Parameter	Statistic	SE for CI and HT	Test Statistic for HT	Critical Value for CI or HT (always $\alpha/2$ for CI)	Sample Size Estimation
μ for σ known and either large sample ($n \geq 30$) or population normally distributed	$\bar{x}$	$\frac{\sigma}{\sqrt{n}}$	$z = \frac{\bar{x}-\mu_0}{SE}$ where μ_0 is the null hypothesized value of μ	$z_{\alpha/2}$ or z_α	$n \geq \frac{z_{\alpha/2}^2\sigma^2}{B^2}$ where B is the desired ME
μ for σ unknown population approximately normally distributed	$\bar{x}$	$\frac{s}{\sqrt{n}}$	$t = \frac{\bar{x}-\mu_0}{SE}$ where μ_0 is null hypothesized value of μ	$t_{\alpha/2,df}$ or $t_{\alpha,df}$ where $df = n - 1$	
$\mu_1 - \mu_2$ for σ known and either large samples (n's ≥ 30) or populations normally distributed	$\bar{x}_1 - \bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	$z = \frac{(\bar{x}_1-\bar{x}_2)-D_0}{SE}$ where D_0 is the null hypothesized difference between μ_1 and μ_2	$z_{\alpha/2}$ or z_α	$n \geq \frac{z_{\alpha/2}^2(\sigma_1^2+\sigma_2^2)}{B^2}$ where B is the desired ME and $n_1 = n_2 = n$
$\mu_1 - \mu_2$ for σ unknown and populations approx. normally distributed, equal variances	$\bar{x}_1 - \bar{x}_2$	Pooled SE: $\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ where $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$t = \frac{(\bar{x}_1-\bar{x}_2)-D_0}{SE}$ where D_0 is the null hypothesized difference between μ_1 and μ_2	$t_{\alpha/2,df}$ or $t_{\alpha,df}$ where $df = n_1 + n_2 - 2$	
$\mu_1 - \mu_2$ for σ unknown and populations approx. normally distributed, not willing to assume equal variances	$\bar{x}_1 - \bar{x}_2$	Unpooled SE: $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$;	$t = \frac{(\bar{x}_1-\bar{x}_2)-D_0}{SE}$ where D_0 is the null hypothesized difference between μ_1 and μ_2	$t_{\alpha/2,df}$ or $t_{\alpha,df}$ where df = Satterthwaite's approximation, but use $df = min(n_1 - 1, n_2 - 1)$ for the written HW and exams	
μ_d	$\bar{d}$	$\frac{s_d}{\sqrt{n}}$ where n = # of pairs	$t = \frac{\bar{x}-\mu_0}{SE}$ where μ_0 is the null hypothesized value of μ_d	$t_{\alpha/2,df}$ or $t_{\alpha,df}$ where $df = n - 1$	$n \geq \frac{z_{\alpha/2}^2\sigma_d^2}{B^2}$ where B is the desired ME

Assumptions

1. Data are random samples representative of the population of interest
2. For two-sample tests the samples are independent (unless a matched-pair design)
3. When pooling variances, you are assuming that $\sigma_1^2 = \sigma_2^2$; use rule-of thumb $\frac{Larger\ s^2}{Smaller\ s^2}$; if ratio < 3 assumption of equal variances probably okay; if > 3, assumption of equal variances probably not a good idea. There are more formal tests for looking at the equality of variances, but we don't cover them in this course. Also, you need to think carefully about pooling if the sample sizes are quite different, especially if the larger sample size is associated with the smallest variance. On the final I will be clear which SE I want you to use, although you should be able to comment on appropriateness.

Parameter	**Statistic**	**SE for CI**	**SE for HT**	**Test Statistic for HT**	**Critical Value for CI or HT (always $\alpha/2$ for CI)**	**Sample Size Estimation**
p	$\hat{p}$	$\sqrt{\frac{pq}{n}} \approx \sqrt{\frac{\hat{p}\hat{q}}{n}}$	$\sqrt{\frac{p_0q_0}{n}}$ where p_0 is the null hypothesized value of p	$z = \frac{\hat{p}-p_0}{SE}$ where p_0 is the null hypothesized value of p	$z_{\alpha/2}$ or z_α	$n \geq \frac{z_{\alpha/2}^2 pq}{B^2}$ where B is the desired ME
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}} \approx$ $\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$	$\sqrt{p_0q_0\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ where p_0 is the pooled sample estimate for p (sometimes denoted as $\bar{p}$ *or* $\hat{p}$ with no subscripts for the two populations) when the null hypothesis is $p_1 = p_2$ or $p_1 - p_2 = 0$ If the null hypothesis is something else, then use SE for CI	$z = \frac{(\hat{p}_1-\hat{p}_2)-D_0}{SE}$ where D_0 is the null hypothesized difference between p_1 and p_2	$z_{\alpha/2}$ or z_α	$n \geq \frac{z_{\alpha/2}^2(p_1q_1+p_2q_2)}{B^2}$ where B is the desired ME and $n_1 = n_2 = n$

Assumptions

1. Data are random samples representative of the population of interest
2. Data can be considered as binomial proportions where $\hat{p} = \frac{x\ total\ successes}{n\ total\ trials}$; for pooled estimate for H_0: $p_1 - p_2 = 0$, $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$
3. For two-sample tests the samples are independent
4. Only use these methods if sample sizes are sufficiently large: $np \geq 10$ and $nq \geq 10$ for one sample; $n_1p_1 \geq 10, n_1q_1 \geq 10, n_2p_2 \geq 10,$ and $n_2q_2 \geq 10$

Chi-square Tests

Test	Hypotheses	Expected Counts	Test Statistic	Critical for HT
Independence: to test for an association between two categorical variables in a single population.	H_0: Variable A and Variable B are independent (not associated). H_a: Variable A and Variable B are dependent (associated).	$E_i = \frac{R_i C_j}{N}$ where R_i is the row total in row *i*, C_j is the column total for column *j*, and *N* is the grand total number of observations.	$\chi^2_{obs} = \sum_{i=1}^{\# \, cells} \frac{(O_i - E_i)^2}{E_i}$ where O_i is the observed count in cell *i* and E_i is the expected count for cell *i* given that the null hypothesis is true. Note that the cells are laid out in a contingency table, so technically, the test statistic is a double sum over rows *i*, and columns *j*.	$\chi^2_{\alpha,df}$ where *df* = (*r* – 1)(*c* – 1) and *r* = # rows and *c* = # columns in the contingency table of two categorical variables. *p*-value = $P(\chi^2_{\alpha,df} > \chi^2_{obs})$
Homogeneity: to test for equality of proportions for one variable across two or more populations.	H_0: The distribution of proportions for Variable A in population 1 is the same as that in population 2. ($p_{A1,1} = p_{A1,2}; p_{A2,1} = p_{A2,2};\ p_{A3,1} = p_{A3,2}; \ldots;\ p_{An,1} = p_{An,2}$) where *A* is a categorical variable with *n* levels and the second subscript are two populations, such a male/female or years 2015/2016. H_a: At least one pair of proportions for Variable A between the two populations is not equal.	Same as above	Same as above	χ^2_{obs} where *df* = (*r* – 1)(*c* – 1) and *r* = # rows and *c* = # columns in the contingency table of two categorical variables. *p*-value same as above
Goodness-of-fit: to test the proportions of a categorical variable against a fixed set of proportions.	H_0: The distribution of proportions for Variable *A* in population 1 is the same as that in population 2. (for a variable with three levels, $p_1 = 0.1; p_2 = 0.3;\ p_3 = 0.6$. The fixed probabilities must sum to 1 over the *n* levels of Variable *A*). H_a: At least one proportion in the null is not as specified.	$E_i = N(p_i)$ where p_i is the null hypothesized proportion for level *i* of Variable *A* and *N* is the grand total number of observations.	$\chi^2_{obs} = \sum_{i=1}^{\# \, levels \, of \, A} \frac{(O_i - E_i)^2}{E_i}$; a double sum does not apply here.	$\chi^2_{\alpha,df}$ where *df* = *k* – 1 where *k* is the number of possible outcomes for a single categorical variable. *p*-value same as above

Parameter	Statistic	SE for CI	SE for HT	Test Statistic for HT	Critical Value for CI or HT (always $\alpha/2$ for CI)
β_1	b_1	$SE(b_1) = \sqrt{\frac{s_e^2}{\mathrm{SS}_{xx}}} = \frac{s_e}{s_x\sqrt{n-1}}$ where $s_x = $ SD of x, $\left(s_e^2 = \frac{\mathrm{SSE}}{n-2}\right)$ and $\mathrm{SS}_{xx} = \sum(x_i - \bar{x})^2$	Same as CI	$t = \frac{b_1 - 0}{SE(b_1)}$ where 0 is null hypothesized value of β_1, meaning no linear relationship	$t_{\alpha/2,df}$ or $t_{\alpha,df}$ where $df = n - 2$
$E(y\|x = x^*)$ or $y\|x = x^*$	$\hat{y}$	SE for CI of mean, $E(y\|x = x^*)$: $\sqrt{s_e^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\mathrm{SS}_{xx}}\right)} = \sqrt{SE^2(b_1)\times(x^* - \bar{x})^2 + \frac{s_e^2}{n}}$ where $\mathrm{SS}_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$; Note: the formulas in blue are the ones given in your text. SE for PI for a future observation of $y\|x = x^*$: $\sqrt{s_e^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\mathrm{SS}_{xx}}\right)}$ $= \sqrt{SE^2(b_1)\times(x^* - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$	No hypothesis testing		$t_{\alpha/2,df}$ where $df = n - 2$

Assumptions for Simple Linear Regression

1. Relationship between X and Y over the range of the data is linear
2. Y values are independent random sample
3. X values assumed to be error-free (no measurement error)
4. The random error epsilon is such that ε i. i. d. $N(0, \sigma^2)$, where i.i.d. means independent, identically distributed

For regression problems on the final I will provide values of n, SSE, SS_{xx}, SS_{yy}, SS_{xy}, $\bar{x}$, and $\bar{y}$. With these values you can derive the regression equation and the s_e^2 that is required for inference. Some relationships to note are:

$$b_1 = \frac{\mathrm{SS}_{xy}}{\mathrm{SS}_{xx}} = r\frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1\bar{x}, \quad \mathrm{SS}_{Total} = \mathrm{SS}_{yy}, \quad \mathrm{SS}_{Total} = \mathrm{SS}_{Model} + \mathrm{SS}_{Resid}, \quad r^2 = \frac{\mathrm{SS}_{Model}}{\mathrm{SS}_{Total}} = \frac{\mathrm{SS}_{Total} - \mathrm{SS}_{Resid}}{\mathrm{SS}_{Total}}$$

$\mathrm{SS}_{Total} = \sum(y_i - \bar{y})^2$; $\mathrm{SS}_{Model} = \sum(\hat{y}_i - \bar{y})^2$; $\mathrm{SS}_{Resid} = SSE = \sum(y_i - \hat{y}_i)^2$, where SS_{Model} is the sum of squares due to the regression (SSR).