Week 4 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 4 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 4, I also have provided you with a cheat-sheet specifically for Excel formulas ("Excel Functions for Discrete Distributions"). So, be sure to check that out, too!

Probability Distributions

A general probability distribution function is given by $f(x)$. There are two main properties of discrete probability distribution functions. First, each element of the probability distribution must be nonnegative [$f(x_i) \geq 0$]. Second, the area under the distribution is equal to 1 [$\sum f(x_i) = 1$].

We can determine probabilities by using the distribution function. For example, $P[X = 3] = f(3)$. The cumulative distribution function, $F(x)$, is defined as the sum of all $f(x_i)$ up to that value of x:

$$F(x) = P[X \leq x] = \sum_{x_i \leq x} f(x_i)$$

The Binomial Distribution

The binomial distribution is used when we have a fixed number of trials (n) and we wish to determine the probability of obtaining exactly x number of successes given a probability of success, p. The binomial distribution is given by the following equation:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

The term $\binom{n}{x}$ is exactly the same as $_nC_r$:

$$\binom{n}{x} = \quad _nC_r = \frac{n!}{(n-r)!\,r!}$$

We can use Excel's **BINOM.DIST** function to easily calculate the above formula for the binomial distribution.

The Geometric Distribution

When we are interested in the probability of getting exactly 1 success in x number of trials (the final success coming on the last trial), we can use the geometric distribution:

$$f(x) = (1-p)^{x-1} \cdot p$$

Here, p is the probability of success and x is the number of trials required to get that first success. Again, the first success is always on the last trial for the geometric distribution.

There is no function in Excel that performs this calculation, but we can use the **NEGBINOM.DIST** function in Excel (the geometric distribution is a special case of the negative binomial distribution; namely, the number of successes is identically 1).

The Negative Binomial Distribution

When we are interested in the probability of getting exactly the r^{th} success in x number of trials (the final success coming on the last trial), we can use the negative binomial distribution:

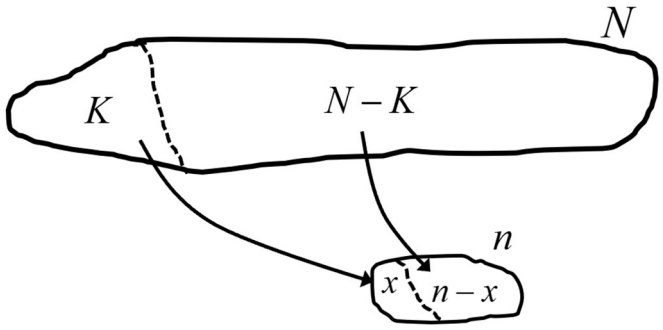
$$f(x) = \binom{x-1}{r-1} \cdot (1-p)^{x-r} \cdot p^r$$

The **NEGBINOM.DIST** function can be used in Excel to calculate non-cumulative and cumulative probabilities associated with the negative binomial distribution.

The Hypergeometric Distribution

When we have sampling without replacement, we must use the hypergeometric distribution. All the above distributions involve Bernoulli trials, which assume that the probability of a success does not change from trial to trial. However, as items are removed from a small sample and not replaced, the underlying probability of success will change.

Consider a population of size N with K successes (note that these could also be defined as defective items – it just depends on your problem of interest and how you define things). A sample of size n is withdrawn from the population, without replacing those items. There is a certain probability that the sample will have x successes, and that's what the hypergeometric distribution predicts.



The hypergeometric distribution function is given by:

$$f(x) = \frac{\binom{K}{x} \cdot \binom{N-K}{n-x}}{\binom{N}{n}}$$

Remember, $\binom{K}{x}$ is the same as ${}_nC_r$ (COMBIN(n,r) in Excel).

Excel has a built-in HYPGEOM.DIST function that can be used to calculate the cumulative and non-cumulative hypergeometric distribution and probabilities related to it.

The Poisson Distribution

When discrete events occur on a continuous interval, the process can be modeled using a Poisson process. The Poisson distribution predicts the probability that there will be x successes in an interval with mean λT, where λ is the long-term average (events/interval) and T is the specific interval of interest.

The Poisson distribution function is given by:

$$f(x) = \frac{e^{-(\lambda T)}(\lambda T)^x}{x!}$$

We can use Excel's POISSON.DIST function to calculate probabilities associated with the Poisson distribution.