STAT 231 — Lecture 3

Bartosz Antczak *Instructor: Michael Wallace* *September 13, 2017*

3.1 Data Summaries

3.1.1 Measures of Location

Let our data set be denoted as $\{y_1, \ldots, y_n\}$. We define three measures of location:

- **Sample Mean:** $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- **Sample Median:** denote the ordered sample as: $y_{(1)}, \ldots, y_{(n)}$, where $y_{(1)} = \min(y_1, \ldots, y_n)$ and $y_{(n)} = \max(y_1, \ldots, y_n)$. We define the median as

$$\widehat{m} = y_{(\frac{n+1}{2})} \quad (n \text{ is odd})$$

$$\widehat{m} = \frac{y_{(\frac{n}{2})} + y_{(\frac{n}{2}+1)}}{2} \quad (n \text{ is even})$$

  Note that the median can be interpreted as the "average" as well.

- **Sample Mode:** the most common value. The mode does *not exist* if this property is shared by multiple values. The mode **must be unique**.

3.1.2 Measure of Variability or Dispersion

- **Sample Variance:** $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n}({y_i}^2) - n\bar{y}\right]$
- **Sample Standard Deviation:** the square-root of the sample variance: s. This means that a larger variance means larger dispersion which means larger uncertainty in data. Approximately 68% of data is found in the interval $(\bar{y} - s, \bar{y} + s)$, and 95% in $(\bar{y} - 2s, \bar{y} + 2s)$.
- **Range:** $y_{(n)} - y_{(1)} = \max(y_1, \ldots, y_n) - \min(y_1, \ldots, y_n)$. The range is very susceptible to outliers (i.e., extreme values).
- **Quantiles:** the value of the pth quantile (or $100p$th percentile) is the value such that a fraction p of the data fall below the value. Denoted as $q(p)$ $(0 < p < 1)$, the pth quantile is determined by:
  - Let $m = (n+1)p$ (n is the sample size)
  - If m is an integer and $1 \leq m \leq n$, then $q(p) = y_{(m)}$
  - Else, determine j such that $j < m < j + 1$ and take $q(p) = \frac{1}{2}\left[y_{(j)} + y_{(j+1)}\right]$

    Example 3.1.1. *7, 11, 13, 17, 19. Find the 0.25th quantile.*

    $m = 0.25(6) = 1.5$. Let $j = 1$ such that $1 < m < 2$. Thus, $q(0.25) = \frac{1}{2}\left[y_{(1)} + y_{(2)}\right] = 9$.
- **Interquartile Range (IQR):** the range defined by $q(0.75) - q(0.25)$