Indian Statistical Institute, Bangalore
B. Math I, Second Semester, 2022-23
Final Examination
Introduction to Statistics and Computation with Data

04.05.23 **Maximum Score 100** **Duration: 3 Hours**

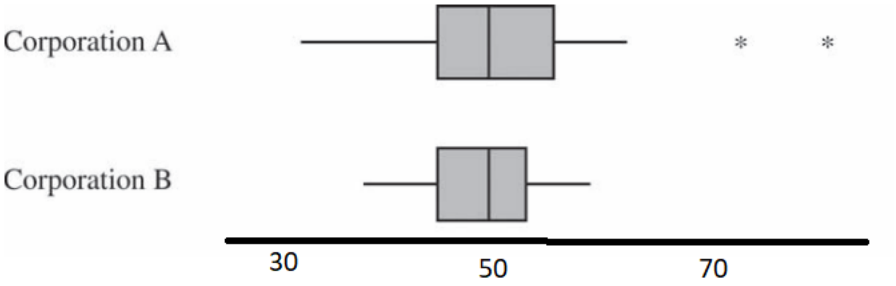
1. (5+5) A college wants to advertise the salary of its alumni 25 years after graduation. In order to collect the data, the college considers two alternatives to contact the batch of 1998.

   - Method 1: Send out an e-mail to all 5000 members of the class asking them to complete an online form. It is expected, based on other similar surveys, that at least 500 members will respond.
   - Method 2: Select a simple random sample of members of the class and contact the selected members directly by phone. Follow up to ensure that all responses are obtained. Because method 2 will require more time than method 1, the staff estimates that only 100 members of the class could be contacted using method 2.

   (a) Which of the two methods would you select for estimating the typical salary of all 5000 members of the class of 1998 ? Explain your reasoning by comparing the two methods and the effect of each method on the estimate.

   (b) Once the data is collected, would you recommend using the mean or the median of the sample as an estimate of the typical income? Justify your answer.

2. (4+3+3) Two large corporations, A and B, hire many new college graduates as accountants at entry-level positions. In 2018 the starting salary for an entry-level accountant position was Rs.36,000 a month at both corporations. At each corporation, data were collected from random samples of 30 employees who were hired in 2018 as entry-level accountants and were still employed at the corporation five years later. The monthly salaries of the 60 employees in 2023 are summarized in the boxplots below. The scale is in Rs 1000 per month. Suppose both corporations offered you a job for Rs 36,000 a month as an entry-level accountant.

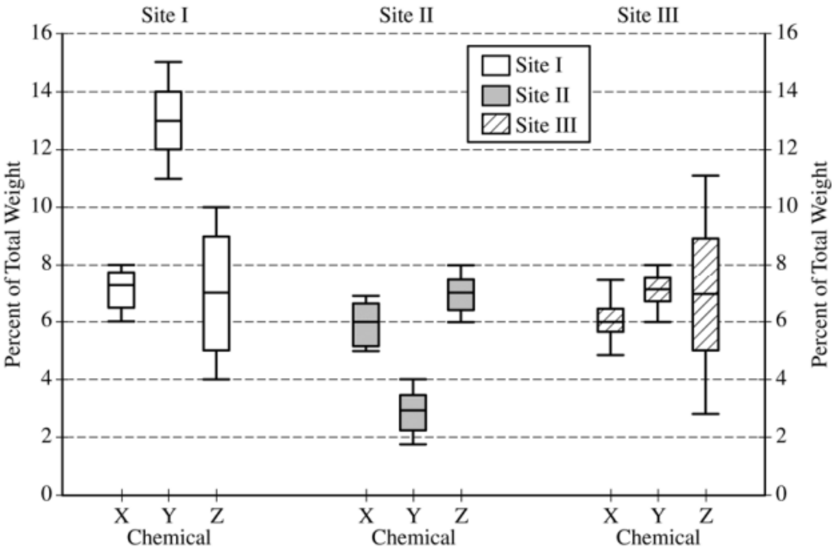


   (a) Compare the distributions of the monthly salaries at the two corporations.

   (b) Based on the boxplots, give one reason why you might choose to accept the job at corporation A.

   (c) Based on the boxplots, give one reason why you might choose to accept the job at corporation B.

3. (5+5) The chemicals in clay used to make pottery can differ based on where the clay originated. The boxplots below summarize the percentage of three chemicals X, Y, Z in pottery originating from three sites I, II, III. Consider a piece of pottery that originated from one of the sites, but the actual site is unknown.



   (a) Suppose the analysis reveals that the sum of proportions of the three chemicals in the piece of pottery is 20.5%. Which is the most likely site of origin?

   (b) Suppose only one chemical can be analysed. Whis one would be most useful in identifying the origin?

4. (12+3) Let $X_1, X_2 \cdots X_m$ be a random sample from Binomial(n, p) distribution where n, p are both unknown.

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

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

5. (10) Explain what the following R code and output is doing. The data is on hair and eye color of 592 individuals. State the model, hypotheses, data, assumptions, test statistic, its distribution and conclusion.

```
> data
       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16
> chisq.test(data)

Pearson's Chi-squared test

data:  data
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

6. (7+3) In a study of the relationship between birth order and college success, an investigator found that 140 in a sample of 200 college graduates were firstborn or only children. In a sample of 120 non-graduates of comparable age and socioeconomic background, the number of firstborn or only children was 66.

   (a) Estimate the difference between the proportions of firstborn or only children in the two populations from which these samples were drawn. Use a 95% confidence interval.

   (b) Can we conclude that the proportions are different in the two populations?

7. (5+7+8) Consider the regression model

$$y_i = \beta x_i + \epsilon_i, \quad 1 \leq i \leq n,$$

   where ϵ_i are iid with mean zero and variance σ^2 and x_i are fixed.
   Consider estimators of β of the form $T_{\mathbf{a}} = \sum_{i=1}^{n} a_i y_i$, for $\mathbf{a} = (\mathbf{a_1}, \cdots, \mathbf{a_n})^{\mathbf{T}} \in \mathbb{R}^{\mathbf{n}}$.

   (a) Find the least squares estimator T of β.

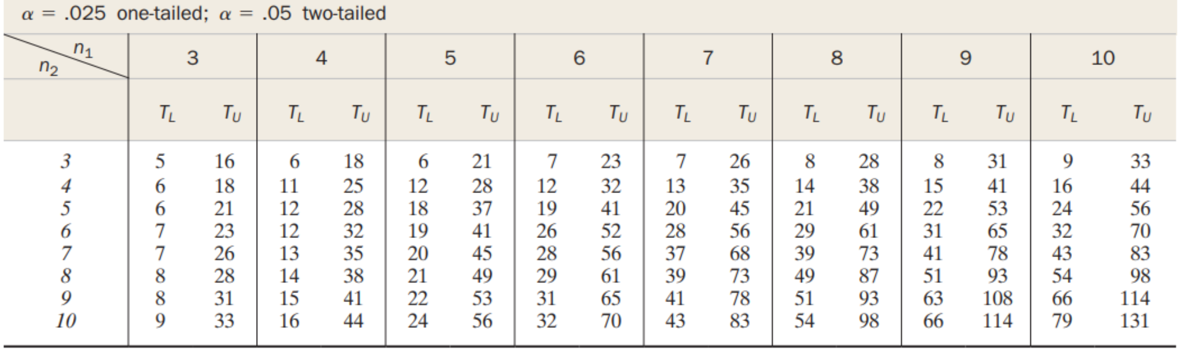
   (b) Find the mean and variance of $T_{\mathbf{a}}$ in terms of a_i's and the parameters β and σ.

   (c) Show that estimator T minimizes the variance of $T_{\mathbf{a}}$ over all possible $\mathbf{a} \in \mathbb{R}^{\mathbf{n}}$ such that $T_{\mathbf{a}}$ is unbiased.

8. (15) Polyethylene terephthalate (PET) bottles are used for carbonated beverages. A critical property of PET bottles is their bursting strength (i.e., the pressure at which bottles filled with water burst when pressurized). In the Journal of Data Science (May 2003), researchers measured the bursting strength of PET bottles made from two different designs: an old design and a new design. The data (in pounds per square inch) for 10 bottles of each design are shown in the following table.

old	210	212	211	215	190	213	212	218	164	209
new	216	217	162	137	229	216	149	153	182	217

   Carry out a nonparametric test at $\alpha = .05$ to compare the median of the distribution of bursting strengths for the two designs. The table of critical values is given below.



α = .025 one-tailed; α = .05 two-tailed

n_2 \ n_1	3		4		5		6		7		8		9		10	
	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U	T_L	T_U
3	5	16	6	18	6	21	7	23	7	26	8	28	8	31	9	33
4	6	18	11	25	12	28	12	32	13	35	14	38	15	41	16	44
5	6	21	12	28	18	37	19	41	20	45	21	49	22	53	24	56
6	7	23	12	32	19	41	26	52	28	56	29	61	31	65	32	70
7	7	26	13	35	20	45	28	56	37	68	39	73	41	78	43	83
8	8	28	14	38	21	49	29	61	39	73	49	87	51	93	54	98
9	8	31	15	41	22	53	31	65	41	78	51	93	63	108	66	114
10	9	33	16	44	24	56	32	70	43	83	54	98	66	114	79	131