DESCRIPTIVE STATISTICS

Dr Alina Gleska

Institute of Mathematics, PUT

April 20, 2018

We consider the following statistical measures:

- measures of locations (also called central tendency) - the descriptive measures that indicate where the center or the most typical value of the variable lies in collected set of measurements;
- measures of statistical dispersion (also called measures of variation) - they numerically measure the extent of variation around the center. Two data sets of the same variable may exhibit similar positions of center but may be remarkably different with respect to variability;
- measures of the shape of the distribution like skewness (asymmetry);
- measures of concentration (kurtosis).

We consider two types of measures of central tendency: classical and depending on the position in the series.

Calculating classical measures we use values of all observations. These measures consist of: the arithmetic mean (average), the geometric mean and the harmonic mean.

Positional measures cover: the mode (the dominant value), the median and the quartiles (specially the first quartile, the second quartile (it is a median) and the third quartile).

ARITHMETIC MEAN - it is the sum of all observations divided by n. We distinguish different types of the arithmetic averages:

(I) the simple arithmetic mean for simple series:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

where x_i is a value of i-th observation, and n - the total number of observations,

(II) the weighted average, for grouped data, which is defined by the formula:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i,$$

where

a) for discrete grouped data x_i is the value of the i-th observation with the frequency n_i,

b) for the continuous data, grouped in the intervals x_i is the center of i-th interval (denoted by x_i^0) with the frequency n_i, and k is the number of intervals.

The properties of the arithmetic mean:

- the sum of the values of all observations is equal to the product of the arithmetic mean by the total number of observations:

$$\sum_{i=1}^{n} x_i = n\overline{x},$$

- the mean is the only single number for which the residuals (deviations from the estimate) sum to zero:

$$\sum_{i=1}^{n} (x_i - \overline{x}) = 0,$$

- the sample mean is also the best single predictor in the sense of having the lowest root mean squared error:

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \min.$$

Remark

- The arithmetic mean for data grouped in the intervals is only the approximation of the true average value because all observations are just represented by the centers of the intervals.
- The center of the interval IS NOT the average value - the differences between values and the center of the interval can be sometimes quite big.

Remark

- The mean value we calculate for closed intervals. But if we have outliers we have to leave last (or first) interval open. In such a case, if outliers are less than 5% of the total number of observations, we can close these open intervals and calculate the arithmetic mean.
- The arithmetic mean is not a robust statistic (a statistic is said to be robust if it is not sensitive to outliers).

RESUME:

- we calculate the arithmetic mean if:
  a) the population is homogeneous; there are no outliers; the distribution is symmetric or a little asymmetric;
  b) the distribution is unimodal (there is only one maximum);
- we do not use the arithmetic mean if:
  a) the distribution is very strong asymmetric;
  b) the distribution is bimodal or multimodal;
  c) the distribution is U-shape.

The geometric mean is a classical measure used in special cases, mainly in time series analysis. We calculate it using the formula:

$$\overline{x_g} = \sqrt[n]{x_1 x_2 \dots x_n},$$

where $x_1, x_2, \dots, x_n$ denote the observations (so the geometric mean is defined as the nth root of the product of n numbers). It is often used for a set of numbers whose values are meant to be multiplied together or are exponential in nature, such as data on the growth of the human population or interest rates of a financial investment.
The geometric mean of growth over periods yields the equivalent constant growth rate that would yield the same final amount.

Example

Suppose an orange tree yields 100 oranges one year and then 180, 210 and 300 the following years, so the growth is 80%, 16.6666% and 42.8571% for each year respectively. Growing with 80% corresponds to multiplying with 1.80, so we take the geometric mean of 1.80, 1.166666 and 1.428571, i.e. $\sqrt[3]{1.80 \times 1.166666 \times 1.428571} = 1.442249$; thus the 'average' growth per year is 44.2249%. If we start with 100 oranges and let the number grow with 44.2249% each year, the result is 300 oranges.

The harmonic mean (sometimes called the subcontrary mean) – can be expressed as the reciprocal of the arithmetic mean of the reciprocals of the given set of observations:

$$\overline{x_H} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

This measure is used rather seldom.
For positive values of observations we have a relation:

$$\overline{x_H} \leq \overline{x_g} \leq \overline{x}.$$

Definition

For discrete data obtain the frequency of each observed value of the variable in a collection and note the greatest frequency.

1. If the greatest frequency is 1 (i.e. no value occurs more than once), then the variable has no mode.
2. If the greatest frequency is 2 or greater, then any value that occurs with that greatest frequency is called a sample mode of the variable.

For continuous data at first we find the interval with the greatest frequency and then we use the interpolating formula:

$$Mo = x_{ld} + \frac{(n_s - n_{s-1})}{(n_s - n_{s-1}) + (n_s - n_{s+1})} \cdot d,$$

where:

- s - the number of the interval with the greatest frequency,
- x_{ld} - the left end of the s-th interval,
- d - the width of the interval,
- n_s - the frequency of the s-th interval.

The practical rules of calculating the mode:

- finding of the mode is legitimate only in case of unimodal distributions (with one clear maximum);
- the interval of the mode and two successive intervals should be the same width;
- we do not calculate the mode for multimodal distributions.

How to find the mode graphically?

- Do a histogram;
- Plot two lines from the vertices of the highest rectangular joining them with the vertices of the successive rectangulars;
- Find the cross-point of those lines and throw it on the X-axis;
- Read an approximation of the mode on the X-axis.

The sample median of a quantitative variable is that value of the variable in a data set that divides the set of observed values in half, so that the observed values in one half are less than or equal to the median value and the observed values in the other half are greater or equal to the median value. To obtain the median of the variable, we arrange observed values in a data set in increasing order and then determine the middle value in the ordered list.

How to find the median?

Arrange the observed values of variable in a data in increasing order.

1. If the number of observation is odd, then the sample median is the observed value exactly in the middle of the ordered list.
2. If the number of observation is even, then the sample median is the number halfway between the two middle observed values in the ordered list.

In both cases, if we let n denote the number of observations in a data set, then the sample median is at position $\frac{n+1}{2}$ in the ordered list.

Remark

1. The median is the value of the middle observation, not its frequency.
2. The median is a robust statistic - it does not depend on outliers.
3. In a case of asymmetric distributions the median gives more information than the arithmetic mean.

How to find the median in a case of grouped data? For the category data:

1) find the position of the middle observation as $\frac{n}{2}$ (or $\frac{n+1}{2}$ for the series with an odd number of observations),
2) find the class where is this middle observation; this is the class in which the cumulative frequency reaches $\frac{n}{2}$ for the first time,
3) read the value of the proper category data.

For the continuous data we use the interpolating formula:

$$Me = x_{lm} + \frac{\frac{n}{2} - n_{m-1}^{cum}}{n_m} \cdot d_m$$

where Me - the median, x_{lm} - the left end of the median interval, n - the total number of observations, n_m - the frequency of the median interval, n_{m-1}^{cum} - the cumulative frequency of the interval preceding the median interval, d_m - the width of the median interval.

How to find the median graphically?

1. Do a histogram for cumulative frequency and plot the line of cumulative frequency.
2. Mark on the Y-axis $\frac{n}{2}$.
3. Plot from this point the horizontal line. Mark the point where this line crosses the line of the cumulative frequency.
4. Throw this cross-point on the X-axis. This is the approximated median.

We have some relations between the arithmetic mean, the mode and the median.

1. For symmetric distributions all those measures are equal:

$$\overline{x} = Mo = Me.$$

2. For right-skewed (asymmetric) distributions the mode is the smallest measure and the arithmetic mean is the greatest one:

$$Mo < Me < \overline{x}.$$

3. For left-skewed (asymmetric) distributions the mode is the greatest measure and the arithmetic mean is the smallest one:

$$\overline{x} < Me < Mo.$$

We have also so called Pearson equation: $Mo = 3Me - 2\overline{x}$. This relation is valid only for symmetric distributions or asymmetric distributions which are skewed in a very small extent. We will say more about it during the lecture on the measures of the skewness.

The quartiles of a ranked set of data values are the three points that divide the data set into four equal groups, each group comprising a quarter of the data. A quartile is a type of quantile. The first quartile Q_1 is defined as the middle number between the smallest number and the median of the data set. The second quartile Q_2 is the median of the data. The third quartile Q_3 is the middle value between the median and the highest value of the data set.

For discrete distributions, there is no universal agreement on selecting the quartile values.

- For simple series Q_1 is the value at $n/4$ position, and Q_3 is the value at $3n/4$ position (if some of them is between the observations we take their arithmetic mean),
- For grouped series with categories we do the same as for the median – we find the class in which the cumulative frequency reaches $n/4$ and $3n/4$ for the first time, and then we read the quartiles,
- For continuous data, grouped in the intervals, we use the interpolating formulas:

$$Q_1 = x_{lq_1} + \frac{\frac{n}{4} - n^{cum}_{q_1-1}}{n_{q_1}} \cdot d_{q_1}$$

where Q_1 - the first quartile, x_{lq_1} - the left end of the first quartile interval , n - the total number of observations, n_{q_1} - the frequency of the first quartile interval, $n^{cum}_{q_1-1}$ - the cumulative frequency of the interval preceding the first quartile interval, d_{q_1} - the width of the first quartile interval.

$$Q_3 = x_{lq_3} + \frac{\frac{3n}{4} - n^{cum}_{q_3-1}}{n_{q_3}} \cdot d_{q_3}$$

where Q_3 - the third quartile, x_{lq_3} - the left end of the third quartile interval, n - the total number of observations, n_{q_3} - the frequency of the third quartile interval, $n^{cum}_{q_3-1}$ - the cumulative frequency of the interval preceding the third quartile interval, d_{q_3} - the width of the third quartile interval.