Practice Final 1 – Math 17/ ENST 24

Name: Solutions

Math 17/ Enst 24 – Introduction to Statistics

Final Exam

PRACTICE 1

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 the actual exam)*
5. Good luck!

(Some data taken from Utts/Heckard, #2 used with permission UofM)

Practice Final 1 – Math 17/ ENST 24

1. A 1987 study of women in three different occupational groups examined the testosterone level in 46 women who were either unemployed (1), employed but whose job did not require an advanced degree (2), or employed and whose job did require an advanced degree (3).

a. There are two variables in the study –occupational group and testosterone level. Occupational group is an example of a categorical (explanatory) variable and testosterone level is an example of a quantitative (response) variable.

b. If you wanted to test to see if testosterone levels were the same across all three occupational groups on average, what procedure would you use and what hypotheses would you be testing (define your parameter(s))?

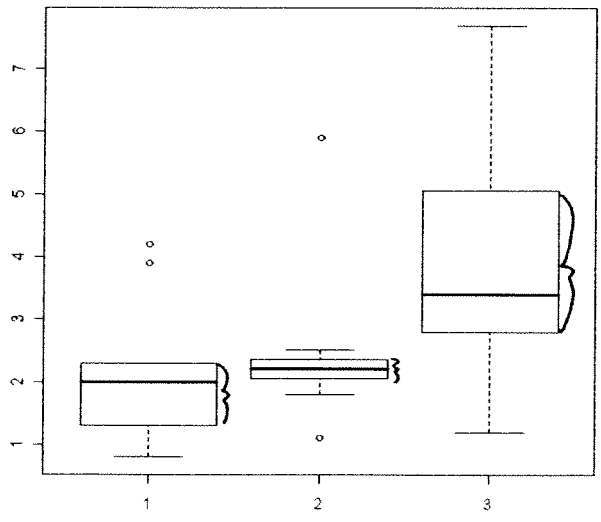
Procedure: ANOVA

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

Alternative hypothesis: At least one μ_i is not the same.

Where μ_i is the population average testosterone level for women in the i^{th} occupational group

c. The testosterone levels were compared using boxplots. Does it look like all the assumptions for the procedure you selected in b are valid? If not, which one(s) are violated?



One of the (four) ANOVA assumptions is clearly violated. The populations should have equal variances, but in the boxplots we see very different IQRs. Therefore, this assumption is not met.

(The populations should also be normally distributed and the outliers may indicate problems with that assumption).

Practice Final 1 – Math 17/ ENST 24

2. Farmer Jed and Farmer Joe are both turkey breeders. Unfortunately, they don't get along very well and they always fight over who breeds the fattest turkeys. One day, the two farmers finally agree on something. They agree to have some of their turkeys weighed and count the number of turkeys that are "big" (i.e. that weigh over 35 pounds). For Farmer Jed, 20 of the 50 randomly selected turkeys are "big". For Farmer Joe, 21 of the 65 randomly selected turkeys are "big".

a. State the hypotheses needed to test that there is a difference between the population proportions of "big" turkeys for the two farmers at a significance level of .05. $\alpha = .05$

$H_0: p_1 = p_2$ where p_1 = pop. prop. of Jed's turkeys which are big

$H_A: p_1 \neq p_2$ p_2 = " " " of Joe's " "

b. Verify any conditions necessary to perform the test.

1. 2 ⊥ RS → yes this is stated and reasonable to assume Jed's turkeys don't affect Joe's & vice versa

2. Check sample size condition. $n_1\hat{p}_1, n_1(1-\hat{p}_1), n_2\hat{p}_2, n_2(1-\hat{p}_2)$ all ≥ 10 ?
20, 30, 21, 44 yes. this is met.

c. Carry out the test. What is the value of your test statistic and corresponding p-value?

$$\hat{P} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} = \frac{20+21}{50+65} = \frac{41}{115} = .3565 \qquad \hat{p}_1 = .4 \quad \hat{p}_2 = .3231$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}} = \frac{.4 - .3231}{\sqrt{\frac{.3565(.6435)}{50} + \frac{(.3565)(.6435)}{65}}} = \frac{.0769}{.0900972} = .8535 \Leftarrow \text{test stat}$$

$$p\text{-value} = 2P(z > .85) = 2P(z < -.85) = 2(.1977) = .3954$$

d. Based on your p-value, choose one: **Reject the null hypothesis** **Do not reject the null hypothesis**

e. What is your conclusion regarding the turkeys of Farmer Jed and Farmer Joe?

We do not have evidence to conclude that there is a difference in the proportions of "big" turkeys for the 2 farmers.

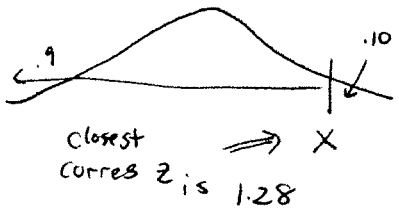
f. Interpret your p-value in context.

If there really was no difference in the proportions of big turkeys, we would obtain a test stat of .85 or larger or -.85 or smaller in almost 40% of replications of the study.

Practice Final 1 – Math 17/ ENST 24

3. Many people get frustrated if they have to wait in line for an extended period of time and then have to wait while their order is completed. Assume that the total amount of time it takes at a particular supermarket for the meat department to take and fill an order is normally distributed with a mean of 4 minutes and a standard deviation of 1 minute.

a. 10% of customers will have a total service time longer than 5.28 minutes. (Show all work).



$$z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + \sigma z$$

$$x = 4 + 1.28(1)$$

$$= 5.28$$

b. What is the probability a randomly selected customer will have a total service time longer than 4.5 minutes? $X \sim N(4,1)$

$$P(X > 4.5) = P\left(z > \frac{4.5-4}{1}\right) = P(z > .5) = P(z < -.5) = .3085$$

c. What is the probability that a random sample of n=16 customers will have an average total service time longer than 4.5 minutes?

$$X \sim N(4,1) \Rightarrow \bar{X}_{16} \sim N\left(4, \frac{1}{\sqrt{16}} = \frac{1}{4}\right)$$

$$\Rightarrow P(\bar{X} > 4.5) = P\left(z > \frac{4.5-4}{.25}\right) = P(z > 2) = P(z < -2)$$

$$= .0228$$

d. If you did not know that the service times were normally distributed, would you have been able to answer c. by relying on the Central Limit Theorem? Why or why not?

No. The sample size is n=16 which is too small to apply the CLT.

e. Assume that same supermarket has 3674 employees in the state. Each employee fills out a mandatory form about possible changes to health care benefits. Of the 3674 employees, 1403 are in favor of the change. 1403/3674 is .3818, or roughly 38%.

Is the 38% described a population parameter or a statistic? population parameter

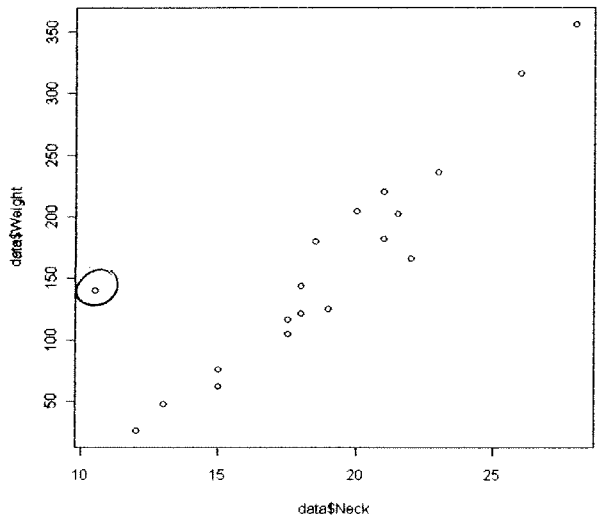
What notation would be used to denote it? p

(mandatory form ⇒ like a census, the entire population responded).

Practice Final 1 – Math 17/ ENST 24

4. In some national parks, park rangers must deal with meandering bears. Some studies attempting to track bears and observe their habits have tagged animals and taken measurements of the bears during the short time they are held and the tag applied. A study of 19 female bears resulted in measurements of neck girth, weight, length, and chest width for each bear. The park rangers are interested in predicting weight (lbs.) using neck girth (in.) for future tagged female bears (it is much easier to measure neck girth than to get a tranquilized bear on a scale).

a. A scatterplot of neck girth vs. weight was generated (at right). Does the plot suggest that a linear regression is appropriate to investigate the relationship between these two variables? Yes, all points lie along a line with + slope except 1.



b. The reported correlation coefficient is .89. Interpret this value. $r = .89$ suggests that there is a strong (or very strong) positive linear relationship between weight and neck girth for female bears.

Selected R output from a linear regression is provided below.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-158.78	40.46	-3.924	0.00109 **
Neck	16.95	2.10	8.071	3.24e-07 ***

Residual standard error: 40.13 on 17 degrees of freedom
Multiple R-squared: 0.793, Adjusted R-squared: 0.7809
F-statistic: 65.14 on 1 and 17 DF, p-value: 3.235e-07

c. What is the equation of the least square regression line of predicting female bear weight from neck girth?

$\hat{y} = -158.78 + 16.95x$

$\hat{y}$: weight; x: neck girth

d. Obtain a 95% confidence interval for the population slope. Is it reasonable to say that the population slope is less than 20? Explain.

t^* based on 17 df for 95% is 2.110

$b_1 \pm t^* se(b_1)$

$16.95 \pm 2.110(2.10)$

$\Rightarrow (12.519, 21.381)$

The CI includes 20 and values > 20, hence we cannot conclude that the slope is less than 20.

Practice Final 1 – Math 17/ ENST 24

4 continued.

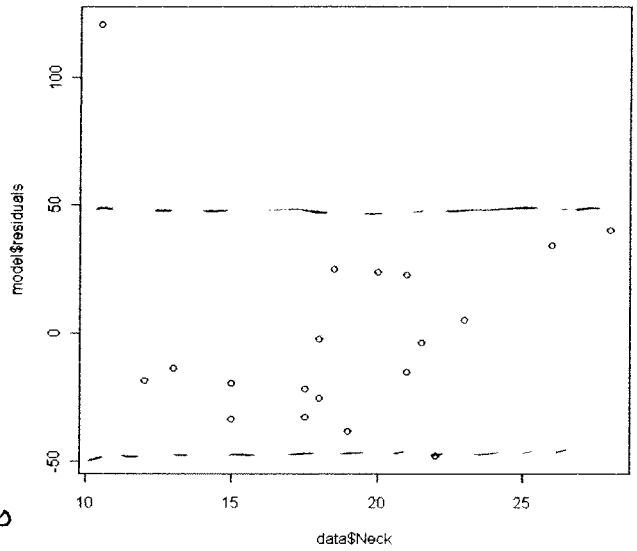
e. Predict the bear weight for a female bear with neck girth of 18 inches.

$\hat{y} = -158.78 + 16.95(18) = 146.32$ lbs.

f. Would you be able to make a prediction for a female bear with a neck girth of 33 inches or a male bear with a neck girth of 18 inches? Why or why not?

No. The data is on females, not males and based on the scatterplot, 33 is out of the data range of neck girth for the female bears.

g. In order to check the assumptions for linear regression, several plots were made. One plot is shown at right.



This is an example of a residual plot.

One assumption this plot is used to check is:

whether or not the population error terms have a constant standard deviation (or variance)

Does that assumption appear valid? Explain.

The assumption looks okay except for 1 pt. b/c there is a rough horizontal band. (The "bad" pt does however make the plot ∨ shaped which is not good, for the linearity assumption.)

5. True/False.

a. Power is the probability of making the correct decision of rejecting the null hypothesis when in fact the alternative is true. **(True)** False

b. If we fail to reject the null hypothesis that a population mean is equal to 10 at the 5% level, then we can say that there is a 95% chance that the population mean is equal to 10.

True **(False)**

c. Increasing the sample size, n, results in a decrease in the power of a hypothesis test.

True **(False)**

Practice Final 1 – Math 17/ ENST 24

6. A variant on the infamous "half-full vs. half-empty" water glass experiment requires study participants to try to draw the water line in a tilted glass when provided with a picture of the level glass. A psychologist studied the success of participants in drawing the line within five different colleges in the University of South Carolina on 8 different example pictures. A score of "pass" was given if the correct line was drawn on at least 4 of the 8 pictures. 50 participants were selected from each school and additionally, exactly half of the 50 were female and half were male from each school. The results are summarized below.

College	Business	Language Arts	Social Sciences	Natural Sciences	Engineering	Total
Pass	33	32	25	38	43	171
Fail	17	18	25	12	7	79
Total	50	50	50	50	50	250

a. Which college has the highest success rate in correctly determining the water line?

Engineering $43/50$

b. To explore the relationship between college and pass status and see whether or not pass status depends on college, what is the appropriate inference procedure? (Be specific). Provide hypotheses for your chosen procedure.

Procedure: χ^2 Test of Independence

Null Hypothesis: College and pass status are not related (independent).

Alternative Hypothesis: College and pass status are related (dependent).

c. The chi-square statistic for the 2-way table above is 16.915, with a corresponding p-value of .002. At a significance level of .01, what is your conclusion for your hypothesis test in b.?

There is evidence that college and pass status are dependent.

d. Interpret the test statistic. The sum of differences between the observed and expected counts squared when standardized by the expected counts was 16.915, much larger than the 0 that would occur if they were equal.

Practice Final 1 – Math 17/ ENST 24

7. A researcher is interested in whether the mean weight of second babies is different than the mean weight of first babies. She asks a representative sample of 40 women with at least two children for the weights of the oldest two at birth.

a. What is the benefit of asking women about the weights of their oldest two children at birth compared to taking two samples and asking the first group for the weight of their first child and the second group for the weight of the second child?

You control for differences in weight expected btw mothers by asking a mother about her 2 oldest children.
(i.e. control extra sources of variability)

b. Define the parameter of interest to the researcher in statistical notation and in words.

μ_d = population mean difference in weight of children (oldest - second born)

(order is subjective) (but this matches below)

c. The researcher found that the mean of the sample differences (first – second) was 5 ounces with a standard deviation of 7 ounces. Obtain a 90% confidence interval for the parameter you defined in b. use 35

90% CI for μ_d $n = 40$ $\bar{d} = 5$ $s_d = 7$ $n-1 = 39$ df $\Rightarrow t^* = 1.690$

$$\bar{d} \pm t^*\left(\frac{s_d}{\sqrt{n}}\right) \Rightarrow 5 \pm 1.69\left(\frac{7}{\sqrt{40}}\right) \Rightarrow 5 \pm 1.87$$

$$\Rightarrow (3.13, 6.87)$$

d. The researcher's question was whether or not there was a difference in the mean weights. Use your confidence interval to answer that question and explain how you reached that conclusion. Be sure you specify what significance level you are able to use thinking about it as a hypothesis test compared to the confidence interval.

Significance level: .10 for 2 sided test (just asked for "difference")

Conclusion: We have evidence to conclude that there is a difference in population mean weight between the first 2 children born to a woman.

Reasoning:

0 is NOT in the CI, so there is a difference.

Practice Final 1 – Math 17/ ENST 24

8. On a college campus (not Amherst), suppose that 30% of students drive to campus, 50% bike to campus, and 20% get to campus some other way each day (includes bus, walking, etc). The college sponsors a "spare the air day" and hopes that fewer students drive to campus that day. To see the results of the event, a random sample of 300 students were asked how they got to campus on that day. College officials want to know if the results differ from the normal transportation patterns at the college.

Method	Drive	Bike	Other	Total
Frequency	80	200	20	300
Expected	$.3(300) = 90$	$.5(300) = 150$	$.2(300) = 60$	300

a. What inference procedure is appropriate to address the college officials' question? (Be specific).

χ^2 Goodness of Fit

b. Determine hypotheses for the procedure you selected in a.

Null Hypothesis: H_0: $p_1 = .3$, $p_2 = .5$, $p_3 = .2$

Alternative Hypothesis: H_A: Not H_0

p_1 = prop who drive to campus
p_2 = prop who bike and p_3 = prop who use another method

c. Compute expected counts for the different methods of transportation and fill in the table above.

$np_1 = .3(300) = 90$ $\quad np_3 = .2(300) = 60$

$np_2 = .5(300) = 150$

d. Compute the relevant test statistic for your hypotheses.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(80-90)^2}{90} + \frac{(200-150)^2}{150} + \frac{(20-60)^2}{60} = 44.\overline{4}$$

Huge χ^2 test stat

e. The corresponding p-value is less than .005, which was found by looking at a χ^2 distribution with 2 degrees of freedom. What is your conclusion at a .01 level?

There is very strong evidence that the spare the air day did have a different transportation pattern than a normal day.

Note: We did not check assumptions, but if you check, you will find they are met.

Practice Final 1 – Math 17/ ENST 24

9. A nursing student is trying to address some common patient concerns with a study on a fairly new drug. Although the drug has been approved by the FDA, she is still interested in the side effects and determining the effective dosage level. For each situation described, determine if the nursing student should perform a hypothesis test, or make a confidence interval, or perform some other procedure (specify other procedure – regression, ANOVA, chi-square test).

a. Interested in the estimated percentage of patients who suffer from nausea as a side effect

Hypothesis Test **Confidence Interval** **Other:** _____

b. Interested in whether 250 milligrams is an effective dose of the drug or if the dosage needs to be increased. ↳ specific value yes or no

Hypothesis Test **Confidence Interval** **Other:** _____

c. Interested in the relationship between days on the drug and concentration of drug in the blood for a daily dose of 250 milligrams.

Hypothesis Test **Confidence Interval** **Other:** Regression
b/c both variables appear quantitative

10. An ANOVA was performed to examine differences in mean GPAs in a large introductory class at a college in California based on preferred seat location during class. Students were asked whether they preferred sitting in the back (1), middle (2), or front (3) of the large lecture hall and what their GPA was for the most recent semester. The ANOVA revealed that there were significant differences in mean GPAs based on seating preference.

Partial R output from the analysis is provided.

TukeyHSD(model,"seat") ★ ↙ it's right here!
Tukey multiple comparisons of means
95% family-wise confidence level
Fit: aov(formula = GPA ~ seat)

	diff	lwr	upr	
2-1	.0659	-.1028	.2347	← 0
3-1	.2835	.0846	.4824	← no 0
3-2	.2176	.0561	.3791	← no 0

In order to determine where the differences are, after an ANOVA has indicated the null hypothesis should be rejected, you need to perform multiple comparisons ★ procedures.

Summarize the differences in mean GPAs that are revealed by the output.

The output shows that the population mean GPA for front (3) students is different than the mean GPA for middle and back sitting students.

OR:

1 2 3