Week 2 Cheat Sheet

Statistics and Data Analysis with Excel, Part 1

Charlie Nuttelman

Here, I provide the mathematical equations and some of the important Excel functions required to perform various calculations in Week 2 of the course. The headings represent the screencasts in which you will find those calculations and concepts. Not all screencasts are referenced below – just the ones that have complex mathematical formulas or Excel formulas that are tricky to use.

Difference Between Population and Sample

The population size is denoted by N and the sample size is denoted by n. Population mean (μ) can be estimated using the sample mean ($\bar{x}$). Population variance (σ^2) can be estimated using the sample variance (s^2). Standard deviation is the square root of variance.

Formulas for these parameters are:

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

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

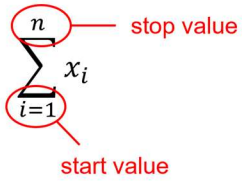
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

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

Population and sample mean (average) can be calculated using the **AVERAGE** function in Excel. Population and sample variances can be calculated using the **VAR.P** and **VAR.S** formulas, respectively, and the population and sample standard deviations can be calculated using the **STDEV.P** and **STDEV.S** formulas, respectively. The **COUNT** function is useful in counting the number of observations.

The Summation Symbol

The summation symbol, Σ, or Greek letter sigma, is used as an indicator to sum over the expression that follows the symbol. The integer below the symbol (typically written as some index variable equal to 1, or other number) is the start value for which iteration and summation will occur. Above the summation symbol is the stop value, or the number at which iteration and summation will occur:



For example, if x = {1, 2, 3, 4, 5}, then:

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + \cdots + x_5$$

$$= 1 + 2 + 3 + 4 + 5 = 15$$

The summation symbol is used in the definition and calculation of average and variance (see below).

Descriptive Statistics

Another common measure of spread in a set of data is the range of the data, which is just the maximum value in the data set minus the minimum value. We can calculate the maximum value of a set of data in Excel using the **MAX** function and the minimum value using the **MIN** function; the range is simply the difference between those two values.

Skewness and kurtosis are sometimes used to describe the asymmetry of a set of data when compared to the normal distribution. The **SKEW** and **KURT** functions in Excel can determine these parameters. For more information on how to interpret these values, please visit support.microsoft.com.

Quartiles and Percentiles

For either quartiles or percentiles, we first determine a rank, k, by using one of the formulas below. We can either include the median or exclude the median (it is more common to exclude the median). The parameter p is the desired percentile; for quartiles, the first quartile is the same as the 25th percentile ($p = 0.25$) and the third quartile is the 75th percentile ($p = 0.75$).

Including the median: $k = p \cdot (n - 1) + 1$

Excluding the median: $k = (n + 1) \cdot p$

Once we have the rank, we can linearly interpolate between ordered values in our data. For example, if our (ordered) data is: 5, 9, 12, 14, 17, 18, 21, 22, 25 ($n = 9$) and we wish to find the first quartile including the median, we would calculate the rank as $k = 0.25 \cdot (9 - 1) + 1 = 3$. Therefore, the first quartile in this case is 12. Similarly, the third quartile would be calculated to be 21 ($k = 7$).

For the same data set, if we wished to find the first quartile excluding the median, we would calculate the rank as $k = (9 + 1) \cdot 0.25 = 2.5$. Therefore, we linearly interpolate 50% of the way between the 2nd and 3rd values of the ordered data, and the first quartile is 9 + 0.5 x (12 – 9) = 10.5. Similarly, the third quartile would be calculated to be 21.5 ($k = 7.5$).

Percentiles are calculated exactly the same but p can be any continuous value between 0 and 1. For example, for the above data set if we wanted to calculate the median-excluded 13th percentile, we calculate the rank: $k = (9 + 1) \cdot 0.13 = 1.3$. The 13th percentile is then 30% of the way between the 1st and 2nd of the ordered values = 5 + 0.3 x (9 – 5) = 6.2.

In Excel, we can use the **QUARTILE(data,q)**, **QUARTILE.INC(data,q)**, and **QUARTILE.EXC(data.q)** to calculate quartiles, where **q** = 1 for the 1st quartile and **q** = 3 for the 3rd quartile. We can use the **PERCENTILE(data,p)**, **PERCENTILE.INC(data,p)**, and **PERCENTILE.EXC(data,p)** functions in Excel to calculate the 100pth percentile (for example, for the 67th percentile p would be 0.67).

Histograms

The best way to visualize the distribution of univariate data is the use of a histogram. In a histogram, the data are sorted into "bins" of constant width and frequencies of each bin are plotted as a column chart. We typically estimate a lower bound and an upper bound for the number of bins:

$$n_{bins,lower} = INT\big(LOG_2(n)\big) - 1$$

$$n_{bins,upper} = \sqrt{n}$$ (typically rounded to the nearest integer)

Here, n is the number of observations or experimental measurements. I like to choose the actual number of bins to be somewhere between the lower and upper estimates for number of bins.

Excel's histogram tool (Data → Data Analysis → Histogram) is great for parsing the data into the bins, but the user must provide the bin boundaries.