Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 4. Data visualization for exploratory analysis. (22 November)
Alla Tambovtseva

Random variables: recall

A **random variable** describes a result of a random experiment. For instance, we can take a random variable that describes the number of scores we get after a single throw of a dice, a random variable that describes the amount of money a person can win in a lottery, etc. In the probability theory course you discussed discrete and continuous random variables. **Discrete random variables** are variables that take a finite set of values. **Continuous random variables** are variables that take an infinite set of values.

Question 1. *Provide your examples of discrete and continuous random variables.*

A **distribution** of a variable is a correspondence between values of this variable and their probabilities.

Example 1. We can present a distribution of a discrete random variable using a table:

X	-1	1	2	3	3.5	4	5	8
p	0.1	0.1	0.2	0.05	0.05	0.3	0.15	0.05

This table is enough to know everything about the variable X: we can calculate any probabilities, compute the expected value and the variance, find the most probable value, and so on.

Example 2. For continuous random variables we cannot make a distribution table because of two features: firstly, the set of values is infinite, secondly, the probability that a variable takes a particular value equals 0. Hence, continuous variables are usually defined by cumulative distribution functions (cdf) [1] and probability density functions (pdf).

However, instead of going further to functions, now we will describe a variable splitting values into groups and calculating the probability of falling into each group. We will plot a **histogram**, a graph that depicts the correspondence between values on a certain interval and the probability of being there.

Consider the discrete variable X defined above. To plot a histogram we should pay attention to two things: a starting point and a size of grouping intervals. Usually a minimal value is taken as a starting point or a value that is slightly less than a minimal one. A size of intervals is often adjusted by a statistical packages automatically, but it good when an interval width is informative, not too small and not too large. Sometimes researchers choose intervals of a size equal to one standard deviation of a variable (a

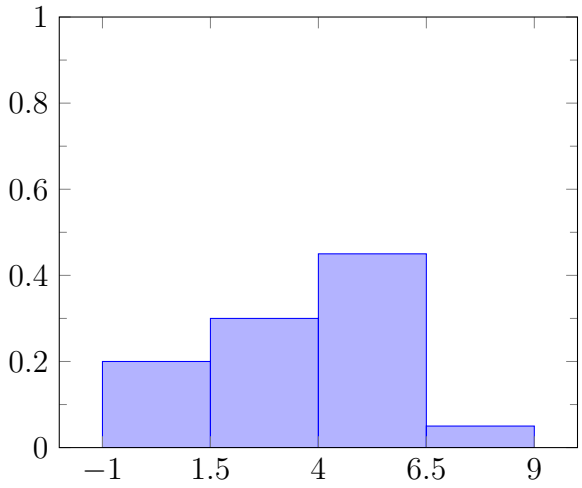
[1] Cumulative distribution functions are also called probability distribution functions, but the term with 'cumulative' is more common since it is convenient to distinguish between cdf and pdf.

sample). That is particularly sensible in case of normal distribution (recall the three-sigma rule).

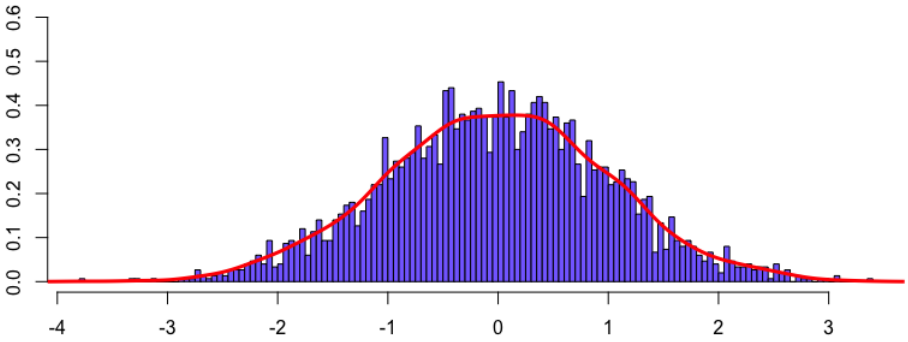
Let's take four intervals[2] of width=2.5: [−1, 1.5), [1.5, 4), [4, 6.5), [6.5, 9). Now we can calculate the probabilities of falling into each interval:

$$\begin{array}{ll} [-1, 1.5) & p = 0.2 \\ [1.5, 4) & p = 0.3 \\ [4, 6.5] & p = 0.45 \\ [6.5, 9 & p = 0.05 \end{array}$$

So, now we are ready to plot a histogram:



The same thing we can perform with continuous variables. And if we go further, we will see that it is possible to regard a density function plot as a smoothed line contouring the histogram with a number of columns approaching infinity (see below).



[2]They are half-open intervals; a square bracket means that an endpoint is included in an interval and a round one means that an endpoint is excluded.

Further we will plot histograms for sample distributions and the logic will be the same. However, instead of probabilities we will use relative frequencies or absolute frequencies (just counts).

Data visualisation[3]

Some principles of good visualisation (common sense, but still important):

- Clarity: illustration should facilitate understanding. If a picture hinders understanding, it is better to improve it or get rid of it.
- Readability: visualisation should be clear and readable. It ought to have a title, axes names and units of measurement.
- Adequacy: correspondence between an aim and a graph type.
- Correctness: correspondence between a scale and a graph type, correct scaling of axes.

Visualisation of nominal data

Ways to visualise a distribution of a nominal variable:

- Table
- Bar chart
- Pie chart

Visualisation of quantitative data

Quantitative data: data in ratio or interval scales. Ways to visualise a distribution of a quantitative variable:

- Histogram
- Box plot
- Violin plot (bean plot)

Visualisation of ordinal data

Depends on the number of categories: if it is large, an ordinal variable can be treated as a numeric one, if it is small (3 or 4, for instance), it should be handled as a nominal variable.

[3]For examples of graphs see lectures on R programming.