Review sheet

1 T-test procedure

1. Calculate your **t statistic** exactly as you did for the z test (whether comparing one sample against a fixed hypothesis or two samples against each other).
2. But you now have to use the **degree of freedom** to find the critical value:
   - $df = n - 1$ for a one-sample test.
   - $df = n_1 + n_2 - 2$ for a two-sample test with equal variances in each sample
   - $df = \frac{se_{ab}^2}{se_a^4/(n_a-1)+se_b^4/(n_b-1)}$ for unequal variances (but you don't need to know this!)
3. Use the t-table to find the critical value for your df and your chosen p-value. Eg, look under $T_{0.05}$ for $p = 0.05$ or to construct a 90% CI, or $T_{0.025}$ to construct the 95% CI.
4. If your t statistic is greater than the critical value, or the null (usually 0) lies outside the CI, reject the null.

2 Chi-square test procedure

1. H_0 = variables are independent
   H_a = variables not independent
2. Calculate f_e (expected quantity) for each cell: $f_e = \frac{\text{(row total)(column total)}}{\text{overall total}}$
3. Calculate you **chi squared statistic** $\chi^2 = \sum \frac{(f_o-f_e)^2}{f_e}$
4. Calculate $df = (r-1)(c-1)$
5. Find the critical χ^2 value for your degree of freedom and p-value (usually $p = 0.05$; always one-tailed).
6. If your calculated statistic from (3) is greater than the critical value, reject the null.

3 ANOVA (f-test) procedure

1. G groups, N total observations. Are the means of these G groups the same?
2. Calculate your **f-statistic** $f = \frac{\text{average variance between groups}}{\text{average variance within groups}}$
   - Between variance $= \frac{n_1(\bar{y}_1-\bar{y})^2+...+n_G(\bar{y}_G-\bar{y})^2}{G-1}$
   - Within variance $= \frac{(n_1-1)s_1^2+...+(n_G-1)s_G^2}{N-G}$
3. Determine your degrees of freedom: $df_1 = G - 1$; $df_2 = N - G$.
4. Find your critical value in the f-table using your df values and chosen p-value (eg. 0.05).
5. If your f-statistic is greater than the critical value, reject the null

4 Regression and correlation (second half of the semester)

Covariance of x and y: $\text{Cov}(x,y) = \frac{1}{(n-1)} \sum_i (x_i - \bar{x})(y_i - \bar{y})$

Correlation of x and y: $r = \frac{\text{Cov}(x,y)}{s_x s_y}$

Regression of y on x_1, x_2, x_3...: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$

where y is the **dependent variable**, the x's are the **independent variables**, β_0 is the **intercept** or **constant** (the value of y when all the x's are 0), β_1 is the **effect** of a one-unit increase in x_1 on y, **controlling for** all the other variables, and so on for the other β values.

To predict a value of $\hat{y}$ for some specific set of **x** values, just plug them into the above equation:

$$\hat{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + ...$$

For a **bivariate** (single x variable) regression:

$$\beta_1 = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

The β values are too complicated to calculate by hand when there is more than one x variable.

R^2 (**R squared**) is the proportion of the variation in y explained by the regression model. R^2 can range between 0 (none) and 1 (all the variation is explained).

For a *bivariate regression only*, $R^2 = r^2$.

But for all regressions (no matter how many x variables):

$$R^2 = \frac{TSS - SSE}{TSS}$$

where TSS = **Total Sum of Squares** and SSE = **Sum of Squared Errors**

$$TSS = \sum_i (y_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

Remember, to calculate the SSE, you will have to calculate $\hat{y}_i$ for every observation in your data set, so you need your β coefficients first.

To test whether a given variable x_i has a **statistically significant effect** on y, you must test whether the β_i associated with that x_i is significantly different from 0:

- This is a two-tailed test, where we must know the degree of freedom ($df = n - 2$), the standard error of β_i (provided by a computer, usually), and of course our p-value threshold (usually 0.05).
- We can either construct a 95% CI around β_i: $\beta_i \pm$[critical value from t-table] * [standard error of β_i] and see whether this includes 0.
- Or we can calculate our test statistic for β_i: $t = \beta_i /$[standard error of β_i], and compare that against a critical value from the t-table.
- The computer output often provides all of these things: the 95% CI for each β, the test statistic, and the exact p-value. The rule of thumb is that if β_i is at least twice the standard error se_{β_i}, then it is significant. If the exact p-value (or "significance") is reported, then you just need to see whether that is less that 0.05, in which case x_i's effect on y is significant.

The **standardized beta** is how much a one-standard-deviation change in x changes y, also measured in standard deviations. This just changes the units of x and y. It allows us to directly compare the relative effects of multiple x variables on y against each other; a bigger standardized beta means a more significant effect (though not necessarily one that is more substantively important).

For a bivariate regression, the standardized beta is just the correlation r.

The difference between a **bivariate regression** (one independent variable, x) and a **multivariate regression** (two or more independent variables) is in what the effect sizes β mean. In the bivariate case, the effect of x_1 (β_1) is the total effect of x_1, including indirect effects (eg, Age might boost Education which in turn boosts Income). In a multlivariate regression, the β_1 for that same x_1 variable is only the *direct effect*, holding all the other variables constant. That is, if Age increases Education which in turn increases Income, the direct effect of Age on Income is what happens when we look at only people of the same Education ("holding Education constant" or "controlling for Education"), so there's no indirect effect, just the direct effect of Age on Income (eg, via the gradual raises people get over time).

Statistical testing

Y variable	X variable	Research question (in casual English)	Null hypothesis	R. hypoth. (two tailed)	Test statistic	Degree of freedom	Test type
Continuous	NA	Is the population mean ($\bar{y}$) equal to some number (μ_0)?	$\bar{y} = \mu_0$	$\bar{y} \neq \mu_0$	$t = \frac{\bar{y} - \mu_0}{se_y}$	$n - 1$	T test
Continuous	Categorical (= 2)	Does the mean of group 1 ($\bar{y}_1$) equal the mean of group 2 ($\bar{y}_2$)?	$\bar{y}_1 = \bar{y}_2$	$\bar{y}_1 \neq \bar{y}_2$	$t = \frac{\bar{y}_1 - \bar{y}_2}{se_{12}}$	$n_1 + n_2 - 2$ (equal variances)	T test
Continuous	Categorical (> 2)	Are the means of all the groups equal?	$\bar{y}_1 = \bar{y}_2 = ... = \bar{y}_G$	At least one mean unequal	$f = \frac{\text{Between var.}}{\text{Within var.}}$	$df_1 = G - 1$ $df_2 = N - G$	f test (ANOVA)
Categorical	Categorical	Are the two variables independent?	$f_o = f_e$ for all cells	At least one cell differs from f_e	$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$	$(r - 1)(c - 1)$	χ^2 test
Continuous	Continuous	Is the "effect" of X on Y statistically significant in a regression: $y = \beta_0 + \beta_1 x$?	$\beta_1 = 0$	$\beta_1 \neq 0$	$t = \frac{\beta_1}{se_{\beta_1}}$	$n - 2$	T test
Categorical	Continuous	Y is binary: Logistic regression (logit). Y is ordinal: ordered logit. Y is categorical (>2): multinomial logit. The tests are the same as in the regular linear regression, however.					

There are three mathematically equivalent methods for testing your hypothesis. I recommend method 1 because it is the easiest and most general.

1. Choose your p value threshold (usually p (or α) $= 0.05$), and use your df to find the critical value in the appropriate table (t, f, χ^2).
   (Remember to divide your p-value by 2 if you are doing a two-tailed test, as is common. But recall that the f and χ^2 are always one-tailed tests.)
   If your test statistic is greater than this critical value, reject the null in favor of your research hypothesis.

2. (For T-test only.) Generate your 95% confidence interval around $\bar{y}$, $\bar{y}_1 - \bar{y}_2$, or β_1.
   CI = mean ± critical value (from t table) * standard error.
   If that CI does not include the null value (μ_0, which = 0 for the difference test or β_1), reject the null.

3. Put your test statistic and the degree of freedom(s) into a f, t, or χ^2 calculator to get an exact p-value.
   If this p-value is less than 0.05, reject the null.