

STAT 100 Lecture 3: Measures of Variation

Nate Strawn

<http://www.math.umd.edu/~nstrawn/>

- 1 We introduced sigma-notation:

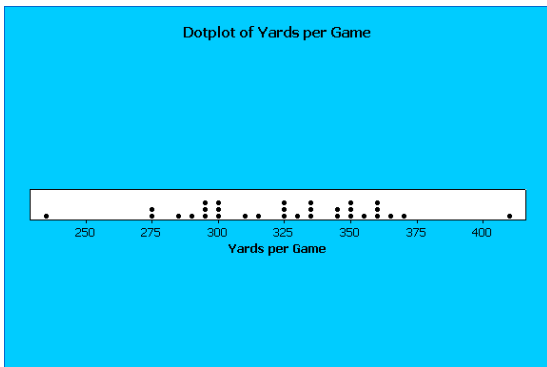
$$\sum_{i=1}^N x_i = x_1 + x_2 + x_3 + \cdots + x_N$$

- 2 We defined the Mean, Median, Percentiles, and Quantiles

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

Quartiles Again!

For the following data, we have $Q_1 = 295.48$, $Q_2 = 329.40$, and $Q_3 = 350.70$.



Today's Agenda

- 1 Deviation from the Mean
- 2 Measures of Variation
 - The Variance and Standard Deviation
 - The Range, Interquartile Range, and Boxplots
- 3 Bell Curves

Deviation from the Mean

Given a set of data $\{x_i\}_{i=1}^N$, we set $d_i = x_i - \bar{x}$ to be the **deviation** of the i^{th} data point.

Proposition

$$\sum_{i=1}^N d_i = 0$$

Proof using sigma-notation.

Using sigma-notation, the proof looks like this:

$$\sum_{i=1}^N d_i = \sum_{i=1}^N (x_i - \bar{x}) = \sum_{i=1}^N x_i - \sum_{i=1}^N \bar{x} = \frac{N}{N} \sum_{i=1}^N x_i - N\bar{x} = N\bar{x} - N\bar{x} = 0$$



Expanded Proof

Proof using expanded notation.

In expanded notation, we have:

$$\begin{aligned}\sum_{i=1}^N d_i &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_N - \bar{x}) \\ &= (x_1 + x_2 + \cdots + x_N) - (\bar{x} + \bar{x} + \cdots + \bar{x}) = \\ &= N \left(\frac{x_1 + x_2 + \cdots + x_N}{N} \right) - N\bar{x} \\ &= N\bar{x} - N\bar{x} \\ &= 0.\end{aligned}$$



Degrees of Freedom

Since $\sum_{i=1}^N d_i = 0$, we have that $d_k = -\sum_{i \neq k} d_i$ for any k between 1 and N .

This means that any deviation can be reconstructed if we know the other $N - 1$ deviations. To indicate this, we say that $\{d_i\}_{i=1}^N$ has $N - 1$ **degrees of freedom** (abbreviated dof).

Definition

The **Variance** of a set of N measurements $\{x_i\}_{i=1}^N$ is

$$s^2 = \frac{\text{sum of squared deviations}}{\text{degrees of freedom}} = \frac{\sum_{i=1}^N d_i^2}{N-1} = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}.$$

The **Standard Deviation** is the square root of the Variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}}$$

A Quicker Formula for the Variance

The defining formula for the Variance requires computation of the deviations. It can be shown that

$$s^2 = \frac{1}{N-1} \left(\sum_{i=1}^N x_i^2 - N\bar{x}^2 \right),$$

so we can compute the Variance with less computations.

The Range and Interquartile Range

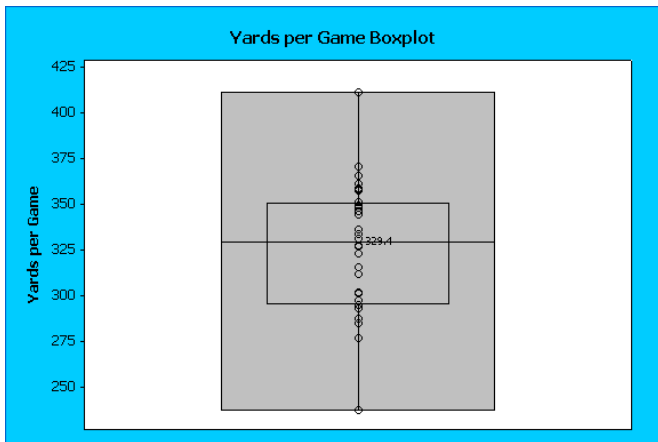
Definition

The **Range** of a data set $\{x_i\}_{i=1}^N$ is the difference between the largest and smallest data value.

Definition

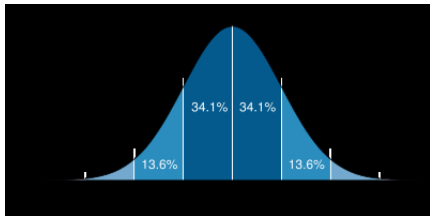
The **Interquartile Range** of a data set is the difference between the third and second quartile.

Boxplots



Bell Curves

Bell Curves (a.k.a. Gaussians, Normal Distributions) are all over statistics. For data that is Normally Distributed, we have that 68 percent of the data lies within one standard deviations of the mean, 95 percent lies within 2 standard deviations of the mean, and 99.7 percent of the data lie within three standard deviations of the mean.



For Next Time

- Read Section 3.4 and 3.6 from Johnson and Bhattacharyya
- Online homework 2.5: 2.57, 2.67, 2.75, 2.77, 2.85
 - Section 0101:
<http://edugen.wiley.com/edugen/class/cls73726/>
 - Section 0201:
<http://edugen.wiley.com/edugen/class/cls73727/>
- Group presentations: 2.63, 2.68, 2.71, 2.76