

COMMERCE MENTORSHIP PROGRAM

# FINAL REVIEW SESSION

## COMM 191



**PREPARED BY**  
Rebekah Redlich

# TABLE OF CONTENTS



T-Testing

Comparison Testing

Power

Chi Squared Test

Linear Regression

Multiple Regression



@ubccmp



@ubccmp



[cmp.cus.ca](http://cmp.cus.ca)

# 0) Hypothesis Testing **Review**

All hypothesis tests have four components:

- 1.) Null (assumed true, default) and alternative hypotheses
- 2.) Test statistic, previously seen in the form:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

- 3.) P-value (compatibility with  $H_0$  vs  $H_a$ , the smaller the value the LESS likely the null is true).
- 4.) Conclusion Statement



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

# 1) Testing for a Mean

From a sampling distribution of  $\bar{x}$  we get the following test statistic:

Sample mean

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Population Mean

Gives us an approximation with the SAMPLE standard deviation. However, this follows the T-DISTRIBUTION not the normal.

Confidence Interval:  $(100)(1-\alpha)\%$  :  $\mu: \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$

Reminder:

Confidence Interval	$z^*$ (Critical Value)
90%	1.645
95%	1.96
99%	2.576



@ubccmp



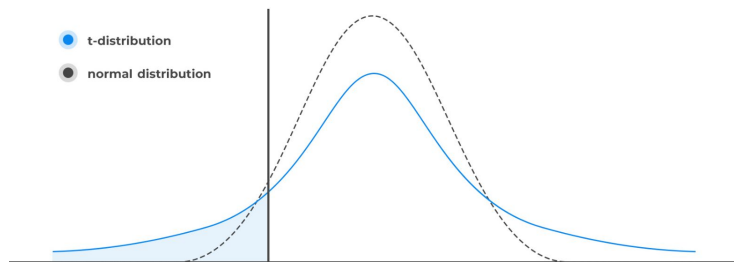
@ubccmp



cmp.cus.ca

# 1) The T-Distribution

The difference between z and t is greatest for small n because this is where s and sigma are furthest.



Do I need to know why d.o.f. Is n-1 this is to do well on the exam? No. Do I need to know when to use n-1 v n-2 to do well? I *would recommend it yes.*

So.... our **excel** formulas go from NORM to T: T.INV, T.DIST.2T, T.DIST.RT !

$$\bar{x} \pm t_{n-1}^* \frac{s}{\sqrt{n}}$$

With n-1 “degrees of freedom.”

As with z-tests, reject  $H_0$  if  $|t\text{-stat}| > t^*$



@ubccmp

@ubccmp

<http://cmp.cus.ca>

## 2) Comparing 2 Independent Means

Unpooled:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (\bar{x}_1 - \bar{x}_2) \pm t^*_{\min(n_1-1, n_2-1)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

**Hypotheses:**

$$H_0: \mu_1 - \mu_2 = \Delta_0 \quad H_A: \mu_1 - \mu_2 \neq \Delta_0$$

**Degrees of Freedom:**

1) df = minimum of  $n_1 - 1$ ,  $n_2 - 1$

2) Sum of  $n_1 - 1$ ,  $n_2 - 1$  (pooled t-testing)

Pooled:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$
$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{(n_1 + n_2 - 2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

**Pooling:**

Assume Population variances are equal. Yields an EXACT confidence interval and t-distribution with  $n_1 + n_2 - 2$  degrees of freedom.

When can I use pooled variances? *If larger s is not much more than 2x the smaller s.*



@ubccmp

@ubccmp

cmp.cus.ca

## 2) Comparing 2 Dependent Means

**Dependent:** Connection/matching of two measurements.

These measurements are PAIRED, link in some way.

$$H_0: \mu_d = \Delta_0 \quad H_A: \mu_d \neq \Delta_0 \quad d_i = x_{1i} - x_{2i}$$

$$t = \frac{\bar{d} - \Delta_0}{\frac{s_d}{\sqrt{n}}}$$

$$\mu_d: \bar{d} \pm t_{n-1}^* \frac{s_d}{\sqrt{n}}$$


Dependence refers to a connection or linkage or **matching** of two measurements.

EX: Before and After.



 @ubccmp

 @ubccmp

 <http://cmp.cus.ca>

## 2) Comparing 2 Independent Proportions

$$\text{Mean } (\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

$$\text{SD}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

$$\text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

$$\begin{aligned} H_0: p_1 - p_2 &= \Delta_0 \\ H_A: p_1 - p_2 &\neq \Delta_0 \end{aligned} \quad Z = \frac{(\hat{p}_1 - \hat{p}_2) - \Delta_0}{\sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}}$$

For testing whether two proportions are equal you can compute a pooled  $p$ :

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2} \quad Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p} \hat{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$





# Practice - Independent Means

Loafe wants to compare baked goods sold / hr at two locations. Two random samples of /hr purchases are taken at each and the sales/hr is recorded.

## Sauder

Mean: 120    SD: 30    Size: 25

## Alumni

Mean: 98    SD: 14    Size: 18



@ubccmp



@ubccmp



cmp.cus.ca

# Practice - Dependent Means

Aritzia wants to test 2 super puff ads. 20 stores of various sizes are chosen. Each store gets Ad 1 for one month and Ad 2 for one month. Sales for each month are recorded:

Mean Difference: 4.23      Sample STDEV: 1.99      Sample: 20



@ubccmp



@ubccmp



cmp.cus.ca

# Practice - Independent Proportions

A survey is done in 2 fashion companies to determine what proportion of employees have fashion degrees. A sample of 300 at Araz Co. shows 166 have degrees compared with 72 of 210 at M&H.



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

# Decisions, Power and Significance

If  $H_0$  is TRUE, conclusion is Do Not Reject  $\rightarrow$  Correct

If  $H_0$  is FALSE, conclusion is Reject  $\rightarrow$  Correct

Power =  $1 - B$  =  
Probability of  
Correct Rejection

	Decision	
TRUE STATE	Do Not Reject Null	Reject Null
Null True	Correct	Type I Error
Null False	Type II Error	Correct

Type I = False Positive (Saying significance without any)  $Pr = \text{Alpha}$

Type II = False Negative (Saying insignificance when it IS)  $Pr = \text{Beta}$



# More on Power

Power =  $1 - B$  = Probability of Correct Rejection

- The p-value is **NOT** the probability that  $H_0$  is correct.

When one is very small, the other is very large, so we compromise by trying to make them both small:

$\alpha < .05$ , the lower the better  $\beta < .20$ , which means power  $> 80\%$ .

## Power is Affected By:

- 1) Size of observed difference (t-stat)- bigger difference = more power
- 2) Size of variation - bigger = less power
- 3) Size of samples - bigger = more power



@ubccmp



@ubccmp



cmp.cus.ca

# Practice

True or False:

- 1) A type II error can't be made if the null is rejected.
- 2) The smaller the level of significance, the the less likely you are to reject the null hypothesis.
- 3) Power will decrease when we consider an alternative further from the null value.



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

# Chi Squared Tests

Examining discrepancies between observed and expected frequencies of categorical variables. Are the **discrepancies** statistically significant?

- An extension of the two sample z-test to more than 2 categories and samples.

$$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$$

- One sided ONLY therefore P-Value =  $\Pr(\chi^2 > \chi^2 \text{ stat})$
- Hypothesis: X and Y are or are not independent.

**CHISQ.DIST.RT (test stat, df) | CHISQ.INV.RT**



@ubccmp



@ubccmp



cmp.cus.ca

# Chi Squared Tests

$$\text{Expected Counts} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Overall Total}}$$

Critical value can be determined using CHI.INV, alpha and df or Table-X.

Test Statistic:  $\chi^2 = \sum \frac{(Obs-Exp)^2}{Exp}$  With (# of rows - 1) x (# columns - 1) degrees of freedom.

Assumptions:

- 1) Data are counts
- 2) Randomly Sampled
- 3) Expected count > 5.

The best part? No confidence interval!

Standardized Residual:  $\frac{Obs-Exp}{\sqrt{Exp}}$



@ubccmp



@ubccmp



cmp.cus.ca



# Practice

Find the value of the test statistic, degrees of freedom, critical value and p-value for the following:

	Taken	Single	Total
First Year	75	225	300
Second Year	70	230	300
Third Year	299	101	400
Total	444	556	1000

# Statistical Models

Representation of a phenomenon with some sort of randomness, variability or error.

$$\text{Outcome} = \text{Systematic} + \text{Random}$$

★ Grand Unified Theory of Statistics - Identifies relationships between outcome and predictor variables based on data type.



@ubccmp



@ubccmp



cmp.cus.ca

# Inference for Linear Regression

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

## Assumptions:

- Unbiased, average error = 0,  $Y_i$  on average is  $B_0 + B_1 X_i$
- Constant variance (error and  $Y_i$ )
- Must be independent (error and  $Y_i$ )
- Must be normally distributed (error and  $Y_i$ )

## Interpretations:

Regression line passes through the average value of  $Y$  for each  $X$ .

Point 1 SD of  $X$  above  $X$  mean is on average  $r$  SD of  $Y$  above the mean of  $Y$ .

# Inference for Linear Regression

**Standard Deviation of Residuals:** The typical vertical distance from a data point to the estimated regression line.

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}}$$

$$b_1 = r \frac{s_y}{s_x} \quad b_0 = \bar{y} - b_1 \bar{x}$$

$B_1$  has normal sampling distribution with mean  $B_1$  and SE  $\frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}}$

$B_0$  has normal sampling distribution with mean  $B_1$  and SE  $s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$



# Confidence Interval for Linear Regression

$$\beta_1 : b_1 \pm t_{n-2}^* \frac{s_e}{s_x \sqrt{n-1}}$$

$$\beta_0 : b_0 \pm t_{n-2}^* s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}}$$

For intercept is this relevant? Not really. For slope, asses deviation from horizontal or 0.

# Confidence ( $u_x^*$ ) and Prediction ( $y_x^*$ )

Confidence for the mean value of  $y$  at a specific value of  $x$ :

$$(b_0 + b_1 x^*) \pm t_{n-2}^* \sqrt{SE^2(b_1) \times (x^* - \bar{x})^2 + \frac{s_e^2}{n}}$$

Prediction for the single value of  $y$  at a particular  $x$ :

$$(b_0 + b_1 x^*) \pm t_{n-2}^* \sqrt{SE^2(b_1) \times (x^* - \bar{x})^2 + \frac{s_e^2}{n} + s_e^2}$$

Confidence is narrower than prediction because it is easier to estimate the mean than a precise observation

**Extrapolation Penalty:**

The further  $x^*$  gets from the mean the worse the precision

# Practice

Suppose a linear regression is used to predict the BMI of athletes based on distance ran per week. The predicted BMI for an athlete who runs 100 km per week is 20. The 95% confidence interval is (18, 22) and the prediction interval is (16, 24).

We are 95% confident that athletes who run 100 km a week \_\_\_\_\_ land between \_\_ and \_\_ .

We are 95% confident that a \_\_\_\_\_ athlete who runs 100 km a week lands between \_\_ and \_\_ .



@ubccmp



@ubccmp



cmp.cus.ca

# Analysis of Variance For Regression (ANOVA)

Is there another, more generalizable way to test whether  $\beta_1 = 0$ ?

Is the X-variable a useful predictor of Y?

Is the model worthwhile?

How can the sources of variation in the response variable be summarized?



@ubccmp



@ubccmp



cmp.cus.ca



# Variance Explained

Sum of Squares Total = Sum of Squares Model + Sum of Squares Error

$$SST = SSM \text{ (regression)} + SSE \text{ (residual)}$$

**Total Variation**  $= \sum (y_i - \bar{y})^2$

**Explained Variation**  $= \sum (\hat{y}_i - \bar{y})^2$

**Unexplained Variation**  $= \sum (y_i - \hat{y}_i)^2$

Sum of Squares Total (SST)  $= (n - 1)s_y^2$

Sum of Squares Model (SSM)  $= SST - SSE$

Sum of Squares Error (SSE)  $= (n - 2)s_e^2$



# ANOVA Testing

Variation	Sum of Squares	DoF	Mean Square	F-Stat
Model	SSM	1	MSM	MSM/MSE
Error	SSE	n-2	MSE	
Total	SST	n-1		

$$MSE = SSE / (n-2) = s_e^2$$

F.DIST.RT (f-stat, df1, df2)

$$R^2 = 1 - SSE / SST = SSM / SST$$



@ubccmp

@ubccmp

<http://cmp.cus.ca>

# Multiple Regression

K = # of variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

$$s_e^2 = \frac{\sum e_i^2}{n-k-1}$$

One quantitative outcome with **MANY** predictors, quantitative and categorical. What set of independent variables provides a “good” explanation of the variation in Y?

Assumptions of linearity, homoscedasticity, normality and independence all hold.

$H_a$  : At least one  $B_1$  is not 0. There is SOME value.

Test Statistic :  $F = \text{MSM} / \text{MSE}$

F.DIST.RT (F-stat, df1, df2)



@ubccmp



@ubccmp



cmp.cus.ca

# ANOVA - Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F-stat
Model/Regression	SSM	k	$MSM = SSM/k$	$MSM/MSE$
Error/Residual	SSE	n-k-1	$MSE = SSE/(n-k-1)$	
Total	SST	n-1		

# Multiple Regression

**Principle of Parsimony:** Use the simplest model that still provides an adequate fit. Don't overfit the model.

**Dangers of Multicollinearity:** You can't maximize  $R^2$ .

**F Tests:** Reduced models uses subsets of the full model.

$H_0$ : Reduced model is adequate  $H_A$ : Reduced model is not adequate.

$$F = \left( \frac{n - k - 1}{q} \right) \left( \frac{R^2(\text{full}) - R^2(\text{reduced})}{1 - R^2(\text{full})} \right) \quad \text{Adjusted } R^2 = 1 - (1 - R^2) \left( \frac{n-1}{n-k-1} \right)$$

**DF1 = q (variables dropped) DF2 = n-k-1**



@ubccmp



@ubccmp



cmp.cus.ca

# Practice

In a multiple regression analysis with 36 data points,  $s_e^2 = 1.6$  and  $SSE = 48$ . How many variables are there?



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

# Practice

In a multiple regression analysis with 2 independent variables,  $SST=1020$ ,  $MSM = 300$ ,  $F\text{-Stat} = 20$ . What is the sample size  $n$ ?



@ubccmp



@ubccmp



cmp.cus.ca

# Practice

A stats prof hired a survey firm to randomly survey 30 students and determine which factors influence final exam performance. The survey provided the following variables: FINAL EXAM SCORE, HOURS OF SLEEP, MIDTERM SCORE, HOURS STUDYING / WEEK.

	<i>Df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Regression	3	81972.00	27324.00	
Residual	26	43550.00	1675.00	
Total	29	125522.00		

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	840.50	225.15	3.733	.0009
HOURS OF SLEEP	-168.00	96.86	-1.734	.0947
MIDTERM SCORE	5.66	0.85	6.659	.0000
HOURS STUDYING	-1.25	1.18	-1.059	.2992

Identify the null and alternative hypotheses.

What is the value of the test statistic for the overall model?

What is the  $s_e$  value?

What proportion of variation is explained by all 3 variables?

What is the adjusted  $R^2$ ?



@ubccmp



@ubccmp



<http://cmp.cus.ca>



# Practice Ctd

An F test for hours of sleep and hours studying is performed. The F-Stat is 1.644.

What are the degrees of freedom for the test?

What is the p-value?



@ubccmp



@ubccmp



cmp.cus.ca

# Practice

Suppose a multiple linear regression analysis of a model involving 10 independent variables and 190 observations gives an  $R^2$  of 80%. A reduced model is fit with only 5 of the original independent variables retained. The  $R^2 = 62\%$

What is the value of the F-statistic?



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca

# Practice - Guess that Test !

- 1) Tim Hortons conducts research on if amount spent depends on minutes spent waiting in line.
- 2) Apple wants to determine if there is a significant relationship between gender and colour of iphone chosen.
- 3) John takes a random sample of 500 swipes on tinder, are at least 20% swiping right?
- 4) A swim instructor wants to know if a new form makes you faster and asks her athletes to swim a lap twice, once with the new form and once with the regular.
- 5) Is the Sauder admission GPA average higher than 90%?



@ubccmp



@ubccmp



[http://](http://cmp.cus.ca) cmp.cus.ca