Summary of Lecture 1

Optional readings: Sections 1.2, 1.6, and 1.7 of OpenIntro Statistics

Businesses often use statistical tools to solve particular problems at hand. As you learn more tools, you will be able to address a wider range of problems and you will also be able to identify opportunities for successful applications of different statistical tools. For example, we have already discussed empirical rule in action for outlier detection. However, businesses also use statistical tools to explore available data without trying to solve any particular problem. The idea is to obtain data understanding and identify business opportunities.

Focus of the first lecture was on data exploration and descriptive statistics.

We started the lecture by discussing different types of data we may encounter. For example, we have numeric, categorical and ordered categorical data. The main thing to remember is that different statistical tools work with certain types of data, but not with others. Numeric data are quantitative and have units. It makes sense to average a numeric variable. Categorical data are qualitative and do not have units. It does not make sense (in general) to average categorical data.

We introduced histograms and time series plots as simple visualization tools that enable us to visualize one variable. For independent (we make this formal later) data ordering of observations (rows) does not matter. In other cases, observations are dependent and the ordering matters. A time series is a sequence of observations that are taken over time. These observations are dependent. For example, if a country is in recession this month it is more likely to stay in recession the following month.

Sample mean and sample median are descriptive statistics used to characterize central tendency or typical value of a variable. The sample mean is an average of observations. We use $\bar{x}$ (bar x) to denote the sample mean of a variable x (for example, "x = age"). The sample median is the middle value of data (50% of values are smaller and 50% are larger) and is robust to outliers.

Sample variance and sample standard characterize dispersion or variability of data around the sample mean. The sample variance is the average squared distance of data from the mean, given as

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2.$$

The units of the sample variance are not in the original units. The sample standard deviation is given as

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

and characterizes the dispersion in the same units as the original data.

Range: the difference between the larges and smallest values - conveys less information than standard deviation and is less commonly used.

Median absolute deviation: the median distance of observation from the median distance, that is MAD = median($|x_i - \text{median}(x_1, \ldots, x_n)|$). This is a robust alternative to sample standard deviation. (**note:** this is an aside)

The empirical rule allows us to connect the summary statistics back to data that was used to create the summary statistics. In particular, for mound shaped data, the empirical rule tells us that

- Roughly 68% of data lies in the interval $(\bar{x} - s_x, \bar{x} + s_x)$
- Roughly 95% of data lies in the interval $(\bar{x} - 2 \cdot s_x, \bar{x} + 2 \cdot s_x)$
- Roughly 99.7% of data lies in the interval $(\bar{x} - 2 \cdot s_x, \bar{x} + 3 \cdot s_x)$

Using the empirical rule, we can guess where future observations are likely to be or we can identify outliers.

We discussed scatter plots as a way to visualize two variables simultaneously. Sample covariance and sample correlation summarize how strong a **linear** relationship is between two variables. The sample covariance is given by

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

and the sample correlation is the sample covariance divided by the sample standard deviation of x and y, that is,

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

The sample correlation:

- is always between -1 and 1;
- does not have units;
- magnitude is easy to interpret;
- has the same sign as the sample covariance.

The sample covariance:

- is affected by change in units of x and y;
- has units that are not easy to interpret (in general);
- magnitude is not easy to interpret;
- sign tells us which quadrant we should expect to see data in relation to the sample mean.

The sample correlation being close to zero **does not** imply that variables are not related. It just tells us that there is no linear relationship.