Week 4 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 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.

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.

Comparison of Means, Variance Known (Parts 1 and 2)

In contrast to Week 3, where we compared a hypothesized mean, variance, or binomial proportion to a fixed value (for example, $H_1: \mu = 10$ or $H_1: \sigma^2 = 3.2$), in this week we perform hypothesis tests for comparison of means, variances, or binomial proportions of two populations.

We can set up a one-tailed (upper or lower) or two-tailed hypothesis test on the comparison of means of two populations when variance is known for both populations, for example a lower-tailed test:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 < \mu_2$$

As with previous hypothesis testing, we can test the hypothesis using confidence intervals, test statistics, or P-values (which use test statistics). To perform the test using the test statistic approach, we can calculate a test statistic for a comparison of means with known variance as:

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

This test statistic can simplify a bit when we have equal sample size ($n_1 = n_2 = n$) or if variances are equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$; σ can be brought out of the square root sign in the denominator). See "Comparison of Means Flowchart" for simplifications of the above test statistic for when variances are known.

We then compare our test statistic to critical z-values. To accept a lower-tailed alternate hypothesis ($H_1: \mu_1 < \mu_2$), we must show that $z_0 < -z_\alpha$. To accept an upper-tailed alternate hypothesis ($H_1: \mu_1 > \mu_2$), we must show that $z_0 > z_\alpha$. And to accept a two-tailed alternate hypothesis ($H_1: \mu_1 \neq \mu_2$), we must show that $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$.

We can calculate P-values for the above tests. The P-value for a lower-tailed test ($H_1: \mu_1 < \mu_2$) is the area in the tail to the left of z_0: P-value = $P[z_0 < -z_\alpha]$, or NORM.DIST(z0,TRUE) in Excel. The P-value for an upper-tailed alternate hypothesis ($H_1: \mu_1 > \mu_2$) is the area in the tail to the right of z_0: P-value = $P[z_0 > z_\alpha]$, or 1-NORM.DIST(z0,TRUE) in Excel. And the P-value for a two-tailed alternate hypothesis ($H_1: \mu_1 \neq \mu_2$) is either: 1) twice the area in the tail to the left of z_0 if $z_0 < 0$: P-value = $2 \cdot P[z_0 < -z_{\alpha/2}]$ (2*NORM.DIST(z0,TRUE) in Excel), or 2) twice the area in the tail to the right of z_0 if $z_0 > 0$: P-value = $2 \cdot P[z_0 > z_{\alpha/2}]$ (2*(1-NORM.DIST(z0,TRUE)) in Excel).

Comparison of Means, Variance Unknown (Parts 1 and 2)

The analysis for comparison of means when variance is unknown is nearly identical for the case when variance is known except that we base our analysis on the T distribution instead of the standard normal distribution.

We can create a test statistic, depending on whether variances are assumed to be equal or unequal (see below in "The F Distribution and Comparison of Variances" for the F test, which allows us to conclude whether we should assume the "variance equal" case or "variance unequal" case). For the case where variances are assumed to be equal but unknown, the test statistic is:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Here, s_p is the pooled sample standard deviation, and the pooled sample variance (s_p^2) is given by:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The test statistic above simplifies a little bit when the sample sizes are equal (see "Comparison of Means Flowchart").

For the case where variances are assumed to be unequal and unknown, the test statistic is:

$$t_0^* = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Again, the test statistic above simplifies a little bit when the sample sizes are equal (see "Comparison of Means Flowchart").

Like the analysis for variance known, we compare our test statistic with a critical t-value based upon our Type I error (α). To accept a lower-tailed alternate hypothesis ($H_1: \mu_1 < \mu_2$), we must show that $t_0 < -t_{\alpha,\nu}$. To accept an upper-tailed alternate hypothesis ($H_1: \mu_1 > \mu_2$), we must show that $t_0 > t_{\alpha,\nu}$. And to accept a two-tailed alternate hypothesis ($H_1: \mu_1 \neq \mu_2$), we must show that $t_0 < -t_{\alpha/2,\nu}$ or $t_0 > t_{\alpha/2,\nu}$.

In all of these cases, ν represents the degrees of freedom. For cases where variances are equal but unknown (see above, any case that we calculate a pooled variance, s_p^2), the degrees of freedom (ν) are equal to $n_1 + n_2 - 2$. For cases in which variances are unequal and unknown (when we calculate t_0^*), degrees of freedom (ν) are equal to the following:

$$\nu = INT\left[\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}\right]$$

We can calculate P-values for the above tests. The P-value for a lower-tailed test ($H_1: \mu_1 < \mu_2$) is the area in the tail to the left of t_0: P-value = $P[t_0 < -t_{\alpha,\nu}]$, or T.DIST(t0,DOF,TRUE) in Excel (DOF is ν, the degrees of freedom). The P-value for an upper-tailed alternate hypothesis ($H_1: \mu_1 > \mu_2$) is the area in the tail to the right of t_0: P-value = $P[t_0 > t_{\alpha,\nu}]$, or T.DIST.RT(t0,DOF) in Excel. And the P-value for a two-tailed alternate hypothesis ($H_1: \mu_1 \neq \mu_2$) is either: 1) twice the area in the tail to the left of t_0 if $t_0 < 0$: P-value = $2 \cdot P[t_0 < -t_{\alpha/2,\nu}]$ (2*T.DIST(t0,DOF,TRUE) in Excel), or 2) twice the area in the tail to the right of t_0 if $t_0 > 0$: P-value = $2 \cdot P[t_0 > t_{\alpha/2,\nu}]$ (2*T.DIST.RT(t0,DOF) in Excel).

Paired T-Tests

A paired T-test is when it makes more sense to pair increases or decreases in a parameter rather than to keep data independent. For example, if you were studying the increase in effectiveness of a skin cream to alleviate rashes, every person responds differently so you would look at the left arm (control) vs. right arm (treatment). It makes more sense to look at the average of the differences for individuals rather than differences in the averages. Or we have two temperature measuring devices and we record the daily high temperature on different days for each device – it's the difference in measurements that we are interested in.

The hypothesis test that we set up is a lower-tailed, upper-tailed, or two-tailed test. Here is an example of an upper-tailed test:

$$H_0: \mu_D = 0$$

$$H_1: \mu_D > 0$$

Note that we are interested in the average of the differences, μ_D. We can create a test statistic based on the average of the differences, $\bar{x}_D$:

$$t_0 = \frac{\bar{x}_D}{s_D/\sqrt{n}}$$

Here, s_D is the sample standard deviation of the differences and n is the sample size (the number of paired observations).

The analysis is just like it is for other hypothesis tests involving the T distribution; we use $n - 1$ degrees of freedom.

The F Distribution and Comparison of Variances

The F test is used to determine whether we can assume the "variance equal" case or "variance unequal" case in the comparison of means (see above) of two populations when variance is unknown.

Like other hypothesis tests, we can perform a lower-tailed, upper-tailed, or two-tailed test. For example, an upper-tailed test:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

Here, σ_1^2 is the variance of population 1 and σ_2^2 is the variance of population 2.

To perform the test, we calculate the following test statistic:

$$f_0 = \frac{s_1^2}{s_2^2}$$

Next, we compare f_0 to the F distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom. To accept a lower-tailed alternate hypothesis ($H_1: \sigma_1^2 < \sigma_2^2$), we must show that $f_0 < f_{1-\alpha,n_1-1,n_2-2}$. To accept an upper-tailed alternate hypothesis ($H_1: \sigma_1^2 > \sigma_2^2$), we must show that $f_0 > f_{\alpha,n_1-1,n_2-1}$. And to accept a two-tailed alternate hypothesis ($H_1: \sigma_1^2 \neq \sigma_2^2$), we must show that $f_0 < f_{1-\alpha/2,n_1-1,n_2-2}$ or $f_0 > f_{\alpha/2,n_1-1,n_2-1}$.

We can calculate P-values for the above tests similar to how we do this for the other tests. The P-value for a lower-tailed test ($H_1: \sigma_1^2 < \sigma_2^2$) is the area in the tail to the left of f_0: P-value = $P[f_0 < f_{1-\alpha,n_1-1,n_2-1}]$, or F.DIST(f0,n1-1,n2-1,TRUE) in Excel. The P-value for an upper-tailed alternate hypothesis ($H_1: \sigma_1^2 > \sigma_2^2$) is the area in the tail to the right of f_0: P-value = $P[f_0 > f_{\alpha,n_1-1,n_2-1}]$, or F.DIST.RT(f0,n1-1,n2-1) in Excel. And the P-value for a two-tailed alternate hypothesis ($H_1: \sigma_1^2 \neq \sigma_2^2$) is either: 1) twice the area in the tail to the left of f_0 for a lower-tailed test: P-value = $2 \cdot P[f_0 < f_{1-\alpha,n_1-1,n_2-1}]$ (2*F.DIST(f0,n1-1,n2-1,TRUE) in Excel), or 2) twice the area in the tail to the right of f_0 for an upper-tailed test: P-value = $2 \cdot P[f_0 > f_{\alpha,n_1-1,n_2-1}]$ (2*F.DIST.RT(f0,n1-1,n2-1) in Excel).

Comparison of Proportions

Finally, Week 4 concludes with hypothesis tests regarding comparison of binomial proportions. There is no direct binomial approach as there was with hypothesis tests on a binomial proportion (see Week 3) so we must base our approach on the normal approximation to the binomial distribution. We aim to determine whether the proportion of population 1 (p_1) having some characteristic is the same as or different from that of population 2 (p_2).

The hypothesis test we set up can be a lower-tailed, upper-tailed, or two-tailed test. Shown here is an example of an upper-tailed test:

$$H_0: p_1 = p_2$$

$$H_1: p_1 > p_2$$

The test statistic that we set up is based on the standard normal distribution:

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Here, $\hat{p}_1 = \frac{X_1}{n_1}, \hat{p}_2 = \frac{X_2}{n_2}, \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ where X_1 and X_2 are the number of items with the characteristic of interest from populations 1 and 2, respectively.

The analysis of the test is exactly the same as for the standard normal distribution (see above) as is the calculation of P-values.