Bivariate analysis

2025-02-04

M1 MIDS/MFA/LOGOS
Université Paris Cité
Année 2024
Course Homepage
Moodle

Objectives

In Exploratory analysis of tabular data, bivariate analysis is the second step. It consists in exploring, summarizing, visualizing pairs of columns of a dataset.

Setup

```
stopifnot(
  require(tidyverse),
  require(glue),
  require(magrittr),
  require(lobstr),
  require(arrow),
  require(ggforce),
  require(vcd),
  require(ggmosaic),
  require(httr),
  require(patchwork),
  require(corrr),
  require(gapminder),
  require(slider)
)
```

Bivariate techniques depend on the types of columns we are facing.

For *numerical/numerical* samples

- Scatter plots
- Smoothed lineplots (for example linear regression)
- 2-dimensional density plots

For *categorical/categorical* samples : mosaicplots and variants

For *numerical/categorical* samples

- Boxplots per group

- Histograms per group
- Density plots per group

Dataset

Once again we rely on the `Recensement` dataset.

> Since 1948, the US Census Bureau carries out a monthly Current Population Survey, collecting data concerning residents aged above 15 from 150000 households. This survey is one of the most important sources of information concerning the american workforce. Data reported in file `Recensement.txt` originate from the 2012 census.

Load the data into the session environment and call it `df`. Take advantage of the fact that we saved the result of our data wrangling job in a self-documented file format. Download a `parquet` file from the following URL:

`https://stephane-v-boucheron.fr/data/Recensement.parquet`

> **Question**
>
> Download a `parquet` file from the following URL:
> `https://stephane-v-boucheron.fr/data/Recensement.parquet`

> - Use `httr::GET()` and `WriteBin()`
> - Use `download.file()`
> - Use `fs` to handle files and directories

> **Question**
>
> Load the data contained in the downloaded file into the session environment and call it `df`

```r
df |>
  glimpse()
## Rows: 599
## Columns: 11
## $ AGE        <dbl> 58, 40, 29, 59, 51, 19, 64, 23, 47, 66, 26, 23, 54, 44, 56,~
## $ SEXE       <fct> F, M, M, M, M, M, F, F, M, F, M, F, F, F, F, F, F, M, M, F,~
## $ REGION     <fct> NE, W, S, NE, W, NW, S, NE, NW, S, NE, NE, W, NW, S, S, NW,~
## $ STAT_MARI  <fct> C, M, C, D, M, C, M, C, M, D, M, C, M, C, M, C, S, M, S, C,~
## $ SAL_HOR    <dbl> 13.25, 12.50, 14.00, 10.60, 13.00, 7.00, 19.57, 13.00, 20.1~
## $ SYNDICAT   <fct> non, non, non, oui, non, non, non, non, oui, non, non, non,~
## $ CATEGORIE  <fct> "Administration", "Building ", "Administration", "Services"~
## $ NIV_ETUDES <fct> "Bachelor", "12 years schooling, no diploma", "Associate de~
## $ NB_PERS    <fct> 2, 2, 2, 4, 8, 6, 3, 2, 3, 1, 3, 2, 6, 5, 4, 4, 3, 2, 3, 2,~
## $ NB_ENF     <fct> 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ REV_FOYER  <fct> [35000-40000), [17500-20000), [75000-1e+05), [17500-20000),~

df |>
  head()
## # A tibble: 6 x 11
##     AGE SEXE  REGION STAT_MARI SAL_HOR SYNDICAT CATEGORIE     NIV_ETUDES NB_PERS
##   <dbl> <fct> <fct>  <fct>       <dbl> <fct>    <fct>         <fct>      <fct>
```

```
## 1     58 F     NE      C              13.2 non       "Administrat~ Bachelor   2
## 2     40 M     W       M              12.5 non       "Building "   12 years ~ 2
## 3     29 M     S       C              14   non       "Administrat~ Associate~ 2
## 4     59 M     NE      D              10.6 oui       "Services"    12 years ~ 4
## 5     51 M     W       M              13   non       "Services"    9 years s~ 8
## 6     19 M     NW      C               7   non       "Services"    12 years ~ 6
## # i 2 more variables: NB_ENF <fct>, REV_FOYER <fct>
```

Categorical/Categorical pairs

Question

Project the dataframe on categorical columns

Question

- Explore the connection between `CATEGORIE` and `SEX`.
- Compute the 2-ways contingency table using `table()`, and `count()` from `dplyr`.

- Use `tibble::as_tibble()` to transform the output of `table()` into a dataframe/tibble.
- Use `tidyr::pivot_wider()` so as to obtain a wide (but messy) tibble with the same the same shape as the output of `table()`.
- Can you spot a difference?

Question

Use `mosaicplot()` from base `R` to visualize the contingency table.

Question

Use `geom_mosaic` from `ggmosaic` to visualize the contingency table

- Make the plot as readable as possible
- Reorder `CATEGORIE` acccording to counts

Question

- Collapse rare levels of `CATEGORIE` (consider that a level is rare if it has less than 40 occurrences). Use tools from `forcats`.

Question

Same as above with `vcd::mosaic`

Testing association

Chi-square independence/association test

https://statsfonda.github.io/site/content/ch4_2.html#test-dindépendance

> **Question**
>
> - Compute the chi-square association statistic between `CATEGORIE` and `SEXE`.
> - Display the output of `chisq.test()` as a table, using `broom::tidy()`

Categorical/Numerical pairs

Grouped boxplots

> **Question**
>
> Plot boxplots of `AGE` according to `NIV_ETUDES`

> **Question**
>
> Draw density plots of `AGE`, facet by `NIV_ETUDES` and `SEXE`

> **Question**
>
> Collapse rare levels of `NIV_ETUDES` and replay.

Numerical/Numerical pairs

Scatterplots

> **Question**
>
> Make a scatterplot of `SAL_HOR`with respect to `AGE`

Correlations

- Linear correlation coefficient (Pearson)
- Linear rank correlation coefficient (Spearman ρ, Kendall τ)
- ξ rank correlation coefficient (Chatterjee)

Linear correlation coefficient

> **Question**
>
> Compute the Pearson, Spearman and Kendall correlation coefficients between `AGE` and `SAL_HOR` using function `cor()` from base `R`

Rank based methods

Spearman's rho ( ) and Kendall's tau ( ) are both non-parametric correlation coefficients used to measure the strength and direction of a *monotonic* relationship between two variables.

Spearman's rho ( ) Based on *rank differences*. Defined as the *Pearson correlation coefficient* between the *ranked variables*.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of each pair, and n is the number of observations.

Kendall's tau ( ) Based on *concordant and discordant pairs*. Measures the *proportion of pairs that have the same order* in both variables compared to the total number of pairs.

$$\tau = \frac{(C - D)}{\frac{1}{2} n(n - 1)}$$

where C is the number of *concordant pairs*, and D is the number of *discordant pairs*.

When to Use Which?

Factor	Spearman's rho ( )	Kendall's tau ( )
Large differences in ranks	More sensitive	Less sensitive
Small sample sizes	Less reliable	More reliable
Outlier resistance	Moderate	High
Computational efficiency	Faster	Slower (due to pairwise comparisons)
Interpretation	Similar to Pearson's correlation	More intuitive (proportion of concordance)

Chatterjee's correlation coefficient (Chatterjee's ξ)

> The three most popular classical measures of statistical association are Pearson's correlation coefficient, Spearman's , and Kendall's . These coefficients are very powerful for detecting linear or monotone associations, and they have well-developed asymptotic theories for calculating P-values. However, the big problem is that they are not effective for detecting associations that are not monotonic, even in the complete absence of noise.
>
> Let (X, Y) be a pair of random variables, where Y is not a constant. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. pairs with the same law as (X, Y), where $n \geq 2$. The new coefficient has a simpler formula if the X_i's and the Y_i 's have no ties. This simpler formula is presented first, and then the general case is given. Suppose that the X_i's and the Y_i 's have no ties. Rearrange the data as $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ such that $X_{(1)} \leq \cdots \leq X_{(n)}$. Since the X_i's have no ties, there is a unique way of doing this. Let ri be the rank of $Y_{(i)}$, that is, the number of j such that $Y_{(j)} \leq Y_{(i)}$. The new correlation coefficient is defined as

$$\xi_n(X, Y) := 1 - 3 \sum_{i=1}^{n-1} \frac{|r_{i+1} - r_i|}{n^2 - 1}$$

In the presence of ties, ξ_n is defined as follows. If there are ties among the X_i's, then choose an increasing rearrangement as above by breaking ties uniformly at random. Let r_i be as before, and additionally define l_i to be the number of j such that $Y_{(j)} \geq Y_{(i)}$. Then define

$$\xi_n(X, Y) := 1 - 3n \sum_{i=1}^{n-1} \frac{|r_{i+1} - r_i|}{2\sum_{i=1}^{n} l_i(n - l_i)}$$

When there are no ties among the Y_i 's, $l_1, \ldots, l_n$ is just a permutation of $1, \ldots, n$, and so the denominator in the above expression is just $n(n^2 - 1)/3$, which reduces this definition to the earlier expression.

From Sourav Chatterjee: A new correlation coefficient

> **Question**
>
> Write a `dplyr` pipeline from computing the ξ correlation coefficient between `Y=lifeExp` and `X=gdpPercap` in the `gapminder` dataset, per `year` and `continent`.

Using package `corrr`

https://corrr.tidymodels.org

> **Question**

`pairs` from base `R`

Just as function `skim` from package `skimr` allows us to automate univariate analysis, function `pairs` from base `R` allows us to automate bivariate analysis.

> **Question**

`ggpairs()`

Useful links

- rmarkdown
- dplyr
- ggplot2
- *R Graphic Cookbook*. Winston Chang. O' Reilly.
- A blog on ggplot object
- `skimr`
- `vcd`
- `ggmosaic`
- `ggforce`
- `arrow`
- `httr`