SET A

Sr. No. of Question Paper:

Unique Paper Code : 12273303

Name of the Course : B. A. (H) Economics

Name of the Paper : Data Analysis (SEC)

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) An engineering college has a total of 4000 full-time students. The Registrar of the college intends to take a probability sample of $n = 200$ students to measure students' satisfaction with online classes. Identify the population of interest in this setting. What type of sampling methods can be chosen? Critically analyze each possible method. **(10)**

   (b) Discuss the difference between the following Excel functions: RAND and RANDBETWEEN. Write the Excel function to generate a simple random sample with replacement of size 200 from a population of 4000. **(3)**

   (c) Write R commands for the following operations: **(3)**
   i. Round off 22/7 to the nearest 3 digits after decimal.
   ii. Round off 18/7 to the greatest integer.
   iii. Round off 17/5 to the least integer.

1. (अ) एक इंजीनियरिंग कॉलेज में कुल 4000 पूर्णकालिक छात्र हैं। कॉलेज के रजिस्ट्रार ऑनलाइन कक्षाओं के साथ छात्रों की संतुष्टि को मापने के लिए n = 200 छात्रों का संभाव्यता नमूने लेने का विचार करते हैं। इस स्थापना में रुचि की जनसंख्या निकालिये । किस प्रकार की सैंपल विधियों को चुना जा सकता है? प्रत्येक संभावित विधि का समालोचनात्मक विश्लेषण कीजिये । **(10)**

(ब) निम्नलिखित एक्सेल फ़ंक्शंस के बीच अंतर बताइये : RAND and RANDBETWEEN। 4000 की आबादी से आकार 200 के प्रतिस्थापन के साथ एक साधारण यादृच्छिक नमूना उत्पन्न करने के लिए एक्सेल फ़ंक्शन लिखें। **(3)**

(स) निम्नलिखित के लिए R कमांड बताएं : **(3)**

(i) 22/7 को दशमलव के बाद निकटतम 3 अंकों में पूर्णांकित करें।

(ii) 18/7 को सबसे बड़े पूर्णांक में पूर्णांकित करें।

(ii) 17/5 को सबसे छोटे पूर्णांक में पूर्णांकित करें।

2. (a) The frequency contingency table of 200 employees of a company, characterized by their gender (male or female) and the stress level faced at the workplace (high or low) is given below. Use this information to answer the following questions: **(10)**

	STRESS LEVEL	
Gender	**High**	**Low**
Male	50	70
Female	40	40

i. Prepare percentage contingency tables, based on row total, column total, and overall total. Based on this information, do you think that male employees are at a greater risk of having high stress levels compared to female employees?
ii. Calculate the percentage of employees who are females and have low stress level.

(b) Suppose raw data on gender and stress level faced by the employees is available in an Excel file in column B and C respectively from rows 2 to 200. Write down the steps to construct a frequency contingency table characterized by gender and stress level using COUNTIF or COUNTIFS Excel function. **(3)**

(c) Explain the command(s) used to import the Excel data file described in part (b) into R? Which R command will you use to construct a frequency contingency table of the employees characterized by their gender and stress level? **(3)**

2. (अ) एक कंपनी के 200 कर्मचारियों की आवृत्ति आकस्मिक तालिका, उनके लिंग (पुरुष या महिला) और कार्यस्थल पर सामना किए जाने वाले तनाव स्तर (उच्च या निम्न) की विशेषता नीचे दी गई है। निम्नलिखित प्रश्नों के उत्तर देने के लिए इस जानकारी का प्रयोग करें: **(10)**

	STRESS LEVEL	
Gender	**High**	**Low**
Male	50	70
Female	40	40

(i) पंक्ति योग, कुल कॉलम और कुल योग के आधार पर प्रतिशत आकस्मिक तालिकाएँ तैयार करें। इस जानकारी के आधार पर, क्या आपको लगता है कि महिला कर्मचारियों की तुलना में पुरुष कर्मचारियों को उच्च तनाव स्तर होने का अधिक जोखिम होता है?

(ii) उन कर्मचारियों के प्रतिशत की गणना करें जो महिलाएं हैं और जिनका तनाव का स्तर कम है।

(ब) मान लीजिए कि कर्मचारियों के लिंग और कर्मचारियों द्वारा सामना किए जाने वाले तनाव के स्तर पर डेटा क्रमशः कॉलम B और C में पंक्तियों 2 से 200 तक उपलब्ध है। COUNTIF या COUNTIFS एक्सेल फ़ंक्शन का उपयोग करके लिंग और तनाव स्तर की विशेषता वाली आवृत्ति आकस्मिक तालिका बनाने के चरणों को लिखें। **(3)**

(स) भाग (ब) में वर्णित एक्सेल डेटा फ़ाइल को R में आयात करने के लिए उपयोग की जाने वाली कमांड की व्याख्या करें। आप कर्मचारियों की उनके लिंग और तनाव के स्तर के आधार पर आवृत्ति आकस्मिक तालिका का निर्माण करने के लिए किस R कमांड का उपयोग करेंगे? **(3)**

3. (a) A ball manufacturer wants to compare diameters of three types of balls of different colours: Red, blue and green. A sample of 10 balls of each colour was selected, and the results representing the diameters of the balls (in mm), are as follows:

Red balls: 34,36,44,30,32,34,36,40,42,50

Blue balls: 40,42,34,23,45,36,38,37,30,33

Green balls: 34,35,36,37,38,39,32,31,30,40

i. For each of the three types of balls, compute and compare the mean and standard deviation. How would the mean diameters change if the last value for red balls was 35 instead of 50? Explain. **(10)**
ii. Which of the following Excel charts can represent the mean of the diameters of the three types of balls most efficiently and why: Scatter plot, Line plot, Column bar plot? Why? **(3)**

(b) Write R commands for the following: **(3)**

i. Make a bag of 10 balls of each of the three colours: Red, green and blue.
ii. Draw a sample of five balls, without replacement.

The following question is in lieu of Q3, Part (a)(ii), only for Visually impaired students:

Discuss the difference between discrete and continuous numerical variables with the help of examples. **(3)**

3. (अ) एक गेंद निर्माता विभिन्न रंगों की तीन प्रकार की गेंदों के व्यास की तुलना करना चाहता है: लाल, नीला और हरा। प्रत्येक रंग की 10 गेंदों का एक सैंपल चुना गया था, और गेंदों के व्यास का प्रतिनिधित्व करने वाले परिणाम (मिलीमीटर में), इस प्रकार हैं:

लाल गेंदें: 34,36,44,30,32,34,36,40,42,50

नीली गेंदें: 40,42,34,23,45,36,38,37,30,33

हरी गेंदें: 34,35,36,37,38,39,32,31,30,40

(i) तीन प्रकार की गेंदों में से प्रत्येक के लिए, माध्य और मानक विचलन की गणना और तुलना करें। यदि लाल गेंदों का अंतिम मान 50 के बजाय 35 था, तो माध्य व्यास कैसे बदलेंगे? समझाइये। **(10)**

(ii) निम्नलिखित में से कौन सा एक्सेल चार्ट तीन प्रकार की गेंदों के व्यास के माध्य का सबसे अधिक कुशलता से प्रतिनिधित्व कर सकता है और क्यों: स्कैटर प्लॉट, लाइन प्लॉट, कॉलम बार प्लॉट? **(3)**

(ब) निम्नलिखित के लिए R कमांड क्या हैं: **(3)**

(i) तीन रंगों में से प्रत्येक की 10 गेंदों का एक बैग बनाएं: लाल, हरा और नीला।

(ii) बिना बदले पांच गेंदों का एक नमूना बनाएं।

निम्नलिखित प्रश्न केवल दृष्टिबाधित छात्रों के लिए ***Q3,*** भाग (अ)***(ii)*** के स्थान पर है:

उदाहरणों की सहायता से असतत और सतत संख्यात्मक चर के बीच अंतर पर चर्चा करें। **(3)**

Section B

4. (a) Based on the descriptive statistics for the percentage annual return on the mutual funds given below, comment on the normality of the returns, using: **(6)**

i. Relationship between mean and median
ii. Relationship between interquartile range and standard deviation
iii. Relationship between range and standard deviation

Minimum	Q1	Q2	Q3	Maximum	Mean	Standard deviation	N
6.69	11.95	14.47	26.15	49.66	19.9	12.02	10

(b) Explain the measure of skewness and kurtosis of a distribution. Which Excel functions can you use to calculate these two measures? **(6)**

(c) Write R commands to construct the following matrices: **(4.5)**
(i) 4X4 matrix A using sequence of numbers from 1 to 16.
(ii) Matrix B, which is transpose of matrix A.
(iii) Matrix C, which is multiplication of matrix A with B: A*B

4. (a) नीचे दिए गए म्यूचुअल फंड पर प्रतिशत वार्षिक रिटर्न के लिए वर्णनात्मक आंकड़ों के आधार पर, निम्नलिखित का उपयोग करते हुए रिटर्न की सामान्यता पर टिप्पणी कीजिये: **(6)**

(i) माध्य और माध्यिका के बीच संबंध
(ii) अन्तःचतुर्थक रेंज और मानक विचलन के बीच संबंध
(iii) रेंज और मानक विचलन के बीच संबंध

Minimum	Q1	Q2	Q3	Maximum	Mean	Standard deviation	N
6.69	11.95	14.47	26.15	49.66	19.9	12.02	10

(ब) एक वितरण के वैषम्य और कुकुद्ता के माप की व्याख्या करें। इन दो मापों की गणना के लिए आप कौन से एक्सेल फ़ंक्शन का उपयोग कर सकते हैं? **(6)**

(स) निम्नलिखित मैट्रिक्स के निर्माण के लिए R कमांड क्या हैं: **(4.5)**

(i) 1 से 16 तक की संख्याओं के अनुक्रम का उपयोग करते हुए 4X4 मैट्रिक्स A।
(ii) मैट्रिक्स B, जो मैट्रिक्स A का स्थानान्तरण है।
(iii) मैट्रिक्स C, जो मैट्रिक्स A का B के साथ गुणा है: A * B.

5. A beverage distributor wants to estimate the amount of beverage contained in one-litre bottles purchased from a local beverage manufacturer. The manufacturer's specifications state that the standard deviation of the amount of beverage is equal to 0.33ml. A random sample of 900 bottles is selected, and the sample mean of beverage per one-litre bottle is recorded as 0.927ml. Based on this information, answer the following questions:

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

(b) Explain sampling error. How will the sampling error change when the number of bottles sampled in above part changes to 1089? Which Excel function can you use to calculate sampling error? **(6)**

(c) Suppose you have a data on marks of the students (out of 100). Write R command(s) for constructing a neatly labeled and colourful histogram, with unequal bins. Following are the breakpoints for the bins: 33, 50, 60, and 75 marks. **(4.5)**

The following question is in lieu of Q5, Part (c), only for Visually impaired students:

Explain the use of the following R commands: getwd() and setwd(). **(4.5)**

5. एक पेय वितरक स्थानीय पेय निर्माता से खरीदी गई एक लीटर की बोतलों में निहित पेय की मात्रा का अनुमान लगाना चाहता है। निर्माता के विनिर्देशों में कहा गया है कि पेय की मात्रा का मानक विचलन 0.33ml के बराबर है। 900 बोतलों का एक यादृच्छिक नमूना चुना जाता है, और प्रति लीटर बोतल में पेय का नमूना माध्य 0.927ml दर्ज किया जाता है। इस जानकारी के आधार पर निम्नलिखित प्रश्नों के उत्तर दीजिए:

(अ) जनसंख्या के लिए एक लीटर की बोतल में शामिल पेय की मात्रा के लिए 95% विश्वास अंतराल अनुमान का निर्माण करें। इन परिणामों के आधार पर, क्या आपको लगता है कि वितरक को पेय निर्माता से शिकायत करने का अधिकार है? क्यों? **(6)**

(ब) सैंपलिंग एरर की व्याख्या करें। सैंपलिंग एरर कैसे बदलेगा जब उपरोक्त भाग में सैंपल की गई बोतलों की संख्या 1089 में बदल जाती है? सैंपलिंग एरर की गणना के लिए आप किस एक्सेल फंक्शन का उपयोग कर सकते हैं? **(6)**

(स) मान लीजिए कि आपके पास छात्रों के अंकों (100 में से) का डेटा है। असमान बिन के साथ, साफ-सुथरे लेबल वाले और रंगीन हिस्टोग्राम के निर्माण के लिए R कमांड लिखें। बिन के लिए ब्रेकप्वाइंट निम्नलिखित हैं: 33, 50, 60, और 75 अंक। **(4.5)**

निम्नलिखित प्रश्न केवल दृष्टिबाधित छात्रों के लिए ***Q5, भाग (c) के स्थान पर है:***

निम्नलिखित R कमांड के उपयोग की व्याख्या करें: getwd () और setwd ()। **(4.5)**

6. A firm is operating in two cities: A and B, through its various outlets. The results for two-sample t-tests, assuming equal variances, for the revenue earned (in US$) from the two cities, is given below:

	City A	*City B*
Mean	**85.59**	**41.63**
Variance	**8222.64**	**1903.06**
Observations	**11**	**18**
Pooled Variance	**4243.64**	
Hypothesized Mean Difference	**0**	
df	**27**	
t Stat	**1.76**	
P(T<=t) one-tail	**0.04**	
P(T<=t) two-tail	**0.09**	
t Critical one-tail at 5% Level	**1.70**	
t Critical two-tail at 5% Level	**2.05**	

(a) Write the null and alternative hypotheses to test that the mean revenue earned in City A is greater than mean revenue earned in City B. **(2)**

(b) At the 0.05 level of significance, is there evidence of a difference in the mean of revenue earned in the two cities? Is it justified that the firm should focus more on one city? **(4)**

(c) Test the hypothesis stated in Part (b) again at 0.01 and 0.10 level of significance using p-value approach. Interpret your results. **(2)**

(d) Suppose an R file contains raw data on revenue from city A and B. Explain R command for performing t-test of average revenue earned in the two cities, assuming equal variances, for the hypothesis mentioned in Part (a) above at 5% significance level. Explain how the command will change when the test has to be conducted at 1% significance level and at 10% significance level? **(4.5)**

(e) Explain the Excel functions used for getting the Student's-t distribution and inverse of Student's t- distributions. **(4)**

6. एक फर्म अपने विभिन्न आउटलेट्स के माध्यम से दो शहरों, A और B, में काम कर रही है। दो शहरों से अर्जित राजस्व (अमेरिकी डॉलर में) के लिए समान भिन्नता मानकर दो-सैंपल t -परीक्षणों के परिणाम नीचे दिए गए हैं:

	City A	*City B*
Mean	**85.59**	**41.63**
Variance	**8222.64**	**1903.06**
Observations	**11**	**18**
Pooled Variance	**4243.64**	
Hypothesized Mean Difference	**0**	
df	**27**	
t Stat	**1.76**	
P(T<=t) one-tail	**0.04**	
P(T<=t) two-tail	**0.09**	
t Critical one-tail at 5% Level	**1.70**	
t Critical two-tail at 5% Level	**2.05**	

(ए) परीक्षण करने के लिए शून्य और वैकल्पिक परिकल्पना लिखें कि शहर A में अर्जित औसत राजस्व शहर B में अर्जित औसत राजस्व से अधिक है। **(2)**

(बी) 0.05 महत्व के स्तर पर, क्या दो शहरों में अर्जित राजस्व के माध्य में अंतर का प्रमाण है? क्या यह उचित है कि फर्म को एक शहर पर अधिक ध्यान देना चाहिए? **(4)**

(सी) ***p***-वैल्यू दृष्टिकोण का उपयोग करके 0.01 और 0.10 के महत्व के स्तर पर फिर से भाग (बी) में बताई गई परिकल्पना का परीक्षण करें। **(2)**

(डी) मान लीजिए कि एक **R** फाइल में शहर A और B से राजस्व पर कच्चा डेटा है। उपरोक्त भाग (ए) में उल्लिखित परिकल्पना के लिए 5% महत्व स्तर पर, समान भिन्नता मानकर, दो शहरों में अर्जित औसत राजस्व का ***t***-टेस्ट करने के लिए **R** कमांड की व्याख्या करें। जब परीक्षण 1% महत्व स्तर और 10% महत्व स्तर पर आयोजित किया जाना है तो कमांड कैसे बदलेगा? **(4.5)**

(ई) छात्र के ***t*** वितरण और छात्र के ***t*** वितरण के विपरीत प्राप्त करने के लिए उपयोग किए जाने वाले एक्सेल कार्यों की व्याख्या करें। **(4)**