1.2 Summary Statistics

Sample Mean

The *sample mean* of a sample $X_1, \cdots, X_n$ is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, a measure of the center of the data.

e.g. For the sample 9, 10, 11, $\bar{X} =$

e.g. For the sample 1, 10, 19, $\bar{X} =$

Standard Deviation

The *deviation* of the i^{th} observation from the mean is (observation) - (sample mean) $= X_i - \bar{X}$.

Adding deviations over entire sample gives $\cdots$

The *sample variance* of n observations is their average squared deviation:

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right) \text{ (easier to compute by hand)} \end{aligned}$$

(Note: Divide not by ______, but by ____________, the *degrees of freedom* in the sum. If we knew the population mean, μ, we'd sum squared deviations from μ and divide by n. But, with $\bar{X}$ calculated from the sample, the deviations sum to 0, so any $n-1$ of them determine the last one. This technicality ________________________ for large n.)

(Note: We could call $\sum_{i=1}^{n}(X_i - \bar{X})^2$ a __________________, and we could then refer to s^2 as a __________________.)

Variance is measured in ____________.

The *sample standard deviation* of n observations is

$$\begin{aligned} s &= \sqrt{\textit{sample variance}} \\ &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ &= \sqrt{\frac{1}{n-1}\left(\sum X_i^2 - n\bar{X}^2\right)} \end{aligned}$$

Standard deviation has the same units as the data.

e.g. Find sample standard deviation for 9, 10, 11:

e.g. Find sample standard deviation for 1, 10, 19:

Mean, variance, and standard deviation of a sample transformed by multiplication by a constant and/or addition of a constant:

If $Y_i = a + bX_i$, where a and b are constants, then
$\bar{Y} =$

$s_Y^2 =$

$s_Y =$

e.g. $Y_i = 32 + \frac{9}{5}X_i$

Outliers

An *outlier* is a data point that is much larger or smaller than the others. Check outliers. (Correct or delete them only if certain that they're due to _____________.)

Sample Median

The *sample median*, M, is the midpoint of a sorted sample. To find it,

- sort sample
- $\begin{cases} n \text{ odd} \implies M \text{ is center data point at position } \frac{n+1}{2} \\ n \text{ even} \implies M \text{ is the average of the two central points at positions } \frac{n}{2} \text{ and } \frac{n}{2}+1 \end{cases}$

e.g. Find median of 3, 1, 4, 2, 0

e.g. Find median of 3, 1, 4, 2, 0, 5

e.g. The median resists outliers better than the mean:

Quartiles

The *first quartile*, Q_1, of a sample is the data point at (sorted) position $\frac{1}{4}(n+1)$, or the average of points on either side if this isn't an integer.

The *third quartile*, Q_3, is the point at $\frac{3}{4}(n+1)$, or the average of points on either side if this isn't an integer.

(The second quartile is ______________).

Percentiles

The *pth percentile* is the point at position $\frac{p}{100}(n+1)$, or the average of points on either side if this isn't an integer. About ______________ of the sorted sample data are less than the pth percentile.

e.g. 75th percentile is ____________________________.

Computing

Check the "Computing" section of the syllabus:

- Help with calculators is under the "instructions" link, which leads to a "mean_standardDeviation" folder for Chapter 1 and to a "correlation_regression" folder for Chapter 2.
- Help with R and RStudio software is under the "R Guide" link. So far, it describes the example code for chapter 1 in "1.R".