**SET A**

**Sr. No. of Question Paper:**

**Unique Paper Code : 12275304**

**Name of the Course : B. A. (H) Economics - Generic Elective**

**Name of the Paper : Data Analysis (GE)**

**Semester : III**

**Duration: 3 Hours** **Maximum Marks: 65**

**Instructions:**

- **This question paper has two sections. Attempt any TWO questions from each section.**
- **You do not require the use of R or Excel software to answer any question. Wherever asked, mention/discuss the command/function/syntax, as required in the question.**
- **The questions in which R or Excel is not mentioned, the answers should be based on your own calculations.**

## SECTION A

1. (a) A professor is interested in examining the mental health awareness among students in a college. Out of 1000 students, a sample of 150 students was collected. The demographic data collected from these students include age, gender, caste category, parent's education, parent's occupation and household income.

    i. Identify the population of interest and discuss the steps involved in collecting a sample using systematic random sampling method. **(3)**

    ii. Give an example of each of the following survey errors that the study may face: coverage error, non-response error, sampling error and measurement error. **(4)**

    iii. Explain the difference between categorical and numerical variables, and indicate whether each of the demographic variables mentioned is categorical or numerical. **(3)**

    (b) What is the difference between the following Excel Functions: ISNUMBER and ISBLANK? Explain using examples. **(3)**

(c) Explain R commands to list a deck of 52 cards, both with and without jokers. **(3)**

1. (अ) एक प्रोफेसर कॉलेज में छात्रों के बीच मानसिक स्वास्थ्य जागरूकता की जांच करने में रुचि रखता है। 1000 छात्रों में से 150 छात्रों का नमूना एकत्र किया गया था। इन छात्रों से एकत्र किए गए जनसांख्यिकीय डेटा में आयु, लिंग, जाति श्रेणी, माता-पिता की शिक्षा, माता-पिता का व्यवसाय और घरेलू आय शामिल हैं।

   i. रुचि की आबादी की पहचान करें और व्यवस्थित यादृच्छिक नमूनाकरण विधि का उपयोग करके नमूना एकत्र करने में शामिल चरणों की चर्चा करें। **(3)**

   ii. निम्नलिखित सर्वेक्षण त्रुटियों में से प्रत्येक का एक उदाहरण दें जिसका शोध को सामना करना पड़ सकता है: कवरेज त्रुटि, गैर-प्रतिक्रिया त्रुटि, नमूना त्रुटि और माप त्रुटि। **(4)**

   iii. श्रेणीबद्ध और संख्यात्मक चर के बीच अंतर स्पष्ट करें, और इंगित करें कि उल्लिखित प्रत्येक जनसांख्यिकीय चर श्रेणीबद्ध या संख्यात्मक है या नहीं। **(3)**

(ब) निम्नलिखित एक्सेल फ़ंक्शंस के बीच क्या अंतर है: ISNUMBER और ISBLANK? उदाहरण का उपयोग करके समझाएं। **(3)**

(स) जोकर के साथ और बिना जोकर के 52 कार्डों के एक डेक को सूचीबद्ध करने के लिए R कमांड की व्याख्या करें। **(3)**

2. (a) The weights (in kilogram) collected from a sample of 10 adults are as follows: 62.3, 61.4, 60.7, 61.9, 61.4, 60.8, 60.2, 62.7, 60.0, 59.3. Use this information to answer the following questions:
   i. Is weight a discrete or continuous variable? Explain. **(2)**
   ii. Prepare an ordered stem and leaf plot. **(2)**
   iii. Compute mean and five-number summary. Use it to comment on the shape of the distribution. **(6)**

   (b) Discuss, using examples, the Excel functions used to calculate mean and Z scores. **(3)**

   (c) What is the difference between ls() and list() commands in R? Illustrate using examples. **(3)**

2. (अ)10 वयस्कों के नमूने से एकत्र किए गए वजन (किलोग्राम में) इस प्रकार हैं: 62.3, 61.4, 60.7, 61.9,61.4, 60.8, 60.2, 62.7, 60.0, 59.3। निम्नलिखित प्रश्नों के उत्तर देने के लिए इस जानकारी का प्रयोग करें:

i. वजन एक असतत या निरंतर चर है? समझाएं। **(2)**

ii. एक आदेशित stem एवं leaf का प्लॉट तैयार करें। **(2)**

iii. माध्य और पाँच-संख्या सारांश की गणना करें। वितरण के आकार पर टिप्पणी करने के लिए इसका प्रयोग करें। **(6)**

(ब) माध्य और Z स्कोर की गणना के लिए उपयोग किए जाने वाले एक्सेल फ़ंक्शंस की चर्चा करें। उदाहरण का उपयोग करके समझाएं। **(3)**

(स) R में ls() और list() कमांड के बीच क्या अंतर है? उदाहरण का उपयोग करके समझाएं। **(3)**

3. (a) The population of participants in a Science quiz has a mean score and standard deviation of 65 and 5, respectively. Suppose the shape of the population is unknown.

i. Explain Chebyshev's theorem and use it to determine the percentage of participants having scores between 30 and 100. **(6)**

ii. How likely is it for a participant to have a score less than 40? **(3)**

(b) Differentiate between AND and OR functions used in Excel. Use AND function to return TRUE if Cell B2 contains "BPL" and Cell C2 contains value less than 20000; FALSE otherwise. **(3)**

(c) Suppose you are given the data of height of students in a class, by gender: Male or female. Write R command to make a stem and leaf plot for representing height of the students. How will the command change if you need to represent the height of only male students? **(4)**

3. (अ) विज्ञान क्विज में प्रतिभागियों की आबादी का औसत स्कोर और मानक विचलन क्रमशः 65 और 5 है। मान लीजिए कि जनसंख्या का आकार अज्ञात है।

i. चेबीशेव के प्रमेय की व्याख्या करें और इसका उपयोग 30 और 100 के बीच स्कोर करने वाले प्रतिभागियों के प्रतिशत को निर्धारित करने के लिए करें। **(6)**

ii. किसी प्रतिभागी के लिए 40 से कम अंक प्राप्त करने की कितनी संभावना है? **(3)**

(ब) एक्सेल में उपयोग किए जाने वाले AND एवं OR फंक्शन के बीच अंतर करें। यदि सेल B2 में "BPL" है और सेल C2 में 20000 से कम का मान है, तो TRUE, अन्यथा FALSE, लौटाने के लिए AND फ़ंक्शन का उपयोग करें। **(3)**

(स) मान लीजिए कि आपको कक्षा में छात्रों की ऊंचाई और जेंडर (पुरुष या महिला) का डेटा दिया गया है। विद्यार्थियों की ऊँचाई को निरूपित करने के लिए stem और leaf का प्लॉट बनाने के लिए R कमांड लिखिए। यदि आपको केवल पुरुष छात्रों की ऊंचाई का प्रतिनिधित्व करने की आवश्यकता है तो कमांड कैसे बदलेगी? **(4)**

## SECTION B

4. (a) Consider a sample of 25 runners selected for a racing competition, drawn from a normal distribution with mean running time and standard deviation of the population equal to 30 minutes and 9 minutes, respectively. Use this information to answer the following questions. Also illustrate using a figure of normal distribution and mark the relevant areas.

    i. Find the probability that the average running time for the sample selected is: **(6)**

        (I) less than 25 minutes

        (II) more than 35 minutes

    ii. The middle 70% of all sample means will fall between which two values? **(4)**

(b) Distinguish between sample standard deviation and population standard deviation. Also explain the excel functions used to calculate the two measures. **(3.5)**

(c) Explain the use of cbind() and rbind() commands in R. **(3)**

4. (अ)एक रेसिंग प्रतियोगिता के लिए चुने गए 25 धावकों के एक नमूने पर विचार करें, जो सामान्य वितरण से लिए गए हैं, जिसमें औसत चलने का समय और जनसंख्या का मानक विचलन क्रमशः 30 मिनट और 9 मिनट के बराबर है। निम्नलिखित प्रश्नों के उत्तर देने के लिए इस जानकारी का प्रयोग करें। सामान्य बंटन के चित्र का उपयोग करके भी स्पष्ट कीजिए और संबंधित क्षेत्रों को चिह्नित कीजिए।

i. प्रायिकता ज्ञात कीजिए कि चयनित नमूने के लिए औसत चलने का समय है: **(6)**

(I) 25 मिनट से कम

(II) 35 मिनट से अधिक

ii. सभी प्रतिदर्श माध्यों का मध्य 70% किन दो मानों के बीच होगा? **(4)**

(ब) नमूना मानक विचलन और जनसंख्या मानक विचलन के बीच अंतर करें। दो उपायों की गणना के लिए उपयोग किए जाने वाले एक्सेल फ़ंक्शन की व्याख्या करें। **(3.5)**

(स) R में cbind () और rbind () कमांड के उपयोग की व्याख्या करें। **(3)**

5. (a) A sugar distributor wants to estimate the weight of sugar contained in 5-Kg bag of sugar purchased from a local sugar manufacturer. The manufacturer's specifications state that the standard deviation of the amount of sugar is equal to 100gms. A random sample of 1000 5-Kg bags is selected, and the sample mean of sugar per 5-Kg bag is recorded as 4.90 Kg. Based on this information, answer the following questions:

    i. Construct a 95% confidence interval estimate for the population mean amount of sugar included in a 5-Kg bag. On the basis of these results, do you think that the distributor has a right to complain to the sugar manufacturer? Why? **(6)**

    ii. What are the crucial assumption(s) required for the population distribution in order to construct the 95% confidence interval in Part (i)? **(3)**

    (b) Determine the sample size needed to estimate, with 95% confidence, that the mean per-individual bill amount in a cafe is within ± Rs. 20. Assume population standard deviation is Rs. 100. Which Excel function will you use to calculate the Z value used in the sample size calculation? **(4.5)**

    (c) Suppose you are given the following data: 5,3,7,10,-2,9,-1,2,-8,-2. Write R command to figure out the positions at which the numbers are greater than zero. **(3)**

5. (अ) एक चीनी वितरक एक स्थानीय चीनी निर्माता से खरीदी गई चीनी के 5 किलो बैग में निहित चीनी के वजन का अनुमान लगाना चाहता है। निर्माता के विनिर्देशों में कहा गया है कि चीनी की मात्रा का मानक विचलन 100 ग्राम के बराबर है। 1000 5-किलोग्राम बैग का एक यादृच्छिक नमूना चुना जाता है, और प्रति 5-किग्रा बैग चीनी का नमूना माध्य 4.90 किलोग्राम के रूप में दर्ज किया जाता है। इस जानकारी के आधार पर निम्नलिखित प्रश्नों के उत्तर दीजिए:

i. 5-किग्रा बैग में शामिल चीनी की मात्रा के जनसंख्या माध्य के लिए 95% विश्वास अंतराल अनुमान की रचना करें। इन परिणामों के आधार पर, क्या आपको लगता है कि वितरक को चीनी निर्माता से शिकायत करने का अधिकार है? क्यों? **(6)**

ii. भाग (i) में 95% विश्वास अंतराल का निर्माण करने के लिए जनसंख्या वितरण के लिए आवश्यक महत्वपूर्ण धारणाएं क्या हैं? **(3)**

(ब) अनुमान लगाने के लिए आवश्यक नमूना आकार निर्धारित करें, 95% विश्वास के साथ, कि एक कैफे में औसत प्रति व्यक्ति बिल राशि ± 20 रुपये के भीतर है। मान लें कि जनसंख्या मानक विचलन रु.100 है| नमूना आकार गणना में प्रयुक्त Z मान की गणना के लिए आप किस एक्सेल फ़ंक्शन का उपयोग करेंगे? **(4.5)**

(स) मान लीजिए आपको निम्नलिखित डेटा दिया गया है: 5,3,7,10,-2,9,-1,2,-8,-2। उन स्थितियों का पता लगाने के लिए R कमांड लिखें जिन पर संख्याएँ शून्य से अधिक हैं। **(3)**

6. (a) A market research project aims to estimate the percent of adults living in a city who use HP laptops. Of the 500 adult people randomly sampled, 180 own HP laptop. Use a 90% confidence level to construct a confidence interval estimate for the population proportion of adult individuals of this city who use HP laptops. Interpret your result. **(4)**

(b) A researcher wants to compare the satisfaction of customers of the telecom operator A and telecom operator B. He selected a random sample of 180 customers of each telecom operator. Each survey respondent is asked to rate their operator on a scale of 1 to 5 where 1 stands for very dissatisfied and 5 stands for very satisfied; and mean customer satisfaction rating is calculated from the two samples. Sample from customers of operator A results in sample mean = 4.2 and sample standard deviation = 0.8. Sample from customers of operator B results in sample mean = 4.1 and sample standard deviation = 0.75. Assume that the random samples are independently selected from two populations and that the populations are normally distributed and have equal variances. Write the null and alternative hypotheses to test whether there is a difference in the customer satisfaction level between the two populations. For this hypothesis, compute the value of the pooled sample variance and pooled variance t test statistic. **(6)**

(c) What are the determinants of sampling error? Which Excel function can you use to calculate sampling error? **(3.5)**

(d) What is the use of dput() and dget() commands in R? **(3)**

6. (अ) एक बाजार अनुसंधान परियोजना का उद्देश्य एक शहर में रहने वाले वयस्कों के प्रतिशत का अनुमान लगाना है जो HP लैपटॉप का उपयोग करते हैं। यादृच्छिक ढंग से लिए गए 500 वयस्कों में से 180 के पास HP का लैपटॉप है। HP लैपटॉप का उपयोग करने वाले इस शहर के वयस्क व्यक्तियों के जनसंख्या अनुपात के लिए एक विश्वास अंतराल अनुमान बनाने के लिए 90% विश्वास स्तर का उपयोग करें। अपने परिणाम की व्याख्या करें। **(4)**

(ब) एक शोधकर्ता दूरसंचार ऑपरेटर A और दूरसंचार ऑपरेटर B के ग्राहकों की संतुष्टि की तुलना करना चाहता है। उसने प्रत्येक दूरसंचार ऑपरेटर के 180 ग्राहकों का यादृच्छिक सैंपल चुना। प्रत्येक सर्वेक्षण उत्तरदाता को अपने ऑपरेटर को 1 से 5 के पैमाने पर रेट करने के लिए कहा जाता है जहां 1 का अर्थ बहुत असंतुष्ट और 5 का अर्थ बहुत संतुष्ट होता है; और माध्य ग्राहक संतुष्टि रेटिंग की गणना दो नमूनों से की जाती है। ऑपरेटर A के ग्राहकों से सैंपल माध्य = 4.2 और सैंपल मानक विचलन = 0.8 में परिणाम देता है। ऑपरेटर B के ग्राहकों से सैंपल माध्य = 4.1 और सैंपल मानक विचलन = 0.75 में परिणाम देता है। मान लें कि यादृच्छिक नमूने स्वतंत्र रूप से दो आबादी से चुने गए हैं और आबादी सामान्य रूप से वितरित की जाती है और समान भिन्नताएं होती हैं। दो आबादी के बीच ग्राहक संतुष्टि के स्तर में अंतर है या नहीं, इसका परीक्षण करने के लिए शून्य और वैकल्पिक परिकल्पना लिखें। इस परिकल्पना के लिए, पूल किए गए सैंपल विचरण और पूलित विचरण t परीक्षण आंकड़ों के मूल्य की गणना करें। **(6)**

(स) नमूना त्रुटि के निर्धारक क्या हैं? नमूना त्रुटि की गणना के लिए आप किस एक्सेल फ़ंक्शन का उपयोग कर सकते हैं? **(3.5)**

(द) R में dput () और dget () कमांड का क्या उपयोग है? **(3)**