National University of Computer and Emerging Sciences, Lahore Campus

Course:	Advanced Statistics	Course Code:	DS-2003
Program:	BS (Data Science)	Semester:	Spring 2022
Duration:	60 Minutes	Total Marks:	40
Paper Date:	06-May-22	Weight	15%
Section:	4A	Page(s):	5
Exam:	Midterm II	Roll No.	

Instruction/Notes: Attempt all questions on the question paper. Rough sheets can be used. If you think some information is missing then assume it and mention it clearly.

$$SSG = \sum_{i=1}^{k} n_i(\overline{x_\iota} - \bar{x})^2$$

$$SST = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SSE = SST - SSG; F = \frac{MSG}{MSE}$$

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O - E)^2}{E}$$

$$t = \frac{\bar{x} - \mu}{SE} \ (for\ hypothesis\ testing); where\ SE = \frac{s}{\sqrt{n}}$$

Confidence Interval: $point\_estimate \pm z^* * SE$

Questions 1 – 6 are worth 1 mark each, questions 7 – 8 are worth 2 marks each

1. Confidence intervals are useful when trying to estimate _________________
   a. Unknown population parameters**
   b. Known population parameters
   c. Unknown statistics
   d. Known statistics

2. The one-sample z-statistic is used instead of the one-sample t-statistic when _________
   a. µ is known
   b. µ is unknown
   c. σ is known**
   d. σ is unknown

3. The probability you reject the null hypothesis when in fact the null hypothesis is true is called?
   a. A Type I error**
   b. A Type II error
   c. A Type III error
   d. A power

4. For a least squares regression line, the slope is calculated as:
   a. $\hat{\beta}_1 = \frac{SP_{xy}}{SS_{xx}}$ **

   b. $\hat{\beta}_1 = \frac{SS_{yy}}{SS_{xx}}$

   c. $\hat{\beta}_1 = \frac{SS_{xx}}{SS_{yy}}$

   d. $\hat{\beta}_1 = \bar{y} - \hat{\beta}_0 x$

5. What statistic is considered as the proportion of the variability in response variable y that is attributable to the linear relationship with the explanatory variable x?
   a. r/n (where r is the coefficient of correlation, and n is the sample size)
   b. r^2 (where r is the coefficient of correlation) **
   c. $\hat{\beta}_1$ (where this is the slope in the regression model)
   d. $p - value$

6. Suppose we were interested in determining if there were differences in the average prices among two local supermarkets. We randomly pick six (possibly different) items to compare at both supermarkets. Which statistical procedure would be best to use for this study?
   a. One-sample t-test
   b. Two-sample t-test**
   c. Two-sample z-test
   d. None of the above

7. Perform a one-sample t-test using the following statistics: n = 5; $\bar{x}$ = 3.871; s = 0.679

   Note the following results in R for help:

   ```
   qt(0.95, df=4) = 2.131847              qt(0.99, df=4) = 3.746947
   ```

   The null hypothesis μ = 5.0 is:

   a. Not rejected at the 5% level; not rejected at the 1% level
   b. Not rejected at the 5% level; rejected at the 1% level
   c. rejected at the 5% level; not rejected at the 1% level** t=(3.871-5)/(0.679/sqrt(5))
   d. rejected at the 5% level; rejected at the 1% level

8. You buy a package of 122 Smarties and 19 of them are red. What is a 95% confidence interval for the true proportion of red Smarties?
   a. (0.092, 0.220)** -> (19/122) ± (1.96*(sqrt((19/122)*((122-19)/122)/122)))
   b. (0.103, 0.230)
   c. (0.085, 0.199)

9. [**15 marks**]

a. (12m) Students were given different drug treatments before revising for their exams. Some were given a memory drug, some a placebo drug and some no treatment. The exam scores (%) are shown below for the three different groups:

	Memory Drug	Placebo	No Treatment
	70	37	3
	77	43	10
	83	50	17
	90	57	23
	97	63	30
Mean	83.40	50.00	16.60
Variance	112.30	109.00	112.30

Carry out a 1-way ANOVA by hand to test the hypothesis that the treatments will have different effects. Note, the following result in R is of importance, assuming a 5% significance level:

```
qf(0.95, df1=2, df2=12) = 3.885294
```

Groups	***Count***	***Sum***	***Average***	***Variance***
Memory Drug	***5***	***417***	***83.4***	***112.3***
Placebo	***5***	***250***	***50***	***109***
No Treatment	***5***	***83***	***16.6***	***112.3***

ANOVA

Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Between Groups	***11155.6***	***2***	***5577.8***	***50.16007***	***1.49E-06***	***3.885294***
Within Groups	***1334.4***	***12***	***111.2***			
Total	***12490***	***14***				

H_0: Different treatments have the same effect
H_A: Different treatments have different effects

Reject Null Hypothesis, 50.1 > 3.88. Evidence suggests that the treatments will have different effects. Scores of each group are not the same.

b. (3m) What is the Bonferroni correction, and when is it used?

The Bonferroni correction suggests that a more stringent significance level is more appropriate for these tests, i.e., the scenario of testing many pairs of groups): $\alpha^* = \alpha/K$
Where: K = is the number of comparisons being considered.

10. [**15 marks**] A random sample of 395 people were surveyed and each person was asked to report the highest education level they obtained. The data that resulted from the survey is summarized in the following table:

	High School	Bachelors	Masters	PhD	Total
Female	60	54	46	41	201
Male	40	44	53	57	194
Total	100	98	99	98	395

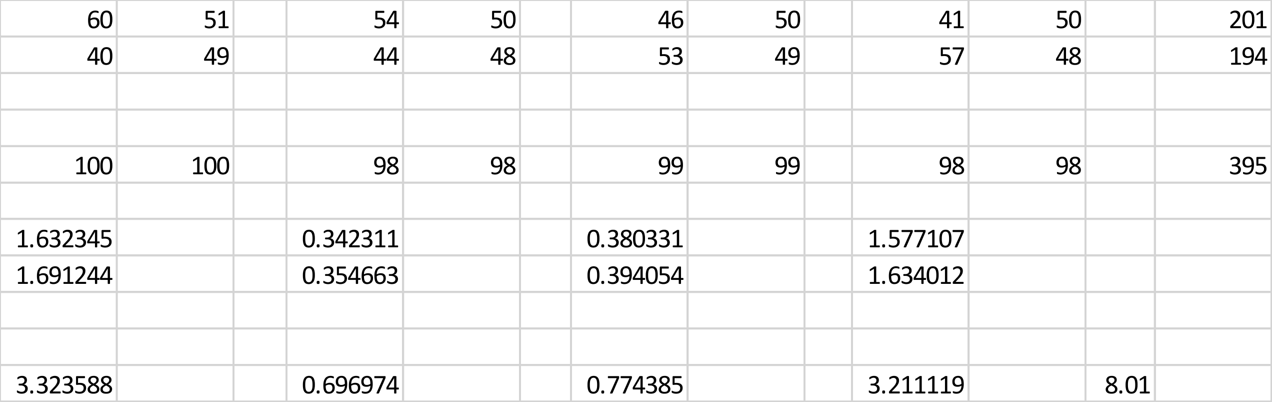
Are gender and education level dependent at 5% level of significance? In other words, given the data collected above, is there a relationship between the gender of an individual and the level of education that they have obtained?

Note: the critical value for χ^2 with 3 degrees of freedom is 7.8147 (at 5% significance), i.e., in R:

```
qchisq(0.95, df=3) = 7.8147
```

$$\chi^2_{df} = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

	High School		Bachelors		Masters		PhD		Total
Female	60		54		46		41		201
Male	40		44		53		57		194
Total	100		98		99		98		395



60	51		54	50		46	50		41	50		201
40	49		44	48		53	49		57	48		194
100	100		98	98		99	99		98	98		395
1.632345			0.342311			0.380331			1.577107			
1.691244			0.354663			0.394054			1.634012			
3.323588			0.696974			0.774385			3.211119		8.01	

H_0: Education level is independent of gender
H_A: Education level is dependent on gender
Reject Null Hypothesis, since 8.01 > 7.81 (i.e., calculated value exceeds critical value)