STAT 231 — Lecture 4

Bartosz Antczak *Instructor: Michael Wallace* *September 15, 2017*

Notes

Assignment 1 due on September 25th. Should be straightforward, but it may take a while to set up R, so don't leave until last minute!

4.1 More Data Summaries

4.1.1 Skewness

Graphs have a wide variety of shapes. Today we will look at measuring a graph's shape. Sample skewness measures the asymmetry of data, defined by:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right]^{\frac{3}{2}}}$$

We'll learn more about why skewness is defined as this when we cover *Gaussian Response Models*. We interpret skewness as:

- Symmetric graphs have skewness $\in [-1, 1]$
- Graphs with a long left tail (i.e., skewed to the left) have a **negative** skewness
- Graphs with a long right tail (i.e., skewed to the right) have a **positive** skewness

Graphs with the same standard deviation, mean, and symmetry can still have different shapes. Another measurement we'll focus is *kurtosis*.

4.1.2 Kurtosis

Kurtosis measures whether data are concentrated in a cenreal peak or in the tail. Sample kurtosis is defined by:

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2\right]^{2}}$$

We interpret kurtosis as:

- Gaussian graphs have sample kurtosis $\in [2, 4]$
- Data with large tails (i.e., more prone to extreme values) have sample kurtosis $>> 3$

- Data with short tails (i.e., most points fall near the mean than in a normal distribution) have a sample kurtosis $<< 3$
- Data that looks uniform have sample kurtosis close to 1.2

Kurtosis means "curved". Higher kurtosis means more curvature; lower kurtosis means less curvature. In order to create a reasonable assumption that a particular data set is Gaussian model, the following characteristics must be met:

- Mean $\approx$ median
- Sample skewness $\in [-1, 1]$
- Sample kurtosis $\in [2, 4]$
- 95% of data found within range $\in [\overline{y} - 2s, \overline{y} + 2s]$

We've learned a bunch of data summary measurements. A useful data summary is a combination five measurements we've learned, called the **five number summary**:

- Minimum $y_{(1)}$
- $q(0.25)$, $q(0.5)$, $q(0.75)$
- Maximum $y_{(n)}$