$  et que les données de la sous-population 2 sont telles que  $(X_{i2})_{1 \leq i \leq n_2} \sim \mathcal{N}(\mu_2, \sigma^2) \Rightarrow$  (pour l'instant) seules les moyennes peuvent être différentes

- $H_0$  hypothèse nulle,  $H_0 : \mu_1 = \mu_2$  vs **H1 hypothèse alternative**,  $H_1 : \mu_1 \neq \mu_2$

- **Statistique de test:**  $T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_2}}}$

- Sous  $H_0$ , T suit une loi de Student  $n_1 + n_2 - 2$

Si les variances sont inégales, le test est différent  $\implies$  tester l'égalité des variances avant de faire le test de comparaison de moyennes

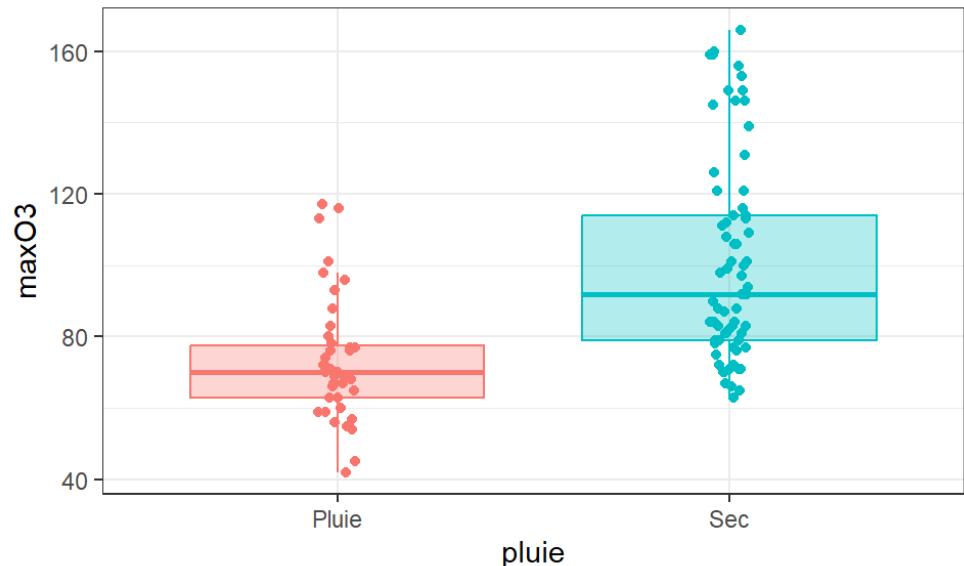
## Test de comparaison de 2 variances

- $H_0$  hypothèse nulle,  $H_0 : \sigma_1^2 = \sigma_2^2$  vs  $H_1$  hypothèse alternative,  $H_1 : \sigma_1^2 \neq \sigma_2^2$

$$\iff H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \text{ contre } H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

- Statistique de test:  $F = \frac{S_1^2}{S_2^2}$

- Sous  $H_0$ , F suit une loi de Fisher à  $n_1$  et  $n_2$  ddl



# Mise en pratique

```
1 var.test(maxO3 ~pluie, data = ozone, alternative = "two.sided")
```

```
F test to compare two variances

data: maxO3 by pluie
F = 0.35906, num df = 42, denom df = 68, p-value = 0.0005659
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
0.2108847 0.6338374
sample estimates:
ratio of variances
0.3590605
```

Ici on rejette  $H_0$  au seuil de 5%, les variances sont différentes

```
1 # NOTEZ LE VAR.EQUAL
2 t.test(maxO3 ~pluie, data = ozone, alternative = "two.sided", var.equal = FALSE)
```

```
Welch Two Sample t-test

data: maxO3 by pluie
t = -6.3362, df = 109.88, p-value = 5.321e-09
alternative hypothesis: true difference in means between group Pluie and group Sec is not equal to 0
95 percent confidence interval:
-36.02936 -18.86110
sample estimates:
mean in group Pluie mean in group Sec
73.39535 100.84058
```

On rejette  $H_0$ , on affirme que les moyennes sont significativement différentes

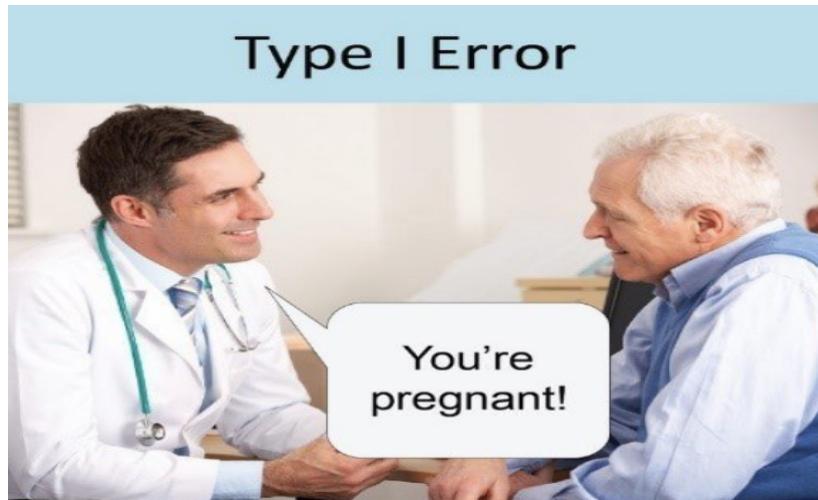
# Erreur et puissance d'un test

Décision d'un test prise pour la population ("il y a plus d'ozone les jours sec") alors qu'on n'observe qu'un échantillon

⇒ Décision incertaine, et deux erreurs possibles

- **Erreur de première espèce:** rejeter  $H_0$  alors que celle-ci est vraie: faux positif
- **Erreur de deuxième espèce :** ne pas rejeter  $H_0$  alors que celle-ci est fausse: faux négatif

En image (? qu'est-ce que  $H_0$  ici ?):



un test est dit puissant si son erreur de 2ème espèce est petite. La **puissance d'un test** est la probabilité de rejeter  $H_0$  et d'avoir raison

## Puissance d'un test (1/2)

**Question souvent posée** Combien de données faut-il pour tester une différence entre 2 moyennes ?

Par ex.: combien de patients pour tester s'il y a une diff entre 2 médicaments pour faire baisser le taux de cholestérol ?

Rappel: de quoi est-ce que ça dépend ?

- ?
- ?
- ?

### Reformuler la question

Combien faut-il de patients pour mettre en évidence une différence d'au moins 0.2g/l entre les 2 médicaments ?

Toujours pas possible 😞

MAAAAAAIS ☺

## Puissance d'un test (2/2)

MAAAAAIS 😊 si;

- on connaît la variance de la variable d'intérêt (ex. expériences antérieures montrent un écart-type de 0.4g/l) (**sd**)
- on veut détecter une différence donnée par ex toute diff > 0.2g/l (**delta**)
- on veut détecter cette différence avec proba de 80% (**power**)

alors c'est possible 😊

```
1 power.t.test(delta= 0.2, sd = 0.4, power = .80)
```

Two-sample t test power calculation

```
n = 63.76576
delta = 0.2
sd = 0.4
sig.level = 0.05
power = 0.8
alternative = two.sided
```

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

# Analyse de variance à un facteur

# Test de l'effet d'un facteur

Par ex. *la direction du vent a-t-elle un effet sur le maximum d'ozone ?*

**Variable réponse quantitative**, notée Y: maxO3

**Variable explicative qualitative à I modalités (groupes)**: direction du vent

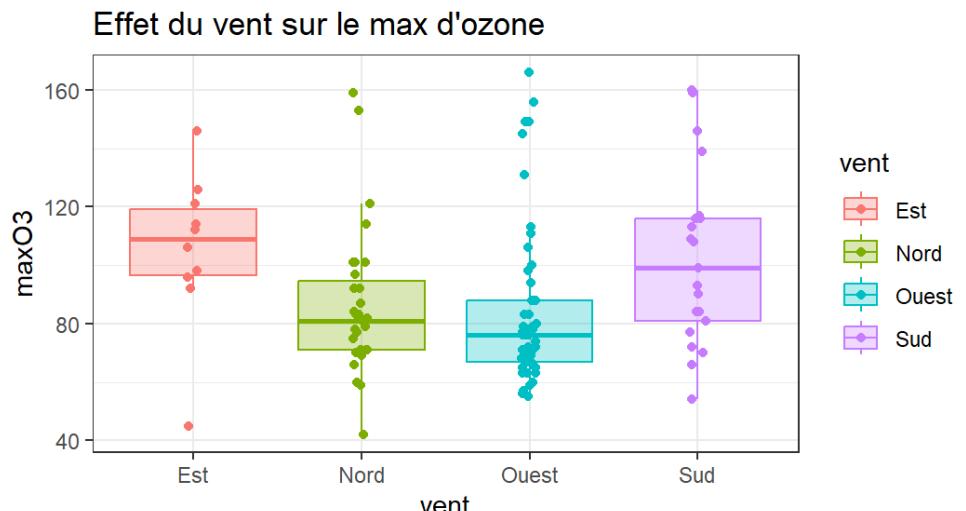
C'est le cadre de l'ANOVA (ANalysis Of VAriance)

| maxO3   | T9     | T12  | T15   | Ne9   | Ne12 | Ne15 | Vx9     | Vx12    |
|---------|--------|------|-------|-------|------|------|---------|---------|
| Vx15    | maxO3v | vent | pluie |       |      |      |         |         |
| 87      | 15.6   | 18.5 | 18.4  | 4     | 4    | 4    | 0.6946  | -1.7101 |
| -0.6946 |        | 84   | Nord  | Sec   |      |      |         |         |
| 82      | 17.0   | 18.4 | 17.7  | 5     | 5    | 5    | -4.3301 | -4.0000 |
| -3.0000 |        | 87   | Nord  | Sec   |      |      |         |         |
| 92      | 15.3   | 17.6 | 19.5  | 2     | 5    | 5    | 2.9544  | 1.8794  |
| 0.5209  |        | 82   | Est   | Sec   |      |      |         |         |
| 114     | 16.2   | 19.7 | 22.5  | 1     | 1    | 1    | 0.9848  | 0.3473  |
| -0.1736 |        | 92   | Nord  | Sec   |      |      |         |         |
| 94      | 17.4   | 20.5 | 20.4  | 8     | 8    | 8    | -0.5000 | -2.9544 |
| -4.3301 |        | 114  | Ouest | Sec   |      |      |         |         |
| 80      | 17.7   | 19.8 | 18.3  | 6     | 6    | 6    | -5.6382 | -5.0000 |
| -6.0000 |        | 94   | Ouest | Pluie |      |      |         |         |

Réflexe ... ???

Visu!

► Code



Question de base modifiée en : la moyenne du maximum d'ozone est-elle la même pour chaque direction du vent ?

# Problématique qui revient souvent

- L'effet de cultures intermédiaires est-il le même selon l'espèce/variété ?
- La présence de haies a-t-elle une influence sur la population d'auxiliaires de cultures ?
- Y a-t-il un effet de la variété sur le rendement du blé dans telle expérimentation ?
- etc. etc. etc.

De manière générale, l'ANOVA est le cadre d'analyse de **l'effet d'une variable qualitative** à  $I$  modalités (par ex. sol nu / moutarde / phacélie) sur une **variable quantitative**

Mais alors pourquoi parler d'Analyse de Variance si on veut comparer des moyennes ?

La réponse viendra avec les explications (et les mains dans le cambouis)

# Mettons les mains dans le cambouis

## Données & notations

| vent  | maxO3 | Notation |
|-------|-------|----------|
| Est   | 92    | $y_{11}$ |
| Est   | 121   | $y_{12}$ |
| Nord  | 87    | $y_{21}$ |
| Nord  | 82    | $y_{22}$ |
| Ouest | 94    | $y_{31}$ |
| Ouest | 80    | $y_{32}$ |
| Sud   | 90    | $y_{41}$ |
| ...   | ...   | ...      |

$y_{ij}$  valeur de  $Y$  pour l'individu  $j$  du groupe  $i$

Moyenne du groupe  $i$ :  $y_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$

Moyenne générale :  $y_{\bullet\bullet} = \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^{n_i} y_{ij}$

## Comparaison de I moyennes

- Dans chaque groupe, on considère que les données suivent une loi  $\mathcal{N}(\mu_i, \sigma^2)$

On considère le **MEME**  $\sigma^2$ .

$Y_{1j} \sim \mathcal{N}(\mu_1, \sigma^2)$  peut s'écrire  $Y_{1j} = \mu_1 + \varepsilon_{1j}$  avec  $\varepsilon_{1j} \sim \mathcal{N}(0, \sigma^2)$

$Y_{2j} \sim \mathcal{N}(\mu_2, \sigma^2)$  peut s'écrire  $Y_{2j} = \mu_2 + \varepsilon_{2j}$  avec  $\varepsilon_{2j} \sim \mathcal{N}(0, \sigma^2)$

...

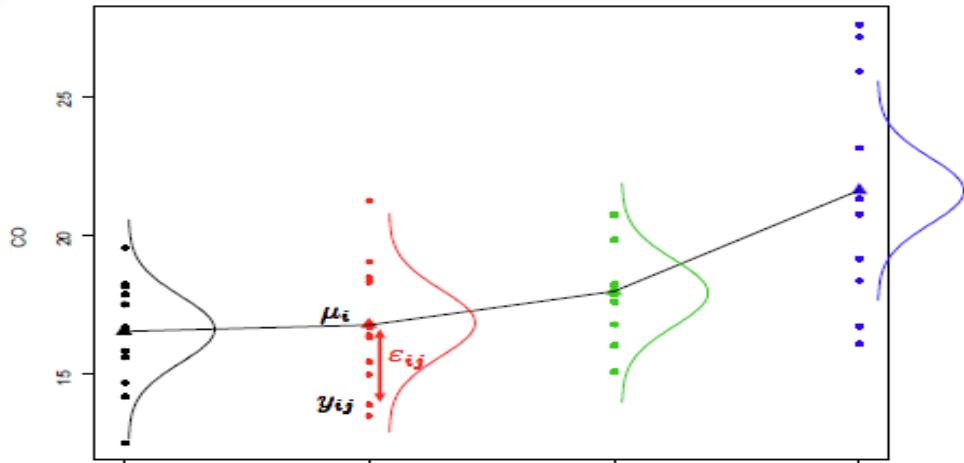
$Y_{Ij} \sim \mathcal{N}(\mu_I, \sigma^2)$  peut s'écrire  $Y_{Ij} = \mu_I + \varepsilon_{Ij}$  avec  $\varepsilon_{Ij} \sim \mathcal{N}(0, \sigma^2)$

⇒ On peut résumer tout ça dans un seul modèle:

$$\begin{cases} \forall i, j \quad Y_{ij} = \mu_i + \varepsilon_{ij} \\ \forall i, j \quad \mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0, \sigma^2) \\ \forall (i, j) \neq (i', j') \quad \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

# Deux définitions du modèle

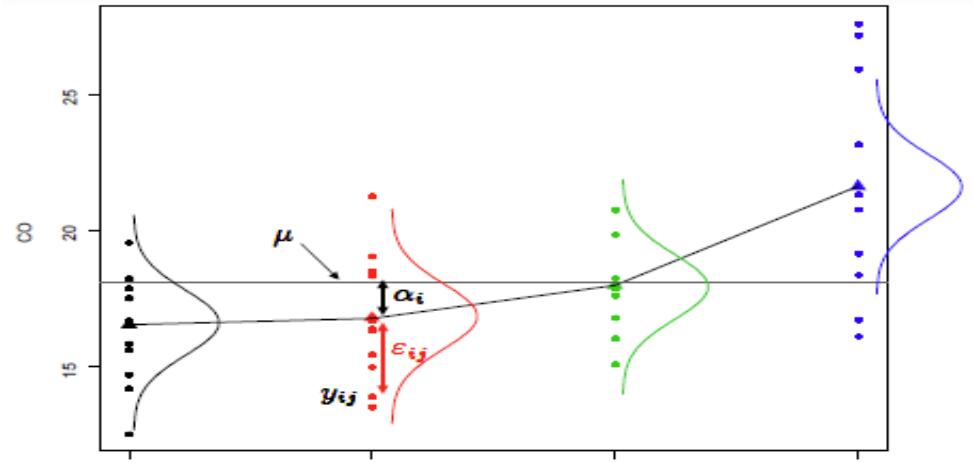
Si  $Y_{ij}$  est la valeur de la réponse du  $j^{eme}$  individu ( $j = 1, \dots, n_i$ ) du  $i^{eme}$  groupe ( $i = 1, \dots, I$ ):



$$\begin{cases} \forall i, j \quad Y_{ij} = \mu_i + \varepsilon_{ij} \\ \forall i, j \quad \mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0, \sigma^2) \\ \forall (i, j) \neq (i', j') \quad cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

$I$  paramètres: les  $\mu_i$

On doit avoir  $\forall i, \mu + \alpha_i = \mu_i$  besoin de **contraintes**



$$\begin{cases} \forall i, j \quad Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \\ \forall i, j \quad \mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0, \sigma^2) \\ \forall (i, j) \neq (i', j') \quad cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

$I + 1$  paramètres: l'effet moyen  $\mu$  et les  $I$  coefficients  $\alpha_i$  (effet du niveau i)

## Estimer les paramètres du modèle

On minimise les écarts entre observations ( $Y_{ij}$ ) et prédictions par le modèle  $\hat{Y}_{ij}$ .

On minimise le **critère des moindres carrés ordinaires** (? : pourquoi carré<sup>2</sup>)

$$SCER = \sum_{i=1}^I \sum_{j=1}^{n_i} \left( Y_{ij} - \hat{Y}_{ij} \right)^2 = \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij} - (\hat{\mu} + \hat{\alpha}_i))^2$$

$$\text{SCER minimal quand } \forall i, \hat{\mu} + \hat{\alpha}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = Y_{i\bullet}$$

Contraintes classiques:

- Un niveau particulier comme référence:  $\alpha_1 = 0 \Rightarrow \hat{\mu} = Y_{1\bullet}$  et  $\forall i, \hat{\alpha}_i = Y_{i\bullet} - Y_{1\bullet}$
- la moyenne des moyennes par groupe comme référence:  $\sum_{i=1}^I \alpha_i = 0 \Rightarrow \hat{\mu} = \frac{1}{I} \sum_{i=1}^I Y_{i\bullet}, \forall i, \hat{\alpha}_i = Y_{i\bullet} - Y_{1\bullet}$

## Exemple pour clarifier

| groupe 1           | groupe 2           | groupe3                     |
|--------------------|--------------------|-----------------------------|
| $y_{11} = 6$       | $y_{21} = 2$       | $y_{31} = 3$                |
| $y_{12} = 9$       | $y_{22} = 4$       | $y_{32} = 1$                |
| $y_{1\bullet} = 7$ | $y_{2\bullet} = 3$ | $y_{3\bullet} = 2$          |
|                    |                    | $y_{\bullet\bullet} = 4.25$ |

- Avec traitement 1 comme référence :  $\hat{\mu} = 7$  ;  $\hat{\alpha}_1 = 0$  ;  $\hat{\alpha}_2 = -4$  et  $\hat{\alpha}_3 = -5$

- $\sum_{i=1}^I \alpha_i = 0$  :  $\hat{\mu} = \frac{1}{3}(7 + 3 + 2) = 4$  ;  $\hat{\alpha}_1 = 3$  ;  $\hat{\alpha}_2 = -1$  et  $\hat{\alpha}_3 = -2$

L'interprétation d'un coefficient  $\alpha_i$  dépend de la contrainte choisie !

## Variance résiduelle

Valeurs prédictes:  $\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i = y_{i\bullet}$

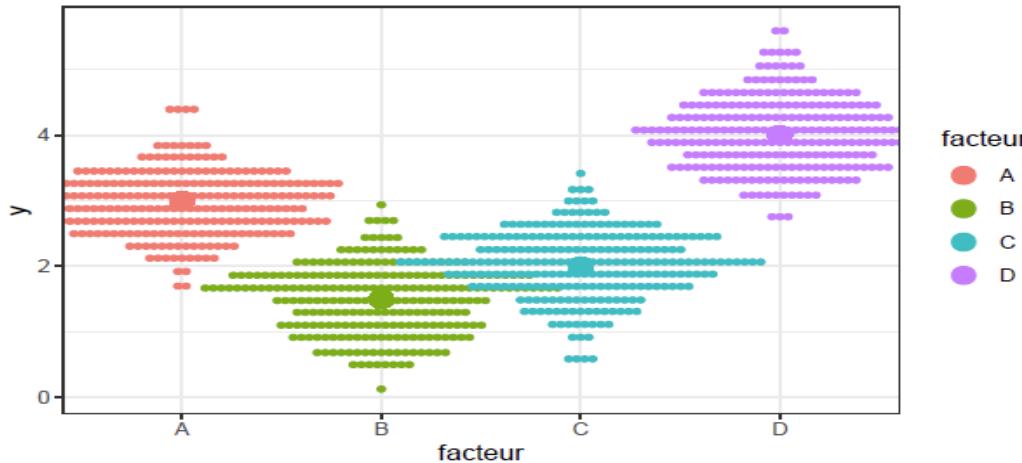
Erreurs d'ajustement ou résidus :  $\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - y_{i\bullet}$

Estimateur de la variabilité résiduelle  $\sigma^2$ :

$$\hat{\sigma}^2 = \frac{\sum_{ij} (Y_{ij} - Y_{i\bullet})^2}{n - I} = \frac{\sum_{ij} \hat{\varepsilon}_{ij}^2}{n - I} \quad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

$n - I$  degrés de liberté sont associés à la somme des carrés des résidus du modèle.

# Décomposition de la variabilité



## Equation d'analyse de la variance

$$\underbrace{\sum_{i,j} (Y_{ij} - Y_{\bullet\bullet})^2}_{SC_T} = \underbrace{\sum_{i,j} (Y_{i\bullet} - Y_{\bullet\bullet})^2}_{SC_F} + \underbrace{\sum_{i,j} (Y_{ij} - Y_{i\bullet})^2}_{SC_R}$$

| Variabilité | totale  | modèle  | résiduelle |
|-------------|---------|---------|------------|
| ddl         | $n - 1$ | $I - 1$ | $n - I$    |

- Vous pouvez commencer à comprendre d'où vient le nom ANalysis Of VAriance

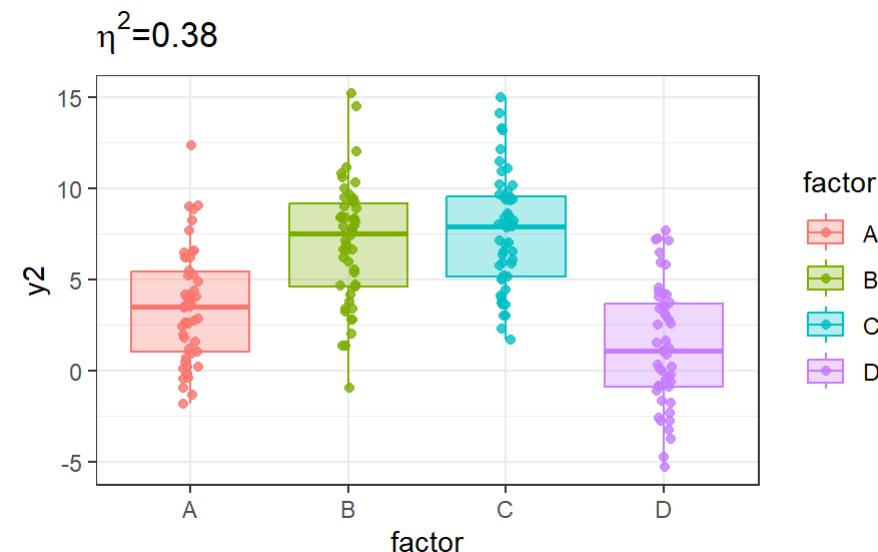
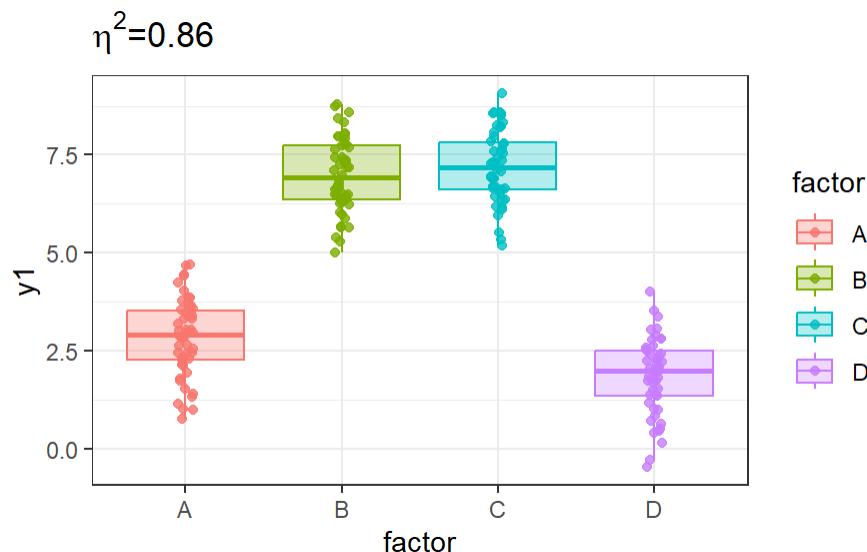
# Indicateur de liaison : rapport de corrélation

$$\eta^2 = \frac{SC_{\text{modèle}}}{SC_{\text{total}}} = 1 - \frac{SC_{\text{résiduelle}}}{SC_{\text{total}}}$$

Expliqué à mon grand-père (ou presque): la proportion de variabilité du phénomène que mon modèle explique

Propriétés:

- $0 \leq \eta^2 \leq 1$  (obvious)
- $\eta^2 = 0 \Leftrightarrow SC_{\text{modèle}} = 0$
- $\eta^2 = 1 \Leftrightarrow SC_{\text{modèle}} = SC_{\text{total}}$



Deux rapports de corrélation différents pour des  $Y_{i\bullet}$  identiques.

# Inférence: test global

Rappel de l'objectif: y a-t-il un effet du facteur sur Y ?

→ La variabilité de Y est-elle expliquée par le facteur groupe ? Ou bien peut-on considérer que les données proviennent d'une même loi  $\mathcal{N}(\mu, \sigma^2)$  ?

Hypothèses:

$$H_0 : \forall i, \mu_i = \mu \quad \Leftrightarrow$$

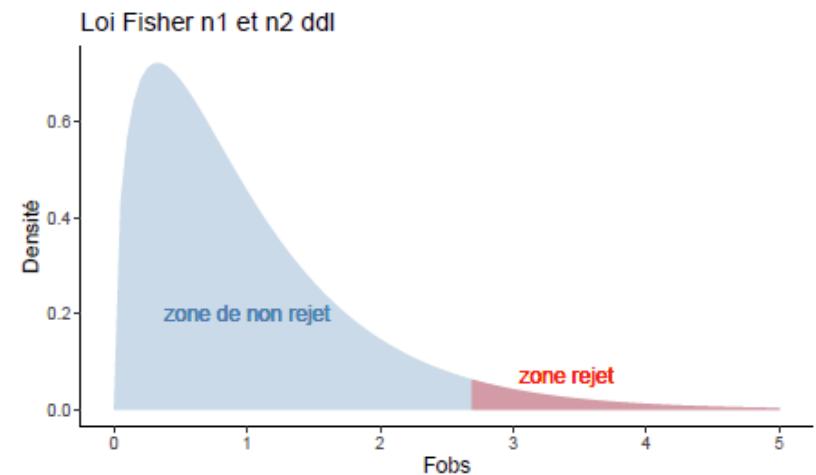
$$H_0 : \forall i, \alpha_i = 0$$

$$H_1 : \exists i / \mu_i \neq \mu$$

$$H_1 : \exists i / \alpha_i \neq 0$$

On a :  $\mathbb{E}\left(\frac{SC_{mod}}{I-1}\right) = \sigma^2 + \frac{1}{I-1} \sum_{i=1}^I n_i \alpha_i^2$     $\mathbb{E}\left(\frac{SC_R}{n-I}\right) = \sigma^2$

- Statistique de test :  $F_{obs} = \frac{SC_{mod}/(I-1)}{SC_R/(n-I)}$
- Loi de  $F_{obs}$  sous  $H_0$  :  $\mathcal{L}(F_{obs}) = \mathcal{F}_{n-I}^{I-1}$



# Table d'analyse de variance: décomposition de variabilité et test

- Ce qui est souvent reporté dans le cadre d'une ANOVA

| Variabilité | Somme Carrés                                       | ddl     | Carré moyen          | $F_{obs}$           |
|-------------|--|---------|----------------------|---------------------|
| Facteur     | $\sum_i n_i (Y_{i\bullet} - Y_{\bullet\bullet})^2$ | $I - 1$ | $\frac{SC_F}{I - 1}$ | $\frac{CM_F}{CM_R}$ |
| Résiduelle  | $\sum_{i,j} (Y_{ij} - Y_{i\bullet})^2$             | $n - I$ | $\frac{SC_R}{n - I}$ |                     |
| Totale      | $\sum_{i,j} (Y_{ij} - Y_{\bullet\bullet})^2$       | $n - 1$ |                      |                     |

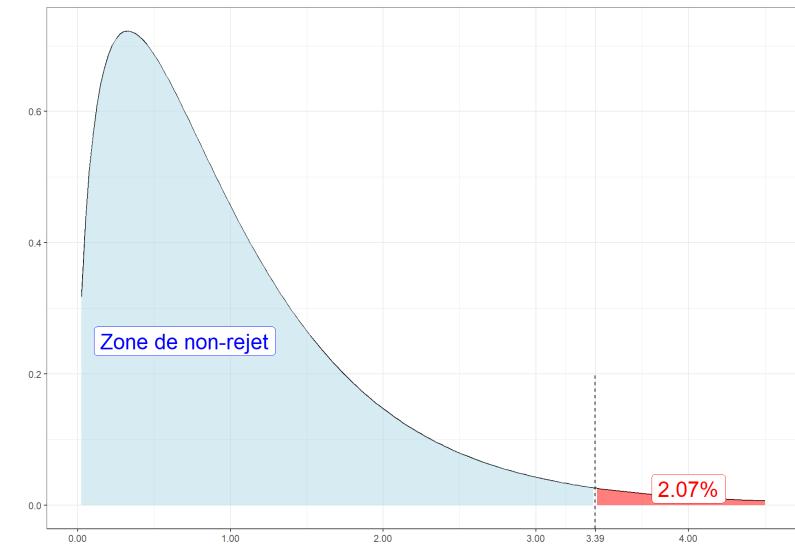
```
1 library(FactoMineR)
2 LinearModel(maxO3 ~ vent, data = ozone)
```

```
Call:
LinearModel(formula = maxO3 ~ vent, data = ozone)

Residual standard error: 27.32 on 108 degrees of freedom
Multiple R-squared:  0.08602
F-statistic: 3.388 on 3 and 108 DF,  p-value: 0.02074
AIC = 744.8      BIC = 755.7
```

```
Ftest
      SS   df   MS F value Pr(>F)
vent    7586   3 2528.69  3.3881 0.02074
Residuals 80606 108   746.35
```

```
Ttest
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.7382   3.0535 31.0265 < 2e-16
vent - Est   10.8618   6.8294  1.5904 0.11466
vent - Nord  -8.6092   4.6219 -1.8627 0.06522
vent - Ouest -10.0382  4.0972 -2.4500 0.01589
vent - Sud     7.7856   5.2052  1.4957 0.13764
```



## Inférence: test de conformité d'un coefficient (1/2)

Si on rejette l'hypothèse  $H_0 : \forall i, \alpha_i = 0$ , on veut savoir quels  $\alpha_i$  sont différents de 0

La valeur de  $\hat{\alpha}_i$  dépend de l'échantillon de données et donc l'estimateur  $\hat{\alpha}_i$  est une variable aléatoire

$$\begin{aligned}\mathcal{L}(\hat{\alpha}_i) &= \mathcal{N}\left(\alpha_i, \sigma_{\hat{\alpha}_i}^2\right) \iff \mathcal{L}\left(\frac{\hat{\alpha}_i - \alpha_i}{\sigma_{\hat{\alpha}_i}}\right) = \mathcal{N}(0, 1) \\ \mathcal{L}\left(\frac{\hat{\alpha}_i - \alpha_i}{\hat{\sigma}_{\hat{\alpha}_i}}\right) &= \mathcal{T}_{n-I}\end{aligned}$$

On peut donc construire le test de nullité d'un coefficient ( $\alpha_1$  par exemple):

**Hypothèses :**  $H_0 : \alpha_1 = 0$  contre  $H_1 : \alpha_1 \neq 0$

**Statistique de test**  $\frac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}}$

**Loi de la statistique de test sous  $H_0$**   $\mathcal{L}\left(T_{obs} = \frac{\hat{\alpha}_1}{\hat{\sigma}_{\hat{\alpha}_1}}\right) = \mathcal{T}_{\nu=n-I}$

**Décision** par la p-value

Rq : connaissant la loi de  $\hat{\alpha}_1$ , on peut construire un intervalle de confiance:

$$\alpha_1 \in [\hat{\alpha}_1 - \hat{\sigma}_{\hat{\alpha}_1} \times t_{0.975}(n - I); \hat{\alpha}_1 + \hat{\sigma}_{\hat{\alpha}_1} \times t_{0.975}(n - I)]$$



## Inférence: test de conformité d'un coefficient (2/2)

```
Estimate Std. Error t value Pr(>|t|)  
(Intercept) 94.738210 3.053461 31.026504 1.290980e-55  
vent - Est 10.861790 6.829425 1.590440 1.146587e-01  
vent - Nord -8.609178 4.621850 -1.862713 6.522029e-02  
vent - Ouest -10.038210 4.097207 -2.450013 1.589160e-02  
vent - Sud 7.785599 5.205173 1.495743 1.376372e-01
```

```
1 library(FactoMineR)  
2 res <- LinearModel(maxO3 ~vent, data = ozone)  
3  
4 res$Ttest
```

```
1 Estimate Std. Error t value Pr(>|t|)  
2 (Intercept) 94.738210 3.053461 31.026504 1.290980e-55  
3 vent - Est 10.861790 6.829425 1.590440 1.146587e-01  
4 vent - Nord -8.609178 4.621850 -1.862713 6.522029e-02  
5 vent - Ouest -10.038210 4.097207 -2.450013 1.589160e-02  
6 vent - Sud 7.785599 5.205173 1.495743 1.376372e-01
```

# Test de comparaison 2 à 2

Autre stratégie : comparer toutes les paires de moyennes

Pb : on effectue beaucoup de tests  $\Rightarrow$  risque de multiplier les erreurs en rejetant des hypothèses  $H_0$ .

$\Rightarrow$  correction des tests: modifier le seuil  $\alpha = 5\%$  et prendre  $\alpha = \frac{5\%}{(\text{nb tests})}$

$\Rightarrow$  les tests sont peu puissants

```
1 library(FactoMineR)
2
3
4 res <- LinearModel(maxO3 ~vent, data = ozone)
5
6 meansComp(res, ~vent, adjust ="Bonferroni", graph = FALSE)

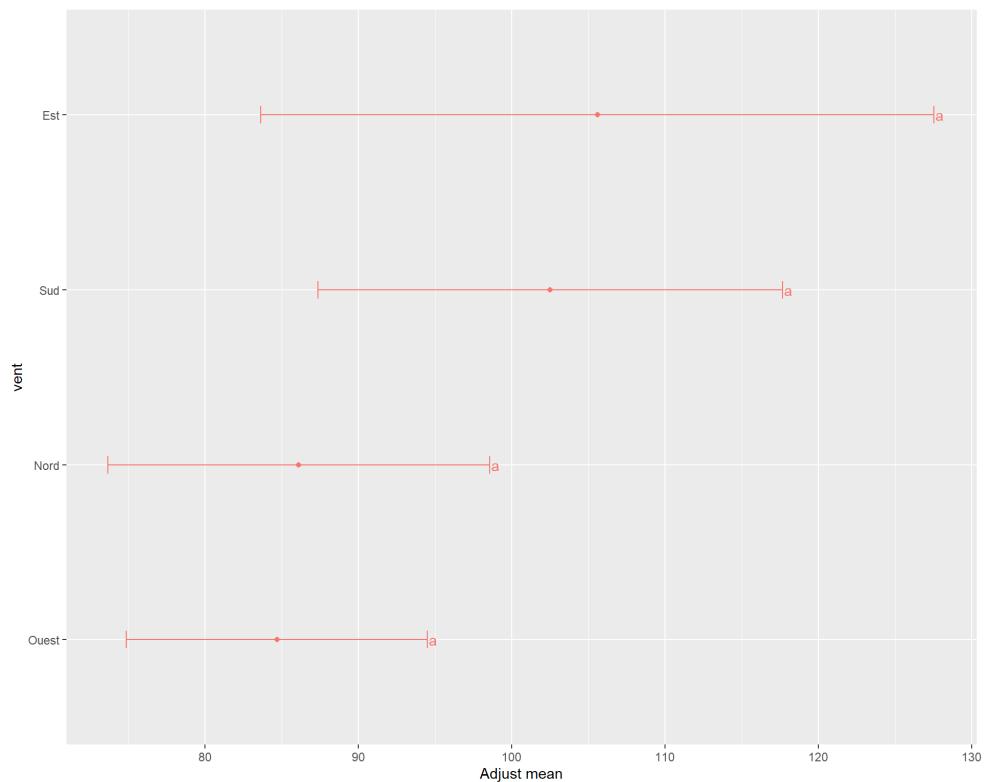
$adjMean
  vent   emmean    SE  df lower.CL upper.CL
Est     105.6  8.64 108     83.7   127.5
Nord    86.1  4.91 108     73.7    98.6
Ouest   84.7  3.86 108     74.9    94.5
Sud     102.5  5.96 108    87.4   117.7

Confidence level used: 0.95
Conf-level adjustment: bonferroni method for 4 estimates

$groupComp
$groupComp$Letters
Ouest   Nord   Sud   Est
  "a"    "a"    "a"    "a"

$groupComp$LetterMatrix
      a
Ouest TRUE
Nord  TRUE
Sud   TRUE
Est   TRUE

attr(),"class")
[1] "meansComp"
```



# Analyse des résidus du modèle

Rappel du modèle:

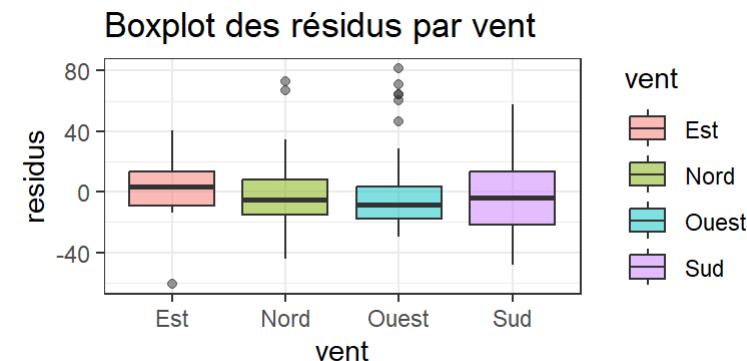
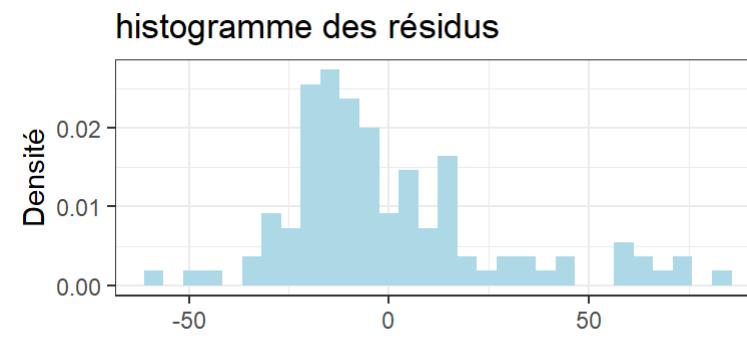
$$\begin{cases} \forall i, j \quad Y_{ij} = \mu_i + \varepsilon_{ij} \\ \forall i, j \quad \mathcal{L}(\varepsilon_{ij}) = \mathcal{N}(0, \sigma^2) \\ \forall (i, j) \neq (i', j') \quad \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$

- On a des hypothèses sur les *résidus* du modèle
- Les résidus, c'est ce qui n'est pas expliqué par le modèle. Donc pour les estimer, il faut *ajuster* le modèle

On peut ensuite vérifier les hypothèses du modèle

```
1 residus <- data.frame(residus=res$lmResult$residuals, vent = ozone$vent)
2 residus %>%
3   ggplot(aes(x=residus)) +
4   geom_histogram(fill="lightblue", aes(y = after_stat(density))) +
5   theme_bw() +
6   labs(x = "Résidu", y = "Densité", title = "histogramme des résidus")
```

```
1 residus %>%
2   ggplot(aes(y=residus,x=vent,fill=vent)) +
3   geom_boxplot(alpha = .5) +
4   labs(title = "Boxplot des résidus par vent") +
5   theme_bw()
```

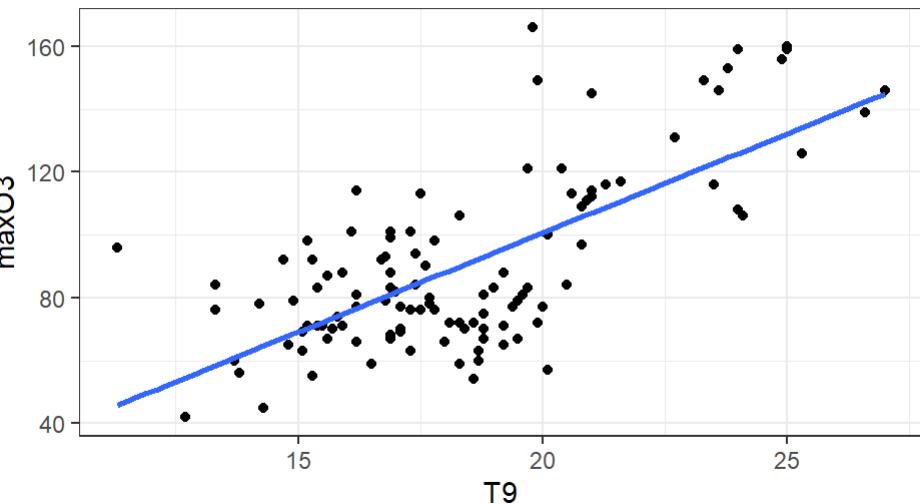


# Régression linéaire simple

# Liaison linéaire

```
1 ozone %>%
2   ggplot(aes(x = T9, y= maxO3)) +
3   geom_point() +
4   geom_smooth(method = "lm", se = FALSE) + # se= FALSE pour virer les
# intervalles de confiance, method = "lm" ajustement linéaire
5   labs(title = "Lien entre température à 9°C et maximum d'ozone") +
6   theme_bw()
```

Lien entre température à 9°C et maximum d'ozone



## Questions ?

- Influence de la température sur le max d'ozone ?
- A quel max d'ozone peut-on s'attendre s'il fait 19.5°C à 9h ?

## Objectifs ⓘ

- Etudier qualitativement et quantitativement la dépendance d'une variable réponse quantitative Y en fonction d'une variable quantitative x
- La variable x permet elle d'expliquer la variabilité de la variable Y?
- Prédire Y à partir de x

# Un autre indice de liaison: le coefficient de corrélation linéaire

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- $-1 \leq r_{xy} \leq 1$
- $r_{xy} = 0 \Leftrightarrow$  pas de liaison *linéaire* entre  $X$  et  $Y$
- $r_{xy} \approx 1 \Leftrightarrow$  relation *linéaire* croissante entre  $X$  et  $Y$
- $r_{xy} \approx -1 \Leftrightarrow$  relation *linéaire* décroissante entre  $X$  et  $Y$

```
1 cor.test(ozone$maxO3, ozone$T9)
```

```
Pearson's product-moment correlation

data: ozone$maxO3 and ozone$T9
t = 10.263, df = 110, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.5904575 0.7832906
sample estimates:
cor
0.6993865
```

- $r_{xy} = 0.70 \Rightarrow$  corrélation assez forte (et significativement différente de 0, mais on verra ça plus tard)

## Corrélation, causalité etc. (1/2)

- Une corrélation est un indicateur utile, mais à utiliser avec précaution

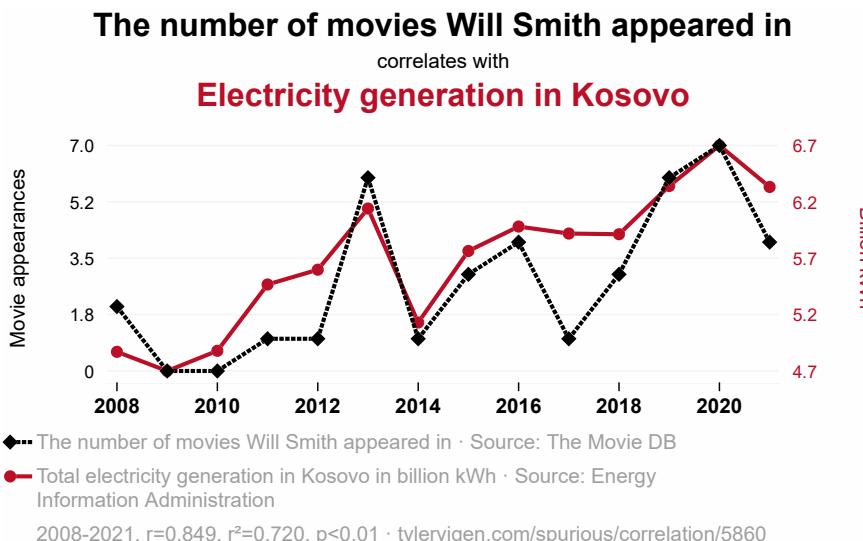
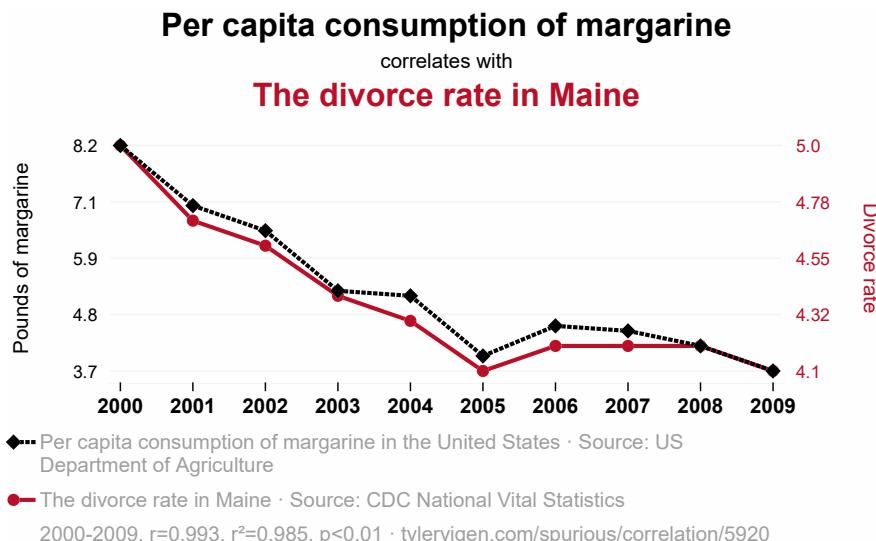
“Corrélation n’implique pas causalité”

Oui, mais... la présence d'une corrélation suppose généralement un lien

- Ex. on observe une corrélation entre A et B. Plusieurs cas existent.
- $A \Rightarrow B$  (ex.   $\Rightarrow$  cancer 
- $B \Rightarrow A$  (idem)
- $A \Rightarrow B$  et  $B \Rightarrow A$  (ex.  et  $CO_2$ )
- Une variable C a été omise, et influence A et B. (ex.  (C)  $\Rightarrow$   (A),  (C)  $\Rightarrow$   (B),   $\Leftrightarrow$  
- Corrélation purement fortuite (site [spurious correlations](#))

## Corrélation, causalité etc. (2/2)

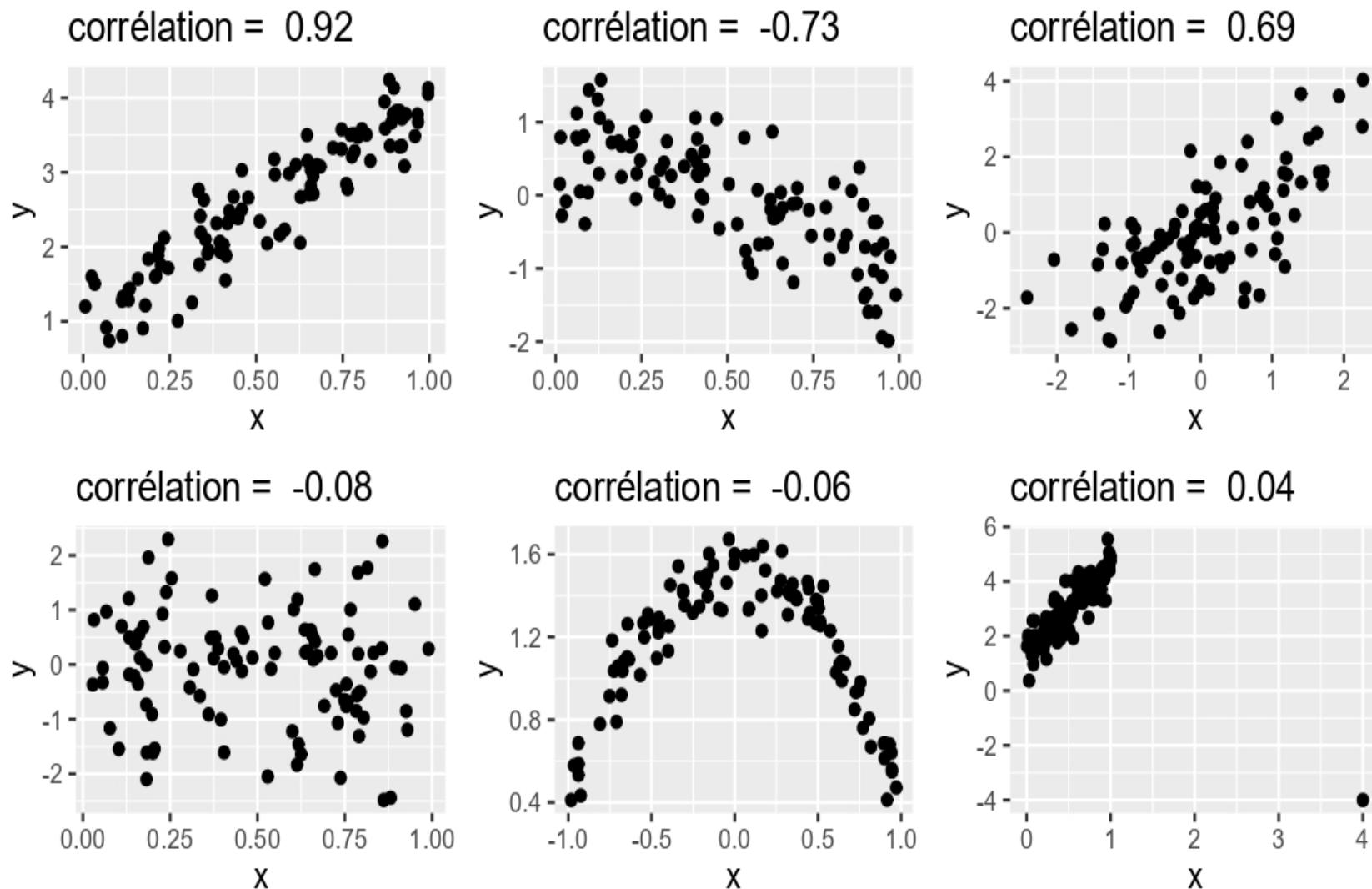
Ex. pour rigoler:



“Corrélation n’implique pas causalité”

Ok, mais cette affirmation ne doit pas servir à réfuter toute association constatée dans des données.

## Quelques valeurs de corrélations



## (Re)Mettons les mains dans le cambouis

Le modèle de régression linéaire simple s'écrit ainsi:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \stackrel{ind}{\sim} \mathcal{N}(0, \sigma^2),$$

avec

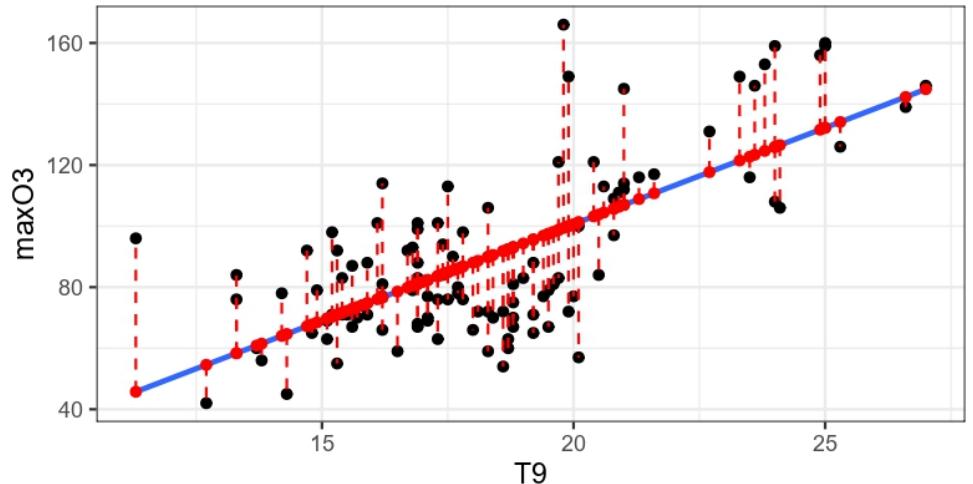
- $x_i$  la valeur de la variable explicative pour l'observation  $i$
- $i = 1, \dots, n$  le numéro d'individu,  $n$  le nombre total d'individus
- $\beta_0$  l'ordonnée à l'origine
- $\beta_1$  la pente de la droite, mesure de l'effet de la variable  $x$
- $\sigma^2$  la variance

# Estimation des paramètres

$$SCER = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Paramètres d'espérance:

- $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$



```
1 res <- LinearModel(maxO3 ~ T9, data = ozone)
2 res$Ttest
```

|             | Estimate   | Std. Error | t value   | Pr(> t )     |
|-------------|------------|------------|-----------|--------------|
| (Intercept) | -25.607608 | 11.4551179 | -2.235473 | 2.740572e-02 |
| T9          | 6.312999   | 0.6151378  | 10.262740 | 9.729952e-18 |

Variance résiduelle:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad ddl_{\text{résiduelle}} = n-2 \quad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

## Test de conformité

$$\begin{aligned}\mathcal{L}(\hat{\beta}_1) &= \mathcal{N}(\beta_1, \sigma_{\beta_1}^2) \text{ avec } \sigma_{\beta_1}^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \implies \mathcal{L}\left(\frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}\right) &= \mathcal{N}(0, 1) \implies \mathcal{L}\left(\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}}\right) = \mathcal{T}_{n-2}\end{aligned}$$

On peut donc construire le test de nullité de  $\beta_1$ :

Hypothèses :  $H_0 : \beta_1 = 0$  contre  $H_1 : \beta_1 \neq 0$

Statistique de test :  $\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$  Loi de la statistique de test sous  $H_0$ :  $\mathcal{L}\left(T_{obs} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\right) = \mathcal{T}_{\nu=n-2}$

Rq 1. : comme dans l'ANOVA, connaissant la loi de  $\hat{\beta}_1$ , on peut construire un intervalle de confiance:

$$\beta_1 \in \left[ \hat{\beta}_1 - \hat{\sigma}_{\hat{\beta}_1} \times t_{0.975}(n-2); \hat{\beta}_1 + \hat{\sigma}_{\hat{\beta}_1} \times t_{0.975}(n-2) \right]$$

Rq 2. : on peut aussi tester  $\beta_0$  mais cela a moins d'importance en pratique

# Décomposition de la variabilité

$$\begin{aligned}\sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ SCT &= SCM + SCR \\ n-1 &= 1 + n-2\end{aligned}$$

## Table d'analyse de la variance

| Source    | Somme de Carrés | Degrés de liberté | Carré moyen       | F                 |
|-----------|-----------------|-------------------|-------------------|-------------------|
| variation | carrés          | liberté           | moyen             |                   |
| Modèle    | SCM             | 1                 | $\frac{SCM}{1}$   | $\frac{CMM}{CMR}$ |
| Erreur    | SCR             | n-2               | $\frac{SCR}{n-2}$ |                   |
| Total     | SCT             | n-1               |                   |                   |

Comparaison de  $SCM$  et  $SCT$  par le critère  $R^2 = \frac{SCM}{SCT}$

Propriétés :

- $0 \leq R^2 \leq 1$
- $R^2 = 0 \Leftrightarrow SC_{\text{modèle}} = 0$
- $R^2 = 1 \Leftrightarrow SC_{\text{modèle}} = SC_{\text{total}}$

Explication claire du concept sur le [site](#) de Science Etonnante

```
1 res <- LinearModel(maxO3 ~T9, data= ozone)
2 res$Ftest

SS          df     MS   F value    Pr(>F)
T9        43138     1 43138 105.32 < 2.2e-16 ***
Residuals 45053 110    410
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Test du modèle

## Hypothèses

- $H_0 : \beta_1 = 0, \Leftrightarrow$  le modèle n'a pas d'intérêt (x n'explique pas Y)
- $H_1 : \beta_1 \neq 0, \Leftrightarrow$  le modèle a un intérêt (x explique Y)

## Stat de Fisher

$$F = \frac{SCM/1}{SCR/(n-2)}$$

Loi de F sous  $H_0$  :  $\mathcal{L}(F) = \mathcal{F}_{n-2}^1$

# Prédiction et intervalle de confiance

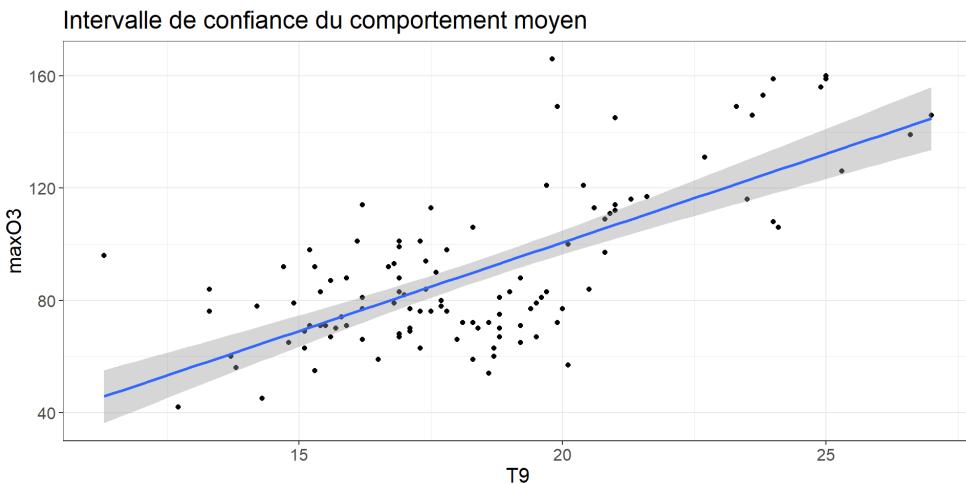
Pour une valeur de  $x_0$  particulière, on peut maintenant prédire  $Y$ :

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Prédiction de la valeur moyenne de  $Y$  pour un  $x_0$  particulier

$$\mathbb{E}(\hat{Y}_0 | x_0) \sim \mathcal{N}(Y_0, \sigma \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}})$$

► Code

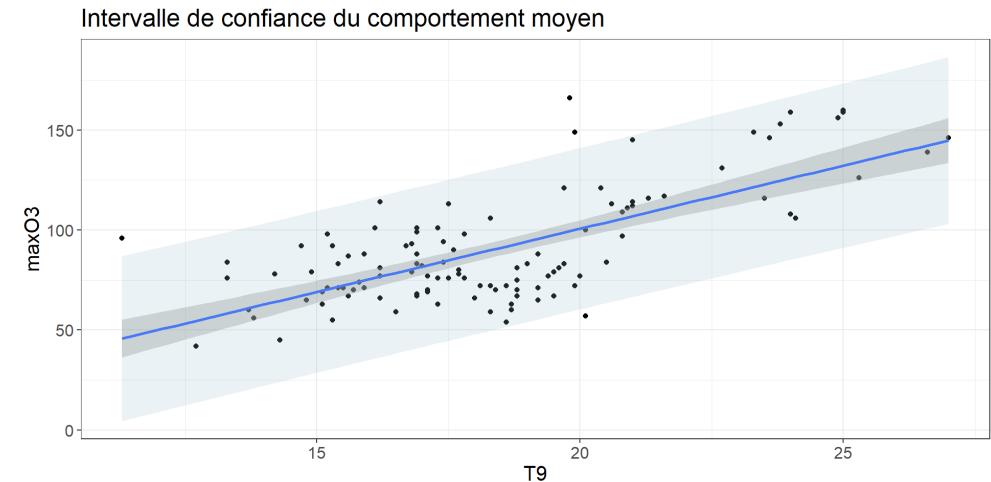


Différent de:

Prédiction d'une nouvelle valeur de  $Y$  pour un  $x_0$  donné (*notez le 1 supplémentaire dans la variance*).

$$\hat{Y}_0 \sim \mathcal{N}(Y_0, \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}})$$

► Code

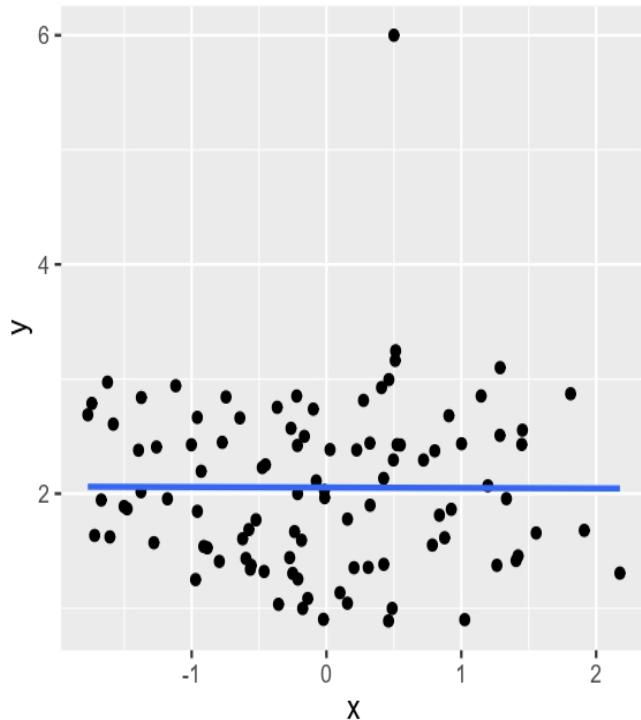


# Validité du modèle

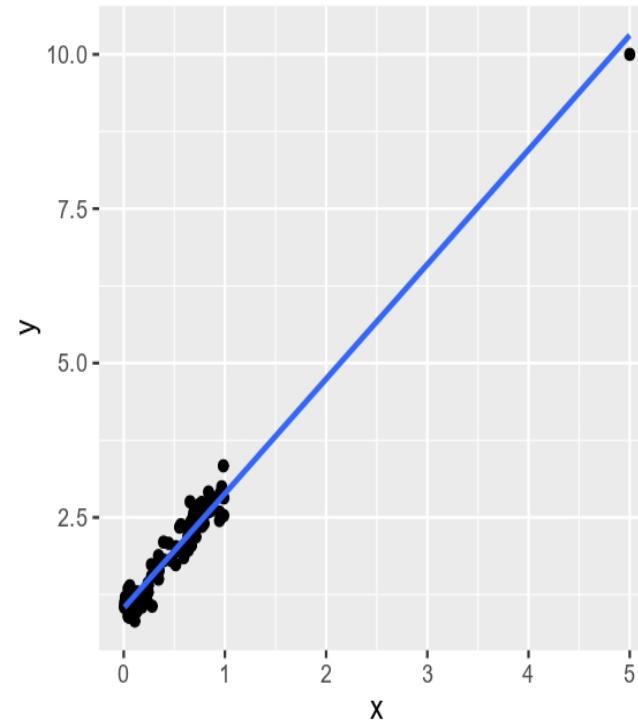
Vérification des hypothèses de départ:

- Normalité des résidus, stabilité de la variance, indépendance des résidus

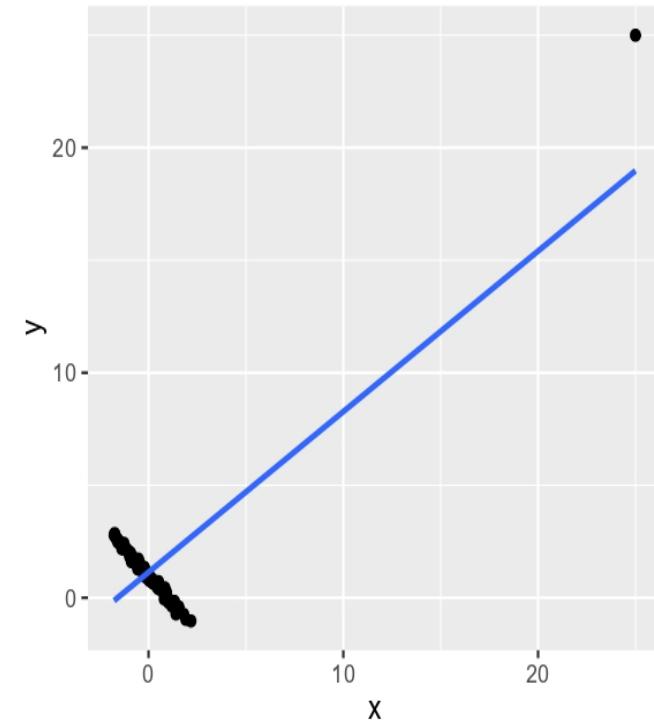
Point aberrant



Point levier avec faible résidu



Point levier très problématique



# **Construction et sélection de modèle**

# Modéliser = Comprendre et prévoir

? Des exemples de modélisation que vous effectuez tous les jours ?

- Budget pour évènement  
-   vs  
- 

? Comment faites-vous ?

- Ex. estimation du budget d'une soirée d'anniversaire
- Vous listez les variables (     etc.)
- Vous éliminez celles qui sont négligeables (ex.  )
- Vous quantifiez l'effet des variables restantes

Les statistiques permettent de faire cela avec des phénomènes complexes, en partant de données collectées.

# Exemple de problématiques

- Prévoir la sévérité d'une maladie fongique en fonction de la météo (P/ETP/T°) et de l'itk (gestion des résidus, sensibilité variétale etc;)
- Prédire les potentiels de rendements futurs du soja en France
- Potentiel de production éolien en fonction des conditions météorologiques

## Objectifs:

- *Comprendre* quelles variables ont une influence sur une variable quanti
- *Prévoir* les valeurs de la variable réponse dans de nouvelles conditions

## Régression simple vs multiple

- En régression simple, on ne considérait *qu'une* seule variable explicative
- En régression multiple, on en considère *plusieurs* (p)

$$\text{Réponse} = f(\text{var1}, \textcolor{red}{var_2}, \dots, \textcolor{red}{var_p}) = \underbrace{\text{var1} + \textcolor{red}{var_2} + \dots + \textcolor{red}{var_p}}_{\text{régression linéaire}}$$

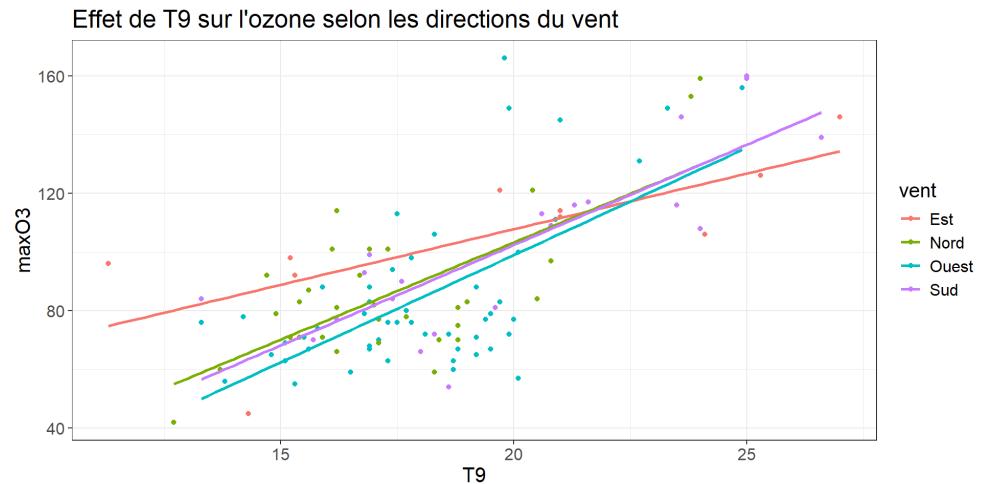
$$\begin{cases} \forall i = 1, \dots, n \quad \varepsilon_i \text{ i.i.d. , } \mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad \text{cov}(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

# Une variable quanti peut avoir un effet différent selon différentes modalités d'une quali

```

1 ozone %>%
2   ggplot(aes(x = T9, y = maxO3, color = vent)) +
3   geom_point() +
4   geom_smooth(method = "lm", se = FALSE) +
5   labs(title = "Effet de T9 sur l'ozone selon les directions du vent",
6   theme_bw() +
7   theme(text = element_text(size = 15))

```



Cela s'écrit ainsi:

$$MaxO3 \sim vent + T9 + vent:T9$$

$$MaxO3_{ij} \sim \mu + \left\{ \begin{array}{l} \alpha_1 \text{ si vent d'est} \\ \alpha_2 \text{ si vent du nord} \\ \alpha_3 \text{ si vent d'ouest} \\ \alpha_4 \text{ si vent du sud} \end{array} \right\} + \left( \beta + \left\{ \begin{array}{l} \gamma_1 \text{ si vent d'est} \\ \gamma_2 \text{ si vent du nord} \\ \gamma_3 \text{ si vent d'ouest} \\ \gamma_4 \text{ si vent du sud} \end{array} \right\} \right) \times T9_{ij} + alea_{ij}$$

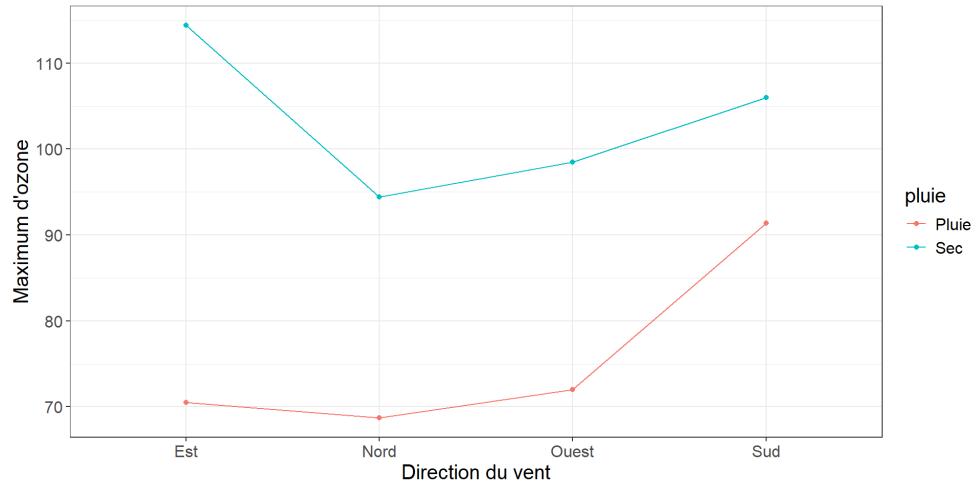
$$\left\{ \begin{array}{l} \forall i, j \quad Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i) \times x_{ij} + \varepsilon_{ij} \\ \forall i, j \quad \varepsilon_{ij} \text{ i.i.d.}, \quad \mathbb{E}(\varepsilon_{ij}) = 0, \quad \mathbb{V}(\varepsilon_{ij}) = \sigma^2 \\ \forall i, j \quad cov(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{array} \right.$$

# Interaction entre deux variables qualitatives

~~Définition courante: réaction réciproque de deux phénomènes l'un sur l'autre~~

Définition statistique : l'effet d'un facteur sur Y diffère selon les modalités de l'autre facteur

```
1 library(dplyr)
2
3 ozone %>%
4   group_by(vent, pluie) %>%
5   summarize(MOY = mean(maxO3)) %>%
6   ggplot(aes(x=vent, y=MOY, col=pluie, group=pluie)) +
7   geom_line() +
8   geom_point() +
9   labs("Interaction pluie:vent sur maxO3", x = "Direction du vent", y =
10 theme_bw() +
11 theme(text = element_text(size = 15))
```



Cela s'écrit ainsi:

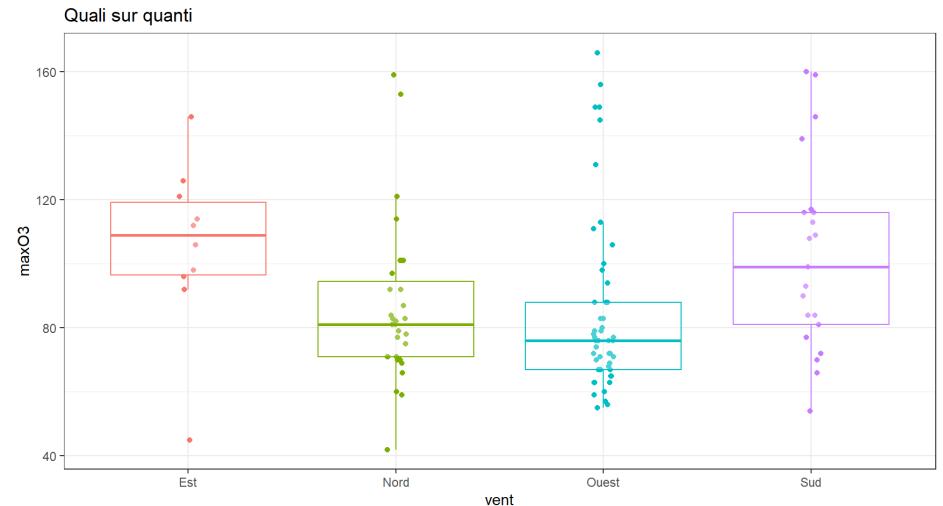
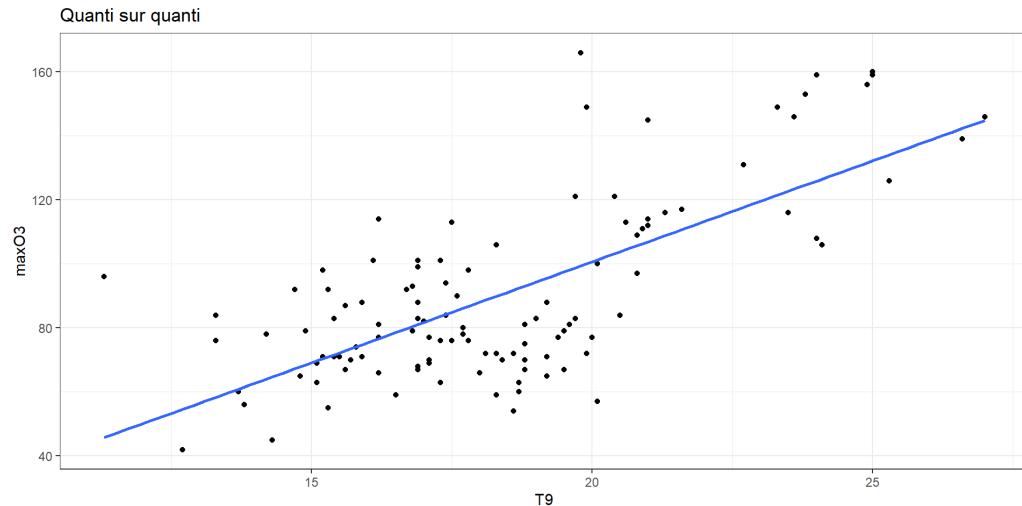
$$MaxO3 \sim vent + pluie + vent:pluie$$

$$MaxO3_{ijk} \sim \mu + \left\{ \begin{array}{l} \alpha_1 \text{ si vent d'est} \\ \alpha_2 \text{ si vent du nord} \\ \alpha_3 \text{ si vent d'ouest} \\ \alpha_4 \text{ si vent du sud} \end{array} \right\} + \left\{ \begin{array}{l} \beta_1 \text{ si pluie} \\ \beta_2 \text{ si sec} \end{array} \right\} + \left\{ \begin{array}{l} \alpha\beta_{11} \text{ si vent d'est ET pluie} \\ \alpha\beta_{12} \text{ si vent d'est ET sec} \\ \alpha\beta_{21} \text{ si vent du nord ET pluie} \\ \dots \alpha\beta_{42} \text{ si vent du sud ET sec} \end{array} \right\} + alea_{ijk}$$
$$\left\{ \begin{array}{l} \forall i, j, k \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \\ \forall i, j, k \quad \varepsilon_{ijk} \text{ i.i.d. , } \mathbb{E}(\varepsilon_{ijk}) = 0, \quad \mathbb{V}(\varepsilon_{ijk}) = \sigma^2 \\ \forall i, j, k \quad cov(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \end{array} \right.$$

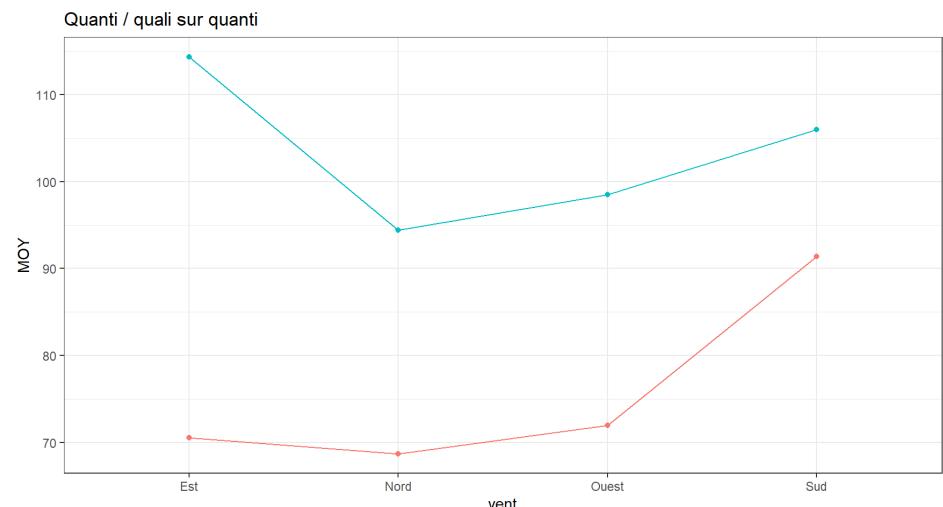
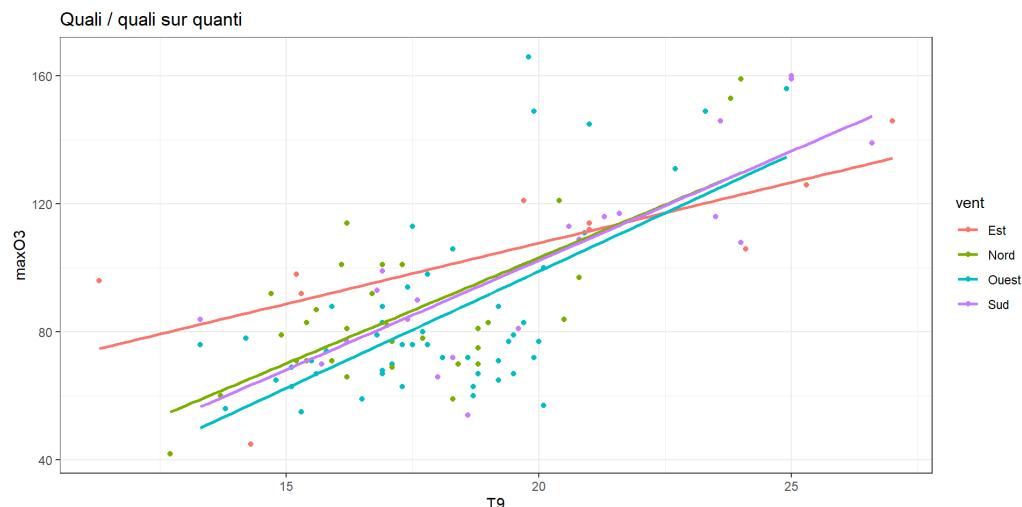


## Quatres types d'effet possibles

► Code



► Code





# Le(s) modèle(s) linéaire(s)

Le modèle linéaire peut souvent être adapté à des problèmes issus du vivant, modulo quelques points de vigilance

| Réponse                            | Variable(s) explicative(s)         | Méthode   |
|------------------------------------|------------------------------------|---|
| Var. quantitative                  | 1 var. quantitative                | régression linéaire simple  |
| Var. quantitative                  | 1 var. qualitative à $I$ modalités | analyse de variance à 1 facteur (rq: si $I = 2$ équivaut à comparaison de 2 moyennes) |
| Var. quantitative                  | $p$ var. quantitatives             | régression linéaire multiple  |
| Var. quantitative                  | $K$ var. qualitatives              | analyse de variance à $K$ facteurs  |
| Var. quantitative                  | var. quantitatives et qualitatives | analyse de covariance   |
| Var. qualitative (2 catégories)    | var. quantitatives et qualitatives | régression logistique   |
| Var. qualitative ( $K$ catégories) | var. quantitatives et qualitatives | régression multinomiale   |

## (Re-Re)mettons les mains dans le cambouis

$$\begin{cases} \forall i = 1, \dots, n \quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_p x_{ip} + \varepsilon_i \\ \forall i = 1, \dots, n \quad \varepsilon_i \text{ i.i.d.}, \quad \mathbb{E}(\varepsilon_i) = 0, \quad \mathbb{V}(\varepsilon_i) = \sigma^2 \\ \forall i \neq k \quad \text{cov}(\varepsilon_i, \varepsilon_k) = 0 \end{cases}$$

(p+1) paramètres à estimer + 1 paramètre de variance  $\sigma^2$

Matriciellement :  $Y = X\beta + E$  avec  $\mathbb{E}(E) = 0, \quad \mathbb{V}(E) = \sigma^2 Id$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & \beta_0 & \beta_1 & \beta_2 & \dots & \beta_p \\ \vdots & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{i1} & x_{i2} & & x_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Rq: ANOVA et ANCOVA peuvent aussi s'écrire sous cette forme.

### ANOVA

$$\begin{cases} \forall i, j, k \quad Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \varepsilon_{ijk} \\ \forall i, j, k \quad \varepsilon_{ijk} \text{ i.i.d.}, \quad \mathbb{E}(\varepsilon_{ijk}) = 0, \quad \mathbb{V}(\varepsilon_{ijk}) = \sigma^2 \\ \forall i, j, k \quad \text{cov}(\varepsilon_{ijk}, \varepsilon_{i'j'k'}) = 0 \end{cases}$$

### ANCOVA

$$\begin{cases} \forall i, j \quad Y_{ij} = \mu + \alpha_i + (\beta + \gamma_i) \times x_{ij} + \varepsilon_{ij} \\ \forall i, j \quad \varepsilon_{ij} \text{ i.i.d.}, \quad \mathbb{E}(\varepsilon_{ij}) = 0, \quad \mathbb{V}(\varepsilon_{ij}) = \sigma^2 \\ \forall i, j \quad \text{cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) = 0 \end{cases}$$



# Estimation des paramètres du modèle

**Critère des moindres carrés:** estimer les paramètres en minimisant la somme des carrés des écarts entre observations et prévisions par le modèle

$$Y \approx X\beta$$

$$X'Y \approx X'X\beta$$

$$\hat{\beta} = (X'X)^{-1}X'Y \text{ si } X'X \text{ est inversible}$$

**Propriétés**  $\mathbb{E}(\hat{\beta}) = \beta$ ;  $\mathbb{V}(\hat{\beta}) = (X'X)^{-1}\sigma^2$

La variance des résidus  $\sigma^2$  est estimée par:

$$\hat{\sigma}^2 = \frac{\sum_i (Y_i - \hat{Y}_i)^2}{\text{nb données} - \text{nb paramètres estimés é partir des données}} \quad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

## Décomposition de la variabilité (bis)

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

---

|             |        |        |            |
|-------------|--------|--------|------------|
| Variabilité | totale | modèle | résiduelle |
|-------------|--------|--------|------------|

Pourcentage de variabilité de  $Y$  expliquée par le modèle:  $R^2 = \frac{SC_{modèle}}{SC_{totale}}$

Propriétés:  $R^2 \in [0, 1]$ .

La variabilité du modèle peut être décomposée par variable de 2 façons:

- en calculant la variabilité expliquée par chaque variable les unes après les autres (pb : la variabilité d'une variable dépend de l'ordre d'introduction des variables)
- en calculant la variabilité exclusivement par une variable (pb : la somme des variabilités de toutes les variables n'est pas égale à la variabilité du modèle)

Dans certains cas (données équilibrées), la variabilité du modèle se décompose parfaitement et ces 2 calculs donnent les mêmes résultats.

# Exemple sur ozone

```
1 library(FactoMineR)
2 LinearModel(maxO3~T9+T12+T15+Ne9+Ne12+Ne15+Vx9+Vx12+Vx15+maxO3v+vent+pluie, data =ozone)
```

```
Call:
LinearModel(formula = maxO3 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +
    Vx9 + Vx12 + Vx15 + maxO3v + vent + pluie, data = ozone)
```

```
Residual standard error: 14.51 on 97 degrees of freedom
Multiple R-squared:  0.7686
F-statistic: 23.01 on 14 and 97 DF,  p-value: 8.744e-25
AIC =   613      BIC = 653.8
```

```
Ftest
      SS df   MS F value   Pr(>F)
T9       0.2  1   0.2  0.0011 0.97325
T12     376.0  1 376.0  1.7868 0.18445
T15      30.3  1   30.3  0.1439 0.70526
Ne9     1016.5  1 1016.5  4.8312 0.03033
Ne12     37.9  1   37.9  0.1803 0.67208
Ne15      0.1  1   0.1  0.0003 0.98680
Vx9      50.2  1   50.2  0.2388 0.62619
Vx12     35.7  1   35.7  0.1697 0.68127
Vx15     122.6  1 122.6  0.5826 0.44715
maxO3v   5560.4  1 5560.4 26.4261 1.421e-06
vent     297.8  3   99.3  0.4718 0.70267
pluie    182.9  1 182.9  0.8694 0.35344
Residuals 20410.2 97  210.4
```

```
Ttest
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.762962 15.461879 1.3428 0.18245
T9          0.039170  1.164957 0.0336 0.97325
T12         1.972574  1.475705 1.3367 0.18445
T15         0.450308  1.187073 0.3793 0.70526
Ne9        -2.109755  0.959855 -2.1980 0.03033
```

```
1 # LinearModel(maxO3 ~ . , data= ozone) ## Ecriture simplifiée
```

## Test de l'effet d'une ou plusieurs variables

L'ensemble de variables  $\mathcal{V}$  apporte-t-il des informations complémentaires intéressantes sachant que les autres variables sont déjà dans le modèle ?

**Hypothèses :**  $H_0$ : "tous les coefficients associés aux variables de  $\mathcal{V}$  sont égaux à 0" contre  $H_1$  : "au moins un coefficient des variables  $\mathcal{V} \neq 0$

**Statistique de test:**  $F_{obs} = \frac{SC_{\mathcal{V}}/ddl_{\mathcal{V}}}{SC_R/ddl_R} = \frac{CM_{\mathcal{V}}}{CM_R}$

**Loi de la statistique de test:** Sous  $H_0$ ,  $\mathcal{L}(F_{obs}) = \mathcal{F}_{ddl_R}^{ddl_{\mathcal{V}}}$

**Décision :**  $\mathbb{P}(\mathcal{F}_{ddl_{\mathcal{V}}}^{ddl_R} > F_{obs}) < 0.05 \implies$  Rejet de  $H_0$

- Revient à choisir entre le sous-modèle sans les variables  $\mathcal{V}$  ou le modèle complet
- Si  $\mathcal{V}$  contient tous les effets : revient à tester si  $R^2$  est significativement différent de 0, i.e. si toutes les variables sont inutiles (versus au moins une utile)
- On somme les degrés de liberté associés à l'ensemble  $\mathcal{V}$  sachant qu'1 variable quanti à 1 ddl, 1 variable quali à  $I - 1$  ddl et une interaction a comme ddl le produit des ddl de chaque facteur

# Sélection de variables

Sélection de modèle = trouver un compromis entre un modèle qui s'ajuste bien aux données et qui n'a pas "trop" de paramètres  
(Rappel:  $\$ \ddot{\text{m}}$ )

- Enorme littérature sur le sujet



- Souvent utilisé: sélection du modèle qui minimise l'AIC / BIC: compromis entre maximisation de la vraisemblance  $L$  (à quel point le modèle s'ajuste bien aux données) et le nb de paramètres

$$AIC = 2p - 2\ln(\hat{L})$$

$$BIC = pln(n) - 2\ln(\hat{L})$$

? A votre avis, quelle est la différence entre les 2 critères (📝 !)

## Plusieurs stratégies

- Construction exhaustive de tous les sous-modèles (long et même impossible si trop de variables)
- Méthode descendante (backward): construire le modèle complet; supprimer la variable explicative la moins intéressante et reconstruire le modèle sans cette variable; itérer jusqu'à ce que toutes les variables explicatives soient intéressantes
- Méthode ascendante (forward): partir du modèle avec la variable la plus intéressante; ajouter la variable qui, connaissant les autres variables du modèle, apporte le plus d'information complémentaire; itérer jusqu'à ce qu'aucune variable n'apporte d'information intéressante
- Méthode stepwise: compromis entre les 2 méthodes ci-dessus

# Exemple sur ozone: sélection de variables

```
1 library(FactoMineR)
2 LinearModel(maxO3 ~ ., data = ozone, selection = "bic")

Results for the complete model:
=====
Call:
LinearModel(formula = maxO3 ~ ., data = ozone, selection = "bic")

Residual standard error: 14.51 on 97 degrees of freedom
Multiple R-squared:  0.7686
F-statistic: 23.01 on 14 and 97 DF,  p-value: 8.744e-25
AIC =   613      BIC = 653.8

Results for the model selected by BIC criterion:
=====
Call:
LinearModel(formula = maxO3 ~ T12 + Ne9 + Vx9 + maxO3v, data = ozone,
selection = "bic")

Residual standard error: 14 on 107 degrees of freedom
Multiple R-squared:  0.7622
F-statistic: 85.75 on 4 and 107 DF,  p-value: 1.763e-32
AIC =   596      BIC = 609.6

Ftest
      SS  df  MS F value    Pr(>F)
T12     6650.4  1 6650.4 33.9334 6.073e-08
Ne9     2714.8  1 2714.8 13.8522 0.0003172
Vx9     903.4   1 903.4  4.6094 0.0340547
maxO3v  7363.5  1 7363.5 37.5721 1.499e-08
Residuals 20970.2 107 196.0

Ttest
      Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.631310 11.000877 1.1482 0.2534427
T12         2.764090  0.474502 5.8252 6.073e-08
```

# Démarche en modélisation

1. Lister les variables potentiellement explicatives / prédictives
2. Visualiser
3. Sélectionner le sous-modèle (minimisation de l'AIC ou du BIC)
4. Interpréter les résultats (quelles variables ressortent ? Est-ce surprenant ? Est-ce en accord avec les connaissances sur le sujet ? Des confusions possibles ?)
5. Interpréter les coefficients (signe, valeur etc.)
6. (Prédire pour de nouvelles valeurs)

# Codage et contraintes pour variables quali

La matrice de design  $X$  (dans  $Y = X\beta + E$ ) a autant de lignes que d'individus. Pour les colonnes, c'est variable...

|                          | Constante ( $\mu$ ou $\beta_0$ ) | Variable quanti       | Variable Quali  | Interaction quali -quali   |
|--------------------------|----------------------------------|-----------------------|---|--|
| Représentation dans $X$  | Une colonne de 1                 | Valeur de la variable | (I-1) colonnes par modalités de chaque variable à I modalités                     | (I-1)*(J-1) colonnes   |
| Degrés de libertés (ddl) | 1                                | 1                     | I-1   | (I-1)*(J-1)  |
| Commentaires             |                                  |                       | Contrainte à poser sur les paramètres (le mieux est $\sum_{i=1}^n \alpha_i = 0$ ) | Contrainte à poser sur les paramètres (le mieux est $\forall i, \sum_j \alpha \beta_{ij} = 0$ et $\forall j, \sum_i \alpha \beta_{ij} = 0$ ) |

## Important

Le choix de la contrainte impacte FORTEMENT l'interprétation

1.  $\sum_i \alpha_i = 0$ , la comparaison est faite par rapport à la moyenne des moyennes par modalité

2.  $\alpha_1 = 0$ , la comparaison est faite par rapport à un niveau de ref (*moins intuitif, et pas pratique quand présence d'interactions*)

Certaines fonctions de R par défaut utilisent la contrainte 2. !

Comparez les sorties de ces différentes lignes de code...

```
1 #lm par défaut
2 coef( lm(maxO3 ~ pl
3      data= ozon
(Intercept)    pluieSec
73.39535     27.44523
```

```
1 #lm en spécifiant le contraste
2 coef(lm(maxO3 ~ pluie,
3         data = ozone,
4         contrasts = list(pluie= "contr.sum")))
(Intercept)    pluie1
87.11796     -13.72262
```

```
1 # LinearModel prémâche le travaille
2 LinearModel(maxO3 ~ pluie,
3             data= ozone)$Ttest
Estimate Std. Error   t value   Pr(>|t|)
(Intercept)  87.11796  2.419555 36.005777 1.066479e-62
pluie - Pluie -13.72262  2.419555 -5.671545 1.156980e-07
pluie - Sec   13.72262  2.419555  5.671545 1.156980e-07
```

# Interprétation des résultats

- Modèle sélectionné  $\Rightarrow$  Interprétation des résultats

*2 situations*

1. Idéale: données équilibrées
2. Difficile: données déséquilibrées

Remarques:

- Les plans d'expériences permettent d'avoir données équilibrées
- En ANOVA, on a souvent des données équilibrées

## Décomposition de la variabilité : variables quali, données équilibrées

Quand les données sont équilibrées, la décomposition de la variabilité est parfaite:

$$\begin{aligned} \sum_{i,j,k} (y_{ijk} - y_{\dots\dots})^2 &= \sum_{i,j,k} \underbrace{(y_{i\dots\dots} - y_{\dots\dots})^2}_{\hat{\alpha}_i^2} + \sum_{i,j,k} \underbrace{(y_{\dots j\dots} - y_{\dots\dots})^2}_{\hat{\beta}_j^2} \\ &\quad + \sum_{i,j,k} \underbrace{(y_{ij\dots} - y_{i\dots\dots} - y_{\dots j\dots} + y_{\dots\dots})^2}_{\widehat{\alpha\beta}_{ij}^2} + \sum_{i,j,k} \underbrace{(y_{ijk} - y_{ij\dots})^2}_{\varepsilon_{ijk}^2} \end{aligned}$$

Quand les données sont équilibrées, les coefficients s'estiment simplement:

$$\begin{aligned} \hat{\mu} &= y_{\dots\dots} & \forall i, \hat{\alpha}_i &= y_{i\dots\dots} - y_{\dots\dots} \\ \forall j, \hat{\beta}_j &= y_{\dots j\dots} - y_{\dots\dots} & \forall i, j \quad \widehat{\alpha\beta}_{ij} &= y_{ij\dots} - y_{i\dots\dots} - y_{\dots j\dots} + y_{\dots\dots} \end{aligned}$$

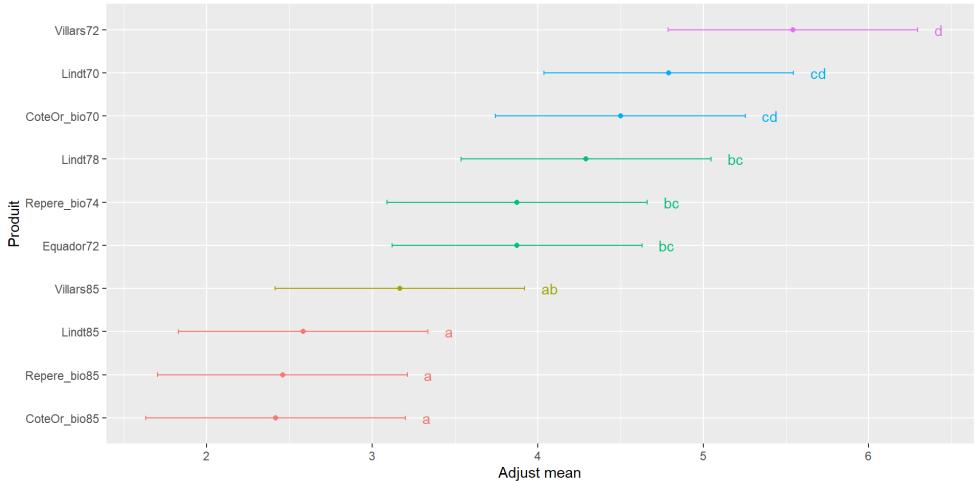
⇒ On quantifie parfaitement ce qui est expliqué par chaque variable ou interaction

# Comparaison de moyennes ajustées

Moyennes ajustées =  $\hat{\mu} + \hat{\alpha}_i$ : Permet de s'affranchir de l'effet des autres variables

Comparaisons possibles 2 à 2, moyennant une correction de Bonferroni par ex.

```
1 data_choco <- read.csv("data/chocolat_2022.csv", sep = ";")
2 mod <- LinearModel(Sucree~Produit+Juge+Produit:Juge, data=data_choco)
3 meansComp(mod, ~Produit, adjust="bonferroni")
```



```
$adjMean
Produit      emmean     SE  df lower.CL upper.CL
CoteOr_bio70  4.50 0.264 118    3.75   5.25
CoteOr_bio85  2.42 0.275 118    1.63   3.20
Ecuador72    3.88 0.264 118    3.12   4.63
Lindt70       4.79 0.264 118    4.04   5.55
Lindt78       4.29 0.264 118    3.54   5.05
Lindt85       2.58 0.264 118    1.83   3.34
Repere_bio74  3.88 0.275 118    3.09   4.66
Repere_bio85  2.46 0.264 118    1.70   3.21
Villars72     5.54 0.264 118    4.79   6.30
Villars85     3.17 0.264 118    2.41   3.92

Results are averaged over the levels of: Juge
Confidence level used: 0.95
Conf-level adjustment: bonferroni method for 10 estimates

$groupComp
CoteOr_bio85 Repere_bio85      Lindt85      Villars85      Ecuador72 Repere_bio74
          "a"        "a"        "a"        "ab"        "bc"        "bc"
          Lindt78 CoteOr_bio70      Lindt70      Villars72      "cd"        "d"

attr(),"class")
[1] "meansComp"
```



# Prédictions

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip}$$

Prédire Y pour : (T12=19, Ne9=8, Vx9=1.2, maxO3v=70) et (T12=23, Ne9=10, Vx9=0.9, maxO3v=95)

Sur PC:

```
1 model <- LinearModel(maxO3 ~ ., data = ozone, selection = "bic")
2
3 xnew <- data.frame(T12=c(19,23), Ne9=c(8,10), Vx9=c(1.2,0.9), maxO3v=c(70,95))
4 predict(model,xnew,interval="pred")
```

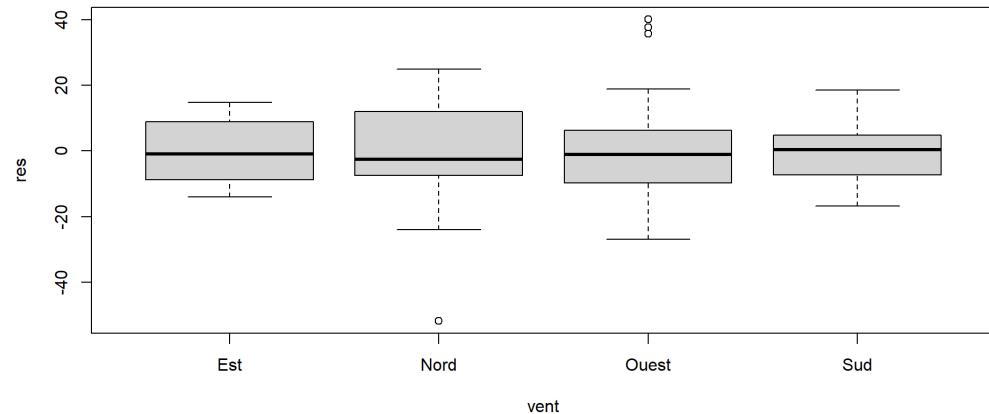
|   | fit      | lwr      | upr       |
|---|----------|----------|-----------|
| 1 | 71.41544 | 42.90026 | 99.93063  |
| 2 | 85.92393 | 56.76803 | 115.07983 |

```
1 predict(model,xnew,interval="confidence")
```

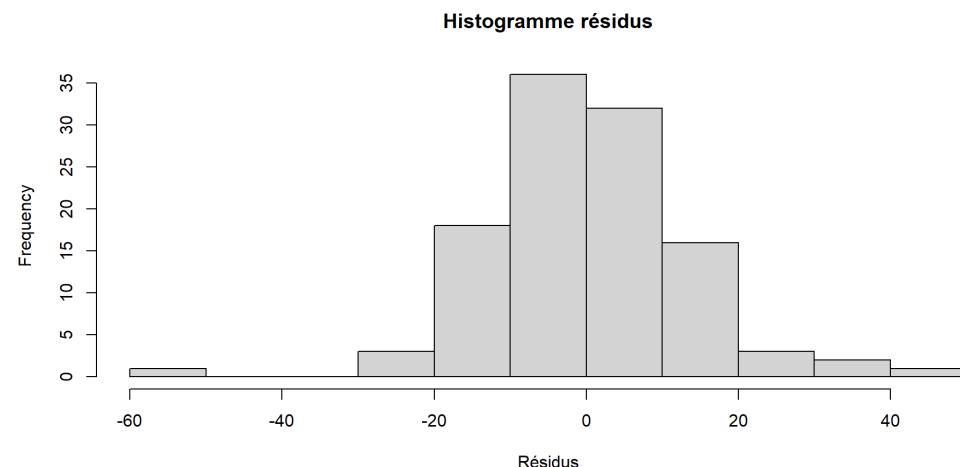
|   | fit      | lwr      | upr      |
|---|----------|----------|----------|
| 1 | 71.41544 | 64.86327 | 77.96761 |
| 2 | 85.92393 | 76.98627 | 94.86159 |

# Analyse des résidus

- Les résidus sont la part du phénomène non-expliquée par le modèle
- On peut vérifier les hypothèses initiales du modèle



- Test de Shapiro-Wilk de normalité des Résidus (Rq : la non-normalité n'est pas un pb tant que la distribution est symétrique)



Shapiro-Wilk normality test

```
data: res  
W = 0.97117, p-value = 0.01587
```

## Retour sur l'exemple ozone

- Maximum d'ozone : variable réponse
- Les variables de températures, nébulosité, vitesse de vent (quanti), et la direction et la pluie (quali) sont prises en compte
- On ne sait pas si les interactions entre variables quali et entre les variables quali et quanti sont négligeables  $\Rightarrow$  on les met dans le modèle

```
LinearModel(max03 ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 + max03v + pluie + vent) *  
(pluie + vent), data=ozone, selection="bic")
```

Ce qui revient à écrire :

```
max03 ~ T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 + max03v + pluie + vent + T9:pluie +  
T12:pluie + T15:pluie + Ne9:pluie + Ne12:pluie + Ne15:pluie + Vx9:pluie + Vx12:pluie + Vx15:pluie +  
max03v:pluie + vent:pluie + T9:vent + T12:vent + T15:vent + Ne9:vent + Ne12:vent + Ne15:vent + Vx9:vent +  
Vx12:vent + Vx15:vent + max03v:vent
```

## Résultat et sortie

Results for the complete model:

```
Call: LinearModel(formula = max03 ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 + max03v + pluie + vent) * (pluie + vent), data = ozone, selection = "bic")
```

Residual standard error: 14.54 on 56 degrees of freedom  
Multiple R-squared: 0.8658 F-statistic: 6.569 on 55 and 56 DF, p-value: 2.585e-11  
AIC = 634 BIC = 786.2

Results for the model selected by BIC criterion:

```
Call: LinearModel(formula = max03 ~ T9 + T15 + Ne12 + Vx9 + max03v + vent + T9:vent + T15:vent, data = ozone, selection = "bic")
```

Residual standard error: 13.8 on 97 degrees of freedom  
Multiple R-squared: 0.7906 F-statistic: 26.16 on 14 and 97 DF, p-value: 8.082e-27  
AIC = 601.8 BIC = 642.6

# Interprétation des résultats

|                | SS      | df    | MS     | F value | Pr(>F)        |     |      |     |     |     |   |
|----------------|---------|-------|--------|---------|---------------|-----|------|-----|-----|-----|---|
| T9             | 606.2   | 1     | 606.2  | 3.1841  | 0.0774866 .   |     |      |     |     |     |   |
| T15            | 252.8   | 1     | 252.8  | 1.3278  | 0.2520220     |     |      |     |     |     |   |
| Ne12           | 2107.0  | 1     | 2107.0 | 11.0664 | 0.0012425 **  |     |      |     |     |     |   |
| Vx9            | 1657.3  | 1     | 1657.3 | 8.7046  | 0.0039790 **  |     |      |     |     |     |   |
| maxO3v         | 5160.8  | 1     | 5160.8 | 27.1060 | 1.078e-06 *** |     |      |     |     |     |   |
| vent           | 12.5    | 3     | 4.2    | 0.0218  | 0.9956079     |     |      |     |     |     |   |
| T9:vent        | 2587.1  | 3     | 862.4  | 4.5294  | 0.0051335 **  |     |      |     |     |     |   |
| T15:vent       | 3722.1  | 3     | 1240.7 | 6.5165  | 0.0004613 *** |     |      |     |     |     |   |
| Residuals      | 18468.3 | 97    | 190.4  |         |               |     |      |     |     |     |   |
| ---            |         |       |        |         |               |     |      |     |     |     |   |
| Signif. codes: | 0       | '***' | 0.001  | '**'    | 0.01          | '*' | 0.05 | '.' | 0.1 | ' ' | 1 |

On peut dire (à partir des effets significatifs) qu'il y a, sur le max d'O3:

- des effets de nébulosité de vitesse de vent, du maximum d'O3 de la veille
- des effets de la direction du vent et des températures mais à travers les interactions : la direction du vent modifie l'effet de la T° (i.e. amplifie l'effet de la T° ou la diminue selon la direction du vent) sur max O3

On peut aussi dire (à partir des absences d'effets significatifs):

- la pluviométrie n'est pas un facteur déterminant qui influe sur le max d'03
- une seule nébulosité (Ne12) est conservée dans le modèle : cela ne veut pas dire que les autres nébulosités n'ont pas d'effet (elles peuvent avoir un effet similaire). Idem pour la vitesse de vent.
- pour la T°, on a besoin des T° à 9h et à 15h pour mieux prévoir le maximum d'ozone. L'effet de la T° n'est pas exactement le même entre 9h et 15h (mais 12h n'est pas utile comme info)
- l'effet de la nébulosité est le même quelle que soit la direction du vent. Idem pour la vitesse du vent.

## Interprétation des résultats (2/2)

|                    | Estimate   | Std. Error  | t value     | Pr(> t )     |
|--------------------|------------|-------------|-------------|--------------|
| (Intercept)        | 18.2063080 | 13.70594596 | 1.32835107  | 1.871790e-01 |
| T9                 | 1.8269164  | 1.02383156  | 1.78439159  | 7.748658e-02 |
| T15                | 1.0200174  | 0.88518659  | 1.15231903  | 2.520220e-01 |
| Ne12               | -3.0131694 | 0.90577505  | -3.32661998 | 1.242541e-03 |
| Vx9                | 2.2212608  | 0.75287821  | 2.95035871  | 3.978950e-03 |
| maxO3v             | 0.3466961  | 0.06659106  | 5.20634637  | 1.078363e-06 |
| vent - Est         | 0.4734242  | 17.01021918 | 0.02783175  | 9.778535e-01 |
| vent - Nord        | 3.2973579  | 15.03561957 | 0.21930310  | 8.268748e-01 |
| vent - Ouest       | -1.1250216 | 14.00661570 | -0.08032073 | 9.361477e-01 |
| vent - Sud         | -2.6457606 | 15.14886158 | -0.17465079 | 8.617181e-01 |
| vent - Est : T9    | -4.7442269 | 2.14177382  | -2.21509241 | 2.909369e-02 |
| vent - Nord : T9   | 6.1407678  | 1.70138426  | 3.60927739  | 4.879524e-04 |
| vent - Ouest : T9  | -0.9505390 | 1.34553370  | -0.70644010 | 4.816079e-01 |
| vent - Sud : T9    | -0.4460019 | 1.52215003  | -0.29300782 | 7.701421e-01 |
| vent - Est : T15   | 3.6809644  | 1.79674431  | 2.04868573  | 4.319389e-02 |
| vent - Nord : T15  | -5.2279430 | 1.21533621  | -4.30164338 | 4.045444e-05 |
| vent - Ouest : T15 | 0.9796886  | 0.93082643  | 1.05249333  | 2.951881e-01 |
| vent - Sud : T15   | 0.5672899  | 1.11114058  | 0.51054741  | 6.108279e-01 |

- pour T9 et T15,  $\beta \neq 0$
- + nébulosité, — max d'O3 ()
- + maximum d'O3 de la veille est grand, + maximum d'O3 du jour
- les jours de vents d'est et surtout du nord ont des max d'O3 supérieur aux jours de vents d'ouest et du sud
- les jours de vent du nord, l'effet de la  $T^o$  à 9h est plus important (coef = 6.14); au contraire quand le vent vient de l'est. C'est l'inverse pour la  $T^o$  à 15h
- Ainsi, quand le vent du nord, l'effet de la  $T^o$  à 9h est particulièrement fort (donc s'il fait chaud à 9h quand le vent vient du nord, le max d'O3 risque d'être important)
- etc.

## Démarche en modélisation (bis)

1. Lister les variables potentiellement explicatives / prédictives
2. Visualiser
3. Sélectionner le sous-modèle (minimisation de l'AIC ou du BIC)
4. Interpréter les résultats (quelles variables ressortent ? Est-ce surprenant ? Est-ce en accord avec les connaissances sur le sujet ? Des confusions possibles ?)
5. Interpréter les coefficients (signe, valeur etc.)
6. Prédire pour de nouvelles valeurs

# Travaux dirigés

# Liste de messages d'erreurs communs

- Erreur dans `file(file, "rt")` : impossible d'ouvrir la connexion:

- Nom du fichier mal spécifié
- Chemin mal spécifié (vérifiez le projet dans lequel vous êtes en haut à droite)

```
1 read.table("https://r-stat-sc-donnees.github.io/ozone.tx",
2             header=TRUE, stringsAsFactors = TRUE)
```

Erreur dans `file(file, "rt")` : impossible d'ouvrir la connexion vers '<https://r-stat-sc-donnees.github.io/ozone.tx>'

- objet xxx introuvable ⇒ objet non chargé (vérifiez votre environnement)

```
1 head(ozone, n= 1)
2 ozone <- read.table("https://r-stat-sc-donnees.github.io/ozone.txt",
3                      header=TRUE, stringsAsFactors = TRUE)
4 head(ozone, n= 1)
```

- impossible de trouver la fonction xxx ⇒ Package non chargé / fonction mal orthographiée (R esT SeNSiBle A La CAsSe !)

```
1 LinearModel(maxO3 ~pluie, data = ozone) # package pas chargé
Error in LinearModel(maxO3 ~ pluie, data = ozone): impossible de trouver la fonction "LinearModel"

1 library(FactoMineR)
2 Linearmodel(maxO3 ~pluie, data = ozone) # oubli majuscule à model
Error in Linearmodel(maxO3 ~ pluie, data = ozone): impossible de trouver la fonction "Linearmodel"
1 #LinearModel(maxO3 ~pluie, data = ozone) # marche
```

- Pas de message mais un + qui apparaît dans la console ⇒ Vous n'avez pas fermé chaque parenthèse ouverte

- Appuyer sur Echap après s'être mis dans la console
- Corriger code

```
1 # (mean(c(1:3))
2
3 # Ceci s'affiche dans la console
4 # > (mean(c(1:3))
5 #+
6
7 # marche
8 mean(c(1:3))
```

[1] 2

- Souvenez-vous...

