

Practice Final 2 – Math 17/ENST 24

Name: *Solutions*

Math 17/ Enst 24 – Introduction to Statistics

Final Exam

PRACTICE 2

Instructions:

1. Show all work. You may receive partial credit for partially completed problems.
2. You may use calculators and a two-sided sheet of reference notes, as well as the provided tables. You may not use any other references or any texts.
3. You may not discuss the exam with anyone but me.
4. Suggestion: Read all questions before beginning and complete the ones you know best first. Point values per problem (separate page for each) are displayed below if that helps you allocate your time among problems. (*would be done for actual exam*)
5. Good luck!

(Some data taken from Utts/Heckard, #1 used with permission UofM)

## Practice Final 2 – Math 17/ENST 24

1. Manatees are large, gentle sea creatures that live along the Florida coast (among other places). Many manatees are killed or injured by powerboats. Data on the number of manatees killed by powerboats and the number of registered powerboats (in thousands of boats) in Florida for the period 1977 to 1990 were analyzed by an introductory statistics student. Selected R output is given below:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.265	6.361	.199	.846
boats	.05004	.011	4.527	.001

Residual standard error: 17.51 on 12 degrees of freedom

Multiple R-squared: 0.631, Adjusted R-squared: 0.600

F-statistic: 20.49 on 1 and 12 DF, p-value: .001

a. Give the equation of the least squares line for using number of registered powerboats (in thousands) to predict the number of manatees killed.


$$\hat{y}_{\text{manatees}} = 1.265 + .05004x \rightarrow \# \text{ registered boats (1000s)}$$

b. Based on the analysis, about 63.1 % of the variation in the number of manatees killed can be explained by the linear relationship with the number of registered powerboats.

c. With each increase of 100,000 powerboat registrations, we would estimate the mean number of manatees killed to increase by about 5.004. (watch units  $\Rightarrow$  100,000  $\Rightarrow$  100 in 1000s)

d. Is there strong evidence that the mean number of manatees killed increases as the number of powerboat registrations increases? Give the appropriate null and alternative hypotheses and the p-value. Assume the necessary assumptions hold.

$$H_0: \beta_1 = 0 \quad H_A: \beta_1 > 0$$


 $\Rightarrow p\text{-value} = .0005$

Are the results significant at the .05 level?

(Yes)

No

(b/c we reject  $H_0$ )

e. For the test performed in d., there are several assumptions that must hold. Name two plots that could be used to assess the validity of the assumptions and which assumptions they are used to assess.

(3 possible plots)

1. Scatterplot of X vs. Y - checks for a linear relationship
2. QQ plot of residuals - check for normally distributed error terms
3. Residual plot - checks for constant variance of error terms (also, linear relationship)

Practice Final 2 – Math 17/ENST 24

2. Until recently, M&M's used to publish the percentages of each color of M&M that were in the common M&M mix (non-holiday). A sample of 4 one-pound M&M bags were obtained and the number of each color counted in order to compare to the most recent published percentages claimed by the company. The data are summarized below.

Color	Brown	Red	Yellow	Blue	Orange	Green	Total
Observed	602	396	379	227	242	235	2081
Alleged (%)	.3	.2	.2	.1	.1	.1	1
Expected	624.3	416.2	416.2	208.1	208.1	208.1	2081

a. What is the appropriate inference procedure to test if the sample matches the alleged distribution of color provided by the company? Provide appropriate hypotheses.

Procedure:  $\chi^2$  Goodness of Fit

Null Hypothesis:  $H_0: p_1 = .3, p_2 = .2, p_3 = .2, p_4 = p_5 = p_6 = .1$

Alternative Hypothesis:  $H_A: \text{Not } H_0$

b. Compute expected counts for each color and fill in the table above. Are the necessary conditions met to perform your hypothesis test?  $np_i, n = 2081, p_i \text{ in } H_0$

**Yes** No, because all expected counts are  $\geq 5$ .  
(also, this sample seems representative of bags of M&M's).

c. Compute the appropriate test statistic to test your hypotheses in a.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(602-624.3)^2}{624.3} + \dots + \frac{(235-208.1)^2}{208.1} = 15.81804$$

d. The test statistic can be used to find the p-value based on a  $\chi^2$  distribution with  $\overset{k-6 \Rightarrow k-1=5}{5}$  degrees of freedom. The corresponding p-value is between .005 and .01. (In fact, it was .007).

e. What is your conclusion using a .01 significance level?

We have significant evidence to conclude the % of each color M&M does NOT match the manufacturer % based on our sample.

3. Trident bubblegum now contains xylitol, a sweetener. However, xylitol has other uses besides bubblegum. A group of Finnish researchers thought that regular use of the sweetener might prevent ear infections in preschool children. In a randomized experiment, two groups of children took either 5 daily doses of placebo or 5 daily doses of xylitol. At the end of the study, it was found that 68 of the 165 children in the placebo group had an ear infection during the study time, while only 46 of the 159 children on xylitol had an ear infection during the study time. Determine if xylitol reduces the risk of ear infection by estimating the difference between the proportions of children who had ear infections (give more than just the point estimate). Be sure to check any necessary conditions.

CI for  $p_1 - p_2$  % unspecified  $\Rightarrow$  Use 95% as example

Check conditions: 1.  $2 \perp RS \rightarrow$  yes it was a randomized experiment

2.  $n_1 \hat{p}_1, n_1(1-\hat{p}_1), n_2 \hat{p}_2, n_2(1-\hat{p}_2)$  all  $\geq 10 \Rightarrow 68, 97, 46, 113$  all  $\geq 10 \checkmark$

$\hat{p}_1 = .412$   $\hat{p}_2 = .2893$   $z^*$  for 95% is 1.96

$$\hat{p}_1 - \hat{p}_2 \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \Rightarrow (.4122 - .2893) \pm 1.96 \sqrt{\frac{(.4122)(.5878)}{165} + \frac{(.2893)(.7107)}{159}}$$

$$\Rightarrow .1229 \pm 1.96(.05255) \Rightarrow (.0199, .2259) \text{ for } \begin{matrix} \text{Placebo} \\ \text{Xylitol} \end{matrix}$$

Since the entire CI is +, placebo % is  $>$  Xylitol %  $\Rightarrow$  yes, Xylitol reduces the risk of ear infection.

4. Determine the appropriate analysis technique for each part below (example: ANOVA or 2 sample t-test).

a. Is there a significant mean difference between gas prices on Wednesday and Saturday gas prices if we check 20 stations on both days?

paired t-test ( $\mu_d$ )

b. Is there a relationship between the region of the country and the size of vehicles driven (small car, large car, truck, SUV)?

$\chi^2$  Independence

c. Is there a significant difference between the average Northampton gas price and average Hadley gas price today?

2 sample t-test ( $\mu_1 - \mu_2$ )

d. Is there a difference in mean miles per gallon for the four populations of vehicle size (small cars, large cars, trucks, and SUVs)?

ANOVA

5. A doctor is trying to determine which cholesterol lowering drug to administer to a patient. For three different drugs, recent studies showed reductions in cholesterol of 6,4, and 2 for the first drug, 10,14, and 9 for the second drug, and 9,12, and 6 for the last drug for a random sample of 9 patients which was split into three groups to test the three drugs. The doctor decides to perform an ANOVA to look for differences in mean cholesterol reductions for the three drugs at a .01 significance level.

a. State the appropriate hypotheses for the doctor's test.

Null hypothesis:  $\mu_1 = \mu_2 = \mu_3$

$\mu_i$  = population cholesterol reduction average for group  $i$ .

Alternative hypothesis: At least 1  $\mu_i$  is different

b. What assumptions need to hold in order for the ANOVA to be valid?

1. Need 3  $\perp$  RS

2. The 3 populations must be normally distributed.

3. The 3 populations must have the same variance.

c. What assumption(s) will be hard to check and why?

Sample sizes are 3.  $\Rightarrow$  It will be hard to check assumptions 2 & 3 above b/c there isn't much data.

d. Assuming the assumptions hold, the following partial output was obtained via R.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
drug	2	78.000	39.000	?	0.03895
Residuals	6	40.000	6.667		

What is the missing value of the test statistic?  $F = \frac{39.000}{6.667} = 5.849708$

e. What is the p-value for the hypothesis test? .03895

f. What is your conclusion?  $\alpha = .01$ !

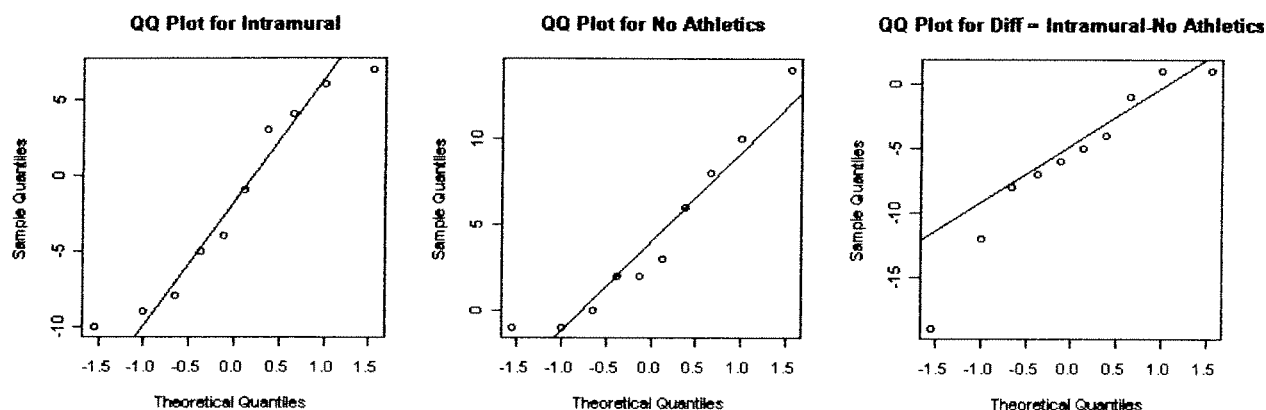
We do not have evidence to conclude that any of the drugs resulted in a different mean cholesterol reduction.

g. Interpret your p-value in context.

If there really was no difference in mean cholesterol reductions we would obtain an F test stat of 5.85 or larger in 3.895% of repetitions of this study.

## Practice Final 2 – Math 17/ENST 24

6. College health officials are studying the impact of additional intramural athletic programs on starting students at a Midwest college. For the freshman class, a random sample of 10 students who participate in intramural athletics is selected. For each student in that sample, a student who does not participate in athletics (at all) is selected whose weight is within 2 pounds of the intramural student. At the end of the year, weight gains (can be negative) for the intramural group and the no athletics group are recorded. The college wants to know if participating in intramural sports decreased weight gain on average. However, the person hired to do the analysis doesn't know which analysis to do! Having had a basic statistics course, they generate the following graphs and output in R:



```
t.test(intramural,noathletics,alternative="less")
Welch Two Sample t-test
data: intramural and noathletics
t = -2.3273, df = 17.09, p-value = 0.01624
alternative hypothesis: true difference in means is less than 0
```

```
t.test(intramural,noathletics,alternative="less",paired=T)
Paired t-test
data: intramural and noathletics
t = -3.0961, df = 9, p-value = 0.006401
alternative hypothesis: true difference in means is less than 0
```

Summaries:

```
summary(intramural)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-10.00  -7.25  -2.50   -1.70   3.75    7.00

summary(noathletics)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -1.0   0.5    2.5     4.3   7.5    14.0

sd(intramural);sd(noathletics)
6.395311; 5.056349
```

(Note: This output looks a little different because it was from R, not Rcmdr, but it is descriptive statistics output. The standard deviations are in the last line.)

a. Should the analyst have performed a two independent samples t-test or a paired t-test? Why?

Two independent samples t-test

Paired t-test

because...

*the participants were matched based on starting weights*

b. Setup appropriate hypotheses to see if participating in intramural sports decreased weight gain on average, and define your parameter(s) of interest.

Null hypothesis:  $H_0: \mu_d = 0$

*(direction depends on parameter definition)*

Alternative hypothesis:  $H_A: \mu_d < 0$

Where  $\mu_d$  is the mean difference in weight gain between intramural – no athletics

*(this order is consistent with the R output)*

c. One assumption for your selected procedure has to do with normal distribution(s). Check that assumption based on the provided plots. Be sure to specify what needs to be normal as well as which plot(s) you are looking at. Does the assumption appear satisfied?

*Need the population of differences to be normally distributed. Use only the last QQ plot.*

*Looks pretty good with all pts. along the line except for one low point.*

d. What is the test statistic and p-value for your hypotheses? (Get from the output)

Test statistic:  $t = -3.0961$

p-value:  $.006401$  *(it is already the p-value we want!)*

e. Is it appropriate to say that participating in intramural sports caused the reduction in weight gain that was observed in the study? Why?

*No. This was not an experiment, so we cannot show causation.*

f. What would be the standard error of the sample mean for the intramural group? Interpret this standard error.

$$se(\bar{x}) = s/\sqrt{n} = \frac{6.395311}{\sqrt{10}} = 2.022$$

*For repeated samples of size 10, we estimate that sample means for weight gain by students in intramurals will differ from the true weight gain by roughly 2.022 lbs, on average.*

7. A survey of students in England asked students their height and whether or not they had ever been bullied while in secondary school. The researchers took the height variable and modified it to be either "short" or "not short". They want to see if there is a relationship between height as "short" or "not short" and bullied ("yes" or "no"). Of 92 short students, 42 had been bullied, and of 117 not short students, only 30 had been bullied.

a. Both the modified height variable and bullied variable are categorical variables.

b. To see if there is a relationship between the two variables, the most appropriate analysis is a chi-square test of Independence, where the hypotheses would be:

Null Hypothesis: *Height and bullied are not related ( $\perp$ ).*

Alternative Hypothesis: *Not  $H_0$ .*

*(Height and bullied are related (dependent)).*

c. What assumptions need to hold in order to perform the test?

*Need expected counts  $\geq 5$  and a random sample of students*

d. For this data set, the chi-square test statistic is 9.133 on 1 degree of freedom, with a corresponding p-value less than .005. Choose a significance level and give a conclusion.

Significance level: .01

*We have evidence that height and bullied do appear to be related.*

e. Suppose the researcher had instead been interesting in whether or not "short" students were bullied more often than "not short" students. Is a chi-square analysis still appropriate? Why? What other analysis would be appropriate if the chi-square one is not appropriate?

*No. You could use a  $p_1 - p_2$  test with a one-sided alternative but not  $\chi^2 \perp$  b/c its alternative is "2-sided" always (i.e. not directed).*



8. After returning to campus after the holiday break, you will need to purchase textbooks for the coming semester. On many college campuses, faculty committees have started investigating the rising textbook costs and brainstorming ways (beyond used texts) to help keep costs down. Suppose a committee has found that scientific textbooks at local textbooks stores cost an average of 115 dollars and a standard deviation of 10 dollars if purchased new, but only an average of 65 dollars and a standard deviation of 5 dollars used.

a. A random sample of 50 new textbooks is examined. What is the distribution of the average price of the random sample of 50 new textbooks? (Give all features of the distribution).

$$\approx N\left(115, \frac{10}{\sqrt{50}}\right) \Rightarrow \approx N(115, 1.414214)$$

b. What is the probability that the average price for the 50 new textbooks selected will be greater than 120 dollars?

$$\begin{aligned} P(\bar{X} > 120) &= P\left(Z > \frac{120 - 115}{1.414214}\right) = P(Z > 3.54) \\ &= P(Z < -3.54) = .0002 \end{aligned}$$

c. A random sample of 50 used textbooks is taken. What is the probability that the average price of the 50 used textbooks will be greater than 67.5 dollars?

$$\bar{X} \approx N\left(65, \frac{5}{\sqrt{50}}\right) \Rightarrow \approx N(65, .7071068)$$

$$P(\bar{X} > 67.5) = P\left(Z > \frac{67.5 - 65}{.7071068}\right) = P(Z > 3.54) = .0002$$

d. In order to do most of this problem, you have been relying on the Central Limit

Theorem which says that if the sample size is large, then the distribution of the sample mean

will be approximately normal.

e. If the sample size is not large, but the original population that the sample was drawn from had a normal distribution, then the sample mean has a normal distribution.

True

False

Practice Final 2 – Math 17/ENST 24

9. A chemist is interested in modeling the weight loss of a particular compound (in pounds) as a function of the amount of time (in hours) that the compound is exposed to the air. Data was collected for 12 samples and a linear model for the relationship between variables (weight loss and exposure time) appears reasonable. Note: weight loss = 4 pounds indicates 4 pounds lost (i.e. it is not recorded as -4).

a. For the data, the response variable is weight loss and the explanatory variable is exposure time (time)

b. An example of a confounding/lurking <sup>extraneous</sup> variable for this problem would be air temperature.

c. Partial R output is provided. Use the output to provide the estimated regression line for predicting the weight loss from the exposure time.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-1.733	1.165	-1.488	.168
exposure	1.317	.208	6.342	.00008

Residual standard error: .804 on 10 degrees of freedom

Multiple R-squared: 0.801

F-statistic: 40.223 on 1 and 10 DF, p-value: .00008

Equation of LS line:  $\hat{y} = -1.733 + 1.317x$   
 $\uparrow$  weight loss  $\uparrow$  time (in hrs.)

d. Predict the weight loss for this compound if it has been exposed to the air for 240 minutes (assume 240 is within the range of the data). 240 minutes = 4 hrs.

$$\hat{y} = -1.733 + 1.317(4) = 3.535 \text{ lbs. lost}$$

e. If the 6<sup>th</sup> residual was .35 and the predicted value for the 6<sup>th</sup> observation was 4.85, what was the actual weight loss for the 6<sup>th</sup> observation?

$$e_6 = y_6 - \hat{y}_6 \Rightarrow y_6 = 4.85 + .35 = 5.2$$

$$y_6 = e_6 + \hat{y}_6$$

f. The chemist thinks that on average the compound loses two pounds in weight for each additional hour it is exposed to the air. Is his thinking consistent with evidence from this sample? Why or why not?

Chemist thinks  $\beta_1 = 2$ . Use CI or test  $\Rightarrow$  CI easier.

$$95\% \text{ CI for } \beta_1 \quad n = 12 \quad n - 2 = 10 = df \quad 95\% t^* = 2.228$$

$$b_1 \pm t^* se(b_1) \Rightarrow 1.317 \pm 2.228(.208)$$

$$1.317 \pm .46342 \Rightarrow (.85358, 1.78042)$$

2 is NOT in the CI so his thinking is not consistent with our evidence from this sample.