Indian Statistical Institute, Bangalore
B. Math II, First Semester, 2020-21
Final Examination, Statistics I

21.12.20 **Maximum Score 100** **Duration: 3.5 Hours**

Students are allowed to consult the book Statistics by McClave and Sincich.
Values from the normal distribution qnorm(0.9)=1.281552, qnorm(0.995)=2.575829, qnorm(0.95)=1.644854
Instructor: Rituparna Sen rsen@isibang.ac.in ritupar.sen@gmail.com 9176620249

1. (10+2+3) Let $X_1, X_2 \cdots X_n$ be a random sample from Beta distribution with pdf given by

$$f(x \mid a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\beta(a,b)} \quad \text{if} \quad x \in (0,1),$$

   where $a, b \in \mathbb{R}^+$ are both unknown.

   (a) Obtain the method of moments (MoM) estimator for a.

   (b) Is this the only MoM estimator, or is it possible to obtain other estimators based on the method of moments?

   (c) Is the estimator consistent? Justify your answer.

2. (5+5+5) Let $U \sim Unif(0,1)$ and $Y = -\lambda \log(U)$. Let X be the integer part of Y.

   (a) Find the distribution of Y.

   (b) Find the distribution of X.

   (c) How will you use the above results to generate a random number from the negative binomial distribution with parameters r and p, given a uniform(0,1) random number?

3. (5+5) Answer any two of the following questions related to the class presentations.

   (a) What is likert scale? What is divergent flow chart and when to use it?

   (b) Explain the problem of collinearity in multiple regression. How is principal component analysis useful in this situation?

   (c) How can you test the assumptions of simple linear regression model through diagnostic plots?

4. (5+2+3+5)From a bivariate dataset $(x_1, y_1), \cdots, (x_n, y_n)$, we obtain two least squares regression lines 2y=3x+5 and 3y=5x+7. One line is for the regression of Y on X and the other is for the regression of X on Y.

   (a) What are the sample means of X and Y?

   (b) Is the correlation coefficient between X and Y positive or negative? Why?

   (c) What is the value of the correlation coefficient between X and Y?

   (d) Which one of the two lines is the regression line of Y on X? Explain.

5. (2+5+5+5+3) Proportion of faulty computers built by Byte Computer Corporation is 0.15. In an attempt to lower the defective rate, the owner ordered some changes made in the assembly process. After the changes were put into effect, a random sample of 42 computers were tested revealing a total of 4 defective computers. Using $\alpha = 0.1$ perform the appropriate test of hypothesis to determine whether the proportion of defective computers has been lowered. In particular, answer the following questions.

   (a) State the null and alternative hypotheses.

   (b) State the test statistic and find its distribution under the null hypothesis.

   (c) Is the distribution in the previous part exact or approximate? In the latter case, why does the approximation work?

   (d) Find the critical region and compute the value of the test statistic.

   (e) Is the null hypothesis rejected? What is the conclusion that the owner can draw regarding the defective rate of computers?

6. (5+3+2) It is of interest to know if the average time it takes police to reach the scene of an accident differs from that of an ambulance to reach the same accident. The summary statistics listed below is obtained where time is in minutes.

	Sample size	Mean	Variance
Police	60	4.2	0.08
Ambulance	55	4.5	0.10

   (a) Estimate the difference in the average times between the police and the ambulance using a 99% confidence interval.

   (b) State the assumptions.

   (c) Can we conclude that the average times taken to reach the scene of accident by police and ambulance are different?

7. (4+4) The iris data consists of 4 characters (sepal length, sepal width, petal length, petal width) measured on 50 flowers from each of 3 species (setosa, versicolor, virginica). We run the following command in R.

```
summary(aov(formula = Petal.Width ~ Species, data = iris))
```

   (a) Complete the table of output.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species		80.41			<2e-16
Residuals			0.04		

   (b) Carry out the ANOVA test using the above output stating the null and alternative hypotheses, assumptions and conclusions.

8. (7) Explain what the following R code and output is doing. State the model, hypotheses, data, assumptions, test statistic, its distribution and conclusion.

```
> x<-c(45,92,287,98)
> chisq.test(x,p=c(0.08,0.17,0.55,0.2))
Chi-squared test for given probabilities
data:  x
X-squared = 0.76351, df = 3, p-value = 0.8582
> qchisq(.99,3)
[1] 11.34487
```