Week 5 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 5 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.

NOTE: For Week 5, I also have provided you with a cheat-sheet specifically for Excel formulas ("Excel Functions for Continuous Distributions"). So, be sure to check that out, too!

Continuous Variables and Distributions

A continuous variable is one that can take on a range of values, not just constrained to discrete values we discussed in the previous week. For example, x can range from $-\infty$ to $+\infty$ ($-\infty < x < \infty$), can be constrained to only positive values ($x > 0$), or can be constrained to a finite range (e.g., $1 < x < 3$). For $f(x)$ to be a probability density function, all values of the function must be non-negative [$f(x) \geq 0$] and the area underneath the function must sum to unity [$\int f(x)dx = 1$]. For continuous density functions, the probability that the random variable takes on any exact value is equal to zero (e.g., $P[X = 4] = 0$).

Like for discrete density functions, we can calculate probabilities by taking differences in cumulative probability density functions:

$$P[a \leq x \leq b] = \int_a^b f(x) \cdot dx = F(b) - F(a)$$

We typically don't need to perform the integration (the term between the two equal signs) since we can either look up cumulative density functions or obtain them from Excel (or other computing tool).

The Uniform Distribution

The uniform distribution is defined by $f(x) = \frac{1}{b-a}$, where x is constrained to $a \leq x \leq b$. The function is a flat line with slope zero. The RAND() function in Excel will output a uniformly distributed variable between 0 and 1 (and it can be modified to output a uniformly distributed variable between any two integers).

The Normal Distribution

The most common continuous distribution that models real world phenomenon very well is the normal distribution, also known as the Gaussian distribution or the "bell curve". As always, we can use the

equation above in "Continuous Variables and Distributions" section to determine probabilities, where we can obtain the cumulative density function from tables or from Excel using the NORM.DIST function. If we have a cumulative probability for the normal distribution, we can always determine the corresponding value of x using the NORM.INV function in Excel.

Some useful properties of the normal distribution are that about 68% of the distribution lies within $\pm\sigma$ (one standard deviation) of the mean, about 95% of the distribution lies within $\pm 2\sigma$ (two standard deviations) of the mean, and 99.7% of the distribution lies within $\pm 3\sigma$ (three standard deviations) of the mean.

Standardizing and Z-Values

The standard normal distribution is a normally distributed variable with mean zero and standard deviation equal to 1. Standardizing was useful before computing tools (Excel being one) became mainstream, and they enabled one to convert a normally distributed variable to the standard normal distribution, and use the z-tables for probability calculations. Nevertheless, it's still important to be able to convert between normally distributed variables and the standard normal distribution.

The standardization equation is given by: $Z = \frac{X-\mu}{\sigma}$, where X is the normally distributed variable of interest, μ is the population mean, and σ is the population standard deviation from which the samples are derived.

Cumulative standard normal distribution tables can be helpful in calculating probabilities associated with standard normal distributions. The NORM.S.DIST and NORM.S.INV functions are useful Excel functions for dealing with the standard normal distribution.

Inverse Normal Distribution Calculations

Oftentimes, we need to know what the z-value (for standard normal distribution) or x-value (for normally distributed variables) is that corresponds to a given probability. We can use inverse normal calculations by looking up data in cumulative standard normal distribution tables. Alternatively, we can use the NORM.INV or NORM.S.INV functions in Excel to do the same.

Exponential Distribution

Another important continuous distribution is the exponential distribution. The exponential distribution is commonly associated with Poisson processes. It provides the probability that there will be a certain interval (time, length, area, volume) between two events of a Poisson process. It is common to use the cumulative exponential distribution to calculate probabilities. The cumulative exponential distribution is given by:

$$F(x) = 1 - e^{-\lambda}$$

Here, λ is the long-term average rate (events per interval). The EXPON.DIST function in Excel is useful for probability calculations related to the exponential distribution.

Other Continuous Distributions

Other continuous distributions that you might encounter in statistics and applied data analysis include: the gamma distribution (GAMMA.DIST in Excel, constrained to $x > 0$); the Weibull distribution (WEIBULL.DIST in Excel, also constrained to $x > 0$); the triangular distribution (constrained to a finite range in x); the beta distribution (BETA.DIST in Excel, constrained to $0 \leq x \leq 1$); and the beta-PERT distribution (can use the BETA.DIST function in Excel, constrained to a finite range in x).

Probability Plots

A probability plot is used to determine whether a set of sample data likely came from a normally distributed population. Typically, in order to show that the data were derived from a normally distributed population, the P-value of the AD statistic should be greater than 0.05. If the P-value of the AD statistic is less than 0.05, then it can be concluded that the data do not come from a normally distributed population.

Many sophisticated software tools will provide probability plots and P-values of the AD statistic. Excel, unfortunately, won't do this, but in the course, I provide a file "Probability plot.xlsx" that will do this for you. Please use this file to explore normality of the data – this will be much more important in Part 2 of the course.