STT 1000 - STATISTIQVES

ARTHVR CHARPENTIER

Définitions

- Individus : objets décrits par un ensemble de données. Un individu peut être une personne, un termomètre, 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 appareilléess (défini ensuite)

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 $\{y_1^{(1)},\cdots,y_{n_1}^{(1)}\}$ et $\{y_1^{(2)},\cdots,y_{n_2}^{(2)}\}$
- Échantillons appareillés : on obtient deux échantillons, pour deux variables, mais les mêmes individus $\{x_1,\cdots,x_n\}$ et $\{y_1,\cdots,y_n\}$, aussi noté $\{(x_1,y_1),\cdots,(x_n,y_n)\}$

(rapide) Typologie

- Variable catégorielle (facteur): les individus sont partitionnés entre plusieurs groupes (en prenant une modalité, et une seule)
  - catégorielle nominale e.g.
    genre $\in \{\text{homme}, \text{femme}\}$ 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 discrète e.g.
    nombre d'enfants, étage

Données pour illustrer

taille d'élèves dans un groupe de 200 personnes,

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

- `sex`: genre de la personnne, catégorielle nominale
- `weight`: poids de la personnne (kg), quantitative continue
- `height`: taille de la personnne (cm), quantitative continue

Données pour illustrer

Pour les variables catégorielles,

```
> mean(Davis$sex)
[1] NA
Warning message:
In mean.default(Davis$sex) :
  argument is not numeric or logical: returning NA
> table(Davis$sex)

  F   M
112  88
```

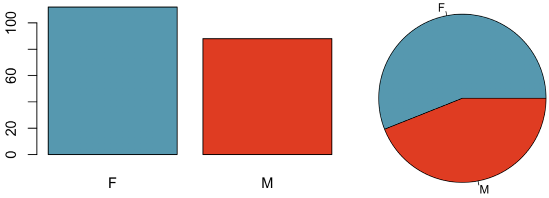
Le genre est nominal, mais on peut avoir des variables ordinales

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

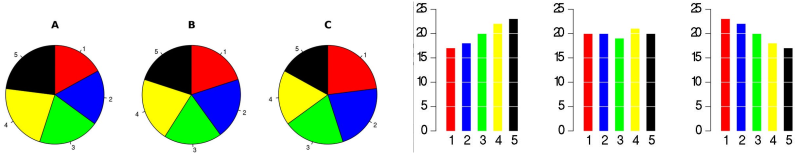
Données pour illustrer

Pour les variables catégorielles,

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



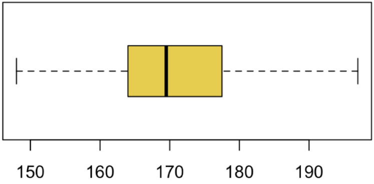
petite note: Pie Charts Are The Worst Charts In The World

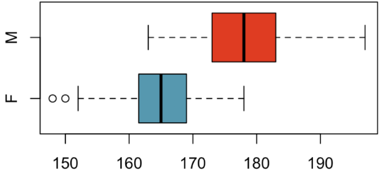


Données pour illustrer

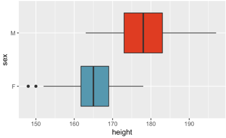
Pour les variables continues,

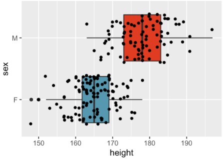
```
> boxplot(Davis$height, horizontal = TRUE)
> boxplot(Davis$height~Davis$sex, horizontal = TRUE,)
```

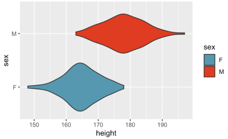





```
> library(ggplot2)
> ggplot(aes(x=height,y=sex, fill=sex), data=Davis)+
    geom_boxplot()
```



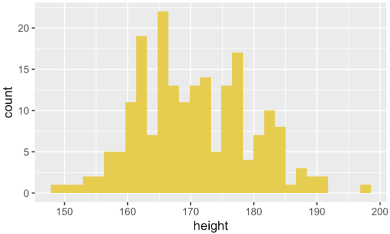


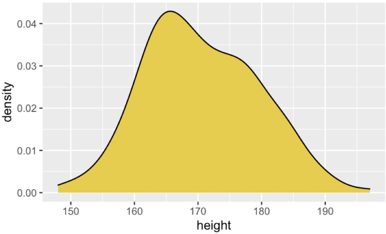


Données pour illustrer

Pour les variables continues

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



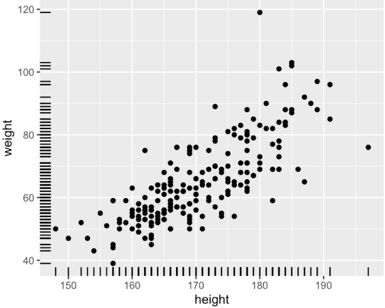


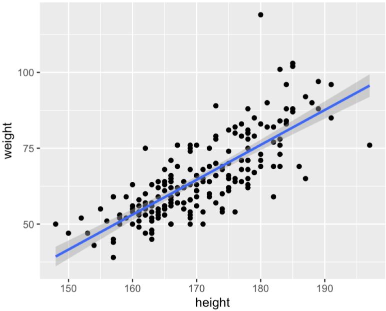

```
> mean(Davis$height)
[1] 170.565
```

Données pour illustrer

Pour les variables continues bivariées appareillées

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

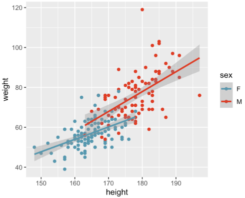


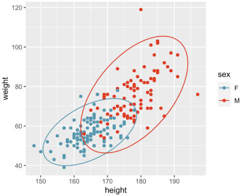


Données pour illustrer

Pour les variables continues bivariées appareillées

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





Values manquantes (NA)

via données ouvertes Montréal: comptages de vélos de 2011 à 2017 pour de nombreuses pistes cyclables

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

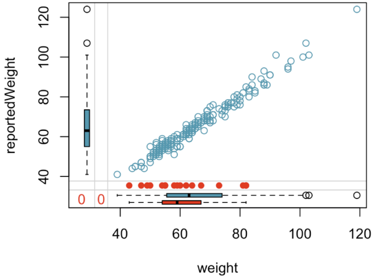
on demande de déclarer une taille et un poids (en plus de les mesurer): on observe 17 NA

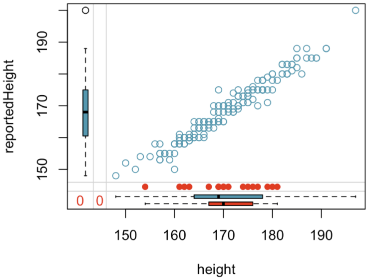
```
> mean(Davis$reportedHeight)
[1] NA
> mean(Davis$reportedHeight, na.rm=TRUE)
[1] 168.4973
```

Values manquantes (NA)

Pour les variables continues bivariées appareillées

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



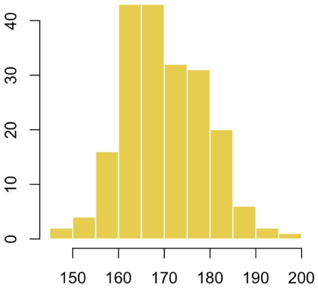


Histogramme

Pour faire l'histogramme d'une variable X,

1. on divise l'étendue de X en classes (intervalles) disjointes
2. on compter le nombre d'observations pour chacune de ces classes
3. on représente un rectangle dont la largeur correspond à l'étendue de la classe et dont la hauteur est proportionnelle au nombre ou au pourcentage d'observations dans cette classe

```
> height_class = cut(Davis$height,
    breaks = seq(145,200,by=5))
> table(height_class)
height_class
(145,150] (150,155] (155,160]
        2         4        16
(160,165] (165,170] (170,175]
       43        43        32
(175,180] (180,185] (185,190]
       31        20         6
(190,195] (195,200]
        2         1
```



Caractérisation d'une distribution

On dit qu'une distribution est unimodale si elle ne possède qu'un pic majeur.
Quand une distribution n'est pas symétrique, elle est dite asymétrique ; on dit qu'une distribution est asymétrique à droite le mode est à droite de la valeur moyenne
Considérons un échantillon $x_1, \cdots, x_n$ ordonné $(x_1 \leq x_2 \leq \cdots \leq x_n)$
Parmi les mesure de la tendance centrale d'un ensemble de nombres

- La moyenne $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$
- La médiane

$$\text{médiane}[\boldsymbol{x}] = \begin{cases} x_{(n+1)/2} \text{ si } n \text{ impair} \\ \dfrac{1}{2}x_{n/2} + \dfrac{1}{2}x_{n/2+1} \text{ si } n \text{ pair} \end{cases}$$

Caractérisation d'une distribution

Plus généralement, considérons les quantiles d'ordre $\alpha \in [0,1]$)
Considérons un échantillon $x_1, \cdots, x_n$ <u>ordonné</u> ($x_1 \leq x_2 \leq \cdots \leq x_n$)
Le quantile d'ordre α, noté q_α de la série d'observations est la valeur telle qu'une proportion α des données sont plus petites ou égales à q_α et une proportion $1-\alpha$ des données sont plus grandes ou égales. Nous le définissons par

$$q_\alpha = (1-f)x_k + fx_{k+1}$$

où $k = \lceil n\alpha \rceil$ et $f = n\alpha - \lfloor n\alpha \rfloor$.
Example $\boldsymbol{x} = \{1, 2, \cdots, 17\}$, on veut $q_{90\%}$

```
> x = 1:17
> quantile(x,.9,type=4)
 90%
15.3
> (ceiling(17*.9)-17*.9)*x[floor(17*.9)]+(17*.9-floor
    (17*.9))*x[ceiling(17*.9)]
[1] 15.3
```

Caractérisation d'une distribution

La variance est

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ ou bien } \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2?$$

(on reviendra longuement sur le $n-1$)
L'étendue pour un échantillon $\{x_1, x_2, \cdots, x_n\}$ est

$$e = \max\{x_i\} - \min\{x_i\}$$

L'intervalle interquartile est

$$IQ = q_{75\%} - q_{25\%}$$

(utilisé dans les box-plot / boîte à moustaches)

Centrer & Réduire

Pour les variables continues, il est parfois intéressant de les centrer et réduire

$$\tilde{x}_i = \frac{x_i - \overline{x}}{\sqrt{s^2}}$$

La série $\tilde{x}$ n'a plus d'unité.

- centrée: $\displaystyle\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i = 0$ (moyenne)
- réduite $\displaystyle\frac{1}{n-1}\sum_{i=1}^{n}\tilde{x}_i^2 = 1$ (variance empirique)

Cette transformation n'a pas d'incidence sur les profils de variation (on ne change pas la forme de la distribution).

Si deux variables sont les mêmes à une transformation près alors leurs deux versions centrées-réduites sont les mêmes.

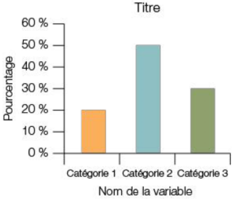
Notons qu'environs 95% des observations devraient avoir une cote comprise entre -2 et 2. En effet,

```
> pnorm(2)- pnorm(-2)
[1] 0.9544997
```

Présentation

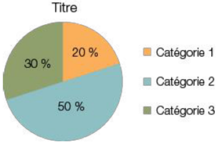
- **Pour représenter la distribution d'une variable qualitative**

Diagramme à rectangles (verticaux ou horizontaux)



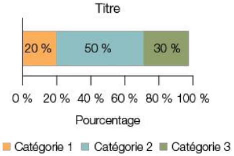
Favorise la comparaison des catégories entre elles.

Diagramme circulaire



Favorise la comparaison de chaque catégorie par rapport à l'ensemble des données.

Diagramme linéaire



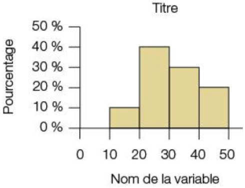
Favorise la comparaison de chaque catégorie par rapport à l'ensemble des données. Facilite la comparaison de plusieurs distributions.

(via Simard (2015))

Présentation

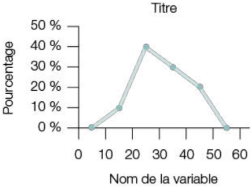
- **Pour représenter la distribution d'une variable quantitative continue**

Histogramme



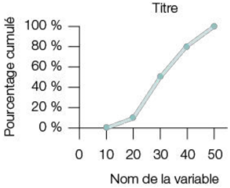
Attention, si les classes n'ont pas la même amplitude, il faut effectuer une rectification de fréquences.

Polygone de fréquences



Facilite la comparaison de plusieurs distributions ayant les mêmes classes.

Ogive ou courbe de fréquences cumulées



Pour représenter une distribution de fréquences cumulées.

(via Simard (2015))