Section 3.2 - Measures of Variation

While we have studied several useful measures of center, these measures are not sufficient to describe the distribution of a data set. Another important quantity is the variation.

The **range** of a set of data values is the difference between the maximum and minimum values:

Range = Max value - Min value

The most commonly used measure of variation is the standard deviation, which measures the extent of the deviations from the mean.

The **sample standard deviation**, denoted by *s*, is obtained from either one of the following formulas:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

or

$$s = \sqrt{\frac{n\ \sum(x^2) - (\sum x)^2}{n\ (n - 1)}}$$

- The standard deviation is a measure of variation of all values from the mean.
- The standard deviation is never negative, and it is zero only when all data values are all the same.
- Small *s* indicates little variation. Large *s* indicates great variation.
- The standard deviation is very sensitive to extreme values.

When comparing data sets, you should compare sample standard deviations only when sample means are approximately the same.

The **population standard deviation**, denoted by σ, should be used when measuring variation in the population and is obtained from the following formula:

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

where μ is the population mean.

Warning: Calculators and software normally are capable of computing both standard deviations. Be sure you know which one you're computing!

The **sample or population variance** is the square of the corresponding standard deviation.

Interpretations and rules of thumb...

- For many data sets, the vast majority of sample values lie within two standard deviations of the mean.

- Minimum "usual" value ≈ mean - 2 × (standard deviation)

- Maximum "usual" value ≈ mean + 2 × (standard deviation)

- $s \approx \frac{\text{Range}}{4}$

For data sets that are approximately normally distributed...

- About 68% of all values lie within 1 standard deviation of the mean, i.e. between $\bar{x} - s$ and $\bar{x} + s$.

- About 95% of all values lie within 2 standard deviations of the mean, i.e. between $\bar{x} - 2s$ and $\bar{x} + 2s$.

- About 99.7% of all values lie within 3 standard deviations of the mean, i.e. between $\bar{x} - 3s$ and $\bar{x} + 3s$.

Chebyshev's Theorem

For **any** data set, the fraction of values lying within *K* standard deviations of the mean is always at least $1-1/K^2$.

The sample standard deviation is an unbiased estimator of the population standard deviation---sample standard deviations obtained from different samples within a single population tend to target the population standard deviation.

Using the mean absolute deviation (page 103), or dividing by the sample size instead of *(n-1),* results in a biased estimator.

Another way to think about the denominator of the formula for *s* is that (*n-1*) is the number of degrees of freedom.

When comparing variation in significantly different data sets, such as those involving different units of measurement or with very different means, the **coefficient of variation** should be used. The **CV** is a standard deviation relative to the mean.

$$CV = \frac{s}{\bar{x}} * 100\% \text{ (Sample CV)}$$

$$CV = \frac{\sigma}{\mu} * 100\% \text{ (Population CV)}$$