

**SEMESTER 1 EXAMINATIONS 2022/2023**

**MODULE:** CA660/A - Statistical Data Analysis

**PROGRAMME(S):**

MCM M.Sc. in Computing

ECSA Study Abroad (Engineering & Computing)

CAPT PhD-track

ECSAO Study Abroad (Engineering & Computing)

GCAI Grad Cert in Artificial Intelligence

EEPT PhD-track **YEAR OF STUDY:** 1,2,O,X

|  |  |  |
| --- | --- | --- |
| **EXAMINER(S):** |  |  |
| Dr. Marija Bezbradica | (Internal) | (Ext:5260) |
| Prof. Giorgio Giacinto | (External) | External |
| Dr. Katarina Domijan | (External) | External |

**TIME ALLOWED:** 3 Hours

**INSTRUCTIONS: Answer 4 questions. All questions carry equal marks.**

**PLEASE DO NOT TURN OVER THIS PAGE UNTIL YOU ARE INSTRUCTED TO DO SO.**

The use of programmable or text storing calculators is expressly forbidden.

Please note that where a candidate answers more than the required number of questions, the examiner will mark all questions attempted and then select the highest scoring ones.

***Requirements for this paper:***

***1. Statistical Tables***

***QUESTION 1 [TOTAL MARKS: 25]***

**Q 1(a) [10 Marks]**

i. Explain the difference between *descriptive* and *inferential statistics data* and how they are used in statistical analysis. Give an example of each.

[4 marks]

Descriptive and Inferential Statistics

Descriptive Statistics describe and summarize data using measures like mean, median, standard deviation, histograms, and boxplots. These techniques provide a clear picture of the data's characteristics without making inferences about a larger population [1, 2].

Example: In an examination results dataset, descriptive statistics can calculate the average score, the range of scores, and create a histogram showing the distribution of grades [1, 3]. This gives insights into the overall performance of the class.

Inferential Statistics use sample data to make generalizations or predictions about a larger population. They involve hypothesis testing, confidence intervals, and statistical models like ANOVA to draw conclusions beyond the observed data [4, 5].

Example: To determine if a new drug is effective in lowering blood pressure, inferential statistics analyze data from a sample of patients taking the drug. Techniques like t-tests or confidence intervals assess whether the observed decrease in blood pressure is statistically significant, allowing researchers to infer the drug's effectiveness for a larger population [6, 7].

ii. Identify all scales of measurement that apply to the following data:

[4 marks]

**Nominal**（名义）、**Ordinal**（顺序）、**Interval**（区间）、**Ratio**（比例）

a. Country of birth (Ireland, India, UK)

b. Daily amount of water intake

c. Average height of a student

d. Kitchen material (pans, cutlery, cutting board, plates)

iii. Worldwide production of grain (in million metric tons), by country, was recorded as in the below table. Can classifications be considered

independent or mutually exclusive or neither? Explain your reasoning.

[2 marks]

|  |  |
| --- | --- |
| **Country** | **Grain Production**  **(million metric tons)** |
| United States | 5,338 |
| China | 3,611 |
| Spain | 2,233 |
| Germany | 1,872 |
| India | 1,524 |
| Canada | 1,042 |
| UK | 797 |
| France | 676 |
| Italy | 671 |

**ndependent events** mean that the occurrence of one event does not affect the probability of the other event occurring. In this case, grain production in one country can potentially influence grain production in other countries. For example, global trade policies, climate patterns, or technological advancements can create dependencies between grain production in different regions.

●

**Mutually exclusive events** cannot happen at the same time. It's clear that a country can have grain production and also belong to a specific continent or region. For instance, the United States produces grain and is located in North America. Therefore, these classifications are not mutually exclusive.

**Q 1(b) [3 Marks]**

Tickets numbered 1 to 20 are mixed up and then a ticket is drawn at random. What is the probability that the ticket drawn has a number which is a multiple of 3 or of 5? (You can leave your answers in a fraction form).

**Q 1(c) [6 Marks]**

Consider a box containing 12 red balls, 8 green balls and 6 blue balls. Let 2 balls

be drawn at random without replacement from the box. Please give the probabilities of the following events. Please show your reasoning in each case. (You can leave your answers in a fraction form).

|  |  |  |
| --- | --- | --- |
| i. | None of the balls drawn is blue? | [3 Marks] |
| ii. | Neither of the balls is red nor green? | [3 Marks] |

**Q 1(d) [6 Marks]**

The Figure below is showing two distributions (marked with a full line and a dashed line). Comment on what you can conclude about these distributions. In short bullet points write all the information as we can get from observing these two plots.

If you are an investor who is prone to risk, would you rather invest in shares represented with a full line or with a dashed line?

Observations on Distributions and Risk Preference

●

**The dashed line distribution is more concentrated around its mean compared to the full line distribution.**1 This indicates that investments represented by the dashed line tend to have returns closer to the average, with less variation.

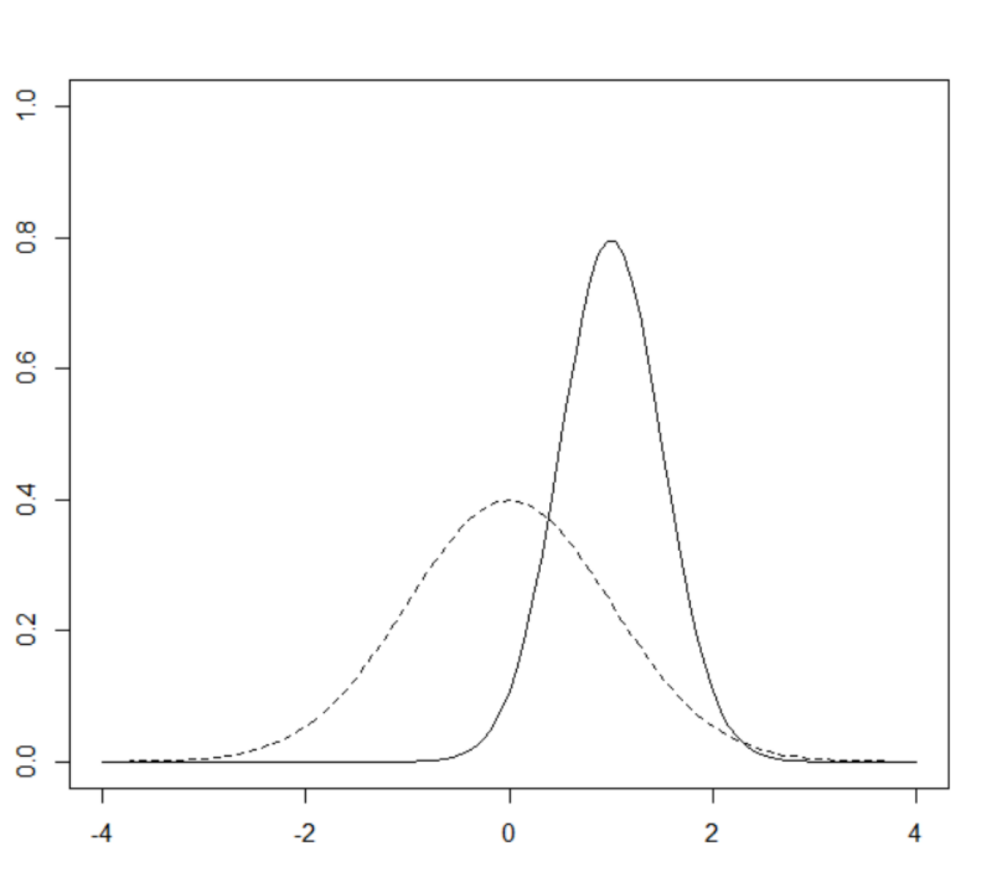
●

**The full line distribution has a wider spread, with its tails extending further to both sides.**1 This suggests a higher probability of both larger gains and larger losses, making the full line investment riskier but also potentially more rewarding.

●

**A risk-prone investor would likely favor the shares represented by the full line.**1 The wider spread indicates higher volatility and a greater chance of extreme outcomes, aligning with a risk-seeking investment strategy.

Additional Considerations for Investors



*Figure for question Q 1(d)*

***[End of Question 1]***

***QUESTION 2 [TOTAL MARKS: 25]***

**Q 2(a) [4 Marks]**

In using probability distributions, indicate when the following might be used for data description and modelling (give a simple example for each distribution below):

### \*\*When to Use Binomial and Student's t-Distribution\*\*

\* \*\*i. Binomial distribution\*\*

The binomial distribution is used when an experiment has the following characteristics:

\* Consists of a fixed number of trials (n).

\* Each trial has only two possible outcomes, typically labelled as "success" and "failure."

\* The probability of success (p) remains constant for each trial.

\* The trials are independent, meaning the outcome of one trial doesn't influence the others.

\*\*Simple Example:\*\* Consider tossing a coin 10 times. The number of heads obtained follows a binomial distribution with n=10 (number of tosses) and p=0.5 (probability of getting heads on a single toss). The binomial distribution can help calculate probabilities like the chance of getting exactly 6 heads, at least 3 heads, etc.

\* \*\*ii. Student's t-distribution\*\*

The Student's t-distribution is used when:

\* The sample size is small (typically n<30).

\* The population standard deviation is unknown and needs to be estimated from the sample.

This distribution is similar in shape to the normal distribution but has heavier tails, accounting for the greater uncertainty associated with small samples.

\*\*Simple Example:\*\* Imagine you want to estimate the average height of students in a university. You collect a random sample of 15 students and measure their heights. Since you don't know the population standard deviation, you would use the Student's t-distribution to construct a confidence interval for the average height or to test a hypothesis about the true average height.

|  |  |  |
| --- | --- | --- |
| i. | Binomial distribution. | [2 marks] |
| ii. | Student’s t-distribution. | [2 marks] |
| **Q 2(b)** |  | **[6 Marks]** |

The mean number of bacteria per millilitre of a liquid is known to be 6. Find the probability that in 1 ml of the liquid, there will be:

|  |  |  |
| --- | --- | --- |
| i. | less than 4. | [3 Marks] |
| ii. | exactly 2 bacteria. | [3 Marks] |

**Q 2(c)**

A coin is tossed 10 times. What is the probability of:

i. Getting exactly 6 heads? ii. Getting up to 3 heads?

**Q 2(d)**

**[6 Marks]**

[3 Marks] [3 Marks]

**[9 Marks]**

A manufacturer fills jars with M&Ms. The weight of M&Ms in ajar can be modelled by a normal distribution with mean 252 grams and standard deviation 5 grams. Find the probability that ajar has:

i. Less than 242 grams. [3 marks]

ii. Between 250 and 262 grams. [3 marks]

iii. More than 260 grams. [3 marks]

**[End of Question 2]**

**Q 3(a) [5 Marks]**

Suppose you are walking down the street and encounter an old fallen tree. Your

initial explanation is that it fell of old age. But then you hear on the TV that there has been a series of flash floods in your area, which is a very rare event there. Your

belief that the tree fell by itself decreases.

Please list all the different probabilities and conditional probabilities for this Bayesian network.

To analyze this scenario as a Bayesian network, we identify the events and their relationships. The key variables are:

1. AAA: Tree fell by itself (old age).
2. FFF: Flash floods occurred (rare event).
3. TTT: Observation of the fallen tree.

The Bayesian network establishes the relationships between these events. Here are the probabilities and conditional probabilities involved:

### ****1. Prior Probabilities****

These represent the baseline beliefs about events without additional information:

* P(A)P(A)P(A): Probability that the tree fell by itself (old age).
* P(F)P(F)P(F): Probability that flash floods occurred.

### ****2. Likelihood Probabilities****

These describe the probability of observing the tree fallen given the causes:

* P(T∣A)P(T | A)P(T∣A): Probability of observing the tree fallen, given it fell by itself.
* P(T∣F)P(T | F)P(T∣F): Probability of observing the tree fallen, given flash floods occurred.

### ****3. Joint Probabilities****

The joint probability of events occurring together:

* P(A∩T)P(A \cap T)P(A∩T): Probability that the tree fell by itself and it was observed fallen.
* P(F∩T)P(F \cap T)P(F∩T): Probability that flash floods occurred and the tree was observed fallen.

### ****4. Marginal Probabilities****

These describe the total probability of a single event occurring:

* P(T)P(T)P(T): Overall probability of observing the tree fallen, regardless of cause.
  + This is calculated using the **law of total probability**: P(T)=P(T∣A)P(A)+P(T∣F)P(F)P(T) = P(T | A)P(A) + P(T | F)P(F)P(T)=P(T∣A)P(A)+P(T∣F)P(F)

### ****5. Posterior Probabilities****

The posterior probabilities update beliefs about the causes of the tree falling after observing it fallen:

* P(A∣T)P(A | T)P(A∣T): Probability that the tree fell by itself, given it was observed fallen.
  + Using Bayes' theorem: P(A∣T)=P(T∣A)P(A)P(T)P(A | T) = \frac{P(T | A)P(A)}{P(T)}P(A∣T)=P(T)P(T∣A)P(A)​
* P(F∣T)P(F | T)P(F∣T): Probability that flash floods caused the tree to fall, given it was observed fallen.
  + Using Bayes' theorem: P(F∣T)=P(T∣F)P(F)P(T)P(F | T) = \frac{P(T | F)P(F)}{P(T)}P(F∣T)=P(T)P(T∣F)P(F)​

### ****Impact of New Evidence****

Hearing about flash floods on TV changes the prior probability of FFF. The updated probabilities will shift:

* If P(F)P(F)P(F) increases, P(A∣T)P(A | T)P(A∣T) will decrease because P(T)P(T)P(T) is influenced more by P(T∣F)P(F)P(T | F)P(F)P(T∣F)P(F).

This Bayesian framework allows us to reason systematically about how observing the tree and hearing about floods affects our belief about why the tree fell.

**Q 3(b) [10 Marks]**

Approximately 2% of population aged 25-34 had a COVID-19 infection during the

recent peak. A person with COVID-19 has a 95% chance of a positive antigen test, while a person without the infection has a 12% chance of a false positive result.

What is the probability that a person does not have COVID-19 given that he just had a positive antigen test?

We can solve this problem using **Bayes' Theorem**, which states:

P(No COVID∣Positive Test)=P(Positive Test∣No COVID)⋅P(No COVID)P(Positive Test)P(\text{No COVID} | \text{Positive Test}) = \frac{P(\text{Positive Test} | \text{No COVID}) \cdot P(\text{No COVID})}{P(\text{Positive Test})}

### ****Definitions and Given Probabilities****

Let:

* CC: Person has COVID-19.
* CcC^c: Person does not have COVID-19.
* T+T^+: Positive antigen test result.

From the problem:

* P(C)=0.02P(C) = 0.02 (2% of the population had COVID-19).
* P(Cc)=1−P(C)=0.98P(C^c) = 1 - P(C) = 0.98.
* P(T+∣C)=0.95P(T^+ | C) = 0.95 (true positive rate).
* P(T+∣Cc)=0.12P(T^+ | C^c) = 0.12 (false positive rate).

### ****Step 1: Find the Total Probability of a Positive Test (****P(T+)P(T^+)****)****

Using the **law of total probability**:

P(T+)=P(T+∣C)P(C)+P(T+∣Cc)P(Cc)P(T^+) = P(T^+ | C)P(C) + P(T^+ | C^c)P(C^c)

Substitute the given probabilities:

P(T+)=(0.95)(0.02)+(0.12)(0.98)P(T^+) = (0.95)(0.02) + (0.12)(0.98) P(T+)=0.019+0.1176=0.1366P(T^+) = 0.019 + 0.1176 = 0.1366

### ****Step 2: Apply Bayes' Theorem****

Now, calculate P(Cc∣T+)P(C^c | T^+):

P(Cc∣T+)=P(T+∣Cc)P(Cc)P(T+)P(C^c | T^+) = \frac{P(T^+ | C^c)P(C^c)}{P(T^+)}

Substitute the known values:

P(Cc∣T+)=(0.12)(0.98)0.1366P(C^c | T^+) = \frac{(0.12)(0.98)}{0.1366} P(Cc∣T+)=0.11760.1366≈0.8607P(C^c | T^+) = \frac{0.1176}{0.1366} \approx 0.8607

### ****Final Answer****

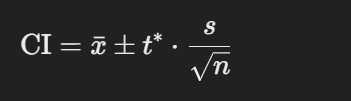
The probability that a person does **not** have COVID-19 given a positive antigen test is approximately **86.07%**.

**Q 3(c) [10 Marks]**

A domestic natural gas distribution company provides and maintains gas connections for several hundred households. During a three-month period, the number of serious technical problems for a sample of 12 of the households had a total mean of 7.5 with a standard deviation of 5.12:

i. Obtain 95% and 99% confidence intervals for the average number of gas supply issues in a three-month period. Why is the t-distribution used here?

[6 marks]

The t-distribution is used because the sample size is small, and the population standard deviation is unknown.

ii. Suppose the mean and standard deviation remain as given but that the

sample size is 50 rather than 12. Test the hypothesis that the average number of problems per household is 6 at the 5% level of significance. [4 marks]

***QUESTION 4 [TOTAL MARKS: 25]***

**Q 4(a) [6 Marks]**

Contrast when each of the following is used, based on type of the data and principal hypotheses:

|  |  |
| --- | --- |
| i. ANOVA | [3 marks] |
| ii. Chi-squared (Contingency Table). | [3 marks] |

**Summary Table Contrasting ANOVA and Chi-Squared**

|  |  |  |
| --- | --- | --- |
| Feature | ANOVA | Chi-Squared (Contingency Table) |
| Data Type | Continuous | Categorical |
| Principal Hypothesis | Difference in means across multiple groups | Association between two categorical variables |
| Test Statistic | F-statistic (ratio of between-group variance to within-group variance) | Chi-squared statistic (measures difference between observed and expected frequencies) |
| Assumptions | Independence, normality, homogeneity of variance**Independence of Observations:** The data points within each group should be independent of each other. This means that the value of one observation should not influence the value of another observation within the same group. Violations of this assumption can occur if there are hidden relationships or confounding variables that are not accounted for.  2.  **Normality of the Response Variable:** The response variable should follow a normal distribution within each group. ANOVA is robust to moderate departures from normality, especially with larger sample sizes. However, severe departures from normality can affect the reliability of the results.  3.  **Homogeneity of Variance:** The variances of the response variable should be similar across all the groups being compared. This assumption ensures that the groups have comparable levels of variability. If the variances are significantly different, it may indicate that ANOVA is not appropriate, and alternative methods like Welch's ANOVA or non-parametric tests may be considered. | Independence of observations |
| Applications | Comparing mean crop yields across different soil types, analyzing the effects of different teaching methods on student test scores | Examining the relationship between gender and political affiliation, testing if there's an association between smoking and lung cancer |

**Q 4(b) [10 Marks]**

A frequency modulation device is used to generate 60 random frequencies in the range 1 - 100 MHz. Test, at the 0.025 level of significance, if the results are

consistent with the hypothesis that the outcomes are randomly distributed.

[hint: use Chi-Square Test of Goodness of Fit here].

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Range | 1-25 | 26-50 | 51-75 | 76-100 | Total |
| Observed Number | 13 | 13 | 16 | 18 | 60 |
| Expected number | 15 | 15 | 15 | 15 | 60 |

To test if the observed frequencies are consistent with the hypothesis that the outcomes are randomly distributed, we will use a **Chi-Square Test of Goodness of Fit**.

### ****Step-by-step solution:****

#### ****1. State the hypotheses:****

* Null hypothesis (H0H\_0): The observed frequencies are consistent with the expected frequencies, i.e., the outcomes are randomly distributed.
* Alternative hypothesis (HaH\_a): The observed frequencies are not consistent with the expected frequencies, i.e., the outcomes are not randomly distributed.

#### ****2. Chi-Square test statistic:****

The formula for the Chi-Square test statistic is:

χ2=∑(Oi−Ei)2Ei\chi^2 = \sum \frac{(O\_i - E\_i)^2}{E\_i}

where:

* OiO\_i = Observed frequency for the ii-th category.
* EiE\_i = Expected frequency for the ii-th category.

#### ****3. Calculate the Chi-Square statistic:****

The observed frequencies and expected frequencies are provided in the table:

| **Range** | **Observed Number (OiO\_i)** | **Expected Number (EiE\_i)** | **(Oi−Ei)2(O\_i - E\_i)^2** | **(Oi−Ei)2Ei\frac{(O\_i - E\_i)^2}{E\_i}** |
| --- | --- | --- | --- | --- |
| 1-25 | 13 | 15 | (13−15)2=4(13 - 15)^2 = 4 | 415=0.267\frac{4}{15} = 0.267 |
| 26-50 | 13 | 15 | (13−15)2=4(13 - 15)^2 = 4 | 415=0.267\frac{4}{15} = 0.267 |
| 51-75 | 16 | 15 | (16−15)2=1(16 - 15)^2 = 1 | 115=0.067\frac{1}{15} = 0.067 |
| 76-100 | 18 | 15 | (18−15)2=9(18 - 15)^2 = 9 | 915=0.600\frac{9}{15} = 0.600 |

Now sum the Chi-Square contributions for each category:

χ2=0.267+0.267+0.067+0.600=1.201\chi^2 = 0.267 + 0.267 + 0.067 + 0.600 = 1.201

#### ****4. Degrees of freedom (df):****

The degrees of freedom for a Chi-Square test of goodness of fit is:

df=k−1df = k - 1

where kk is the number of categories. Here, we have 4 categories (ranges), so:

df=4−1=3df = 4 - 1 = 3

#### ****5. Find the critical value:****

At the 0.025 significance level and 3 degrees of freedom, the critical value for the Chi-Square distribution can be found from the Chi-Square table or using statistical software. For df=3df = 3 and α=0.025\alpha = 0.025, the critical value is approximately **9.348**.

#### ****6. Decision rule:****

* If the calculated χ2\chi^2 statistic is greater than the critical value, reject the null hypothesis.
* If the calculated χ2\chi^2 statistic is less than or equal to the critical value, do not reject the null hypothesis.

#### ****7. Conclusion:****

* The calculated χ2=1.201\chi^2 = 1.201.
* The critical value is 9.3489.348.

Since 1.201<9.3481.201 < 9.348, **we do not reject the null hypothesis**.

### ****Final Conclusion:****

There is not enough evidence at the 0.025 level of significance to conclude that the outcomes are not randomly distributed. Therefore, the observed frequencies are consistent with the hypothesis that the outcomes are randomly distributed.

**Q 4(c) [9 Marks]**

Several different exam formats were given to students to test if there is any

difference in population means for the exams. Students were put into *block*s, as we know that their abilities differ, and the effect of exam format was tested as a *factor*.

The *randomized block design* of the experiment performed gave the following

values: the *degrees of freedom* for the error sum of squares is 12, while that for the factor sum of squares is 3.

i. Calculate the number of blocks used in the experiment. [3 marks]

If the total sum of squares = 6230, the error sum of squares = 720 and the sum of squares due to blocks = 4480:

ii. Construct the appropriate ANOVA table. [3 marks]

iii. Perform appropriate tests of hypotheses (at 0.01 level of significance) and state your conclusions. [3 marks]

### i. ****Calculate the Number of Blocks Used in the Experiment****

In a randomized block design, the degrees of freedom for the **error** sum of squares (SS\_error) is calculated as:

dferror=(b−1)(t−1)df\_{\text{error}} = (b-1)(t-1)

Where:

* bb is the number of blocks.
* tt is the number of treatments.

Given:

* dferror=12df\_{\text{error}} = 12
* dffactor=3df\_{\text{factor}} = 3 (This corresponds to t−1=3t - 1 = 3, so t=4t = 4).

Using the degrees of freedom for the error:

dferror=(b−1)(t−1)df\_{\text{error}} = (b-1)(t-1)

Substituting dferror=12df\_{\text{error}} = 12 and t=4t = 4:

12=(b−1)(4−1)=(b−1)×312 = (b - 1)(4 - 1) = (b - 1) \times 3

Solving for bb:

b−1=123=4b - 1 = \frac{12}{3} = 4 b=5b = 5

Therefore, there are **5 blocks** in the experiment.

### ii. ****Construct the ANOVA Table****

We are given the following information:

* Total Sum of Squares (SS\_total) = 6230
* Error Sum of Squares (SS\_error) = 720
* Sum of Squares due to Blocks (SS\_blocks) = 4480
* The degrees of freedom for the **error** (df\_error) = 12
* The degrees of freedom for the **factor** (df\_factor) = 3

Let's calculate the remaining values and construct the ANOVA table.

#### Step 1: ****Sum of Squares for Treatments (SS\_treatments)****

SStreatments=SStotal−SSblocks−SSerrorSS\_{\text{treatments}} = SS\_{\text{total}} - SS\_{\text{blocks}} - SS\_{\text{error}} SStreatments=6230−4480−720=1030SS\_{\text{treatments}} = 6230 - 4480 - 720 = 1030

#### Step 2: ****Degrees of Freedom (df)****

* dftotal=n−1df\_{\text{total}} = n - 1 (where nn is the total number of observations, which can be calculated using the degrees of freedom for error and blocks).
* dfblocks=b−1=5−1=4df\_{\text{blocks}} = b - 1 = 5 - 1 = 4
* dftreatments=t−1=4−1=3df\_{\text{treatments}} = t - 1 = 4 - 1 = 3
* dferror=12df\_{\text{error}} = 12
* dftotal=dfblocks+dftreatments+dferror=4+3+12=19df\_{\text{total}} = df\_{\text{blocks}} + df\_{\text{treatments}} + df\_{\text{error}} = 4 + 3 + 12 = 19

#### Step 3: ****Mean Squares (MS)****

The mean square is the sum of squares divided by the respective degrees of freedom.

MStreatments=SStreatmentsdftreatments=10303=343.33MS\_{\text{treatments}} = \frac{SS\_{\text{treatments}}}{df\_{\text{treatments}}} = \frac{1030}{3} = 343.33 MSblocks=SSblocksdfblocks=44804=1120MS\_{\text{blocks}} = \frac{SS\_{\text{blocks}}}{df\_{\text{blocks}}} = \frac{4480}{4} = 1120 MSerror=SSerrordferror=72012=60MS\_{\text{error}} = \frac{SS\_{\text{error}}}{df\_{\text{error}}} = \frac{720}{12} = 60

#### Step 4: ****F-statistics****

Ftreatments=MStreatmentsMSerror=343.3360=5.72F\_{\text{treatments}} = \frac{MS\_{\text{treatments}}}{MS\_{\text{error}}} = \frac{343.33}{60} = 5.72 Fblocks=MSblocksMSerror=112060=18.67F\_{\text{blocks}} = \frac{MS\_{\text{blocks}}}{MS\_{\text{error}}} = \frac{1120}{60} = 18.67

#### ANOVA Table

| **Source** | **Sum of Squares (SS)** | **Degrees of Freedom (df)** | **Mean Square (MS)** | **F-statistic (F)** |
| --- | --- | --- | --- | --- |
| Treatments | 1030 | 3 | 343.33 | 5.72 |
| Blocks | 4480 | 4 | 1120 | 18.67 |
| Error | 720 | 12 | 60 |  |
| Total | 6230 | 19 |  |  |

### iii. ****Perform the Hypothesis Tests (at 0.01 level of significance)****

#### Null and Alternative Hypotheses for Treatments:

* H0H\_0: There is no significant difference in the population means for the different exam formats.
* HAH\_A: There is a significant difference in the population means for the different exam formats.

For the test of treatments, the F-statistic is Ftreatments=5.72F\_{\text{treatments}} = 5.72.

Using an F-distribution table or statistical software, we can compare the calculated F-statistic with the critical F-value at α=0.01\alpha = 0.01, df1=3df\_1 = 3, and df2=12df\_2 = 12.

The critical value for F(3,12)F(3, 12) at the 0.01 significance level is approximately **5.93**.

Since 5.72<5.935.72 < 5.93, we **fail to reject** the null hypothesis. This suggests that there is no statistically significant difference in the population means for the different exam formats at the 0.01 significance level.

#### Conclusion for Treatments:

* We **fail to reject** H0H\_0. There is not enough evidence to conclude that there is a significant difference in the population means for the different exam formats.

#### Null and Alternative Hypotheses for Blocks:

* H0H\_0: There is no significant difference in the blocks (differences in students' abilities).
* HAH\_A: There is a significant difference in the blocks (differences in students' abilities).

For the test of blocks, the F-statistic is Fblocks=18.67F\_{\text{blocks}} = 18.67.

Using an F-distribution table or statistical software, we compare the calculated F-statistic with the critical F-value at α=0.01\alpha = 0.01, df1=4df\_1 = 4, and df2=12df\_2 = 12.

The critical value for F(4,12)F(4, 12) at the 0.01 significance level is approximately **6.94**.

Since 18.67>6.9418.67 > 6.94, we **reject** the null hypothesis. This suggests that there is a significant difference in students' abilities, as indicated by the block effect.

#### Conclusion for Blocks:

* We **reject** H0H\_0. There is enough evidence to conclude that there is a significant difference in students' abilities, which was considered as a block in the experiment.

### Final Summary:

* For treatments (exam formats), there is no significant difference in population means.
* For blocks (students' abilities), there is a significant difference.

**[End of Question 4]**

***QUESTION 5 [TOTAL MARKS: 25]***

**Q 5(a) [7 Marks]**

Describe clearly and consistently when would you use non-parametric (distribution

free) methods when performing statistical data analyses, stating the advantages and disadvantages over using parametric methods. Give an illustration and an example.

**Q 5(b) [8 Marks]**

The average scores when testing games, with two different experimental setups, were obtained as follows:

**Experiment 1** 50 75 92 110

**Experiment 2** 150 180 209 320

i. Use the Wilcoxon-Mann-Whitney to test that the medians are equal at 5% level of significance. [6 Marks]

ii. If data from a further independent experimental method were available, what non-parametric technique would you use? [2 Marks]

**Q 5(c) [10 Marks]**

In a financial survey to assess the effectiveness of different investment strategies on improving portfolio performance, an investment consultancy asked 10 major

investors (the blocks) to rank strategies A, B, C, according to the investment gain.

Perform and report results on an appropriate non-parametric test to determine if

rankings indicate a difference in effectiveness of the different strategies? (Use level of significance α=0.10).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Major investors** | | | | | | | | | |
|  | **a** | **b** | **c** | **d** | **e** | **f** | **g** | **h** | **i** | **j** |
| A | 3 | 2 | 3 | 2 | 3 | 2 | 1 | 2 | 3 | 1 |
| B | 2 | 1 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 |
| C | 1 | 3 | 2 | 3 | 1 | 3 | 3 | 3 | 2 | 3 |

***[End of Question 5]***

***[END OF EXAM]***