

Contents

Sample Papers

- Current Version : Friday 19th April 2013
- End of year exam paper is worth 70% of module grade.
- There will be five questions. You must attempt four.
- This sample paper will have multiple variants for each of the five questions.

Q.1 Descriptive Statistics (Variant 1)

Part A

Data on the durations (measured in months) were collected for a random sample of product development projects. The durations for these development projects were collected and tabulated as follows:

12	11	20	19	18	9	16	15
----	----	----	----	----	---	----	----

- i. (1 Mark) Calculate the mean of the project durations.
- ii. (2 Marks) Calculate the variance for this sample.
- iii. (1 Mark) Calculate the standard deviation for this sample.

Part B

A telecommunications company has two servers from different suppliers. The Operations Director randomly selected 9 months from the previous year's figures. The down time figures for each server is as follows (down time per month in hours)

	Server 1	Server 2
1	5.3	5.4
2	5.4	5.5
3	5.5	6.1
4	5.2	5.0
5	5.1	5.1
6	5.6	5.2
7	5.4	5.3
8	5.5	5.4
9	5.6	5.9
10	5.6	5.2

- i. (8 Marks) You are required to construct a box plot for each server's output and to comment on the salient features of each plot. Is there evidence from the box plots to reject the hypothesis that there is no difference in the down time between the two servers? Use the box plot to justify your comments.
- ii. (2 Marks) Compare and contrast the Mean and the Median as measures of location

Q.1 Descriptive Statistics (Variant 2)

Part A

The exam results for a class of 40 students are tabulated below.

20	36	37	37	41	43	44	44	44	45
49	50	51	52	52	52	52	53	53	53
54	55	55	57	57	59	59	59	62	62
65	66	67	70	71	74	86	87	89	98

- i. (4 Marks) Summarize the data in the above table using a relative frequency table and a cumulative relative frequency table. Use 6 class intervals, with 11 as the lower limit of the first interval.
- ii. (5 Marks) Draw a histogram for the above data. Comment on the shape of the histogram. Based on the shape of the histogram, what is the best measure of centrality and variability?
- iii. (7 Marks) Construct a box plot for the above data. Clearly demonstrate how all of the necessary values were computed.

Part B

Consider the following data set of seven numbers:

29 14 17 30 19 25 13

For this sample, compute the following descriptive statistics:

- i. (1 Mark) The mean,
- ii. (2 Mark) The variance,
- iii. (1 Mark) The standard deviation.

(End of Question)

Q.1 Descriptive Statistics (Variant 3)**Part A**

The mass of 30 individuals (in kg) was observed. The results are given below

37	44	48	51	53	56	58	60	62	63
65	67	69	70	72	74	76	77	79	81
83	86	88	91	94	97	101	106	113	127

- i. (6 Marks) Construct a relative frequency and cumulative relative frequency table for this data set.
- i. (6 Marks) Draw a histogram for these data. Comment on the shape of the histogram.
- ii. (8 Marks) Draw the boxplot for this sample.

(End of Question)

Q.2 Probability (Variant 1)

Part A

The probability distribution of discrete random variable X is tabulated below. There are 5 possible outcomes of X , i.e. 1, 2, 3, 5, 10 and 20.

x_i	1	2	5	10	20
$P(x_i)$	0.10	0.25	0.30	0.20	0.15

- i. (3 Marks) Determine the expected value $E(X)$.
- ii. (3 Marks) Evaluate $E(X^2)$.
- iii. (2 Marks) Compute the variance of random variable X .

Part B

An electronics assembly subcontractor receives its entire supply of resistors from two suppliers. Company A provides 70% of the subcontractor's resistors, while company B supplies the remainder. The additional information has also been made available.

- 2% of the resistors provided by company A failed the final test,
- 3% of company B's resistors also fail final test.

Answer the following questions:

- i. (3 Marks) What is the probability that a resistor fails the final test?
- ii. (3 Marks) What is the probability that a resistor fails the final test given that the resistor in question came from company A?

Q.2 Probability (Variant 2)

Part A

The probability distribute of discrete random variable X is tabulated below. There are 5 possible outcome of X , i.e. 1, 2, 4, 6 and 8.

x_i	1	2	4	6	8
$p(x_i)$	0.50	0.15	0.20	0.05	0.10

- (2 Mark) What is the expected value of X ?
- (2 Mark) Given that $E(X^2) = 12.5$, compute the variance of X .

Part B

On completion of a programming project, four programmers from a team submit a collection of subroutines to an acceptance group. The following table shows the percentage of subroutines each programmer submitted and the probability that a subroutine submitted by each programmer will pass the certification test based on historical data.

Programmer	A	B	C	D
Proportion of subroutines submitted	0.10	0.20	0.30	0.40
Probability of acceptance	0.55	0.60	0.95	0.75

- (3 Marks) What is the proportion of subroutines that pass the acceptance test?
- (3 Marks) After the acceptance tests are completed, one of the subroutines is selected at random and found to have passed the test. What is the probability that it was written by Programmer A?

Part c

A doctor treating a patient issues a prescription for antibiotics and provides for two repeat prescriptions. The probability that the infection will be cleared by the first prescription is $p_1 = 0.6$. The probability that successive treatments are successful, given that previous prescriptions were not successful are $p_2 = 0.5$, $p_3 = 0.4$. Calculate the probability that:

- (2 Marks) a patient will require the third prescription,
- (2 Marks) the patient is still infected after the third prescription,
- (2 Marks) the patient is cured by the second prescription, given that the patient is eventually cured.

Q.3 Probability Distributions (Variant 1)

Part A

Suppose that a student is taking a multiple-choice exam in which each question has four choices. Suppose that she has no knowledge of the correct answer to any of the questions. Furthermore suppose that she selects one of the possible choices at random as her answer.

- i. (2 Marks) If there are five multiple-choice questions on the exam, what is the probability that she will answer four questions correctly.
- ii. (2 Marks) What is the probability that she will answer none of the questions correctly?
- iii. (2 Marks) What is the probability that she will answer at least two questions correctly?

Part B

A power supply unit for a computer component is assumed to follow an exponential distribution with a mean life of 1,400 hours. What is the probability that the component will:

- i. (2 Marks) fail in the first 700 hours?
- ii. (2 Marks) survive more than 1,750 hours?
- iii. (2 Marks) last between 1,050 hours and 1,750 hours?

Part C

Assume that the diameter of a critical component is normally distributed with a Mean of 50mm and a Standard Deviation of 2mm.

NB You must draw a rough sketch of the normal curve and estimate the approximate probability of the following measurements occurring on an individual component.

- i. (2 Marks) Between 50 and 51.2mm
- ii. (2 Marks) Less than 48.5 mm
- iii. (2 Marks) Between 48.2 and 101.6 mm

Use the normal tables to get the exact probabilities for the above.

Q.3 Probability Distributions (Variant 2)

Part A

Telephone calls arrive at a switchboard at a rate of 30 per hour. Assume that the switchboard operators take 2 minutes to deal with a customer query. Calculate the following:

- i. (2 Marks) The probability of 2 or more calls arriving in any 4 minute period,
- ii. (2 Marks) The probability of no phone calls arriving in a 4 minute period,
- iii. (2 Marks) The probability of exactly three phone call arriving in a 10 minute period.

Part B

Assume that the length of injected moulded plastic components are normally distributed with a mean of 12.5mm and a standard deviation of 2.5mm. Calculate the corresponding probability for the following measurements occurring on an individual component.

- i. (2 Marks) Between 12.5 and 15mms,
- ii. (2 Marks) Less than 10 mms,
- iii. (2 Marks) Between 12 and 15 mms,
- iv. (2 Marks) Less than 10.3 mms.

Illustrate each of your answers with a sketch.

Part C

It is estimated by a particular bank that 25% of credit card customers pay only the minimum amount due on their monthly credit card bill and do not pay the total amount due. 50 credit card customers are randomly selected.

1. (3 marks) What is the probability that 9 or more of the selected customers pay only the minimum amount due?
2. (3 marks) What is the probability that less than 6 of the selected customers pay only the minimum amount due?
3. (3 marks) What is the probability that more than 5 but less than 10 of the selected customers pay only the minimum amount due?

Q.3 Probability Distributions (Variant 3)

Part A

Flaws occur in an LCD display at the rate of 0.5 per square mm. Calculate the probability that:

- i. (2 Marks) exactly 2 flaws will occur in a square mm section,
- ii. (2 Marks) exactly 3 flaws will occur in a 5 square mm section,
- iii. (2 Marks) 5 or more flaws will occur in a 10 square mm section.

Part B

The average lifespan of a PC monitor is 6 years. You may assume that the lifespan of monitors follows an exponential probability distribution.

1. (2 marks) What is the probability that the lifespan of the monitor will be at least 5 years?
2. (2 marks) What is the probability that the lifespan of the monitor will not exceed 4 years?
3. (2 marks) What is the probability of the lifespan being between 5 years and 7 years?

Part C

A machine is used to package bags of potato chips. Records of the packaging machine indicate that its fill weights are normally distributed with a mean of 455 grams per bag and a standard deviation of 10 grams.

1. (3 marks) What proportion of bags filled by this machine will contain more than 470 grams in the long run?
2. (3 marks) What proportion of bags filled by this machine will contain less than 445 grams in the long run?
3. (2 marks) What proportion of bags filled by this machine will be between 465 grams and 475 grams in the long run?

Q4. Inference Procedures (Variant 1)

Part A

- i. (2 Marks) In the context of hypothesis testing, explain what a p-value is, and how it is used. Support your answer with a simple example.
- ii. (2 Marks) What is meant by Type I error and Type II error?

Part B

A well-known polling company estimates that 57% of Irish voters support a new constitutional amendment. 800 people were randomly surveyed and asked about their voting preferences. 482 of the 800 people responded positively to the amendment. You are required to:

- i. (1 Mark) Obtain a point estimate of the proportion of people supporting the constitutional amendment.
- ii. (2 Marks) Construct a 95% confidence interval for the proportion of people in favour of the constitutional amendment.

Part C

The standard deviations of data sets X and Y are 10 and 9 respectively. An inference procedure was carried out to assess whether or not X and Y can be assumed to have equal variance.

- i. (1 Mark) Formally state the null and alternative hypothesis.
- ii. (1 Mark) The Test Statistic has been omitted from the computer code output. Compute the value of the Test Statistic.
- iii. (2 Marks) What is your conclusion for this procedure? Justify your answer.

F test to compare two variances

```
data: X and Y
F = ....., num df = 13, denom df = 11, p-value = 0.7349
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.3639938 3.9475262
sample estimates:
ratio of variances
  .....
```

Part D

Two samples of students are randomly selected from two IT training companies; Echelon and Deltatech. The mean and the standard deviation of the number of marks obtained in a well known IT competency exam by both sets of students are described below:

	Number	Mean	Std. Dev.
DeltaTech	14	24	10
Echelon	12	22.5	9

Test the hypothesis that Echelon students and DeltaTech students, on average, obtain the same mark in the IT certification exam. Use a significance level of 5%. You may assume that any required assumptions have been validated.

- i. (2 Marks) Formally state the null and alternative hypotheses.
- ii. (3 Marks) Compute the Test Statistic.
- iii. (2 Marks) State the appropriate Critical Value for this hypothesis test.
- iv. (2 Marks) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.

Q4. Inference Procedures (Variant 2)

Part A

- i. (2 Marks) In the context of hypothesis testing, explain what a p-value is, and how it is used. Support your answer with a simple example.
- ii. (2 Marks) What is meant by Type I error and Type II error?

Part B

The mean and the standard deviation of the number of marks obtained in the biology leaving certificate exam by randomly selected male and female pupils are described below:

	Number	Mean	Std. Dev.
Male	10	57	12
Female	12	61	11

Test the hypothesis that males and females on average obtain the same mark in the biology leaving certificate exam. Use a significance level of 5%. You may assume that all required assumptions have been validated.

- i. (2 Marks) Formally state the null and alternative hypotheses.
- ii. (3 Marks) Compute the Test Statistic.
- iii. (2 Marks) State the appropriate Critical Value for this hypothesis test.
- iv. (2 Marks) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.

Part C

Consider the following inference procedure performed on data set X .

```
> shapiro.test(X)
```

```
Shapiro-Wilk normality test
```

```
data: X
```

```
W = 0.9619, p-value = 0.6671
```

- i. (1 Mark) Describe what is the purpose of this procedure.
- ii. (2 Marks) What is the null and alternative hypothesis?
- iii. (1 Mark) Write the conclusion that follows from it.

Q4. Inference Procedures (Variant 3)

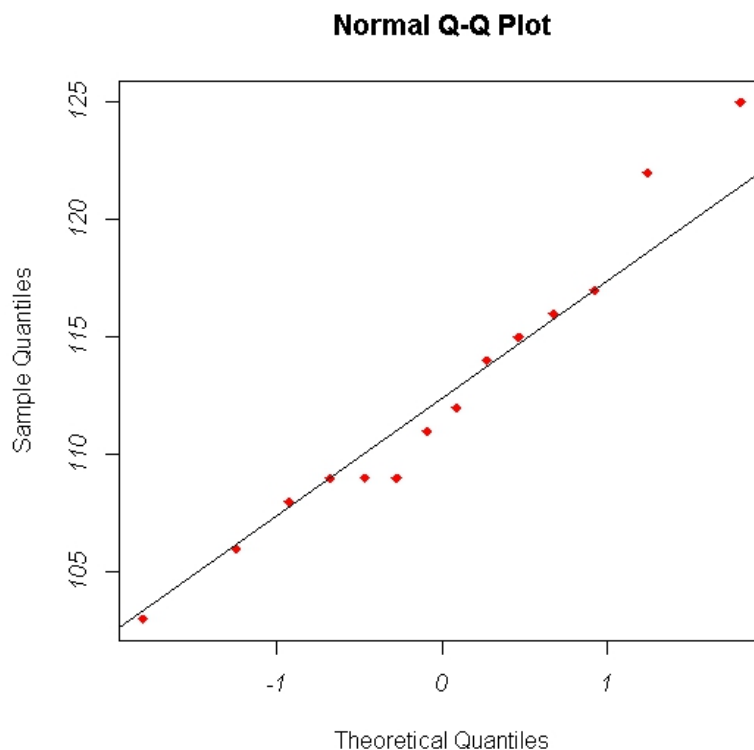
Part A

The strength of dosage of a plant growth enhancement chemical is often measured by the proportion of plants that grow faster. A particular dosage of the chemical is fed to 115 plants of these plants, 94 actually show faster growth.

- Calculate a point estimate \hat{p} for the proportion of plants that grow faster due to the dosage.
- What is the standard error of the estimate?
- Find a 95% confidence interval for p .
- What size sample is required to give the estimate in (c) to within $\pm 1\%$.
- Carry out a testing procedure to investigate the claim that the chemical is 90% effective.

Part B

The data set Y is both assumed to be normally distributed. A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set Y . Consider the Q-Q plot in the figure below.



- (2 marks) Provide a brief description on how to interpret this plot.
- (1 marks) What is your conclusion for this procedure? Justify your answer.

Part C

6.98	8.49	7.97	6.64
8.80	8.48	5.94	6.94
6.89	7.47	7.32	4.01

```
> grubbs.test(x, two.sided=T)
```

```
      Grubbs test for one outlier
```

```
data:  x
```

```
G = 2.4093, U = 0.4243, p-value = 0.05069
```

```
alternative hypothesis: lowest value 4.01 is an outlier
```

- i. (2 marks) Describe what is the purpose of this procedure.
- ii. (2 marks) Write the conclusion that follows from it.

Q4. Inference Procedures (Variant 4)

Part A

- (2 Marks) In the context of hypothesis testing, explain what a p-value is, and how it is used. Support your answer with a simple example.
- (2 Marks) What is meant by Type I error and Type II error?

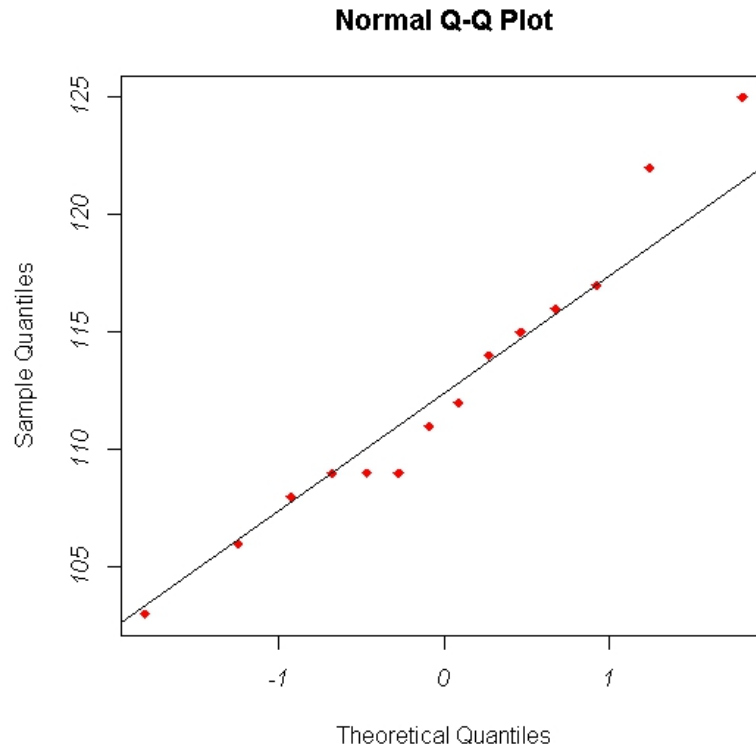
Part B

A well-known polling company estimates that 57% of Irish voters support a new constitutional amendment. 800 people were randomly surveyed and asked about their voting preferences. 482 of the 800 people responded positively to the amendment. You are required to:

- (1 Mark) Obtain a point estimate of the proportion of people supporting the constitutional amendment.
- (2 Marks) Construct a 95% confidence interval for the proportion of people in favour of the constitutional amendment.

Part C

A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set Y. Consider the Q-Q plot in the figure below.



- (2 marks) Provide a brief description on how to interpret this plot.

- ii. (1 marks) What is your conclusion for this procedure? Justify your answer.

Q5. Correlation and Linear Regression (Variant 1)

A fire insurance company wants to relate the amount of fire damage in major residential fires to the distance between the residence and the nearest fire station. The study is to be conducted in a large suburb of a major city; a sample of 15 recent fires in this suburb is selected. The amount of damage (thousands of pounds) and the distance (in Kilometers) between the fire and the nearest station are recorded for each fire.

Incident	Distance (X)	Fire Damage (Y)	Incident	Distance (X)	Fire Damage (Y)
1	3.4	21	9	2.6	15
2	1.8	13	10	4.3	26
3	4.6	26	11	2.1	19
4	2.3	18	12	1.1	12
5	3.1	23	13	6.1	38
6	5.5	31	14	4.8	31
7	0.7	9	15	3.8	21
8	3	17			

$$\begin{aligned} \sum X &= 49.2 & \sum Y &= 320 & \sum XY &= 1219.2 \\ \sum X^2 &= 196.16 & \sum Y^2 &= 7722 \end{aligned}$$

- i. (5 Marks) Illustrate the data in a scatter diagram.
- ii. (5 Marks) Compute the value of the correlation coefficient.
- iii. (5 Marks) Find the equation of the regression line and plot the regression line on the scatter diagram. Interpret the equation.
- v. (5 Marks) Using a significance level of $\alpha = 0.05$, test the claim that there is no linear correlation between distance and fire damage.

Q5. Correlation and Linear Regression (Variant 2)

A wood scientist wanted to establish if there was a relationship between the adhesive strength of laminated wood and the dwell time in press machine. A random sample of 9 different times and their corresponding adhesive strengths in pounds per square inch (PSI) were recorded as follows:

Sample	Time (Mins) (X)	Pull Strength (PSI) (Y)
1	5.0	3.5
2	4.8	3.3
3	5.6	3.9
4	4.3	2.7
5	4.2	3.2
6	5.4	4.1
7	5.5	4.3
8	4.0	2.8
9	4.7	3.7

$$\begin{aligned} \sum X &= 43.5 & \sum Y &= 31.5 & \sum XY &= 154.61 \\ \sum X^2 &= 213.03 & \sum Y^2 &= 112.71 \end{aligned}$$

- i. (5 Marks) Draw a scatter-plot and comment on its features.
- ii. (5 Marks) Calculate the correlation coefficient. Interpret your answer.
- iii. (5 Marks) Calculate the equation of the least squares regression line and interpret the value of the slope.
- iv. (3 Marks) Using this regression model, estimate the adhesive strength for a piece of wood that has been in the press machine for 7 minutes.
- v. (2 Marks) Is such an estimate reliable? Briefly explain why.

Formulae

Descriptive Statistics

- Sample Variance

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Probability

- Conditional probability:

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}.$$

- Bayes' Theorem:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)}.$$

- Binomial probability distribution:

$$P(X = k) = {}^n C_k \times p^k \times (1 - p)^{n-k} \quad \left(\text{where } {}^n C_k = \frac{n!}{k!(n-k)!} \right)$$

- Poisson probability distribution:

$$P(X = k) = \frac{m^k e^{-m}}{k!}.$$

- Exponential probability distribution:

$$P(X \leq k) = \begin{cases} 1 - e^{-k/\mu}, & k \geq 0, \\ 0, & k < 0. \end{cases} \quad \left(\text{where } \mu = \frac{1}{\lambda} \right)$$

Confidence Intervals

One sample

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

$$S.E.(\hat{P}) = \sqrt{\frac{\hat{p} \times (100 - \hat{p})}{n}}.$$

Two samples

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

$$S.E.(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\hat{p}_1 \times (100 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 \times (100 - \hat{p}_2)}{n_2}}.$$

Hypothesis tests

One sample

$$S.E.(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

$$S.E.(\pi) = \sqrt{\frac{\pi \times (100 - \pi)}{n}}$$

Two large independent samples

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

$$S.E.(\hat{P}_1 - \hat{P}_2) = \sqrt{(\bar{p} \times (100 - \bar{p})) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

Two small independent samples

$$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}.$$

Paired sample

$$S.E.(\bar{d}) = \frac{s_d}{\sqrt{n}}.$$

Standard Deviation of case-wise differences (computational formula)

$$s_d = \sqrt{\frac{\sum d_i^2 - n\bar{d}^2}{n - 1}}.$$

Regression estimates

$$S_{XY} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$S_{XX} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{YY} = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

Slope Estimate

$$b_1 = \frac{S_{XY}}{S_{XX}}$$

Intercept Estimate

$$b_0 = \bar{y} - b_1\bar{x}$$

Pearson's correlation coefficient

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \times S_{YY}}}$$

Standard error of the Slope

$$S.E.(b_1) = \sqrt{\frac{s^2}{S_{XX}}}$$

where $s^2 = \frac{SSE}{n-2}$ and $SSE = S_{YY} - b_1 S_{XY}$