Indian Statistical Institute
Mid Sem Exam
Linear Models
B.Stat. III-2018-2019

Date: 03. 09. 2018 | Full Marks. 50 | Time 2:30 - 4:30

1. Consider the linear model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a vector of random errors with $N(\mathbf{0}, \sigma^2 I)$. Obtain the distribution of the sum of squares due to error as well as the sum of squares due to regression and show that these two sums of squares are independent. [12]

2. Consider the following linear model:

$$
\begin{aligned}
Y_1 &= \theta_1 + \theta_2 + \epsilon_1 \\
Y_2 &= \theta_2 + \theta_3 + \epsilon_2 \\
Y_3 &= \theta_2 + \theta_4 + \epsilon_3 \\
Y_4 &= \theta_4 + \theta_5 + \epsilon_4 \\
Y_5 &= \theta_3 + \theta_6 + \epsilon_5 \\
Y_6 &= \theta_4 + \theta_6 + \epsilon_6
\end{aligned}
$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_6)'$, with $E(\boldsymbol{\epsilon}) = 0$, and $V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$.

(a) Find the necessary and sufficient condition for estimablity of a linear parametric function $c'\theta$.

(b) Find four unbiased estimators of $\theta_3 - \theta_4$.

(c) Find the BLUE of $\theta_3 - \theta_4$.

(d) Find the sum of squares due to error under the above model.

(e) Find the appropriate test statistic to test $H_0 : \theta_3 - \theta_4 = 0$ and its distribution under the null hypothesis.

[4+ 6+ 4 + 4+ 6= 24]

3. Consider the usual multiple linear regression model :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, ..., 6.$$

with the design matrix X as

$$
X = \begin{vmatrix} 1 & 0 & 1 & 2 \\ 1 & 0 & 2 & 1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 0 & 2 \end{vmatrix}
$$

Obtain in terms of $y_i's$, the sum of squares due to the hypothesis $H_0 : \beta_1 = 2$. [14]

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination: 2018-2019

B. Stat. (Hons.) 3rd Year. 1st Semester

Parametric Inference

Date: September 05, 2018 Maximum Marks: 40 Duration: 2 hours

- Answer all the questions.
- **You should present all your arguments while answering a question.**
- You must state clearly any result stated (and proved) in class you may need in order to answer a particular question.

---

1. Let $X = (\min(T, C), I(T \leq C))$ where T, C are independent with

$$P(T = j) = p(j),\ j = 0, 1, 2, 3, \text{ and } P(C = j) = r(j),\ j = 0, 1, 2, 3,$$

where $(p, r) \in \{(p, r) : p(j) > 0, r(j) > 0, 0 \leq j \leq 3, \sum_{j=0}^{3} p_j = 1, \sum_{j=0}^{3} r_j = 1\}$. Suppose $X_1, \ldots, X_n$ are observed i.i.d. according to the distribution of X. Show that $\{p(j) : j = 0, 1, 2, 3\}$, $\{r(j) : j = 0, 1, 2, 3\}$ are identifiable. [10]

2. Suppose $Y_1, \ldots, Y_n$ are independent with $Y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$, $i = 1, \ldots, n$, where the quantities $x_1, \ldots, x_n$ are fixed (non-random) and are **not** all equal. Let $\boldsymbol{\theta} \equiv (\beta_0, \beta_1, \sigma)$, $\beta_0, \beta_1 \in \mathbb{R}, \sigma > 0$. Find a minimal sufficient statistic for $\boldsymbol{\theta}$. [10]

3. Consider the linear model $(\boldsymbol{Y}, \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{X}$ is an $n \times p$ matrix with entries 0, 1 and -1. Assume that $\rho(\boldsymbol{X}) = p$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_p)^{\mathrm{T}}$ be the BLUE of $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^{\mathrm{T}}$. Show that $\mathrm{Var}(\widehat{\beta}_j) \geq \sigma^2/n$, $j = 1, \ldots, p$. [8]

4. Suppose $X_1, \ldots, X_n$ are i.i.d. with pdf given by

$$f(x|\theta) = \frac{1}{2\theta} \exp\left(-\frac{|x|}{\theta}\right), \quad -\infty < x < \infty, \quad \theta > 0.$$

Find the UMVUE of $\psi(\theta) := \theta^2$. [12]

Indian Statistical Institute

Mid-semester of First Semester Examination: 2018-19

Course Name: BSTAT III

Subject Name: Economic and Official Statistics

Date: 6 September, 2018 Maximum Marks: 30

Duration : 1 hour

Answer any one question from Group A and any two question from Group B

Group A

1. State and Explain how Optimal demand for a commodity is determined? 10

2. State the forms of Tornqvist's demand function in case of Necessities, Luxuries, relative luxuries and inferior goods. 2.5*4= 10

Group B

1. State the main features of Consumer Expenditure Survey undertaken by NSSO with special reference to 68^{th} round. 10

2. Provide an overview of the Sample Registration System in India. 10

3. Mention in brief the instruments of monetary policy of the Reserve Bank of India. 10

Indian Statistical Institute

Mid-semester of First Semester Examination: 2018-19

Course Name: BSTAT III

Subject Name: Economic and Official Statistics

Date: 6 September, 2018 **Maximum Marks: 30**

Duration : 1 hour

Answer any one question from Group A and any two question from Group B

Group A

1. State and Explain how Optimal demand for a commodity is determined? 10

2. State the forms of Tornqvist's demand function in case of Necessities, Luxuries, relative luxuries and inferior goods. 2.5*4= 10

Group B

1. State the main features of Consumer Expenditure Survey undertaken by NSSO with special reference to 68th round. 10

2. Provide an overview of the Sample Registration System in India. 10

3. Mention in brief the instruments of monetary policy of the Reserve Bank of India. 10

INDIAN STATISTICAL INSTITUTE

Mid-Semestral Examination : 2018-19

Course Name : B.Stat. 3rd Year
Subject Name : Sample Surveys
Date : Sept 07, 2018 **Duration : 2 hrs.**
Total Marks : 40 (Written test = 30 + Assignment = 10)

Answer Q4 and any two from Q1 to Q3. Each question carries marks 10.

Notations are as usual.

1. (a) In SRSWOR of sample size n out of N population units, for two variables of interest y and x, derive $Cov(\bar{y}, \bar{x})$ and an unbiased estimator of it.

   (b) A survey is to be made to estimate the prevalence rate of TB in Kolkata (with an estimated population of 11.5 million). The proportion of persons affected by TB is to be estimated with a maximum discrepancy of 0.001 at the 95% probability level. What sample size of a SRSWOR is needed?

   If you are given an additional information that the proportion of people affected by TB cannot exceed 0.005, re-determine the sample size?

2. (a) Write down the unbiased estimator for population mean in a stratified random sampling with SRSWOR for within stratum sampling and the variance of that. Also write down the expression of the variance estimator.

   State with justification, when a stratified random sampling will perform better than usual SRSWOR.

   (b) Derive Neyman's optimum allocation and Bowley's proportional allocation formulae for startified random sampling.

3. (a) Show that the variance of the linear systematic sample mean is

$$\frac{\sigma^2}{n}\left[1 + (n-1)\rho_c\right],$$

   where σ^2 is the population variance and ρ_c is the intraclass correlation coefficient.

   (b) Show with proper proof how the units in the population should be arranged in order that systematic sampling will be much more efficient than SRSWOR.

4. For a certain study carried out in Ajmer division of Rajasthan during 1980-81, a **two stage cluster sampling** scheme was adopted. In first stage, **4** tehsils were selected by **SRSWR** from the **12** tehsils of Ajmer Division. In second stage, from each selected tehsil a certain number of villages was selected by **SRSWOR**. Finally, the sheep population in each selected village was counted. The necessary data are presented below. Also, given that the total number of villages in the entire Ajmer Division is **1488**.

(i) Estimate the total sheep population in the entire Ajmer Division and obtain an estimate of the standard error.

(ii) Also estimate the mean sheep population per village in the entire Ajmer Division along with the standard error.

Selected Tehsil	Total No. of villages	No.of sample villages	Sheep population in the selected villages
Behrar	102	10	266, 890, 311, 46, 174, 31, 17, 186, 224, 31
Bairath	105	12	129, 57, 64, 11, 163, 77, 278, 50, 26, 127, 252, 194
Kishangarh	200	16	247, 622, 225, 278, 181, 132, 659, 403, 281, 236, 595, 265, 431, 190, 348, 232
Bansur	85	9	347, 362, 34, 11, 133, 36, 34, 61, 249

INDIAN STATISTICAL INSTITUTE
Semestral Exam
Linear Statistical Models
B.Stat. III-2018-2019
Full Marks. 100
Answer Question 5 or 6 and all other questions

Date: 12. 11. 2018 Time 2:30 p.m-6.00p.m

1. Consider the Gauss Markov set up ($\underline{Y}$, $X\underline{\beta}$, σ^2 I). Let H_0 : $C\underline{\beta} = \underline{d}$ be a testable hypothesis.
   (a) Write down the test statistic derived from the confidence ellipsoid of $\underline{\psi}=C\underline{\beta}$ and also the likelihood ratio test statistic for testing H_0 (no derivation is needed).
   (b) Show that these two test statistics are equivalent. [(3+3) + 6 = 12]

2. Consider the following one way fixed effects model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, a, \quad j = 1, \ldots, n_i,$$

   where ϵ_{ij} 's are independent and identically distributed as $N(0, \sigma^2 \text{ I })$, $n_1 = n_2 = n_3 = 4, n_4 = 5$.
   (a) Split the between group sum of squares into the sums of squares due to appropriate contrasts of α_i's, each carrying one degree of freedom. Justify why the contrasts chosen by you serve the purpose.
   (b) Develop Scheffe's method of multiple comparison procedure in this context and apply it to the contrasts (i) $\alpha_1 + \alpha_2 + \alpha_3 - 3\alpha_4$ and (ii) $\alpha_1 - 2\alpha_2 + \alpha_3$.

   [6 + 6 = 12]

3. The effective life (in hours) of batteries is thought to be influenced by material type (1, 2 or 3) and operating temperature: Low (-10°C), Medium (20°C) or High (45°C). An experiment is run with twelve batteries from each of the material type and four batteries from each material are randomly allocated to each temperature level.
   (a) Suggest a suitable linear model for the experiment.
   (b) Write down the explicit form of the design matrix and hence for any linear parametric function of the model parameters, derive the necessary and sufficient condition on the coefficients of the model parameters such the parametric function is estimable.
   (c) Logically justify how you can frame a hypothesis to test for the presence of interaction of material type and operating temperature? Derive explicitly an appropriate test statistic for testing the hypothesis.
   (d) Suggest a hypothesis and a test statistic for testing 'no difference in material effect' (no derivation is required).

   [2+ (3+5)+(4+11) + (3+2) =30]

4. In an experiment machines in an assembly process are evaluated for assembly times. There are three machines with two configurations unique to each of the machines. Three power levels are of interest. It turns out that each machine with any of its configuration can be operated at each of the three power levels twice.

   (a) Suggest a suitable model for this experiment.

   (b) Write down the complete ANOVA table showing the expected mean squares explaining the notations you use.

[4+16= 20]

5. (a) Suppose that an experimenter is studying the effects of three different formulations(denoted by F1, F2 and F3) of a rocket propel used in aircrew escape systems on the observed burning rate. Furthermore the formulations are prepared by four operators (denoted by O1, O2, O3 and O4) whose skill and experience differ. The table below shows the number of times each formulation is being handled by the operators.

	O1	O2	O3	O4
F1	1	1	1	0
F2	1	1	0	0
F3	0	0	1	1

   i. Is it possible to test the equality of the formulation effects? Justify your answer.

   ii. Do you think that the adjusted and unadjusted sum of squares due to the formulation effects will be the same in this scenario? Justify your answer.

   (b) A study is performed to determine if there is a difference in the strength(y) of a monofilament fibre produced by three different machines. It is expected that the strength of the fibre is affected by its thickness(x). The data are given below.

Machine1		*Machine2*		*Machine3*	
y	x	y	x	y	x
36	20	40	22	35	21
41	25	48	28	37	23
39	24	39	22	42	26
42	25	45	30	34	21
49	32	44	28	32	15
—	—	—	—	—	—
207	126	216	130	180	106

Suggest a suitable model to analyse the above data clearly specifying all the assumptions you need. Compute the error sum of squares with appropriate degrees of freedom. Frame the hypothesis to test that 'there is no effect of thickness on the strength of fibre' and compute the sum of squares due to this hypothesis with appropriate degrees of freedom.

[(4+6)+(2+(9+1)+1+3) =26]

6. (a) Consider a two factor random effects model with interaction. Write down the model explicitly with all assumptions. What are the entries of the complete ANOVA table (no derivation is required)? How can the variance components be estimated? Specify the hypotheses of interest and indicate the corresponding test statistics.

(b) For the following data on fabric strength where four looms are selected at random form the collection of all looms in the textile company and for each loom four strength measurements are taken, assume a single factor random effects model. Obtain a 95 % confidence interval of the proportion of the variance of any observation which is the result of variation between the looms.

Looms	*Observations*				
	1	2	3	4	$y_{i.}$
1	98	97	99	96	390
2	91	90	93	92	366
3	96	95	97	95	383
4	95	96	99	98	388
					1527

[(2+6+4+4)+10=26]

See the back page for F table

IV. Percentage Points of the F Distribution (*continued*)

$F_{0.025,\nu_1,\nu_2}$

Degrees of Freedom for the Numerator (ν_1)

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	1.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Degrees of Freedom for the Denominator (ν_2)

INDIAN STATISTICAL INSTITTUE

First Semestral Examination 2018-19

B.Stat (Hons.) 3rd Year

Subject: Economic & Official Statistics and Demography

Group B: Demography & Economic Statistics

Date: 14.11.2018 | Maximum Marks: 50 | Duration: 2 hours

*(**Instruction**: Standard notation are followed.)*

(Answer any five questions.)

1(a) Define the following with full explanation of all symbols:

i) Order specific fertility rate

ii) Order age specific fertility rate

iii) Parity specific fertility rate

iv) Age parity specific fertility rate

(b) Establish a relation between birth order and parity.

[2 + 2 + 2 +2 +2 = 10]

2(a) Derive the general relation between m-type and q-type mortality rates. How does Greville's relation differ from this relation and how is it more advantageous in the construction of a life table.

(b) Show that the crude birth rate in a stationary population corresponding to a life table is equal to ($1/e_0$) where e_0 is the life expectation at birth.

(c) If the crude birth rate in a country remains constant over a number of years, but the general fertility rate increases steadily, what does this tell you about the country's population? [3 + 2 + 2 + 3 = 10]

3. Describe a method of evaluating birth and death registration by using age distribution and child survivorship data. [10]

4. What is the purpose of El-Badry's procedure? Explain how the purpose is served. All symbols must be defined clearly. [1 + 9 = 10]

5. The table below gives the parity progression ratios for a number of recent birth cohorts in a country.

(i) Assuming that no woman in any of these birth cohorts had a fifth child, calculate

(a) the proportion of women in each birth cohort who had exactly 0, 1, 2, 3 and 4 children,

(b) the total fertility rate for women in each birth cohort,

(ii) Comment on your results.

Calendar years of birth	Parity Progression Ratios			
	0 -1	1 – 2	2 – 3	3 – 4
1931-33	0.861	0.804	0.555	0.518
1934-36	0.885	0.828	0.555	0.489
1937-39	0.886	0.847	0.543	0.455
1940-42	0.890	0.857	0.516	0.416
1943-45	0.892	0.854	0.458	0.378
1946-48	0.885	0.849	0.418	0.333

[(6 + 3) + 1 = 10]

6. Write down the important properties of the Lorenz curve with all used symbols fully explained. Show how Lorenz ratio is related to the Gini Mean Difference.

(5+5 = 10)

==END==

First Semester Examination 2018-19

B. STAT (Hons) III year

Economic and Official Statistics

Maximum Marks: 50 Duration: 2hours

Group A: Economic Statistics

Part I: ***Answer any three of the following questions***

1. Elucidate the problem of pooling cross section and time series data in Engel Curve Analysis. 10

2. a) Explain the notion of Homogeneous Production functions b) Show that the elasticity of CES production function is constant over a given range of values. 3+7=10

3. Do you agree that Laspeyres' and Paasche's Index Numbers **do not provide** any bound each for the true index number of prices? Justify. 10

4. State the various steps for the construction of Cost of Living Index following the Aggregate Expenditure Method and Method of Weighted relatives. 10

5. State and explain any four statistical issues in Engel Curve Analysis. 10

Part II: ***Answer any two of the following questions***

1. Describe how Human Development Index is computed by UNDP through creation of dimension indices. 10

2. Mention in brief the sampling design and the method for estimation of crop yields as adopted by Ministry of Agriculture, Government of India. 10

3. State and explain the mathematical expressions on input relations, output relations, industry cost equations used by CSO ,Govt of India for the construction of Input-output table. 10

4. State how the featured indictors for Environment (any ten) are measured by World Bank. 10

INDIAN STATISTICAL INSTITUTE

Sem Examination

Bachelor of Statistics(Hons.) 3rd year 2017-18 (Semester-I)

Design and Analysis of Algorithms

Date: **November 19, 2018** Maximum Marks:**100** Duration: **3 hours 30 minutes**

Note: The question paper carries a total of **120** marks. You can answer as much as you can, but the maximum ou can score is **100**.

You need to prove the correctness of your algorithms.

1. Find the asymptotic behaviour of the function $T(n)$ defined by the recurrence relation $T(n) = T(\frac{n}{2}) + T(\lfloor\sqrt{n}\rfloor) + n$, $T(1) = 1$, $T(2) = 2$, $n = 2^k$ for some integer ≥ 2.

[8]

2. Let B be a binary tree. The distance between two nodes in B is the length of the path connecting these two nodes. The diameter of B is the maximal distance over all pairs of nodes. Design a linear time algorithm to find the diameter of B.

[10+5=15]

3. (a) Show that a comparison based sorting algorithm requires $\Omega(n \log n)$ comparisons.
   (b) The input is a sequence of integers with many duplications, such that the number of distinct integers in the sequence is $O(\log n)$. Design an algorithm to sort such sequences using at most $O(n \log \log n)$ comparisons. Why is the lower bound of $\Omega(n \log n)$ not satisfied here?

[7+((7+3)+3)=20]

4. Write down Kruskal's algorithm for finding the Minimum Cost Spanning Tree of an undirected connected weighted graph. Describe how it may be implemented efficiently. Derive the worst-case time complexity of your implementation.

[(5+6)+5+5=21]

5. Given an array, an element is called a peak element, if it is not less than any of its immediate neighbours. Prove that every array has at least one peak element. Write a sub-linear time algorithm to find a peak element in a given array. Derive the time complexity of your algorithm.

[4+(7+3)+4=18]

6. Let G be an undirected graph such that each vertex has even degree. Design a linear-time algorithm to direct the edges of G such that for each vertex, the indegree is equal to the outdegree. Derive the time complexity of your algorithm.

[(7+3)+5=15]

7. (a) Define polynomial time reduction. When do you call a problem NP-complete?
   (b) The 3-coloring problem tries to find if a given undirected graph is 3-colorable or not. Show that the 3-coloring problem is NP-complete.

[(4+4)+15=23]

INDIAN STATISTICAL INSTITUTE

First Semester Examination : 2018-19

Course Name : B.Stat. 3rd Year
Subject Name : Sample Surveys
Date : 22.11.18, 2018 Total Marks : 60 Duration : 3 hrs.

Answer Q.4 and any 2 from Q.1 to Q.3.
Symbols and notations are as usual.

1. (a) Show how the population total Y of a variable of interest y can be estimated unbiasedly based on a probability proportional to size with replacement (PPSWR) sample of n draws made from N units having the size measure variable x.

   (b) Derive the variance of this estimator and obtain an unbiased estimator of the variance.

   (20)

2. (a) Write down the form of Horvitz and Thompson's (HT) estimator for the population total Y of a variable of interest y and show that it is unbiased. Clearly state the condition which should be satisfied for using this estimator.

   (b) Derive the variance of this estimator and an unbiased estimator of the variance. Also clearly state the condition which should be satisfied for using this variance estimator.

   (20)

3. (a) For estimating the population ratio $R = \frac{Y}{X}$ through SRSWR, derive Hartley and Ross's unbiased estimator. In addition, obtain an unbiased estimator for Y utilizing Hartley and Ross's ratio estimator.

   (b) Explain how Politz and Simmon's 'at-home-probabilities' technique can be utilized in the non-response situation to estimate the population mean $\bar{Y}$ by SRSWR scheme. Also obtain an estimator for the variance of this estimator.

   (20)

4. A garment manufacturer has **N = 90** plants located throughout the United States and wants to estimate the average number of hours that the sewing machines were down for repairs in the past months. Because the plants are widely scattered, and each plant contains many machines, and checking the repair record for each machine would be time-consuming, the manufacturer decides to use a **stratified two-stage cluster sampling**. For this the entire USA is divided into two starta as West and East regions, and each plant is considered as a cluster of machines. Two-stage cluster sampling is used in each stratum. Sufficient time and money are available to sample $n_1 = 5$ and $n_2 = 5$ plants by **PPSWR** with the area size of the plants as size measure respectively from each of the west and east regions. And also from every selected plant, approximately 20% of the machines are selected by **SRSWOR**. The resulting data are given in the table below. And it is known that the total number of machines in all plants in entire USA is $M_0 = \mathbf{4500}$.

Based on the following data, give an estimate of the **average downtime per machine for the entire USA**. And also obtain its standard error ($\widehat{SE}$) and the relative standard error ($RSE = \frac{\widehat{SE}}{|\hat{Y}|}$) as a percentage.

(20)

Downtime in hours of the sample machines

Stratum	Sample plant	Inverse of probability ($1/p_i$)	Total no. of Machines (M_i)	No. of sample machines (m_i)	Downtime of sample machines in hours
1. West	1	440.21	50	10	5, 7, 9, 0, 11, 2, 8, 4, 3, 5
	2	660.43	65	13	4, 3, 7, 2, 11, 0, 1, 9, 4, 3, 2, 1, 5
	3	31.50	45	9	5, 6, 4, 11, 12, 0, 1, 8, 4
	4	113.38	48	10	6, 4, 0, 1, 0, 9, 8, 4, 6, 10
	5	21.00	52	10	11, 4, 3, 1, 0, 2, 8, 6, 5, 3
2. East	1	67.68	58	12	12, 11, 3, 4, 2, 0, 0, 1, 4, 3, 2, 4
	2	339.14	42	8	3, 7, 6, 7, 8, 4, 3, 2
	3	100.00	66	13	3, 6, 4, 3, 2, 2, 8, 4, 0, 4, 5, 6, 3
	4	68.07	40	8	6, 4, 7, 3, 9, 1, 4, 5
	5	24.76	56	11	6, 7, 5, 10, 11, 2, 1, 4, 0, 5, 4

INDIAN STATISTICAL INSTITUTE

First Semester Examination: 2018–2019
B. Stat. (Hons.) 3rd Year. 1st Semester
Parametric Inference

November 26, 2018 Maximum Marks: 60 Duration: 3 hours

- This question paper carries 65 points. Answer as much as you can. However, the maximum you can score is 60.
- You should provide as much details as possible while answering a question. You must state clearly any result stated (and proved) in class you may need in order to answer a particular question.

---

1. Suppose $X_1, \ldots, X_n$ $(n \geq 2)$ are i.i.d. $\text{Bin}(1, \theta)$, $0 < \theta < 1$.

(a) Find the UMVUE of $\psi(\theta) := \theta(1 - \theta)$.

(b) Decide, with adequate justification, if the variance of the estimator in (a) attains the corresponding Cramér-Rao lower bound for some $\theta \in (0, 1)$. [6+6=12]

2. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ $(n \geq 2)$ be a random sample from the uniform distribution $\text{U}(\theta, \theta + 1)$, $\theta \in \mathbb{R}$. We wish to test "$H_0 : \theta \leq 0$" *against* "$H_1 : \theta > 0$" at level $\alpha \in (0, 1)$. Show that there exists a UMP test which rejects H_0 when $\min(X_1, \ldots, X_n) > C(n, \alpha)$ or $\max(X_1, \ldots, X_n) > 1$ for suitable $C(n, \alpha)$. [12]

3. A random sample $X_1, \ldots, X_n$, is drawn from a Pareto population with pdf

$$f(x|\theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}}, \quad x \geq \nu, \quad \theta > 0, \nu > 0.$$

(a) Show that the LRT of "$H_0 : \theta = 1$, ν *unknown*" *against* "$H_1 : \theta \neq 1$, ν *unknown*" has critical region of the form $\{\boldsymbol{x} : T(\boldsymbol{x}) \leq c_1 \text{ or } T(\boldsymbol{x}) \geq c_2\}$, where $0 < c_1 < c_2$ and

$$T(\boldsymbol{x}) := \sum_{i=1}^{n} \log(x_i) - n \log\left(\min_i x_i\right), \quad \boldsymbol{x} := (x_1, \ldots, x_n).$$

(b) Show that under H_0, $2T(\boldsymbol{X})$ has a chi-square distribution with degrees of freedom to be obtained by you. [8+6=14]

[P.T.O.]

4. Let $X_i = \theta t_i^2 + \epsilon_i$, $i = 1, \ldots, n$, where the ϵ_i's are i.i.d. $\mathrm{N}(0, \sigma_0^2)$ variables, and σ_0 is assumed to be a known quantity. The t_i's are known constants, not all zero.

(a) Using a pivot based on the MLE $\widehat{\theta}_n$ of θ, find, with adequate reasons, the level $1-\alpha$ confidence interval for θ which has smallest length among all level $1-\alpha$ confidence intervals for θ which are based on the pivot.

(b) If $0 \leq t_i \leq 1$, $i = 1, \ldots, n$, but we may otherwise choose the t_i's freely, what values should we use for the t_i's so as to make the interval in (a) as short as possible for a given α? Give reasons. [9+4=13]

5. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(θ), $\theta > 0$. Write $\boldsymbol{X} = (X_1, \ldots, X_n)$.

(a) Construct, with adequate reasons, a family of conjugate priors for θ.

(b) With the prior as in (a), construct, with reasons, the $100(1-\alpha)\%$ highest posterior density (HPD) credible interval for θ. [7+7=14]

INDIAN STATISTICAL INSTITUTE

Backpaper Examination: 2018–2019
B. Stat. (Hons.) 3rd Year. 1st Semester
Parametric Inference

January 21, 2019 Maximum Marks: 100 Duration: 3 hours

- Answer all the questions.
- You should provide as much details as possible while answering a question. You must state clearly any result stated (and proved) in class you may need in order to answer a particular question.

---

1. Suppose $X_1, \ldots, X_n$ are i.i.d. with pdf given by $f(x;\theta) = \frac{1}{2}\exp(-\mid x-\theta \mid)$, $x \in \mathbb{R}$, $\theta \in \mathbb{R}$. Show the set of order statistics $T(X_1, \ldots, X_n) = (X_{(1)}, \ldots, X_{(n)})$ is minimal sufficient for θ but not complete. [12+6=18]

2. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(θ), $\theta > 0$. Let $\psi(\theta) := P_\theta(X_1 = 0) = \exp(-\theta)$. Define $\widehat{\theta}_n = \sum_{i=1}^n X_i/n$.

(a) Show that $\exp(-\widehat{\theta}_n)$ is a biased estimate of $\exp(-\theta)$.

(b) Find the UMVUE of $\psi(\theta)$. [6+12=18]

3. Suppose $X_1, \ldots, X_n$ are i.i.d. exponential variables with location parameter θ, $\theta > 0$. Using a pivot based on the MLE $\widehat{\theta}_n$ of θ, find, with adequate reasons, the level $1-\alpha$ confidence interval for θ which has smallest length among all level $1-\alpha$ confidence intervals for θ which are based on the pivot. [10]

4. Let $X_1, \ldots, X_n$ be the times in months until failure of n similar equipments. If the equipment is subject to wear, a model often used is the one where $X_1, \ldots, X_n$ is a random sample from a Weibull distribution with density $f(x, \lambda) = \lambda c x^{c-1} e^{-\lambda x^c}$, $x > 0$. Here c is a known positive constant and $\lambda > 0$ is the parameter of interest. We wish to test the hypothesis "$H_0 : 1/\lambda \leq 1/\lambda_0$" *against* "$H_1 : 1/\lambda > 1/\lambda_0$". Prove the following.

(a) The critical region of the UMP level-α test is given by $\{\sum_{i=1}^n X_i^c \geq k\}$, where $k := x_{2n}(1-\alpha)/2\lambda_0$, $x_{2n}(1-\alpha)$ being the $(1-\alpha)$-th quantile of the χ^2_{2n} distribution.

(b) The power function of the test in (a) is strictly decreasing in λ. [10+8=18]

[P.T.O.]

5. Suppose that we have two independent random samples: $X_1, \ldots, X_m$ from a Beta$(\theta, 1)$ population and $Y_1, \ldots, Y_n$ from a Beta$(\mu, 1)$ population. Let $\boldsymbol{X} = (X_1, \ldots, X_m)$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$.

(a) Show that the LRT of "$H_0 : \theta = \mu$" *against* "$H_1 : \theta \neq \mu$" is based on the statistic

$$T(\boldsymbol{X}, \boldsymbol{Y}) := \frac{\sum_{i=1}^{m} \log X_i}{\sum_{i=1}^{m} \log X_i + \sum_{j=1}^{n} \log Y_j}.$$

(b) Find the null distribution of $T(\boldsymbol{X}, \boldsymbol{Y})$. [10+8=18]

6. Suppose $X_1, \ldots, X_n$ are i.i.d. Bin$(1, \theta)$, $0 \leq \theta \leq 1$, and the prior distribution Π of Θ is given by a Beta(a, b) distribution. Write $\boldsymbol{X} = (X_1, \ldots, X_n)$.

(a) Find the posterior distribution of Θ given $\boldsymbol{X}$.

(b) Find the Bayes estimator of $\theta(1 - \theta)$ under squared error loss.

(c) Find the $100(1 - \alpha)\%$ highest posterior density (HPD) credible interval for θ.

[6+6+6=18]

INDIAN STATISTICAL INSTITUTE
Back Paper Exam
Linear Models
B.Stat. III-2018-2019
Full Marks. 100
Answer all questions

Date: 21.01.2019 Time 3.5 hour.

1. Consider the Gauss Markov set up ($\underline{Y}$, $X\underline{\beta}$, σ^2 I). Let H_0 : $C\underline{\beta} = \underline{0}$ be a testable hypothesis. Derive the likelihood ratio test statistic for testing H_0 and its distribution. [10+ 8 = 18]

2. The maximum output voltage of a particular type of storage battery is thought to be influenced by the material used in the plates and the temperature in the location at which the battery is installed. An experiment is run in the laboratory with one observation for each of the 4 material and 3 temperature combinations.
   (a) Suggest a suitable linear model for the experiment.
   (b) Write down the explicit form of the design matrix and hence for any linear parametric function of the model parameters, derive the necessary and sufficient condition on the coefficients of the model parameters such that the parametric function is estimable. Exhibit a complete set of independent parametric functions.
   (c) Frame the hypotheses to test 'no difference in material effect' and derive the test statistic. Hence write down the complete ANOVA table clearly explaining your notations and showing the column of expected mean squares.
   (d) Is it possible to test interaction between the material effect and temperature? Justify your answer. [2+ (3+5+3)+(2+8+7) + 2 =32]

3. In an experiment purity of three batches of raw materials, supplied by four different suppliers is being tested. One measurement of purity is taken for each combination of the supplier and the batch of raw materials supplied by him.
   (a) Suggest a suitable model for this experiment.
   (b) What are the hypotheses of interest in this scenario?
   (c) Write down the complete ANOVA table, explaining the notations you use and indicate the appropriate test statistics for the hypotheses you mentioned above.

[3+4+13= 20]

P.T.O

4. In an unbalanced two way fixed effects model without interaction involving factor A and factor B, derive the different sum of squares to be included in the ANOVA table. [18]

5. For the following data on fabric strength where three looms are selected at random form the collection of all looms in the textile company and for each loom three strength measurements are taken, assume a single factor random effects model. Estimate the different variance components in the model.

Looms	*Observations*		
	1	2	3
1	98	97	99
2	91	90	93
3	96	95	97

[12]

INDIAN STATISTICAL INSTITUTE
Mid-Semester Examination 2018-19

B.Stat. - 3rd Year
Design of Experiments

18th February, 2019

Maximum Marks: 60 Time: 3 hours

[Note: Notations are as used in the class. Answer as much as you can. Best of Luck!]

1. What are Uniformity Trials? Which purpose it is used for? Which design is often used in a Uniformity Trial and why? [2+1+2]=5

2. A soap manufacturer wants to find out if there is any significant difference in weight loss due to dissolution among three particular types of their soap when allowed to soak in water for the same length of time. As a statistician working for the manufacturer, your job is to design an appropriate experiment to be performed for the above purpose.

(a) Identify all sources of variation, including treatment factors and their levels, experimental units and blocking or noise factors, if any.

(b) Which design will you suggest for the above experiment? Justify.

(c) Specify the measurements to be made and the experimental procedures.

(d) Do you anticipate any difficulty in performing the actual experiment? What will be your plan to tackle these difficulties?

[4+2+(2+2)+5]=15

3. In a balanced completely randomized design with v treatments, explicitly construct $(v-1)$ mutually orthogonal treatment contrasts, each carrying one degrees of freedom, such that their corresponding sum-of-squares add up to the overall treatment sum-of-squares. [10]

4. Construct, if possible, a general connected block design with $v = 6$ treatments and $b = 3$ blocks of sizes $\boldsymbol{k} = (2, 3, 2)^T$. Justify your answer. [5]

PTO

5. Derive the A-efficient design among block designs with v treatments and b blocks each of size $k = v$. [5]

6. Consider a randomized block design with v treatments and b blocks. A treatment contrast $\mathbf{l}^T\boldsymbol{\tau}$ is estimable if $\mathbf{l}^T\mathbf{1} = 0$.

(a) Derive a simultaneous confidence interval for all estimable treatment contrasts based on their BLUEs.

(b) Write down the ANOVA table for testing $H_0 : \tau_1 = \tau_2 = \cdots , \tau_v$.

(c) Find out the expectations of the mean-squares due to treatments and blocks in the above ANOVA table.

(d) Based on (c), or otherwise, derive the non-null distribution of the F-statistics in (b). Explain how it can be used to determine optimal sample size for the experiment.

[5+5+(3+1)+(2+3)]=20

7. Consider a randomized block design with v treatments and b blocks. If $v < b$ and observations under treatment i in block i are missing for $i = 1, 2, \ldots, v$, is the resulting design still (i) connected and/or (ii) orthogonal and/or (iii) balanced? Justify your answers.

[5+5+2]=12

Date: 19.2.2018
Time: 2 hours

INDIAN STATISTICAL INSTITUTE

Statistical Methods in Genetics
B-Stat (3rd Year) 2018-2019
Final Examination

This paper carries 30 marks. Calculators are allowed.
Answer all questions.

1. Consider the following genotype data at a triallelic locus on 500 randomly chosen individuals in a population:

Genotype	Frequency
AA	182
AB	170
AC	66
BB	48
BC	34

(a) Test whether the locus is in Hardy-Weinberg Equilibrium.
(b) Obtain an approximate 95% confidence interval for the difference in the frequencies of the alleles A and B. [8 + 7]

2. Consider a biallelic X-linked locus with alleles *A* and *a*. If the initial frequency of *A* is 0.6 among males and 0.75 among females, in how many generations will the frequency of *A* be 0.7000122 among males? [7]

3. Consider a dominant disorder controlled by an autosomal biallelic locus. If the prevalence of the disease is 0.19, find the probability that the mating between an uncle and his niece produces an affected offspring? [8]

MID-SEMESTER EXAMINATION - NUMBER THEORY
ISI KOLKATA - B.STAT. III YEAR - 2018-19
DATE : FEBRUARY 19, 2019, TIME : 150 MINS. (2:30-5:00 PM)
MAXIMUM MARKS : 30

Instructions: Answer as much as you can. Course materials and calculators are not allowed.

(1) Show that the harmonic sum $H_n = \sum_{k=1}^{n} \frac{1}{k}$ is not an integer for $n \geq 2$. 3 marks

(2) Determine the values of positive integer n, for which $d(n)$ is odd. 2 marks

(3) For a multiplicative function h, can we have two non-zero arithmetic functions f, g which are not multiplicative such that $f * g = h$? Justify. 3 marks

(4) Prove that as $x \to \infty$,

$$\sum_{n \leq x} \frac{\mu(n)}{n} \left[\frac{x}{n}\right] = cx + O(\log x),$$

for some absolute constant c. 3 marks

(5) Let $d_3(n)$ denote the number of ways of writing n as a product of 3 positive integers. Then prove that as $x \to \infty$,

$$\sum_{n \leq x} d_3(n) = \frac{x(\log x)^2}{2} + O(x \log x).$$

Further, refine this estimate to obtain a secondary main term of the form $cx \log x$ for some constant c and an error term which is $o(x \log x)$. 6(=2+4) marks

(6) Prove that for any integer $a \geq 2$, the sum $\sum_{n \geq 0} \frac{1}{a^{n!}}$ is a Liouville number. 3 marks

(7) Prove that the set of Liouville numbers $\mathcal{L}$ is dense in $\mathbb{R}$. 4 marks

(8) Prove that the reciprocal of a Liouville number is a Liouville number. 4 marks

(9) Let $\xi \in \mathbb{R} \setminus \mathbb{Q}$ and for $n \geq 0$, let p_n/q_n be its n-th convergent and a_n be its n-th partial quotient. If we set $p_{-1} = 1, q_{-1} = 0$ and $p_0 = a_0, q_0 = 1$, prove that for $n \geq 0$,

$$q_n p_{n-1} - p_n q_{n-1} = (-1)^n.$$

Further, for $n \geq 1$,

$$q_n p_{n-2} - p_n q_{n-2} = (-1)^{n-1} a_n.$$

4 marks

(10) Let ξ be an irrational number and its continued fraction expansion be $[a_0; a_1, a_2, a_3, \ldots]$. Prove that after finitely many terms the partial quotients of $-\xi$ are $a_3, a_4, \ldots$.
(Hint: First treat the case when $a_1 \geq 2$. Then show that if $a_1 = 1$, in that case $-\xi = [-a_0 - 1; a_2 + 1, a_3, \ldots]$.) 5 marks

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2018 – 19

B.Stat 3rd Year

Random Graphs

Date: 19/02/2019 **Maximum Marks: 40** **Duration: 2 hr**

Anybody caught using unfair means will immediately get 0. Please try to explain every step. Only class notes are allowed in the exam.

(1) Show that the threshold for the existence of a cycle in $ER_n(p)$ is $1/n$. [10 marks]

(2) (a) For a branching process with i.i.d. offspring X with mean $\mu < 1$, show that

$$\mathsf{E}[T_p] = \frac{1}{1-\mu}$$

where T_p is the total progeny size.

(b) Let $|C(1)|$ be the size of connected cluster of 1 in $ER_n(\lambda/n)$ with $\lambda < 1$, show that $\mathsf{E}[|C(1)|] \leq 1/(1-\lambda)$.

[5+5=10 marks]

(3) Consider $ER_n(\lambda/n)$ with $\lambda > 1$. Show that

$$\mathsf{P}(1 \leftrightarrow 2) = \zeta_\lambda^2(1 + o(1)).$$

Here ζ_λ is the survival probability of a Galton-Watson tree with $Poi(\lambda)$ offspring. [10 marks]

(4) Fix $\lambda = a \log n$ and let M denote the number of isolated edges in $ER_n(\lambda/n)$, that is, the edges that are occupied but for which the vertices at either end have no other neighbours.

(a) Show that $M \xrightarrow{P} \infty$ when $a > 1/2$.

(b) Show that $M \xrightarrow{P} 0$ when $a < 1/2$.

[5+5=10 marks]

Date: 19.2.2018
Time: 2 hours

INDIAN STATISTICAL INSTITUTE

Statistical Methods in Genetics
B-Stat (3rd Year) 2018-2019
Final Examination

This paper carries 30 marks. Calculators are allowed.
Answer all questions.

1. Consider the following genotype data at a triallelic locus on 500 randomly chosen individuals in a population:

Genotype	Frequency
AA	182
AB	170
AC	66
BB	48
BC	34

(a) Test whether the locus is in Hardy-Weinberg Equilibrium.
(b) Obtain an approximate 95% confidence interval for the difference in the frequencies of the alleles A and B. [8 + 7]

2. Consider a biallelic X-linked locus with alleles A and a. If the initial frequency of A is 0.6 among males and 0.75 among females, in how many generations will the frequency of A be 0.7000122 among males? [7]

3. Consider a dominant disorder controlled by an autosomal biallelic locus. If the prevalence of the disease is 0.19, find the probability that the mating between an uncle and his niece produces an affected offspring? [8]

INDIAN STATISTICAL INSTITUTE

Bachelor of Statistics (Hons.) Third Year

Special Topics on Algorithm

Date: **February 19, 2019** Maximum Marks:**60** Duration: **2 hours**

Note: The question paper carries a total of **74** marks. You can answer as much as you can, but the maximum you can score is **60**.

You need to prove the correctness of your algorithms.

1. (a) State and prove the Max-Flow-Min-Cut Theorem.

   (b) Show that the Ford-Fulkerson Method may not always work.

   [8+7=15]

2. Suppose you are given an undirected graph $G = (V, E)$ and a positive integer b_v for every vertex $v \in V$. A b-matching is a subset M of edges such that each vertex v is incident to at most b_v edges of M. (The standard matching problem corresponds to the case where $b_v = 1$ for every $v \in V$.) Prove that the problem of computing a maximum-cardinality b-matching reduces to the problem of computing a (standard) maximum-cardinality matching in a bigger graph. Your reduction should run in time polynomial in the size of G and in $\sum_{v \in V} b_v$.

   [12]

3. Encode the maximum flow problem as a linear program. The size of your linear program should be linear in that of the maximum flow instance.

   [12]

4. Write an algorithm to compute the Voronoi Diagram of a set of points on the 2D plane. Establish the time and space complexities of your algorithm.

   [20+5=25]

5. Given an undirected graph $G = (V, E)$ and an integer k, the vertex cover problem tries to determine whether G has a vertex cover containing $\leq k$ vertices. Show that the vertex cover problem is NP-complete.

   [10]

INDIAN STATISTICAL INSTITUTE

Mid-semester Examination : (2018-2019)

B.Stat. 3rd Year

NONPARAMETRIC AND SEQUENTIAL METHODS

Date: 20 February, 2019 Max. Marks: 60 Duration: 2 Hours

1. Let $X_1, \dots, X_n$ be a random sample from a population with unknown continuous distribution function F.

(a) Describe the Kolmogorov-Smirnov test for the hypothesis $H_0 : F = F_0$ (for various possible alternatives) where F_0 is a specified distribution function.

(b) Show that the Kolmogorov-Smirnov statistic D_n^- is distribution free.

(c) Show that the Kolmogorov-Smirnov statistics D_n^+ and D_n^- are identically distributed under H_0. [6+9+8=23]

2. Let $X_1, \dots, X_n$ be a random sample from a population with a continuous distribution F which is symmetric about its unknown median θ. Consider the use of Wilcoxon signed rank statistic T_+ for testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$.

(a) Express T_+ as a weighted sum of two U-statistics.

(b) Find the (non-degenerate) asymptotic distribution of suitably normalized and centred T_+ under H_0.

(c) Find an asymptotic level α $(0 < \alpha < 1)$ test based on T_+ for this problem and show that it is consistent for any alternative under which $P(X_1 + X_2 > 0) > \frac{1}{2}$. [10+9+8=27]

3. Let $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ be two random samples drawn independently from two populations with continuous distribution functions F and G respectively. Assume that $G(x) = F(x - \theta)$ for all x and some θ.

Consider the Mann-Whitney U test of level α for testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Show that this test is unbiased. Also show that this test is of level α for testing the null hypothesis $H : \theta \leq 0$ against $H_1 : \theta > 0$. [10]

Indian Statistical Institute
Statistics Comprehensive
B-III, Midsem

Date: Feb 21, 2019 Duration: 3hrs.

This paper has two groups. Attempt all questions of both groups. The maximum you can score in each group is 50. Use separate answer scripts for the two groups. This is an open book, open note examination. You may use your own calculator. Justify all your steps.

Group A$_{\text{rnab}}$

1. We have 100 values for each of 3 variables, X_1, X_2 and Y. We want to use MLE to fit the multiple regression model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

where $\epsilon \sim N(0, \sigma^2)$ for some unknown $\sigma^2 > 0$. Here the values of X_i's are all nonrandom. What problems (if any) will you face if the covariance matrix of $V(X_1 + X_2)$ is very small? As an alternative to MLE it is suggested that we estimate $\beta_0, ..., \beta_2$ as

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X + \lambda I)^{-1} X' \mathbf{y},$$

for some given $\lambda > 0$. Here X is the design matrix and $\mathbf{y}$ is the vector of Y values. Express the bias vector and covariance matrix of this estimator as a function of λ.

[3+5+7]

2. We have a random variable that takes only the values $1, 2, 3, 4$ with probabilities $p, q, p+q, r$, respectively. Here (p, q, r) is our parameter. The parameter space is $\{(p, q, r) \; : \; 2(p+q) + r = 1, \;\; p, q, r > 0\}$. Based on a single observation, X, from this distribution, construct

   (i) a complete, sufficient statistic.

   (ii) a complete statistic that is not sufficient.

   (iii) a sufficient statistic that is not complete.

   If any of these is/are impossible, prove the impossibility. [5+5+5]

3. A lake contains 2 types of fish. We want to estimate the numbers of fish of both types using the capture-recapture idea. Suggest a method to do so. Clearly write down the steps and formulae. Suggest a method (analytic or numerical) to estimate/approximate the covariance matrix for your estimator.

[5+7]

4. A paint program is to be written for a touchscreen device. For this we need a method to draw a thin smooth curve (1 pixel thickness) by sweeping one finger over the screen. When the thick finger touches the screen many pixels are activated. Thus, we have a cloud of pixels that have been activated during the sweep. The thin curve to be drawn is like a central tendency measure for this cloud. Suggest some technique to find such a curve. What if you are also given the time stamps for each activated pixel (i.e., when each particular pixel was activated)?

[8+5]

[P.T.O.]

Group B

1. Suppose that a batsman scored as given below in 40 innings ($*$ implies a "not out"):

   53, 21, 7*, 0, 105, 97*, 42, 61, 18*, 73, 93, 63*, 41, 28, 49, 51, 21*, 49, 100, 41, 33, 50, 25*, 5, 61, 61*, 43, 26, 17, 19, 4, 0, 10, 101, 40*, 25, 32*, 29, 37*, 65

   (a) What is the usual batting average?

   (b) Can you suggest a better way to define the average score with justification. Compute the average with your notion.

   [15]

2. Consider the following binary sequence:

   0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 1 1 0 0 1 0 0 1 0 0 1 0 0 0 0 1 1 1 1 0 0

   (a) Compute the autocorrelation of lags 1 and 2.

   (b) Write down the frequency distribution of runs. On the basis of your results above can you say whether the sequence is nonrandom or not. Justify your answer.

   [20]

3. A pathological test to identify a particular virus, whose prevalence is about 1 in 10,000, may make a wrong inference with probability p. What should the value of p be, so that both the following error probabilities are less than 0.01 :

   - P(the sample has the virus | test says negative)
   - P(the sample does not have the virus | test says positive)

   Justify your answer.

   [10]

4. Write down your version of "truth", with justification, in the movie ***Roshomon***.

   [10]

INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination: 2018-2019

B. Stat. (Hons.) III Year

Subject: SQC & OR

Full Marks: 100 **Time: 3 hours** **Date of Examination: 22nd Feb 2019**

Books and class-notes are NOT allowed. Calculators are allowed.
Relevant Tables are attached with the question paper.

NOTE: *This paper carries 110 marks. You may answer any part of any question; but the maximum you can score is 100.*

1. A company has two grades of inspectors **1** and **2**, who are to be assigned for a quality control inspection. It is required that at least 2,000 pieces be inspected per 8-hour day. (That is the inspectors work for only 8 hours on a working day.) Grade 1 inspector can check pieces at the rate of 40 per hour with an accuracy of 97%. Grade 2 inspector can check pieces at the rate of 30 per hour with an accuracy of 95%.

   The wage rate of Grade 1 inspector is Rs 5 per hour while that of Grade 2 inspector is Rs 4 per hour. An error made by an inspector costs Rs 3 to the company. There are only nine grade I inspectors and eleven Grade 2 inspectors available in the company.

   The company wishes to assign work to the available inspectors so as to minimize the total cost to the company. Formulate this as an LP model.

   [20]

2. The system $Ax = b, \quad x \geq 0$ is given by

$$x_1 + x_2 - 8x_3 + 3x_4 = 2$$
$$-x_1 + x_2 + x_3 - 2x_4 = 2$$

   Find

   a) a non-basic feasible solution

   b) a basic solution which is not feasible

   c) a BFS which corresponds to more than one basis matrix. Write all the corresponding basis matrices.

   d) a solution which is neither basic nor feasible.

   [2+2+(1+6)+2=13]

3. The starting and current tableaux of a given problem are given below:

Starting Tableau

	y_0	$-x_1$	$-x_2$	$-x_3$
x_0	0	$-A$	-1	$+3$
x_4	6	B	C	D
x_5	1	-1	2	E

Current Tableau

	y_0	$-x_2$	$-x_3$	$-x_4$
x_0	-4	$1/3$	$-H$	I
x_1	F	$2/3$	$2/3$	$1/3$
x_5	3	G	$-1/3$	$1/3$

Find the values of the unknowns A through I. Show your working completely.

[27]

4. a) The diameter of a shaft is subject to statistical control using $\bar{X}$ and R charts. After 30 subgroups of 5 shafts each have been examined, $\sum \bar{X} = 34{,}290$ and $\sum R = 330$. Assume that the quality characteristic is normally distributed.

i. Estimate the mean μ and the standard deviation σ of the process assuming that it is in statistical control.
ii. Calculate control limits for $\bar{X}$ and R charts.
iii. What are the natural tolerances of this process?

b) An item is made in lots of 200 each. The lots are given 100% inspection. The record sheet for the first twenty-five lots inspected showed that a total of 75 items did not conform to specifications.

i. Determine the trial limits for the np chart.
ii. Assume that all points fall within the control limits and are randomly distributed. What is your estimate of the process average fraction non-conforming μ_p?
iii. If this μ_p remains unchanged, what is the probability that the twenty-sixth lot will contain exactly 7 non-conforming units?

[(4+6+4)+(3+2+5)=24]

5. a) Define quality costs. Give their broad categories for a manufacturing organization.
b) Distinguish between ***control limits*** and ***specification limits***.
c) Distinguish between ***engineering quality control*** and ***statistical quality control.***

[(3+2)+4+4=13]

6. Choose the best answer. **You need not copy the statements.**

i. Pick out the appraisal quality cost from the following:
   a. Fees for an outside auditor to audit the quality management system.
   b. Time spent to review customers' drawing before contract.
   c. Time spent in concurrent engineering meetings by a supplier.
   d. Salary of a metrology lab technician who calibrates the instruments.
   e. None of the above.

ii. Which of the following will be considered a failure quality cost?

a. Salaries of personnel testing repaired products.
b. Cost of test equipment.
c. Cost of training workers to achieve production standards
d. Incoming inspection to prevent defective parts coming into stores.
e. All of the above.

iii. Which of the following is not a quality cost:
a. Cost of inspection and test.
b. Cost of routine maintenance of plant and machinery.
c. Cost of routine maintenance of test instruments.
d. Salary of SPC analysts
e. None of the above.

iv. When measurements show a lack of statistical control, the standard error of the average:
a. Is related to the confidence limits.
b. Is a measure of process variability.
c. Is simple to compute.
d. Has no meaning.

v. The PDCA wheel is attributed to
a. Box.
b. Juran.
c. Dodge and Romig.
d. Deming.

vi. Which of the following is **not** a benefit of SPC charting?
a. Charting helps in evaluating of system quality.
b. It helps identify unusual problems that might be fixable.
c. It encourages people to make continual adjustment to processes.
d. Without complete inspection, charting still gives a feel for what is happening.

vii. Who is reputed to have said: *All models are wrong; but some are useful.*
a. P C Mahalanobis
b. G Taguchi
c. G E P Box
d. J M Juran

viii. If in any simple iteration the minimum ratio rule fails, then the LPP has
a. non-degenerate BFS
b. degenerate BFS

c. unbounded solution
d. infeasible solution

ix. In Phase-I of the two phase method an artificial variable turns out to be at positive level in the optimal table of Phase-I, then the LPPP has

a. No feasible solution
b. Unbounded solution
c. Optimal solution
d. Alternative optima

x. For $x_1, x_2 \geq 0$, consider the system

$$x_1 + 2x_2 - x_3 - 2x_4 - 3x_5 = -1$$
$$2x_2 + x_3 + 5x_4 - 3x_5 = -1$$

Its solution $x_1 = 0,\ x_2 = 1, x_3 = 0, x_4 = 0, x_5 = 1$ is

a. A basic solution
b. A basic feasible solution
c. Feasible solution
d. Not a basic feasible solution

xi. The most important parameter in statistical quality control is

a. Mean
b. Range
c. Variance
d. Proportion

xii. Specification limits can be drawn on the

a. $\bar{X} - R$ charts
b. $p -$ chart
c. $c -$ chart
d. None of the above

xiii. If in a LPP , the solution of a variable can be made infinity large without violating the constraints the solution is ______________

a. Infeasible
b. Unbounded
c. Alternative optima exists
d. None of the above

[1 × 13 = 13]

Factors for Constructing Variables Control Chart → 5

Standard Normal Distribution → 6

Appendix VI Factors for Constructing Variables Control Charts

Observations in Sample, n	Chart for Averages: Factors for Control Limits			Chart for Standard Deviations: Factors for Center Line		Chart for Standard Deviations: Factors for Control Limits				Chart for Ranges: Factors for Center Line		Chart for Ranges: Factors for Control Limits				
	A	A_2	A_3	c_4	$1/c_4$	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.575
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.115
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

For $n > 25$

$$A = \frac{3}{\sqrt{n}}, \quad A_3 = \frac{3}{c_4\sqrt{n}}, \quad c_4 \simeq \frac{4(n-1)}{4n-3},$$

$$B_3 = 1 - \frac{3}{c_4\sqrt{2(n-1)}}, \quad B_4 = 1 + \frac{3}{c_4\sqrt{2(n-1)}},$$

$$B_5 = c_4 - \frac{3}{\sqrt{2(n-1)}}, \quad B_6 = c_4 + \frac{3}{\sqrt{2(n-1)}}.$$

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
-0.8	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
-0.7	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
-0.6	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
-0.5	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
-0.4	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
-0.3	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
-0.2	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
-0.1	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
-0.0	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414

INDIAN STATISTICAL INSTITUTE

Semestral Examination, 2nd Semester, 2018-2019

B.Stat. 3rd Year

NONPARAMETRIC AND SEQUENTIAL METHODS

Date: 18 April, 2019 Maximum Marks: 100 Duration: 3 Hours

Answer all questions.

1. Consider Wald's sequential probability ratio test (SPRT) for simple hypotheses with target strength (α, β) in the case of i.i.d. observations where α and β are small positive numbers with $\alpha + \beta < 1$.

(a) Show that Wald's approximations for the boundaries of an SPRT are conservative with respect to error probabilities.

(b) Find approximate expressions for average sample number (ASN) under H_0 using Wald's approximation and Wald's first equation. [12+7=19]

2. Let $X_i, i = 1, 2, \ldots$ be i.i.d $N(\theta, \sigma^2)$ where σ^2 is known. Consider the SPRT for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ $(\theta_1 \neq \theta_0)$ where the boundaries satisfy $0 < B < 1 < A < \infty$. Show, without using Stein's lemma, that the SPRT terminates with probability one under all $\theta \neq \frac{1}{2}(\theta_0 + \theta_1)$. [13]

3. Let $X_1, X_2, \ldots$ be i.i.d $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Describe Stein's two-stage sampling procedure for obtaining a bounded length confidence interval for μ with confidence coefficient $(1 - \alpha)$. Prove the results you state. [14]

4. (a) Describe the concept of Pitman's asymptotic relative efficiency of tests.

(b) Let $X_1, \ldots, X_n$ be i.i.d $N(\mu, 1)$ where μ is unknown. We want to test $H_0 : \mu = 0$ against $H_1 : \mu > 0$. Find the asymptotic relative efficiency of the sign test relative to the UMP test based on the sample mean. Assume that the distribution of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Sign(X_i)$ under $\mu = \delta/\sqrt{n}$ converges to $N(\sqrt{2/\pi}\delta, 1)$.

[6+12=18]

P.T.O

5. Consider a U-statistic U_n for unbiased estimation of $\theta = \theta(F)$ based on a kernel $h(x_1, \ldots, x_m)$ and $n(\geq m)$ i.i.d. observations from a distribution F.

Define projection $\hat{U}_n$ of the U-statistic U_n and find its expression. Show that under suitable conditions $\sqrt{n}(U_n - \hat{U}_n) \xrightarrow{P} 0$ and hence find the asymptotic distribution of $\sqrt{n}(U_n - \theta)$. (Assume that $n\mathrm{Var}(U_n) \to m^2\sigma_1^2$ where σ_1^2 has its usual meaning.) [8+14=22]

6. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two random samples drawn independently from two populations with continuous distribution functions F and G respectively. Assume that $G(x) = F(x - \theta)$ for all x and some θ.

Construct an upper confidence bound for θ with confidence coefficient $(1 - \alpha)$ $(0 < \alpha < 1)$ based on the differences $Y_j - X_i$, $i = 1, \ldots, m$, $j = 1, \ldots, n$. [14]

INDIAN STATISTICAL INSTITUTE
B. Stat. (Hons.) III Year 2018-2019

Second Semester Examination

Subject : SQC & OR

Date : **26th April 2019** Full Marks : **100** Duration : **3 hrs.**

This paper carries 115 marks. You may answer as much as you can; but the maximum you can score is 100.

This is a closed book closed notes examination.
Calculators are allowed during the examination.

1. Consider the following problem $\mathcal{P}$:

$$\mathcal{P}: \text{ Maximize } Z = x_1 + 2x_2 + 3x_3 + 4x_4$$

St

$$x_1 + 2x_2 + 2x_3 + 3x_4 \leq 20$$
$$2x_1 + x_2 + 3x_3 + 2x_4 \leq 20$$

$$x_i \geq 0 \quad i = 1, \cdots, 4$$

Solve $\mathcal{P}$ using the results of Duality Theory. Mention clearly the results that you use.

[15]

2. Two companies X and Y sell two brands of soap. Both the companies advertise using radio, TV, newspaper and the Web. Depending on the cleverness and intensity of the advertisement campaign, each company can capture a portion of the market from the other. The following matrix summarizes the percentage of the market captured or lost by company X.

		Company Y			
		Radio	TV	Newspaper	Web
Company X	Radio	1	2	−1	2
	TV	3	1	2	4
	Newspaper	−1	3	3	1
	Web	−2	2	1	−1

Determine the optimal strategies for both companies and the value of the game.

[15]

3. In a study to isolate both gage repeatability and gage reproducibility, three operators use the same gage to measure 20 parts twice each. The data are given in Table 1. The specification limits are USL= 60, LSL=5.

Table 1: Data for Question No. 3

Part Number	Operator I Measurements		Operator II Measurements		Operator III Measurements	
	1	**2**	**1**	**2**	**1**	**2**
1	21	20	20	20	19	21
2	24	23	24	24	23	24
3	20	21	19	21	20	22
4	27	27	28	26	27	28
5	19	18	19	18	18	21
6	23	21	24	21	23	22
7	22	21	22	24	22	20
8	19	17	18	20	19	18
9	24	23	25	23	24	24
10	25	23	26	25	24	25
11	21	20	20	20	21	20
12	18	19	17	19	18	19
13	23	25	25	25	25	25
14	24	24	23	25	24	25
15	29	30	30	28	31	30
16	26	26	25	26	25	27
17	20	20	19	20	20	20
18	19	21	19	19	21	23
19	25	26	25	24	25	25
20	19	19	18	17	19	17
Average ($\bar{\bar{X}}$)	22.30		22.28		22.60	
Mean Range ($\bar{R}$)	1.00		1.25		1.20	

a) Estimate the gage repeatability, gage reproducibility, total variability and product variability.
b) Find the precision-to-tolerance ratio for this gage. Is it adequate?
c) Do you think training to the operators is called for? Give reasons for your answer.

[8+(2+1)+4=15]

4. An agency has three people in charge of customer services consisting of (i) purchase and sales (ii) documentation, and (iii) accounts. Each person takes a mean time of 10 min to serve a customer. Assume that customers arrive at a rate of 15 per hour.
   a) What is the probability that a customer has to wait for service?

b) What is the mean number of customers in the queue?
c) Find the expected waiting time in the system.
d) A new manager joins the company and re-structures the customer services into three units. One person is assigned to each unit. She feels the arrival rate of customer at each service unit is 5 per hour. The three employees are assigned to one service unit each. Do you think that the new re-structuring is advisable? Give reasons for your answer.

[4+3+3+10=20]

5. Answer the following questions briefly:-

a) Why should we be interested in obtaining the optimal solution of the primal by solving the dual?
b) What does the notation $(M/E_3/4):(12/36/SIRO)$ mean in the context of queuing theory?

[5+5=10]

6. Answer the following questions:

a) If a quality characteristic is centered and normally distributed, then find the expected proportion of Non-Conformance (NC) in terms of the process capability index C_p.
b) A consumer has received a special consignment of lot size 100. Find the probability of accepting the lot if it contains 2% defective and he decides to use a sampling plan with $n = 10$ and $c = 1$.
c) Describe the operation of a double sampling plan.
d) Distinguish between Type A and Type B OC curve of an acceptance sampling plan.

[7+8+5+5=25]

7. Fill in the blanks. You need not copy the statements. The first one is done for you.

i. Customer is the ____________. (Answer: *king*)

ii. The j^{th} constraint in the dual of an LPP is satisfied as strict inequality by the optimal solution. The j^{th} variable of the primal will assume a value ____________.

iii. If the j^{th} constraint in the primal is an equality, then the corresponding dual variable is __.

iv. AOQL means ____________ .

v. Consumer's risk of 10% means that ____________ .

vi. If ***nothing*** is known concerning the pattern of variation of a set of numbers, we can calculate the standard deviation of this set of numbers and state that the sample mean$\pm$3 the calculated standard deviation will include ***at least*** _____% of all the numbers.

vii. Customer behavior in which she moves from one queue to another (in the hope of reducing her waiting time) in a multiple channel situation is called __________.

viii. World Quality Day is celebrated on _________.

ix. World Standard Day is celebrated on _________.

x. The simplex algorithm of solving a linear programming problem was developed by _____.

xi. The process capability index which is appropriate when a given target is to be met is defined by _________.

xii. The quality tool used to separate the vital few from the trivial many is called ________.

xiii. An appropriate control chart which might be used to monitor the viscosity of paint in the paint shop of a car manufacturing company is __________.

xiv. If X is exponentially distributed, then $Y = f(X) =$ _________ is a Weibull random variable, which (i.e. Y) can be approximated by the normal distribution.

xv. In the context of $\bar{X} - R$ charts, the $2\sigma -$ limits placed on the $\bar{X} -$ chart is called _________, while the $3\sigma -$ limit is called _________.

[15]

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Appendix VI Factors for Constructing Variables Control Charts

Observations in Sample, n	Chart for Averages: Factors for Control Limits			Chart for Standard Deviations: Factors for Center Line		Chart for Standard Deviations: Factors for Control Limits				Chart for Ranges: Factors for Center Line		Chart for Ranges: Factors for Control Limits				
	A	A_2	A_3	c_4	$1/c_4$	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.575
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.115
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

For $n > 25$

$$A = \frac{3}{\sqrt{n}}, \qquad A_3 = \frac{3}{c_4\sqrt{n}}, \qquad c_4 \cong \frac{4(n-1)}{4n-3},$$

$$B_3 = 1 - \frac{3}{c_4\sqrt{2(n-1)}}, \qquad B_4 = 1 + \frac{3}{c_4\sqrt{2(n-1)}},$$

$$B_5 = c_4 - \frac{3}{\sqrt{2(n-1)}}, \qquad B_6 = c_4 + \frac{3}{\sqrt{2(n-1)}}.$$

Date: 29.4.2019
Time: 3 hours

INDIAN STATISTICAL INSTITUTE

Statistical Methods in Genetics
B-Stat (3rd Year) 2018-2019
Semester Examination

This is an open notes examination. The paper carries 50 marks. Calculators are allowed.

1. Consider an autosomal locus with alleles $A_1, A_2, \ldots, A_{10}$.

(a) If the locus is in Hardy-Weinberg Equilibrium, show that the heterozygosity at this locus can be at most 0.9.

(b) Suppose the population practises self-mating. Obtain the equilibrium frequencies of the different genotypes at this locus. [4 + 6]

2. Consider a disorder controlled by an autosomal biallelic locus. Suppose 100 nuclear families each comprising two unaffected parents and one offspring and 60 nuclear families each comprising exactly one parent affected and one offspring are randomly selected from the population. It was found that in the first group of families, there were 10 affected offspring while in the second group of families, there were 18 affected offspring. Are the above data consistent with a recessive mode of inheritance? [10]

3. Suppose the fitness values corresponding to the genotypes A_1A_1, A_1A_2 and A_2A_2 at a biallelic locus are f_1, f_2 and f_3 (not all same), respectively.

(a) If $f_2 < \min\{f_1, f_3\}$, show that the allele frequencies do not reach non-trivial equilibrium values.

(b) For each of the following fitness models, determine whether the allele frequencies reach non-trivial equilibrium values:
(i) f_1, f_2 and f_3 are in H.P.
(ii) $f_1 = 1-\beta\mu_{12}$, $f_2 = 1$, $f_3 = 1-\beta\mu_{21}$ where μ_{ij} is the mutation rate from A_i to A_j i,j=1,2 and β is a positive real number. [4 + 6]

4. Suppose individual *1* is a paternal aunt of individual *2*.

(a) If X_i denotes the number of minor alleles possessed by individual i, i=1,2 at an autosomal biallelic locus, compute the correlation coefficient between X_1 and X_2.

(b) Suppose individual *3* is the paternal grandfather of individual *2*. If individual 1 is affected with a genetic disease controlled by this locus, who between individuals *2* and *3* is more likely to be affected? [6 + 4]

5(a) Consider an allele-based case-control association test at an autosomal biallelic locus where the true Odds Ratio (OR) is 2. What should be the minimum <u>combined</u> sample size such that the large sample test based on OR at level 0.05 has power of at least 0.8?

(b) Suppose data on a quantitative trait are collected on k individuals for each of the genotype combinations corresponding to two autosomal loci: one biallelic and the other triallelic. Stating your assumptions clearly, suggest a suitable linear model framework to model the marginal and the interaction effects of the two loci. Explain how you would test for these effects. [5 + 5]

INDIAN STATISTICAL INSTITUTE

Bachelor of Statistics (Hons.) Third Year

Special Topics on Algorithm

Date: **April 29, 2019** Maximum Marks:100 Duration: **3 hours**

Note: The question paper carries a total of **120** marks. You can answer as much as you can, but the maximum you can score is **100**.

You need to prove the correctness of your algorithms.

1. You are given flow network $G = (V, E)$ is a graph with a source vertex $s \in V$ and target vertex $t \in V$. Each edge $e \in E$ has a positive integral capacity c_e. Suppose f is a maximum in G, that has been computed. Now, someone changes the capacity of a particular edge $e = (u, v)$ from c_e to c'_e, where c'_e is also an integer and $|c_e - c'_e| = r$. Design an algorithm to compute the new max-flow in $O(r(|V| + |E|)$ steps.
[Note that, the algorithm may behave differently depending on whether $c_e > c'_e$ or $c_e < c'_e$.]
[20]

2. k trucking companies, $c_1, \ldots, c_k$, want to use a common road system, which is modeled as a directed graph, for delivering goods from source locations to a common target location. Each trucking company c_i has its own source location, modeled as a vertex s_i in the graph, and the common target location is another vertex t. (All these $k+1$ vertices are distinct.) The trucking companies want to share the road system for delivering their goods, but they want to avoid getting in each others way while driving. We assume that there is no problem if trucks of different companies pass through a common vertex. Design an algorithm to determine k such paths, if possible, and otherwise return "impossible".
[20]

3. Show that for any $\alpha > 1$, there does not exist an α-approximation algorithm for the traveling salesman problem on n cities, provided $P \neq NP$. Describe a 2-approximation algorithm for the metric traveling salesman problem.
[6+14=20]

4. In the Knapsack problem, we are given a set $A = \{a_1, \ldots, a_n\}$ of n items, where each a_i has a specified positive integer size s_i and a specified positive integer value v_i. We are also given a positive integer knapsack capacity B. Assume that $s_i \leq B$ for every i. The problem is to find a subset of A whose total size is at most B and for which the total value is maximized.

   (a) Consider the following greedy algorithm Alg_1 to solve the Knapsack problem: Order all the items a_i in non-increasing order of their density, which is the ratio of value to size, $\frac{v_i}{s_i}$. Make a single pass through the list, from highest to lowest density. For each item encountered, if it still fits, include it, otherwise exclude it. Prove that algorithm Alg_1 does not guarantee any constant approximation ratio. That is, for any positive integer k, there is an input to the algorithm for which the total value v of the set of items returned by the algorithm is at most $\frac{v}{Opt}$.

   (b) Consider the following algorithm Alg_2. If the total size of all the items is $\leq B$, then include all the items. If not, then order all the items in non-increasing order of their densities. Without loss of generality, assume that this ordering is the same as the ordering of the item indices. Find the smallest index i in the ordered list such that the total size of the first i items exceeds B (i.e., $\sum_{j=1}^{i} s_j > B$ but $\sum_{j=1}^{i-1} s_j \leq B$). If $v_i > \sum_{j=1}^{i-1} v_j$, then return $\{a_i\}$. Otherwise return $\{a_1, \ldots, a_{i-1}\}$. Prove that Alg_2 always yields a 2-approximation to the optimal solution.

[10+10=20]

5. Consider the following experiment that proceeds in a sequence of rounds. For the first round, we have n balls, which are thrown independently and uniformly at random into n bins. After round i, for $i \geq 1$, we discard every ball that ended up in a bin by itself (i.e., it is the only ball in that bin) in round i. The remaining balls are retained for round $i + 1$, in which they are again thrown independently and uniformly at random in to the n bins.

   (a) Suppose that in some round we have $k = \epsilon n$ balls. How many balls should you expect to have in the next round?

   (b) Assuming that everything proceeded according to expectation, prove that we would discard all the balls within $O(\log \log n)$ rounds.

[8+12=20]

6. Describe the Perceptron learning Algorithm. Establish the correctness of the algorithm.

[6+14=20]

INDIAN STATISTICAL INSTITUTE

Semester Examination: 2018 – 19

B.Stat 3rd Year

Random Graphs

Date: 29-04-2019 Maximum Marks: 60 Duration: 3 hr

Anybody caught using unfair means will immediately get 0. Please try to explain every step. **Only class notes are allowed in the exam.**

(1) Consider the generalized random graph model $GRG(\mathbf{w})$ on n vertices. Suppose $\mathbf{w} = (w_i)_{i\in[n]}$ satisfies the standard assumptions: let W_n be the weight of an uniformly chosen vertex.

(A) Suppose there exists a random variable W such that $W_n \xrightarrow{d} W$.

(B) Suppose $\mathsf{E}[W] > 0$ and $\mathsf{E}[W_n] \to \mathsf{E}[W]$.

(C) $\mathsf{E}[W_n^2] \to \mathsf{E}[W^2]$.

Show that

(a) $\max_{i\in[n]} w_i = \mathrm{o}(\sqrt{n})$. [5]

(b) Let F be a distribution function such that

$$F(x) = \begin{cases} 0 & \text{if } x \leq 1 \\ 1 - x^{-\alpha} & \text{if } x \geq 1 \end{cases}$$

for some $\alpha > 0$. Fix $w_i = (1-F)^{\leftarrow}(i/n)$ where $(1-F)^{\leftarrow}$ is the generalized inverse:

$$(1-F)^{\leftarrow}(z) = \inf\{x : (1-F)(x) \leq z\}.$$

Find conditions on α such that the above standard assumptions hold. [10]

(2) Consider the degree sequence $(d_i)_{i\in[n]}$ such that $\ell_n = \sum_{i\in[n]} d_i$ is even. Suppose that $(d_i)_{i\in[n]}$ satisfy the standard assumptions: let D_n be the degree of an uniformly chosen vertex.

(A) Suppose there exists a random variable D such that $D_n \xrightarrow{d} D$ with $P(D \geq 1) = 1$.

(B) Suppose $\mathsf{E}[D_n] \to \mathsf{E}[D]$.

(C) $\mathsf{E}[D_n^2] \to \mathsf{E}[D^2]$.

Show that the number of simple graphs with degree sequence $(d_i)_{i\in[n]}$ is equal to

$$e^{-\frac{\nu}{2}-\frac{\nu^2}{4}} \frac{(\ell_n - 1)!!}{\prod_{i\in[n]} d_i!}(1 + o(1))$$

where $\nu = \frac{\mathsf{E}[D(D-1)]}{\mathsf{E}[D]}$. Using above find out asymptotically the number of r-regular graphs when nr is even. [10]

(3) Consider the Erdős-Rényi random graph $ER_n(p)$. Let $(S_k)_{k\geq 0}$ denote the random walk exploration process of a connected cluster. Show that for $t \geq 1$,

$$S_t + (t-1) \stackrel{d}{=} \text{Bin}(n-1, 1-(1-p)^t).$$

[10]

(4) Consider T_n, the number of occupied 4-cycles in $ER_n(\lambda/n)$. Recall that the distinct vertices (i, j, k, l) form an occupied 4-cycle when the edges $\{i, j\}, \{j, k\}, \{k, l\}, \{l, i\}$ are all occupied. Note that the 4-cycles (i, j, k, l) and (i, k, j, l) are different. Show that T_n asymptotically follows a Poisson distribution.

[Hint: Use method of factorial moments] [15]

(5) Show that a random variable X is stochastically smaller than Y if and only if there exists a coupling $(\widehat{X}, \widehat{Y})$ of (X, Y) such that $\mathsf{P}(\widehat{X} \leq \widehat{Y}) = 1$. [5+5=10]

END SEMESTER EXAMINATION - NUMBER THEORY
ISI KOLKATA - B.STAT. III YEAR - 2018–19
DATE : APRIL 29, 2019, TIME : 180 MINS. (2:30–5:30 PM)
MAXIMUM MARKS : 50

Instructions: Answer as much as you can. Maximum you can score is 50. If you are using any standard result, you must mention it. Calculators and course materials are not allowed.

(1) Let $\gcd(a,b)=1$. Then show that $\gcd(a+b,a-b)$ is at most 2. [3]

(2) Prove that there exist infinitely many primes of the form $3n+2$, $n\in\mathbb{N}$. [4]

(3) Show that there are infinitely many primes, by using the fact that $\sum_{n=1}^{\infty}\frac{1}{n}$ diverges. [5]

(4) Show that $x^2-5y^2=13$ has no solution for $x,y\in\mathbb{Z}$. [5]

(5) Classify all odd primes p such that the Legendre symbol $\left(\frac{7}{p}\right)=1$. Do the same for odd primes p such that the Legendre symbol $\left(\frac{7}{p}\right)=-1$. [3+2=5]

(6) Prove that if $p\geq 7$ is a prime then there exist at least two consecutive quadratic residues mod p. [6]

(7) Show that $\sin\alpha$ is a transcendental number for any non-zero algebraic number α. Using this show that $\tan\alpha$ is also transcendental. [3+2=5]

(8) Prove that the arithmetic function f defined by $f(n)=[\sqrt{n}]-[\sqrt{n-1}]$ is multiplicative, but not completely multiplicative. [4]

(9) For a positive integer q, prove that as $x\to\infty$, [5]

$$\sum_{\substack{n\leq x\\(n,q)=1}}\frac{1}{n}=\frac{\varphi(q)}{q}\log x+c+O\left(\frac{1}{x}\right),$$

for some constant c.

(10) Show that for $\Re(s)>1$, [4]

$$\frac{\zeta(s)}{\zeta(2s)}=\sum_{n=1}^{\infty}\frac{|\mu(n)|}{n^s},$$

where μ denotes the Möbius function.

(11) Prove that

$$\frac{\zeta(s)\zeta(s-a)\zeta(s-b)\zeta(s-a-b)}{\zeta(2s-a-b)}=\sum_{n=1}^{\infty}\frac{\sigma_a(n)\sigma_b(n)}{n^s},$$

in the region of absolute convergence where $\sigma_\alpha(n)=\sum_{d|n}d^\alpha$. [5]

(12) Prove that $\mathbb{Z}$ is dense in the set of p-adic integers $\mathbb{Z}_p$. [4]

P.T.O.

(13) Find the p-adic expansion of $1/2$ where p is an odd prime. [4]

(14) Let p be a prime. Prove that in $\mathbb{Q}_p$, [3+3=6]

$$\sum_{n\geq 1} n! \cdot n = -1 \text{ and } \sum_{n\geq 1} n^2 \cdot (n+1)! = 2.$$

INDIAN STATISTICAL INSTITUTE
Semester Examination 2018-19

B.Stat. - 3rd Year
Design of Experiments

3rd May, 2019

Maximum Marks: 100 Time: 3 hours

Note: Notations are as used in the class. Answer as much as you can. Best of Luck!

1. We define a general block design (GBD) to be *Special* if the BLUEs of all normalized estimable treatment contrasts have the same variance.

(a) Show that a GBD having C-matrix C is *Special* if and only if θC is idempotent for some $\theta > 0$.

(b) Show that a Connected GBD having C-matrix C is *Special* if and only if all diagonal elements of C are the same and all off-diagonal elements of C are also the same.

[10+5]=15

2. Consider a 4×4 square numbered as follows:

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Construct a GBD with 16 treatments and 8 blocks by considering the 4 rows and 4 columns of the above square as blocks. Prove that the resulting design is not orthogonal. [5]

3. Is it possible to construct a GBD involving more than 4 treatments which is design orthogonal but not connected? Justify your answer. [6]

4. Consider a set of five treatments numbered as $\{1, 2, 3, 4, 5\}$. Take all possible triplets as blocks to construct a GBD with 10 blocks of size 3 each.

(a) Show that the resulting design is A-efficient in the class of all proper block designs with $v = 5$, $b = 10$ and $k = 3$. Clearly state the result(s) which you are using.

(b) Verify if the resulting design is *Special* as per the definition given in Question 1.

[(6+3)+6]=15

5. Answer the following questions for Latin Square (LS) designs as per the class notation.

(a) Construct any three mutually orthogonal LSs of order $v = 9$.

(b) Choose any one of your three LSs constructed in part (a) for analysis. Suppose that, unfortunately after performing the experiment, we have two observations missing there, in cells $(1, 1)$ and $(8, 9)$. Explicitly estimate these observations through the Missing-Plot Technique. Clearly write down the main result that you need to use here.

(c) Derive the ANOVA test for treatment comparison in the case discussed in Part (b).

[12+(5+3)+9]=29

6. (a) Construct a Balanced Confounded Design (BCD) for $(3^3, 3^2)$ factorial experiment with 4 replications such that each of the 4 three-way interactions are confounded in exactly one of the 4 replications. You need to specify all the blocks for each replication.
(b) Write down the ANOVA Table (without the expectations) for your design constructed in Part (a). [15+5]=20

7. Consider a split-plot design with two factors having p and q levels, respectively, for the whole-plot and subplot factors and with r replications.

(a) Clearly write down the model with all assumptions.

(b) Write down the BLUEs of all the effect parameters along with estimates of the variance components.

(c) Derive a test for pairwise comparison between the interaction effects using the estimates from (b). You can assume normality.

[5+5+10]=20

INDIAN STATISTICAL INSTITUTE

Backpaper Examination : (2018-2019)

B.Stat. 3rd Year

NONPARAMETRIC AND SEQUENTIAL METHODS

10.07.2019

Date: ????, 2019 Max. Marks: 100 Duration: 3 Hours

Answer all questions.

1. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two random samples drawn independently from two populations with continuous distribution functions F and G respectively. Assume that $G(x) = F(x - \theta)$ for all x and some θ.

(a) Consider the problem of testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Describe the Mann-Whitney U test and the Wilcoxon rank sum test for this problem and show that these two tests are equivalent.

(b) Consider the Mann-Whitney U test of level α for testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Show that this test is unbiased. Also show that this test is of level α for testing the null hypothesis $H : \theta \leq 0$ against $H_1 : \theta > 0$.

(c) Construct a confidence interval for θ with confidence coefficient $(1-\alpha)$ $(0 < \alpha < 1)$ based on the differences $Y_j - X_i$, $i = 1, \ldots, m$, $j = 1, \ldots, n$.

[12+10+13=35]

2. Let $X_1, \ldots, X_n$ be a random sample from a population with a continuous distribution having unknown (unique) median θ. Consider the problem of testing $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Find an asymptotic level α $(0 < \alpha < 1)$ test for this problem and show that this test is consistent for any alternative $\theta > 0$. [6+9=15]

3. Consider Wald's sequential probability ratio test (SPRT) for a simple hypotheses H_0 and a simple alternative H_1 in the case of i.i.d. observations $X_1, X_2, \ldots$with target strength (α, β) and boundaries A and B satisfying $0 < B < 1 < A < \infty$. Let n be the stopping time of the SPRT and $Z_i =$

PTO

$\log(f_1(X_i)/f_0(X_i))$, $i \geq 1$, $f_j(\cdot)$ denoting the density of a single observation under H_j, $j = 0, 1$.

(a) State the *Fundamental Identity of Sequential Analysis* and describe how it can be used to find approximate expressions for the OC and ASN functions of an SPRT. Prove your results.

(b) Let H be a hypothesis under which $X_1, X_2, \ldots$ are i.i.d. and $P_H(Z_1 = 0) < 1$. If $E_H(Z_1) = 0$, show that the OC under H is approximately equal to $a/(a-b)$ where $a = \log A$ and $b = \log B$. [(3+20)+6=29]

5. Let $X_1, X_2, \ldots$ be i.i.d $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Describe Stein's two-stage sampling procedure for obtaining a bounded length confidence interval for μ with confidence coefficient $(1-\alpha)$. Prove the results you state.

Let n be the stopping time of Stein's procedure. Prove that n is finite with probability one. Also show that the sample mean $\bar{X}_n$ is an unbiased estimator of μ. [13+4+4=21]

INDIAN STATISTICAL INSTITUTE
B. Stat. (Hons.) III Year 2018-2019

Back Paper Examination

Subject : SQC & OR

Date : 12.07.2019 Full Marks : **100** Duration : **3 hrs.**

1) Consider the following LP:

$$\mathcal{P}: \quad \text{Minimize} \quad Z = 2x_1 + x_2 - x_3$$

subject to

$$x_1 + x_2 - x_3 = 1$$
$$x_1 - x_2 + x_3 \geq 2$$
$$x_2 + x_3 \leq 3$$
$$x_1 \geq 0;\ x_2 \leq 0;\ x_3 \text{ unrestricted in sign.}$$

a) Write down the dual $\mathcal{D}$ of the above problem.
b) Find (if it exists) a feasible solution for the primal $\mathcal{P}$.
c) Give a feasible solution (if it exists) for the corresponding dual problem.
d) State with reasons whether the Weak Duality Theorem holds for this pair of primal - dual problems.
e) State with reasons whether this pair of primal - dual problems admits of an optimal solution.

[10+2+2+3+3=20]

2) A computer network shares a printer. Jobs arrive at a mean rate of 2 jobs per minute and follow a Poisson process. The printer prints 10 pages per minute and the mean number of pages per job is 4. There is a 3-second idle time between one job and the next. The service time follows an exponential distribution. Calculate the following:
a) The percentage of time that the printer is available.
b) The mean queue length.
c) The service rate required for the mean time in the system to be under 3 minutes.

[5+5+5=15]

3) Two breakfast food manufacturers A and B are competing for an increased market share. The payoff matrix, shown in the following table, describes the increase in market share for A and decrease in market share of B.

A	*B*			
	Give Coupons	*Decrease Price*	*Maintain Present Strategy*	*Increase Advertising*
Give Coupons	1	2	−2	2
Decrease Price	3	1	2	3
Maintain Present Strategy	−1	3	2	1
Increase Advertising	−2	2	0	−3

Determine the optimal strategies for both the manufacturers and the value of the game.

[15+5=20]

4) Control charts for $\bar{X}$ and R are in use with the following parameters:

	$\bar{X}$ chart	R chart
UCL	363.0	16.81
Central Line	360.0	8.91
LCL	357.0	1.64

The sample size is $n = 9$. Both charts exhibit control. Assume that the quality characteristic is normally distributed.

a) What is the $\alpha -$ risk associated with the $\bar{X} -$ chart?

b) Specifications on this quality characteristics are 358 ± 6. Find the C_{pk} index for this process.

c) Find the proportion of non-conformance being produced, if any.

d) Suppose the mean shifts to 357. What is the probability that the shift will not be detected on the first sample following the shift?

e) What would be the appropriate control limits for the $\bar{X} -$ chart if the probability of type I error were to be 0.01?

f) Since both charts exhibit control, the supervisor suggested that the sample size be reduced to $n = 5$. What will be the new control limits (and the central lines) for the two charts?

[3+2+3+3+3+6=20]

5) Consider the following acceptance sampling plans:-

Plan	N	n	c
I	1000	240	2
II	1000	170	1
III	1000	100	0

Suppose that the AQL has been fixed at 2.2%.

i. As a consumer which plan would you prefer to use?

ii. Find the probability of accepting lots of 1% defective for each of these plans.

iii. Using (ii) above and without any further calculations, state which plan would you prefer to use if you are the producer?

[6+6+3=15]

6) Answer the following questions briefly:-

a) Distinguish between Type A and Type B OC curves of acceptance sampling plans.

b) Why should we be interested in obtaining the optimal solution of the primal by solving the dual?

c) Sis Proseguer is a company which describes itself as ***Leader in Cash Logistics***. It provides comprehensive range of Cash in Transit and ATM Services for banks and commercial establishments. You have been recruited at a hefty salary to help the company improve its services. Mention three OR problems that you can take up to help the company.

[4 + 3 + 3 = 10]

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

Appendix VI Factors for Constructing Variables Control Charts

Observations in Sample, n	Chart for Averages: Factors for Control Limits			Chart for Standard Deviations: Factors for Center Line		Chart for Standard Deviations: Factors for Control Limits				Chart for Ranges: Factors for Center Line		Chart for Ranges: Factors for Control Limits				
	A	A_2	A_3	c_4	$1/c_4$	B_3	B_4	B_5	B_6	d_2	$1/d_2$	d_3	D_1	D_2	D_3	D_4
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.575
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.115
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

For $n > 25$

$$A = \frac{3}{\sqrt{n}}, \quad A_3 = \frac{3}{c_4\sqrt{n}}, \quad c_4 \cong \frac{4(n-1)}{4n-3},$$

$$B_3 = 1 - \frac{3}{c_4\sqrt{2(n-1)}}, \quad B_4 = 1 + \frac{3}{c_4\sqrt{2(n-1)}},$$

$$B_5 = c_4 - \frac{3}{\sqrt{2(n-1)}}, \quad B_6 = c_4 + \frac{3}{\sqrt{2(n-1)}}.$$

INDIAN STATISTICAL INSTITUTE

Semester Examination: 2018 – 19

B.Stat 3rd Year

Random Graphs-Back/Supplementary Paper

Date: 15-07-2019 Total Marks: 100 Duration: 3 hr

Anybody caught using unfair means will immediately get 0. Please try to explain every step.

(1) Consider the configuration model $CM_n(\mathbf{d})$ with degree sequence $\mathbf{d} = (d_i)_{i\in[n]}$ satisfying the assumptions:

(a) Let U be an uniformly chosen vertex and $D_n = d_U$ the degree corresponding to this vertex. Assume there exists a D such that $D_n \xrightarrow{d} D$.

(b) Also assume $\lim_{n\to\infty} E[D_n] = E[D] > 0$.

Let $(D_i^{er})_{i\in[n]}$ be the degree sequence of the erased configuration model. Define

$$p_k^n = \frac{1}{n}\sum_{i\in[n]} \mathbf{1}_{\{d_i=k\}} \text{ and } P_k^n = \frac{1}{n}\sum_{i\in[n]} \mathbf{1}_{\{D_i^{er}=k\}}.$$

If $p_k = P(D = k)$ then show that under assumptions (a) and (b) that $(P_k^n)_{k\geq 1}$ converges in probability to $(p_k)_{k\geq 1}$.

[20 pts]

(2) Let $i \leftrightarrow j$ denote that there exists a path in the $ER_n(p)$ between i and j. Show that

$$\mathsf{P}(i \not\leftrightarrow j \,||C(i)| = l) = 1 - \frac{l-1}{n-1}.$$

[15 pts]

(3) Let $(X_i)_{i\geq 1}$ be iid $Poi(\lambda)$. Let I be the rate function. Show that for $a > \lambda$.

$$I(a) = a\log(\frac{a}{\lambda}) - a + \lambda.$$

Show that $I(a) > 0$ for all $a \neq \lambda$.

[15pts]

(4) The *clustering coefficient* of a random graph $G = (V, E)$ with V the vertex set and E the edge set is defined to be

$$\mathtt{CC}_G := \frac{E[\Delta_G]}{E[W_G]}$$

where

$$\Delta_G = \sum_{i,j,k \in V} \mathbb{1}_{\{ij,\, jk,\, ki \text{ occupied}\}}, \quad W_G = \sum_{i,j,k \in V} \mathbb{1}_{\{ij,\, ik \text{ occupied}\}}.$$

Thus (since we are not restricting to $i < j < k$ in Δ_G and to $i < k$ in W_G), Δ_G is six times the number of triangles in G, W_G is two times the number of wedges in G, and $\mathtt{CC}_G$ is the ratio of the number of expected closed triangles to the expected number of wedges. Compute $\mathtt{CC}_G$ for $ER_n(\lambda/n)$. [20 pts]

(5) Prove that $p(n) = \log n/n$ is the threshold function for the event A:="the Erdős-Rényi graph $ER_n(p(n))$ has at least an isolated vertex". [15 pts]

(6) Show that for (X, Y) random variables, $P(Y \le z) \le P(X \le z)$ for all z if and only if there exists a coupling $(\widehat{X}, \widehat{Y})$ of (X, Y) such that $P(\widehat{X} \le \widehat{Y}) = 1$.

Let $\lambda, \mu \ge 0$ such that $\lambda \le \mu$. Let $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$. Then show that $P(Y \le z) \le P(X \le z)$. [15pts]

INDIAN STATISTICAL INSTITUTE

Bachelor of Statistics (Hons.) Third Year

Special Topics on Algorithm

Date: **July 15, 2019** Maximum Marks:**100** Duration: **3 hours**

Note: The question paper carries a total of **100** marks.

You need to prove the correctness of your algorithms.

1. i) State and prove the Max-Flow-Min-Cut Theorem.

   ii) Show that the Ford-Fulkerson Method may not always work.

   [10+10]

2. Write an algorithm to compute the Voronoi Diagram of a set of points on the 2D plane. Establish the time and space complexities of your algorithm.

   [20]

3. Given an undirected graph $G = (V, E)$ a *Vertex Cover* of G is a set of vertices such that each edge in G is incident to at least one of these vertices. The vertex cover problem tries to find a minimum size vertex cover of a given graph. Show that the vertex cover problem is NP-complete. Give a 2-approximation algorithm for vertex cover.

   [10+10=20]

4. Show that for any $\alpha > 1$, there does not exist an α-approximation algorithm for the traveling salesman problem on n cities, provided $P \neq NP$. Describe a 2-approximation algorithm for the metric traveling salesman problem.

   [10+10=20]

5. Describe the Perceptron learning Algorithm. Establish the correctness of the algorithm.

   [6+14=20]

INDIAN STATISTICAL INSTITUTE
Back-paper Examination 2018-19 Date: 17-7-19

B.Stat. - 3rd Year
Design of Experiments

Maximum Marks: 100 Time: 3 hours

[Note: Notations are as used in the class. Answer as much as you can. Best of Luck!]

1. Prove or disprove the following as per the class notation.

   (a) In a completely randomized design with v treatments, the overall treatment sum-of-squares (SS) can be written as the sum of the SS due to $(v-1)$ mutually orthogonal treatment contrasts, each carrying one degree of freedom.

   (b) Structural Connectedness of a general block design is equivalent to its Rank Connectedness.

   (c) Design Orthogonality of a general block design implies its Data Orthogonality.

   (d) It is possible to construct a Balanced Confounded Design for $(3^5, 3^2)$ factorial experiment where no main effects and no two-way interaction effects are confounded.

   (e) A complete set of Mutually Orthogonal Latin Squares of order v can always be constructed if v is a prime or prime power. [10×5]=50

2. Consider a randomized block design with one missing observation.

   (a) Estimate the missing observation through the Missing-Plot Technique. Clearly write down the main result that you need to use here.

   (b) Derive the ANOVA test for treatment comparison with the observation estimated as in part (a) above. [(3+2)+10]=15

3. Explain confounding in factorial designs with an example. Suppose a block of $(2^7, 2^4)$ factorial experiment is given by (in standard notation)

$$g, acg, efg, acefg, ab, bc, abef, bcef, de, acde, df, acdf, abdeg, bcdeg, abdfg, bcdfg.$$

Find out the confounded factorial effects within this block. [5+15]=20

4. Assuming Hierarchical principal of factorial experiment, identify the better plan among the following two designs. Each design is given in terms of factorial effects confounded in each Replication.

Plan 1:	*Rep I*	AB	BC	AC	Plan 2:	*Rep I*	AB	CD	$ABCD$
	Rep II	CD	AD	AC		*Rep II*	AC	BD	$ABCD$
	Rep III	BC	BD	CD		*Rep III*	AD	ABC	BCD
	Rep IV	AD	BD	AB		*Rep IV*	BC	ABD	ACD

Write down the ANOVA Table (without the expectations) for the better design.

[7+8]=15

5. Consider a split-plot design with two factors having p and q levels, respectively, for the whole-plot and subplot factors and with r replications.

(a) Clearly write down the model with all assumptions.

(b) Write down the full ANOVA Table.

(c) Show that the pairwise contrasts of the whole plot treatments are estimated with less precision than those for the subplot treatments, under appropriate assumption on the error correlation.

[5+10+10]=25

INDIAN STATISTICAL INSTITUTE

Mid-Sem Examination

Bachelor of Statistics(Hons.) 3rd year 2017-18 (Semester-I)

Design and Analysis of Algorithms

Date: **September 4, 2018** Maximum Marks:**50** Duration: **1 hour 45 minutes**

Note: The question paper carries a total of **59** marks. You can answer as much as you can, but the maximum you can score is **50**.

1. (a) Write an algorithm to merge three sorted arrays into one.

   (b) Mergesort can be implemented as a three-way mergesort, where the input is split into three and three-way mergesort is applied recursively to each segment. Write an algorithm for three-way merge sort. Derive an expression for the worst case complexity (in terms of number of comparisons) for your algorithm.

   [4+(4+4)=12]

2. Clearly describe an algorithm, strictly better than $O(n^2)$, that takes a positive integer s and a set A of n positive integers and returns a Boolean answer to the question whether there exist two distinct elements of A whose sum is exactly s. Evaluate its complexity.

   [8+4=12]

3. Explain the data structure known as a *heap* and describe how a heap can be implemented using a simple linear block of memory. Write an algorithm to sort a set of numbers using a heap. Derive the complexity of your algorithm.

   [2+3+5+2=12]

4. Show how one may compute the square of a 2×2 matrix using only five multiplications.

   [6]

5. (a) The input is a set of n rectangles all of whose edges are parallel to the axes. Design an efficient algorithm to find the intersection of the all the rectangles. Derive the complexity of your algorithm.

   (b) Write the *Graham's Scan* algorithm for finding the convex hull of a set of points. Derive the complexity of the algorithm.

   [(4+3)+(5+5)=17]