Commerce Mentorship Program

COMM 291

Final Exam Review Package

Prepared by Angelica Cabrera

1. THE NORMAL CURVE

Problem 1. Final exam grades in a COMM 291 section are normally distributed. Bill's grade was 90%, corresponding to a z-score of 2. Steve's grade was 60%, corresponding to a z-score of -1. What are the mean and standard deviation of the whole section (whole numbers)?

Bill

$2 = \frac{90 - \mu}{\sigma}$

Steve

$-1 = \frac{60 - \mu}{\sigma}$

Set above equations equal to one another and get μ=70 and σ=10.
Remember that a z-score of -1 or 1 lies one standard deviation away from the mean, which explains Bill's score (70-10=60).

Problem 2. The distribution of weights of granola bars is claimed to be normal with a mean of 250g and a standard deviation of 24g. If a random sample of 16 granola bars is taken, what is the probability that the sample mean weight of the 16 bars will be less than 232g?

T= B1 + B2 +…+B16 (Total weight = Weight of Bar 1, etc.)

Mean (T) = 16(250) = 4000

Var (T) = $16(24^2)$=9216

SD (T) = √9216=96

Pr (x<3712) = Pr(z< $\frac{3712 - 4000}{96}$) =Pr (z<-3.00) = 0.13%

Problem 3. Mike challenges Harvey to a round of golf. Record-keeping from previous years show that Mike's scores are normally distributed with mean 110 and standard deviation 10, and that Harvey's scores are normally distributed with mean 100 and standard deviation 8. What is the probability that Mike will beat Harvey?

Let X = Mike's score and Y = Harvey's score.

Mean(X–Y) = 110 – 100 =10

Var(X–Y) = 10^2 +8^2 = 164

SD(X–Y) = √164 = 12.81

Pr(X–Y<0) = Pr(Z < [0–10]/12.81)= Pr(Z < -0.78) = 0.22
Remember that in golf, the low score wins.

2. COMPARING TWO MEANS

A) Two-Sample T-Test

- Answers the question: "Is there a difference in means between two _____ groups"?
- We use the Pooled (Equal) Variances Version
  - _____ df and t-distribution

TWO-SAMPLE T-TEST (Pooled Variances Version)

X is binary, Y is quantitative

1. Hypotheses

Two-sided:	*One-sided:*
H_o: $\mu_1 - \mu_2 = 0$ or Δ_0	H_o: $\mu_1 - \mu_2 > 0$ or Δ_0
H_a: $\mu_1 - \mu_2 \neq 0$ or Δ_0	H_a: $\mu_1 - \mu_2 < 0$ or Δ_0

2. Find pooled variances

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

3. Test statistic

$$\frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

4. df

n_1+n_2-2

5. Reject H_0 if *(use Table T)*

Two-sided:
2 x Pr ($t^*_{n1+n2-2}$) > |t-stat|) < α

One-sided:
Pr ($t^*_{n1+n2-2}$ > |t-stat|) < α

100(1 – α)% CONFIDENCE INTERVAL

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{(n1+n2-2)}\, s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

ASSUMPTIONS

1. Independent groups and data
2. $\sigma_1^2=\sigma_2^2$ if _____

Problem 4a. Car appraisers examine vehicles that have been involved in accidents to assess the cost of repairs. An experiment was done to examine whether different appraisers produce significant different assessments of minor collisions. The experiment was designed so that 20 cars were used, 10 were shown to Appraiser 1 and 10 to Appraiser 2.

Car	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Appraiser 1	1650	360	660	1050	890	750	470	1270	550	730
Car	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Appraiser 2	1440	380	600	920	930	650	410	1080	480	770

OUTPUT 1

t-Test: Two-Sample Assuming Unequal Variances

	Appraiser 1	*Appraiser 2*
Mean	838	762
Variance	154618	105507
Observations	10	10
Hypoth. Mean Diff.	0	
df	17	
t Stat	0.471	
P(T<=t) one-tail	0.322	
t Critical one-tail	1.740	
P(T<=t) two-tail	0.643	
t Critical two-tail	2.110	

i. What is your conclusion?

ii. Construct a 95% CI for the difference in mean appraisals between the two appraisers (whole numbers). Assume that the pooled variance is 130321. Does it support your answer in part i?

A) Paired T-Test

- Answers the question: "Is there a difference in means between two _____ groups?"
- connection between two samples → usually two measurements on one subject

PAIRED T-TEST
X is binary, Y is quantitative

1. Hypotheses

Two-sided: H_o: $\mu_d = 0$ or Δ_0 ; H_a: $\mu_d \neq 0$ or Δ_0

One-sided: H_o: $\mu_d > 0$ or Δ_0 ; H_a: $\mu_d < 0$ or Δ_0

2. Find sample mean of differences	3. Test statistic	4. df	5. Reject H_0 if *(use Table T)*
$\bar{d}$	$\frac{\bar{d} - \Delta_0}{s_d / \sqrt{n}}$	n-1	*Two-sided:* 2 x Pr (t^*_{n-1} > |t-stat|) < α *One-sided:* Pr (t^*_{n-1} > |t-stat|) < α

100(1 – α)% CONFIDENCE INTERVAL

$\bar{d} \pm t^*_{n-1} s_d / \sqrt{n}$

ASSUMPTIONS

1. $n_1 = n_2$
1. Data is matched (not independent)
2. Individuals are independent
3. Population of differences is Normal

Problem 4b. Consider the experiment from Problem 4a. Suppose that it was instead designed so that a total of 10 cars were used and each car was assessed by each appraiser.

Car	1	2	3	4	5	6	7	8	9	10
Appraiser 1	1650	360	660	1050	890	750	470	1270	550	730
Appraiser 2	1440	380	600	920	930	650	410	1080	480	770

OUTPUT 2

t-Test: Paired Two Sample for Means

	Appraiser 1	*Appraiser 2*
Mean	838	762
Variance	154618	105507
Observations	10	10
Hypoth. Mean Diff.	0	
df	9	
t Stat	2.496	
P(T<=t) one-tail	0.017	
t Critical one-tail	1.833	
P(T<=t) two-tail	0.034	
t Critical two-tail	2.262	

i. What are the null and alternative hypotheses?

ii. What is the appropriate P-value?

iii. What is your conclusion at the 5% significance level.?

iv. What would be a Type I error in this situation?

6. COMPARING TWO PROPORTIONS

A) Two-Sample z-test

- Answers the question: "Is there a difference in proportions between two independent groups?"

TWO-SAMPLE Z-TEST
X is binary, Y is binary

1. Hypotheses

Two-sided:	One-sided:
H_o: $p_1 - p_2 = \Delta_0$	H_o: $p_1 - p_2 > \Delta_0$
H_a: $p_1 - p_2 \neq \Delta_0$	H_a: $p_1 - p_2 < \Delta_0$

2. Test statistic

$$\frac{(\hat{p}_1 - \hat{p}_2) - D_0}{\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

3. Critical value

Z	CL
1.645	90%
1.96	95%
2.576	99%

4. Reject H_0 if *(use Table Z)*

Two-sided:
2 x Pr (z*> |z-stat|) < α

One-sided:
Pr (z* > |z-stat|) < α

100(1 – α)% CONFIDENCE INTERVAL
$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

ASSUMPTIONS
1. $n_1\hat{p}_1$, $n_1\hat{q}_1$, $n_2\hat{p}_2$, and $n_2\hat{q}_2$ to be greater than 10. 2. Populations are independent

B) Chi-Square Goodness of Fit Test

- Answers the question: "Is there a difference between observed and expected frequencies for _____ categorical variable?"
- "goodness of fit" = how well hypothesized proportions fit the observed proportions
- If this test has one df, equivalent to one-sample z-test for proportion

CHI-SQUARE GOODNESS OF FIT TEST			
X is binary, Y is binary			
1. Hypotheses	**2. Test statistic**	**3. df**	**4. Reject H_0 if** *(use Table F)*
H_o: $p_1=p_2=\ldots=p_n=X$ H_a: at least one p is not X	$\chi^2 = \sum \frac{(Obs - Exp)^2}{Exp}$	#cells-1	Pr ($\chi^2_{\#cells-1} > \chi^2$ -stat) < α

ASSUMPTIONS
1. Data are in counts 2. Data in sample are independent 3. Populations are Normal

C) Chi-Square Test of Independence

- Answers the question: "Are 2+ categorical variables independent?"
- Two-sample z-test = 2 x 2 tables, chi-square test of independence = R x C tables
  - A 2x2 table can be tested using either
- Works best when each Eij is 5
- Perfectly independent when $\chi^2 = 0$ or ratio of column %s is the same in all columns

CHI-SQUARE GOODNESS OF FIT TEST			
X categorical (2+), Y is categorical (2+)			
1. Hypotheses	**2. Test statistic**	**3. df**	**4. Reject H_0 if** *(use Table F)*
H_o: Row and column variables are independent H_a: Row and column variables are not independent	$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$ *where* O_{ij} *= given* E_{ij} *= (i^{th} Row Total) x (j^{th} Column Total) / Overall Total*	(#rows–1) x (#columns–1)	Pr ($\chi^2_{df} > \chi^2$ -stat) < α

ASSUMPTIONS
1. Data are in counts 2. Data in sample are independent 3. Sample is sufficiently large

Problem 5. Buy More, an electronics retail chain, wants to determine if coupon redemption for purchases on its online website is independent of gender.

		Coupon Redeemed?		
		Yes	**No**	
Gender	**Male**	65	65	
	Female	125	75	

a) Calculate the appropriate row or column percentage.
   i. What is the probability that a coupon was redeemed given that the online shopper was female?

   ii. What percentage of all online purchases have coupon redemptions?

b) What are the null and alternative hypotheses?

c) Calculate the expected frequencies for the above cells (2 decimal places) and put them in brackets beside the observed values.

d) How much does the top right cell contribute to the χ^2 test statistic (2 decimal places)?

e) The χ^2 test statistic is 5.04. At the 5% significance level, what is your conclusion?

f) Calculate the 95% CI for the difference in proportions of males and females who redeem coupons online. Does this support your conclusion in part d)?

7. SIMPLE LINEAR REGRESSION

- What do we use to assess the relationship between:
  - Categorical variables? _____
  - Quantitative variables? _____

A) The Simple Linear Regression Model

- For two quantitative variables: one _____ and one _____
- Previously: least-squares regression → $\hat{y} = b_0 + b_1x$

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

Response = Predictor + Random unexplained errors
Result = Systematic Component + Random Component

β_0 = true intercept
β_1 = true slope
Y = Unknown value of the dependent variable
X = known value of the independent variable
ε = "random error" term

- Assumptions → from now on until One-way ANOVA! (Simple Regression – Multiple Regression)
  1. Linearity: relationship between x and y is linear
  2. Constant Variance (Equal Spread): variability of errors is constant
  3. Normality: At any given X, Y's or errors follow a normal model
  4. Independence: Y's or errors are independent

B) Test of Slope

- Answers the question: "Is the slope different from zero?"

TEST OF SLOPE			
X is quantitative, Y is quantitative			
1. Hypothesis	**2. Test statistic**	**3. df**	**4. Reject H_0 if** *(use Table T)*
H_o: $\beta_1 = 0$ H_a: $\beta_1 \neq 0$	$t = \frac{b_1}{SE(b_1)}$	n-2	Pr (t^*_{n-2}>|t-stat|) < α

100(1 – α)% CONFIDENCE INTERVAL

$b_1 \pm t^*_{n-2}$ **SE(**b_1)

C) Test of Correlation

- Answers the question: "Is the correlation different from zero?"

TEST OF CORRELATION			
X is quantitative, Y is quantitative			
1. Hypothesis	**2. Test statistic**	**3. df**	**4. Reject H_0 if** *(use Table T)*
H_0: ρ =0 H_a: ρ ≠ 0	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	n-2	Pr (t^*_{n-2}>|t-stat|) < α

D) Predicting Y

	Confidence Interval for a Mean Response	**Prediction Interval for a Future Observation**
Answers the question:	What is the mean Y ($\hat{\mu}x$) for all units having a particular x-value (x*)?	What is the single value of Y ($\hat{y}$x*) for a single unit having a particular x-value (x*)?
Point Estimate	$\hat{\mu}x^* = b_0 + b_1x^*$	$\hat{y}x^* = b_0 + b_1x^*$
SE	SE($\hat{\mu}x^*$) = $\sqrt{SE^2(b_1) \cdot (x^* - \bar{x})^2 + \frac{{S_e}^2}{n}}$	SE($\hat{y}$x*)=$\sqrt{SE^2(b_1) \cdot (x^* - \bar{x})^2 + \frac{{S_e}^2}{n} + {S_e}^2}$
CI or PI	$\hat{\mu}x^* \pm t^*_{n-2}$ SE($\hat{\mu}x^*$)	$\hat{y}x^* \pm t^*_{n-2}$ SE($\hat{y}$x*)
Approximate PI	n/a	$\hat{y}x^*$±2 x S_e If: 1. n is large 2. ExtrapolatIon Penalty is small $\frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}$

- PI is _____ than CI because it is _____ to estimate precisely a single observation than the mean

E) F-Test from ANOVA Table

- Answers the question: "Is x a helpful predictor of Y?"
- Why are Y-values different from mean value of Y?
  Total variation = Explained variation + Unexplained variation
  SST = SSM + SSE
- ANOVA table

Source of Variation	**Sum of Squares**	**df**	**Mean Square (MS)**	**F-stat**
Model	SSM	1	MSM	$\frac{MSM}{MSE}$
Error	SSE	n-2	MSE	
Total	SST	n-1		

S_e	**Coefficient of Determination**
$\sqrt{MSE}$	$R^2 = \frac{SSM}{SST}$

F-TEST FOR SIMPLE REGRESSION			
X is quantitative, Y is quantitative			
1. Hypothesis	**2. Test statistic**	**3. df**	**4. Reject H_0 if** *(use Table F)*
H_o: $\beta_1 = 0$ H_a: $\beta_1 \neq 0$	$F = \frac{MSM}{MSE}$	1,n-2	Pr ($F^*_{1,n-2}$ > F-stat) < α or F-stat > $F^*_{1,n-2}$

- F-test gives same _____ as Test for Slope in Simple Regression
  - F-stat = _____

Problem 6a. Gimbels, a retail department store, has incurred high losses due to shoplifting. Data were collected for a random sample of 17 months from the past 10 years.

Y = SHOPLIFT, monthly dollar loss due to shoplifting
X2 = SALETRAN, number of sales transactions

OUTPUT 1. Simple Linear Regression (SALETRAN)

ANOVA

	Df	*SS*	*MS*	*F*	*Sig. F*
Regression	1	1270172	1270172	9.9105	.0066
Residual	15	1922459	128164		
Total	16	3192632			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2273.088	338.70	6.711	0.0000	1551.173	2995.003
SALETRAN	0.0809	0.0254				

a) Carry out a hypothesis test to determine if the model is worthwhile. Use a 5% significance level.

b) Compute an EXACT 95% CI for the slope of the regression line (2 decimal places).

c) What percent of total variation in SHOPLIFT is explained by SALETRAN (whole number)?

d) Interpret the coefficient of SALETRAN.

e) Fill in the missing values in the bottom row of the table.

(1)

(2)

(3)

(4)

f) What is the standard deviation of the residuals?

g) What is the approximate 95% prediction interval for monthly dollar loss in a month with 1000 transactions?

MULTIPLE REGRESSION

A) The Multiple Regression Model

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

- Signs of coefficients shows how they affect Y "given" other predictors
  - Ex. Average Selling Price of Home = $20{,}000 - 7{,}500_{\#Bedrooms} + 95_{SquareFeet}$

B) F-Test from ANOVA Table

- Answers the question: "Are the X's a helpful predictor of Y?"

Source of Variation	Sum of Squares	df	Mean Square (MS)	F-stat
Model	SSM	k	MSM	$\frac{\text{MSM}}{\text{MSE}}$
Error	SSE	n-k-1	MSE	
Total	SST	n-1		

S_e	Coefficient of Multiple Determination
$\sqrt{MSE}$	$R^2 = \frac{SSM}{SST}$

F-TEST FOR MULTIPLE REGRESSION USING ANOVA			
X is quantitative/categorical (2+), Y is quantitative			
1. Hypothesis	**2. Test statistic**	**3. df**	**4. Reject H_0 if** *(use Table F)*
H_o: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ H_a: at least one β_j is not 0	$F = \frac{MSM}{MSE}$	k,n-k-1 *Where* *k = # x-variables* *n = sample size*	$Pr(F^*_{k,n-k-1} > \text{F-stat}) < \alpha$ or $\text{F-stat} > F^*_{k,n-k-1}$

C) Principle of Parsimony: find the _____ set of predictors that adequately fit the data

- Multicollinearity: "predictor variable redundancy"
  - When two x-variables have a very high _____, so much of information in X_1 is "contained" in X_2.
  - This does not create a good model because each x-variable does not add unique predictive information given the variables already in the model
  - If use regular R^2, adding more x-vars will _____, but this does not mean it is a better model
- Use adjusted $R^2 = 1 - (1 - R^2)(\frac{n-1}{n-k-1})$
  - Imposes a "penalty" for each new x-variable added.
  - Increases and decreases depending on whether a predictor is contributing

D) F-Test to Compare Full and Reduced Models

- Answers the question: "What x-variables make the model worthwhile?"
- Full model: uses all available x-variables
  Reduced model: drops a # of x-variables
- Two ways to drop:
  1. Remove _____ x-variable at a time based on _____ until all x-variables left are significant
  2. Remove _____ x-variables at a time, using _____ to compare Full to Reduced Model

F-TEST TO COMPARE FULL AND REDUCED MODELS
X is quantitative/categorical (2+), Y is quantitative

1. Hypothesis

H_o: The reduced model is adequate
H_a: The reduced model is not adequate

2. Test statistic

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

3. df

q,n-k-1

4. P-value *(use Table F)*

Pr ($F^*_{q,n-k-1}$ > F-stat)

or

F-stat > $F^*_{q,n-k-1}$

Where
k = number of X-variables in the Full Model
q = number of variables dropped from the Full Model to get the Reduced Model

Problem 6b. Consider Gimbels, the same department store from Problem 6a. However, it is now a multiple regression model with three more independent variables:

X1 = TEMP, average monthly temperature
X3 = HOLIDAY, dummy variable for one holiday month
X4 = EMPLOYEE, number of employees on store's monthly payroll

OUTPUT 2. Multiple Regression (Four predictors)

Regression Statistics	
Multiple R	0.8106
R Square	0.6571
Adj. R Square	0.5428
Standard Error	302.0344
Observations	17

ANOVA

	Df	*SS*	*MS*	*F*	*Sig. F*
Regression	4	2097934	524483	5.749	0.0080
Residual	12	1094698	91225		
Total	16	3192632			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4875.365	1543.275	3.159	0.0082
TEMP	6.263	5.963	1.050	0.3142
SALETRAN	0.239	0.073	3.275	0.0066
HOLIDAY	-190.259	202.545	-0.939	0.3661
EMPLOYEE	-27.322	11.974	-2.282	0.0415

a) What are the null and alternative hypotheses?

b) At a 5% significance level, can we say that the regression equation is significant?

c) What is the estimated difference in losses between two months that are identical in all respects, except one month has 10 more employees?

d) At a 5% significance level, test whether HOLIDAY can be dropped from the model.

What would happen to R^2 as a result?

e) The R^2 of a model with only SALETRAN and EMPLOYEE as predictors is 0.5742. At α = 0.05, test whether HOLIDAY and TEMP can both be dropped from the model.

A) COMPARING TWO OR MORE MEANS

A) One-Way ANOVA

- Answers the question: "Is there a difference among means from 2+ groups that are classified according to one categorical variable?"
- Compares variation across the means to variation between the samples
- ANOVA Table

Source of Variation	Sum of Squares	df	Mean Square (MS)	F-stat
Factor	SSFactor	k-1	MSFactor	$\frac{\text{MSFactor}}{\text{MSE}}$
Error	SSE	N-k	MSE	
Total	SST	N-1		

S_e	Coefficient of Determination
$\sqrt{MSE}$	$R^2 = \frac{SSFactor}{SST}$

ONE-WAY ANOVA TEST
X is categorical, Y is quantitative

1. Hypothesis	2. Test statistic	3. df	4. Reject H_0 if *(use Table F)*
H_0: $\mu_1 = \mu_2 = \cdots = \mu_k$ H_a: at least one μ_i is different from the others	$F = \frac{MSFactor}{MSE}$	k-1, N-k *Where* *k = # groups* *n = total sample size*	Pr ($F^*_{k-1,N-k}$ > F-stat) or F-stat > $F^*_{k-1,N-k}$

ASSUMPTIONS

1. Normality
2. Constant Variance
3. Independence

Problem 7. Pizza Planet, a pizza chain, randomly selected restaurants in the North, South, East, and West regions of the United States. The mean quarterly sales (in $1000s) for each sample were recorded. Pizza Planet's CEO is curious if there is a significant difference in the mean sales between the regions.

Region	Sample Size	Sample Mean	Sample SD
North	20	100	1.20
South	17	99	1.30
East	21	105	1.10
West	20	102	1.40

a) What are the null and alternative hypotheses?

b) Fill in the blanks in the following ANOVA table.

Source of Variation	Sum of Squares	df	Mean Square (MS)	F-stat
Factor			137.75	
Error				
Total	527.13			

c) Compute the overall standard deviation of the total sample (2 decimal places)?

d) The P-value for this statistic turns out to be <0.001. At α=0.05, what is your conclusion?