Politics. Economics. Philosophy, 2018-2019
Data Analysis in the Social Sciences
Lecture 3. Data types. (15 November)
Alla Tambovtseva

Data: basic terms

In the social sciences these terms are often interchangeable: *a table*, *a data set*, *a data frame*, *a database*. In computer science or in data analysis *a database* is usually understood as a set of tables somehow related to each other. However, in this course we will use this term as a synonym to a single table.

In data analysis rows of a table are called *observations*[1] and columns are called *variables* (or *features* as it is defined in machine learning). When we want to describe a table, we indicate the number of rows and the number of columns. These numbers are called *dimensions*, and the pair (# rows, # columns) is the *dimensionality* of a table.

id	age	gender	male
1	26	F	0
2	33	M	1
3	19	M	1
4	45	F	0
5	21	M	1

Table 1: Let's call it table `T`

The dimensions of the table 1 are (5, 4) or 5×4. If we choose a value 19 from this table, we can say that it is `T[3,2]` since again we indicate the row's number first, and then the column's number.

In data analysis data sets are usually presented in the way described above. However, sometimes data can look differently, so indicators can go by rows and observations – by columns. It may cause difficulties since in statistical packages columns are regarded as variables and rows are regarded as observations. To avoid it, we need *transposition*, an operation that switches rows and columns. This operation "reverts" indices of rows and columns, so if we choose a cell `T[3,2]`, in the table `S` that is a transpose of `T` it will be `S[2,3]`. The example of a transposition can be seen below.

1	2	4
0	8	9

$\xrightarrow{\text{transpose}}$

1	0
2	8
4	9

Table 2: Transposition

[1] Sometimes observations are called *cases*, but this term is more applicable in the context of qualitative research rather than quantitative one.

Data types (scales)

Data scales should not be confused with data types in R or in programming in general. They are related, but data types in programming are defined by their form (numbers, texts, logical values) while data scales cover substantial features, what do variables actually mean. In other words, a variable `gender` in R can be numeric and take values 0 and 1, but its scale is definitely nominal (qualitative) because these numbers are just our artificially constructed codes, not quantities that we can compare.
There are four data scales:

- Nominal (qualitative) scale
- Ordinal scale
- Interval scale
- Ratio (absolute) scale

Although this classification is sometimes criticised, it is still popular.

Nominal scale

Corresponds to data that cannot be measured, i.e. qualitative information, non-numeric categories. The distinctive feature of a nominal variable is that we cannot compare its values to each other. This feature stems from the fact that qualitative categories cannot be arranged in an ascending or descending order. To understand it, let us consider several examples.

Example 1. A type of a political regime (authoritarian, democratic, hybrid) is a nominal variable. Although we can assign numeric values to every type, for instance, autocracy – 1, democracy – 2, hybrid – 3, it is impossible to conclude that 3 is greater than 2, and 2 is greater than 1. These numbers are just our codes replacing regime types (texts) and there is no sense in arranging them from the minimum value to the maximum one.

Example 2. Other examples of nominal variables: gender, professions, bird species, regions, forms of government (monarchy, republic), master specialisations, colours of cars.

Ordinal scale

Corresponds to data that can be organised in categories comparable to each other, but the differences between every pair of subsequent values are not equal. In other words, for an ordinal variable we often can decide which value is greater, but we cannot be sure that one unit increase is the same for all values. Consider the following example as an illustration.

Example 3. In the HSE students' grades vary from 1 to 10. Grades from 1 to 3 are unsatisfactory, 4 and 5 are satisfactory, 6 and 7 are good, and 8, 9, 10 are excellent. It is obvious that 10 is better than 9, and 3 is worse than 4. However, there is a drastic difference between these two pairs of values. The difference between 9 and 10 is one score, but it not so significant as the same difference between 3 and 4. If a student gets 10, it

stands for "fantastic, brilliant, genius" and if he gets 9, it is "still excellent, but there are some outstanding people who got 10". The difference between 3 and 4 is more dramatic: if a student gets 3, he fails, if 4 – passes.

Example 4. Other examples of ordinal variables: extent of agreement with a statement (from *completely disagree* to *completely agree*), rating of countries based on the Corruption Perception Index, Doing business rating.

Sometimes it is not clear how to distinguish between ordinal and nominal scales. Consider the variable *type of economy* that takes values "not developed", "developing" and "well-developed". As we can arrange values from the worst to the best, it can be regarded as an ordinal variable. However, if the variable *type of economy* takes values "traditional economy", "command economy", "market economy", 'mixed economy" , it will be a nominal one since it is impossible to place these categories in a certain order.
The following two scales are often combined into one more general type called *quantitative scale*. Usually in data analysis there is no crucial difference in handling these two scales, but sometimes their features can be important, e.g. in interpretation of model results.

Ratio (absolute) scale

Corresponds to results of direct measurements, so values of ratio variables can be treated as numbers in mathematics. We can add, subtract, divide and multiply them, and these operations will make sense. The distinctive feature of the ratio scale is the presence of a natural zero that corresponds to the absence of some quantity (or feature).

Example 5. Take the variable *income* (in $), for example. It is clear that income is the result of direct measurements. A person can simply count the amount of money he or she gets, and the number obtained will be "physical", not a result of a convention [2] If income equals 0, it definitely means that the person does not get money (unemployed, pensioner, under the working age).

Another important feature of this scale stems from its name. The ratios calculated from two values are meaningful. For instance, if my income is 200$ and my neighbour's income is 100$, it is correct to say that I earn twice as large as my neighbour ($200/100 = 2$, and this ratio makes sense).

Example 6. Other examples of ratio variables: age, GDP, population of a country, number of children in a family.

Interval scale

Usually corresponds to artificially constructed scales that refer not to "objective" numbers gained from measurements, but to results of a particular convention. Unlike ratio variables, if an interval variable takes value of 0, it does not mean the absence of a feature or a quantity. Most indices proposed by researchers are interval.

[2] For instance, such a convention: from -100 to 0 if I earn less than my neighbour, from 0 to 100 if I earn more than my neighbour, and the exact value depends on the size of difference between our incomes.

Example 7. Polity2 index from Polity IV takes values from -10 to 10, where greater values correspond to a higher level of democracy. The value of 0 does not mean the absence of democracy, it is just the value in the middle of the scale.

Ratios for interval variables can be computed, but they are meaningless. If we know that for Armenia Polity2 index equals 5 and for Spain it is 10, it does not mean that Spain is twice more democratic since we cannot verify it or measure directly.

Example 8. Other examples of interval variables: time (0 year is the result of convention, A.D. and B.C.), Corruption Perception Index, temperature in Celsius scale.

Descriptive statistics

Nominal variables

What descriptive statistics are applicable for nominal variables?

As there is no sense in treating nominal variables as numeric ones, it is not meaningful to calculate means or variances. However, there are still characteristics that can be used for description.

- Frequencies: frequencies can be absolute (just counts) or relative (shares or %).
- Mode: the most frequent value of a variable.

Quantitative variables

What descriptive statistics are applicable for quantitative variables?

Basic statistics

- Minimum value: min;
- Maximum value: max;
- Range: $\max - \min$;

Measures of central tendency

- Average: $\bar{x} = \dfrac{sum(x)}{n}$, where n is the sample size;
- Median (see explanation below)

Variability in data

- Sample variance: $s^2 = \dfrac{sum(x - \bar{x})^2}{n - 1}$;
- Sample standard deviation: $s = \sqrt{s^2}$;
- Coefficient of variation: used to evaluate a level of variability in data (judging by variance and standard deviation we cannot conclude whether the variability is high or not).

  $$CV = \frac{sd(x)}{\bar{x}}$$

  Interpretation: if $|CV| < 0.3$, the variability in data is low, if $0.3 \leqslant |CV| \leqslant 0.7$, the variability is medium, and if $|CV| > 0.7$, the variability is high.

Quantiles [3]

Quantile of the level p is the value that other values in a sample do not exceed with the probability p (the probability here is considered as a relative frequency).

Example 9. There is a sample X:

8 7 3 0 1 2 6 9 12 9

So as to calculate quantiles by hand, we have to arrange all values in an ascending order:

0 1 2 3 6 7 8 9 9 12

Now let's find the quantile of the level 0.2 or $q_{0.2}$. Here $q_{0.2} = 1$ since 20% of values in the sample (2 out of 10) do not exceed 1.[4]

Example 10. If we know that 32 is the quantile of level 0.4 of the variable *age* in our data set, we can conclude that 40% of people in a data set are not older than 32.

There are quantiles of specific levels (25%, 50%, 75%, 100%) that are called **quartiles**. This term stems from the fact that quartiles divide a sample into four equal parts (see figure 1).

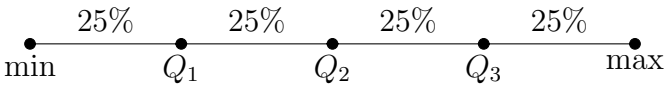

Figure 1: Quartiles

$\boldsymbol{Q_1}$ is **the lower quartile** (1st quartile), a value that separates the first 25% of values, $\boldsymbol{Q_3}$ is **the upper quartile** (3rd quartile), a value that separates the first 75% of values. $\boldsymbol{Q_2}$ is not usually called the 2nd quartile, it is called **the median** since it divides a sample into two parts, the first 50% and the last 50% of observations.

Example 11. If we know that for the variable *salary*:

- 1st quartile: 18000 RUR
- median: 35000 RUR
- 3rd quartile: 52000 RUR,

we can conclude that 25% of respondents earn no more than 18000 roubles, 50% of respondents earn no more than 35000 roubles, and 75% of respondents earn no more than 52000 roubles (or 25% of people earn more than 52000 roubles.

[3] Here we discuss sample quantiles that are estimates of population quantiles evaluated on the data given.

[4] Of course, sometimes we might have samples where it is not so definite because of the sample size. These cases are handled differently, they will can be found in extra materials for the course.

Using quartiles we can calculate **the interquartile range**, the measure of variability that is more sustainable to extremely large/small values in a sample compared to an "ordinary" range. The interquartile range is computed as follows:

$$IQR = Q_3 - Q_1.$$

Example 12. Consider a sample (already sorted):

2 2.5 2.8 3 3.4 4.8 5.2 5.3 7.1 8.2 8.8 100

If we make conclusions based on the range, we will decide that our data variate significantly ($range = max - min = 98$). However, this result is due to the only value that is too large. If we calculate the interquartile range, the result will reflect the situation in a more correct way ($IQR = Q_3 - Q_1 = 5.6$). The interquartile range is not too high, and the values do not vary significantly if we exclude 100.

Ordinal variables are usually handled depending on the number of levels (categories, unique values) they have. If a number of categories is high, a variable can be treated as a numeric one, otherwise, it should be treated as nominal. There are no certain conventions on which number of levels is high or low, it is different for different sciences, but 10 categories and more can be enough to handle an ordinal variable as numeric.