3/29/2022

Continuing with Chapter 2
Measures of Location (in Excel)
Measures of Center
Skewness
Measures of Spread

Last time we discussed: the 5-number summary: minimum, the first quartile (Q1), the median, the third quartile (Q3), and maximum.

Don't do calculations by hand!

See Excel for methods for calculating these values.

Measures of Center

Average = Mean (Excel calls it an average, Statistics calls it the mean)
The mean of a population (or theoretical distribution) is given the symbol μ (Greek letter mu).
The mean of a sample is given the English abbreviation/symbol $\bar{x}$.
Sometimes you will see $\hat{\mu}$, the hat means it's an estimate.

Normally in statistics, theoretical values (parameters) are given Greek letters, and the statistics we can calculate are given English equivalents.

Median is another method of estimating the "center" of the data. 50% at or below that value and 50% at or above the value. The symbol is $\tilde{x}$, but this is uncommon. Often you just get m or med, or M.

Sometimes measures of center are thought of as "typical" values. The mode is another concept of "typical", it's the most common value in the data set (the value that appears the model number of times). This also doesn't have a consistent symbol (sometimes M).

"Modal class" is the class in a histogram (or bar graph) that has the most observations.

Sometimes there is more than one value that appears the same number of times as the most common value. Excel can calculate multiple modes. Normally on stats exams, we don't count them unless there are two or fewer modes. If there are three or more, then normally we say "no mode".

Data with two modes is called "bimodal".

Skewness

The name of the measure of skewness is called kurtosis.
Look at skewness from a visual standpoint using histograms and boxplots. Our discussion is going to a little imprecise, but this is one of the ways real statisticians think about data.

Shapes of distributions (distribution is how the data is spread out)

Uniform distribution: most or all of the bars in histogram are basically the same height. In a theoretical case, the bars are all exactly the same height. When we use real data, this almost never happens. (Everything is approximate).

Normally distributed data looks like a “bell curve”. It’s tallest in the middle and falls off equally on each side.

Symmetric – looks like a distribution where if you fold it in half along the center line, the two sides look the same (roughly).

Skewed distributions: have a long tail on one side (only)

Right-skewed distribution: has a long tail on the right side (extreme large values)

Left-skewed distribution: has a long tail on the left side (extreme small values)

The skewness is determined by the tail, not by where most of the data is.

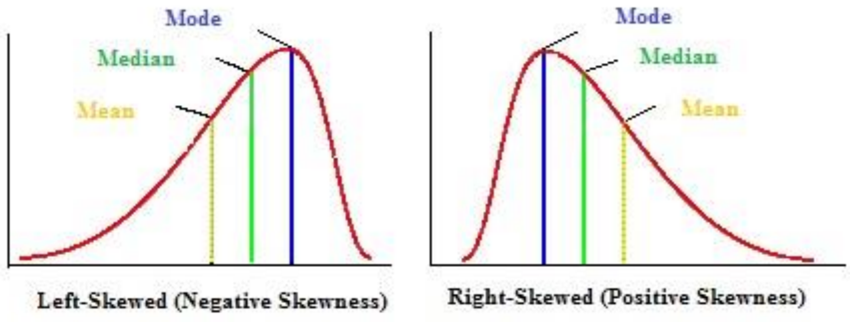
Skewed distributions will impact the location of mean, median and mode relative to each other. (setting aside the bimodal case)

In symmetric distributions (normal or uniform), the mean, median and mode are all about the same value.

In a skewed distribution, the mean is most affected by the skewed, and the model will be at the peak.

In a right-skewed distribution: mode, median, mean

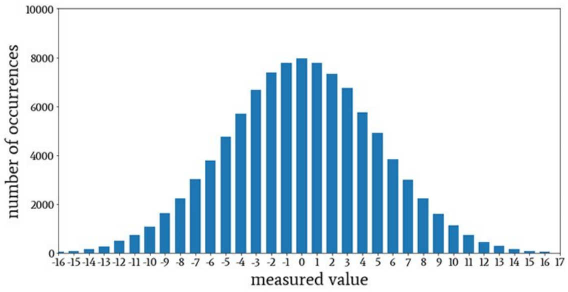
In a left-skewed distribution: mean, median, mode


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

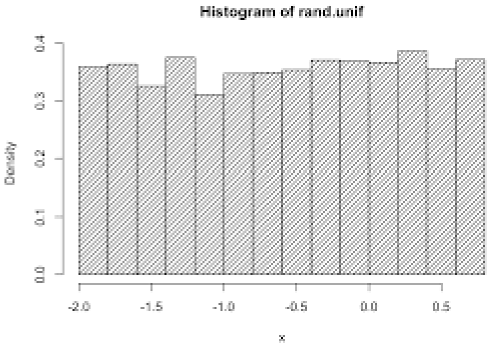
Things that often right-skewed:
Time (you can get longer, as long as you like, but there is a hard 0)
Salaries are often right-skewed

Things that are often left-skewed:
The dates of coins/money

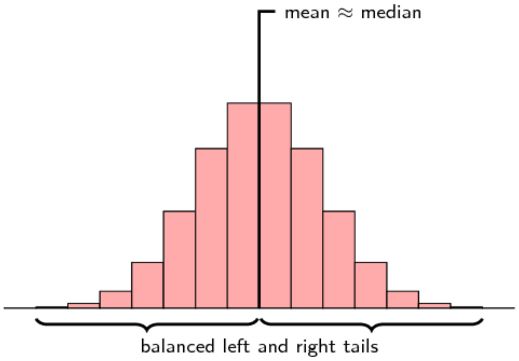
Things that are normally distributed:
Heights of people (of a single gender)
Test scores (SATs, ACTs, IQ tests, exam scores)

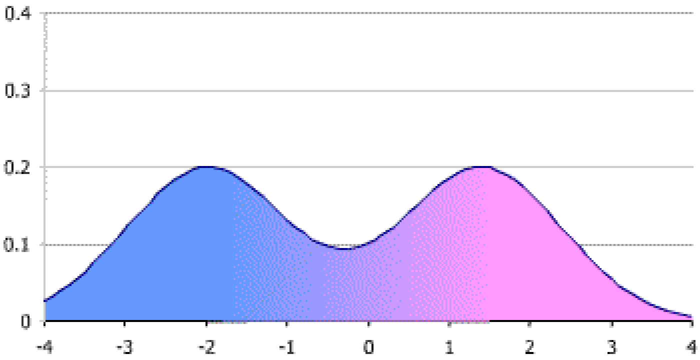


Uniformly distributed:
Fair die roll.



Symmetric





(the latter example is bimodal)

Measures of Spread
Standard Deviation, Variance, Range, Interquartile Range

Range: is the difference between the maximum value and the minimum value

Interquartile Range: is the difference between the third quartile and the first quartile $IQR = Q3 - Q1$ (the middle 50% of the data)

Standard deviation: approximately the average distance of the data to the mean.
(a similar calculation with absolute values: mean absolute deviation MAD)

Variance is the square of the standard deviation

(The variance and the standard deviation have two versions: one is for calculating the value for a population, and one is for calculating the value for a sample. We are always going to use the sample formula. Don't ever use the population version of the formula unless the problem specifically tells you to.)

Population versions:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

This says find the mean, subtract from every value. Square to make positive. Add them up. Take the average.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Again, do not use these unless the problem says explicitly "population"!!!!!!

The sample versions (the ones you want):

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

The $n - 1$ is because the population version slightly underestimates the value from a sample.

The normal distribution can be described entirely by the mean and the standard deviation.

Fences/Extreme values in a box plot.

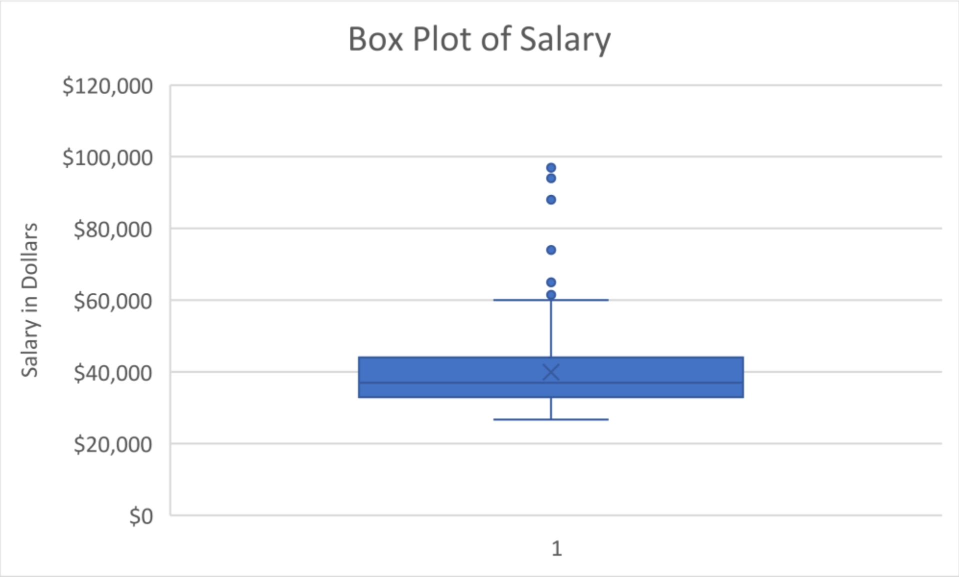
"These fences" mark the boundary in a box plot between the "normal" values and the extreme values. The distance to the "fence" is measured from each quartile. The distance is calculated as $1.5 \times IQR$. More extreme values use $3.0 \times IQR$.

Minimum	26700	
1st Quartile	33000	
Median	37000	
3rd Quartile	44000	
Maximum	$97,000	
IQR		11000

Lower Fence: $Q1 - 1.5 \times IQR$
Upper Fence: $Q3 + 1.5 \times IQR$

$$1.5 \times IQR = 1.5 \times 11{,}000 = 16{,}500$$

Lower Fence: $33{,}000 - 16{,}500 = 16{,}500$, there are no values less than this so no dots on the low end
Upper Fence: $44{,}000 + 16{,}500 = 60{,}500$, there values higher than this, so the whisker stops here, and plots larger values as dots.



Data Analysis Tool Pack