DESCRIPTIVE STATISTICS

Dr Alina Gleska

Institute of Mathematics, PUT

25 lutego 2018

We consider two main types of statistics:

- descriptive statistics (one- and multidimensional),
- mathematical (inferential) statistics.

DESCRIPTIVE STATISTICS provides methods for:

- design: planning and carrying out research studies,
- description: summarizing and exploring data.

Descriptive statistics includes the construction of graphs, charts, and tables, and the calculation of various descriptive measures such as averages, measures of variation, and percentiles.

MATHEMATICAL (INFERENTIAL) STATISTICS provides methods for inference: making predictions and generalizing about phenomena represented by the data. Inferential statistics includes methods like point estimation, interval estimation and hypothesis testing which are all based on probability theory.

Begin

↓

Formulate the research problem

↓

Define the population and the sample

↓

Collect the data

↓

Do descriptive data analysis

↓

Use appropriate statistical methods to solve the research problem

↓

Report the results

↓

End

POPULATION - the collection of all individuals or items under consideration in a statistical study. (Weiss, 1999) It can be:

- finite, or
- infinite.

SAMPLE - that part of the population from which information is collected. (Weiss, 1999)
VARIABLE - a characteristic that varies from one individual member of the population to another. Examples of variables for humans are height, weight, number of siblings, sex, marital status, and eye color.

Classification of variables:

- measurable (quantitative, numerical) and qualitative (non-numerical),
- continuous, quasi-continuous and discrete,
- simple and compound,
- stimulants, destimulants and nominants (neutral).

Scales:

- nominal scale (for qualitative variables without the natural ordering),
- order scale (also for qualitative variables but with the ordering; e.g. education),
- interval scale (the classic example is the Likert's scale: Strongly agree; Agree somewhat; I am not sure; Disagree somewhat; Strongly disagree),
- ration scale.

Quantitative variables, whether discrete or continuos, are defined either on an interval scale or on a ratio scale. If one can compare the differences between measurements of the variable meaningfully, but not the ratio of the measurements, then the quantitative variable is defined on interval scale. If, on the other hand, one can compare both the differences between measurements of the variable and the ratio of the measurements meaningfully, then the quantitative variable is defined on the ratio scale. In order to the ratio of the measurements being meaningful, the variable must have natural meaningful absolute zero point, i.e, a ratio scale is an interval scale with a meaningful absolute zero point. For example, temperature measured on the Centigrade system is an interval variable and the height of person is a ratio variable.

Classification of statistical series by SOBCZYK (1998)
Statistical series:

- count data
- collected (grouped) data
  - measurable variable
    - points (category)
    - interval
  - non-measurable variable
- time series (dynamical)
  - by moment
  - by period
- dimensional (geographically)

How to construct grouped data series for continuous variables:

1) first we have to find two observation: one with the minimal value x_{min} and one with the maximum value x_{max},
2) then we calculate the range $R = x_{max} - x_{min}$,
3) we establish the number of class k using one of the following formulas: $k \approx \sqrt{n}$, $k \approx \frac{3}{4}\sqrt{n}$, $k \leq 5\log(n)$, $k \approx 1 + 3,322\log(n)$ (H. A. Sturges 1926),
4) we calculate the width of the interval d by the formula: $d \approx R/k$ (we do it rather without overlapping, but sometimes it is necessary to take a little longer interval (we round up the width to the nearest integer),

5) we take as a lower limit x_{min} or a fixed number (but it is necessary that x_{min} belongs to the first interval), and we take as an upper limit x_{max} or a fixed number (but it is necessary that x_{max} belongs to the last interval),
6) we construct k intervals with the width d, and then we count observations belonging to the proper intervals.

RULES FOR GROUPED SERIES CONSTRUCTING:

- let us group observed values of numerical variable in data into 5 to 15 class intervals. A smaller number of intervals is used if the number of observations is relatively small; if the number of observations is large, the number of intervals may be greater than 15,
- classes should cover all data,
- classes should be separated (nonoverlapping),
- classes should be no empty,
- the width of the each interval should be the same; in case of very large observation (we call them outliers) we just leave the first and the last interval open,
- observations in the same class should be homogeneous.

Notations: 20-49,9; 50-79,9; 80-109,9 etc. mean that we have closed-open intervals $[\ldots;\ldots)$. Respectively: 20,1-50; 50,1-80; 80,1-110 etc. mean that we have open-closed intervals $(\ldots;\ldots]$.

The number x_i^0 in the middle of the i-th class is called the class mark of the i-th class.

The number of observations that fall into particular class (or category) of the qualitative variable is called the frequency (or count) of that class and denoted by n_i. A table listing all classes and their frequencies is called a frequency distribution.
A cumulative frequency (count) is obtained by summing the frequencies of all classes up to the specific class:

$$n_i^{cum} = \sum_{j=1}^{i} n_j$$

for $i = 1, 2, \ldots, k$. In a case of qualitative variables, cumulative frequencies make sense only for ordinal variables, not for nominal variables.

In addition of the frequencies, we are often interested in the percentage of a class. We find the percentage by dividing the frequency of the class by the total number of observations and multiplying the result by 100. The percentage of the class, expressed as a decimal, is usually referred to as the relative frequency of the class.

$$\text{Relative frequency of the class} = \frac{\text{Frequency in the class}}{\text{Total number of observation}},$$

$$\omega_i = \frac{n_i}{n}.$$

A table listing all classes and their relative frequencies is called a relative frequency distribution. The relative frequencies provide the most relevant information as to the pattern of the data. Relative frequencies sum to 1 (100%).

Cumulative relative frequency we calculate as

$$\omega_i^{cum} = \sum_{j=1}^{i} \omega_j.$$

SOME MISTAKES

MAKING DURING GROUPED SERIES CONSTRUCTION

1) If the population is nonhomogeneous and we have many observations in some particular classes then using intervals with the same width is not correct.

2) The comparison of two populations when they have different numbers of observations.

Graphical representation of data

The qualitative data are presented graphically either as a pie chart or as a horizontal or vertical bar graph.

A pie chart is a disk divided into pie-shaped pieces proportional to the relative frequencies of the classes. To obtain angle for any class, we multiply the relative frequencies by 360 degrees, which corresponds to the complete circle.

A horizontal bar graph displays the classes on the horizontal axis and the frequencies (or relative frequencies) of the classes on the vertical axis. The frequency (or relative frequency) of each class is represented by vertical bar whose height is equal to the frequency (or relative frequency) of the class. In a bar graph, its bars do not touch each other. At vertical bar graph, the classes are displayed on the vertical axis and the frequencies of the classes on the horizontal axis.

Nominal data is best displayed by pie chart and ordinal data by horizontal or vertical bar graph.

A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable) and was first introduced by Karl Pearson. It is a kind of bar graph. To construct a histogram, the first step is to divide the entire range of values into a series of intervals — and then count how many values fall into each interval. The bins are usually specified as consecutive, non-overlapping intervals of a variable. The bins (intervals) must be adjacent, and are often (but are not required to be) of equal size.

If the bins are of equal size, a rectangle is erected over the bin with height proportional to the frequency — the number of cases in each bin. A histogram may also be normalized to display relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the sum of the heights equaling 1.
As the adjacent bins leave no gaps, the rectangles of a histogram touch each other to indicate that the original variable is continuous.
Histograms are sometimes confused with bar charts. A histogram is used for continuous data, where the bins represent ranges of data, while a bar chart is a plot of categorical variables. Some authors recommend that bar charts have gaps between the rectangles to clarify the distinction.

When a variable is continous, one can choose class intervals in the frequency distribution and for the histogram as narrow as desired. Now, as the sample size increases indefinitely and the number of class intervals simultaneously increases, with their width narrowing, the shape of the sample histogram gradually approaches a smooth curve. We use such curves to represent population distributions.

Distributions