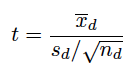
Cheat Sheet Exam 3

**Paired Sample Test**

Just use the one sample T test on the differences.



df = n-1

Two Sample Tests for mean x and mean y

When to Use:

We want to compare two population means.

We have independent samples.

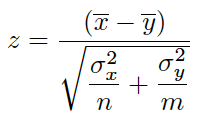
Assumptions: The populations are normally distributed or both sample sizes are at least 30.

Null Hypothesis: 

Test Statistic:

**Two Sample z Test:**

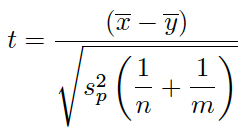
If you know the population variances sigma squared (x and y) use the test statistic



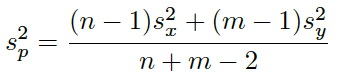
and the standard normal distribution.

**Two Sample t Test (Equal Variances):**

If you don't know the population variances sigma squared (x and y), but you think that they are the same, use the test statistic



where the pooled variance is

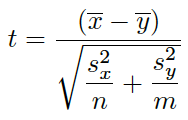


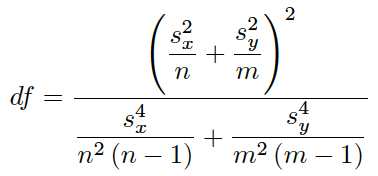
and the t distribution with degrees of freedom df = n + m - 2.

**Two Sample t Test (Unequal Variances):**

If you don't know the population variances sigma squared (x and y), but you think that they are not the same, use the test

statistic



and the t distribution with the Satterthwaite approximation of degrees of freedom

**The ANOVA F Test**

When to Use: We want to compare several population means.

conditions: (We won't check these)

Null Hypothesis: H0 : all the population means are equal

Alternative Hypothesis: Ha : at least one of the population means is different

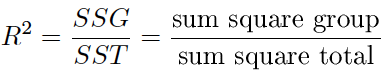
Test Statistic: F = MSG/MSE =variation between groups/

variation within groups

P-Value: The p-value is the area under the F curve to the right of the test statistic.

Remember that the alternative hypothesis only tells us that at least one of the population means is different. If we decide that at least one of the population means is different, we could go on to do a POST-HOC test to compare the different population means to each other to try to pinpoint which of the means is different from the others.

**Coefficient of Determination**



R2 tells us the percentage of the total variation that is explained by differences between the groups. The rest of the variation comes from the natural variation from person to person within a group.

**Chi-Square Test for Goodness of Fit**

When to Use:

We have a proposed distribution and we want to know if it is a good fit for a data. (We want to know if our theoretical probabilities are correct).

The total sample size is n.

The sample items are classified into one of k groups.

Conditions: All the expected cell counts are at least 5.

Null Hypothesis: H0 : all our theoretical probabilities are correct **or**

H0 : the proposed distribution is a good fit for our data

Alternative Hypothesis: Ha : at least one of the theoretical probabilities is incorrect **or**

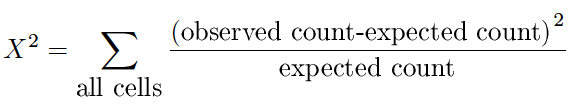
Ha : the proposed distribution is not a good fit for our data

Expected Count:

expected count=n \* pi

expected count=(total sample size) \* (theoretical probability for that group)

Test Statistic:



Degrees of Freedom: df = k - 1

P-value: The area under the chi-square curve to the right of X2.

**Chi Square Test of Independence**

When to Use:

We have a contingency table with two variables.

We want to see if two variables affect each other.

Conditions: All the expected cell counts are at least 5.

Null Hypothesis: H0 : the two variables are independent **or**

H0 : there is no relationship between the two variables

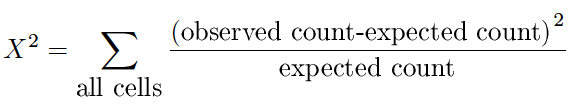
Alt Hypothesis: Ha : the two variables are dependent **or**

Ha : there is a relationship between the two variables

Expected Count:

expected cell count=(row total\*column total)/total sample size

Test Statistic:



Degrees of Freedom: df = (r - 1) (c - 1)

P-value: The area under the chi-square curve to the right of X2.

**T Test for Slope**

When to Use: We want to know the value of the population slope.

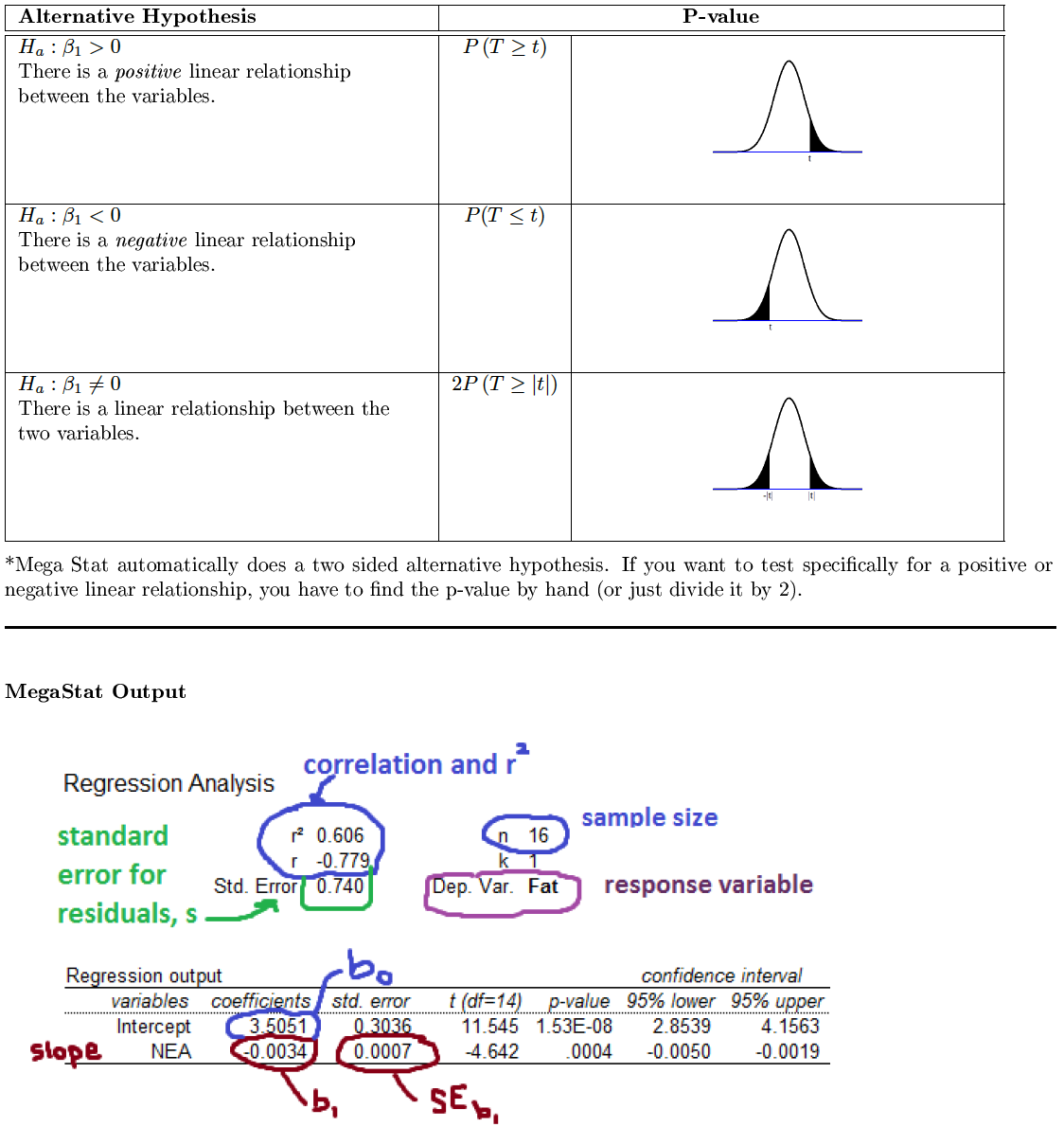
conditions: Just check the three plots to see if linear regression is appropriate.

Null Hypothesis: H0 :  = 0 (There is no linear relationship between the variables.)

Test Statistic:



Degrees of Freedom: n – 2



**Linear Regression**

Different textbooks use different symbols. We use 0 and  for our slope and intercept we find from the sample data.

But my other class uses b0 and b1.

Response Variable (Y): Variable you want to predict

Explanatory Variable (X): Variable you use to explain the response variable

Correlation (r): Measures the strength of the linear relationship between two variables.

Correlation is between -1 and 1.

Correlation close to -1 or 1 means points are close to line.

Correlation close to 0 means the relationship is very weak and the points are not close to the line.

Intercept: If x = 0, then y =...

Slope: If x goes up 1, then y goes up ...

Residual: difference between observed and predicted y value

i = yi 􀀀 ^yi

Extrapolation: This is when we use a regression line for predictions, but the x values are far outside the range of the x values used to obtain the line.

These predictions are often inaccurate. We can't guarantee that the relationship between x and y remains the same for x values outside our original data range.

Association vs Causation: We can use our regression line to use x to predict y.

But we **cannot show** that a change in an explanatory variable

causes a change in the dependent variable.

