

MATH 203: Principles of Statistics 1

Practice Final Examination 2

Fall 2024

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.

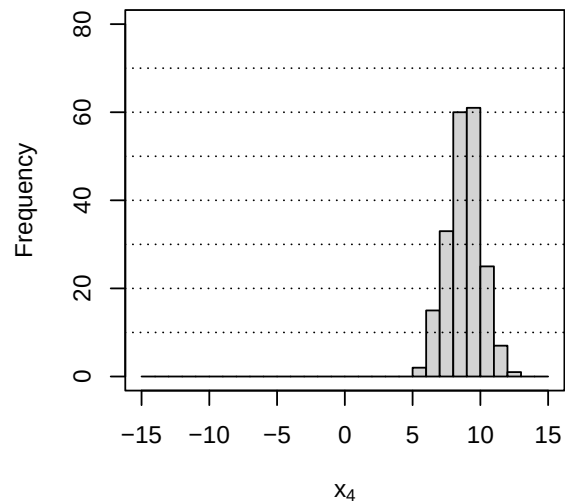
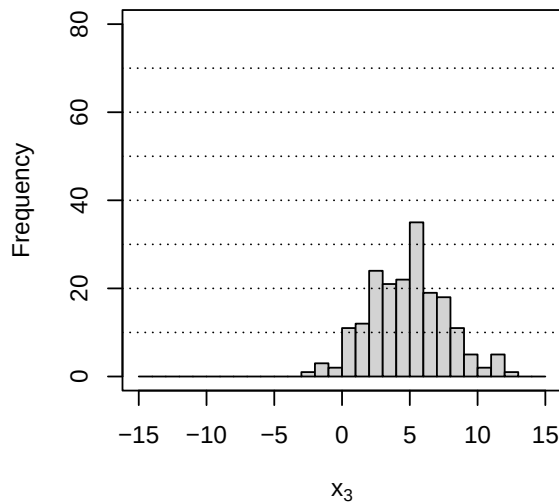
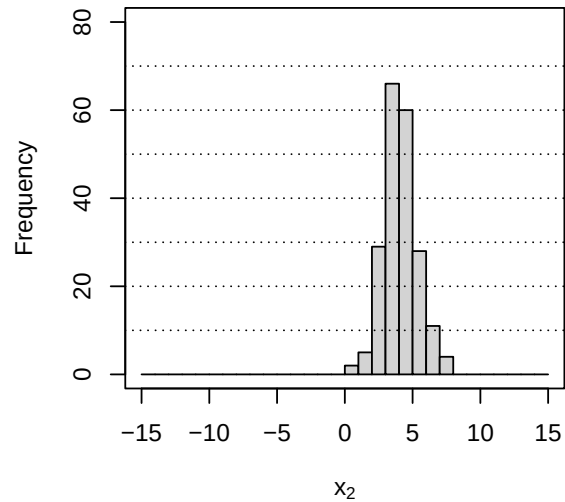
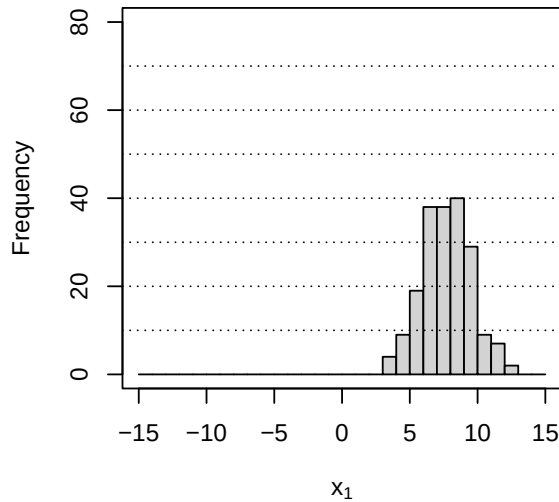
1. Héma-Québec is the organization that manages blood products in Québec. It carries out routine analysis of the blood samples of blood donors for the presence of antibodies demonstrating exposure to viruses such as SARS-COV2. In one month in 2022, a sample of 78 donors were selected and their blood tested, with the presence or absence of antibodies being recorded, along with their age, and a response to a question as to whether or not they had tested positive for COVID in the last six months.
- (a) What is the study population? 2 MARKS
 - (b) What are the experimental units for the study 2 MARKS
 - (c) Is this an experimental study? 2 MARKS
 - (d) Are the three variables recorded qualitative or quantitative? 3 MARKS
 - (e) If the question concerning testing positive was amended to ask 'How many times have you tested positive for COVID in the last six months?' would the response be a qualitative or quantitative variable ? 1 MARK

Justify your answers in each case.

Q1

- (a) The study population is blood donors in Québec.
- (b) The experimental units are individual blood donors.
- (c) This is not an experimental study, there is no experimental intervention.
- (d) Presence of antibodies (Yes/No) is qualitative, Age is quantitative, Answer to tested positive question (Yes/No) is qualitative.
- (e) The response is now quantitative (as it is a count).

2. The graphical display presents the samples obtained for four variables x_1, x_2, x_3 and x_4 ; horizontal dotted lines are added for ease of interpretation and comparison.



- Approximately how many values in the x_1 sample lie between 5 and 10 ? 2 MARKS
- Assess which of the variables has the largest sample variance; explain your assessment. 2 MARKS
- Would you expect the sample median and the sample mean to be approximately equal for the four variables? Justify your answer. 2 MARKS
- The graphical summaries indicate that each of the variables is approximately Normally distributed. Identify one other graphical method that could be used to verify Normality, and explain in general terms how Normality would be verified using this method. 4 MARKS

Q2

- (a) From the first histogram, we can assess this as approximately $19 + 38 + 38 + 40 + 29 = 164$.
- (b) As the spread in the histogram is largest, it seems that x_3 has the largest variance.
- (c) Yes, the sample median and the sample mean should be approximately the same, as the histograms are fairly symmetric, there is no skewness.
- (d) The best method would be to use a QQ plot: this plots the (standardized) sample quantiles against the theoretical quantiles from a (standard) Normal distribution. If the data are truly Normally distributed, the plot will closely follow a straight line, although we might expect a small amount of variability.

3. Waiting times (in hours) of people visiting a hospital emergency room are recorded. The sample data, sorted into ascending order, are

0.09, 0.20, 0.33, 0.35, 0.55, 0.60, 0.65, 0.66, 0.76, 0.76, 1.04, 1.14, 1.34, 2.01, 5.73

with summary statistics $\bar{x} = 1.081$ and $s = 1.375$.

- (a) Identify the median of the sample 2 MARKS
- (b) Compute the sample range. 2 MARKS
- (c) Is there any evidence in the data that suggests the data may not be Normally distributed? Justify your answer. 3 MARKS
- (d) Does the Empirical Rule for the interval formed as $(\bar{x} - 2s, \bar{x} + 2s)$ match the observed behaviour of this sample? Justify your answer. 3 MARKS

Q3

- (a) As $n = 15$, the median is the observation in position 8 in the ordered sample, so this is 0.66.
- (b) The sample range is $5.73 - 0.09 = 5.64$.
- (c) Yes, there is an indication that the data may not be Normally distributed, as the sample mean is much larger than the sample median, indicating a skewed distribution. However, the sample size is quite small, so we do not have definitive evidence.
- (d) Here we have

$$(\bar{x} - 2s, \bar{x} + 2s) = (1.081 - 2 \times 1.375, 1.081 + 2 \times 1.375) = (-1.669, 3.831).$$

and according to the Empirical Rule, this interval should contain approximately 95% of the sample. In this case it contains 14 out of 15, which is approximately 93.3%, so, given the sample size, this is a reasonable approximation. However, the interval itself extends well below the lowest value observed. Notice that the largest observation is well beyond the upper end of the interval, indicating that this outlying observation is likely inflating the sample standard deviation appreciably.

4. A survey of 2000 commuters in five boroughs of a large city was carried out. Each commuter was asked to assess their experience of the public transport system in the last month, and asked to rate whether they were Dissatisfied, Neutral or Satisfied. The data are summarized in the following table.

	1	2	3	4	5
Dissatisfied	71	412	146	73	70
Neutral	54	290	121	68	47
Satisfied	53	311	150	64	70

A commuter is selected from the sample, with all commuters equally likely to be selected. Find the probability that the commuter is

- (a) Satisfied; 2 MARKS
 (b) from borough 5; 2 MARKS
 (c) Dissatisfied, given that they are from borough 1 or borough 2; 3 MARKS
 (d) from borough 4 or borough 5, given that they are not Satisfied 3 MARKS

Use probability notation. You may leave your answers as fractions.

Q4

Let the commuter response events be C_1, C_2, C_3 and the borough origins B_1, B_2, B_3, B_4, B_5 .

- (a) This is

$$P(C_3) = \sum_{j=1}^5 P(C_3 \cap B_j) = \frac{53 + 311 + 150 + 64 + 70}{2000} = \frac{648}{2000} = 0.324.$$

- (b) This is

$$P(B_5) = \sum_{i=1}^3 P(B_5 \cap C_i) = \frac{70 + 47 + 70}{2000} = \frac{187}{2000} = 0.0935.$$

- (c) By the partitioning results and the definition of conditional probability, this is

$$P(C_1 | B_1 \cup B_2) = \frac{P(C_1 \cap (B_1 \cup B_2))}{P(B_1 \cup B_2)} = \frac{P(C_1 \cap B_1) + P(C_1 \cap B_2)}{P(B_1 \cup B_2)} = \frac{71 + 412}{178 + 1013} = \frac{483}{1191}.$$

- (d) By the partitioning results and the definition of conditional probability, this is

$$P(B_4 \cup B_5 | C_3^c) = \frac{P((B_4 \cup B_5) \cap (C_1 \cup C_2))}{P(C_1 \cup C_2)} = \frac{P(C_1 \cap B_4) + P(C_1 \cap B_5) + P(C_2 \cap B_4) + P(C_2 \cap B_5)}{P(C_1 \cup C_2)}$$

that is,

$$P(B_4 \cup B_5 | C_3^c) = \frac{73 + 70 + 68 + 47}{772 + 580} = \frac{258}{1352}$$

Note: For (c) and (d), the full mathematical derivation is not needed provided the calculation is laid out numerically.

5. A fair die is rolled once:

- if the die shows 1 or 2, a score of -2 is recorded,
- if the die shows 3, a score of -1 is recorded,
- if the die shows 4, a score of +1 is recorded,
- if the die shows 5 or 6, a score of +2 is recorded.

The recorded score, Y , is a discrete random variable taking values on $\{-2, -1, 1, 2\}$.

- (a) Write down the pmf, $p(y)$, that records $P(Y = y)$ for each y . 3 MARKS
- (b) Compute the expected value of Y , $E(Y)$. 2 MARKS
- (c) Compute the variance of Y , $\text{Var}(Y)$. 3 MARKS
- (d) The die is to be rolled 100 times independently, with the recorded scores calculated in the same way as Y , resulting in random variables $X_1, \dots, X_{100}$. Write down the approximate distribution of the sample mean random variable, $\bar{X}$. 2 MARKS

Q5

(a) Due to equally likely outcomes in the fair die, we have the pmf

y	-2	-1	1	2
$p(y) = P(Y = y)$	1/3	1/6	1/6	1/3

(b) We have

$$E(Y) = (-2 \times 1/3) + (-1 \times 1/6) + (1 \times 1/6) + (2 \times 1/3) = 0.$$

Note that the pmf is symmetric around zero, so we can deduce that the expectation is zero immediately.

(c) For the variance, as $E(Y) = 0$,

$$\text{Var}(Y) = E(Y^2) - \{E(Y)\}^2 = E(Y^2) = (4 \times 1/3) + (1 \times 1/6) + (1 \times 1/6) + (4 \times 1/3) = 3$$

(d) By the Central Limit Theorem, we have that

$$\bar{X} \dot{\sim} \text{Normal}(\mu, \sigma^2/n) = \text{Normal}(0, 0.03).$$

that is, the variance is 0.03.

6. A small car rental company owns 12 cars; the probability that any car needs maintenance in one month is 0.05, independently of other cars. At the end of each month, every car undergoes routine servicing, so that maintenance events in successive months can be also considered independent. Compute
- (a) the expected number of cars that need maintenance in one month; 2 MARKS
 - (b) the probability that in one month, the total number of cars that need maintenance is two or fewer; 3 MARKS
 - (c) the probability that zero cars require maintenance in a six month period; 3 MARKS
 - (d) the expected number of months until a specific car needs maintenance. 2 MARKS

Q6

- (a) This comes from the fact that if Y is the number of cars needing maintenance in one month, then $Y \sim \text{Binomial}(12, 0.05)$, and so the expectation is $12 \times 0.05 = 0.6$.
- (b) This comes from the $\text{Binomial}(12, 0.05)$, and we have the probability as

$$P(Y = 0) + P(Y = 1) + P(Y = 2) = \binom{12}{0} 0.95^{12} + \binom{12}{1} 0.95^{11} 0.05 + \binom{12}{2} 0.95^{10} 0.05^2 = 0.9804317.$$

- (c) This comes from the $\text{Geometric}(p)$ distribution, as months are independent, where

$$p = P(Y = 0) = \binom{12}{0} 0.95^{12} = 0.5403601.$$

Therefore, if X is the rv recording the first month that any maintenance is required, we have

$$P(X > 6) = P(\text{No maintenance in any of the first six months}) = p^6 = 0.0248943$$

- (d) This comes from the $\text{Geometric}(0.05)$ distribution, and the expectation is $1/0.05 = 20$ months.

7. Movie trailers are designed to entice audiences by showing scenes from coming attractions. The duration of a movie trailer can be assumed to be Normal with mean 150 seconds and standard deviation 30 seconds.

- (a) What is the probability that a randomly selected trailer lasts less than 2 minutes? 2 MARKS
- (b) Any movie trailer that lasts beyond 4 minutes and 30 seconds is considered too long. What percentage of movie trailers is considered too long? 2 MARKS
- (c) Ten percent of movie trailers have durations above what length? 3 MARKS
- (d) For a sample of 16 randomly selected movie trailers, find the probability that the sample mean duration is greater than 4 minutes and 30 seconds. 3 MARKS

Show your working.

Q7

Let Y be the trailer length, and Z be a standard Normal random variable.

- (a) This is

$$\begin{aligned} P(Y < 120) &= P\left(\frac{Y - 150}{30} < \frac{120 - 150}{30}\right) \\ &= P(Z < -1) = 1 - P(Z < 1) = 1 - 0.8413 = 0.1587 \end{aligned}$$

by symmetry, and using tables.

- (b) This is

$$\begin{aligned} P(Y > 270) &= P\left(\frac{Y - 150}{30} > \frac{270 - 150}{30}\right) \\ &= P(Z > 4) = 1 - P(Z \leq 4) = 1 - 1.0000 = 0.0000 \end{aligned}$$

to four decimal places, so the percentage is zero.

- (c) We want to find the value y such that $P(Y > y) = 0.1$, or equivalently $P(Y \leq y) = 0.9$. This is available from the table by looking up 0.9000 in the body of the table, and noting the row and the column; we can see the two closest values as 0.8997 and 0.9015 for values 1.28 and 1.29 respectively. So the closest value is 1.28. This says that 10% of trailers have lengths more than 1.28 standard deviations above the mean, so the required length is $150 + 1.28 \times 30 = 188.4$ seconds.
- (d) Here we have that $\bar{X} \sim \text{Normal}(150, (30/\sqrt{n})^2)$ exactly, even though n is small, by properties of the Normal distribution. Thus with $n = 16$, we have that $\bar{X} \sim \text{Normal}(150, (7.5)^2)$. Given the result in (b), we can deduce that to four decimal places, the required probability is zero, as the sample mean random variable is even more tightly distributed around 150.

8. An airline public relations department says that the airline rarely loses passengers' luggage. It further claims that on those occasions when luggage is lost, 90% is recovered and delivered to its owner within 24 hours. A consumer group that surveyed a large number of air travellers found that only 103 of 122 people who lost luggage on that airline received their missing items by the next day.

- (a) Do the data provide sufficient evidence to conclude that the percentage of people whose lost luggage is recovered and delivered back to them within 24 hours is lower than what the airline claims? Carry out an appropriate hypothesis test.

6 MARKS

- (b) Compute the p-value for this test.

4 MARKS

Q8

- (a) Here we have a one sample test for parameter p , with $\hat{p} = 103/122 = 0.8442623$ and hypothesized value $p_0 = 0.9$. We wish to test

$$H_0 : p = 0.9$$

$$H_a : p < 0.9$$

and α is not specified, so we may choose $\alpha = 0.05$. The test statistic is

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \quad \text{with} \quad T_{obs} = \frac{0.8442623 - 0.9}{\sqrt{\frac{0.9 \times 0.1}{122}}} = -2.0521436$$

The Rejection Region for this one sided alternative is $\{T \leq -z_\alpha\}$, that is, $\{T \leq -1.6449\}$ from the t-table final line where $\nu = \infty$. Thus, T_{obs} lies in the Rejection Region, so we reject H_0 in favour of H_a at this value of α .

Here $n\hat{p} = 103$ and $n(1 - \hat{p}) = 19$ but $np_0 = 109.8$ and $n(1 - p_0) = 12.2$, so the sample size is perhaps not quite large enough to apply the test according to the standard rules of thumb. We must note this potential limitation.

- (b) For this test, the p-value is

$$P_0(T \leq T_{obs}) = P_0(T \leq -2.0521436) \approx P_0(T \leq -2.05) = 1 - P_0(T \leq 2.05)$$

which is available from the Normal table as $1 - 0.9798 = 0.0202$. The actual p-value must be even smaller than this, as T_{obs} is a little smaller than -2.05.

Note that we can also use the t-table to bound the p-value; we see from the table final row that the p-value must be between 0.025 and 0.02 as the T_{obs} lies between 1.9600 and 2.0537; in fact we can deduce that the p-value must be very close to 0.02.

9. In a study to investigate whether vitamin C is an effective treatment for colds, volunteers who caught a cold after having been exposed to various cold viruses are assigned at random to one of two groups. The 10 volunteers in the experimental group were given tablets containing 1 gram of vitamin C to be taken four times a day. The 12 volunteers in the control group were given placebo tablets. The length of time (in days) the cold lasted for each volunteer was then recorded: the mean length of time the cold lasted for volunteers in the experimental group was found to be 6.45 with a standard deviation of 0.76, while the mean length of time the cold lasted for volunteers in the control group was found to be 7.13 with a standard deviation of 0.88.
- (a) Conduct an appropriate test of hypotheses of whether taking vitamin C daily reduces the population mean length of time in days that a cold lasts, and state the conditions required for the test to be statistically valid. 6 MARKS
- (b) Construct a 95% confidence interval for the difference in the means for the experimental and control groups. 4 MARKS

Q9

- (a) This is a two sample test for unpaired samples, with the experimental group having population mean μ_1 say, with m and n small. We wish to test

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 < \mu_2$$

and α is not specified so we can take $\alpha = 0.05$. Here the test statistic uses the pooled sample variance

$$s_p^2 = \frac{(m-1)s_1^2 + (n-1)s_2^2}{m+n-2} = 0.68584,$$

so that $s_p = 0.8281546$, and is given by

$$T = \frac{\bar{x} - \bar{y}}{S_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad \text{with} \quad T_{obs} = \frac{6.45 - 7.13}{0.8281546 \sqrt{\frac{1}{10} + \frac{1}{12}}} = -1.9176829$$

The Rejection Region for this one sided alternative is $\{T \leq -t_{\alpha, m+n-2}\}$, that is, $\{T \leq -1.7247\}$ from the t-table line where $\nu = 20$. Thus, T_{obs} lies in the Rejection Region, so we reject H_0 in favour of H_a at this value of α .

In this small sample case, we require that the original data are Normally distributed, with equal population variances, as well as the usual assumption of independence of the two random samples.

- (b) The 95% confidence interval in this case is

$$(\bar{x} - \bar{y}) \pm t_{\alpha/2, m+n-2} s_p \sqrt{\frac{1}{m} + \frac{1}{n}} = (-1.4196843, 0.0596843).$$

as $t_{\alpha/2, m+n-2} = 2.0860$ from tables.

10. A new drug designed to reduce fever is being tested for efficacy and side effects. Ten patients entering a hospital with a high fever were selected at random. The temperature (in °F) of each patient was measured, the drug was administered, and two hours later, the temperature was measured again. The data and summaries are provided in the table below.

Patient	Before Drug	After Drug	Before - After
1	102.6	99.8	2.8
2	99.2	97.5	0.4
3	102.3	97.5	4.8
4	101.1	100.3	0.8
5	102.7	99.6	3.1
6	102.6	102.8	-0.2
7	100.5	99.0	1.5
8	103.5	101.8	1.7
9	105.7	97.1	8.6
10	104.3	99.2	5.1
Mean	102.45	99.59	2.86
Variance	3.50	3.06	7.18

- (a) Conduct the appropriate hypothesis test to determine whether there is any evidence that the new drug reduces the mean patient temperature after two hours. Use $\alpha = 0.05$. 6 MARKS
- (b) What assumptions are necessary for this test to be statistically valid? 2 MARKS
- (c) Data from 12 patients in a second hospital were recorded. Explain briefly how you would test to assess whether the effect of the drug was different in the two hospitals. 2 MARKS

- (a) Here we have a two sample paired problem, with $n = 10$ small. Suppose that the first measurement has population mean μ_1 and the second measurement has population mean μ_2 . We test the hypotheses

$$H_0 : \mu_1 = \mu_2$$

$$H_a : \mu_1 > \mu_2$$

The test statistic is based on the sample difference **standard deviation** with observed value

$$s_d = \sqrt{s_d^2} = \sqrt{7.18} = 2.6795522$$

and takes the form

$$T = \frac{\bar{d}}{S_d/\sqrt{n}} \quad \text{with} \quad T_{obs} = \frac{2.86}{2.6795522/\sqrt{10}} = 3.3752334$$

The Rejection Region for this one sided alternative is $\{T \geq t_{\alpha, n-1}\}$, that is, the interval $\{T \geq 1.8331\}$ from the t-table line where $\nu = 9$. Thus, T_{obs} lies in the Rejection Region, so we reject H_0 in favour of H_a at this value of α .

- (b) We need the sample differences to be Normally distributed (and the usual assumption of independence of measurements across individuals).
- (c) You would have two samples of differences which were not paired, so you could carry out a hypothesis test for unpaired data for the equality of means of differences against the two-sided alternative hypotheses. The data would be the two samples of **differences**. As the sample sizes are both small, you would need to assume Normality of the two populations of differences, and also that the variances of the two populations of differences were the same. Finally, you would need the usual assumption of independent measurements in the two hospitals.