Statistiques descriptives et Introduction à R

Partie 1 : Statistiques univariées

Edmond Sanou

Université d'Evry Val d'Essonne

Mars 2023

Sommaire

Sommaire

Sommaire

Définition

Comme pour un caractère qualitatif, on appelle **distribution** d'un caractère quantitatif X, sa répartition par effectif ou fréquence selon les différentes valeurs de X

- Quand le caractère est continu, il est habituel de regrouper les valeurs en classes
- On crée ainsi une nouvelle variable de nature qualitative ordinale
- Regrouper en classes fait perdre de l'information
- La fréquence cumulée devient dans ce cas utile

Notations

- Soit X une variable quantitative
- Soit $(x_1, \ldots, x_n)$ la série statistique associée
- On note $(x_{(1)}, \ldots, x_{(n)})$ la série statistique ordonnée associée
- On note C_k, $k \in \mathbb{N}^\star$ les classes d'intervalles disjoints où se répartissent les obervations x_i. On a $C_j = [e_j, e_{j+1}[$ ou $C_j =]e_j, e_{j+1}]$
- $n_j = \#\{i = 1, \ldots, n : x_i \in C_j\}, \quad \forall j = 1, \ldots, n$, avec $\sum_{j=1}^{k} n_j = n$.

La distribution de X (relativement aux classes choisies) est la suite des effectifs $(n_j), j = 1, \ldots, k$ ou des fréquences $(f_j), j = 1, \ldots, k$ avec $f_j = \frac{n_j}{n}$.

Représentation à l'aide d'un tableau

Répartition des notes dans une matière

Classe	Effectif	Effectif cumulé	Fréquence (%)	Fréquence cumulée (%)
[0,6]	3	3	2,5	2,5
[6,10]	38	41	32,2	34,7
[10,20]	77	118	65,3	100

En pratique

- Afficher une description du jeu de données `Mesures5` du package `BioStatR`

```
str(Mesures5)
```

- Construire le tableau des effectifs de la variable `graines`

```
table_graines <- table(Mesures5$graines)
table_graines
```

- Calculer le tableau des effectifs cumulés

```
effcum_graines <- cumsum(table_graines)
effcum_graines
```

- Calculer le tableau des fréquences

```
## 1ere façon
freq_table_graines<-table_graines/sum(table_graines)
freq_table_graines
## 2e façon
freq_table_graines<-prop.table(table(Mesures5$graines))
freq_table_graines
```

En pratique

- Afficher une description du jeu de données `Mesures5` du package `BioStatR`

```
str(Mesures5)
```

- Construire le tableau des effectifs de la variable `graines`

```
table_graines <- table(Mesures5$graines)
table_graines
```

- Calculer le tableau des effectifs cumulés

```
effcum_graines <- cumsum(table_graines)
effcum_graines
```

- Calculer le tableau des fréquences

```
## 1ere façon
freq_table_graines<-table_graines/sum(table_graines)
freq_table_graines
## 2e façon
freq_table_graines<-prop.table(table(Mesures5$graines))
freq_table_graines
```

En pratique

- Afficher une description du jeu de données `Mesures5` du package `BioStatR`

```
str(Mesures5)
```

- Construire le tableau des effectifs de la variable `graines`

```
table_graines <- table(Mesures5$graines)
table_graines
```

- Calculer le tableau des effectifs cumulés

```
effcum_graines <- cumsum(table_graines)
effcum_graines
```

- Calculer le tableau des fréquences

```
## 1ere façon
freq_table_graines<-table_graines/sum(table_graines)
freq_table_graines
## 2e façon
freq_table_graines<-prop.table(table(Mesures5$graines))
freq_table_graines
```

En pratique (2)

- Calculer le tableau des fréquences cumulées

```
freqcum_table_graines<-cumsum(freq_table_graines)
freqcum_table_graines
```

Représentation à l'aide d'un histogramme

On appelle **histogramme** d'une variable X un graphique d'occurences comportant en ordonnées les effectifs (respectivement fréquences) **relatifs** (respectivement relatives) associées à des classes de valeurs représentées en abscisses.

- En règle générale, les classes sont de même amplitude
- Dans le cas contraire, il faut respecter la contrainte de proportionnalité entre l'aire des rectangles et les effectifs
- Le choix des classes (amplitudes et valeurs seuils) est souvent subjectif et il se base sur des classes ayant un sens dans la réalité
  Exemple : Moins de 6 : note éliminatoire, 6 à 10 : note compensable, 10 ou plus : matière validée
- 2 méthodes pour définir le nombres de classes (règle de Sturge $v = 1 + \log_2(n)$ et de Yule $v = 2.5\sqrt[4]{n}$)

Exemple : Age des athlètes dans un club sportif

Identifiant	Âge	Identifiant	Âge	Identifiant	Âge	Identifiant	Âge
01	15	11	16	21	22	31	18
02	18	12	19	22	20	32	17
03	21	13	21	23	21	33	24
04	24	14	24	24	23	34	15
05	21	15	23	25	28	35	23
06	23	16	28	26	22	36	25
07	28	17	25	27	28	37	27
08	22	18	18	28	22	38	26
09	18	19	15	29	20	39	16
10	19	20	27	30	25	40	21

- Quel est la nature du caractère observé ?
- Proposer un nombre de classes et une largeur de classe.
- Calculer les effectifs et fréquences pour chaque classe.
- Construire à main levée et interpréter l'histogramme du caractère observé.

En pratique

- Tracer l'histogramme de la variable `masse` du jeu de données `Mesures`

```
histo <- hist(Mesures$masse)
histo
```

Représentation à l'aide d'une courbe de fréquences cumulées

Soit $f_1, \ldots, f_k$ la suite des fréquences associées aux k classes $C_j = [e_j, e_{j+1}[$ d'une série statistique. On appelle **fonction de répartition empirique** d'une variable quantitative X la fonction F_X définie de $\mathbb{N}^\star$ dans $[0, 1]$ par

$$F_X(x) = \begin{cases} 0 \text{ si } x < e_1 \\ \sum_{i=1}^{j} f_i \text{ si } e_j \leq x \leq e_{j+1} \quad \forall j = 1, \ldots, k-1 \\ 1 \text{ si } x \geq e_{k+1} \end{cases}$$

La courbe des fréquences cumulées est la représentation graphique de la fonction de répartition empirique.

Représentation à l'aide d'une courbe de fréquences cumulées (2)

- Tracer l'histogramme de la variable `masse` dans le jeu de données `Mesures`

```
## 1ere façon
histo <- hist(Mesures$masse, plot = FALSE)

barplot <- barplot(cumsum(histo$counts),
                   ylab = "Effectif",
                   xlab = "Masse",
                   main = "Histogramme des effectifs cumulés")
## 2e façon
plot(ecdf(Mesures$masse))
```

Sommaire

Sommaire

Objectifs

- Résumer l'information contenue dans une série statistique en quelques valeurs numériques
- Comparer plusieurs séries statistiques
- Uniquement pour les variables quantitatives (excepté le mode)
- Deux types de caractéristiques : statistiques de tendance centrale et de position (moyenne, mode, quantiles, mode), statistiques de dispersion (variance, écart-type, écart inter-quantiles).

Sommaire

Définition

- Une statistique est dite **résistante** si elle est peu influençable par les valeurs extrêmes de la série statistique à partir de laquelle elle est obtenue.
- La notion de robutesse est à une méthode statistique, ce que la notion de résistance est à une statistique.
- Par exemple, la regression linéaire n'est pas un méthode robuste comparativement à la regression quantile qui est une méthode robuste
- Dans les échantillons de petite taille, une seule valeur extrême peut modifier fortement les résultats.

Sommaire

Moyenne généralisée

Soient

- $x = (x_1, \ldots, x_n)$ une série statistique d'une variable quantitative X
- $\omega = (\omega_1, \ldots, \omega_n)$ un vecteur de poids relatifs
- vérifiant $\sum_{i=1}^{n} w_i = 1$

La **moyenne pondérée d'ordre** $r \in \mathbb{R}^\star$ est la quantité

$$\bar{x} = \left(\sum_{i=1}^{n} w_i x_i^r \right)^{1/r}$$

- Le plus souvent $w_i = \frac{1}{n}, \forall i = 1, \ldots, n$
- La moyenne est une statistique non résistante

Moyenne arithmétique

On appelle **moyenne arithmétique pondérée** la quantité

$$\bar{x} = \bar{x}_1 = \left(\sum_{i=1}^{n} w_i x_i \right)$$

- Si $\forall i = 1, \ldots, n$, on a $w_i = \frac{1}{n}$ on parle alors de moyenne arithmétique simple.
- La moyenne arithmétique simple vérifie

$$0 = \sum_{i=1}^{n} (x_i - \bar{x})$$

$$\bar{x} = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^{n} (x_i - a)^2$$

Moyenne géométrique

On appelle **moyenne géométrique pondérée** la quantité

$$\bar{x}_G = \bar{x}_0 = \lim_{x \to 0} \bar{x}_r = \prod_{i=1}^{n} x_i^{wi}$$

- $\ln(\bar{x}_G)$ est la moyenne arithmétique pondérée des $\ln$ −valeurs de la série
- Ce type de moyenne est utilisée pour calculer des taux de croissances moyens

Moyenne géométrique (2)

Exemple

On considère une action boursière présentant les rendements suivants : $+10\%$ la première année et -5% les deux dernières années suivantes.

- On cherche le taux de croissance moyen de l'action sur les trois années ;

Moyenne harmonique

On appelle moyenne harmonique pondérée la quantité

$$x_H = \bar{x}_{-1} = \left(\sum_{i=1}^{n} \frac{w_i}{x_i} \right)^{-1}$$

- Cette moyenne est utilisée pour calculer des moyennes de grandeurs dont l'unité est elle-même un ratio d'unités (km/h, médécins par habitant ...)

Moyenne harmonique (2)

Exemple

Considérons 4 investissements de 2000 euros chacun sur quatre périodes différentes et sur le même actif : Le premier investissement est réalisé lorsque le cours d'un actif coûte 10 € ; Le deuxième lorsque le cours est de 8 € ; Le troisième lorsque le cours est de 5 € et le dernier investissement lorsque le cours de l'actif est de 4 €.

- Calculer le cours moyen du portefeuille.

Moyenne quadratique

On appelle moyenne quadratique pondérée la quantité

$$\bar{x}_Q = \bar{x}_2 = \sqrt{\sum_{i=1}^{n} w_i x_i^2}$$

- Cette moyenne est surtout utilisée pour mesurer des écarts de mesure

Remarques

- Il est possible d'ordonner les différents moyennes

$$\bar{x}_H \leq \bar{x}_G \leq \bar{x}_A \leq \bar{x}_Q$$

- Si les valeurs $(x_1, \ldots, x_n)$ sont peu différentes les unes des autres, les valeurs des différentes moyennes sont proches.
- Quand une série comporte des valeurs trop extrêmes qui travestissent la réalité du phnomène étudié, on utilise parfois des moyennes tronquées
- On appelle moyenne tronquée à $\alpha\%$ la moyenne calculée sur $(100 - \alpha)\%$ de la taille de la série. La troncature se fait en queue de distribution (unilatéralement ou bilatéralement, das ce dernier cas la troncature est symétrique).

Médiane

On appelle **médiane** la valeur qui divise la série en deux sous-populations de tailles égales. Ainsi, la médiane M_e vaut

$$M_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)} \text{ si } n \text{ est impair} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \text{ par convention si } n \text{ est pair} \end{cases}$$

- Si la série est partitionnée en classes, on parlera alors de classe médiane
- Quelle est la médiane de la série $(1, 5, 8, 12; 13, 15, 25, 30, 32)$? Que se passe-t-il si on rajoute la valeur 10^5?
- La médiane est une statistique résistante

Médiane

On appelle **médiane** la valeur qui divise la série en deux sous-populations de tailles égales. Ainsi, la médiane M_e vaut

$$M_e = \begin{cases} x_{\left(\frac{n+1}{2}\right)} \text{ si } n \text{ est impair} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \text{ par convention si } n \text{ est pair} \end{cases}$$

- Si la série est partitionnée en classes, on parlera alors de classe médiane
- Quelle est la médiane de la série $(1, 5, 8, 12; 13, 15, 25, 30, 32)$? Que se passe-t-il si on rajoute la valeur 10^5?
- La médiane est une statistique résistante

Médiane

Interpolation linéaire

- On a le tableau suivant

Taille (cm)	[150, 160[	[160, 165[	[165, 170[	[170, 175[	[175, 180[	[180,190[
Effectifs	2	7	10	6	2	3
Eff. cumulée	2	9	19	25	27	30

- A quelle classe appartient la médiane ?
- Proposer une valeur de la médiane par interpolation linéaire

Le mode

On appelle **mode** d'une distribution, la valeur de la variable X dont l'occurence est la plus élevée.

- Il peut y avoir plusieurs modes (on parle alors de distribution plurimodale ou multimodale)
- Pour une variable quantitative continue, on s'intéressera à la classe modale
- Le mode est une statistique résistante

Sommaire

Les quantiles

On appelle **quantile** ou **fractile** de X d'ordre α avec $\alpha \in [0, 1]$, la valeur $q_x(\alpha)$ telle qu'une proportion α de la population présente une valeur inférieure à $q_x(\alpha)$.

Le quantile d'ordre α vérifie ainsi

$$q_x(\alpha) = F_X^{-1}(\alpha) = \inf\{x / F_X(x) \geq \alpha\}$$

avec F_X la fonction de répartition.

Quelques quantiles usuels

- On note Q_1, Q_2 et Q_3 les premier, deuxième et troisième quartiles. Il s'agit des quantiles d'ordre $0.25, 0.5, 0.75$.
- On note $(D_i)_{i=1,\dots,9}$ les neufs déciles. Il s'agit des quantiles d'ordre 0.1 à 0.2
- On note $(C_i)_{i=1,\dots 99}$ les centiles. Il s'agit des quantiles
- Tout comme la médiane, ces quantiles peuvent se calculer par interpolation linéaire

Valeurs adjacentes supérieure et inférieure

On appelle **VAI** d'une série statistique, la plus petite valeur de la série supérieure à la quantité $Q_1 - 1.5(Q_3 - Q_1)$.

$$VAI = \min\{x_i \in (x_1, \ldots, x_n) / x_i \geq Q_1 - 1.5(Q_3 - Q_1)\}$$

On appelle **VAS** d'une série statistique, la plus grande valeur de la série inférieure à la quantité $Q_3 + 1.5(Q_3 - Q_1)$.

$$VAI = \max\{x_i \in (x_1, \ldots, x_n) / x_i \leq Q_3 + 1.5(Q_3 - Q_1)\}$$

- Une valeur est dite extrême si elle est inférieure à la VAI ou supérieure à la VAS.
- Cette notion de valeur extrême n'est pas universelle.

Boxplot

La boîte à moustache ou boxplot ou boîte de Tukey est un moyen de visualiser rapidement plusieurs statistiques de position dont

- Les limites de la boîte sont Q_1 et Q_3
- Le trait central dans la boîte est $Q_2 = M_e$
- Les moustaches de la boîte sont VAI et VAS (en général). Cela peut aussi être D_1 et D_9.
- Eventuellement les valeurs extrêmes

Les boxplots sont utiles pour comparer des distributions sur différents sous-échantillons.

Exemple de Boxplot

- Construire la boîte à moustaches de la variable `masse`

```
boxplot(Mesures$masse)
title("Boite à moustaches de la variable masse")
```

Sommaire

Variance

On appelle **variance pondérée**, la statistique σ_X^2 (respectivement s_X^2) lorsqu'il s'agit d'une population (respectivement d'un échantillon). On a

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

- En général $w_i = \frac{1}{n} \forall i = 1, \ldots, n$
- La plupart des logiciels calculent la **variance corrigée** qui vaut

$$\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{au lieu de} \quad \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- Voir cours de Statistiques inférentielles pour la différence entre les deux

Variance

- La décomposition de Konig-Huygenes permet de calculer la variance autrement

$$\sigma_X^2 = \sum_{i=1}^{n} w_i x_i^2 - \bar{x}^2$$

- La variance n'est pas une statistique résistante

Ecart-type

On appelle écart-type noté σ_X la racine carée de la variance.

Tout comme la variance, l'écart-type n'est pas une statistique résistante.

Limites de la variance et de l'écart-type

La variance et l'écart-type présentent un inconvénient majeur, ils ne permettent pas de comparer la dispersion de deux séries dont les moyennes sont différentes. De plus, ils sont impactés par l'unité choisie.

- Soit la série $x = (996, 997, \ldots, 1003, 1004)$

```
x <- 996:1004; mean(x); var(x); sd(x)
```

```
## [1] 1000
```

```
## [1] 7.5
```

```
## [1] 2.738613
```

- Soit la série $z = 0.996, \ldots, 1.004$

```
z <- x/1000; mean(z); var(z); sd(z)
```

```
## [1] 1
```

```
## [1] 7.5e-06
```

```
## [1] 0.002738613
```

Coefficient de variation

On appelle **coefficient de variation**, l'écart-type rapporté à la moyenne. On a alors

$$CV(x) = \frac{\sigma_x}{\bar{x}}$$

- Le coefficient de variation est un nombre sans unité (on parle alors de paramètre de dispersion relative)
- Cette statistique n'a de sens que si les moyennes des distributions ne sont pas proches de 0
- Le coefficient de variation est une statistique non résistante
- Avec les données de la slide précédente, on a :

$CV(x) = 2.6 \times 10^{-3}, CV = 2.6 \times 10^{-3}$

Les écarts interquantiles

On appelle écart interquantile (resp. écart inerquantile relatif), la différence entre le dernier et le premier quantile (resp. cette différence rapportée à la médiane)

- Ecart interquartile $Q_3 - Q_1$
- Ecart interquartile relatif $\frac{Q_3 - Q_1}{Q_2}$
- Ecart interdécile $D_9 - D_1$
- Ecart interdécile relatif $\frac{D_9 - D_1}{D_5}$
- Ecart intercentile $C_{99} - C_1$
- Ecart intercentile relatif $\frac{C_{99} - C_1}{C_{50}}$

Les écart interquantiles relatifs sont sans unité et permettent de comparer des distributions n'ayant pas la même médiane.