Démarche Statistique 1

Statistiques descriptives

Pierre Neuvial, http://stat.genopole.cnrs.fr/~pneuvial
Evry, M1 SGO, automne 2014

Introduction

Variables quantitatives

Définition: variable mesurant une **quantité**

Propriété: variable sur laquelle on peut faire des **opérations** arithmétiques (somme, moyenne, ...)

Variable continue

peut prendre (en théorie) un nombre infini de valeurs

→ taille, pression artérielle, ...

Variable discrète

ne peut prendre qu'un nombre fini de valeurs

→ nombre d'individus dans un foyer, nombre de lectures (reads) s'alignant à un endroit particulier du génome dans une expérience de RNAseq, ...

Variables qualitatives

Définition: variable dont chacune des "modalités" (valeurs possibles) décrivent une **qualité**, au sens étymologique de: "une nature possible"

Propriété: variable sur laquelle on **ne peut pas** faire des opérations arithmétiques !
On peut en revanche calculer des effectifs et des fréquences

Variable nominale

dont les modalités sont décrites par un "nom"

→ sexe, groupe sanguin

Variable ordinale

dont les modalités traduisent un ordre

→ taille de vêtements, échelle de préférence

Quel type de variable ?

Code postal

→ variable nominale ou ordinale ?
→ variable continue ou discrète ?

Génotype

→ variable nominale ou ordinale ?

Niveau d'expression d'un gène

→ variable continue ou discrète ?

Variables qualitatives

Exemple: échelle de douleur

Description

- étude post-opératoire
- 60 patients
- échelle ordinale de douleur: "none", "mild", "moderate", "severe"

Données brutes (6 premiers patients)

	PT_NUM	PAIN_LEVEL
1	1	moderate
2	2	none
3	3	mild
4	4	none
5	5	severe
6	6	none

Effectif, fréquence, et leurs versions cumulées

Exemples de questions

- Certains niveaux de douleurs sont-ils plus fréquents que d'autres ?
- La douleur sévère est-elle rare ?

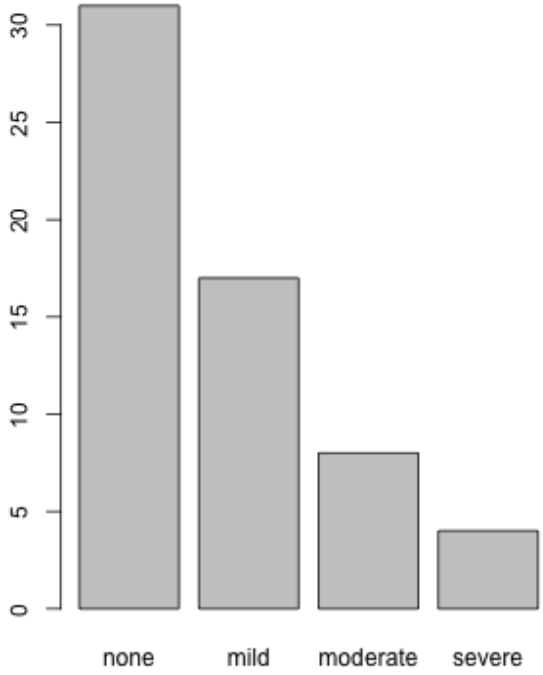
Indicateurs permettant de résumer ces données

	EFFECTIF	FRÉQUENCE	EFF. CUM.	FRÉQ. CUM.
none	31.00	0.52	31.00	0.52
mild	17.00	0.28	48.00	0.80
moderate	8.00	0.13	56.00	0.93
severe	4.00	0.07	60.00	1.00

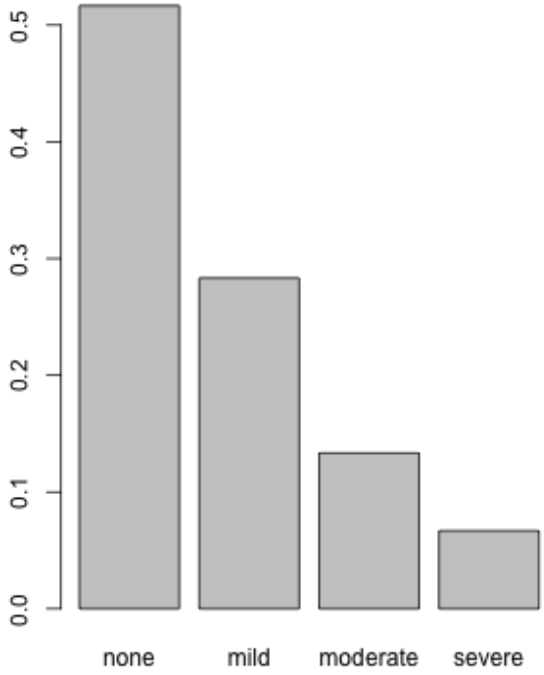
On a représenté la **distribution** des effectifs et des fréquences de l'échantillon.

Représentations graphiques: barplot

Effectifs



Fréquences



Variables continues

Résumer un échantillon de données continues

Une variable numérique (ex: la taille) peut prendre un nombre **infini** de valeurs. On ne peut donc pas résumer un échantillon de données continues à l'aide d'effectifs ou de fréquences comme dans le cas des données discrètes.

Indicateurs numériques

- mesures de **position** (moyenne, médiane, ...)
- mesures de **dispersion** (étendue, variance, écart-type, ...)
- quantiles

*Ces indicateurs pourront être définis sur la **population**, ou sur un **échantillon***

Représentations graphiques

- histogrammes
- boîtes à moustache ("boxplot")

Notation $\sum$ pour la somme

Somme des éléments

Si A est un ensemble de nombres on notera $\sum_{a \in A} a$ la somme des éléments de A.

$\rightarrow$ Si $A = \{3, 5, 4, 8, 7\}$, alors

$$\sum_{a \in A} a = 27$$

En particulier si A est un ensemble contenant les k éléments $a_1, a_2, \dots a_k$ on pourra aussi écrire cette somme comme $\sum_{i=1}^{k} a_i$.

Lorsqu'il n'y a pas d'ambiguité sur l'ensemble A considéré, on pourra noter $\sum a$ cette somme.

Somme des carrés

$$\sum_{a \in A} a^2 = 163$$

Mesures de position

Moyenne

On considère un échantillon noté x de n valeurs $x_1, x_2, \dots x_n$

Définition

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Exemple

$\rightarrow$ Si l'échantillon est $3, 5, 4, 8, 7$, alors $\bar{x} = \frac{27}{5} = 5.4$

Propriété

La moyenne est influencée par **toutes** les observations de l'échantillon

Médiane

On considère un échantillon noté x de n valeurs $x_1, x_2, \dots x_n$

On note $x_{(1)}, x_{(2)}, \dots x_{(n)}$ l'échantillon ordonné correspondant: $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$

La médiane est la valeur telle que

- **la moitié** des observations de l'échantillon sont plus grandes qu'elle
- (et donc aussi) **la moitié** des observations de l'échantillon sont plus petites qu'elle

Formule

- si n est impair, $\text{med}(x) = x_{((n+1)/2)}$
- si n est pair, $\text{med}(x) = \frac{x_{(n/2)} + x_{((n+1)/2)}}{2}$

Propriété

La médiane n'est **pas influencée** par les observations extrêmes de l'échantillon

Moyenne et médiane

→ Si l'échantillon est $3, 5, 4, 8, 7$, alors $\bar{x} = 5.4$ et $\text{med}(x) = 5$

→ Si l'échantillon est $3, 5, 4, 8, 7, 2, 11$, alors $\bar{x} = 5.7142857$ et $\text{med}(x) = 5$

→ Si l'échantillon est $3, 5, 4, 8, 7, 2, 161$, alors $\bar{x} = 27.1428571$ et $\text{med}(x) = 5$

Quantiles

La valeur telle qu'une proportion p des observations sont plus petites qu'elle (et donc $(1 - p)$ sont plus grande qu'elle) est appelée **quantile d'ordre p** de l'échantillon.

PROPORTION	NOM DU QUANTILE
10%	premier décile
25%	premier quartile
50%	médiane

PROPORTION	NOM DU QUANTILE
75%	dernier quartile
90%	dernier décile
p %	centile d'ordre p

Mesures de dispersion

Étendue

On considère un échantillon noté x de n valeurs $x_1, x_2, \dots x_n$

On note $x_{(1)}, x_{(2)}, \dots x_{(n)}$ l'échantillon ordonné correspondant: $x_{(1)} \leq x_{(2)} \leq \dots x_{(n)}$

Définition

L'étendue (*range* en anglais) est la distance séparant la plus petite et la plus grande valeur:

$$\text{Range}(x) = x_{(n)} - x_{(1)}$$

Exemple

$\rightarrow$ Si l'échantillon est $3, 5, 4, 8, 7$, alors $\text{Range}(x) = 5$

Propriété

L'étendue dépend uniquement des observations extrêmes de l'échantillon.
C'est donc une mesure très **instable** de la dispersion

Variance, écart-type

On considère un échantillon noté x de n valeurs $x_1, x_2, \dots x_n$

La **variance** est la moyenne des carrés des **écarts** de chaque observation x_i **à la moyenne** $\bar{x}$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

On la note s^2, où s est appelé **écart-type**

Propriétés

- comme son nom l'indique, s traduit l'**écart typique** entre deux observations de l'échantillon.
- contrairement à la variance, l'écart type a la même **échelle** que les variables de l'échantillon.
  $\rightarrow$ si x est un échantillon de tailles en cm, alors s est aussi en cm alors que s est en cm^2.

Exemple

$\rightarrow$ Si l'échantillon est $3, 5, 4, 8, 7$, alors $s = 2.0736441$

Autres mesures de variabilité

Moindres écarts absolus (median absolute deviation, mad)

La **mad** est la médiane des **valeurs absolues des écarts** de chaque observation x_i **à la médiane** $\bar{x}$

$$\text{mad} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Intervalle inter-quartile (inter-quartile range, IQR)

L'**IQR** est la distance entre le premier et le troisième quartile

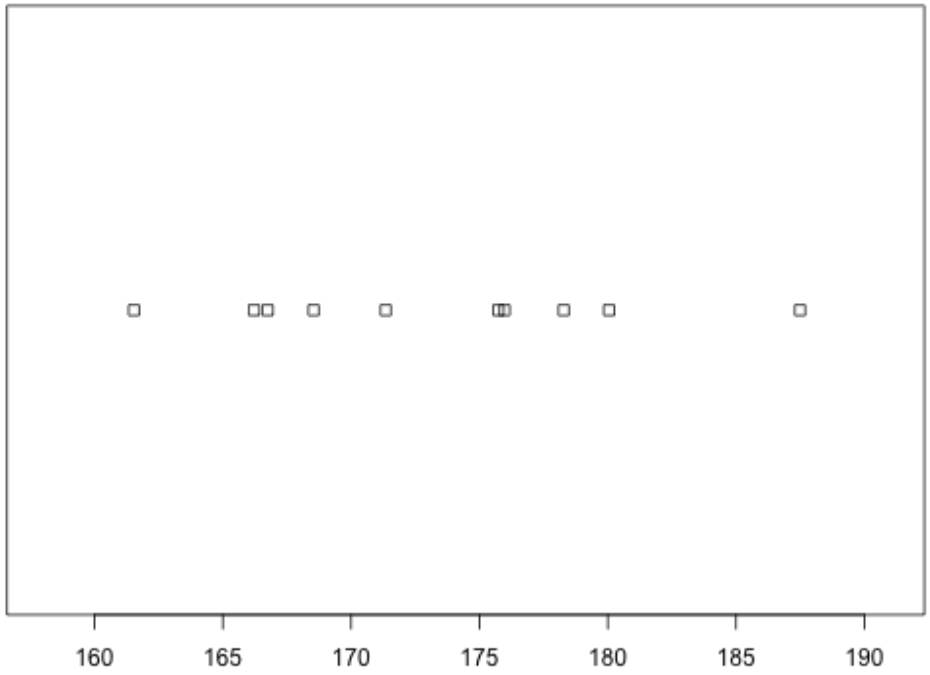
Comparaison des mesures de dispersion

	SD	MAD	IQR
3,5,4,8,7	2.07	2.97	3.00
3,5,4,8,7,2,11	3.15	2.97	4.00
3,5,4,8,7,2,161	59.06	2.97	4.00

Représentations graphiques

Tracé en bande ("stripchart")

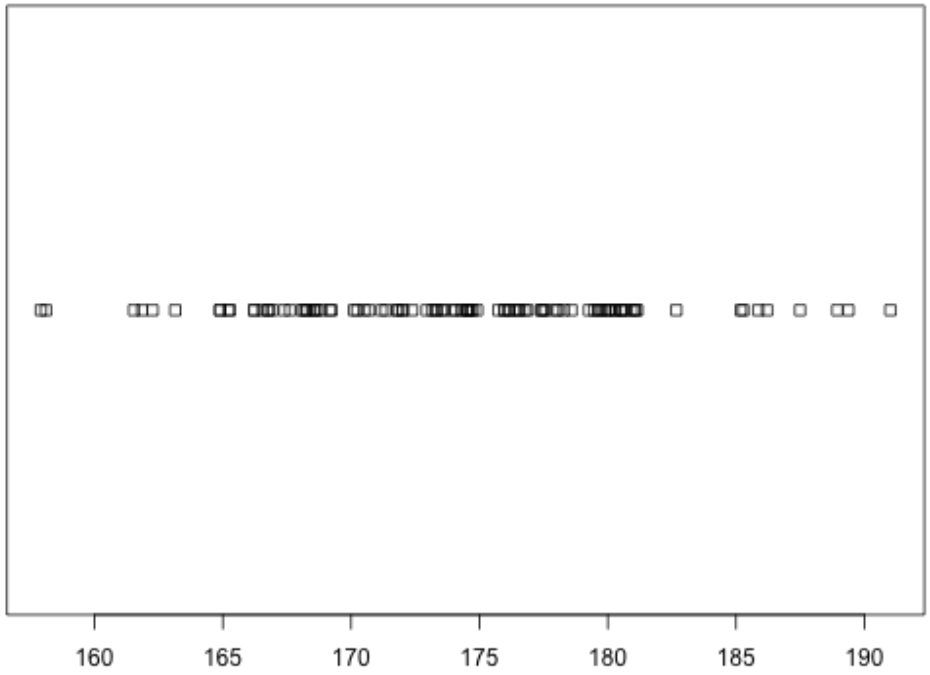
Taille de 10 hommes français:



- toutes les valeurs observées sont représentées
- le graphique donne une idée de leur **distribution**

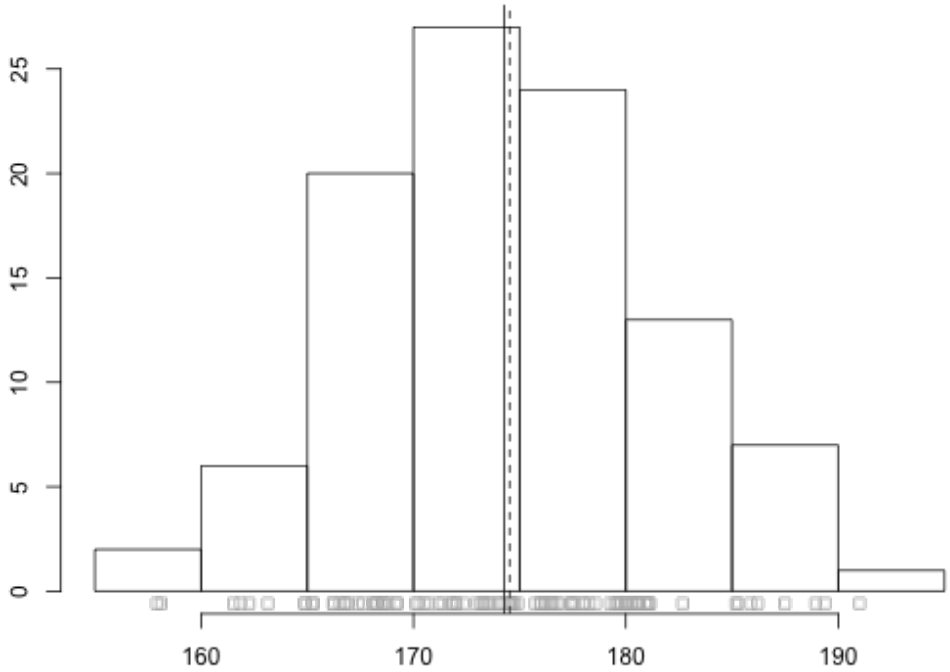
Limites du stripchart

Taille de 100 hommes français



- on ne voit plus grand chose !
- nécessité de regrouper les valeurs similaires pour **résumer** l'information

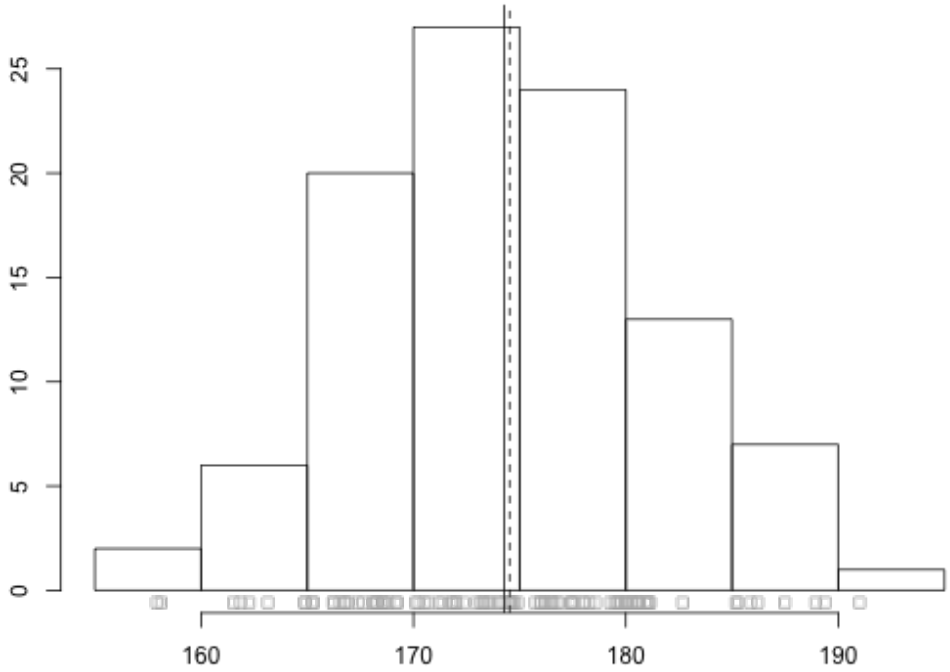
Histogramme

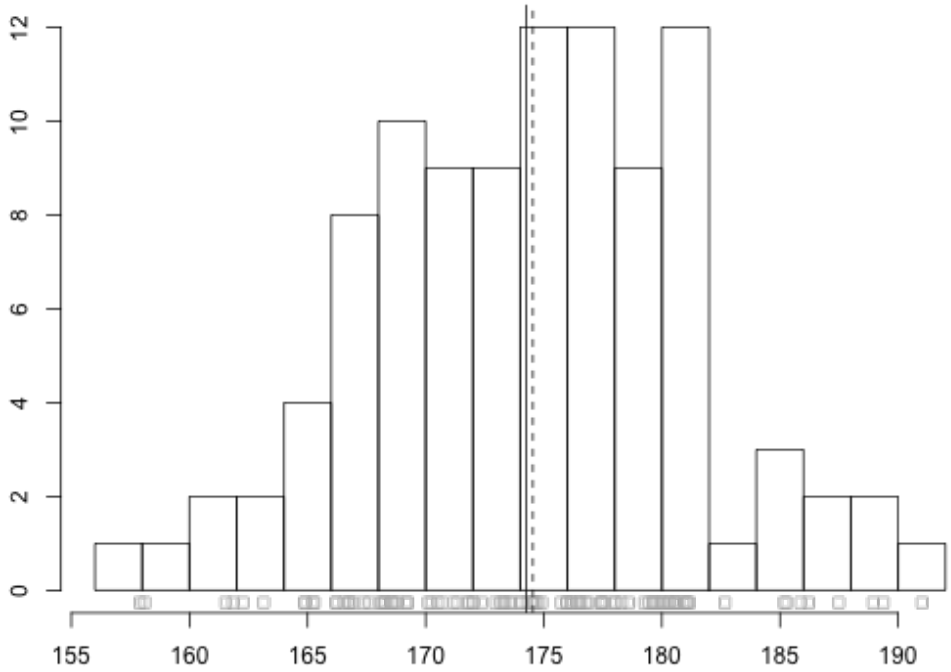


- les valeurs de l'échantillon sont regroupées en **classes**
- chaque classe est représentée par un rectangle
- la surface du rectangle est proportionnelle à la fréquence de la classe

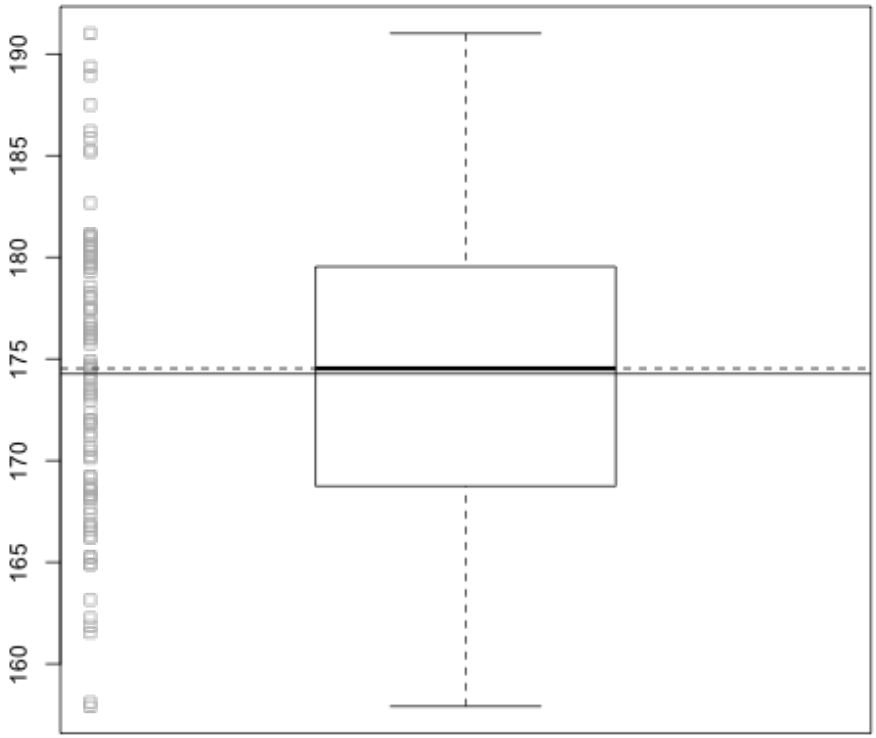
Histogramme

L'allure de l'histogramme peut dépendre du nombre de classes considéré



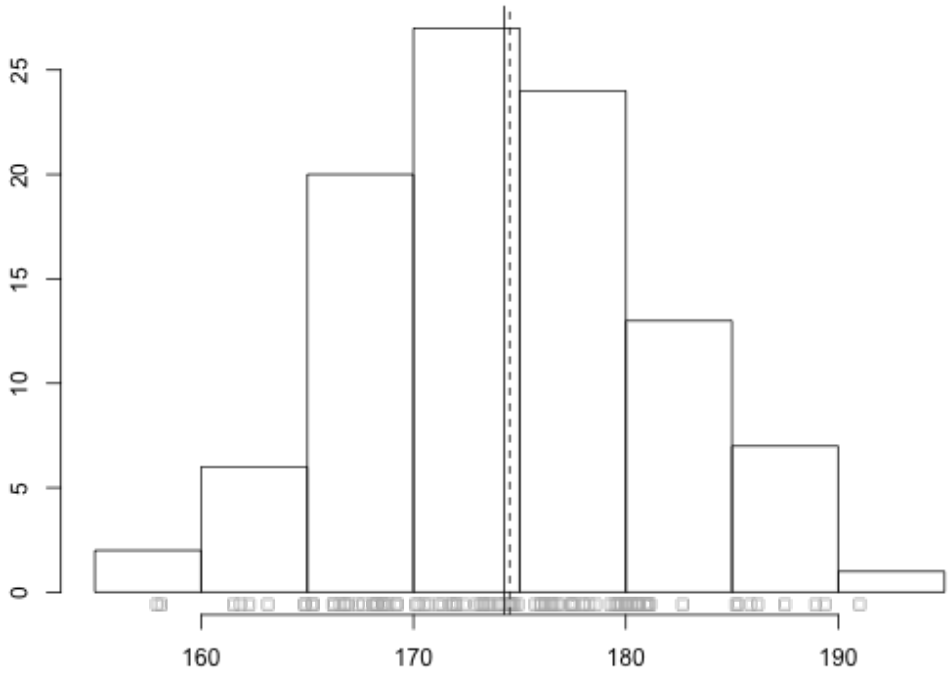


Boîte à moustaches ("boxplot")

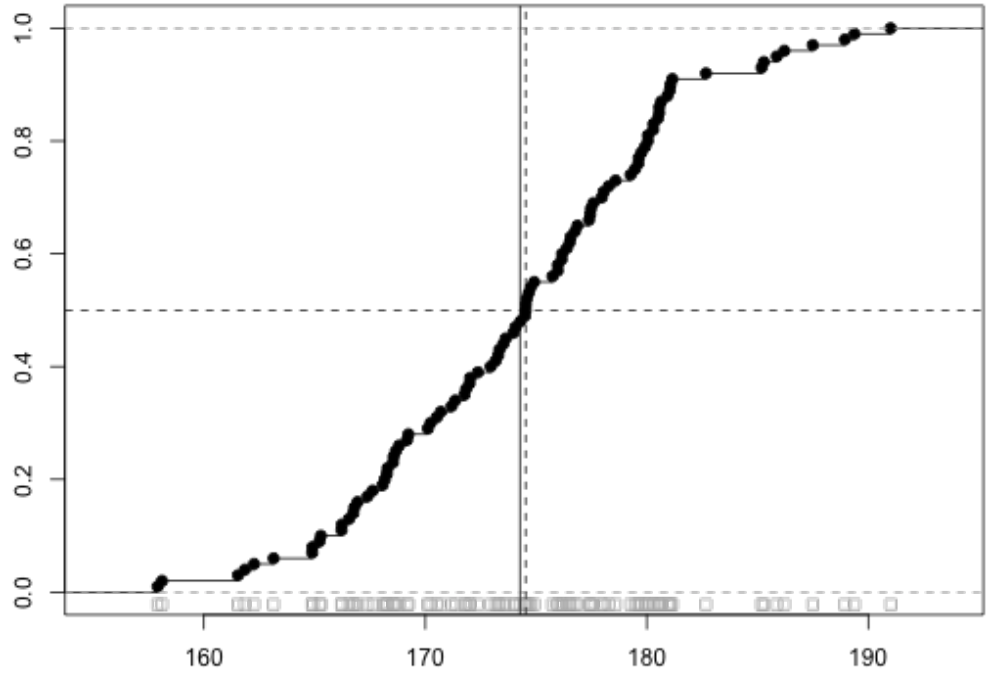


Histogramme et fonction de répartition empirique

Histogramme

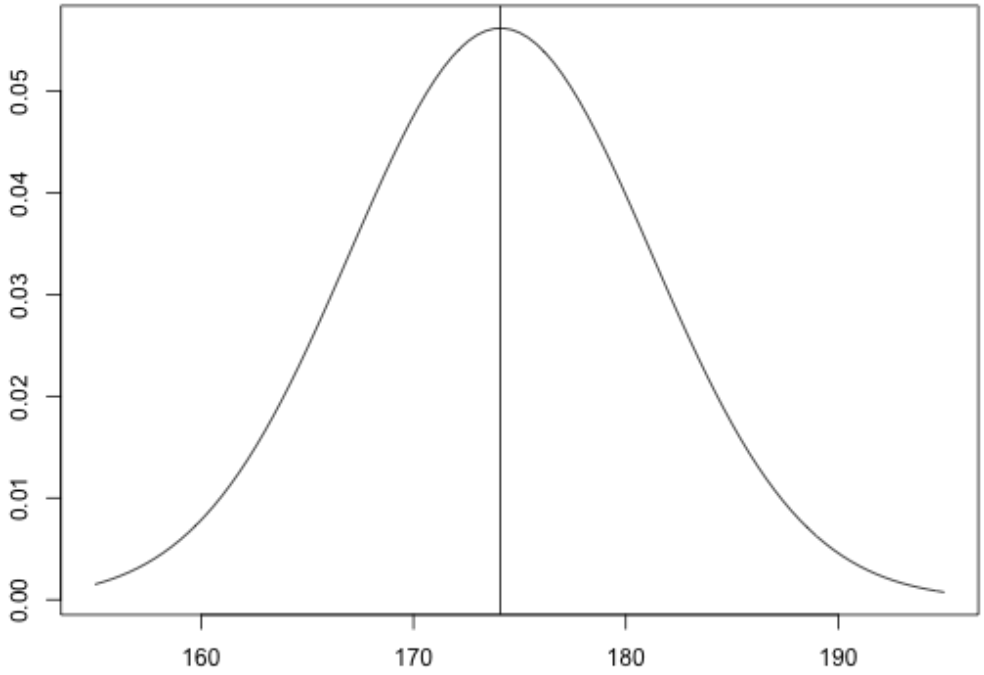


Fonction de répartition empirique

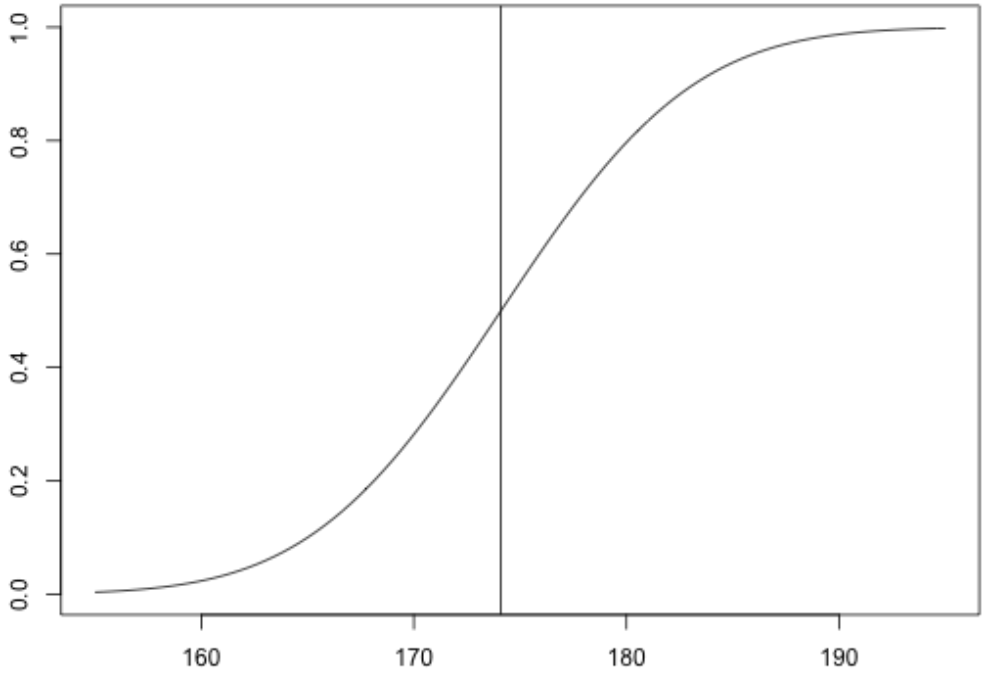


Loi normale

Densité

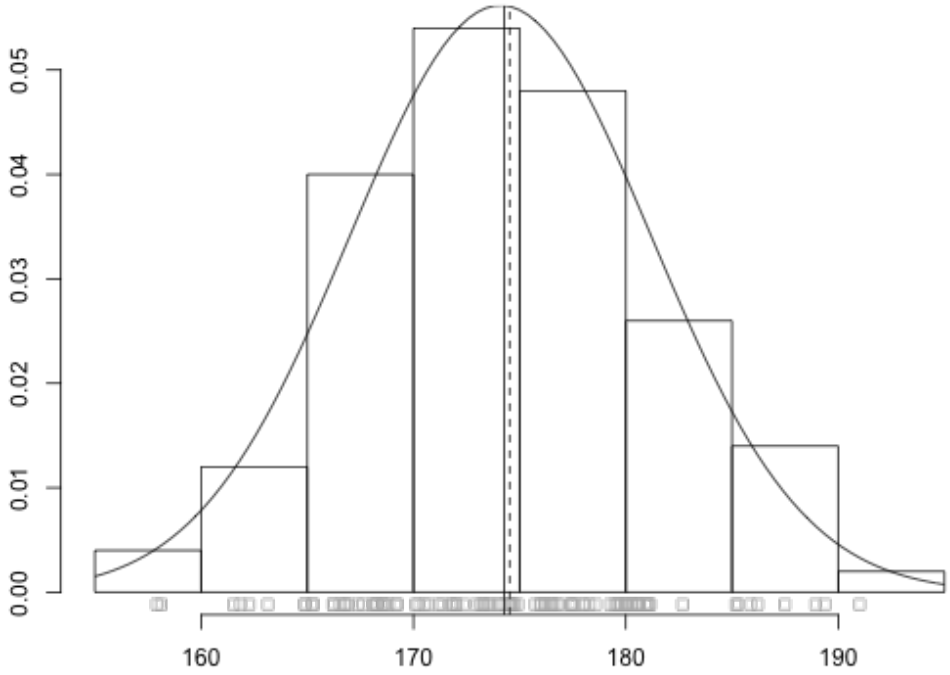


Fonction de répartition



Loi empirique et loi théorique

Densités



Fonctions de répartition