tuto-R

Tutoriel sur l'analyse d'enquêtes avec R & le *tidyverse*

Joseph Larmarange

2022-09-01T00:00:00+02:00

Table des matières

Préface

> **Site en construction**
>
> Le présent site est en cours de construction et sera complété dans les prochains mois.
> En attendant, nous vous conseillons de consulter le site analyse-R.

Ce document est un tutoriel dédié à l'analyse d'enquêtes avec R, logiciel libre de traitement et d'analyse de données, en ayant notamment recours à une série d'extensions regroupées sous l'appellation *tidyverse*.

Il est basé sur R version 4.1.3 (2022-03-10).

Il s'agit d'une extension/refonte du site analyse-R et a pour ambition d'en réorganiser, simplifier et compléter les contenus.

L'objectif premier de **tuto-R** est de présenter comment réaliser des analyses statistiques et diverses opérations courantes (comme la manipulation de données ou la production de graphiques) avec R. Il ne s'agit pas d'un cours de statistiques : les différents chapitres présupposent donc que vous avez déjà une connaissance des différentes techniques présentées. Si vous souhaitez des précisions théoriques / méthodologiques à propos d'un certain type d'analyses, nous vous conseillons d'utiliser votre moteur de recherche préféré. En effet, on trouve sur internet de très nombreux supports de cours (sans compter les nombreux ouvrages spécialisés disponibles en librairie).

Ce site est généré avec quarto, et le code source est disponible sur GitHub.

Pour toute suggestion ou correction, vous pouvez ouvrir un ticket GitHub.

Remerciements

Ce document a bénéficié de différents apports provenant notamment de l'*Introduction à R* de Julien Barnier et d'analyse-R : introduction à l'analyse d'enquêtes avec R et RStudio.

Merci donc à Julien Barnier, Julien Biaudet, François Briatte, Milan Bouchet-Valat, Ewen Gallic, Frédérique Giraud, Joël Gombin, Mayeul Kauffmann, Christophe Lalanne & Nicolas Robette.

Licence

partie I

Section

1 Statistique univariée

On entend par statistique univariée l'étude d'une seule variable, que celle-ci soit quantitative ou qualitative. La statistique univariée fait partie de la statistique descriptive.

Nous utiliserons dans ce chapitre les données de l'enquête *Histoire de vie 2003* fournies avec l'extension `{questionr}`.

```
library(questionr)
data("hdv2003")
d <- hdv2003
```

1.0.1 Principaux indicateurs

Comme la fonction `str()` nous l'a indiqué, notre tableau `d` contient plusieurs variables numériques ou variables quantitatives, dont la variable *heures.tv* qui représente le nombre moyen passé par les enquêtés à regarder la télévision quotidiennement. On peut essayer de déterminer quelques caractéristiques de cette variable, en utilisant les fonctions `mean()` (moyenne), `sd()` (écart-type), `min()` (minimum), `max()` (maximum) et `range()` (étendue) :

```
mean(d$heures.tv)
```

```
[1] NA
```

```
mean(d$heures.tv, na.rm = TRUE)
```

```
[1] 2.246566
```

```
sd(d$heures.tv, na.rm = TRUE)
```

```
[1] 1.775853
```

```
min(d$heures.tv, na.rm = TRUE)
```

```
[1] 0
```

```
max(d$heures.tv, na.rm = TRUE)
```

```
[1] 12
```

```
range(d$heures.tv, na.rm = TRUE)
```

```
[1]  0 12
```

On peut lui ajouter la fonction `median()` qui donne la valeur médiane, `quantile()` qui calcule plus généralement tout type de quantiles, et le très utile `summary()` qui donne toutes ces informations ou presque en une seule fois, avec en prime le nombre de valeurs manquantes (`NA`) :

```
median(d$heures.tv, na.rm = TRUE)
```

```
[1] 2
```

```
quantile(d$heures.tv, na.rm = TRUE)
```

```
  0%  25%  50%  75% 100%
   0    1    2    3   12
```

```
summary(d$heures.tv)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   1.000   2.000   2.247   3.000  12.000       5
```

La fonction `summary()` est une fonction générique qui peut être utilisée sur tout type d'objet, y compris un tableau de données. Essayez donc `summary(d)`.

1.0.2 Histogramme

Tout cela est bien pratique, mais pour pouvoir observer la distribution des valeurs d'une variable quantitative, il n'y a quand même rien de mieux qu'un bon graphique.

On peut commencer par un histogramme de la répartition des valeurs. Celui-ci peut être généré très facilement avec la fonction `hist` :

```
hist(
  d$heures.tv,
  main = "Nombre d'heures passées devant la télé par jour",
  xlab = "Heures",
  ylab = "Effectif"
)
```

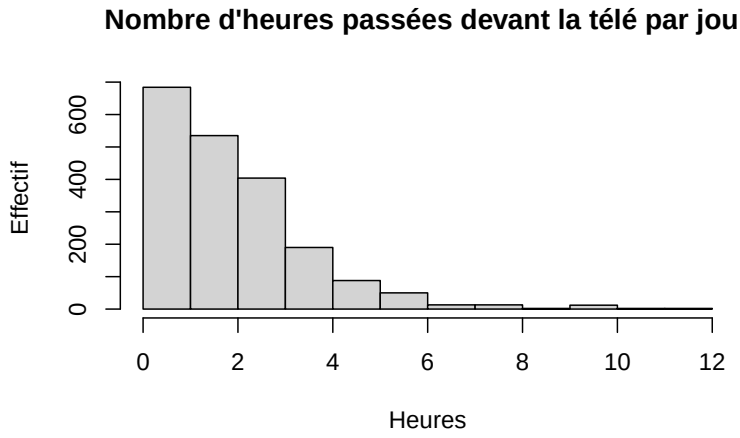

Figure 1.1: Exemple d'histogramme

Sous **RStudio**, les graphiques s'affichent dans l'onglet *Plots* du quadrant inférieur droit. Il est possible d'afficher une version plus grande de votre graphique en cliquant sur *Zoom*.

Ici, les options `main`, `xlab` et `ylab` permettent de personnaliser le titre du graphique, ainsi que les étiquettes des axes. De nombreuses autres options existent pour personnaliser l'histogramme, parmi celles-ci on notera :

- `probability` si elle vaut `TRUE`, l'histogramme indique la proportion des classes de valeurs au lieu des effectifs.
- `breaks` permet de contrôler les classes de valeurs. On peut lui passer un chiffre, qui indiquera alors le nombre de classes, un vecteur, qui indique alors les limites des différentes classes, ou encore une chaîne de caractère ou une fonction indiquant comment les classes doivent être calculées.
- `col` la couleur de l'histogramme[1].

Voir la page d'aide de la fonction `hist()` pour plus de détails sur les différentes options. Les deux figures ci-après sont deux autres exemples d'histogramme.

[1] Il existe un grand nombre de couleurs prédéfinies dans **R**. On peut récupérer leur liste en tapant simplement `colors()` dans la console, ou en consultant le document suivant : http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf.

```
hist(
  d$heures.tv,
  main = "Heures de télé en 7 classes",
  breaks = 7,
  xlab = "Heures",
  ylab = "Proportion",
  probability = TRUE,
  col = "orange"
)
```

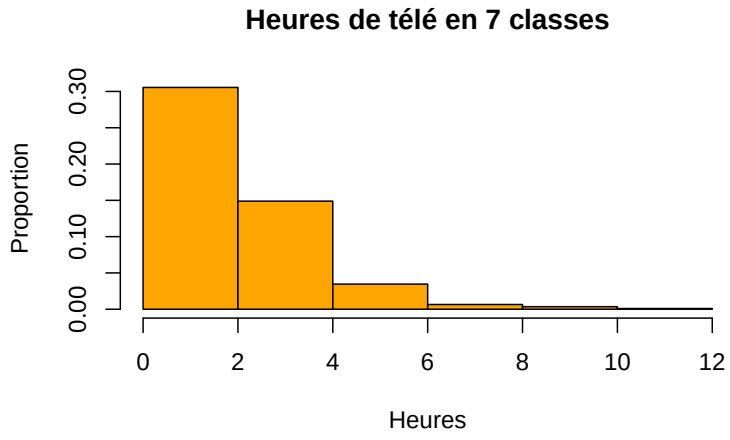

Figure 1.2: un autre exemple d'histogramme

1.1 Aller plus loin

https://youtu.be/oEF_8GXyP5c

Références