Week 1: Descriptive statistics

Tamás Biró

October 6, 2008

1 Steps of research

- Defining the *phenomenon*, the **population** and the **sample**.
- **Data collection:** see L. Stowe's class; Moore and McCabe chapter 3; and everything you learn within your discipline, in general.
- Result: a huge amount of data, difficult to grasp what is going on. Have always a detailed look at the raw data.
- **Visualization**: visual processing of the data, to help get a first understanding.
  - Bar graph: typically for non-numeric (e.g., nominal/categorical) data.
  - Pie chart: when values are compared not only to each other, but also to the entire sample (when the sum of the values gives a meaningful 100%).
  - Histogram: for numeric data (by creating fewer or more categories, the shape of the histogram can change radically).
- **Descriptive statistics**: describe/summarize data without making further conclusions. Numerical processing of the data, in order to condense the huge amount of data into a few numbers that one is already able to understand.
- **Inferential statistics**: derive conclusions from data collected in the sample that are probably true for the entire population.
- Interpret your data.
- Draw conclusions to your discipline (formulate and refute theories, etc.).

2 Basic notions

- **Variable** $\times$ **case** $\rightarrow$ **value**.
- Now we concentrate on a single variable. Later on: connection between variables (e.g., correlations).
- Observed or empirical **distribution** of a variable: a function that gives for each possible value of the variable the number of cases that have produced this value.
- (Theoretical distribution: if we had the whole population/infinite measurements, what would be the empirical distribution? The distribution observed using a sample is an approximation to the theoretical distribution.)
- Overall **shape** of a distribution: symmetric, tails, skewed, etc.
- Outliers: error in data collection or in data entry? is there a special story behind them? are they integral part of the phenomenon being measured?
- **Density curve**: a curve approximating the shape of the distribution. Always positive, has area exactly 1 underneath it.
- **Cumulative proportion**: proportion of the observations (or of the density curve) that lies at or below a given value.
- Mathematical measures of the shape of the observed distribution, a.k.a. *statistics* derived from the empirical data:
  - Measuring the "center" of the distribution: mode, median, mean.
  - Measuring further characteristic points: minimum and maximum; quartiles, percentiles, etc. $\rightarrow$ *five-number summary* (Min Q_1 Med Q_3 Max; boxplot).
  - Measuring the "spread" of the distribution: range; interquartile range, semi-interquartile range; standard deviation.
- *Normal distribution* (a.k.a. Gaussian distribution): bell-shaped; inflexion points show standard deviation; 68–95–99.7 rule.

3 Definitions

First, let us understand how observed data are distributed:

- You are interested in the distribution of variable X in the population. You collect a sample of n elements (e.g., n individuals), these are your n cases. So you will have n measurements, that is, n values for variable X, one value for each case. Some cases might exhibit the same value, so a value of the variable may occur more times.

- Suppose your observations are: x_1, x_2,... x_n. These together form the *data set*.
- **Number of data**: n.
- **Absolute frequency** of value x: number of cases when the value x was observed in the data set (that is, during the data collection).
- **Relative frequency** of value x: the absolute frequency of value x divided by the number of data n. That is, the proportion of value x. Relative frequency can also be seen as (more or less) the *probability* of obtaining value x whenever you run an experiment to measure the value of variable X.
- (Empirical) **distribution** of the observed data (first visualized as a *histogram*, then approximated by a nicely smooth *density curve*): a function of x that gives you the absolute or relative frequency of the value x in the data set.
  If absolute frequency is used, then the area under the curve is n. If relative frequency is used, then the area under the curve is 1.
- **Cumulative proportion**: a function of x that gives you the absolute or relative frequency of values equal to or less than x in the data set.
  It is the sum of the area under the distribution function (density curve) from the beginning up to x.
  At the "left end", cumulative proportion is 0.
  If absolute frequency is used, then the value of cumulative proportion at the right end is n. If relative frequency is used, then it is 1.

Second, let us find some characteristic points of the distribution:

- **Mode**: the value that has been observed the most frequently. That is the value at which the distribution function is the highest (and the cumulative proportion function the steepest).
- **Minimum**: the lowest value being observed.
- **Maximum**: the highest value being observed.
- **Median**: the value that is just in the "middle" of the observations. Half of the cases are above it, and half of the cases are below it.
- **1st quartile** (Q_1): the median of the lower half of the observations. One quarter of the observations are between the minimum and the 1st quartile, and one quarter of the observations are between the first quartile and the median.
- **3rd quartile** (Q_3): the median of the upper half of the observations. One quarter of the observations are between the median and the 3rd quartile, while another quarter of the observations are between the 3rd quartile and the maximum.

- **Percentiles**: the nth percentile is the value below which (or equal to it) you find $n\%$ of the observations.
  Minimum = 0th percentile, 1st quartile = 25th percentile, median = 50th percentile, 3rd quartile = 75th percentile, maximum = 100th percentile.

A more exact rule for calculating the median: 1. sort your observed values from smallest to largest (some values will be repeated if you measured the same value for several cases); 2. if n is odd (like 3,5,7,9, etc.), take the central value (that is, the $\frac{(n+1)}{2}$th element); 3. if n is even (like 2,4,6,8, etc.), take the average of the two central elements. In other words, if your observations (after having sorted them) are: x_1, x_2,... x_n, then the median is $x_{(n+1)/2}$ if n is odd, and $\frac{x_{n/2}+x_{(n/2)+1}}{2}$ if n is even.

Median – and in typically occurring distributions, also the mode – are some measures of the "center" of the distribution. Yet, the most frequently used measure of the "center" is the mean. (Refer to examples in Moore&McCabe, as well as in John Nerbonne's slides, showing that mean, median and mode can be very different in asymmetric distributions.)

- **Mean**: the average of the values x_1, x_2,... x_n:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Here, the $\sum$ symbol means the sum of whatever follows, as i grows gradually from 1 to (and including) n. That is just a fancy (but very useful) way of abbreviating the long summation.

Ways to measure the spread of the distribution, using these measurements:

- **Range**: maximum − minimum.
- **Interquartile range** (IQR): Q_3-Q_1.
- Semi-IQR: half of IQR.

The **deviation** of a data point x_i from the mean $\bar{x}$ is the difference $x_i - \bar{x}$.

A remark for those liking maths: the sum of the deviations is always 0, because:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \left(\sum_{i=1}^{n} x_i\right) - \left(\sum_{i=1}^{n} \bar{x}\right) = \left(\sum_{i=1}^{n} x_i\right) - n \cdot \frac{\left(\sum_{i=1}^{n} x_i\right)}{n} = 0$$

The average of the deviations could also be a good measure of the width of the distribution. Its meaning is: what is the average distance of the measured data points from the mean of the distribution? So let us take the average of the absolute values of the deviations: $\frac{1}{n}(|x_1 - \bar{x}| + ... + |x_n - \bar{x}|)$. Because of mathematical reasons, however, we prefer replacing the absolute values with squares, and the $\frac{1}{n}$ with $\frac{1}{n-1}$. Note that the square root of the square of any number is the absolute value of that number: $\sqrt{x^2} = |x|$. This is the motivation behind the following definition:

- **Variance**:

$$s^2 = \frac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Standard deviation**:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

This seems to be extremely complicated, but this is the measure that will prove to be most useful in all future statistical techniques.

Note that many software offers you the choice to use either n or $n-1$ in the definition of variance and standard deviation. Using n seems more obvious (that would give you the mean of the squared deviations), but you should always use $n-1$.

The mean and the variance summarize adequately a symmetric distribution (typically a normal distribution), but is not enough to summarize an asymmetric (skewed) distributions or outliers. Luckily, most often we shall encounter symmetric distributions.

4 Further important notions in Chapter 1 (sections 1.1 and 1.2) of Moore&McCabe

Stemplots. Unimodal or multimodal distributions. Time plots, time series, trend, seasonal variation (moving averages: see John Nerbonne's slide, p. 30). The $1.5 \times IQR$ rule to find outliers (and modified boxplots). Linear transformation (changing the unit of measurement).

Why are median, quartiles and the interquartile range more *resistant* measures than mean and standard deviation? Because minor changes in the data set are less likely to influence them. (Why?)