Statistique descriptive unidimensionnelle: Descriptions numériques des variables quantitatives

Antonio Falcó

Dans cette lesson on s'intéresse plus particuliérement aux variables quantitatives avec un grand nombre de modalités, et on considére les caractéristiques suivantes de leur distribution:

- position: *Où se situe la distribution?*
- dispersion: *A quel point la distribution est-elle éparpillée*

Mesures du "milieu" d'une distribution

Pour mesurer le "milieu" d'une distribution, i.e. où se trouvent les données de façon globale, les deux mesures les plus utilisées sont la moyenne arithmétique, souvent appelée simplement moyenne, et la médiane. Une troisième mesure parfoisutilisée est le mode.

Moyenne

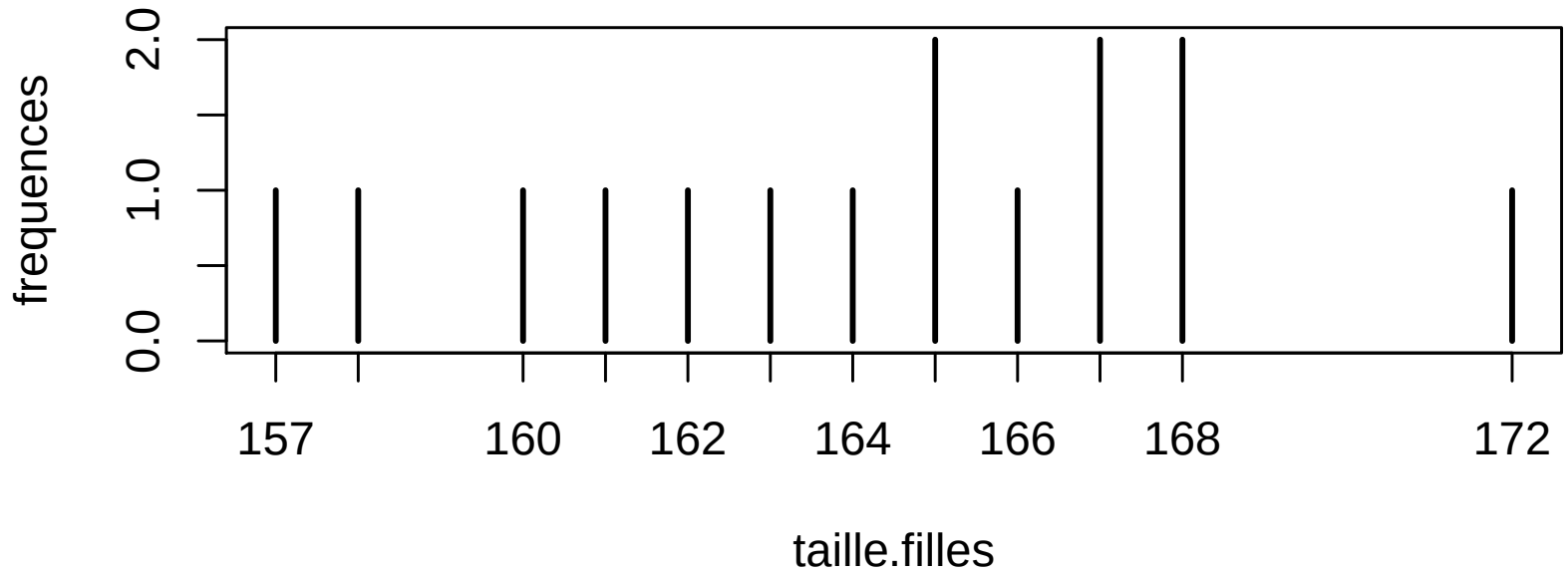
Soient $x_1, \ldots, x_n$ les observations d'une variable X. La moyenne de X, notée $m(X)$, est définie par

$$m(X) = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

Exemple 1. Tailles des filles [cm]: 168, 157, 167, 168, 163, 167, 166, 164, 172, 165, 158, 161, 160, 162, 165

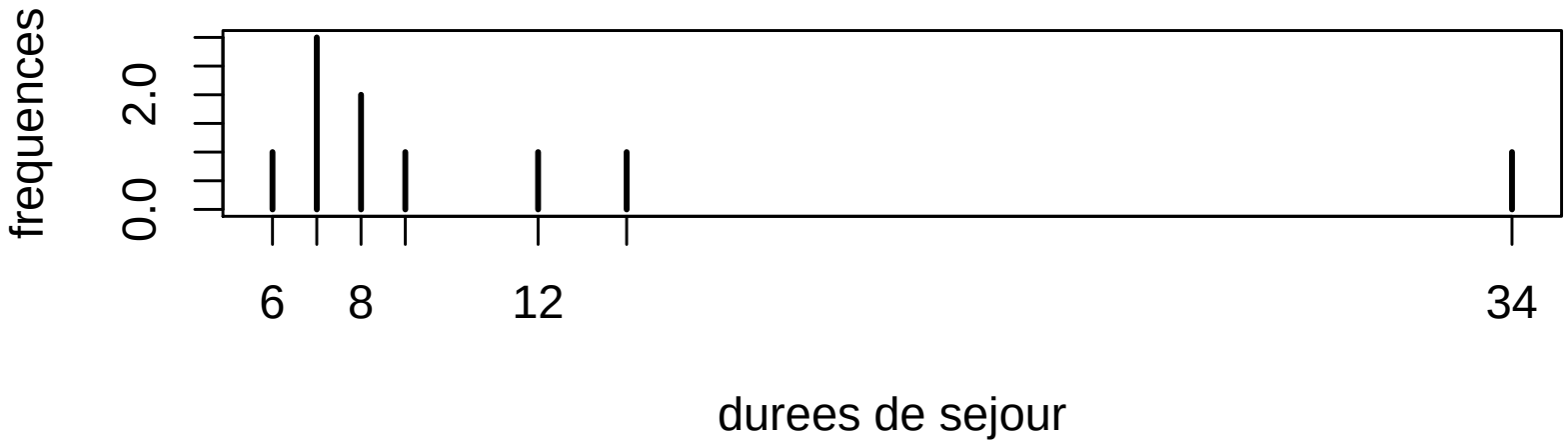
$$m(T) = (168 + 157 + 167 + 168 + 163 + 167 + 166 + 164 + 172 + 165 + 158 + 161 + 160 + 162 + 165)/15 = 164.2$$

La moyenne est un bon résumé du "milieu" de la distribution.



Exemple 2. Durées de séjour dans un hôpital [jours]: 7, 12, 14, 7, 34, 8, 8, 9, 6, 7

$$m(D) = (7 + 12 + 14 + 7 + 34 + 8 + 8 + 9 + 6 + 7)/10 = 11.2$$



La moyenne est un mauvais résumé du "milieu" de la distribution, elle est influencée par quelques valeurs extrêmes.

Propriétés de la moyenne

Soient X et Y deux variables, $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$ leurs observations sur les mêmes individus 1 à n. Soient a, b et c des constantes.

1. Si tous les x_i sont ≥ 0 alors $m(X) \geq 0$

2. $m(aX) = a\, m(X)$	La variable aX est définie comme ayant les observations $ax_1, \ldots, ax_n$. Ex: Si X est une taille en m et que $a = 100$, aX est cette taille en cm.
3. $m(X + a) = m(X) + a$	La variable $X + a$ est définie comme ayant les observations $x_1 + a, \ldots, x_n + a$. Ex: Si X est une température en degrés Celsius et que $a = 273.15$, $X + a$ est cette température en degrés Kelvin.

4. $m(X+Y) = m(X) + m(Y)$	La variable $X+Y$ est définie comme ayant les observations $x_1 + y_1, \ldots, x_n + y_n$. Ex: Si X et Y sont les pts obtenus à deux questions d'examen, $X+Y$ est le total des pts.

5. $m(aX + bY + c) = a\,m(X) + b\,m(Y) + c$ (découle de 2., 3. et 4.)

6. En général, $m(X\,Y) \neq m(X)\,m(Y)$	La variable $X\,Y$ est définie comme ayant les observations $x_1\,y_1, \ldots, x_n\,y_n$.

Médiane

La médiane est une valeur telle que la moitié des observations se trouve à sa gauche et l'autre moitié à sa droite.

Soient $x_1 \ldots, x_n$ les observations d'une variable X. Pour trouver leur médiane, il faut d'abord ordonner les observations.

Notation: on notera $x_{[1]}, \ldots, x_{[n]}$ les observations mises dans l'ordre croissant. Autrement dit, on aura toujours (par définition) que

$$\min(X) = x_{[1]} \leq \cdots \leq x_{[n]} = \max(X).$$

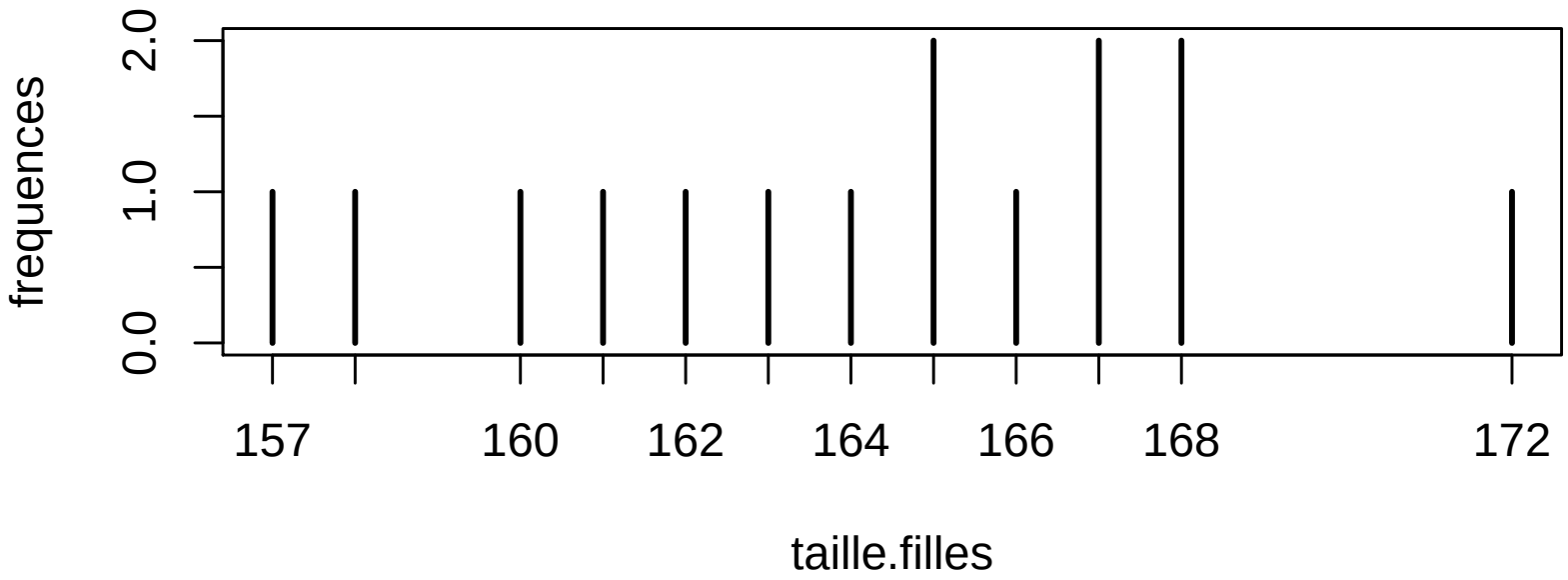
La médiane de X, notée $med(X)$, est alors définie par

$$med(X) = \begin{cases} x_{[\frac{n+1}{2}]} & \text{si } n \text{ est impair,} \\ \frac{1}{2}(x_{[\frac{n}{2}]} + x_{[\frac{n}{2}+1]}) & \text{si } n \text{ est pair.} \end{cases}$$

Exemple 1. Tailles des filles dans l'ordre croissant [cm]:

$t_{[1]}$	$t_{[2]}$	$t_{[3]}$	$t_{[4]}$	$t_{[5]}$	$t_{[6]}$	$t_{[7]}$	$t_{[8]}$	$t_{[9]}$	$t_{[10]}$	$t_{[11]}$	$t_{[12]}$	$t_{[13]}$	$t_{[14]}$	$t_{[15]}$
157	158	160	161	162	163	164	165	165	166	167	167	168	168	172

$n = 15$ est impair et donc $med(T) = t_{[\frac{n+1}{2}]} = t_{[8]} = 165$



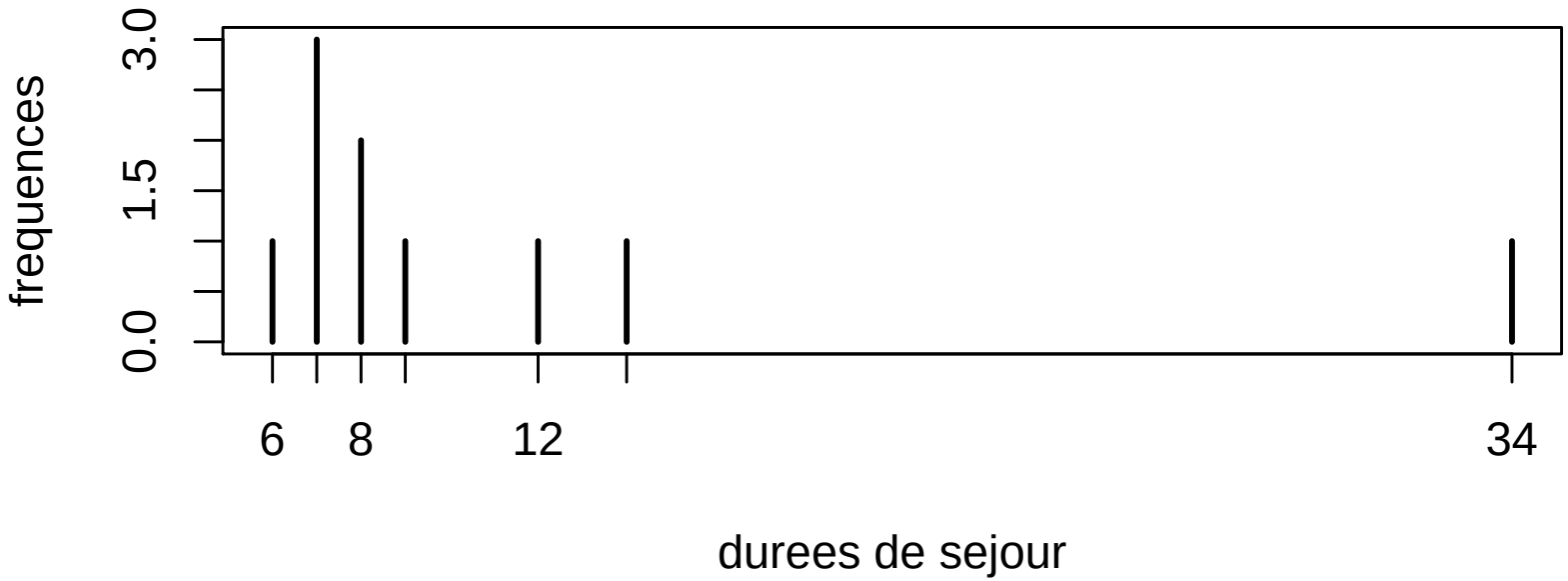
La médiane est un bon résumé du "milieu" de la distribution. Elle est trés proche de la moyenne ($m(T) = 164.2$).

Exemple 2. Durées de séjour dans un hôpital dans l'ordre croissant [jours]:

$d_{[1]}$	$d_{[2]}$	$d_{[3]}$	$d_{[4]}$	$d_{[5]}$	$d_{[6]}$	$d_{[7]}$	$d_{[8]}$	$d_{[9]}$	$d_{[10]}$
6	7	7	7	8	8	9	12	14	34

$n = 10$ est pair et donc

$$med(D) = \frac{1}{2}(d_{[\frac{n}{2}]} + d_{[\frac{n}{2}+1]}) = \frac{1}{2}(d_{[5]} + d_{[6]}) = (8+8)/2 = 8$$



La médiane est un meilleur résumé du "milieu" de la distribution que la moyenne. Elle est peu influencée par les valeurs extrêmes ($med(D) = 8$ et $m(D) = 11.2$).

Propriétés de la mdiane

Soient X et Y deux variables, $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$ leurs observations sur les mêmes individus 1 à n. Soit a constant.

1. Si tous les x_i sont ≥ 0 alors $med(X) \geq 0$

2. $med(aX) = a\, med(X)$	La variable aX est définie comme ayant les observations $ax_1, \ldots, ax_n$. Ex: Si X est une taille en m et que $a = 100$, aX est cette taille en cm.
3. $med(X + a) = med(X) + a$	La variable $X + a$ est définie comme ayant les observations $x_1 + a, \ldots, x_n + a$. Ex: Si X est une température en degrés Celsius et que $a = 273.15$, $X + a$ est cette température en degrés Kelvin.

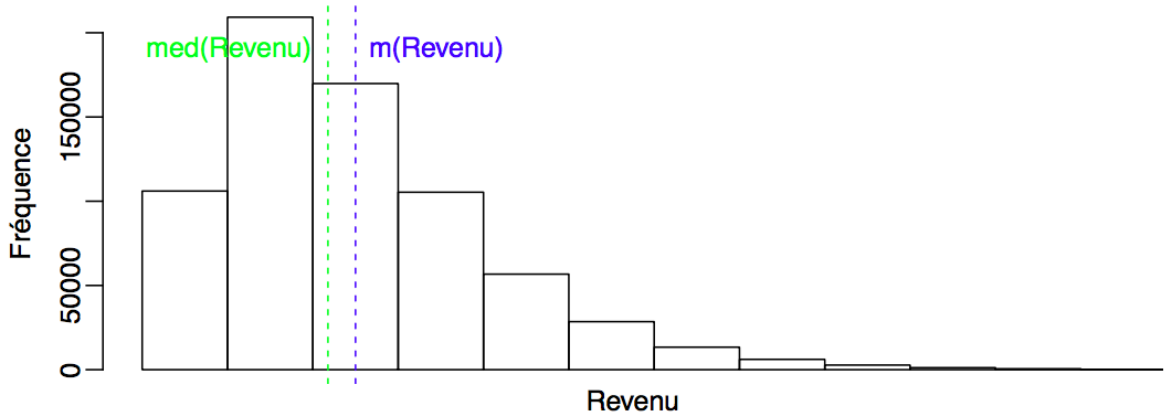
4. En général,b $med(X + Y) \neq med(X) + med(Y)$

5. En général, $med(X\,Y) \neq med(X)\,med(Y)$

Question: Faut-il utiliser la moyenne ou la médiane?

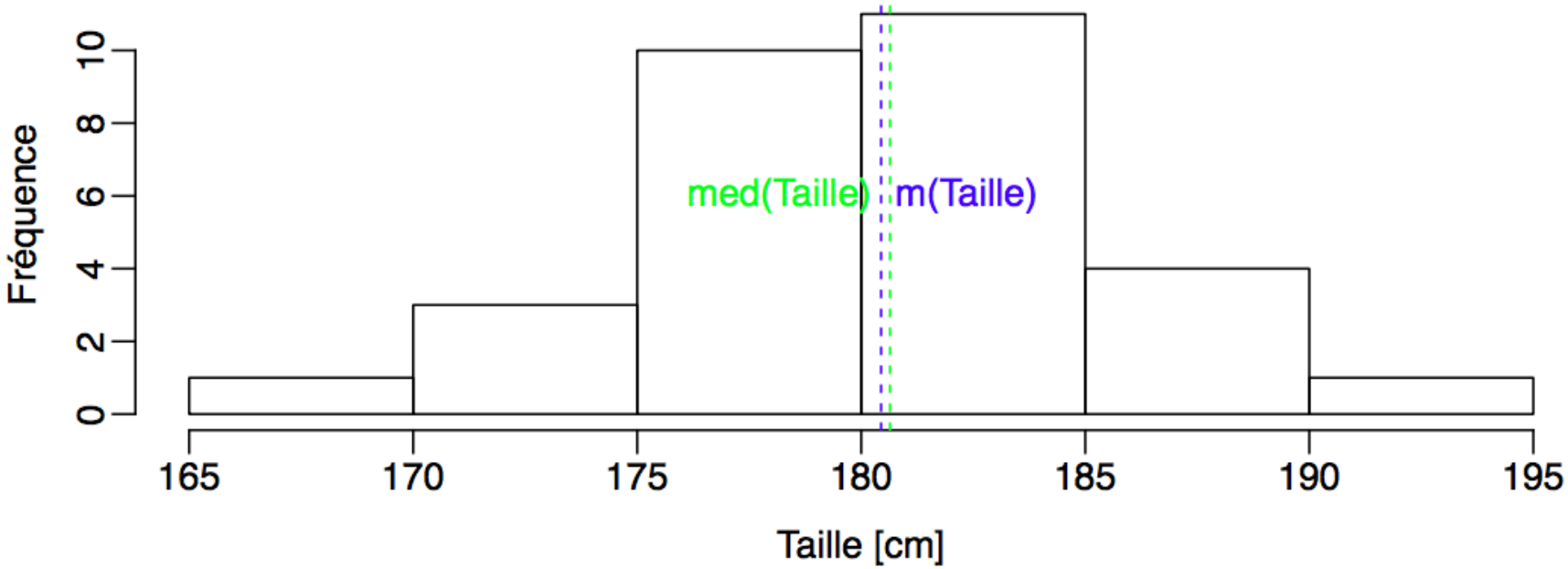
- Cela dépend de ce que l'on veut mesurer.

Exemple: Distribution des revenus. Les distributions de revenus ont typiquement une forme asymétrique.

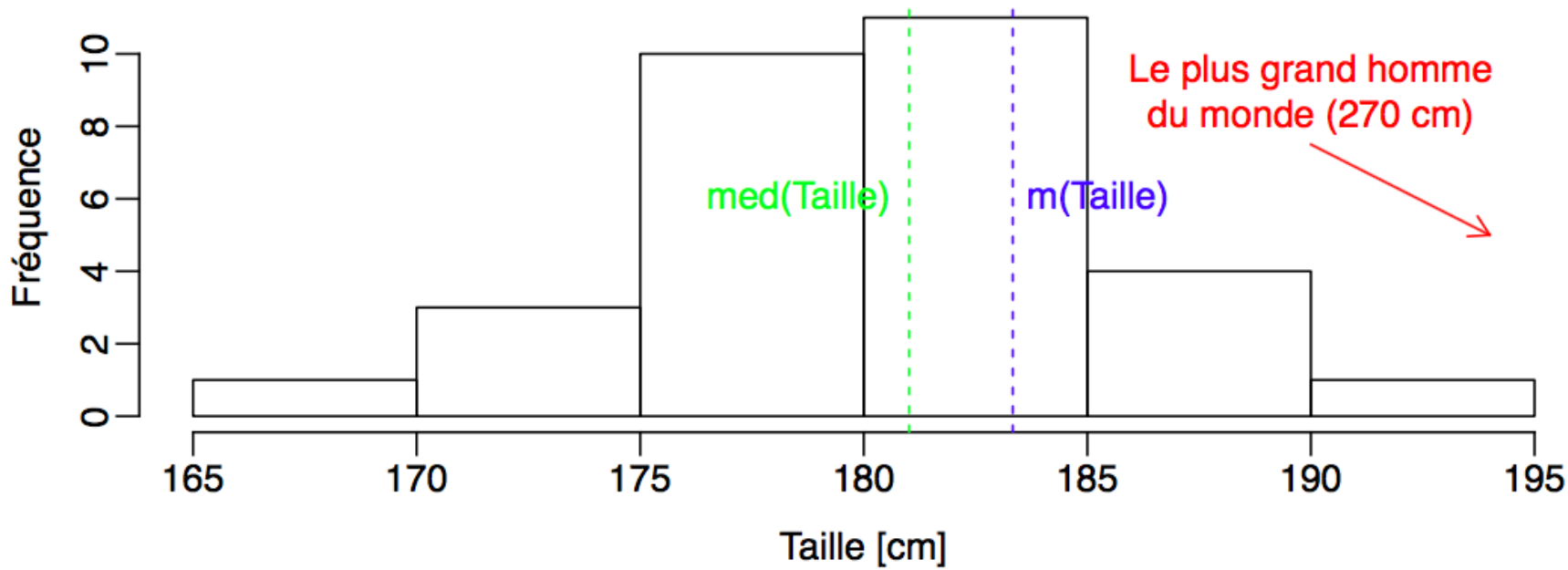


- Pour un habitant, il est plus intéressant de connaître la médiane: elle permet de se situer dans la moitié riche ou la moitié pauvre de la population.
- Pour l'administration des impôts, il est plus utile de connaître la moyenne: elle permet de se faire une idée des rentrées fiscales ($\approx$ revenu moyen $\times$ coefficient moyen $\times$ nb d'habitants). La moyenne est utile lorsqu'on s'intéresse à un total.

Cela dit, lorsque la distribution est symétrique la moyenne et la médiane sont égales.
Exemple: Tailles (simulées) de 30 hommes:



Par contre la moyenne est très sensible aux outliers: si on ajoute le plus grand homme du monde à notre échantillon, elle change d'environ 2 cm alors que la médiane ne change presque pas. Si on a affaire à une distribution symétrique mais qu'on s'attend à ce qu'il y ait des outliers, il vaut donc mieux utiliser la médiane.



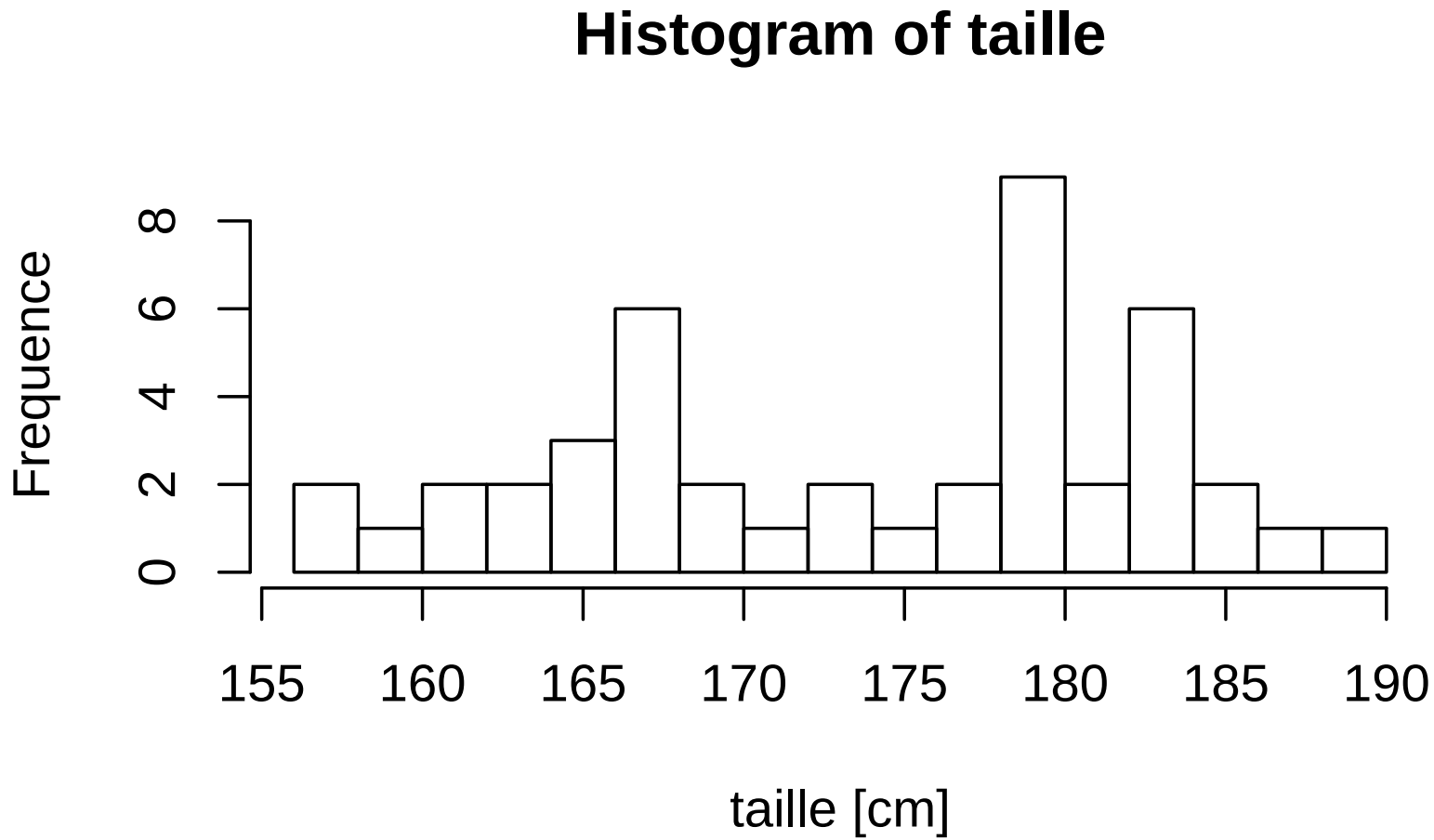
Mode

Le mode d'une distribution est défini comme la modalité qui a la plus haute fréquence.

De façon plus générale, on pourra appeler mode toute valeur où la fréquence atteint un maximum local. On pourra ainsi avoir des distribution bimodales, trimodales, etc.

Pour les variables quantitatives continues, on définit les modes à partir de l'histogramme, comme les milieux des classes de frêquence maximale.

Lorsqu'une distribution a plusieurs modes, c'est souvent le signe que la population est constituée de plusieurs sous-populations distinctes.



On observe deux modes (166 cm et 180 cm), correspondant aux sous-populations des filles et des garçons.

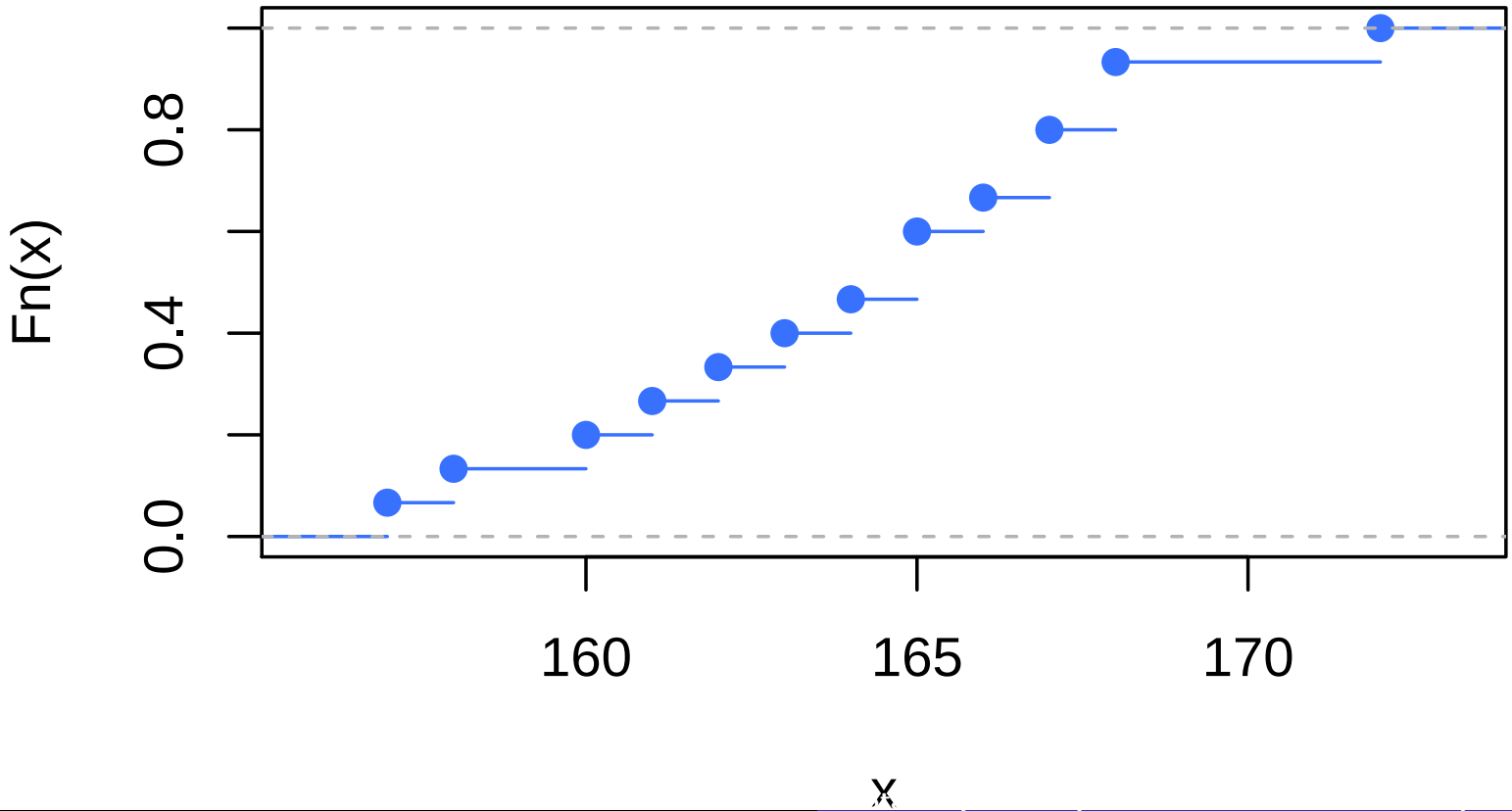
Autres mesures de position: les quantiles

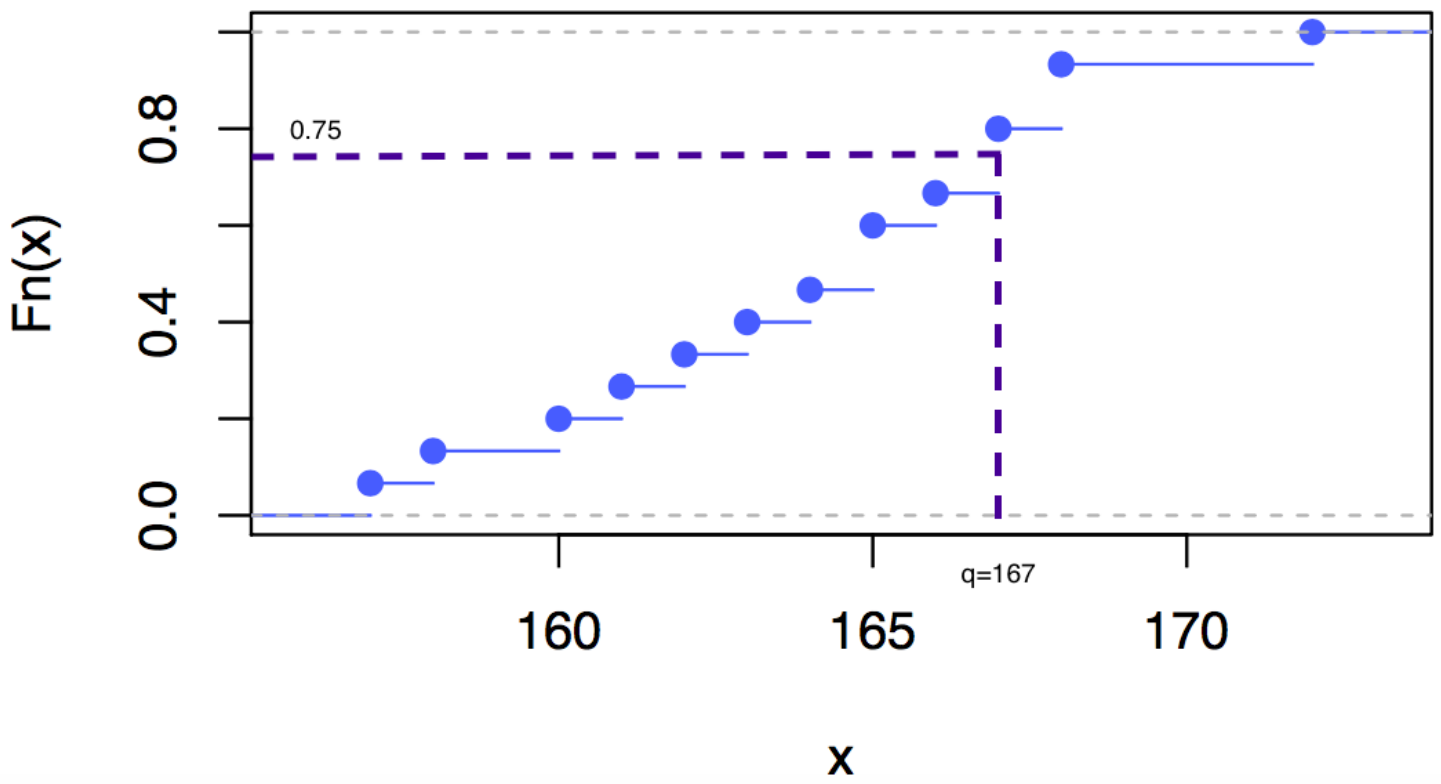
On a vu que la médiane partage la distribution en deux parties, de telle sorte que 50% des données lui sont inférieures et 50% lui sont supérieures.

On peut généraliser ce procédé en demandant qu'une proportion α des données soient dans la premiéere partie et le reste dansla deuxième. La limite entre les deux parties s'appelle alors le quantile d'ordre α et on le note q_α.

Autrement dit, le quantile d'ordre α est une valeur telle qu'une proportion α des observations se trouve à sa gauche et une proportion $1-\alpha$ à sa droite.

Pour définir les quantiles, on se sert de la fonction de distribution cumulative: Que vaut $q_{0.75}(T)$, le quantile d'ordre 75% de la distribution des tailles des filles?

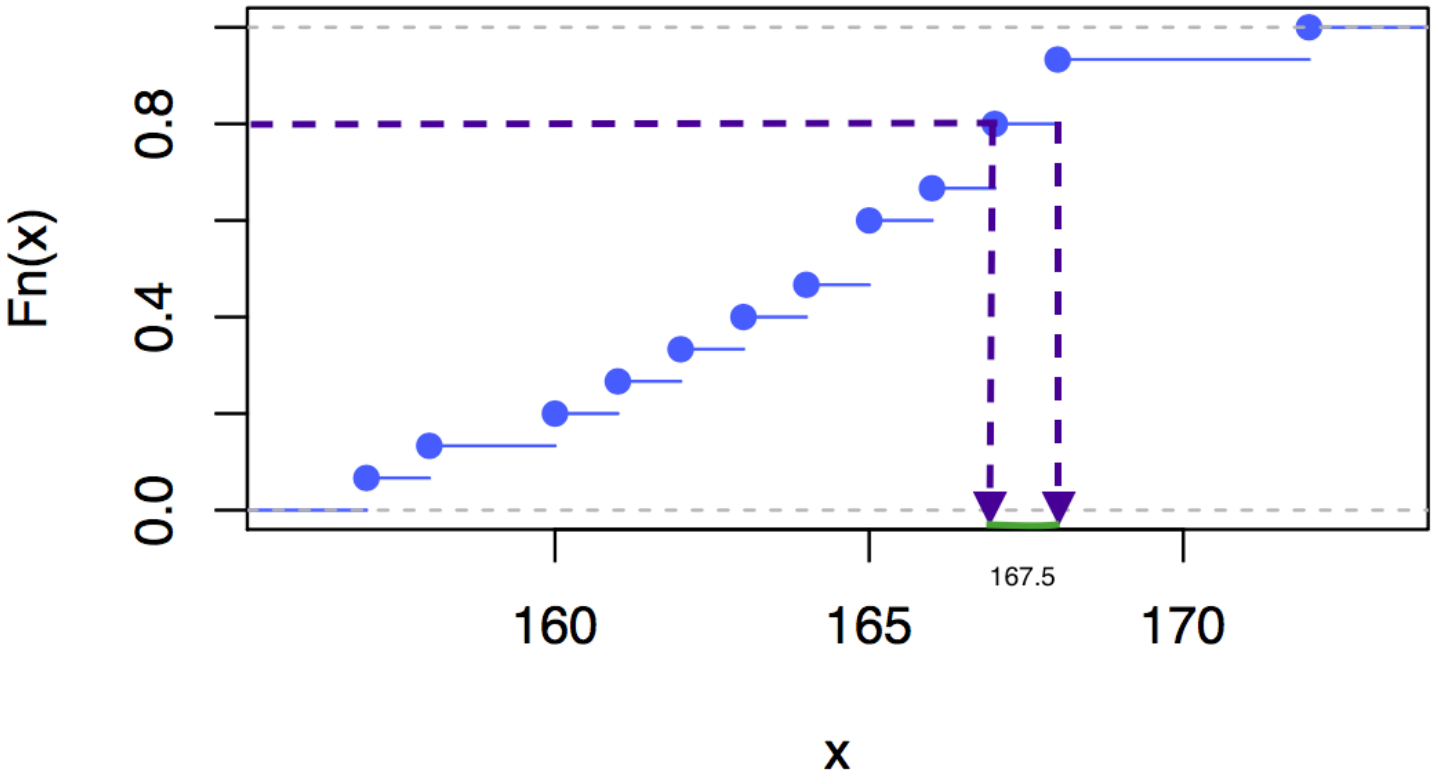




Les quantiles sont obtenus en inversant la fonction de distribution cumulative.

Cas spécial:

Que vaut $q_{0.8}(T)$, le quantile d'ordre 80% de la distribution des tailles des filles?



Lorsqu'on tombe sur un plateau, on prend la moyenne des valeurs extrêmes du plateau.

La mesure de dispersion (ou éparpillement) la plus utilisée est la variance, et plus particuliérement sa racine carrée appelée l'écart-type. Deux mesures alternatives sont le mad (median absolute deviation) et l'écart interquartile.

Variance

Soient $x_1, ..., x_n$ les observations d'une variable X. La variance de X, notée $s^2(X)$, est définie par

$$s^2(X) = m((X - m(X))^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - m(X))^2.$$

En mots, il s'agit de la moyenne des carrés des écarts entre X et sa moyenne. L'écart-type $s(X)$ est défini comme la racine carrée de la variance:

$$s(X) = \sqrt{s^2(X)}.$$

Exemple: Taille des filles

	t_i	$t_i - m(t_i)$	$(t_i - m(t_i))^2$
	168	3.8	14.44
	157	-7.2	51.84
	167	2.8	7.84
	168	3.8	14.44
	163	-1.2	1.44
	167	2.8	7.84
	166	1.8	3.24
	164	-0.2	0.04
	172	7.8	60.84
	165	0.8	0.64
	158	-6.2	38.44
	161	-3.2	10.24
	160	-4.2	17.64
	162	-2.2	4.84
	165	0.8	0.64
moyenne	164.2	0	15.63

On a donc $s^2(T) = 15.63$ cm^2 et $s(T) = \sqrt{15.63}$ cm $= 3.95$ cm.

Contrairement à la variance, l'écart-type est mesuré dans les mêmes unités que la variable.

Propriétés de la variance et de l'écart-type

Soient X et Y deux variables et soient a, b et c des constantes.

1. $s^2(c) = 0$.
2. $s^2(aX + b) = a^2 s^2(X)$
3. $s(aX + b) = |a| s(X)$
4. En général, $s^2(X + Y) \neq s^2(X) + s^2(Y)$
5. La somme des écarts $x_i - m(X)$ est toujours nulle
6. $s^2(X) = m(X^2) - m(X)^2$

La formule 6. est utile pour les calculs á la main, car elle évite de calculer tous les écarts $x_i - m(X)$.

Remarque

On trouve aussi dans certains ouvrages la définition alternative suivante de la variance:

$$S^2(X) = \frac{1}{n-1} \sum_{i=1}^{n} (xi - m(X))^2.$$

La raison pour prendre un dénominateur égal à $n-1$ au lieu de n dépasse le cadre de cette lesson. Notez que la formule 6. ne fonctionne pas avec cette définition alternative.

Variable standardisée

Soit X une variable. La variable Z définie comme

$$Z = \frac{X - m(X)}{s(X)}$$

est appelée la version standardisée ou centrée et réduite de X. En appliquant les propriétés de la moyenne et de la variance, on obtient que

- $m(Z) = 0$
- $s^2(Z) = 1$

Cette opération est utile lorsqu'on veut ramener des variables différentes sur une échelle commune, ou lorsqu'on veut se ramener à une situation standard (lessons suivantes).

mad

Soit X une variable. Le mad (median absolute deviation) de X est défini par

$$mad(X) = med(|X - med(X)|).$$

En mots, il s'agit de la médiane des écarts absolus entre X et sa médiane. De façon analogue à la relation entre moyenne et médiane, l'écart-type est une mesure très sensible aux outliers, alors que le mad est résistant.

Ecart interquartile

Pour une variable X, on définit le premier, le deuxième et le troisième quartile comme

$$q_{0.25}(X), q_{0.5}(X) \text{ et } q_{0.75}(X)$$

respectivement. Ainsi

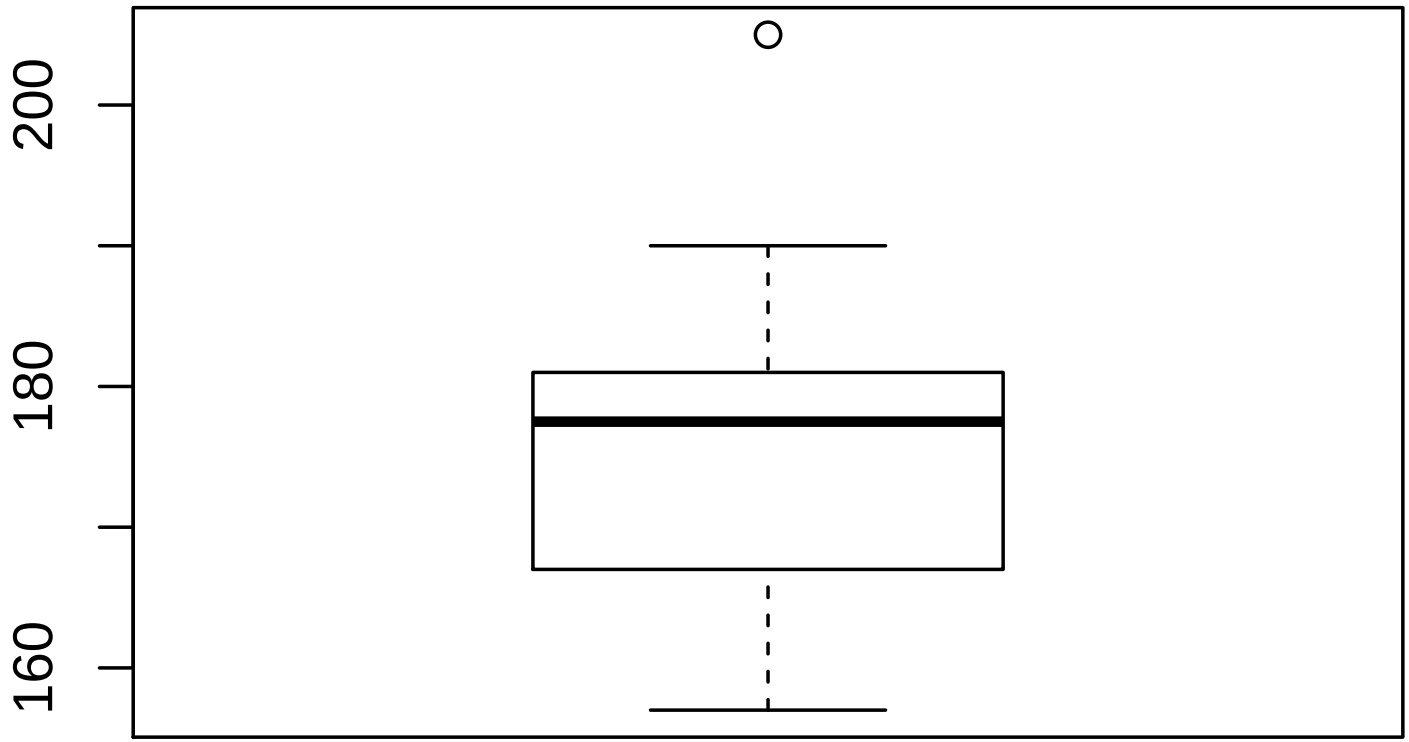
- Les quartiles partagent la distribution en quatre parties contenant chacune 25% des observations
- Le deuxième quartile n'est autre que la médiane.

L'écart interquatile de X, noté $I_q(X)$ est simplement défini comme la différence entre le troisième et le premier quartile de X :

$$I_q(X) = q_{0.75}(X) - q_{0.25}(X).$$

L'écart interquartile est plus résistant aux outliers que l'écart-type. Le mad résiste encore mieux, mais il est plus difficile à interpréter.

Le box-plot, ou box-and-whiskers plot (en français boîte à moustaches) est une représentation graphique simple mais puissante d'un échantillon.



Construction

Le long d'un axe vertical, on trace tout d'abord la box (boîte), qui va du premier au troisième quartile. Ainsi, la box contient approximativement la moitié (50%) centrale des données.

La box est ensuite partagée en deux par un trait horizontal au niveau de la médiane. Ensuite on va définir les inliers, i.e. les observations non extrêmes, comme toutes les observations se trouvant dans un intervalle défini comme suit:

- la borne supérieure est égale au troisième quartile plus $1.5 \times I_q$ (I_q = Intervalle interquartile = hauteur de la box).
- la borne inférieure est égale au premier quartile moins $1.5 \times I_q$.

Cette procédure trouve une justification dans le cadre de la distribution normale. Dans le cadre de ce modéle fréquent dans la nature, la définition ci-dessus conduit à environ 99% d'inliers et 1% d'outliers.

On peut alors tracer les moustaches:

- La moustache supérieure va du sommet de la boîte au plus grand des inliers
- La moustache inférieure va du bas de la boîte au plus petit des inliers

Les données qui ne sont pas des inliers sont marquées individuellement par le symbole 'O' (outlier).

Le box-plot permet en un coup d'oeil d'apprécier les caractéristiques suivantes d'une distribution:

- **Position**: la box indique ou se trouve la moitié centrale des données, et comment elle se répartit autour de la médiane.
- **Dispersion**: la hauteur de la box donne l'écart interquartile. La longueur des moustaches donne une idée de la dispersion des données extérieures a la box.
- **Asymétrie**: la position de la médiane dans la box et la différence de longueur entre les moustaches nous renseigne sur le degré d'asymétrie.
- **Présence d'outliers**: marqués individuellement.

Pour comparer des échantillons, on peut représenter plusieurs box-plots côte à côte.

```
summary(taille)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  157.0   167.0   177.0   174.3   181.0   190.0
```

```
mean(taille)#moyenne
```

```
[1] 174.3111
```

```
var(taille)#variance
```

```
[1] 78.08283
```

```
sd(taille)#ecart-type
```

```
[1] 8.836449
```