Fiche TD avec le logiciel R : tdr205

Quelques paramètres décrivant la variabilité

A.B. Dufour & J.R. Lobry & M. Royer

Cette fiche comprend des exercices portant sur les principaux paramètres de dispersion : l'étendue, la variance, l'écart-type et le coefficient de variation.

Exercice 1

Les données de cet exercice ont été extraites de la fiche de cours `bs3 : Pratique des tests élémentaires`. Dans 3 groupes de TD, les notes de contrôle continu sont les suivantes :

```
groupe1 <- c(14.9, 12, 9.5, 7.3, 8.4, 9.8, 11, 13.8, 14.3, 5, 4.4,
    14.3, 13.7, 18, 12.4)
groupe2 <- c(10.9, 10.1, 10, 12.2, 10, 11.1, 10.3, 9.5, 9.6, 10,
    10.9, 11.2)
groupe3 <- c(13, 12.1, 8.7, 10.9, 12.7, 9.5, 10.5, 12.2, 16, 10.3,
    9.6, 10.9, 7.3, 9.8)
```

1) Enregistrer chaque vecteur de notes dans R puis calculer la moyenne de chaque groupe.

   Que constate-t-on ? Peut-on dire que les enseignants ont noté de la même façon ?

2) Construire les histogrammes des notes des trois groupes - Ne pas oublier d'harmoniser les échelles pour rendre les représentations graphiques comparables (*cf* figure 1).
   Peut-on observer une différence de notation entre les groupes ?

3) On regroupe les trois vecteurs dans un seul appelé `notes` et on construit une variable qualitative `groupes` à trois modalités : 1, 2 et 3.

```
notes <- c(groupe1, groupe2, groupe3)
groupes <- rep(c(1, 2, 3), c(length(groupe1), length(groupe2), length(groupe3)))
groupes <- factor(groupes)
```

   a) Réaliser la représentation graphique montrant les trois boîtes à moustaches des notes selon les groupes.

   b) Utiliser maintenant la commande `stripchart` pour représenter les notes au sein de chaque groupe.

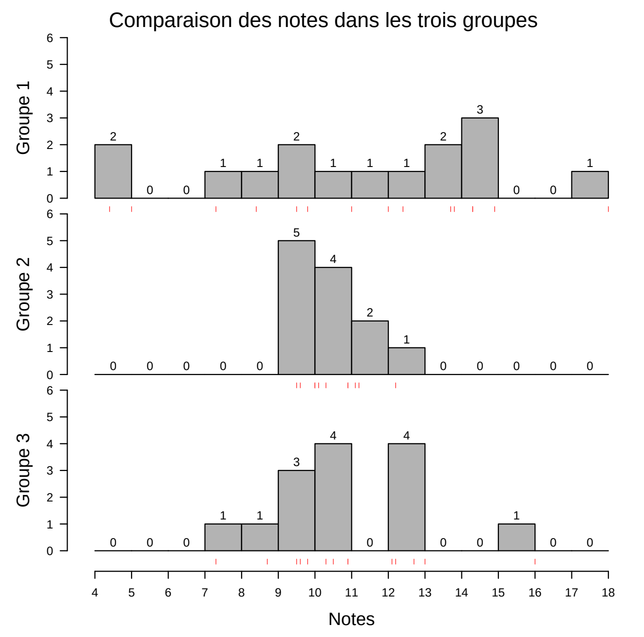

FIG. 1 – Pour permettre une comparaison visuelle les histogrammes doivent être disposés les uns en dessous des autres avec une échelle commune.

c) Peut-on observer une différence de notation entre les groupes ?

4) La dispersion au sein d'un groupe peut être caractérisée par les deux paramètres que sont le minimum et le maximum. Ces informations sont présentes dans le résumé statistique. On peut également les calculer à l'aide de la fonction `range(x)`. La différence $max - min$ s'appelle l'étendue. Calculer l'**étendue** des notes de chaque groupe et commenter.

5) L'étendue est très sensible aux valeurs extrêmes. Ce n'est pas un bon paramètre de dispersion. On lui préfère la **variance empirique**, notée s^2. La variance est la moyenne des carrés des écarts à la moyenne.

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

Elle s'exprime par le carré de l'unité de mesure de la variable.
Calculer la variance empirique de chaque groupe en utilisant les commandes `sum` et `length`.

6) Commenter la variance n'est pas simple. On définit l'**écart-type** (noté s) comme étant la racine de la variance, qui lui, s'exprime dans les unités de la variable.

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

Calculer les écarts type de chaque groupe en utilisant la commande `sqrt`.

7) Analyser les relations entre les variances (ou les écarts type) et les représentations graphiques. Conclure.
8) Utiliser la commande `var` de R pour calculer la variance de la note de chaque groupe. Retrouve-t-on les mêmes valeurs qu'à la question précédente ?
9) En fait, R calcule la quantité suivante :

$$\texttt{var} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Il s'agit de la variance estimée de la population à partir d'un échantillon. Par quel coefficient doit-on multiplier la valeur obtenue par la commande `var` pour obtenir le même résultat qu'à la question 5) ?

Exercice 2

Il existe un autre paramètre décrivant la variabilité : le coefficient de variation :

$$cv = \frac{ecart - type}{moyenne} \quad \text{ou} \quad cv = \frac{s}{\overline{x}}$$

Il présente l'avantage d'être sans unité. Il permet deux types d'étude :

i- comparer la variabilité de plusieurs variables quantitatives mesurées au sein d'un même échantillon

ii- comparer la variabilité de plusieurs échantillons pour une même variable.

Pour bien comprendre ce coefficient de variation, on va partir des données extraites de l'enquête longitudinale française (Sempé *et al.*, "Auxologie, Méthode et Séquence") [1].

1) On donne la moyenne et l'écart-type de quatre mesures anthropométriques réalisées sur un même groupe de garçons âgés de 14 ans.

variable	moyenne	écart-type
taille	159.9	8.10
poids	47.7	8.65
périmètre thoracique	77.3	6
diamètre biacromial	34.3	2.3

Quelle est, de ces quatre mesures, celle qui présente le plus de variabilité à l'âge de 14 ans ?

2) Chez les filles, l'âge de 7 ans correspond au premier pic pubertaire. On s'attend donc à ce qu'une différence de taille entre filles et garçons commence à apparaître. Qu'en est-il ?

taille	moyenne	écart-type
garçons	119.7	4.9
filles	118.2	4.6

3) Chez les garçons, la puberté a lieu entre 12 et 17 ans. A quel âge peut-on observer la plus grande variabilité de la taille ? La question a un sens dans la mesure où là encore, ce sont les mêmes garçons mesurés à des âges différentes.

âge	moyenne	écart-type
12	145.8	6.1
13	152.5	7.3
14	159.9	8.1
15	166.7	7.7
16	171.1	6.4
17	173.5	6

Exercice 3

On connait la taille et le poids de 21 adultes atteints de trisomie 21.

```
poidsT21 <- c(65.5, 64.5, 54.5, 46.9, 72.7, 48.2, 53.1, 61.4, 66.9,
    69.7, 76.1, 71.9, 92.6, 55.4, 52.7, 62.3, 66.4, 61.5, 68.2,
    62.3, 65.1)
tailleT21 <- c(1.55, 1.55, 1.58, 1.49, 1.59, 1.53, 1.46, 1.48, 1.41,
    1.68, 1.69, 1.58, 1.56, 1.58, 1.39, 1.41, 1.35, 1.37, 1.41,
    1.55, 1.48)
```

1) Calculer l'indice de masse corporelle de ces adultes.
2) Quelle est, parmi ces trois variables, taille, poids et indice de masse corporelle, celle qui admet la plus grande variabilité ?
3) On donne les moyennes et les écarts type de la taille et du poids d'adultes issus d'une population ne présentant aucune pathologie.

Variable	moyenne	écart-type
taille	175.3	6.0
poids	65.0	7.1

   a) Réaliser une comparaison "descriptive" de la variabilité de la taille et du poids pour les deux groupes.
   b) Peut-on calculer la moyenne et l'écart-type de l'indice de masse corporelle à partir du tableau précédent ?

4) Représenter graphiquement les données et conclure.

Remarque

Nous avons défini la variance descriptive et le coefficient de variation qui ne sont pas des fonctions pré-existantes dans R . Il est tout à fait possible de les "programmer".
`x` représente le vecteur (variable quantitative) sur lequel nous voulons calculer la variance descriptive `vardes` et le coefficient de variation `cvariation`.

```
vardes <- function(x) {
    res <- var(x) * (length(x) - 1)/length(x)
    return(res)
}
cvariation <- function(x) {
    res <- sqrt(vardes(x))/mean(x)
    return(res)
}
```

Pour ce faire,

- ouvrir dans le menu `Fichier` l'item `Nouveau script`. Apparaît une fenêtre `R-Editor`.
- écrire les deux fonctions
- retourner dans le menu `Fichier`, `Sauver sous` et donner un nom au fichier en prenant soin de bien le mettre dans le répertoire de travail (par exemple `dispersions.R`).
- fermer la fenêtre `R-Editor`
- retourner dans le menu `Fichier` et sourcer le fichier ainsi créé. Si aucune erreur de programmation n'a été commise, le prompt de commande apparaît.
- exécuter les scripts :

```
vardes(poidsT21)
cvariation(poidsT21)
```

Références

[1] M. Sempé, M.P. Roy-Pernot, and G. Pedron. *Auxologie, Méthode et Séquences*. Editions Théraplix, Paris, 1979.