Univariate analysis I

2024-09-02

- **M1 MIDS & MFA**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

Objectives

In Exploratory Data Analysis of tabular data, univariate analysis is the first step. It consists in exploring, summarizing, visualizing columns of a dataset. In this workbook we focus on univariate numerical samples. We explore techniques for:

- Summarizing univariate numerical samples
- Displaying numerical samples

This is also an opportunity to:

- Introduce to the General Social Survey
- Use packages `gssr` and `gssrdoc`

Setup

If the required packages have not (yet) been installed, install them.

```
stopifnot(
  require(skimr),   # Univariate summaries from the shelf
  require(lobstr),  # R introspection
  require(rlang),   # R introspection
  require(glue),    # Like formatted strings
  require(gssr),
  require(gssrdoc),
  require(fs),      # File manipulation
  require(patchwork), # piecing ggplots together
  require(tidyverse) # What else?
)
```

General Social Survey (GSS) dataset

Question

Load the cumulative GSS dataset (`gss_all`). Have a glimpse at the resulting dataframe. Load `gss_dict`.

Question

- In dataset `gss_all`, what do the rows stand for?
- In dataset `gss_all` what do columns `year` and `id` stand for?
- For a given value of `id`, can you find several rows ?
- For a given value of `year`, can you find several rows with the same `id`?
- How many distinct values of `year` can you find in `gss_data`?
- For each value of `year`, how many people were surveyed?
- Why is this dataset called *cumulative*?

Question

Table exploration

Load `gss_sub` which is much smaller than `gss_all`. Have a glimpse.

Question

- How many variables can you find in `gss_sub`?
- How many distinc values for each column?
- Which columns should be considered as categorical/factor?

Caveat

In the sequel, we explore the `age` distribution as is the `age` column was a genuine univariate sample. This is done for teaching purpose. The `age` column is not collected by repeatedly picking individuals uniformly at random from a fixed population.
Indeed the `age` column is a union of samples collected every year or every two years since 1972. The American population has changed thoughout the five decades.
Moreover, yearly samples are not i.i.d. samples from the whole population. The sampling methods have varied over time. Sampling methods rely on multistage stratified sampling and quotas.

Exploring age column

Question

For column age, disregarding any weighting process

- compute the summary.
- compute the range, the IQR, the standard deviation
- compute the Mean Absolute Deviation, the Median Absolute Deviation

Filter out rows with missing data in columns age or sex

```r
gss_fil <- gss_sub |>
  filter(!is.na(age), !is.na(sex))
```

Boxplots

Question

- Build a boxplot for age.
- Equip the plot with a title, a subtitle, a caption
- Annotate the boxplot with summary statistics.

Question

- Build a boxplot of age distribution according to sex.
- What is the impact of argument varwidth=T?
- What is the impact of argument notch=T?
- What is the difference between stat_boxplot() and geom_boxplot()?
- How would you get rid of the useless ticks on the x-axis?

Histograms

Question

- Plot a histogram of the age distribution
- Facet by sex
- Draw the age distribution histograms for each sex on the same plot
- Facet by sex and year
- Build an animated histogram plot where frame is determined by year

Histograms are used to sketch possibly (absolutely) continuous distributions by using piecewise constant approximations of density functions. Histograms can also be viewed as column plots for binned data (that is discretizations of “continuous” data).

Question

- Define *breaks* for `age` data
  - regular breaks with age ranges of length 5
  - irregular breaks `[18-25[`, `[25, 35[`, `[35,50[`, `[50, 65[`, `[65,+∞[`
- Bin `age` according to defined breaks using `cut()`
- Plot the binned data using `geom_bar()` or `geom_col()`

Demographers use *population pyramids* to sketch the age distribution in a population. Population pyramids are special facetted histograms or barplots.

Question

- Plot an age-sex pyramid for the `gss` sample.
- Animate with respect to `year`

Density plots

Histograms deliver piecewise constant estimations/approximations of a population density. If we suspect the population density to be *smooth*, it is sensible to try to build smooth estimates/aproximations of the population density. This is the purpose of density estimates.

Question

- Draw density plots for age distribution
- Use different bandwidths
- Use different kernels
- Facet by `sex`
- Facet by `sex` and `year`
- Overlay histograms and density plots (in `geom_histogram()` use `aes(y=after_stat(density))`)

Use `stat_density()`

Question

Build violine plots for `age` distribution (use `geom_violine()`).

Cumulative Distribution Functions

Not all probability distributions have densities, but all are characterized by their Cumulative Distribution Functions (CDFs). Each sample defines an Empirical Cumulative Distribution Function (ECDF).

Question

- Plot the `age` ECDF using `stat_ecdf()`
- Facet by `sex`, then by `year` and `sex`
- Use base `R` `ecdf()` and `stat_function()` to draw the same plot.

Question ☕

- Compare the `age` distributions for women and men using the Kolmogorov-Smirnov statistic (`ks.test()`)
- How is the Kolmogorov-Smirnov statistic computed?

Quantile plots

The *quantile function* of a probability distribution is the (generalized, left-continuous) inverse of its CDF. Quantile functions are useful devices in EDA and random generation.

Question

- Plot the quantile function of the `age` empirical distribution
- Plot the quantile functions of the `age` empirical distributions for men and women
- Design a function that takes as input a univariate numerical sample and returns the quantile function (in the same way as `ecdf()` does)

Question

- Draw a quantile-quantile plot to compare `age` distribution for women and men with base `R` `qqplot()`
- ☕ Draw a quantile-quantile plot to compare `age` distribution for women and men using `ggplot2`.

How could you comply with the DRY principle ?

Lazy loading and labelled format

Question

- What is `R` *lazy loading*?
- What is the *labelled format* used the GSS data?

References

- rmarkdown
- dplyr

- ggplot2
- *R Graphic Cookbook*. Winston Chang. O' Reilly.
- A blog on ggplot object
- Package `skimr`
- Package `gssr`
- Package `gssrdoc`
- General Social Survey
- Data gathering and processing from Statistics Canada