Mesures de tendance centrale

Nous étudierons quelques mesures de tendance centrale. Notre but sera d'évaluer de façon formelle et quantitative une valeur qui soit, dans un sens ou dans un autre, "représentative" de la distribution au complet. Introduisons d'abord les mesures que nous allons étudier. Le calcul de ces valeurs se fera de façon différente selon les particularités de la variable étudiée, alors nous présenterons chaque cas séparèment.

Définition 1. *Le mode est la valeur, ou classe de valeurs, que la variable prend le plus souvent.*

Définition 2. *La moyenne est la valeur, ou la valeur approximative, telle que si la variable serait constante pour chaque individu la somme de ses valeurs resterait la même.*

Définition 3. *La médiane est la valeur, ou la valeur approximative, telle que la variable est inférieure pour la moitié des individus de l'échantillon et supérieure pour l'autre moitié.*

Il est assez facile à voir qu'il serait difficile de calculer des mesures (centrale ou de dispersion) pour des variables qualitatives. Dans certains cas nous pourrons appliquer certaines de nos méthodes aux variables qualitatives et nous en parlerons d'avantage en classe, mais elles ne seront pas notre objectif principal pour aujourd'hui et la semaine prochaine.

Pour les variables quantitatives, nous devons regarder les trois cas suivants séparèment: le cas d'une variable dont les données ne sont pas organisées en classes et peuvent prendre un grand nombre de valeurs diffèrentes, le cas d'une variable dont les données ne sont pas organisées en classes mais ne peuvent prendre qu'un petit nombre de valeurs et le cas d'une variable dont les données sont organisées en classes.

À remarquer que nous donnerons des détails et des explications des formules ci-dessous en classe, les notes de cours ne sont qu'un résumé de la matière.

Calcul dans le cas d'une variable quantitative continue, ou discrète mais pouvant prendre un nombre extrêmement grand de valeurs

Définissons d'abord nos quantités: pour une variable telle que décrite plus haut nomée X sur un échantillon de taille N, nous appelerons les valeurs que la variable a pris mises en ordre croîssant, possiblement avec répétitions, $x_1, ..., x_N$. Les fréquences ne nous intéresseront pas particulièrement car la plupart de valeurs n'apparaîtront qu'une fois.

Le mode ne nous intéressera pas particulièrement dans ce cas-ci. La moyenne, notée μ_X ou $\overline{X}$, est calculée comme suit:

$$\mu_X = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

La médiane, notée $Md(X)$, sera calculée comme suit:

- Si N est pair, $Md(X) = \frac{x_{N/2} + x_{N/2+1}}{2}$
- Sinon, $Md(x) = x_{(N+1)/2}$

Exemple 1. *On s'intéresse au temps que les utilisateurs d'un site internet passent dessus. Un recensement rapide des 15 derniers utilisateurs nous donne les longueurs suivantes pour les visites, en secondes:*

339 296 307 298 321

306 276 278 317 268

271 278 338 288 295

Ici, nous avons des données non-triées qui peuvent prendre n'importe quelle valeur entière. Nous pouvons calculer le mode, et nous remarquons que le nombre 278 revient deux fois, mais il s'agit simplement de l'effet du hasard et ne nous donne aucune information utile sur notre distribution.

Plus utile, ici, serait la moyenne. Nous appliquons notre formule directement, ce qui nous donne:

$$\mu_X = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{339 + 296 + ... + 288 + 295}{15} = 298,4$$

Pour la médiane, nous nous confrontons à un problème: nous ne pouvons pas appliquer directement notre formule car les nombres ne sont pas en ordre croissant. Commençons donc par les placer en ordre croissant:

268 271 276 278 278

288 295 296 298 306

307 317 321 338 339

Maintenant, nous pouvons évaluer la médiane comme étant la $\frac{15+1}{2} = 8$ème valeur: 296.

Calcul dans le cas d'une variable quantitative discrète pouvant prendre un nombre très limité de valeurs

Avant de continuer, nous aurons besoin de deux définitions plus précises que celles qu'on a donné au premier cours:

Définition 4. *Soit X une variable statistique quantitative mesurée sur un échantillon de taille N pouvant prendre les valeurs $x_1, ..., x_k$ dont les fréquences absolues sont $n_1, ..., n_k$.*

La fréquence relative de chaque valeur sera $f_1 = \frac{n_1}{N}, ..., f_k = \frac{n_k}{N}$, f_i représentant la proportion des individus pour lesquels $X = x_i$.

La fréquence cumulative de chaque valeur sera $F_1 = f_1, F_2 = F_1 + f_2, ..., F_k = F_{k-1} + f_k$, F_i représentant la proportion des individus pour lesquels $X \leq x_i$.

Le ou les mode(s) de X sera la ou les valeur(s) x_i dont la fréquence est la plus élevée. À noter qu'il est parfaitement possible, fréquent même, d'avoir plusiers modes. On dit alors que la variable est multimodale.

La moyenne, notée μ_X ou $\overline{X}$, est calculée comme suit:

$$\mu_X = \frac{x_1 n_1 + ... + x_k n_k}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \sum_{i=1}^{k} x_i f_i$$

La médiane, notée $Md(X)$, sera calculée comme suit:

- Si il existe un i tel que $F_i = 0,5$, alors $Md(x) = \frac{x_i + x_{i+1}}{2}$
- Sinon, $Md(x) = x_i$ tel que x_i est la plus petite modalité de X pour laquelle $F_i > 0,5$

Exemple 2. *Un recensement des 100 nouveaux élèves d'un collège inscrits en sciences de la nature révèle que 5 ont 16 ans, 45 ont 17 ans, 30 en ont 18, 10 en ont 19, 6 en ont 20, trois ont 21 ans et un 22 ans.*

Le mode étant la valeur ayant la plus haute fréquence, il est unique et s'agit de 17 ans.

Pour la moyenne, nous pouvons calculer directement:

$$\mu_X = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \frac{16 \cdot 5 + 17 \cdot 45 + ... + 22 \cdot 1}{100} = 17,8$$

Pour la médiane, nous allons regarder les fréquences cumulatives. La fréquence cumulative de 16 est $\frac{5}{100}$, celle de 17 est $\frac{5+45}{100} = 0,5$. Par notre méthode de calcul, la médiane sera la moyenne de 17 et de la prochaine valeur possible, soit $\frac{17+18}{2} = 17,5$.

Calcul dans le cas d'une variable quantitative dont les valeurs sont regroupées en classes

Soit X une variable quantitative prélevée sur un échantillon de taille N et dont les valeurs ont été regroupées en classes qui sont des intervalles de valeurs réelles.

Nous appelerons les intervalles dans lesquelles on a classé nos données $[b_1, b_2[, [b_2, b_3[, ..., [b_k, b_{k+1}[$ et m_i la valeur au milieu de $[b_i, b_{i+1}[$. On se rapelle que $m_i = \frac{b_i + b_{i+1}}{2}$. La fréquence absolue de chaque classe sera $n_1, ..., n_k$, la fréquence relative de chaque classe sera $f_1, ..., f_k$ et celle cumulative $F_1, ..., F_k$.

Le ou les mode(s) de X sera la ou les classe(s) x_i dont la fréquence est la plus élevée. Il est fréquent que la variable soit multimodale.

La moyenne, notée μ_X ou $\overline{X}$, est calculée comme suit:

$$\mu_X = \frac{m1n1 + ... + m_k n_k}{N} = \frac{\sum_{i=1}^{k} m_i n_i}{N} = \sum_{i=1}^{k} m_i f_i$$

La médiane, notée $Md(X)$, sera calculée comme suit:

- Si il existe un i tel que $F_i = 0,5$, alors $Md(x) = b_{i+1}$
- Sinon, supposons que la première classe telle que la fréquence cumulative soit supérieure à 0,5 est $[b_i, b_{i+1}[$. Alors la formule de la médiane sera la suivante:
$$Md(X) = b_i + \left(\frac{0,5 - F_{i-1}}{f_i}\right)(b_{i+1} - b_i)$$

Exemple 3. *On regarde le montant des achats dans un supermarché au cours d'une journée. Il y a eu 400 clients, et le montant des achats (en dollars) se distribuent comme suit:*

Classe	*Milieu*	*Fréquence absolue*
$[0, 20[$	*10*	*40*
$[20, 40[$	*30*	*100*
$[40, 60[$	*50*	*100*
$[60, 80[$	*70*	*60*
$[80, 100[$	*90*	*40*
$[100, 120[$	*110*	*30*
$[120, 140[$	*130*	*20*
$[140, 160[$	*150*	*10*

La distribution est bimodale: $[20, 40[$ *et* $[40, 60[$ *sont les deux modes.*

Le calcul de la moyenne est direct:

$$\mu_X = \frac{\sum_{i=1}^{k} m_i n_i}{N} = \frac{10 \cdot 40 + 30 \cdot 100 + ... + 130 \cdot 20 + 150 \cdot 10}{400} = 59\$$$

En ce qui a trait à la médiane, nous devons travailler un peu plus.

$$f_1 = F_1 = \frac{40}{400} = 0,1$$

$$f_2 = \frac{100}{400} = 0,25,\ F_2 = F_1 + f_2 = 0,35$$

$$f_3 = \frac{100}{400} = 0,25,\ F_3 = F_2 + f_3 = 0,6$$

La médiane doit donc être dans la troisième catégorie, $[40, 60[$. *Calculons donc avec* $i = 3$*:*

$$Md(X) = b_i + \left(\frac{0,5 - F_{i-1}}{f_i}\right)(b_{i+1} - b_i) = b_3 + \left(\frac{0,5 - F_2}{f_3}\right)(b_4 - b_3)$$

$$= 40 + \left(\frac{0,5 - 0,35}{0,25}\right)(60 - 40) = 52$$

Exercices

1. Les températures du 16 septembres pour les dix dernières années ont été:

   $$16,7 \quad 20,1 \quad 19,2 \quad 13,3 \quad 15,1 \quad 18,2 \quad 21,4 \quad 19,8 \quad 16,7 \quad 19,5$$

   En degrés Fahrenheit, ceci équivaut à:

   $$62,1 \quad 68,2 \quad 66,6 \quad 55,9 \quad 59,2 \quad 64,8 \quad 70,5 \quad 67,6 \quad 62,1 \quad 67,1$$

   Calculez la moyenne et la médiane en Celsius et Fahrenheit.

2. Un sondage rapide auprès de 20 employés d'une entreprise montre que 8 n'ont pas d'enfants, 7 en ont un, 3 en ont deux, un en a trois et un quatre. Évaluez le mode, la moyenne et la médiane des enfants en utilisant la méthode appropriée.

3. Une région possède une municipalité ayant entre 100000 et 1000000 d'habitants, trois entre 10000 et 100000, onze entre 1000 et 10000 et cent-cinq entre 0 et 1000. Approximez la moyenne et la médiane des populations de localités à partir de ces chiffres, puis expliquez pourquoi ces résultats ne sont pas très utiles.

Réponses

2. $\mu_X = 18C^\circ = 64,41F^\circ$, $Md(X) = 18,7C^\circ = 65,7F^\circ$
2. $Mo(X) = 0$, $\mu_X = 1$, $Md(X) = 1$
3. $\mu_X = 6900$, $Md(X) = 571$