STP226 Final Review Problems Chapters 10-12, 4

Topics: Inferences for Two Populations Means, Using Independent Samples; Different Standard Deviations, Inferences for Two Populations Means, Using Paired Samples; Confidence Interval for One Population Proportion, Hypothesis Test for One Population Proportion, The Chi-Square Distribution, Chi-Square Goodness-of-Fit Test, Chi-Square Independence Test, Linear Equations with One Independent Variable, Linear Correlation, The Least Squares Regression Equation, The Coefficient of Determination,

#1 Is autism marked by different brain growth patterns in early life. Studies have linked brain size in infants and toddlers to a number of future ailments, including autism. One study looked at brain size of 30 autistic boys and 12 nonautistic boys (control) who had received MRI scan as toddlers. The average brain volumes and standard deviations in milliliters are given below. Two samples may be regarded as independent SRS from two normal populations with different standard deviations. Is there a difference between the means? Test appropriate hypothesis, use $\alpha = 0.05$. Software gives 25.3 df.

Group	Condition	n	$\bar{x}$	s	Population means
1	Autistic	30	1297.6	88.4	μ_1
2	Control	12	1179.3	70.7	μ_2

#2 Compute 95% CI for $\mu_1 - \mu_2$ in example #1 and state if your CI is consistent with the outcome of the test in ex#1

#3 The number of friends consulted for advice before purchasing a car or a computer was examined by a certain consumer research paper. Two independent samples of consumers were selected. The summary statistics consistent with the information in the paper are given in the following table. Software produced 20.96 df.

Type of purchase	Number of purchases (n)	Mean number of friends consulted ($\bar{X}$)	Standard deviation (S)
(1) car	12	3.65	0.22
(2) computer	15	4.26	0.46

Is mean number of friends consulted before each purchase greater for people purchasing computers? Test appropriate hypothesis at 5% significance level. Assume that populations are normal with different standard deviations.

#4 During a 1998 race for state senator a newspaper conducted a poll and found that 607 of 1200 registered voters sampled would vote for the Republican candidate. Let p be the population proportion of registered voters who would vote for the Republican today.

a. Give a 90 % level confidence interval for p.

b. Based on your CI from part a, can you conclude that if the election was held today a Republican is likely to win if more than 50% is needed for a win. Explain.

c. What sample size is needed to cut margin of error in your interval to 1% Use your estimate from part a as a guess for p-hat?

#5 Use a data from example #4 to answer the following question:
Is there evidence at 10 % significance level that **majority of registered voters** will vote for a Republican if election was held that day? Test appropriate hypothesis.

#6. Each person in two independent large random samples of male and female working adults in Calgary, Canada, was asked how long , in minutes, his or her typical daily commute was. The 99% confidence interval for the difference between true mean commute times for male and female working Calgary residents , ($\mu_1 - \mu_2$), is (15.97 minutes , 26.37 minutes).
Do you think that that mean commute times for male and female working Calgary residents are different? Explain why.

#7 Experiment was conducted to see if wounding a tomato plant would make it improve its defense against insects. Researchers grew larvae of the tobacco hornworm on wounded and unwounded plants, weight in mg after 7 days of growth was recorded.
Summary of the results are given below:

	Control (1)	Wounded (2)
n	18	16
$\bar{x}$	37.96	28.66
s	11.14	9.02
population means	μ_1	μ_2

(μ_1 and μ_2 are mean weights of larvae for wounded and control populations)

Do we have evidence at 1% significance level that wounding is effective in increasing plant's defense against insects (i.e. we are asking if average weight of the larvae will be larger when it was feeding on control plants as compared to wounded plants)
Software produced df=31.8

#8 The offspring produced by a cross between white and red plants can produce red , pink or white plants. A simple inheritance model suggests that the red, pink and white offspring should be in a ratio 1:2:1 respectively, which means we should have 25% each white and red offspring and 50% of pink offspring. An experiment was conducted in which 100 plants were bred by crossing the two parents, one red one white. The genetic classification of offspring are recorded below. Do these data support the hypothesis that the offspring follow the predicted ratio? Test using $\alpha = .05$

Genotype:	Red	Pink	White
O=observed frequency:	18	55	27

#9 In a study of spatial orientation of certain fish 50 individuals were caught in various locations and later tested in artificial pool to see which direction they would choose when released. Use the following data and Chi-square GOF test to test the null hypothesis that directional choice of these fish is random. Use $\alpha = .05$.

Directional choice:	**#of fish=O**
Toward shore	18
Away from shore	12
Along shore (right)	13
Along shore (left)	7

#10 Does pollster's gender have an effect on poll responses by men? *A U.S News & World Report* stated recently: "On sensitive issues, people tend to give "acceptable" rather that honest responses; their answers may depend on the gender of interviewer." To support that claim, following data was provided for an Eagleton Institute poll in which surveyed men were asked if they agreed with a statement: *"Abortion is a private matter that should be left to the women to decide without government intervention".*
Two random samples of males were independently interviewed, one by male interviewers, the other by female interviewers. Data is shown below:

		Gender of Interviewer		
		MALE	FEMALE	total
Response	AGREE	560	308	868
	DISAGREE	240	92	332
	total	800	400	1200

Do we have evidence that response is associated with the gender of the interviewer? Test using $\alpha = .05$ and Chi-square test of independence.

#11. Following data was obtained to determine if person's occupation is independent of weather the cause of death of that person was a homicide or other cause

		Occupation			
		Police	Cashiers	Taxi Drivers	Guards
Cause of Death	Homicide	82	107	70	59
	Other	92	9	29	42

Do we have evidence that cause of death and occupation are associated? Use 5% significance level.

#12
Use following information
Data in the table below represents quiz scores (out of 10 points) of several students and their study time (in hours) for that quiz.

Study time	Quiz score
0	2
1	3
2	5
4	7
5	9

- Calculate the **least squares regression equation**, use Study time as an explanatory variable and Quiz score as a response variable. Round your answers to 2 decimal places. **You may use a calculator or by hand computations.**

- Interpret the slope of the equation you received, be very specific.

- Given that SST= 32.8 and the Error Sum of Squares is 0.42, compute the Regression Sum of Squares and determine the percentage of total variability in the Quiz scores that is explained by your regression line.

#13
Suppose we obtained following least squares linear equation: y = 45.2 – 6.8x , where x and y showed strong linear trend. Coefficient of determination was 0.93. Values of x were form an interval [1, 6]

- Determine linear correlation coefficient, round your answer to 2 decimal places.

A) 0.96 B) -0.96 C) 0.86 D) -0.86 E) none of these

- Use the equation to predict y for x= 5, round your answer to 2 decimal places

A) 5.91 B) 192 C) 219.2 D) 79.2 E) none of these

- Explain why predicting y for x=17 may not give reasonable results

#14 Suppose you are testing $H_0: \mu = 6$ *vs* $H_a: \mu > 6$, how would you classify that test? (2-tailed, left-tailed, right-tailed)

#15 Suppose test in #1 is a z-test and test statistics z=2.45
a) compute p-value for your test include sketch
b) find rejection region for your test at 5% significance level, include sketch.
c) Do you reject null hypothesis at 5% sign. level?
d) Repeat parts a-c if you change alternative to $H_a: \mu \neq 6$

#16 Suppose test in #1 is a t-test and test statistics t=1.45, df=16
a) compute p-value for your test, include sketch
b) find rejection region for your test at 5% significance level, include sketch

#17 Two independent random samples of males and females from a large University we taken and each person was asked how many hours per week they study for their classes. Data is presented below:

$Males: \bar{x}_1 = 13 \quad s_1 = 5 \quad n = 31, \quad Females: \bar{x}_2 = 16 \quad s_2 = 2.1 \quad n = 33$

a) Test appropriate hypothesis to establish if there is a difference between mean weekly study times for males and females at that University.
(decide if pooled or non-pooled test should be used)
b) Suppose 95% CI for the difference between mean weekly study times for males and females is (1.3, 2.4), do you think there is evidence that means are different?

#18 12 sets of identical twins (10 years old) were given a standard math test, then the differences between test scores were computed (test score of older twin- test score of younger twin) giving following data: $\bar{d}=3.4 \quad s_d=2.3$. Is there evidence at 10 % significance level that the mean test scores for all the 10 year old twins are different?

#19 Following table shows grades distribution for a random sample of Mat117 students at ASU last year. Do we have evidence at 5% significance level that gender and grade received are associated? Test by means of Chi-square test. Use p-value method.

	A or B	C	D or E
Males	23	43	14
Females	17	51	6

#20 Suppose in a random sample of 325 students at ASU 231 stated that they work part or full time. Obtain 90% CI for true % of all ASU students that work part or full time.

#21 Suppose out of 1200 cars that passes this morning through certain stretch of 101 freeway following number of cars selected each of the available lanes:

lane1	lane2	lane3	lane4
250	275	310	365

Do we have evidence at 5% significance level that some lanes are preferred over others? Test by means of Chi-square GOF test. Use rejection region method.

#22 Suppose Chi-square test of independence gives a test statistics $\chi^2 > 9$ with 3 degrees of freedom. Do you have evidence of association at 5% sign. level?

#23 Student obtained following two regression lines to fit the data shown in the table below: Line A: $\hat{y}=x$ and Line B: $\hat{y}=0.5x+0.5$

Determine which of the two lines fits given set of data points better **according to the Least Squares Criterion. Clearly show all work for credit.**

x	1	2	3	4
y	2	1	3	3

#24 If you try to rent an apartment or buy a house, you will find that real estate representatives establish apartment rents and house prices on the basis of the square footage of the heated floor space. The following data give the square footage (in 100-s square feet) and sale prices (in $1000) of 12 houses randomly selected from those sold in a small city.

X=sq. feet ($100ft^2$)	14.6	21.08	17.43	15.00	18.64	23.91	19.77	16.1	15.3	17.59	18.21	22.16
Y=price ($1000)	59.7	79.3	71.4	61.1	64.4	85.9	75.4	67.0	62.4	68.2	74.3	81.7

a) Make a **scatter plot** of your data on your calculator. Describe the form of the relationship. Is there a linear pattern? Do you see any unusual observations **(outliers)**? Explain what is an **influential observation.**

b)Fit the **least squares regression line** to your data. You may use you calculator. Give the equation of your line.

c) Explain what is the criterion satisfied by the Least Squares Regression Line.

d)Explain what the **slope** of your equation tells you about the change in y with respect to the change in x. Give units of the slope.

e)Use your equation to **predict** the price of a 1500 ft^2 house . Compute the error you made by using the equation. Did you overestimate or underestimate by using your equation?

f) Use your equation to **predict** the price of a 5500 ft^2 house. Clearly explain why this may not be a reasonable prediction

g)Compute **correlation coefficient r** (use calculator). Explain what r tells you about the strength of linear relationship between x and y.

h)Determine the **% of variation** in y that is explained by the regression line? Does it mean that the regression line fits well? Explain.

Key:

#1 (not pooled t-test)
$H_0: \mu_1 = \mu_2$, $H_a: \mu_1 \neq \mu_2$ t=4.55
p-value: p=1.2*10^{-4} < .05 , df=25.3, round it down to 25
Rejection Region: $t \leq -2.06 \cup t \geq 2.06$
Reject H_0 , evidence for alternative hypothesis . There is a difference between brain volumes.

#2 95% CI for $\mu_1 - \mu_2$ is: (64.7, 171.9), clearly no 0 inside, consistent with rejecting null hypothesis

#3 $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 < \mu_2$ $t = -4.53$ df=20, non-pooled t-test
r**ejection region:** $t \leq -1.725$ Reject H_0 , evidence that more friends are consulted before computer purchase. **p-value:** 9.2*10^{-5} < .05, same conclusion.

p-value from tables: 4.53 > 2.845= $t_{.005}$, so p-value< .005

#4 a. (0.48, 0.53) b. Not conclusive, CI contains numbers below and above 50%.

c. $.5058(1-.5058)\left(\frac{1.645}{.01}\right)^2 = 6764.2$, n=6765

#5 $H_0: p = .5$ $H_a: p > .5$ $\hat{p} = .5058$, z=0.4041. P=0.3431>.10
(rejection region : $z \geq 1.282$, 0.4041 is not in the rejection region)
Do not reject H_0 , no evidence that majority will vote for a Republican

#6 There is evidence of a difference, since CI does not contains 0. Based on our CI we have 99% confidence that 15.97< $\mu_1 - \mu_2$ <29.37

#7 We test $H_0: \mu_1 = \mu_2$ (wounding not effective) vs $H_a: \mu_1 > \mu_2$ (wounding effective), and $\alpha = 0.01$
$t_s = \frac{37.96 - 28.66}{3.46} = 2.69$, use df=31

p-value=tcdf(2.69, 10^6, 31)=.006<.01, so we reject null hypothesis, there is evidence for alternative hypothesis at 1% significance level. Yes, wounding appears to increase plant's defense against insects.

#8 This is GOF Chi-square test, variable=genotype of offspring, 3 classes.
Notice that if predicted ratio is a:b:c, then $p_1 = \frac{a}{a+b+c}, p_2 = \frac{b}{a+b+c}$ and $p_3 = \frac{c}{a+b+c}$
H_0: p_1=1/4, p_2=1/2, p_3=1/4 (data follows predicted ratio)
H_a: not all probabilities are as stated in the null hypothesis (data does not follow predicted ratio)
E= 25, 50 and 25 , $\chi^2 = \frac{(18-25)^2}{25} + \frac{(55-50)^2}{50} + \frac{(27-25)^2}{25} = 2.62$ df=2
p-value = $\chi^2 cdf(2.62, 10^6, 2)$ =.27>.05
(Rejection region: $\chi^2 \geq 5.991$, 2.62 is not in the rejection region)
Do not reject null, data support hypothesis that offspring follow predicted ratio.

#9 GOF test, $H_0: p_1 = p_2 = p_3 = p_4 = .25$ ie. directions are randomly selected (all equally likely)
H_a: not all p_i are as stated in null hypothesis (selections not random, some

directions are preferred over others)

E= np=.25(50) = 12.5 for each category $\chi_s^2 = 4.88$, p= $\chi^2 cdf(4.88, 10^6, 3)$ =0.180, do not reject H_0 , no evidence that choices are not random.

#10 Hypotheses can be formulated as follows:

H_0 : Proportions of agree/disagree responses are equal for male and female interviewers (response independent of the sex of the interviewer)

H_a : Proportions of agree/disagree responses are different for male and female interviewers (response not independent of the sex of the interviewer)

Expected counts are: 579 289
221 111 (rounded to nearest integer)

$\chi^2 = 6.53$ df=1, p-value=0.011<0.05.

Reject H_0 , we have evidence that proportion of agree/disagree responses depends on the gender of the interviewer.

#11 Here we have 4 independent samples from different occupations, one variable: cause of death.

H_0 : Cause of death and occupation are not associated

H_a : Cause of death and occupation are associated

Expected counts are: 113 75 64 66
61 41 35 35

$\chi^2 = 65.52$ df=3, p-value=$3.4*10^{-14}$<0.05.

(Rejection region: $\chi^2 \geq 7.815$, 65.52 is in the rejection region)

Reject H_0 , there is overwhelming evidence that cause of death and occupation are associated.

#12

- $\hat{y} = 1.37x + 1.91$
- Slope= change in y with respect change in x. As study time (x) increases by 1 hour, the grade increases by 1.37 points.
- SSR=32.38 $r^2 = \frac{32.38}{32.8} = 0.987$, 98.7%

#13

- r=-.96
- $\hat{y} = 11.2$
- Since our x-values were from the interval [1,6], x=17 is outside the data range, so it would be extrapolation, we can't be sure that the same trend will hold past x=6 or below x=1.

#14 right-tailed test

#15 a. pv=.0071 pv= area over $Z \geq 2.45$
b. Cv=1.645, rej. Region: $Z \geq 1.645$
c. Null rejected
d. Pv=2(.0071)=.0142 pv=area over $Z \leq -2.45$ or $Z \geq 2.45$ cv= ± 1.96
rej. Region: $Z \leq -1.96$ or $Z \geq 1.96$
Null rejected

#16 a. Pv=.0832 pv=area over $t \geq 1.45$
b. Cv=1.746, rej Region : $t \geq 1.746$

#17 a. Non pooled t-test, two tailed test, t=-3.09, pv=.0036<.05
Null would be rejected, there is evidence that population means are different.
b. Yes, since CI does not contain zero

#18 Paired t-test, two tailed, t=5.12, pv=.00033<.10
Null rejected, we have evidence that mean test scores for all 10 y. old twins differ.

#19 $\chi^2 = 4.55$, pv=.1026>.05, null not rejected, no evidence of association.

#20 CI: (.6694, .75214)

#21 H_o: All lanes equally preferred (p1=p2=p3=p4=.25)
H_a: Some lanes preferred over others
E=.25(1200)=300 $\chi^2 = 24.83$ df=3 , cv=7.815
Null rejected, some lanes are preferred over others

#22 cv=7.815, since test stat. >9, it falls into the rejection region, so null would be rejected, we have evidence of association.

#23 $\hat{y} = x$ is better, sum of squared errors=2<2.5 = sum of squared error for the other line.

#24a) Strong, positive linear trend. Possible outlier (18.64, 64.4).
Influential observation=observation that will greatly change the regression equation when removed from the data.

b) $\hat{y} = 20.83 + 2.73x$

c) $\sum e^2 = \sum (y - \hat{y})^2$ is minimized for LS regression line

d) As x increases by 100 square feet, y increases by $2730

e) $\hat{y} = \$61780$, e=-$680, overestimate

f) $\hat{y} = \$170980$, this is an example of extrapolation, predictions outside of data range may not give reasonable results.

g) r=0.946 indicating very strong positive linear relationship between x and y

h) $r^2 = 0.894$ 89.4% , good fit