Formulas to know for PSYCH 3101 (Keller):

1) Basic descriptive statistics (Chapters 3-4)

When to use: When we want to describe the characteristics of a sample or for use as components of other inferential formulas

Formulas: $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$

$$s^2 = \frac{SS}{df} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

$$s = \sqrt{s^2}$$

Notes: We divide the SS (sums of squared deviations) by *n*-1 (the degrees of freedom) in order to get an unbiased estimate of the variance. If we divided by *n* instead, our estimate of variance would be too low because the individual scores are necessarily closer to the sample mean than they are to the population mean. If we were to use the population mean in the formula above in the place of $\bar{X}$, we could divide by *n*.

2) z-score (Chapter 5)

When to use: When we want to know how far an individual score is above or below the mean, in units of standard deviation.

Formulas: $z = \dfrac{X - \mu}{\sigma}$ when the population mean/variance are known

$z = \dfrac{X - \bar{X}}{s}$ when the population mean/variance are not known

Notes: Don't get a z-score confused with a z-test. A z-score is simply a conversion of individual scores into a new unit – in this case, units of standard deviation. It's like converting from inches to feet – nothing else about the distribution changes (converting scores to z-scores does NOT make the scores normally distributed). There is no hypothesis testing with z-scores since we are not trying to say anything about a population mean.

3) z-test (Chapters 7-8)

When to use: When we want to use our sample mean to know ("infer") if a population mean is different than some hypothesized mean (a "null" mean) and when we know the standard deviation (or variance) of the population.

Inferential Formula: $z = \dfrac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$

Effect size formula: Cohen's d = $\frac{\bar{X}-\mu_0}{\sigma}$

Confidence interval formula: $[\bar{X}-z_{crit}\times\sigma_{\bar{X}}, \bar{X}+z_{crit}\times\sigma_{\bar{X}}]$

Notes: I've denoted the mean that corresponds to the null hypothesis, or the "null" mean, as μ_0. It should be obvious what the null hypothesis is given the problem at hand. The denominator, $\sigma_{\bar{X}}=\sigma/\sqrt{n}$, is called the "standard error of the mean," and is the standard deviation of a (hypothetical) sampling distribution of means taken from samples of size *n*. The z-statistic is normal if the original scores are normally distributed, but becomes normal whenever *n*>30 by the Central Limit Theorem, regardless of the original distribution of scores. If the z-score is > 1.96 or < -1.96, then we reject the null hypothesis (if $\alpha=.05$) that the population mean is equal to the "null" mean.

4) one-sample t-test (Chapter 9)

When to use: When we want to use our sample mean to know ("infer") if a population mean is different than some hypothesized mean (a "null" mean) and when we do not know the standard deviation (or variance) of the population.

Inferential Formula: $t_{df}=\frac{\bar{X}-\mu_0}{s_{\bar{X}}}=\frac{\bar{X}-\mu_0}{s/\sqrt{n}}$

Effect size formula: Cohen's d = $\frac{\bar{X}-\mu_0}{s}$

Confidence interval formula: $[\bar{X}-t_{crit}\times s_{\bar{X}}, \bar{X}+t_{crit}\times s_{\bar{X}}]$

Notes: As above, the denominator of the inferential formula is called also the "standard error of the mean," and is the standard deviation of a (hypothetical) sampling distribution of means taken from samples of size *n*. The t-statistic is t-distributed if the original scores are normally distributed, but becomes t-distributed whenever *n*>30 by the Central Limit Theorem, regardless of the original distribution of scores. The critical values of *t* corresponding to $\alpha=.05$ (or any other alpha) change depending on the degrees of freedom, which is denoted above as t_{df}.

5) independent-samples t-test (Chapter 10)

When to use: When we want to know whether the population means from which two samples were drawn are different from each other and when we do not know the standard deviations (or variances) of the populations.

Inferential Formula: $t_{df_1+df_2} = t_{n_1+n_2-2} = \frac{(\bar{X}_1 - \bar{X}_2)-(\mu_1-\mu_2)}{s_{(\bar{X}_1-\bar{X}_2)}} = \frac{(\bar{X}_1 - \bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}}$

Pooled Variance Formula: $s_p^2 = \frac{SS_1 + SS_2}{df_1 + df_2}$

Effect size formula: Cohen's d = $\frac{(\bar{X}_1 - \bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{s_p^2}}$

Confidence interval formula: $[(\bar{X}_1 - \bar{X}_2) - t_{crit} \times s_{\bar{X}_1-\bar{X}_2}, (\bar{X}_1 - \bar{X}_2) + t_{crit} \times s_{\bar{X}_1-\bar{X}_2}]$

Notes: We are almost always interested in testing the null hypothesis that $\mu_1 = \mu_2$, i.e., that $\mu_1 - \mu_2 = 0$, so that part of the equations above is often left out. The SS_1 and SS_2 are the sum of squared deviations within each group (i.e., the deviations from each group mean, squared and then summed).

6) repeated measures t-test (Chapter 11)

When to use: When the same individual is measured twice and we want to know whether mean for the first set of scores is higher or lower than the mean of the second set of scores. Also, we do not know the standard deviation (or variation) of the population. Can I stop saying that?

Inferential Formula: $t_{df} = \frac{\bar{X}_D - \mu_D}{s_{\bar{X}_D}} = \frac{\bar{X}_D - \mu_D}{s_D / \sqrt{n}}$

Effect size formula: Cohen's d = $\frac{\bar{X}_D - \mu_D}{s_D}$

Confidence interval formula: $[\bar{X}_D - t_{crit} \times s_{\bar{X}_D}, \bar{X}_D + t_{crit} \times s_{\bar{X}_D}]$

Notes: The repeated measures t-test is extremely simple. Start by finding the difference between the scores of the same person (e.g., scores time 1 – scores time 2 for each person). Then conduct a one-sample t-test on those difference scores. The subscript "D" in the statistics above means that those statistics were conducted on difference scores, so $\bar{X}_D$ is the mean of the difference scores, and s_D is the standard deviation of the difference scores. The null hypothesized difference, μ_D, is almost always hypothesized to be 0 (e.g., we're interested in whether time 1 is different than time 2 scores).

7) single factor (or "one-way") ANOVA (Chapter 13)

When to use: When we want to know whether the population means from which 3 or more samples were drawn are different from each other and when we do not know the standard deviations (or variances) of the populations.

Sum of Squares Formulas: $SS_{between} = \sum_{j=1}^{k} n_j (\bar{X}_j - \bar{\bar{X}})^2$

$$SS_{within} = \sum_{j=1}^{k} SS_j = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (X_i - \bar{X}_j)^2$$

$$SS_{Total} = SS_{between} + SS_{within} = \sum_{i}^{N} (X_i - \bar{\bar{X}})^2$$

Degrees of Freedom Formulas: $df_{between} = k - 1$

$$df_{within} = df_1 + df_2 + ...df_k = N - k$$
$$df_{Total} = df_{between} + df_{within} = N - 1$$

Mean Squares Formulas:

$$MS_{between} = SS_{between} / df_{between}$$
$$MS_{within} = SS_{within} / df_{within}$$

Inferential Formula: $F_{df_{between}, df_{within}} = MS_{between} / MS_{within}$

Effect size formula: $r^2 = \frac{SS_{between}}{SS_{Total}} = \frac{SS_{between}}{SS_{between} + SS_{within}}$

Notes: $\bar{\bar{X}}$ is the "grand mean," or the mean of all the scores (not necessarily the same as the mean of the means). *N* stands for the total sample size (all individuals summed across all groups). The "MS" nomenclature stands for "Mean Squared," and can be considered variance estimates (in the case of MS_{within}) or variance estimates upweighted by sample size (in the case of $MS_{between}$). $MS_{between}$ is an unbiased estimator of MS_{within} if the null hypothesis is true; i.e., $MS_{between} \approx MS_{within}$ or $\frac{MS_{between}}{MS_{within}} = F \approx 1$ if the null hypothesis is true. The formula for MS_{within} is an obvious extension of the formula for the pooled variance, seen above for independent samples t-tests. There are two degrees of freedom for an *F* test – the first corresponding to the numerator (the $df_{between}$) and the second to the denominator (the df_{within}). Critical F values, generally corresponding to $\alpha = .05$, must be looked up in a table or in R. F statistics always increase (get bigger than 1) to the degree that the null hypothesis is false (i.e., that the means differ from one another), so you're looking for larger F values only in order to reject the null hypothesis that all the population means are the same (i.e., it isn't a symmetric distribution like a z- or t-distribution).