

Part I

Statistiques univariées

1 Statistiques descriptives

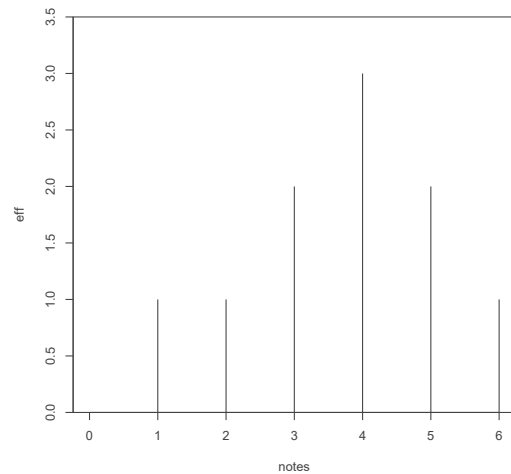
La statistique est une méthode scientifique qui consiste à organiser, analyser et interpréter des observations faites sur un ou plusieurs caractères (*poids, consommation, \dots*) des individus (*chèvres, machines, pièces manufacturées, \dots*) d'une population de taille N .

1.3 Variable discrète quantitative

Prenons l'exemple suivant d'une série de notes d'un examen noté sur dix points.

x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1
f_i	0.1	0.1	0.2	0.3	0.2	0.1

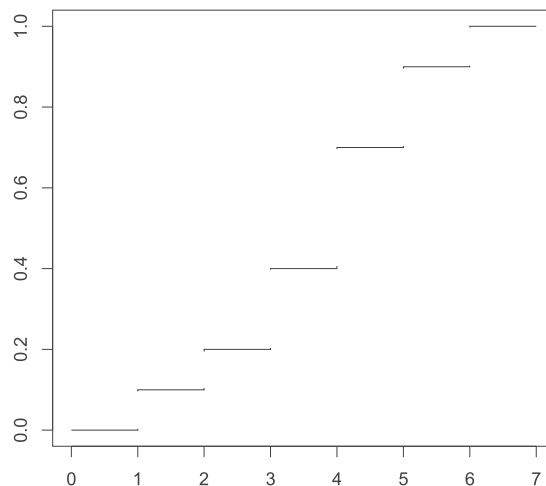
Il est à noter qu'il est totalement équivalent de parler en terme d'effectif n_i ou en terme de fréquences f_i .
On représentera graphiquement la série précédente par un diagramme en bâtons :



On pourra également tracer la fonction de répartition. Pour cela on ajoute la ligne des effectifs cumulés (*eff. c.*) et celle des fréquences cumulées (*f.c.*) au tableau précédent.

x_i	1	2	3	4	5	6
<i>eff. c.</i>	1	2	4	7	9	10
<i>f.c.</i>	0.1	0.2	0.4	0.7	0.9	1

Ces lignes permettent de répondre facilement à la question *combien d'élèves ont eu une note inférieure ou égale à 4*. Le graphe des fréquences cumulées en fonction des valeurs de X est appelé fonction de répartition.



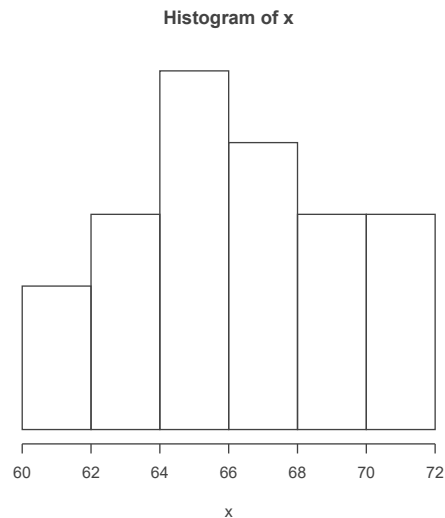
1.4 Variable continue quantitative

On a mesuré le poids de $N = 20$ personnes. Ici le caractère x , le poids, est une variable continue. On regroupe les mesures en classe et on obtient le tableau suivant :

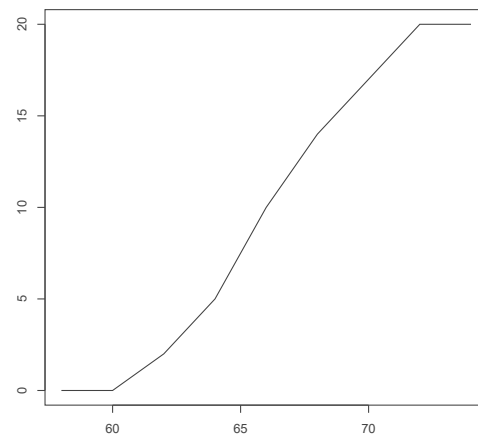
x_i	$[60,62[$	$[62,64[$	$[64,66[$	$[66,68[$	$[68,70[$	$[70,72[$
n_i	2	3	5	4	3	3

Il est également possible de s'exprimer en terme de fréquences. On représente ces données par un histogramme :

Remarque 1.1 *Sur l'histogramme, l'effectif se lit en terme d'aire des rectangles. Dans cet exemple (et c'est souvent le cas), les largeurs des classes sont les mêmes et donc les hauteurs des rectangles sont proportionnelles aux effectifs.*



Dans cette organisation en classe, on suppose qu'au sein d'une classe l'effectif est equiréparti, i.e. il y aura autant de personnes ayant un poids compris entre 60 et 61 kg qu'entre 61 et 62 kg. On a donc la fonction de répartition suivante :



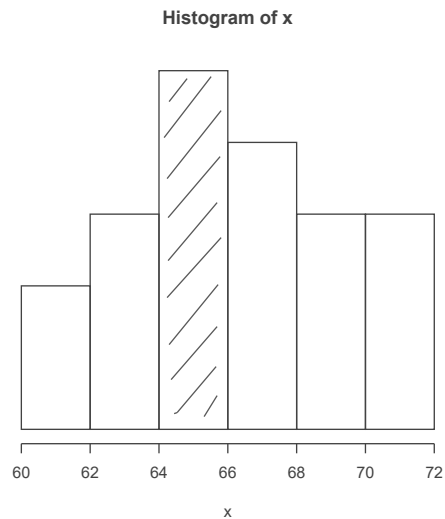
1.6 Caractéristiques de position

1.6.1 Le mode

Le mode, ou classe modale, d'une série statistique correspond intuitivement à la valeur du caractère (modalité.s) qui regroupe le plus grand effectif.

- Pour une variable discrète, c'est la modalité avec le plus grand effectif. Dans l'exemple 1.3, le mode est $x_4 = 4$ avec un effectif de 3.

- Pour une variable continue, c'est la classe $[a_i; a_{i+1}[$ dont la barre sur l'histogramme est la plus haute.



Autrement dit, la classe modale est la classe $[a_i; a_{i+1}[$ qui maximise la quantité

$$\frac{n_i}{l([a_i; a_{i+1}[)},$$

où l est la longueur de l'intervalle. Ici, la classe modale est $[64; 66[$.

1.6.2 Médiane, quartiles, quantiles

La médiane est une valeur du caractère qui sépare le classement en deux groupes de même taille. Le premier quartile (resp. le troisième) partage le classement dans les proportions 0.25 et 0.75 (resp. 0.75 et 0.25). On peut voir les quartiles comme les médianes des demi-groupes séparés par la médiane.

Plus généralement, le quantile d'ordre α , noté q_α est une valeur du caractère qui sépare le classement dans les proportions α et $(1 - \alpha)$.

Proposition 1.2 *La médiane est un antécédant de 0.5 par la fonction de répartition. Le quantile d'ordre α est un antécédant par la fonction de répartition de α .*

1.6.3 Moyenne

La moyenne est la valeur du caractère qu'auraient tous les individus de la population s'ils étaient identiques. Il s'agit du point d'équilibre (centre de gravité) du diagramme en batons ou de l'histogramme. En pratique, il s'agit de la moyenne des modalités pondérée par l'effectif :

$$m(x) = \bar{x} = \sum_i x_i \frac{n_i}{N}.$$

Dans l'exemple 1.3, la moyenne vaut 3.7.

Remarque 1.3 *Dans le cas d'une série continue regroupée en classes, il convient de remplacer les valeurs x_i du caractère par c_i , les valeurs du centre de la classe.*

Proposition 1.4 (Linéarité de la moyenne) *Soit a, b, c des réels, x, y des séries statistiques. On a*

$$m(ax + by + c) = am(x) + bm(y) + c.$$

1.7 Caractéristiques de dispersion

Il arrive que deux séries aient les mêmes caractéristiques de position mais soient totalement différentes. En effet, ces derniers indicateurs (hors quantiles) ne s'intéressent pas aux différences au sein de la population et donnent le stéréotype d'une population où tous les individus sont identiques. Pour aller plus loin dans l'étude statistique il nous faut donc introduire les notions suivantes concernant la dispersion des données.

1.7.1 Étendue

Il s'agit de l'écart entre la plus grande et la plus petite valeur prise par la série : $\max\{x_i\} - \min\{x_i\}$.

1.7.2 Variance et écart-type

La variance et l'écart-type sont des mesures des écarts à la moyenne : $(x_i - \bar{x})$. La variance pondère ces écarts au carré :

$$Var(x) = \sum_i (x_i - \bar{x})^2 \frac{n_i}{N}.$$

La variance est donc très influencée par les valeurs éloignées de la moyenne.

Proposition 1.5 *En pratique, on pourra calculer*

$$\text{Var}(x) = m(x^2) - (\bar{x})^2,$$

où

$$m(x^2) = \sum_i x_i^2 \frac{n_i}{N}.$$

Le résultat se vérifie en développant et en factorisant dans la formule initiale de la variance.

On définit l'écart-type σ par la racine carrée de la variance : $\sigma = \sqrt{Var(X)}$. Ainsi, l'écart-type renvoie à la distance pondérée entre les valeurs de la série et la moyenne. Il en découle la propriété suivante.

Proposition 1.6 *Soit a, b des réels. On a*

$$\sigma_{ax+b} = |a|\sigma_x, \quad \text{Var}(ax+b) = a^2\text{Var}(x).$$

- (i) Tracer un rectangle qui s'étend du premier quartile au troisième.
- (ii) Séparer ce rectangle en deux à la hauteur de la médiane. On obtient alors une boîte.
- (iii) On complète ce rectangles par deux segments. Pour cela, on calcule

$$a = q_{0.25} - 1.5IQ \quad \text{et} \quad b = q_{0.75} + 1.5IQ,$$

avec la distance inter-quartile

$$IQ = q_{0.75} - q_{0.25}.$$

On repère les valeurs :

$$x_a = \min\{x_i : x_i \geq a\} \quad \text{et} \quad x_b = \max\{x_i : x_i \leq b\}.$$

Ces valeurs sont appelées *valeurs adjacentes*. On relie ces valeurs aux cotés de la boîte.

- (iv) Les valeurs qui ne sont pas comprises entre les valeurs adjacentes sont représentées par des points et sont appelées *valeurs extrêmes*.

