CS 7280 - Spring 2016 Olga Vitek

Questionnaire

January 14, 2016

1. You measure the height of every person in the room, and would like to visualize the data in a box plot. What summaries of the data will be reported in the box plot?

   Answer:

   A boxplot (or box-and-whisker diagram) displays the 5-number summary: the minimum value, the first quartile Q_1, the median (or second quartile Q_2), the third quartile Q_3, and the maximum value.

   The scale will include the minimum and maximum of the data values. A box (rectangle) will be constructed that extends from Q_1 to Q_3, with a line drawn in the box at the median value.

   In some implementations, the scale will only span $Q2 \pm 1.5 \times (Q3 - Q1)$ ($Q3 - Q1$ is the inter-quartile range), and the individual data points will be shown as outliers.

2. You are interested in the association between two variables X and Y. A colleague advises you to calculate the correlation, and another advises you to calculate the slope of linear regression. Are these two approaches equivalent? If not, which is better?

   Answer:

   A regression views one variable as response (i.e., random), and the other as predictor (i.e., fixed). It examines the probability distribution of the response conditional on the predictor. Correlation, on the other hand, is a property of two random variables, and views both X and Y as random. If both X and Y are Normally distributed, there is a direct mathematical relationship between the slope of the linear regression and the Pearson coefficient of correlation, and from one we can deduce the other.

3. In a linear regression of X and Y, a software package reports $R^2 = 0.99$. Does this mean that X and Y have a good linear association?

   Answer:

   The coefficient of determination (R^2) is the amount of variation in Y that is explained by the regression line. More precisely:

$$R^2 = \frac{\textit{explained variation}}{\textit{total variation}}$$

   Thus having $R^2 = 0.99$, then we can conclude that the 99% of the total variation in Y can be explained by the X. However, R^2 does not quantify the linearity of fit. We can not conclude the linearity of the association.

4. What is the difference between a standard deviation and a standard error?

   Answer:

   The standard deviation quantifies the variability in the population, and the standard error quantifies the uncertainty in a parameter of a population. The former will not decrease with the increased sample size, while the latter will decrease. For example, if one were to take multiple samples from this population, there will be variations among the means. Then by estimating the variation in the means among samples which is called the standard error of the estimate of the means. There is a relation between the standard error of the sample mean ($\sigma_{\bar{x}}$), which depends on both the standard deviation (σ) and the sample size (n) as follows:

$$\textit{Standard Error of the Mean } (\sigma_{\bar{x}}) = \frac{\textit{Standard Deviation } (\sigma)}{\sqrt{n}}$$

5. Is the following statement correct: "The p-value of testing a hypothesis H_0 against the alternative H_a is the probability that H_0 is true".

   Answer:

   This is incorrect. The p-value is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis H_0 is true. This is not the same as quantifying the probability that H_0 is true.