

MATH 203: Principles of Statistics 1

Final Examination: Solutions

Fall 2024

December 12th 2024: 9.00am – 12.00pm

MATH 203: Specific Instructions.

Please write your answers on the exam in the boxes provided for each question; only answers written inside the correct boxes will be marked. Extra space for answers can be found at the end of the exam paper. Please write clearly.

This paper contains ten questions. Each question carries 10 marks. Credit will be given for all questions attempted. The total mark available is 100 but rescaling of the final mark may occur.

Only non-programmable and non text-storing calculators may be used. Dictionaries and Translation dictionaries are permitted.

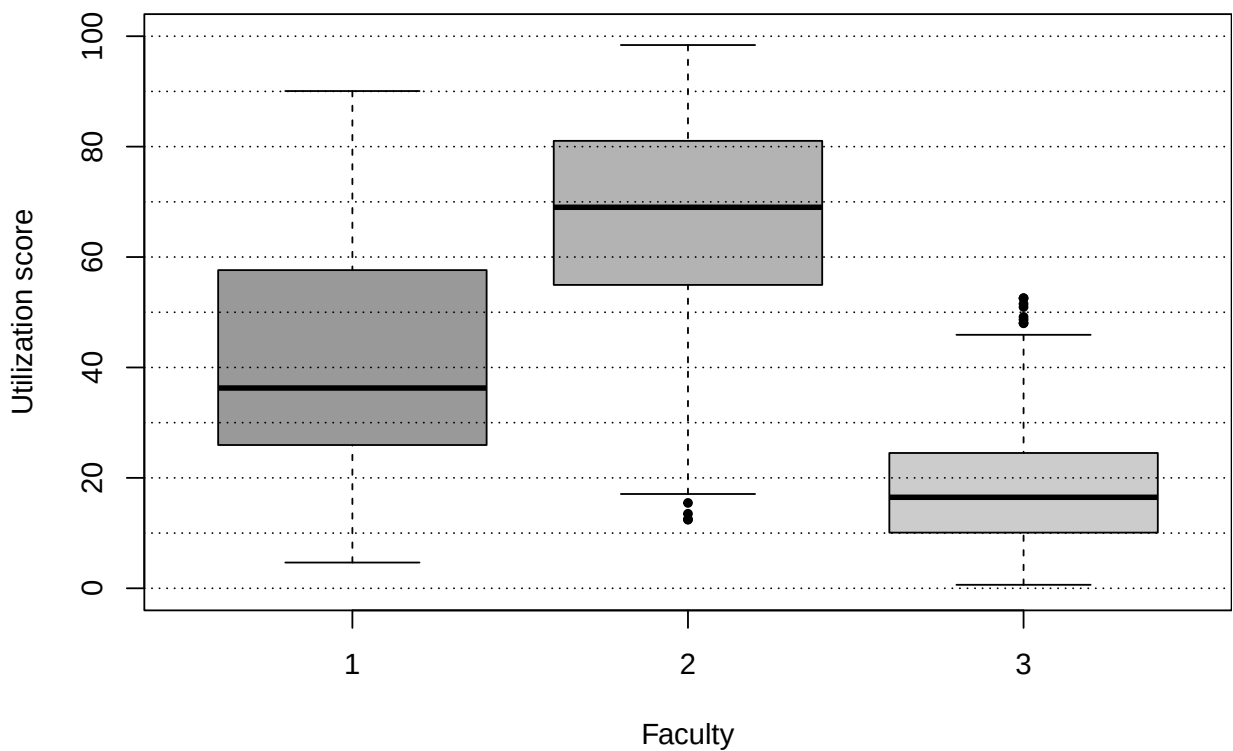
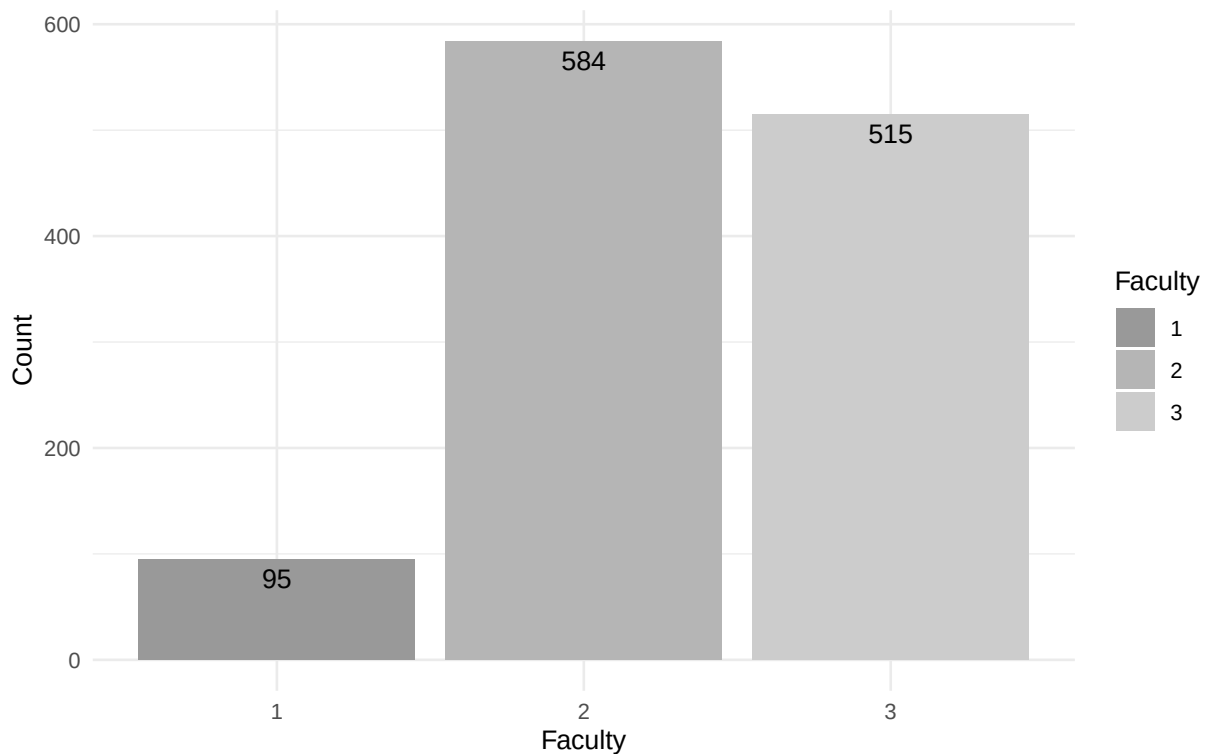
A formula sheet is provided.

1. In a study of the effectiveness of a new non-steroidal anti-inflammatory drug in providing pain relief for adult arthritis sufferers, 130 adult participants who suffered from this condition were recruited at a Montreal clinic between June and December 2022. Each participant was randomly allocated to receive either the new drug or a standard treatment (ibuprofen). After a week of treatment, each participant then reported a pain score on a continuous scale of zero to ten, and also reported whether or not they had experienced significant side effects. For this study, identify
- (a) the population being studied; 2 MARKS
 - (b) the experimental units; 2 MARKS
 - (c) the sample; 2 MARKS
 - (d) the three variables that will be recorded for each participant; 2 MARKS
 - (e) whether this is an observational or experimental study. 2 MARKS

Q1

- (a) Population: adult arthritis sufferers. 2 MARKS
- (b) Experimental units: individual adult participants. 2 MARKS
- (c) Sample: 130 participants recruited in Montreal clinic. 2 MARKS
- (d) Variables: study arm (new drug/ibuprofen), pain score (0–10), side-effects (Yes/No) . 2 MARKS
- (e) This is an experimental study as there is an experimental intervention. 2 MARKS

2. In a study of mental health services at a large US University, first-year students in three Faculties were recruited and asked to fill in a 30 item questionnaire to assess their need for, and utilization of, mental health facilities, leading to a Utilization Score for each participant. The recorded data are summarized in the following graphical displays: in the second plot, horizontal dotted lines are added to ease interpretation.



Address the following questions, giving brief justification of your answers.

- (a) What is the total sample size ? 1 MARK
- (b) Which Faculty has the lowest median utilization score, and what is its approximate value ? 2 MARKS
- (c) Approximately, what is the interquartile range (IQR) of utilization scores for Faculty 2 ? 2 MARKS
- (d) Approximately, what is the range of utilization scores for Faculty 1 ? 2 MARKS
- (e) For Faculty 3, the sample mean and sample standard deviation for utilization score are 18.52 and 10.94 respectively. Use the Empirical Rule to construct an interval that aims to contain 95% of the observed samples for that Faculty, and comment on the validity of this interval to achieve that aim. 3 MARKS

Q2

- (a) Total sample size: $95+584+515 = 1194$. 2 MARKS
- (b) From the boxplot mid-line, lowest median is Faculty 3, at around 16-17. 2 MARKS
- (c) From the boxplot box, the IQR for Faculty 2 is approximately $81-55 = 26$. 2 MARKS
- (d) From the boxplot whiskers (noting the absence of outliers), the range for Faculty 1 is approximately $90-5 = 85$. 2 MARKS
- (e) We have the interval
$$(\bar{x} - 2s, \bar{x} + 2s) = (-3.36, 40.4).$$
which may achieve the aim, but as the distribution is right-skewed, is not perhaps suitable as an interval to report; note that the lower limit is negative. 2 MARKS

3. The body temperature (in Celsius) of 33 adults entering an airport terminal is measured using thermal imaging. The data, sorted into ascending order are

36.33 36.69 36.70 36.95 37.01 37.01 37.02 37.03 37.05 37.09 37.14
 37.15 37.20 37.21 37.30 37.33 37.35 37.36 37.39 37.41 37.44 37.58
 37.58 37.61 37.76 37.80 37.83 37.85 37.96 37.98 38.04 39.90 41.10

Numerical descriptive summaries are

$$\sum_{i=1}^{33} x_i = 1238.15 \quad \sum_{i=1}^{33} x_i^2 = 46479.6981.$$

- (a) Compute the sample mean, $\bar{x}$. 2 MARKS
 (b) Compute the sample standard deviation, s . 3 MARKS
 (c) Compute the sample range. 2 MARKS
 (d) The largest observed value is determined to be a spurious observation and is discarded. Does this increase or decrease the value of s ? 3 MARKS

Q3

- (a) We have 2 MARKS

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{33} \times 1238.15 = 37.519697$$

- (b) We have 3 MARKS

$$s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n(\bar{x})^2 \right) = \frac{1}{32} (46479.6981 - 33(37.519697)^2) = 0.7714155.$$

- (c) The sample range is $R = 41.1 - 36.33 = 4.77$. 2 MARKS

- (d) Removing the largest value will always decrease the variance, and the value of s ; we can verify this by noting that after deletion

$$\bar{x} = 37.4078125 \quad \sum_{i=1}^{32} x_i^2 = 44790.4881 \quad s^2 = 0.3698757$$

so $s = 0.6081741$. 3 MARKS

Note: rounding of the answers to several decimal places is acceptable.

4. A total of 882 wines from four regions of Europe are rated by a panel of tasters as either being Low, Medium, Good or Excellent quality. The results of the rating are displayed in the following table.

	Low	Medium	Good	Excellent
Region 1	56	98	45	9
Region 2	62	41	41	92
Region 3	30	43	60	102
Region 4	10	106	44	43

One of the wines is selected, with all wines being equally likely to be selected. Using probability notation, express and compute the probability that the wine

- (a) originates from Region 3; 2 MARKS
 (b) is rated as Excellent quality; 2 MARKS
 (c) is rated as Low or Medium quality, given that it came from Region 1 or Region 2; 3 MARKS
 (d) comes from Region 4, given that it is rated as Low quality. 3 MARKS

Q4

Let the partition by Region be denoted by events R_1, R_2, R_3, R_4 , and the qualities be denoted C_1, C_2, C_3, C_4 .

- (a) originates from Region 3; this is 2 MARKS

$$P(R_3) = P(R_3 \cap C_1) + P(R_3 \cap C_2) + P(R_3 \cap C_3) + P(R_3 \cap C_4) = \frac{235}{882} = 0.2664399.$$

- (b) is rated as Excellent quality; this is 2 MARKS

$$P(C_4) = P(R_1 \cap C_4) + P(R_2 \cap C_4) + P(R_3 \cap C_4) + P(R_4 \cap C_4) = \frac{246}{882} = 0.2789116.$$

- (c) is rated as Low or Medium quality, given that it came from Region 1 or Region 2; this is

$$P(C_1 \cup C_2 | R_1 \cup R_2) = \frac{56 + 98 + 62 + 41}{444} = \frac{257}{444} = 0.5788288$$

3 MARKS

- (d) comes from Region 4, given that it is rated as Low quality; this is

$$P(R_4 | C_1) = \frac{10}{158} = 0.0632911$$

3 MARKS

5. Let Y be a discrete random variable taking values on the set $\{1, 2, 3, 4, 5\}$ with the following probabilities

y	1	2	3	4	5
$P(Y = y)$	0.10	0.20	0.40	0.20	0.10

- (a) Compute the expected value of Y , $E(Y)$ 2 MARKS
 (b) Compute the variance of Y , $Var(Y)$. 3 MARKS
 (c) A sample of $n = 100$ independent and identically distributed random variables $X_1, \dots, X_n$ are to be collected, each having the same distribution as Y . If

$$S_n = \sum_{i=1}^n X_i$$

compute an approximate value of $P(S_n < 290)$.

5 MARKS

Q5

- (a) We have

$$E(Y) = \sum_{y=1}^5 yp(y) = (1 \times 0.1) + (2 \times 0.2) + (3 \times 0.4) + (4 \times 0.2) + (5 \times 0.1) = 3.$$

which we may also deduce using a symmetry argument.

2 MARKS

- (b) We have

$$E(Y^2) = \sum_{y=1}^5 y^2 p(y) = (1 \times 0.1) + (4 \times 0.2) + (9 \times 0.4) + (16 \times 0.2) + (25 \times 0.1) = 10.2$$

so therefore

$$Var(Y) = E(Y^2) - \{E(Y)\}^2 = 1.2.$$

3 MARKS

- (c) By the Central Limit Theorem,

$$S_n \rightsquigarrow Normal(n\mu, n\sigma^2) = Normal(300, 120).$$

Therefore, if Z denotes a standard Normal random variable

$$P(S_n < 290) = P\left(\frac{S_n - 300}{\sqrt{120}} < \frac{290 - 300}{\sqrt{120}}\right) = P(Z < -0.9128709) \approx P(Z < -0.91)$$

and using the symmetry result and tables, we have

5 MARKS

$$P(Z < -0.91) = 1 - P(Z < 0.91) = 1 - 0.8186 = 0.1814.$$

6. In a specific week, a bank handles 800 transactions involving sums more than \$5000. On the basis of historical data, 0.2% (that is, 2 in 1000) of such transactions are deemed to be fraudulent. Let Y be the number of fraudulent transactions recorded in a specific week under study.
- (a) Compute the probability that the number of fraudulent transactions in a week is at most one, that is, $P(Y \leq 1)$. 3 MARKS
 - (b) Compute the expected number of fraudulent transactions, $E(Y)$. 2 MARKS
 - (c) Use a Poisson approximation to compute $P(Y = 0)$. 2 MARKS
 - (d) The cost to the bank (in hundreds of dollars) of one fraudulent transaction is a random variable, X , which is Normally distributed with $\mu = 5$ and $\sigma = 1.5$. Compute the probability that X exceeds \$700. 3 MARKS

Q6

- (a) We have that $Y \sim \text{Binomial}(800, 0.002)$ so 3 MARKS

$$\begin{aligned}
 P(Y \leq 1) &= P(Y = 0) + P(Y = 1) \\
 &= \binom{800}{0} (0.998)^{800} + \binom{800}{1} 0.002 (0.998)^{799} \\
 &= 0.2015733 + 800 \times 0.000404 = 0.5247369
 \end{aligned}$$

- (b) From the formula sheet, $E(Y) = np = 800 \times 0.002 = 1.6$. 2 MARKS

- (c) We set $\lambda = np = 1.6$, and compute 2 MARKS

$$P(Y = 0) = e^{-\lambda} = e^{-1.6} = 0.2018965.$$

- (d) If $X \sim \text{Normal}(5, 1.5^2)$, we require

$$P(X > 7) = P\left(\frac{X - 5}{1.5} > \frac{7 - 5}{1.5}\right) = P(Z > 1.3333333) \approx 1 - P(Z \leq 1.33)$$

from tables, where $Z \sim \text{Normal}(0, 1)$. Therefore

$$P(X > 7) \approx 1 - 0.9082 = 0.0918.$$

3 MARKS

7. In a study examining how the popularity of crime-scene investigation television shows influences expectations of what evidence should be produced at a trial, a survey of 500 potential jurors found that 281 were regular watchers of at least one crime-scene forensics television series.
- (a) What is the point estimate for the population proportion, p , of all potential jurors who regularly watch at least one crime-scene forensics television series.? 2 MARKS
 - (b) Construct a 95% confidence interval for the proportion of all potential jurors who regularly watch at least one crime-scene forensics television series. 4 MARKS
 - (c) Carry out a hypothesis test of the null hypothesis that $p = 0.5$ against the alternative hypothesis that $p > 0.5$. 4 MARKS

Q7

- (a) The point estimate is 2 MARKS

$$\hat{p} = \frac{281}{500} = 0.562.$$

- (b) The confidence interval is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0.562 + 1.9600 \sqrt{\frac{0.562 \times 0.438}{n}} = (0.5185113, 0.6054887)$$

as $z_{\alpha/2} = z_{0.025} = 1.9600$ 4 MARKS

- (c) Here

$$H_0 : p = p_0$$

$$H_a : p > p_0$$

with $p_0 = 0.5$, and the test statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \quad \text{with} \quad T_{obs} = \frac{0.562 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{500}}} = 2.7727243.$$

The Rejection Region for this alternative takes the form

$$\{T \geq z_{\alpha}\}$$

and we may choose $\alpha = 0.05$, so the interval is $\{T \geq 1.6449\}$. The test statistic lies in the Rejection Region, so we reject H_0 in favour of H_a at this value of α . 4 MARKS

The rejection of H_0 also follows if we choose $\alpha = 0.01$, as then the Rejection Region is $\{T \geq 2.3263\}$.

8. The flight times (in minutes) for 10 randomly selected flights from city A to city B during January 2022 are recorded as 270, 256, 267, 285, 274, 275, 266, 258, 271, 281; these data yield the sample summaries $\bar{x} = 270.3$ and $s = 9.14$ (rounded to two decimal places). It is desired to construct a 99% confidence interval (C.I.) for the population mean time for flights from A to B, denoted μ .
- (a) Construct the C.I. for μ based on the data. State the assumptions that are needed to ensure that this C.I. is statistically valid. 5 MARKS
- (b) Data are subsequently recorded for the whole of 2022; in total, 117 flight times are recorded, and the sample mean and sample standard deviation for the larger sample are 276.4 and 53.28 (minutes) respectively. Construct a 99% C.I. for μ based on the new data. Identify the assumption from (a) that can be relaxed and still leave the interval statistically valid. 5 MARKS

Q8

- (a) We have here a small sample problem, so the 99% CI is

$$\bar{x} \pm t_{0.005, n-1} \frac{s}{\sqrt{n}} = 270.3 \pm 3.2498 \frac{9.14}{\sqrt{10}} = (260.9070323, 279.6929677)$$

as we use the $t_{n-1} = t_9$ distribution. For this interval to be valid, we need that the data are Normally distributed (and the usual assumption that flight times are independent from the same population – this does not need to be mentioned for full marks).

5 MARKS

- (b) We have here a large sample problem, so the 99% CI is

$$\bar{x} \pm z_{0.005} \frac{s}{\sqrt{n}} = 276.4 \pm 2.5758 \frac{53.28}{\sqrt{117}} = (263.7122847, 289.0877153)$$

Here we no longer need to assume Normality, as the sample size is large. 5 MARKS

9. Anxiety test scores were recorded for two samples (of sizes $m = 45$ and $n = 47$) from two populations of parents of very low birthweight babies. Summaries are given in the following table:

	Sample size	Sample mean	Sample variance
Sample 1	45	57.42	137.16
Sample 2	47	52.16	128.81

- (a) Construct a 95% confidence interval for the difference in mean population anxiety scores between the two populations. 6 MARKS
- (b) Test, using suitable hypotheses, whether there is any difference between the two population means using $\alpha = 0.05$. 4 MARKS

State any required assumptions in each case.

Q9

Here we have a two sample unpaired problem, with samples sizes large. For both parts, we simply need independent samples for Sample 1 and Sample 2.

- (a) The 95% CI is

$$(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}} = (57.42 - 52.16) \pm 1.9600 \sqrt{\frac{137.16}{45} + \frac{128.81}{47}}$$

$$= (0.5443205, 9.9756795)$$

6 MARKS

- (b) Here the appropriate hypotheses are

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

and the test statistic is

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{m} + \frac{S_2^2}{n}}} \quad \text{with} \quad T_{obs} = \frac{57.42 - 52.16}{\sqrt{\frac{137.16}{45} + \frac{128.81}{47}}} = 2.1862385.$$

The Rejection Region for this alternative takes the form

$$\{|T| \geq z_{\alpha/2}\} = \{|T| \geq 1.9600\}$$

and as $T_{obs} = 2.1862385$, the test statistic lies in the Rejection Region, so we reject H_0 in favour of H_a at this value of α . 4 MARKS

10. The times (in seconds) taken to complete a complex task for 10 subjects are recorded under two experimental conditions:

- Condition 1: Complete silence.
- Condition 2: A loud alarm going off throughout the task.

The raw data, and the difference in time taken between Condition 2 and Condition 1, are recorded in the following table:

Raw data (in seconds)			
Subject	Condition 1	Condition 2	Difference
1	10.4	15.3	4.9
2	11.7	14.6	2.9
3	8.2	14.4	6.2
4	8.3	17.1	8.8
5	10.4	13.2	2.8
6	8.5	14.6	6.1
7	10.8	12.5	1.7
8	8.4	11.4	3.0
9	10.3	14.6	4.3
10	9.4	15.1	5.7

The summary statistics for the data are included in the following table:

Summary statistics			
	Condition 1	Condition 2	Difference
Mean	9.64	14.28	4.64
Standard deviation	1.24	1.59	2.13

- (a) Carry out a suitable test of hypotheses to assess whether the time taken to complete the task under Condition 2 is greater than the time taken under Condition 1. State the assumptions that are needed to ensure that this test is statistically valid.
6 MARKS
- (b) Construct a 95% confidence interval for the population mean time taken to complete the task under Condition 1. State the assumptions needed to ensure that this confidence interval is statistically valid.
4 MARKS

Q10

Here we have a two sample paired data problem, with small sample size. We label Condition 2 as the population with mean μ_1 .

(a) Here the appropriate hypotheses are

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

and the test statistic is

$$T = \frac{\bar{D}}{S_d/\sqrt{n}} \quad \text{with} \quad T_{obs} = \frac{4.64}{2.13/\sqrt{10}} = 6.8887175.$$

The Rejection Region for this alternative takes the form $\{T \geq t_{\alpha, n-1}\}$, and if we select $\alpha = 0.05$, we have

$$\{T \geq t_{\alpha, n-1}\} = \{T \geq t_{0.05, 9}\} = \{T \geq 1.8331\}$$

and as $T_{obs} = 6.8887175$, the test statistic lies in the Rejection Region, so we reject H_0 in favour of H_a at this value of α . If $\alpha = 0.01$, we have

$$\{T \geq t_{\alpha, n-1}\} = \{T \geq t_{0.01, 9}\} = \{T \geq 2.8214\}$$

and as $T_{obs} = 6.8887175$, the test statistic lies in the Rejection Region, so we also reject H_0 in favour of H_a at this value of α .

Here we need that the sample of differences is from a Normal population. 6 MARKS

(b) The 95% CI is (using the usual notation)

$$\begin{aligned} \bar{y} \pm t_{0.025, 9} \frac{s_2}{\sqrt{n}} &= 9.64 \pm 2.2622 \frac{1.24}{\sqrt{10}} \\ &= (8.7529574, 10.5270426) \end{aligned}$$

Here, in this small sample case, we need that the sample of Condition 1 measurements is from a Normal population. 4 MARKS