

MAT4681 - Statistique pour les sciences

Arthur Charpentier

01 - Données, variables

été 2022

Données et variables aléatoires

En majuscules, X, X_1, \dots, X_n, Y sont des variables aléatoires

En minuscules, x, x_1, \dots, x_n, y sont des observations

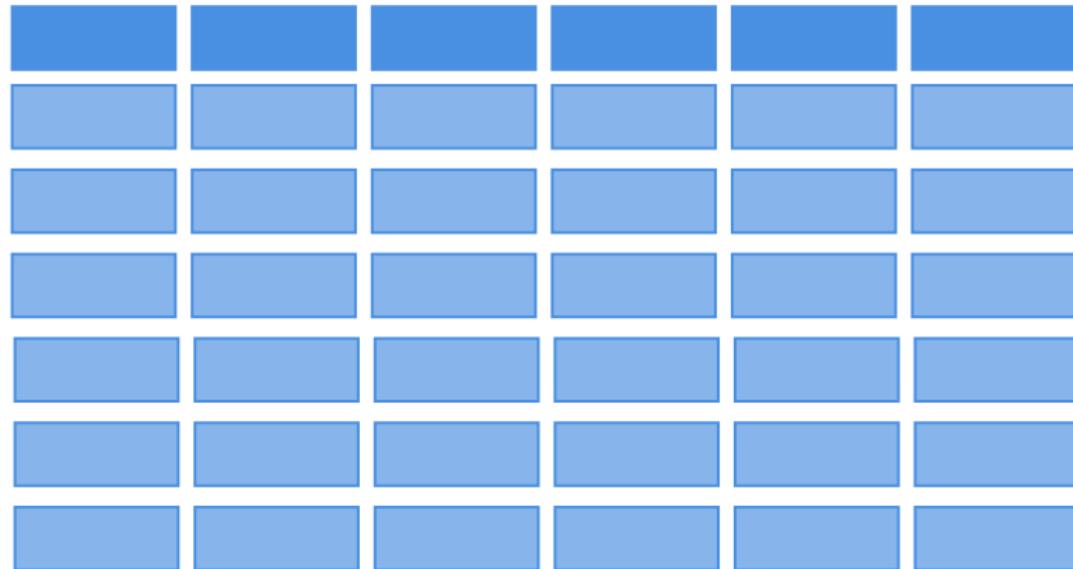
Example On lance 10 fois une pièce, X_3 est l'indicatrice de le 3-ème lancé ait été 'pile'. Plus généralement, X_i est associée au i -ème lancer $X_i \in \{0, 1\}$

Étant donné un échantillon $\{1, 1, 1, 0, 0, 0, 0, 1, 0\}$, $x_3 = 1$.

Définitions

- ▶ **Individus** : objets décrits par un ensemble de données. Un individu peut être une personne, un thermomètre, un pays e.g. Les individus sont notées génériquement i
- ▶ **Variable** : certaine caractéristique d'un individu. Elle prend potentiellement différentes valeurs pour différents individus. Le genre, l'âge, la taille, la température, le revenu médian des individus sont des variables. Les variables sont notées génériquement Y , X ou y , x
- ▶ **Échantillon** : sous-ensemble (de taille n) de la population.
- ▶ **Échantillon aléatoire** : échantillon pigé au hasard dans la population (souvent de telle sorte que tous les éléments ont la même chance d'être pigés).
- ▶ **Base de données**: “matrice” représentant les individus en ligne (i) et les variables en colonne (j), pour des données appareillées

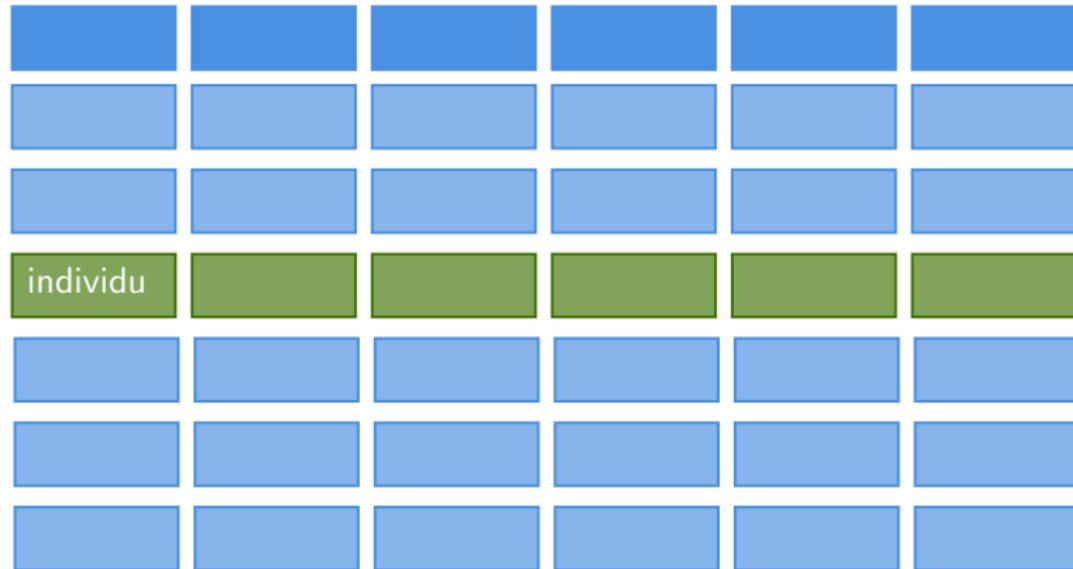
Définitions



Définitions

	variable				

Définitions



Définitions

- ▶ **Série temporelle** : séquence de variables observées à des dates régulièrement espacées dans le temps (quotidienne, hebdomadaire, mensuelle, etc)
- ▶ **Échantillons indépendants** : on obtient deux échantillons à deux dates ou deux endroits différents pour une même variable $\{y_1^{(1)}, \dots, y_{n_1}^{(1)}\}$ et $\{y_1^{(2)}, \dots, y_{n_2}^{(2)}\}$
- ▶ **Échantillons appareillés** : on obtient deux échantillons, pour deux variables, mais les mêmes individus $\{x_1, \dots, x_n\}$ et $\{y_1, \dots, y_n\}$, aussi noté $\{(x_1, y_1), \dots, (x_n, y_n)\}$

Time Series

A **time series** is a sequence of observations (Y_t) ordered in time, at regular dates.

Buy-Ballot (1847, *Les changements périodiques de température, dépendants de la nature du soleil et de la lune, mis en rapport avec le pronostic du temps, déduits d'observations néerlandaises de 1729 à 1846*) - original probably in Dutch.

TABLEAU REPRÉSENTANT LA MARCHE DE LA TEMPÉRATURE PENDANT L'ANNÉE.

Date.	Temp.	Temp. calculée.	Diffr.	Temp. calculée.	Diffr.	Date.	Temp.	Temp. calculée.	Diffr.	Temp. calculée.	Diffr.
10 Janv.	32.55	32.5 8	0	32.55	0	17 Juill.	65.55	65.4 2	-0.15	65.3 6	+ 0.05
19 «	+ 1.12 + 6.7 8	+ 0.23 + 6.8 8	+ 0.12	+ 0.23 + 6.8 8	+ 0.12	20 «	64.30	64.30	+ 0.25	64.63	+ 0.35
20 «	0.69 34.14	+ 0.35 34.34	+ 0.05	+ 0.35 34.34	+ 0.05	27 «	64.57	64.57	+ 0.01	64.37	+ 0.21
25 «	0.82 34.92	+ 0.30 35.22	+ 0.05	+ 0.30 35.22	+ 0.05	1 Août	64.83	64.83	+ 0.02	64.71	+ 0.13
30 «	1.26 35.70	+ 0.78 + 5.5 9	+ 0.67	+ 0.78 + 5.5 9	+ 0.67	6 «	62.06	62.06	0	65.54	0
4 Févr.	- 0.01 36.47	0	36.40	+ 0.07							
9 «	+ 0.39 + 6.5 5	- 0.16 36.99	- 0.13	- 0.16 36.99	- 0.13	11 «	64.45	64.4 2	- 0.05	64.5 0	+ 0.05
14 «	+ 0.39 37.36	+ 0.39 37.36	0	+ 0.39 37.36	0	21 «	64.22	64.22	+ 0.05	64.66	+ 0.40
19 «	+ 0.05 38.13	+ 0.06 38.17		+ 0.06 38.17		26 «	63.80	63.80	+ 0.05	63.55	+ 0.30
24 «	+ 1.53 38.68	+ 0.08 38.76	0	+ 0.08 38.76	0	26 «	63.38	63.38	+ 0.10	63.04	+ 0.15
1 Mars	0.48 39.24	0	+ 0.6 6.9	+ 0.25	1 Sept.	62.96	62.96	0	62.54	+ 0.42	
6 «	0.25 + 6.7 3	- 0.48 40.14	- 0.65	- 0.48 40.14	- 0.65	6 «	62.03	62.03	0		
11 «	0.11 40.69	- 1.10 40.93	- 1.23	- 1.11 40.93	- 1.23	11 «	60.71	60.71	+ 0.05	61.2 6	+ 0.51
17 «	1.83 41.45	+ 0.41 41.52	+ 0.09	+ 0.41 41.52	+ 0.09	16 «	59.59	59.59	+ 0.45	59.21	+ 0.21
22 «	0.93 41.45	+ 0.07 41.52	+ 0.05	+ 0.07 41.52	+ 0.05	22 «	57.88	57.88	+ 0.45	58.25	+ 0.41
27 «	0.67 42.89	+ 0.57 + 1.2 9	- 0.62	+ 0.57 + 1.2 9	- 0.62	26 «	56.33	56.33	0	56.89	+ 0.34
1 Avril	1.38 + 1.3 6	- 0.02 44.80	- 0.57	- 0.02 44.80	- 0.57	1 Oct.	55.74	55.74	+ 0.27	55.74	
6 «	1.90 45.61	+ 0.52 46.09	+ 0.64	+ 0.52 46.09	+ 0.64	6 «	54.69	54.69	- 0.03	1.4 3	+ 0.35
11 «	1.23 46.59	+ 0.37 47.38	- 0.02	+ 0.37 47.38	- 0.02	11 «	53.37	53.37	- 0.20	52.88	+ 0.29
16 «	0.39 48.32	- 0.11 48.67	- 0.42	- 0.11 48.67	- 0.42	16 «	52.05	52.05	- 0.12	51.46	+ 0.48
21 «	1.41 49.40	+ 0.49 49.96	- 0.26	+ 0.49 49.96	- 0.26	21 «	50.73	50.73	- 0.12	50.50	+ 0.70
26 «	1.56 50.1 1	+ 0.25 50.56	- 0.26	+ 0.25 50.56	- 0.26	26 «	50.63	50.63	- 0.50	48.60	
2 Mai	1.56 50.13	+ 0.69 + 1.1 2	- 0.44	+ 0.69 + 1.1 2	- 0.44	1 Nov.	47.44	47.44	- 0.49	1.4 1	- 0.24
7 «	1.50 53.34	+ 0.08 53.50	+ 0.02	+ 0.08 53.50	+ 0.02	6 «	45.82	45.82	- 0.58	45.77	- 0.53
12 «	0.29 54.56	- 0.01 54.62	- 0.07	- 0.01 54.62	- 0.07	11 «	44.20	44.20	0	44.36	+ 0.16
17 «	1.19 55.79	0	+ 0.95	+ 0.95	+ 0.95	21 «	43.76	43.76	+ 0.19	43.56	+ 0.30
22 «	1.28 + 0.9 5	+ 0.34 56.07	+ 0.20	+ 0.34 56.07	+ 0.20	21 «	42.68	42.68	- 1.15	41.52	0
27 «	0.85 57.06	+ 0.20 57.06	- 0.04	+ 0.20 57.06	- 0.04	21 «	42.02	42.02	- 0.97	41.70	+ 0.12
1 Juin	0.34 58.04	- 0.04 58.06	- 0.44	- 0.04 58.06	- 0.44	1 Déc.	41.46	41.46	0	40.13	+ 0.76
6 «	1.37 59.60	+ 0.02 59.32	+ 0.30	+ 0.02 59.32	+ 0.30	6 «	40.85	40.85	- 0.42	39.43	+ 0.37
11 «	0.93 60.55	0	+ 59.98	+ 0.57	+ 59.98	11 «	39.19	39.19	- 0.40	38.73	+ 0.06
16 «	0.57 + 6.5 5	+ 0.02 60.65	+ 0.47	+ 0.02 60.65	+ 0.47	10 «	38.33	38.33	- 0.29	38.04	0
21 «	0.06 61.63	- 0.47 61.31	- 0.13	- 0.06 61.63	- 0.13	21 «	37.49	37.49	0	37.19	+ 0.40
26 «	0.77 62.30	- 0.25 61.97	- 0.02	- 0.25 61.97	- 0.02	20 «	36.74	36.74	- 0.47	35.99	- 0.17
2 Juill.	0.68 63.57	- 0.12 62.63	- 0.51	- 0.68 63.57	- 0.51	5 «	35.12	35.12	- 0.35	34.19	- 0.58
7 «	0.68 63.31	+ 0.02 + 0.3 32	+ 0.35	+ 0.68 63.31	+ 0.35	5 Juin	33.98	33.98	- 0.65	33.58	- 0.58
12 «	0.53 63.86	0	62.33	+ 0.53	62.33	10 «	32.59	32.59	0	32.58	

(rapide) Typologie

- ▶ Variable **catégorielle** (facteur): les individus sont partitionnés entre plusieurs groupes (en prenant une modalité, et une seule)
 - ▶ catégorielle **nominale** e.g.
genre $\in \{\text{homme, femme}\}$ ou
 - ▶ catégorielle **ordinale** e.g.
revenu $\in \{[0, 50], [50 - 100], [100, 200], [200+]\}$
- ▶ Variable **quantitative**
 - ▶ quantitative **continue** e.g.
taille, revenu, température, superficie
 - ▶ quantitative **discrete** e.g.
nombre d'enfants, étage

Visualiser

taille d'élèves dans un groupe de 200 personnes,

```
1 > str(Davis)
2 'data.frame': 200 obs. of 5 variables:
3 $ sex      : Factor w/ 2 levels "F","M": 2 1 1 2 ...
4 $ weight   : int 77 58 53 68 59 76 76 69 ...
5 $ height   : int 182 161 161 177 157 170 ...
6 > summary(Davis)
7 sex           weight           height
8 F:112    Min.   : 39.00   Min.   :148.0
9 M: 88    1st Qu.: 55.00   1st Qu.:164.0
10          Median : 63.00   Median :169.5
11          Mean   : 65.25   Mean   :170.6
12          3rd Qu.: 73.25   3rd Qu.:177.2
13          Max.   :119.00   Max.   :197.0
```

- ▶ sex: genre de la personne, **catégorielle nominale**
- ▶ weight: poids de la personne (kg), **quantitative continue**
- ▶ height: taille de la personne (cm), **quantitative continue**

Visualiser

```
1 > head(Davis$height, 12)
2 [1] 182 161 161 177 157 170 167 186 178 171 175 166
```

Parfois, il sera intéressant de **trier** les données,

```
1 > head(sort(Davis$height), 12)
2 [1] 148 150 152 153 154 155 156 157 157 157 157 157
```

On peut avoir quelques statistiques de base

```
1 > summary(Davis$height)
2   Min. 1st Qu. Median      Mean 3rd Qu.      Max.
3 148.0    164.0   169.5    170.6   177.2    197.0
```

Visualiser

Pour les variables catégorielles,

```
1 > mean(Davis$sex)
2 [1] NA
3 Warning message:
4 In mean.default(Davis$sex) :
5   argument is not numeric or logical: returning NA
6 > table(Davis$sex)
7
8   F     M
9 112   88
```

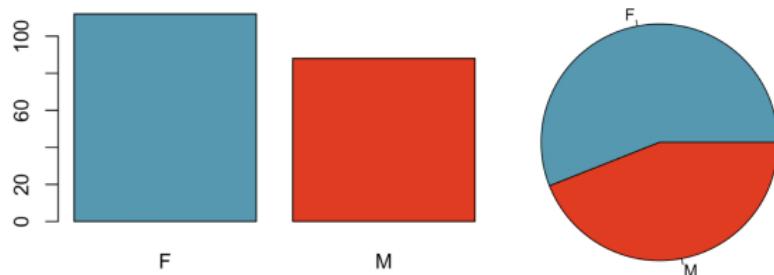
Le genre est nominal, mais on peut avoir des variables ordinaires

```
1 > height_class = cut(Davis$height, breaks = seq
2   (140,200,by=10))
3 > str(height_class)
4 Factor w/ 6 levels "(140,150]","(150,160]",...: 5 3...
5 > table(height_class)
6 height_class
7 (140,150] (150,160] (160,170] (170,180] (180,190]
8               2          20         86         63         26
```

Visualiser

Pour les variables catégorielles,

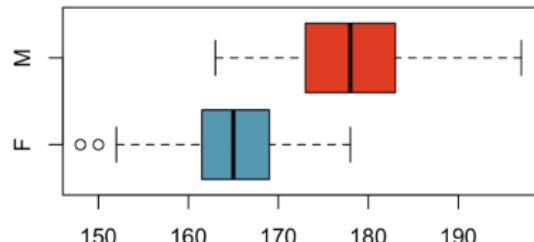
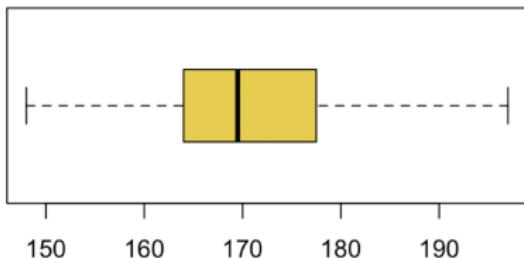
```
1 > barplot(table(Davis$sex))  
2 > pie(table(Davis$sex))
```



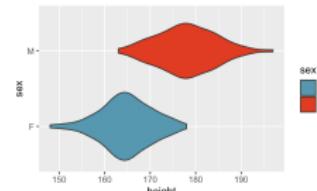
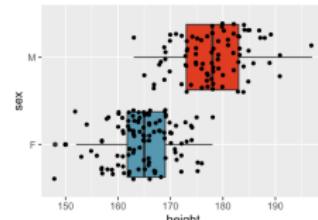
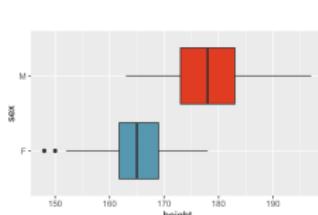
Visualiser

Pour les variables continues,

```
1 > boxplot(Davis$height, horizontal = TRUE)
2 > boxplot(Davis$height~Davis$sex, horizontal = TRUE,)
```



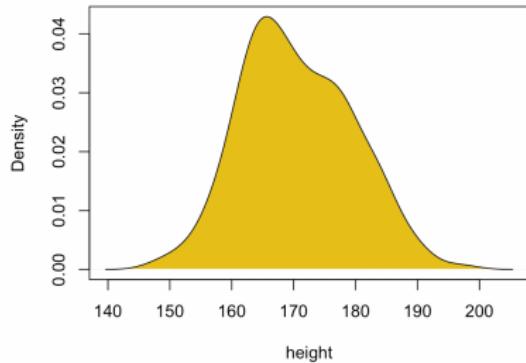
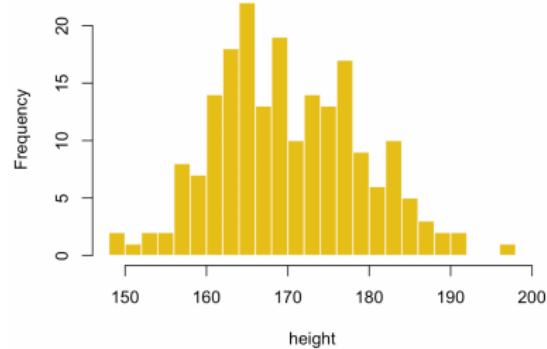
```
1 > library(ggplot2)
2 > ggplot(aes(x=height, y=sex, fill=sex), data=Davis) +
    geom_boxplot()
```



Visualiser

Pour les variables continues

```
1 > hist(Davis$height, breaks = 30)
2 > plot(density(Davis$height))
```

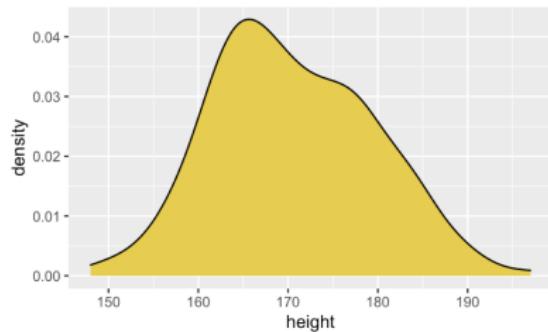
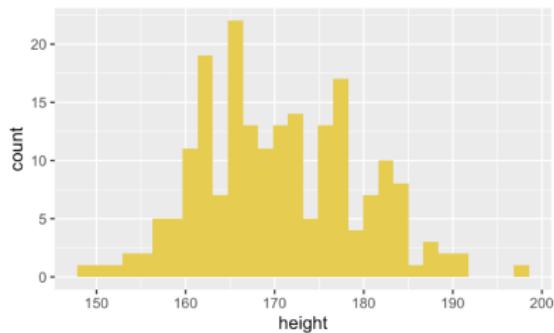


```
1 > mean(Davis$height)
2 [1] 170.565
```

Visualiser

Pour les variables continues

```
1 > ggplot(Davis, aes(x=height)) + geom_histogram()  
2 > ggplot(Davis, aes(x=height)) + geom_density()
```

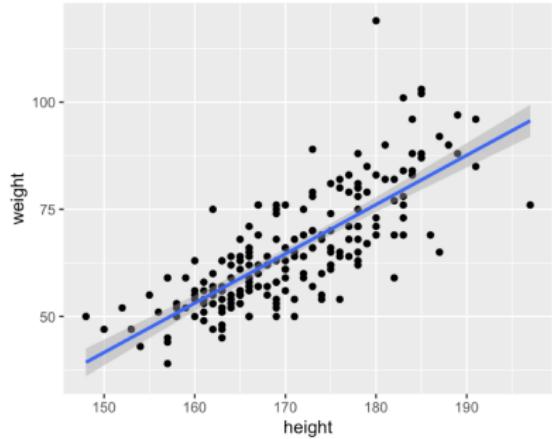
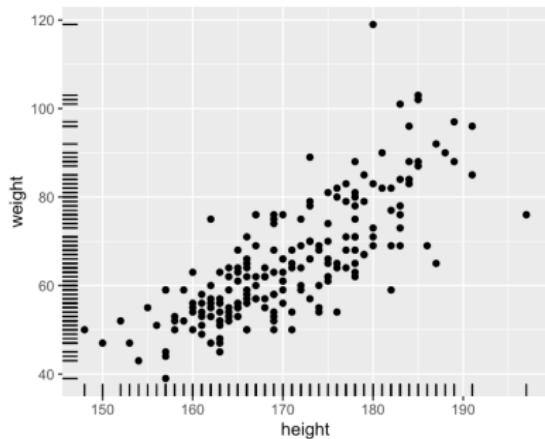


```
1 > mean(Davis$height)  
2 [1] 170.565
```

Visualiser

Pour les variables continues bivariées appareillées

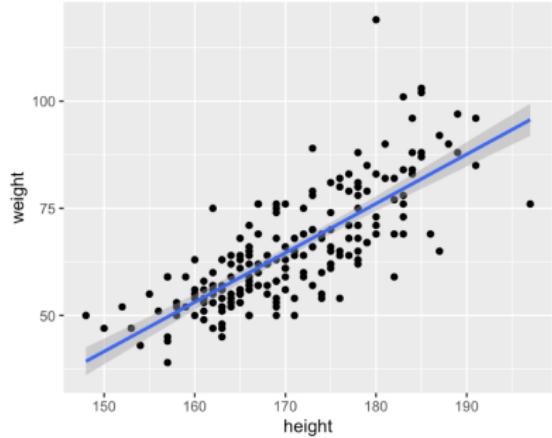
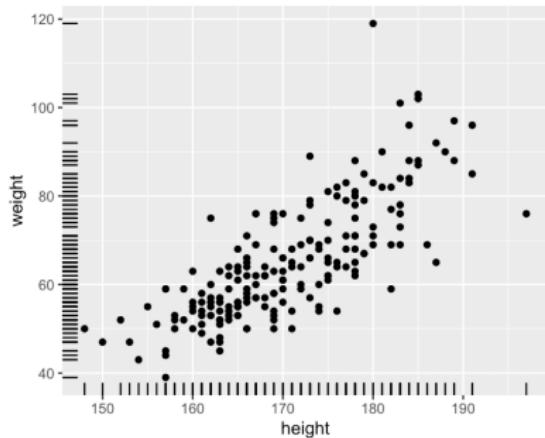
```
1 > ggplot(Davis, aes(x=height, y=weight)) + geom_point()  
2 > ggplot(Davis, aes(x=height, y=weight)) + geom_point()  
   + geom_smooth(method=lm)
```



Visualiser

Pour les variables continues bivariées appareillées

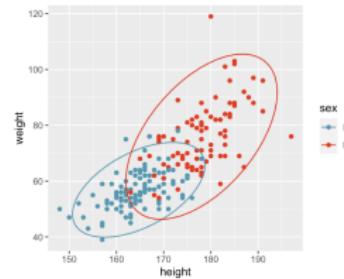
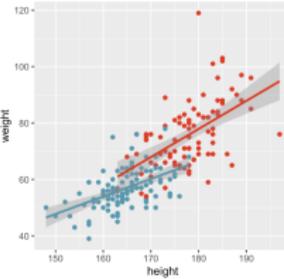
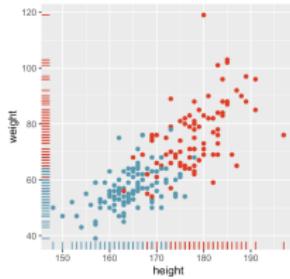
```
1 > ggplot(Davis, aes(x=height, y=weight)) + geom_point()  
2 > ggplot(Davis, aes(x=height, y=weight)) + geom_point()  
   + geom_smooth(method=lm)
```



Visualiser

Pour les variables continues bivariées appareillées

```
1 > ggplot(Davis, aes(x=height, y=weight, color=sex)) +  
   geom_point()  
2 > ggplot(Davis, aes(x=height, y=weight, color=sex)) +  
   geom_point() + geom_smooth(method=lm)  
3 > ggplot(Davis, aes(x=height, y=weight, color=sex)) +  
   geom_point() + stat_ellipse(type = "norm")
```



Values manquantes (NA)

```
1 > summary(Davis)
2   reportedWeight    reportedHeight
3   Min. : 41.00      Min. :148.0
4   1st Qu.: 55.00    1st Qu.:160.5
5   Median : 63.00    Median :168.0
6   Mean   : 65.62    Mean   :168.5
7   3rd Qu.: 73.50    3rd Qu.:175.0
8   Max.   :124.00    Max.   :200.0
9   NA's    :17        NA's    :17
```

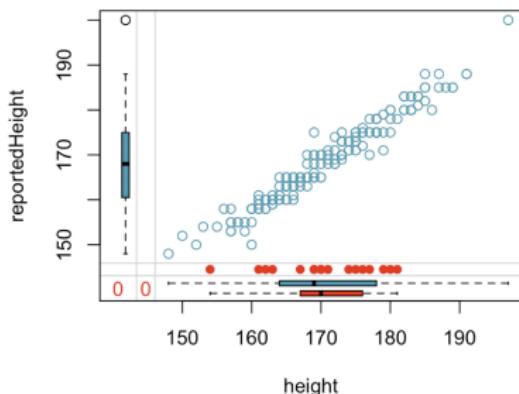
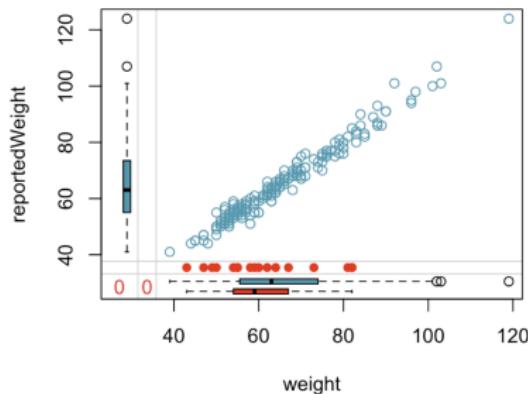
on demande de déclarer une taille et un poids (en plus de les mesurer): on observe 17 NA

```
1 > mean(Davis$reportedHeight)
2 [1] NA
3 > mean(Davis$reportedHeight, na.rm=TRUE)
4 [1] 168.4973
```

Values manquantes (NA)

Pour les variables continues bivariées appareillées

```
1 library(missMDA)
2 library(VIM)
3 marginplot(Davis[,c("weight", "reportedWeight")])
4 marginplot(Davis[,c("height", "reportedHeight")])
```



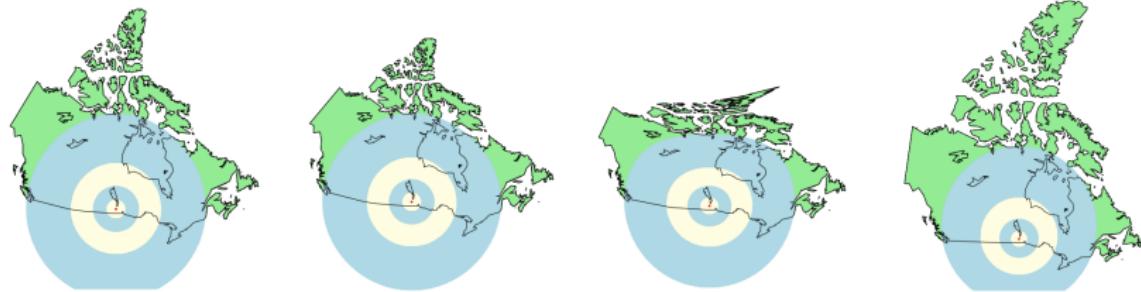
Données textuelles

Il alla sur-le-champ à la prison, il descendit au cabanon du "saltimbanque", il l'appela par son nom, lui prit la main et lui parla. Il passa toute la journée auprès de lui, oubliant la nourriture et le sommeil, priant Dieu pour l'âme du condamné et priant le condamné pour la sienne propre. Il lui dit les meilleures vérités qui sont les plus simples. Il fut père, frère, ami, évêque pour bénir seulement. Il lui enseigna tout, en le rassurant et en le consolant. Cet homme allait mourir désespéré. La mort était pour lui comme un abîme. Debout et frémissant sur ce seuil lugubre, il reculait avec horreur. Il n'était pas assez ignorant pour être absolument indifférent. Sa condamnation, secousse profonde, avait en quelque sorte rompu ça et là autour de lui cette cloison qui nous sépare du mystère des choses et que nous appelons la vie. Il regardait sans cesse au dehors de ce monde par ces brèches fatales, et ne voyait que des ténèbres. L'évêque lui fit voir une clarté.

il	et	lui	la	pour	le	en	du	de	qui	que	priant
10	8	7	7	5	4	3	3	3	2	2	2

Données spatiales et projections

CRS (Coordinate Representation Systems) can get complicated



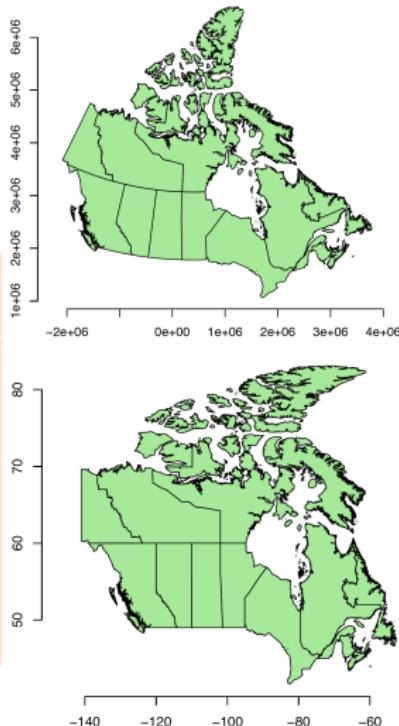
E.g. **EPSG:4326** latitude, longitude in WGS-84 coordinate system
EPSG:900913 and **EPSG:3857**: Google spherical Mercator
ESRI:102718: and **NAD 1983 State Plane NY Long Island**

Données spatiales et projections

Usually spatial data are in one specific CRS
but we can convert to another one

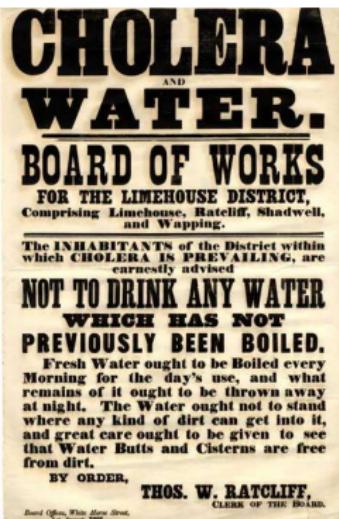
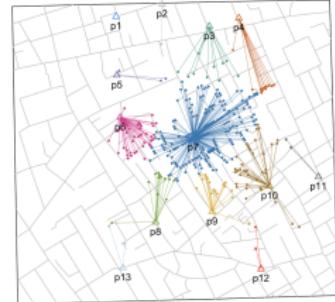
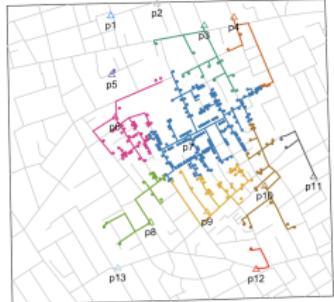
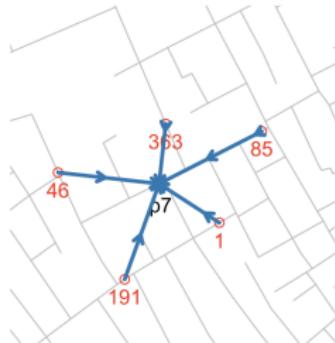
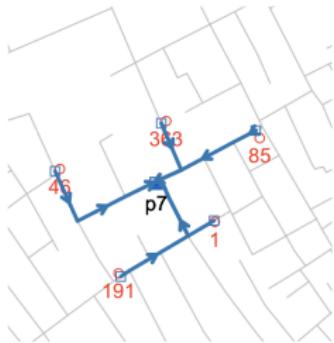
<https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/>

```
1 library(rgdal)
2 library(mapproj)
3 Proj = CRS("+proj=longlat +datum=WGS84")
4 CAN_shp = readShapePoly("CAN_adm1.shp",
                         verbose=TRUE, proj4string=Proj)
5 plot(CAN_shp)
6 new_CAN_shp = spTransform(CAN_shp, CRS
                           ("+init=epsg:26978"))
7 plot(new_CAN_shp)
```

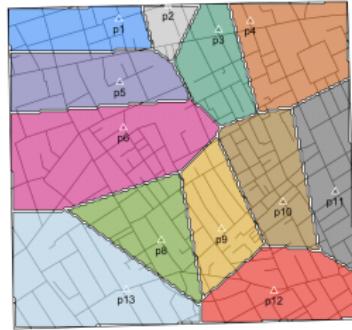
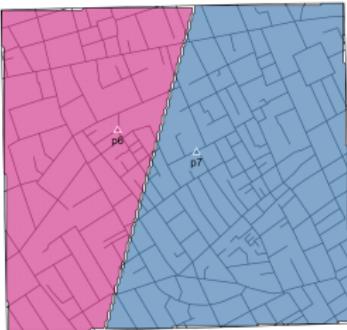
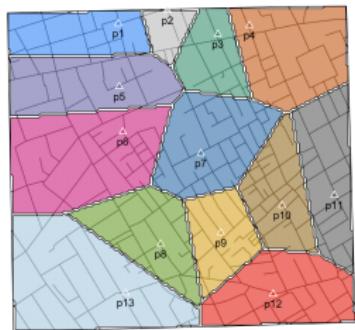
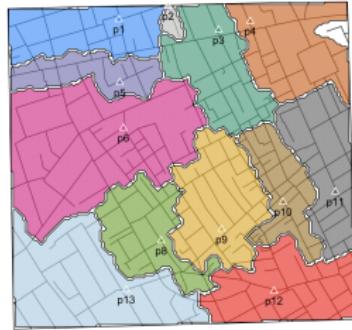
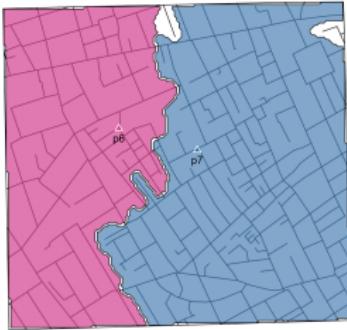
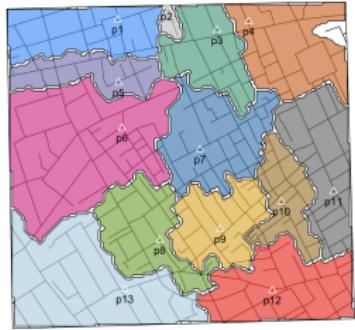


Cholera in London (Snow)

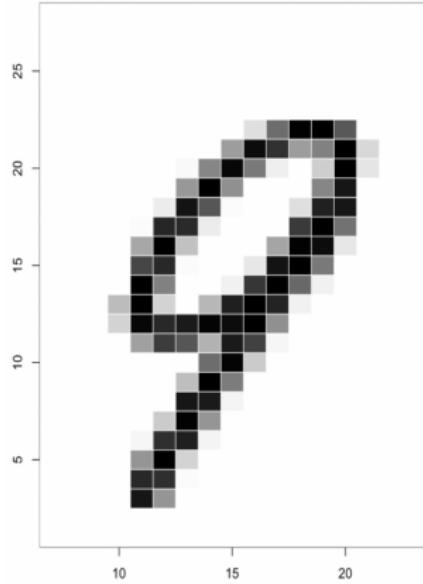
The cholera map that changed the world
(inspired by [lindbrook's R codes](#))



Cholera in London (Snow)



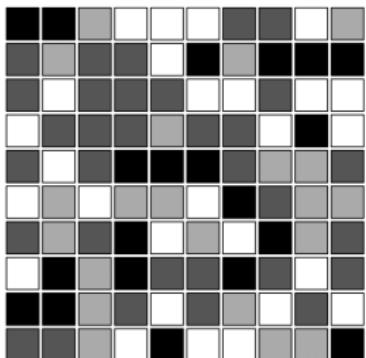
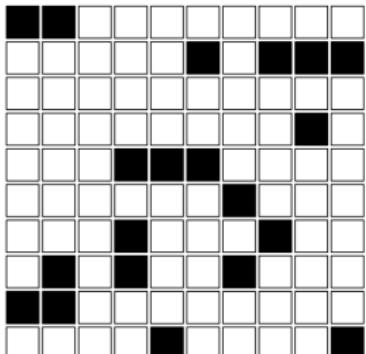
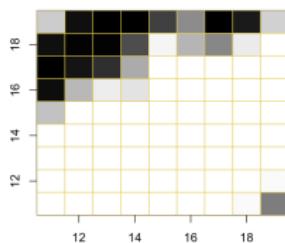
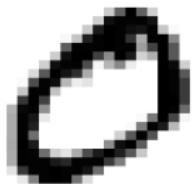
Images I



Images II

A $(w \times h)$ picture, with w pixels for the width and h for the height is a **matrix**

Example: black and white picture, matrix M with $M_{i,j} \in [0, 1]$ the grey level.



Images III

A $(w \times h)$ picture, with w pixels for the width and h for the height is a [tensor](#)

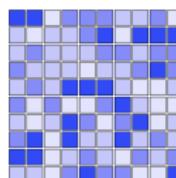
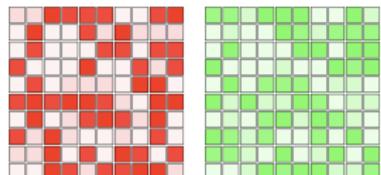
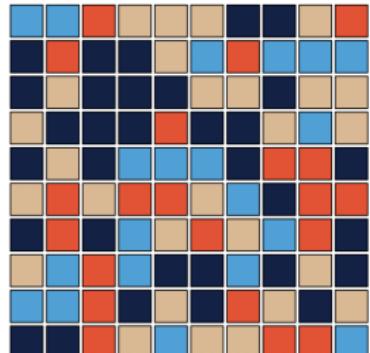
Example: a color picture is a tensor T

$$T = (T[,,r], T[,,g], T[,,b])$$

(RGB decomposition of a color)

with $T_{i,j,c} \in [0, 1]$ the color level.

working with pictures means working with three dimensional matrices



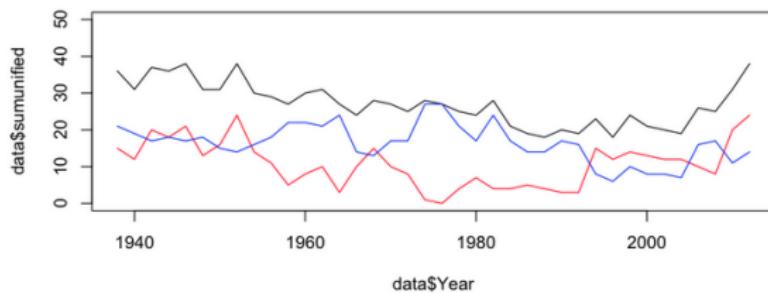
R is based on the S statistical programming language developed by [John Chambers](#) at Bell labs in the 80's

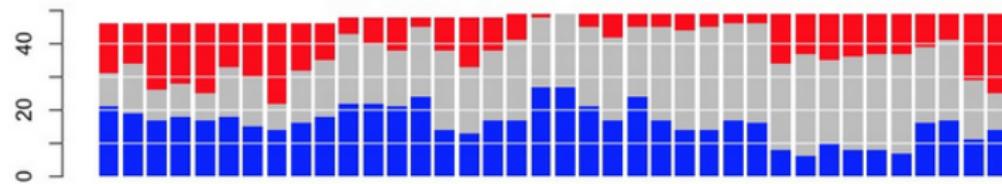


R is an open-source implementation of the S language, developed by [Robert Gentleman and Ross Ihaka](#)

'If you can picture it in your head, chances are good that you can make it work in R. R makes it easy to read data, generate lines and points, and place them where you want them. It's very flexible and super quick. When you've only got two or three hours until deadline, R can be brilliant.' Amanda Cox, a graphics editor at the New York Times. "R is particularly valuable in deadline situations when data is scant and time is precious."

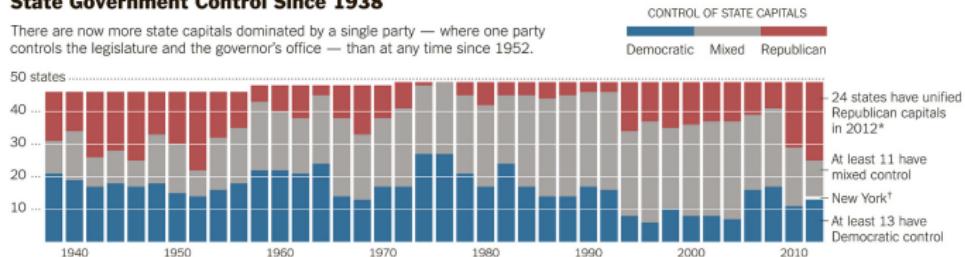
Source: <http://chartsnthings.tumblr.com/post/36978271916/r-tutorial-simple-charts>





State Government Control Since 1938

There are now more state capitals dominated by a single party — where one party controls the legislature and the governor's office — than at any time since 1952.



* Virginia is counted as unified Republican because its State Senate is tied and its tiebreaker, the lieutenant governor, is a Republican.

† Early results appeared to show that New York had unified Democratic control, but votes are still being counted in many races.

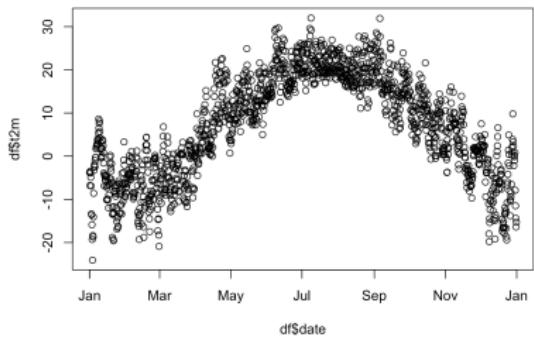
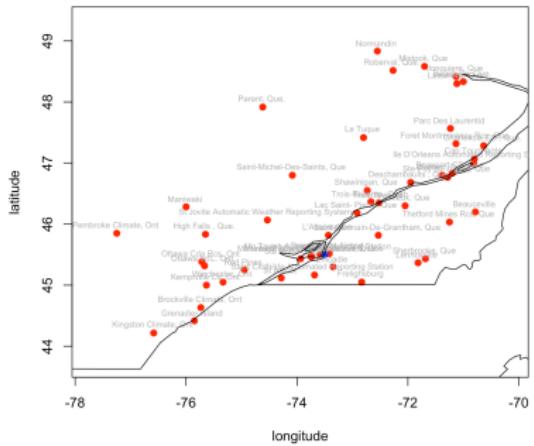
Source: National Conference of State Legislatures

THE NEW YORK TIMES

Exemple de Package : climate

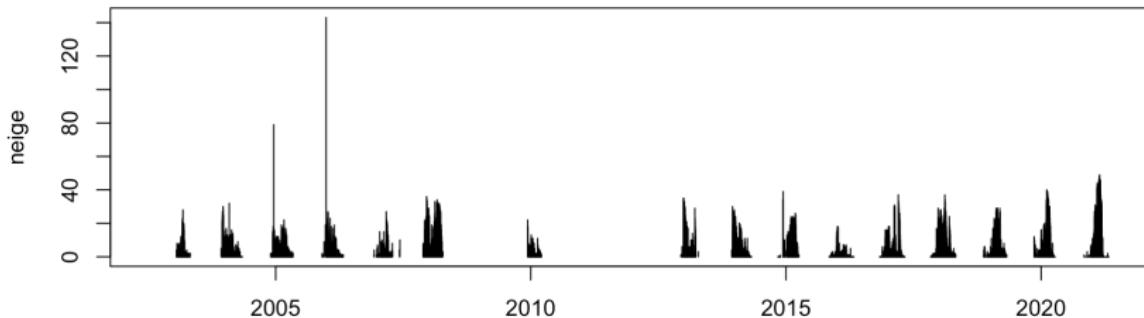
```
1 > library(climate)
2 > ns = nearest_stations_ogimet(
+   country = "Canada", point
+   = c(-73.5, 45.5),
+   no_of_stations = 50,
+   add_map = TRUE)
3 > df = meteo_noaa_hourly(
+   station = "711830-99999",
+   year = sample(1991:2022,
+   1))
4 > plot(df$date, df$t2m)
```

cf [isd-history](#) pour les codes station



Exemple de Package : climate II

```
1 > df <- meteo_ogimet(interval = "hourly", date = c  
2 > ("1991-01-01", "2021-05-01"), station = "71183")  
3 > date = as.Date(df$Date, format = "%m/%d/%Y")  
4 > neige = as.numeric(df$Snowcm)  
4 > plot(d,neige,type="h")
```



Exemple de Package : climate III

Attention, il peut y avoir des soucis lors de l'importation...

```
1 > plot(d,df$TC)
```

