# Week 2 Cheat Sheet

*Statistics and Data Analysis with Excel, Part 2*

*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.

My equations and analysis follow that of Montgomery and Runger in: *Applied Statistics and Probability for Engineers, 7th edition*, Douglas C. Montgomery and George C. Runger, Wiley (2018). This is an excellent text for applied probability and statistics and highly recommended if you need a supplementary text for the course.

## *Sampling Distribution vs. Population Distribution*

The important thing to note with regards to the difference between a sampling distribution and a population distribution is that a population distribution applies when the sampling size is equal to 1 ($n = 1$). The variance of a population distribution is simply $\sigma^2$. A sampling distribution represents the distribution when samples of size $n > 1$ are taken – it represents the distribution of $\bar{x}$ rather than the distribution of $x$. The effective variance of a sampling distribution is $\sigma^2/n$, which makes the sampling distribution narrower than the population distribution (the sampling distribution becomes narrower as $n$ increases). Both the sampling distribution and population distribution are centered around the true population mean, $\mu$.

## *Central Limit Theorem*

The Central Limit Theorem basically just tells us that, even if the population distribution is not normally distributed, the sampling distribution will often be normally distributed. As the sample size increases, the sampling distribution becomes more normally distributed.

## *Variance Known or Unknown?*

When calculating confidence intervals and performing hypothesis tests on the mean, we need to determine if we are working with the “variance known” case or the “variance unknown" case. If you know the variance – for example, from a long history of the process – then you obviously are working with the “variance known” case. For the “variance known” case, we base confidence intervals and test statistics on the standard normal distribution (i.e., we use $z$-values).

If we have small sample sizes ($n \leq 40$) and we don’t know the underlying population variance, we assume that we are working with the “variance unknown” case. For the “variance unknown” case, we base confidence intervals and test statistics on the T distribution (i.e., we use $t$-values). Even if we don't know the variance but have very large sample sizes ($n > 40$), we can assume the "variance known" case since sample variance approaches population variance for large sample sizes.

### *Confidence Interval on the Mean, Variance Known*

When variance is known, we can calculate a two-sided $(1 - \alpha)$% confidence interval on the population mean. What this means is that, given a sample average ($\bar{x}$), we can be $(1 - \alpha)$% sure that the true population mean lies within the following interval:

$$P\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

It is important to remember that $z_{\alpha/2}$ is the z-value that has $(\alpha/2)$% of the distribution to the right of it. $z_{\alpha/2}$ can be looked up in the "Percentage Points of the Standard Normal Distribution" table or, much easier, can be calculated using the NORM.S.INV function (see "Excel Functions for the Standard Normal and T Distribution").

One-sided confidence bounds can also be used, but they are a lot less common.

### *The T Distribution*

When variance is unknown, we use the T distribution instead of the standard normal distribution. In contrast to the standard normal distribution, the T distribution is different for different degrees of freedom. The number of degrees of freedom for the T distribution is equal to $n - 1$, where $n$ is the sample size. When working with the T distribution, $t_{\alpha,n-1}$ refers to the t-value that has $\alpha$% of the distribution to the right of it based on the T distribution with $n - 1$ degrees of freedom.

$t_{\alpha,n-1}$ can be looked up in the "Percentage Points of the T Distribution" table or, much easier, can be calculated using the T.INV function (see "Excel Functions for the Standard Normal and T Distribution").

### *Confidence Interval on the Mean, Variance Unknown*

When variance is unknown, we can calculate a two-sided $(1 - \alpha)$% confidence interval on the population mean and we base this analysis on the T distribution. What this means is that, given a sample average ($\bar{x}$), we can be $(1 - \alpha)$% sure that the true population mean lies within the following interval:

$$P\left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] = 1 - \alpha$$

It is important to remember that $t_{\alpha/2,n-1}$ is the t-value that has $(\alpha/2)$% of the distribution to the right of it. $t_{\alpha/2,n-1}$ can be looked up in the "Percentage Points of the T Distribution" table or, much easier, can be calculated using the T.INV function (see "Excel Functions for the Standard Normal and T Distribution").

One-sided confidence bounds can also be used, but they are a lot less common.

### *Prediction Interval for a Future Observation*

Oftentimes we wish to estimate a prediction interval in which a single future observation ($X_{n+1}$) will lie. This is a different problem entirely from estimating a confidence interval on the mean as it is a single observation and not the mean of a population.

The following formula can be used to calculate a $(1-\alpha)\%$ prediction interval on a single future observation ($X_{n+1}$):

$$\bar{x} - t_{\frac{\alpha}{2},n-1} \cdot s\sqrt{1+\frac{1}{n}} \leq X_{n+1} \leq \bar{x} + t_{\alpha/2,n-1} \cdot s\sqrt{1+\frac{1}{n}}$$

### *The Chi-Squared Distribution and Confidence Intervals on the Variance*

When creating confidence intervals for and performing hypothesis tests on the population variance, we use the chi-squared distribution. The chi-squared distribution is asymmetric and positive for all values of the independent variable. Like the T distribution, the chi-squared distribution depends upon degrees of freedom, which is typically $n-1$, where $n$ is the sample size.

Using sample statistics ($s$ and $n$), we can calculate a $(1-\alpha)\%$ confidence interval on the population variance. What this means is that, based on the sample statistics, we can be $(1-\alpha)\%$ confident that the true population variance lies in that interval.

The $(1-\alpha)\%$ two-sided confidence interval on the variance is given by:

$$P\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right] = 1-\alpha$$

One-sided confidence bounds can also be used, but they are a lot less common.

$\chi^2_{\frac{\alpha}{2},n-1}$ refers to the chi-squared variable underneath the chi-squared distribution with $n-1$ degrees of freedom that has $(\alpha/2)\%$ of the distribution to the right of it.

$\chi^2_{\frac{\alpha}{2},n-1}$ and $\chi^2_{1-\frac{\alpha}{2},n-1}$ can be obtained from the “Percentage Points of the Chi-Squared Distribution” table or, much easier, can be calculated in Excel using the CHISQ.INV and CHISQ.INV.RT functions (see “Excel Functions for the Chi-Squared Distribution”).