Statistiques descriptives avec avec R

Datastorm - B. Thieurmel

1 Représenter une variable qualitative

Nous allons rentrer des données à la main pour une variable qualitative. Cette variable représente l'appartenance à un groupe avec 3 modalités **g1**, **g2** et **g3**. Les 2 premiers individus sont dans le groupe 1, les 3 suivants dans le groupe 2 et le dernier dans le groupe 3 :

```
ybrut <- c("g1","g1","g2","g2","g2","g3")
print(ybrut)
summary(ybrut)
```

Que fait le dernier ordre ci-dessus ? Nous devons transformer ce vecteur (de caractères) en variable qualitative (nommée `factor` sous **R**) :

```
y <- factor(ybrut)
```

Que font les ordres suivants ?

```
levels(y)
nlevels(y)
table(y)
sum(table(y))
table(y) / sum(table(y)) * 100
```

Tracer les effectifs de chaque modalité dans un diagramme en barre :

```
barplot(table(y))
```

Tracer les pourcentages de chaque modalité dans un diagramme en barre :

```
barplot(table(y) / sum(table(y)) * 100, ylab = "pourcentages",
        xlab = "groupes", main = "Répartition de y")
```

Que font les options `xlab`, `ylab` et `main` ?

Que fait le résumé numérique d'une variable qualitative ?

```
summary(y)
```

2 Représenter une variable quantitative continue

Représentons une variable quantitative continue. Ouvrir le fichier **varquant.r** et exécuter son contenu (vous pouvez également utiliser la fonction `source` pour inclure ce fichier et l'éxécuter depuis un autre script)

Que fait le résumé numérique d'une variable quantitative (continue) ?

```
summary(y)
```

Trouver sur les deux graphiques ci-dessous la différence et expliquer la.

```
hist(y, freq = TRUE)
hist(y, freq = FALSE)
```

Que font toutes les options pour ce graphique ?

```
hist(y, freq = FALSE, breaks = 10, xlab = "huile", main = "Histogramme")
```

Que fait cette option ?

```
hist(y, freq = FALSE, breaks = c(15,18,25,30,36))
```

Expliquer tous les ordres ci-dessous :

```
boxplot(y, xlab = "", ylab = "teneur en huile")
mean(y)
abline(h = mean(y))
quantile(y)
median(y)
abline(h = median(y), col = 2)
```

Conclusion : l'histogramme est tracé grâce à `hist` avec l'option `freq = FALSE`.

Un autre estimateur de la densité (estimateur à noyau) est disponible afin d'estimer la densité par une fonction continue :

```
density(x, ...)
```

Retourne les coordonnées `x` et `y` d'un estimateur de la densité du vecteur de données `x`. L'argument `bw` indique la largeur de fenêtre (plus elle est grande plus la courbe est lisse)

```
lnormal <- rnorm(100)
ndens <- density(lnormal, width=1.2)
hist(lnormal, probability = T)
lines(ndens)
```

3 Représenter une variable quantitative discrète

Représentons une variable quantitative discrète : le nombre d'enfant par famille. La première famille ce compose de 5 enfants, la second n'en a pas, la troisième et la quatrième ont 2 enfants et la cinquième n'en a pas.

```
y <- c(5,0,2,2,0)
```

Que font les commandes suivantes ?

```
unique(y)
sort(unique(y))
table(y)
```

Le diagramme en barre des effectifs est le diagramme suivant :

```
plot(sort(unique(y)), table(y), type="h", ylim = c(0, max(table(y))))
```

En général, dès que les valeurs possibles sont assez nombreuses (par exemple 7 ou 10 ou plus) la variable quantitative discrète est assimilée à une variable quantitative continue. La distinction quantitatif discret ou continue n'existe pas sous **R**, les deux sont des variables numériques (`numeric`).

On peut préférer la fonction `barplot` :

```
barplot(table(y))
```

4 Données des tournesols

1. Importer le tableau **tournesol.csv** dans la variable `tpropre`. Il contient les variables décrites ci-dessous.

Code variable	Descriptif variable
ecotype	code plante
plt	numéro du plant d'un écotype donné
etat	état d'origine de la plante (aux USA)
longitude	longitude du lieu de collecte (aux USA)
latitude	latitude du lieu de collecte (aux USA)
haut	hauteur des plants
semflo	jour de floraison (écart en jour par rapport au premier mai)
rambas	note de ramification basale (entre 0 aucune et 4 maximum)
longfeu	longueur du cumulée du limbe et du pétiole (cm ?)
grlon	longueur maxi de la graine (mm, moyenne sur 15 graines minimum)
huile	pourcentage d'huile

Figure 1: Variables mesurées sur les tournesols (dans la station d'essai aux environs de Montpellier).

2. Donner pour chaque variable son type (variable qualitative, quantitative discrète, quantitative continue).
3. Effectuer un résumé numérique du tableau `tpropre` :

```
summary(tpropre)
```

4. Quelles sont les variables qui sont reconnues comme variables quantitatives et comme variables qualitatives ?
5. Donner à chaque variable le type voulu grâce à `factor` ou `as.numeric`

5 Deux variables quantitatives continues

Par defaut **R** trace des points (`type="p"`) aux coordonnées fournies (ci-dessous l'ordonnée est la variable `huile` et l'abscisse la variable `grlon`). Détailler le rôle des options suivantes :

```
plot(huile~grlon, data = tpropre)
plot(huile~grlon, data = tpropre, pch = "+")
plot(huile~grlon, data = tpropre, col = 2, pch = "+")
```

Traçons des lignes :

```
plot(huile~grlon, data = tpropre, type = "l")
```

Qu'a t-on fait ?

```
# Trions les donnees préalablement
ordre <- order(tpropre$grlon)
plot(x = tpropre$grlon[ordre], y =  tpropre$huile[ordre], type = "l",
     lty = 2, lwd = 2, col = "purple")
```

6 Deux variables qualitatives : tableau de contingence

Utilisez l'ordre suivant :

```
table(tpropre[,"ecotype"], tpropre[,"etat"])
```

Que renvoit il ?

7 Données des tournesols (suite)

1. Calculer la moyenne empirique des variables **huile**, **grlon** et **longfeu**.
2. Pour ces mêmes variables donner leurs quartiles empiriques.
3. Pour ces mêmes variables les représenter par un `boxplot`.
4. Pour ces mêmes variables calculer leur variance empirique.
5. Représenter graphiquement chacune des variables et exporter ces représentations graphiques. (bouton d'export dans **RStudio** ou utilisation des fonction `jpeg`, `png`, `pdf`, ...)