Mesures de dispersion

Notre but aujourd'hui sera d'essayer de décrire la distribution d'une variable autour des valeurs centrales déjà introduites.

Définition 1. *La variance, notée σ^2, est la moyenne, possiblement approximative, du carré de la différence entre les valeurs que la variable prend et leur moyenne. Nous verrons plus loin comment calculer ceci dans chaque cas mais la notion de carré des différences est une que nous réverrons lorsqu'on parlera de régression linéaire. Pour une population de taille N, si la variable X prend les valeurs exactes $x_1, x_2, ..., x_k$ avec fréquences absolues $n_1, n_2, ..., n_k$ et moyenne $\overline{X}$ la variance est:*

$$\sigma^2(X) = \frac{(x_1 - \overline{X})^2 n_1 + (x_2 - \overline{X})^2 n_2 + ... + (x_k - \overline{X})^2 n_k}{N} = \frac{\sum_{i=1}^{k} (x_i - \overline{X})^2 n_i}{N}$$

Dans le cas d'une variable dont les valeurs sont données en classes de points milieux $m_1, m_2, ..., m_k$ et fréquences absolues $n_1, n_2, ..., n_k$ la variance est:

$$\sigma^2(X) = \frac{(m_1 - \overline{X})^2 n_1 + (m_2 - \overline{X})^2 n_2 + ... + (m_k - \overline{X})^2 n_k}{N} = \frac{\sum_{i=1}^{k} (m_i - \overline{X})^2 n_i}{N}$$

Définition 2. *Deux mesures dérivées de la variance sont l'écart-type $\sigma = \sqrt{\sigma^2}$ et le coefficient de variation $CV = 100 \cdot \frac{\sigma}{\mu}\%$. L'écart-type est simplement la racine carrée de la variance et le coefficient de variation est la proportion en pourcentage entre l'écart-type et la moyenne.*

Nous continuerons les exemples de la semaine passée pour éviter les calculs inutiles.

Exemple 1. *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. Nous avions calculé la semaine passée que la moyenne est de 17,8 ans et la médiane de 17,5.*

Calculons donc la variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{X})^2 n_i}{N} = \frac{5(16 - 17,8)^2 + 45(17 - 17,8)^2 + ... + (22 - 17,8)^2}{100} = 1,38$$

À partir de ceci, nous pouvons calculer nos autres mesures de dispersion:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1,38} \approx 1,175$$

$$CV = 100 \cdot \frac{\sigma}{\overline{X}}\% \approx 100 \cdot \frac{1,175}{17,8}\% \approx 6,6\%$$

Exemple 2. *On regarde le montant des achats dans un supermarché au cours d'une journée présenté précedemment. Il y a eu 400 clients, l'achat moyen comme calculé la dernière fois se montait à 59$.*

Calculons d'abord la variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(m_i - \overline{X})^2 n_i}{N} = \frac{40(10-59)^2 + ... + 10(150-59)^2}{400} = 1239$$

À partir de ceci, nous pouvons calculer nos autres mesures de dispersion:

$$\sigma = \sqrt{\sigma^2} = \sqrt{1239} \approx 35,2$$

$$CV = 100 \cdot \frac{\sigma}{\overline{X}}\% \approx 100 \cdot \frac{35,2}{59}\% \approx 59,66\%$$

À remarquer que le coefficient de variation est beaucoup plus grand ici que précedemment. On peut d'ailleurs remarquer ceci en regardant la distribution en soi...

Définition 3. *Lorsqu'on essaie de mesurer la valeur de X sur une population à partir d'un échantillon plus petit et que $\overline{X}$ a été calculée à partir de l'échantillon, on doit utiliser la variance échantillonale, notée $s^2(X)$. La formule est la même sauf que le dénominateur utilisé sera $N-1$ au lieu de N.*

L'écart-type échantillonal est défini de façon analogue par $s = \sqrt{s^2}$.

Remarque. *La variance et l'écart-type échantillonals sont utilisés lorsque l'échantillon est utilisé pour généraliser à une population au complet et que la moyenne a été calculée sur l'échantillon. Si on regarde un sous-groupe d'un plus grand groupe* mais que ce sous-groupe est celui qui nous intéresse*, alors ce sous-groupe est notre population au complet et nous n'utiliserons pas les formules échantillonales.*

Nous n'utilisons pas non plus les formules échantillonales si la moyenne a été calculée sur la population au complet et non sur l'échantillon, mais cette situation ne survient que très rarement.

Exemple 3. *Reprenons l'exemple du supermarché mais en essayant de généraliser les nombres pour une journée aux achats faits dans ce supermarché en général. Nous avions trouvé une moyenne de 59\$. Calculons d'abord la variance:*

$$s^2 = \frac{\sum_{i=1}^{N}(m_i - \overline{X})^2 n_i}{N-1} = \frac{40(10-59)^2 + ... + 10(150-59)^2}{400-1} \approx 1242,105$$

À partir de ceci, nous pouvons calculer nos autres mesures de dispersion:

$$s = \sqrt{s^2} \approx 35,24$$

$$CV = 100 \cdot \frac{s}{\overline{X}}\% \approx 100 \cdot \frac{35,24}{59}\% \approx 59,73\%$$

À remarquer que les résultats ne sont pas trop différents car il n'y a pas une grande différence entre 400 et 399. Plus l'échantillon est petit, plus la différence sera significative.

Cote z

Définition 4. *Soit X une variable statistique de moyenne μ et écart-type σ (ou s dans le cas d'un échantillon).*

La cote z d'une valeur spécifique x de X est $z = \frac{x-\mu}{\sigma}$.

Remarque. *La cote z approxime la position d'une valeur spécifique relativement au reste de la population. Nous discuterons la cote z beaucoup plus en détail dans le cas d'une variable distribuée normalement.*

Exercices

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

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

   Estimez la variance, l'écart-type et le coefficient de variation de la température pour le 16 septembre.

2. Un sondage rapide auprès des 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 la variance, l'écart-type et le coefficient de variation du nombre d'enfants des employés de cette compagnie.

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 variance, l'écart-type et le coefficient de variation des population des municipalités de cette région.

4. Quelle serait la cote z d'une valeur étant supérieure à la moyenne de une unité pour les variables des numéros 1, 2 et 3? Expliquez la différence.

Réponses

1. $s^2 = 6,31\overline{3}$, $s \approx 2,513$, $CV \approx 13,96\%$
2. $\sigma^2 = 1,2$, $\sigma \approx 1,095$, $CV \approx 109,544\%$, $z = \frac{2-1}{\sqrt{1,2}} \approx 0,913$
3. $\sigma^2 \approx 2,551 \times 10^9$, $\sigma \approx 50510$, $CV \approx 732\%$