Statistique descriptive univariée

Marie-Ange Remiche

Université de Namur

Conditions d'analyse

Définition

Une population regroupe l'ensemble des individus sur lesquels on désire réaliser une ou plusieurs mesures d'un caractère (encore appelé variable). Les populations pouvant être de taille importante, il est parfois plus simple d'utiliser un échantillon de celle-ci, soit un sous-ensemble de cet ensemble complet qu'est la population.
La statistique est une science dont l'objectif est d'interpréter les données récoltées au sujet d'une population ou d'un échantillon de cette population. Le mot statistique désigne également un ensemble de données, on parle encore de série statistique associée à un caractère particulier.

Définition

Un caractère peut être quantitatif ou qualitatif.
Lorsqu'il est quantitatif, il peut être discret ou continu. Dans ce dernier cas, il peut prendre toutes les valeurs d'un intervalle particulier.
Un caractère peut être simple (ou encore univarié) ou multiple. Lorsqu'il est simple, la mesure du caractère ne produit qu'une seule valeur. Par contre, lorsqu'il est multiple, sa mesure produit une série de valeurs.

Définition

On distingue les échelles quantitatives suivantes

- Echelle ordinale : une notion d'ordre existe entre les valeurs mesurées.
  Exemple : classement de projets par ordre de priorité selon le lecteur.
- Echelle de rapport : la notion de 0 a un sens physique (l'absence du caractère observé).
  Exemple : temps réalisé pour un 100 mètres par des étudiants de 18 ans.
- Echelle d'intervalles : le 0 ne signifie pas l'absence du caractère.
  Permet uniquement la comparaison d'intervalles.
  Exemple : année d'observation de crues exceptionnelles. On peut comparer la longueur de temps qui sépare deux crues successives mais on ne peut pas faire de rapport (soit 2021/ 2029) entre deux années d'observation.

Dans le cas d'une variable qualitative, on parle d'échelle nominale.

UNIVERSITÉ DE NAMUR

Nous travaillons, avec la data.frame `BD` contenant les variables

- `genre` : caractère qualitatif, la valeur 1 désignant un homme.
- `sigar` : caractère quantitatif d'échelle de rapport, indiquant le nombre de cigarettes fumées habituellement par jour.
- `bronchite` : caractère qualitatif indiquant par la valeur 1 si la personne a eu au moins une bronchite les trois derniers mois, 0 sinon.
- `age` : caractère quantitatif prenant ses valeurs dans l'ensemble $\mathbb{N}$ et d'échelle de rapport.
- `maux` : caractère qualitatif à trois modalités, indiquant par la valeur 0 qu'aucun maux de tête n'a été observé lors des sept derniers jours, 1 que seuls des maux de tête légers ont pu être observés, et 2 pour les maux de tête sévères.

Attention à la notion d'individus et de variables

Nom	Note 1	Note 2	Note 3	Note 4	Note 5
Alice	15	12	14	7	16
Bob	7	4	10	6	12
Chris	19	15	16	12	19
Eric	13	12	14	11	14
Fred	13	13	12	14	17

Cas d'un caractère discret

Soit n la taille de notre échantillon $\mathcal{S}$, et un caractère discret X. La suite finie notée

$X = (45, 54, 55, 45, 55)$

$$\underline{X}(\mathcal{S}) = \big(X_1, \quad \ldots \quad , X_n\big)$$

où X_i prend ses valeurs dans l'ensemble $\{x_1, \ldots, x_p\}$, avec $x_1 < x_2 < \ldots < x_p$. Ceux-ci correspondent aux valeurs atteintes par les n différents individus appartenant à notre échantillon. Une autre représentation de cette série statistique correspondante est de considérer

$x(45,1)$ $x(54,1)$ $x(55,2)$

$$\underline{X} = (x_i, n_i)_{i=1,\ldots,p}$$

où n_i est le nombre de fois que la valeur x_i a été observée dans notre échantillon. Plutôt que de série statistique, certains auteurs parlent de distribution observée ou encore de distribution empirique.

Exemple

Parmi les non-fumeurs, conservez uniquement les individus interrogés qui ont eu une bronchite.

```
library(knitr)
kable(subset(BD,BD$sigar==0 & BD$bronchite==1))
```

Si on désire observer la table de contingence des maux pour ces personnes-là,

```
kable(table(subset(BD, BD$sigar==0 & BD$bronchite==1)$maux))
```

Définition

On parle d'effectif de la valeur x_i pour désigner ce nombre n_i.
L'effectif cumulé en x_i est défini comme

$$\text{effectif cumulé en } x_i \stackrel{\text{def}}{=} \sum_{j=1}^{i} n_j,$$

défini pour $i = 1, \ldots, p$.

Propriété

L'effectif cumulé en x_p est tel que

$$\text{effectif cumulé en } x_p = n.$$

Définition

La fréquence de la valeur x_i, noté f_i est le rapport de l'effectif de la valeur x_i avec l'effectif n de l'échantillon, soit

$$f_i \stackrel{\text{def}}{=} \frac{n_i}{n}.$$

La fréquence cumulée F_i en x_i se définit comme

$$F_i \stackrel{\text{def}}{=} \sum_{j=1}^{i} f_j.$$

$F_p = 1$

Exemple

```r
BD1<-(subset(BD,BD$sigar==0 & BD$bronchite==1))
prop.table(table(BD1$maux))
```

$[a, b]$ non compris : "(" "v" "]"

$[a_0, a_1]$ $(a_1, a_2]$ $(a_2, a_3]$ $(a_{p-1}, a_p]$ b

Définition

La règle de Sturges propose

- de déterminer le nombre de classes d'une découpe de la manière suivante

$$\text{nombre de classes} \overset{\text{def}}{=} \lceil \log_2(n) + 1 \rceil$$

  où n la taille de la série statistique, et

- de conserver une longueur de classe constante

Définition

L'effectif de $]\cancel{a_i; a}_{i+1}]$ est le nombre n_i de valeurs observées dans l'intervalle $]a_i; a_{i+1}]$. $]a_{i-1}, a_i]$

L'effectif cumulé en a_i est le nombre de valeurs observées dans l'intervalle $]-\infty, a_i]$.

La distribution statistique groupée est alors notée

$$(]a_i; a_{i+1}], n_i)_{i=\cancel{1}^{\,0},\ldots,\cancel{p}\; p-1}$$

avec $a_i < a_{i+1}$ et p étant le nombre de classes formant la partition de l'intervalle des valeurs possibles.

Définition

La fréquence de $]\cancel{a_i ;} a_{i+1}]$, notée f_i est égale au rapport suivant

$]a_{i-1}, a_i]$

$$f_i \overset{\text{def}}{=} \frac{n_i}{n}.$$

La fréquence cumulée en a_i est égale à la somme

$$\text{fréquence cumulée en } a_i \overset{\text{def}}{=} \sum_{j=1}^{i} f_j.$$

Exemple

La commande `hist` permet d'obtenir un regroupement en classes, classes déterminées par la règle de Sturges.

```
BDEtudeAGE<-hist(BD$age)
BDEtudeAGE
```

```
BDEtudeAGE<-hist(BD$age, breaks=c(20,22,27,37,40,max(BD$age)))
BDEtudeAGE$density
prop.table(table(BDEtudeAGE$breaks))
```

En particulier, on remarque

- le vecteur `breaks` de cette frame.table `BDEtudeAGE`, où sont conservés les différents a_i formant les limites des classes obtenues,
- le vecteur `counts` donnant pour chaque classe, son effectif,
- le vecteur `density` à ne pas confondre avec les fréquences des classes. En effet, ces valeurs seront utilisées pour représenter l'histogramme.

Cas d'un caractère qualitatif

Définition

Le tableau de contingence regroupe les effectifs des différentes modalités de la variable.

Exemple

```
table(BD$age)
```

Définition

Le diagramme en bâtons des effectifs (respectivement des fréquences) d'une série statistique discrète est tel que

- en abscisse, on considère les différentes valeurs possibles x_i, pour $i = 1 \ldots, p$
- en ordonnée, on indique l'effectif (respectivement la fréquence) observé.

Diagramme en bâtons des effectifs et des fréquences

Exemple

```
plot(table(BD$age))
plot(prop.table(table(BD$age))  )
```

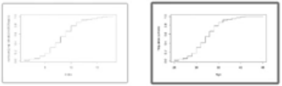
Distribution empirique

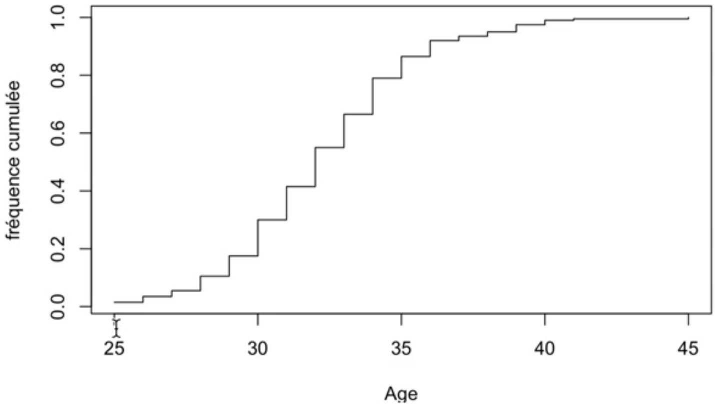
Exemple

```
plot(cumsum(prop.table(table(BD$age))),type="s")
```

Pour la distribution empirique

````
```{r}
plot(cumsum(prop.table(table(BD$age))),type="s")
plot(x=rownames(table(BD$age)),cumsum(prop.table(table(BD$age))),type="s",xlab="Age",ylab="
fréquence cumulée")
```
````





X est observée sur (x_1, x_p)

La distribution empirique d'une variable est une fonction en escalier dont l'équation est la suivante, $\forall x \in \mathbb{R}$

$$F(x) = \begin{cases} 0 & \text{si } x < x_1 \\ \sum_{j=1}^{i} f_i & \text{si } x_i \leq x < x_{i+1}, i = 1, \ldots, p-1 \\ 1 & \text{si } x \geq x_p. \end{cases}$$

Définition

L'histogramme est une représentation graphique d'une série statistique groupée où une barre (ou rectangle) est associée pour chaque classe. La surface de ce rectangle est proportionnelle à l'effectif de la dite classe. On distingue deux cas

- lorsque les amplitudes des classes sont égales, la base du rectangle est égale à cette amplitude et la hauteur est proportionnelle avec le même paramètre K à l'effectif de chaque classe.
- Lorsque les amplitudes sont différentes, il existe un commun diviseur a. Dès lors, la base de chaque rectangle sera proportionnelle à l'amplitude de la classe mais sera un multiple entier de ce diviseur a. La hauteur est proportionnelle avec le paramètre K à l'effectif divisé par le rapport de l'amplitude avec le diviseur a.

UNIVERSITÉ DE NAMUR

Soit l'échantillon discret suivant $\underline{X} = (X_i)_{i=1,\ldots,n}$ où X_i est l'observation réalisée sur l'individu i, sachant que l'échantillon compte n individus.
Ces observations peuvent être présentées de plusieurs manières, dont

- comme nous l'avons vu précédemment, en utilisant les effectifs, soit $\underline{X} = (x_i, n_i)_{i=1,\ldots,p}$ avec $x_i < x_{i+1}$, pour tout $i < p$,
- soit en classant ces observations par ordre croissant et obtenons la suite $(\underline{X}) = (X_{(i)})_{i=1,\ldots,n}$, que nous appelons la statistique d'ordre.

$\underline{X} = (22, 23, 22, 24, 22)$

$\underline{X}$ $(22, 3)$ $(23, 1)$ $(24, 1)$

$X(1)$

↳ la plus petite valeur observée $(\underline{X}) = (22, 22, 22, 23, 24)$

Définition

Le mode de la série statistique discrète $(\underline{X}) = (X_{(i)})_{i=1,\ldots,n}$, noté $\text{Mo}(\underline{X})$ est la valeur x_i dont la fréquence est maximale.

Dès lors, on distinguera

- les distributions unimodales avec un seul mode,
- des distributions plurimodales avec plusieurs modes.

Exemple

```
sort(table(BD$age))
```

Définition

Le quantile d'ordre α, où $\alpha \in]0,1[$, d'une série statistique discrète $(\underline{X}) = (X_{(i)})_{i=1,\dots,n}$, noté $Q_\alpha(\underline{X})$ est tel que

$$Q_\alpha(\underline{X}) \stackrel{\text{def}}{=} X_{(m)} + d\left(X_{(m+1)} - X_{(m)}\right)$$

où

$$m = \lfloor \alpha(n+1) \rfloor$$
$$d = \alpha(n+1) - m.$$

- la médiane ou second quartile, noté $Q_{0.5}(\underline{X})$,
- le premier quartile, noté $Q_{0.25}(\underline{X})$,
- le troisième quartile, noté $Q_{0.75}(\underline{X})$.

UNIVERSITÉ DE NAMUR

Définition

Soit l'échantillon discret $\underline{X} = (x_i, n_i)_{i=1,\dots,p}$, la moyenne arithmétique de cet échantillon, notée $\overline{X}$, est donnée par

$$\overline{X} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{p} n_i x_i.$$

1. Si nous travaillons avec la représentation $\underline{X} = (X_i)_{i=1,\dots,n}$, la moyenne se calcule de la manière suivante
$$\overline{X} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} X_i.$$
2. Il existe d'autres moyennes que la moyenne arithmétique, comme la moyenne géométrique ou quadratique.

Exemple

```
quantile(BD$age, c(0.05))
```

Caractéristiques de dispersion

Définition

L'étendue d'un échantillon $\underline{X}$ est la différence suivante

$$e(\underline{X}) = \max(\underline{X}) - \min(\underline{X}),$$

soit la différence entre la plus grande et la plus petite valeur de l'échantillon.
L'étendue interquartile, notée $\mathrm{EIQ}(\underline{X})$ est la différence suivante

$$\mathrm{EIQ}(\underline{X}) = \mathrm{Q}_{0.75}(\underline{X}) - \mathrm{Q}_{0.25}(\underline{X}).$$

Exemple

```
IQR(BD$age)
```

Définition

La variance empirique, notée $S^2(\underline{X})$ est obtenue de la manière suivante

$$S^2(\underline{X}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{p} \left(x_i - \overline{X}\right)^2 n_i.$$

La variance empirique corrigée, notée $S_c^2(\underline{X})$, se calcule de la manière suivante

$$S_c^2(\underline{X}) \stackrel{\text{def}}{=} \frac{1}{n-1} \sum_{i=1}^{p} \left(x_i - \overline{X}\right)^2 n_i.$$

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

$$s(\underline{x}) \stackrel{\text{def}}{=} +\sqrt{s^2(\underline{x})}$$

Définition

L'écart-type empirique, noté $S(\underline{X})$ s'obtient comme

$$S(\underline{X}) \stackrel{\text{def}}{=} \sqrt{S^2(\underline{X})},$$

tout comme l'écart-type empirique corrigé, noté $S_c(\underline{X})$

$$S_c(\underline{X}) \stackrel{\text{def}}{=} \sqrt{S_c^2(\underline{X})}.$$

UNIVERSITÉ DE NAMUR

Définition

Le moment centré d'ordre r d'une série statistique discrète $\underline{X} = (x_i, n_i)$ pour $i = 1, \ldots, p$, noté $m_r(\underline{X})$ est défini comme

$$m_r(\underline{X}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{p} (x_i - \overline{X})^r n_i.$$

Le moment d'ordre r de la série statistique $\underline{X}$ se calcule comme

$$\overline{X^r} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{p} x_i^r n_i.$$

Propriété - Formule de Huygens

La variance empirique respecte l'identité suivante

$$S^2(\underline{X}) = \overline{X^2} - \overline{X}^2,$$

où $\overline{X^2}$ est le moment d'ordre 2 de $\underline{X}$.

$$S^2(\underline{X}) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i^2 + \overline{X}^2 - 2X_i\overline{X}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \frac{1}{n}\sum_{i=1}^{n} \overline{X}^2 + \frac{1}{n}\sum_{i=1}^{n} \left(-2X_i\overline{X}\right)$$

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i^2}_{\overline{X^2}} + \frac{1}{\cancel{n}}\overline{X}^2 \cdot \cancel{n} - 2\overline{X}\underbrace{\frac{1}{n}\sum_{i=1}^{n} X_i}_{\overline{X}}$$

$$S^2(\underline{X}) = \overline{X^2} + \overline{X}^2 - 2\overline{X}^2 = \overline{X^2} - \overline{X}^2$$

Exemple

```r
var(BD$age)
sd(BD$age)
```

Attention, il s'agit respectivement de la variance empirique corrigée et de l'écart-type (ou *Standard Deviation* en anglais) empirique corrigé.

Exemple

Une même variable est observée sur deux échantillons distincts. En voici la table de contingence.

Valeurs observées	1	2	3	4	10	16	17	18	19
Echantillon 1	1	1	1	0	3	0	1	1	1
Echantillon 2	1	1	1	1	1	1	1	1	1

On obtient alors les caractéristiques numériques suivantes

	Echantillon 1	Echantillon 2
Moyenne	10	10
Amplitude	18	18
Intervalle Interquartile	14	14
Ecart-type	6.55	7.15

UNIVERSITÉ DE NAMUR

Définition

Le coefficient de variation d'un échantillon $\underline{X}$, noté $CV(\underline{X})$ est défini par le rapport entre l'écart-type empirique et la moyenne arithmétique.

$$CV(\underline{x}) = S(\underline{x}) / \bar{x}$$

Définition

Le coefficient d'asymétrie de Fisher d'une série statistique $\underline{X}$, noté $\gamma_1(\underline{X})$ est donné par

$$\gamma_1(\underline{X}) \stackrel{\text{def}}{=} \frac{m_3(\underline{X})}{S^3(\underline{X})}.$$

Le coefficient d'asymétrie de Pearson, noté $\beta_1(\underline{X})$ est le carré du coefficient d'asymétrie de Fisher, soit

$$\beta_1(\underline{X}) \stackrel{\text{def}}{=} \gamma_1^2(\underline{X}).$$

Définition

Le coefficient d'aplatissement de Fisher d'une série statistique $\underline{X}$, noté $\gamma_2(\underline{X})$ est donné par

$$\gamma_2(\underline{X}) \stackrel{\text{def}}{=} \frac{m_4(\underline{X})}{m_2^2(\underline{X})} - 3.$$

Le coefficient d'aplatissement de Pearson, noté $\beta_2(\underline{X})$ est donné par

$$\beta_2(\underline{X}) \stackrel{\text{def}}{=} \frac{m_4(\underline{X})}{S^4(\underline{X})}.$$

UNIVERSITÉ DE NAMUR

Le package `moments` comprend deux commandes `skewness` et `kurtosis` permettant de calculer respectivement le coefficient d'asymétrie de Fisher et le coefficient d'aplatissement de Pearson. En les appliquant, on obtient

Exemple

```r
library(moments)
skewness(BD$age)
kurtosis(BD$age)
```

Le boxplot

Définition

La *boîte à moustache* ou *boxplot* en anglais est un graphique où sont représentés des caractéristiques de position et de dispersion.

Les valeurs extrêmes de la boîte à moustaches ne représentent pas nécessairement les extrêmes des observations mais plutôt les valeurs

$$\text{borne inf} \stackrel{\text{def}}{=} \max(Q_{0.25}(\underline{X}) - 1.5(Q_{0.75}(\underline{X}) - Q_{0.25}(\underline{X})), \min(\underline{X}))$$

$$\text{borne sup} \stackrel{\text{def}}{=} \min(Q_{0.75}(\underline{X}) + 1.5(Q_{0.75}(\underline{X}) - Q_{0.25}(\underline{X})), \max(\underline{X}))$$

Exemple

```
boxplot(BD$age)
```

UNIVERSITÉ DE NAMUR