དཔལ་ལྡན་འབྲུག་གཞུང་།
རྒྱལ་གཞུང་ཞི་གཡོག་ལྷན་ཚོགས།

ROYAL GOVERNMENT OF BHUTAN
ROYAL CIVIL SERVICE COMMISSION

BHUTAN CIVIL SERVICE EXAMINATION 2024

MAIN EXAMINATION

OCTOBER 4, 2024

CATEGORY: TECHNICAL (STATISTICS)

PAPER II: GENERAL SUBJECT KNOWLEDGE

ROYAL CIVIL SERVICE COMMISSION
BHUTAN CIVIL SERVICE EXAMINATION (BCSE) 2024
EXAMINATION CATEGORY: TECHNICAL

PAPER II: GENERAL SUBJECT KNOWLEDGE PAPER FOR STATISTICS

Date : October 4, 2024
Total Marks : 100
Writing Time : 90 minutes (1.5 hours)
Reading Time : 15 Minutes (prior to writing time)

GENERAL INSTRUCTIONS:

1. Write your Registration Number clearly and correctly on the Answer Booklet.
2. The first 15 minutes is to check the number of pages of Question Paper, printing errors, clarify doubts and to read the instructions. You are NOT permitted to write during this time.
3. This paper consists of **TWO Parts: Part I & Part II**

   Part I consists of 70 multiple choice questions of 1 (one) mark each, and

   Part II consists of short answer questions for 30 marks.

4. All questions are COMPULSORY.
5. All answers should be written on the Answer Booklet provided to you. Candidates are not allowed to write anything on the question paper. If required, ask for additional Answer Booklet.
6. **All answers should be written with correct numbering of Part and Question Number in the Answer Booklet provided to you. Note that any answer written without indicating the correct Part and Question Number will NOT be evaluated and no marks will be awarded.**
7. Begin each Part in a fresh page of the Answer Booklet.
8. You are not permitted to tear off any sheet(s) of the Answer Booklet as well as the Question Paper.
9. Use of any other paper including paper for rough work is not permitted.
10. **You must hand over the Answer Booklet to the Invigilator before leaving the examination hall.**
11. This paper has **14 printed pages**, including this instruction page.

GOOD LUCK!

PART I
Multiple Choice Questions (70 marks)

Choose the correct answer and write down the letter of your chosen answer in the Answer Booklet against the question number e.g. 71 (a). Each question carries ONE mark. Any double writing, smudgy answers or writing more than one choice shall not be evaluated.

1. In a standard normal distribution, what is the mean and standard deviation?
   a) Mean = 0, Standard Deviation = 1
   b) Mean = 1, Standard Deviation = 0
   c) Mean = 0, Standard Deviation = 0
   d) Mean = 1, Standard Deviation = 1

2. Which of the following is not a measure of dispersion?
   a) Range
   b) Mode
   c) Standard deviation
   d) Variance

3. The probability of getting at least one tail if a coin is tossed two times is
   a) ½
   b) ¼
   c) 1
   d) ¾

4. Which of the following is represented on the x-axis of a histogram?
   a) Categories or ranges of data values
   b) Frequency of data points
   c) Mean values of data sets
   d) Standard deviation of data points

5. What is a characteristic of official statistics?
   a) They are collected on an ad-hoc basis.
   b) They are subject to rigorous quality control and standardization.
   c) They are exclusively qualitative.
   d) They are typically secret and not released to the public.

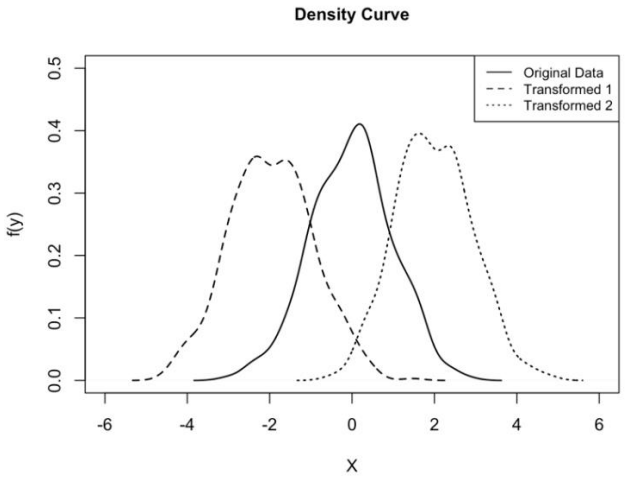
6. Which of the following is not a method for handling missing data?
   a) Clustering
   b) Deletion
   c) Imputation
   d) Substitution

7. What is the main purpose of regression analysis?
   a) To ascertain the distribution of variables
   b) To determine causal relationship between variables
   c) To determine the relationship between variables
   d) To measure the central tendency

8. Which of the following is not required when constructing a confidence interval for a population mean?
   a) Sample mean
   b) Population mean
   c) Sample size
   d) Standard error

9. Which of the following is an example of administrative data?
   a) Data from a market research survey
   b) Responses to a public opinion poll
   c) Results from a controlled experiment
   d) Patient records from a hospital

Use the density curve below to answer **Question 10**.



10. What do you observe in the probability densities of the transformed 1 and 2 as compared to the original data?
    a) The density curves are more flattened
    b) The density curves have different mean
    c) The density curves have different standard deviation
    d) The density curves are more skewed

11. Which of the following represent the distribution of age of death from natural causes (heart disease, cancer, etc.)? Most such deaths happen at older ages, with fewer cases happening at younger ages.
    a) Left-skewed distribution
    b) Right-skewed distribution
    c) Symmetrical distribution
    d) Uniform distribution

12. What does the term "bias" refer to in statistical analysis?
    a) A systematic error introduced into sampling or testing.
    b) The accuracy of the measurements.
    c) The consistency of the measurements.
    d) The variability of the data.

13. Which of the following best shows 'data deprivation' and its impact on public policy?
    a) A country has good economic data but lacks details on specific industries, leading to ineffective sector-specific policies.
    b) A city has demographic data but cannot share it between departments, causing inefficient service delivery and urban planning.
    c) A region frequently hit by natural disasters lacks systematic data on their impacts, leading to poorly informed disaster response strategies.
    d) A university has global health data but faces privacy concerns, limiting its ability to publish detailed studies.

14. Which of the following is a typical challenge in collecting official statistics?
    a) High response rates
    b) Data accuracy
    c) Sufficient funding
    d) Public awareness

15. When reporting statistical results, it is ethical to
    a) Adjust data points to fit the expected outcome
    b) Hide the methodology used to prevent scrutiny
    c) Only report results that support your hypothesis
    d) Present all relevant findings, including those that contradict the hypothesis

16. Which of the following examples would be most appropriately analysed using parametric test?
    a) The heights of adult males in a specific country.
    b) The number of daily visitors to a website.
    c) The median household income in a city.
    d) The ranks of students in a class.

17. In which scenario would a country's real GDP be a better indicator of economic growth compared to nominal GDP?
    a) When assessing changes in population size.
    b) When comparing GDP between different countries.
    c) When evaluating government fiscal policy.
    d) When measuring the impact of inflation on economic growth.

18. A hospital system consolidates patient information from different departments, including radiology, pharmacy, and lab results, into a central repository. What data integration technique is most appropriate for ensuring that all patient records are correctly linked?
    a) Data replication
    b) Data warehousing
    c) Data deduplication
    d) Data federation

19. In official statistics, what does the term "data confidentiality" mean?
    a) Ensuring that individual responses cannot be traced back to respondents
    b) Keeping the data hidden from the public
    c) Publishing only aggregate data
    d) Sharing all data with government agencies

20. Which statement is not true about the sample mean?
    a) A fixed quantity
    b) A random quantity
    c) An estimator of population mean
    d) Has uncertainty

21. A researcher is conducting a study on student performance across different schools and uses a convenience sampling method by choosing the schools that are easiest to access. What is a major limitation of this approach?
    a) It guarantees that the sample is representative of all schools.
    b) It introduces potential bias because the sample may not be representative of the entire population.
    c) It ensures a higher level of precision in estimates.
    d) It eliminates the need for a sampling frame

22. Which of the following situations would likely produce a correlation coefficient closest to zero?
    a) The number of hours studied and exam scores
    b) The height of a person and their shoe size.
    c) The amount of rainfall and the price of bread.
    d) The weight of a car and its fuel efficiency.

23. Which of the following is an example of paradata?
    a) Aggregate statistics from survey results
    b) Survey respondents' answers to demographic questions
    c) Summary of findings from survey analysis
    d) The time taken by respondents to complete the survey

24. Which type of variable best describes the variable "Number of books read per month"?
    a) Nominal
    b) Ordinal
    c) Interval
    d) Ratio

25. In a time series plot, what does a spike at a lag of 12 in the autocorrelation function (ACF) suggest?
    a) The presence of a trend in the data.
    b) A possible seasonal effect occurring every 12 periods.
    c) Random noise affecting the data.
    d) A significant increase in data values every 12 periods.

26. Which of the following is an example of metadata?
    a) The content of an email message
    b) The author, date, and file size of a document
    c) Survey responses from participants
    d) The number of pages in a book

27. A retailer calculates the average sales per day for a month. The mean sales per day is Nu 500, but one day had sales of Nu 5,000 due to a special promotion. What measure of central tendency should the retailer use to avoid the effect of this outlier?
    a) Mean
    b) Mode
    c) Median
    d) Range

28. In a survey of 2,000 households, the average household income per month is found to be Nu 30,000 with a standard deviation of Nu 1,000. Which test should be used to examine if the average household income per month is significantly different from Nu 35,000?
    a) Paired t-test
    b) One-sample t-test
    c) Independent t-test
    d) Z-test for proportions

29. A machine learning model is being trained to classify images using maximum likelihood estimation. During training, if the model frequently misclassifies images, what does this imply about the likelihood function?
    a) The likelihood function is high.
    b) The likelihood function is low.
    c) The likelihood function is constant.
    d) The likelihood function is irrelevant.

30. When analysing a time series, what does stationarity imply?
    a) The time series data have no trend or seasonal components.
    b) The mean and variance of the time series change over time.
    c) The time series has periodic fluctuations.
    d) The statistical properties of the time series remain constant over time.

31. Suppose the nominal GDP of Country Y is Nu 500 billion, and there is an annual inflation rate of 3%. What would be the primary factor if the nominal GDP rises to Nu 515 billion the next year?
    a) Solely the increase in population
    b) Solely the increase in prices due to inflation
    c) Solely the increase in the production of goods and services
    d) A combination of inflation and increased production

32. In what sense the word ‘probability’ is used in statistical theory?
    a) The degree of certainty that a specific event will occur.
    b) The relative frequency of an event's occurrence in a large number of trials.
    c) The likelihood of all possible outcomes occurring equally.
    d) The belief in the occurrence of an event based on personal judgment.

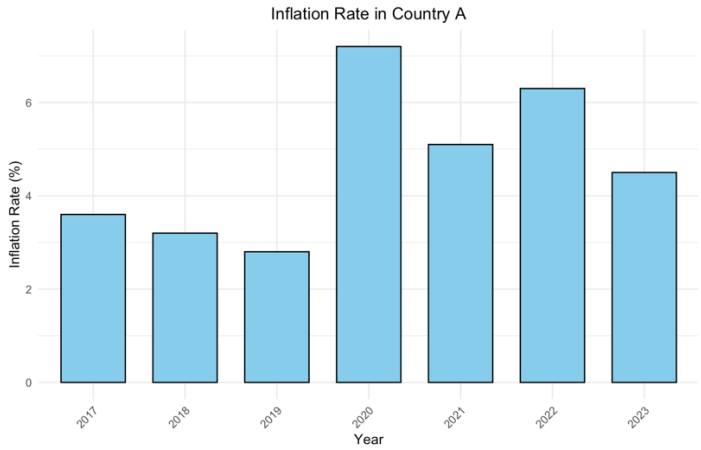
33. Which of the following is an example of a discrete variable?
    a) The weight of apples in a basket
    b) The time it takes to commute to work
    c) The number of pets owned by a family
    d) The height of students in a classroom

34. Which of the following is an example of official statistics?
    a) A scientific research study published in a journal
    b) Data from a social media platform
    c) National census data
    d) Survey results from a private market research firm

35. If a distribution is positively skewed, where is the mean relative to the median?
    a) Mean = Median
    b) Mean < Median
    c) Mean > Median
    d) Mean and median are unrelated

36. A retail store tracks the number of customers over five days: 150, 160, 170, 180, and 300. Which measure of dispersion is most affected by the outlier?
    a) Interquartile Range (IQR)
    b) Median
    c) Mode
    d) Range

Use the following chart to answer **Questions 37, 38 & 39**



37. What type of pattern is indicated in the chart?
    a) Cyclical pattern
    b) Seasonal pattern
    c) Trend pattern
    d) Irregular pattern

38. Which of the following statement about the inflation rate's variability can be inferred if the bar graph shows a large peak and low variability in other years?
    a) The inflation rate is consistently high.
    b) The inflation rate is very stable over time.
    c) The peak year represents an anomaly or outlier in the data.
    d) The inflation rate has a normal distribution.

39. What would have been the best graphical representation of the inflation rates data instead of the bar graph?
    a) Dot/scatter plot
    b) Line graph
    c) Pie chart
    d) Table

40. What is the primary purpose of Analysis of Variance (ANOVA) in statistical analysis?
    a) To compare the means of two groups.
    b) To determine the correlation between two variables.
    c) To assess whether there are statistically significant differences between the means of three or more groups.
    d) To estimate the probability of a single event occurring.

41. What is one benefit of using official statistics for research and analysis?
    a) They are always current and up-to-date
    b) They are collected through informal methods
    c) They provide a reliable and standardized source of data
    d) They are influenced by personal opinions

42. In government agencies, the most common number of staff at the division level is typically around 6. What measure of central tendency does this represent?
    a) Median
    b) Mode
    c) Mean
    d) Standard deviation

43. A market researcher computes the standard error of the mean height from a sample of 50 customers. If the standard error is small, what can be inferred?
    a) The sample mean height is highly variable
    b) The sample mean height is a precise estimate of the population mean
    c) The sample size was too small
    d) The heights of customers are distributed normally

44. For a discrete random variable X, which of the following is true about the cumulative distribution function (CDF) $F(x) = P(X \leq x)$?
    a) $F(x)$ decreases as x increases.
    b) $F(x)$ can have discontinuities at the values of x where $P(X=x)$ is non-zero.
    c) $F(x)$ is always differentiable.
    d) F(x) is equal to the probability density function (PDF) of X.

45. Which of the following scenarios would typically result in a random variable X being classified as discrete?
    a) Measuring the time taken for a chemical reaction to complete.
    b) Counting the number of defective items in a batch of products.
    c) Recording the temperature of a liquid at various time intervals.
    d) Measuring the distance between two points in a city.

46. A healthcare provider integrates patient records from multiple sources, such as electronic health records (EHRs), medical imaging, and genomic data. What big data challenge does this scenario exemplify?
    a) Data integration
    b) Data storage
    c) Data visualization
    d) Data security

47. What does the probability of an event represent in statistical theory?
    a) The number of times the event has occurred.
    b) The chance that the event will occur in a single trial.
    c) The proportion of times the event will occur in repeated trials.
    d) The belief in the occurrence of the event.

48. What is the primary aim of the design of experiments (DOE)?
    a) To minimize the number of trials needed to find the best outcome.
    b) To identify and quantify the effects of different factors on a response variable.
    c) To maximize the number of variables considered in a study.
    d) To ensure that all possible combinations of factors are tested.

49. A researcher uses time series analysis to predict monthly electricity consumption and notices that consumption spikes during the summer and winter months. This pattern is most likely due to
    a) A long-term trend
    b) Seasonal effects
    c) Random noise
    d) Cyclical fluctuations

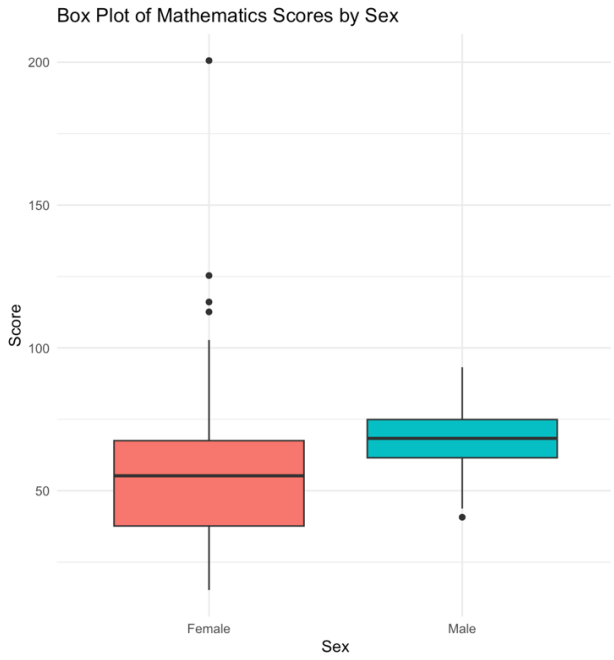
50. A teacher finds that the scores of a test in a class are closely clustered around the mean. Which measure of dispersion would be low for this set of scores?
    a) Interquartile Range (IQR)
    b) Range
    c) Standard Deviation
    d) Variance

51. In a study to estimate the average height of adults in a city, a researcher uses a stratified random sampling method where the city is divided into different age groups. Why might this approach be preferred over simple random sampling?
    a) It ensures that all age groups are represented proportionally in the sample.
    b) It eliminates the need for sampling altogether.
    c) It guarantees that the sample size will be larger.
    d) It reduces the variability within the sample.

52. If X and Y are independent random variables, which of the following statements is false?
    a) $E[X+Y] = E[X] + E[Y]$
    b) $Var(X+Y) = Var(X) + Var(Y)$
    c) $P(X > x \; and \; Y > y) = P(X > x) \times P(Y > y)$
    d) $Cov(X,Y) = 0 \; only \; if \; X \; and \; Y \; are \; normally \; distributed$

53. In a statistical experiment, the sample space S consists of 6 equally likely outcomes. If an event A is defined as a subset of S with 4 outcomes, what is the probability of the complementary event A^c?
    a) 1/3
    b) 1/6
    c) 2/3
    d) 4/6

Use the box plot below to answer **Questions 54 and 55.**



54. When comparing box plots for male and female scores, which of the following statements could indicate that the female scores are more skewed than the male scores?
    a) The median of female scores is higher than that of male scores.
    b) The box for female scores is wider than the box for male scores.
    c) The whiskers for female scores are longer on one side than for male scores.
    d) The box for female scores is shorter than the box for male scores.

55. In the given box plot, if the median score for males is higher than for females but the interquartile range (IQR) is larger for females, what can be inferred?
    a) Males have more variability in their scores than females.
    b) Females have a wider range of middle 50% of scores compared to males.
    c) Males have a higher variance in their scores compared to females.
    d) Females have a lower mean score compared to males.

56. A multinational corporation integrates financial data from subsidiaries located in different countries, each using different currencies. What is a key consideration in this data integration process?
    a) Data encryption
    b) Data latency
    c) Data standardization
    d) Data visualization

57. If two events are mutually exclusive, what is the probability that both events occur?
    a) 0
    b) 1
    c) It depends on the number of trials
    d) It depends on the probability of each event

58. An airline integrates data from its booking system, loyalty program, and flight operations to provide personalized offers to frequent flyers. What is the primary benefit of this data integration?
    a) Improved data security
    b) Enhanced customer experience
    c) Reduced data storage costs
    d) Simplified regulatory compliance

59. Which of the following does not affect the correlation coefficient between two variables?
    a) Changing the scale of measurement for one or both variables
    b) Shifting the variables by adding or subtracting a constant
    c) The presence of outliers
    d) The linearity of the relationship between the variables

60. Sex ratio is a basic measure to explain the sex composition of a population. It is defined as the number of males per 100 females. If a country reports sex ratio of 105, which of the following is correct interpretation?
    a) There are 105 females for every 100 males in the population.
    b) There are 5 more females than females in the population.
    c) There are 105 males for every 100 females in the population.
    d) There are 5 more males than females in the population.

61. A social media platform analyses millions of posts per second to identify trending topics in real-time. What aspect of big data does this scenario best represent?
    a) Volume
    b) Variety
    c) Velocity
    d) Value

62. A company is analysing the effect of advertising spend (X) on sales (Y). They find a positive regression slope. Which of the following conclusions is correct?
    a) Increasing advertising spend will always increase sales
    b) There is a causal relationship between advertising spend and sales
    c) Higher sales can be expected with higher advertising spend, holding other factors constant
    d) Sales are independent of advertising spend

63. If a random variable follows a chi-square distribution with k degrees of freedom, what happens to the shape of the distribution as k increases?
    a) It becomes more skewed to the right.
    b) It becomes more skewed to the left.
    c) It becomes symmetric and approaches a normal distribution.
    d) It remains unchanged regardless of *k.*

64. Suppose a company wants to predict the annual sales of its product based on the amount spent on advertising. However, the sales only increase up to a certain point and then start decreasing as the advertisement becomes overwhelming. Which type of regression model should be used?
    a) Linear Regression
    b) Polynomial Regression
    c) Logistic Regression
    d) Ridge Regression

65. A race has 8 runners, and medals are awarded for 1st, 2nd, and 3rd places. How many different ways can the medals be awarded?
    a) 56
    b) 168
    c) 336
    d) 720

66. If the confidence level increases from 95% to 99%, what happens to the width of the confidence interval?
    a) The interval width decreases.
    b) The interval width remains the same.
    c) The interval width increases.
    d) The effect on interval width cannot be determined without additional information.

67. The correlation between the height and weight of individuals is found to be 0.89. If we convert height from centimetres to inches (where 1 inch = 2.54 cm), the new correlation coefficient will be
    a) 0
    b) 0.35
    c) 0.5
    d) 0.89

68. A researcher increases the sample size from 50 to 200 while keeping the confidence level the same. What will happen to the confidence interval for the mean?
    a) It will become wider.
    b) It will become narrower.
    c) It will stay the same.
    d) It will shift upwards.

69. The mean and standard deviation of 100 observations are 40 and 5, respectively. If 4 is added to each observation, the new mean and standard deviation will be
    a) 35, 5
    b) 40, 0
    c) 45, 5
    d) 45, 0

70. A biologist is estimating the growth rate of a population of bacteria. She assumes that the population grows exponentially and collects data on population size over time. Which method should she use to estimate the growth rate parameter?
    a) Method of Moments
    b) Bayesian Estimation
    c) Maximum Likelihood Estimation
    d) Least Squares Estimation

PART II
Short Answer Questions (30 marks)

Answer all short answer questions. Marks for each question are indicated in the brackets.

1. In what ways can the central limit theorem be applied to real-world data analysis, and what are its limitations? (5 marks)

2. How do confidence intervals provide more information than point estimates in statistical reporting? (5 marks)

3. What role do assumptions play in the application of statistical models, and how can violations of these assumptions affect the results? (5 marks)

4. What are some challenges associated with the use of statistical learning in handling large-scale data sets? (5 marks)

5. State True or False and then support your answer with a brief explanation

   i. In regression analysis, the correlation coefficient between the predictor and the response variable always accurately reflects the strength and direction of the causal relationship between them (2 marks).

   ii. In a large dataset, a correlation coefficient of 0.1 implies a strong relationship between two variables (2 marks).

   iii. Increasing the sample size of a study will always improve the accuracy of parameter estimates (2 marks).

   iv. A high R-squared value in a regression model always indicates that the model has good predictive power (2 marks).

   v. In a dataset with a bimodal distribution, the mean and median will always be different (2 marks).

TASHI DELEK