The table below broadly summarizes all the inference techniques you have learnt in this course. For the specific details (assumptions, hypothesis, etc), please look at the Big_5 handout, the ANOVA overview and your notes on regression and Chi-square tests. There is also the Big Ideas handout posted.

<table>
<thead>
<tr>
<th>Test (Method)</th>
<th>Response (Y)</th>
<th>Predictor (X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small sample binomial test</td>
<td>Categorical</td>
<td>N/A</td>
</tr>
<tr>
<td>One-sample z-test</td>
<td>Categorical</td>
<td>N/A</td>
</tr>
<tr>
<td>2 sample z-test</td>
<td>Categorical</td>
<td>Categorical (2 groups)</td>
</tr>
<tr>
<td>One-sample t-test</td>
<td>Numeric</td>
<td>N/A</td>
</tr>
<tr>
<td>Paired</td>
<td>2 sets Numeric outcomes → looking at differences</td>
<td></td>
</tr>
<tr>
<td>2 sample t-test</td>
<td>Numeric</td>
<td>Categorical (2 groups)</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Numeric</td>
<td>Categorical ( &gt; 2 groups)</td>
</tr>
<tr>
<td>Regression</td>
<td>Numeric</td>
<td>Numeric</td>
</tr>
<tr>
<td>Chi-square</td>
<td>Categorical</td>
<td>Categorical</td>
</tr>
</tbody>
</table>

The rest of this handout focuses on what you have learnt after exam 2. You should refer to your previous review sheets for the previous material.

ANOVA:
- **Purpose:** To compare 3 or more group means
- **Hypotheses:** Ho: \( \mu_1 = \mu_2 = \ldots = \mu_i \); Ha: at least one population mean is different
- **Assumptions:** Response for each population is assumed to be normal; all population variances assumed equal; independent random samples
- **Test Statistic:** \( F \), degrees of freedom = \( (k-1), (N-k) \)

Regression:
- **Model and Hypotheses:**
  - **True Model:** \( E(y) = \beta_0 + \beta_1(x) \)
  - **Estimated Model:** \( \hat{y} = b_0 + b_1(x) \)
  - **Hypothesis to test the significance of the model:**
    - (a) Ho: \( \beta_1 = 0 \) vs. Ha: \( \beta_1 \neq 0 \) (OR \( \beta_1 > 0 \) OR \( \beta_1 < 0 \))
    - ... note that this is equivalent to:
    - (b) Ho: There is not a significant (positive OR negative) linear relationship between the two variables,
      Ha: There is a significant linear (positive OR negative) relationship between the two variables
- **Assumptions:**
  - The residuals \( (e_i = y_i - \hat{y}_i) \) need to be normally distributed with a mean of zero and constant variance.
  - The normality assumption is best checked with a QQ-plot; the constant variance assumption can be checked with a residual plot.
- **Test Statistic:**
  - There are two test statistics you can use to the hypotheses above:
    - (1) \( t \), df = n-2 (p-value found in the Regression Coefficients output in SPSS, the row for x)
    - (2) \( F \), df = \( 1, n-2 \) (p-value found in the Regression ANOVA output in SPSS)
    - * These are testing the same hypothesis, so they will give the same p-value.
- **Interpretations:**
  - Pearson’s correlation coefficient \( (r) \): \( r \) measures the strength and the direction of the **linear** relationship between two quantitative variables. A value close to 1 (or -1) indicates a strong positive (or negative) relationship.

For the interpretations: note that words that appear in bold are key; words that are italics are “blanks” that need to be filled in based on the context of the problem.
* SPSS gives the absolute value of \( r \) in the Model Summary output; you can get its sign from the slope of the line. The value of \( r \) (with the proper sign) can also be found in the Regression Coefficients output under the heading Standardized Coefficients Beta.
* Also note that correlation does not imply causation. There are four interpretations of an observed association:
  1. There is causation. The explanatory variable is indeed causing a change in the response variable.
  2. There may be causation, but confounding factors contribute as well and make this causation difficult to prove.
  3. There is no causation. The association is explained by how the explanatory and response variables are both affected by other variables.
  4. The response variable is causing a change in the explanatory variable.

- Coefficient of Determination (\( r \)-square): \( r \)-square*100% of the observed variation in \( y \) is explained by its linear relationship with \( x \).
  * The value of \( r \)-square can be found directly in the Model Summary in SPSS, and remember that it is the square of the \( r \) value.
- Standard deviation from the regression line (s): On average, the actual observations for \( y \) differs by roughly \( s \) units from the regression line prediction based on \( x \).
  * \( s \) can be thought of as measuring the average size of the residuals.
  * The value of \( s \) can be found in two places in SPSS output. First, it can be read directly from the Model Summary. Second, it is the square root of MSE in the AVOVA.
- Estimated Slope (\( b_1 \)): The mean increase (or decrease) in \( y \) expected for each 1-unit increase in \( x \) is estimated to be \( b_1 \) units.
- Estimated Intercept (\( b_0 \)): \( y \) will take the value \( b_0 \) when \( x \) equals zero.
  * For this quantity, you want to think if it’s possible for \( x \) to actually take the value zero. You don’t want to interpret or examine the significance of a variable that doesn’t make sense.

Intervals:
- There are three main intervals you could be asked to compute:
  1. Confidence Interval for the true slope (\( b_1 \))
    * The interpretation for this interval is the same as before; just adapt it to the specific situation.
  2. Confidence Interval for the mean response
    * Think: Am I being asked about an average for an entire group of people?
  3. Prediction Interval for an individual response
    * Think: Am I being asked about a possible response for a particular type of person?
- Notice that a prediction interval will always be wider than a confidence interval (you can see that the standard errors in the formula for the PI includes \( s^2 \), but the formula for the CI does not). This is because a CI in for a population parameter that is unknown but fixed, but a PI is for a random variable (the individual’s response).
- To calculate \( S_{xx} = \sum (x_i - \bar{x})^2 \) (needed for each of these intervals):
  1. It could be given to you directly.
  2. You could use the sample standard deviation of \( x \), by solving \( s_x = \sqrt{\frac{S_{xx}}{n-1}} \) for \( S_{xx} \)

\[ S_{xx} = (n-1)s_x^2. \]
  3. You could use the standard error of the slope, by solving \( s.e.(b_1) = \frac{s}{\sqrt{S_{xx}}} \) for \( S_{xx} \)

\[ S_{xx} = \frac{s^2}{(s.e.(b_1))^2}. \]

For the interpretations: note that words that appear in bold are key; words that are italics are “blanks” that need to be filled in based on the context of the problem.
Chi-Square:

(1) Test of Independence:
* Think: Are there two variables of interest that are being compared for the entire sample?
  - H_o: The two variables are not related for the population.
  - H_a: The two variables are related for the population.
* The expected counts are counts that would be expected for that combination of factors in the long run if the null hypothesis were true (i.e., if the two variables are not related).

(2) Test of Homogeneity:
* Think: Is there really only one variable of interest that is being compared across two (or more) groups?
  - H_o: The distribution for the response variable is the same for all populations.
  - H_a: The distribution for the response variable is not the same for all populations.
* Expected counts are hypothetical counts that would occur if the null hypothesis were true.

(3) Test for Goodness of Fit:
* Think: Am I given a bunch of numbers and asked to test the proportions for a variable of interest?
  - H_o: p_1=p_{1o}, p_2=p_{2o}, ..., p_i=p_{io} (i.e., all proportions are really what you are told they are).
  - H_a: The distribution of the variable is not given by the above proportions.

For all three tests:
- Assumptions: Large sample (all cells have expected count at least 1; at least 80% have expected count at least 5). Note that this assumption can be verified by the footnote on SPSS Chi-square test output.
- Test Stat: $\chi^2$ (sometimes called the Pearson Chi-Square in SPSS output)
- P-value given in SPSS output or in table A.5
* Note that the degrees of freedom depend on the test:
  (1) df = (# of rows - 1)*(# of columns - 1)
  (2) df = (# of rows - 1)*(# of columns - 1)
  (3) df = (# of groups - 1).
* Also note that the expected value of this test statistic is the degrees of freedom.

* Decision rules for ANOVA, regression and chi-square are the same as always:
  Reject H_o if p-value $\leq \alpha$
* Remember that your conclusion should be given in terms of H_a (based on the specific context of the problem).

For the interpretations: note that words that appear in bold are key; words that are italics are “blanks” that need to be filled in based on the context of the problem.