Sample exam 1

Please note:

- For the actual exam, these questions will be delivered to you in Quiz format via the LMS.
- You will be required to submit handwritten answers to the short answer questions, and you will be required to upload these at the end of the exam.
- Late submissions cannot be accepted.

Because the exam is open book, detailed solutions to the sample exams will not be provided. Answers to selected questions will be provided so that you can check your work.

Question 1. 14 marks

A medical professional is using a machine learning model to assist in diagnosing an unusual but relatively harmless disease. The model may produce a positive result or a negative result. If a patient tests positive for the illness, then they are referred to a specialist for further investigation; otherwise, the patient takes no further action.

It is known that approximately 4.1% of the population has the illness. After training the model and running it on test data, the true positive rate for the test is estimated as 75.1%, and the false positive rate for the test is estimated as 0.8%.

Let T denote the event that a patient tests positive for the illness, and let I denote the event that a patient has the illness.

(a) **(1 mark)** From the information above, state the probabilities $P(T \mid I)$, $P(T \mid I^c)$ and $P(I)$.

(b) **(2 marks)** Use the Law of Total Probability to determine $P(T)$.

(c) **(3 marks)** Determine $P(I \mid T)$ and $P(I^c \mid T^c)$. What names are given to these probabilities?

(d) **(3 marks)** Determine the false omission rate and negative predictive value of the classifier.

(e) **(3 marks)** A patient tests positive for the illness, and based on the true positive rate for the test, comes to believe that there is a 75.1% chance that they have the illness. Is this line of reasoning justified? If so, explain. If not, explain a more suitable line of reasoning, and give the correct probability.

(f) **(2 marks)** To further study this illness, a researcher tests people at random until they find 20 people who test positive for the illness. Let X denote the number of people who test negative before 20 people who test positive are found.

What probability distribution does X follow? Give your answer in the form $X \sim$ ____.

Show all working. Give your answers to at least 3 decimal places.

Question 2. 7 marks

The probability mass function for a joint random variable (X, Y) is shown below.

$P(X = x \cap Y = y)$	$y = 1$	$y = 2$	$y = 6$
$x = 1$	0.01	0.13	0.11
$x = 3$	0.04	0.05	0.09
$x = 4$	0.06	0.19	0.05
$x = 7$	0.02	0.08	0.17

(a) **(2 marks)** Write down the probability mass function for Y.

(b) Determine each of the following:

(i) **(1 mark)** $P(X = 1 \cap Y = 1)$

(ii) **(1 mark)** $P(X = 1 \mid Y = 1)$

(iii) **(1 mark)** $P(Y = 1 \mid X = 1)$

(iv) **(2 marks)** $P(Y = 2 \mid X < 7)$

Question 3. 12 marks

Imagine that you have been asked to monitor a machine that manufactures small parts. The machine builds parts one at a time, but on average, 4% of the parts are faulty and they cannot be used. Assume that each attempt at building a part is independent of any other attempt.

(a) Suppose another company has asked you manufacture 50 of these small parts. These 50 parts must be working (i.e., not faulty).

(i) **(2 marks)** In the context of this problem, write a sentence to define a relevant random variable X with $X \sim \mathrm{NB}(r, p)$. You must also give the values of r and p.

An appropriate sentence might start with, "Let X be the number of...".

(ii) **(2 marks)** What is the expected number of parts in total (faulty plus working) that will be manufactured after completing the company's request?

(iii) **(3 marks)** Determine the probability that 2 or more parts are faulty.

(b) Some time later, you will use the machine to build exactly 120 parts (regardless of whether they are faulty or not).

(i) **(3 marks)** Write a sentence to define a suitable random variable in the context of this problem, and then describe the probability distribution that random variable follows using correct notation for random variables.

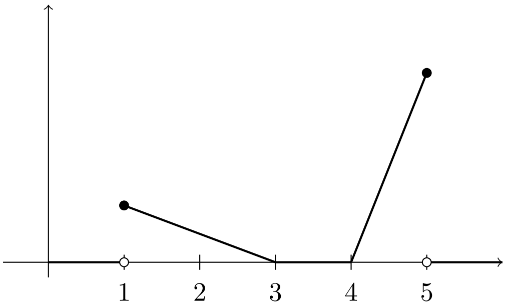
(ii) **(2 marks)** Write one line of R code that will give the probability that 10 or fewer of these parts will be faulty. *You do not need to calculate this probability.*

Question 4. **11 marks**

A random variable X has its probability density function given by

$$f_X(x) = \begin{cases} \frac{9-3x}{16}, & \text{if } 1 \leqslant x < 3, \\ \frac{5x}{4} - 5, & \text{if } 4 \leqslant x \leqslant 5, \\ 0, & \text{otherwise.} \end{cases}$$

Its graph is shown below on the interval $[0, 6]$.



You must use the formula to determine the vertical values.

(a) **(2 marks)** Write code to implement this function in R. Take care to use correct syntax and symbols that would be used when typing code.

(b) **(3 marks)** Explain why $f_X(x)$ is a valid probability density function. Show details of any relevant calculations.

(c) **(3 marks)** Determine $P(X \leqslant \frac{9}{2})$.

(d) **(3 marks)** Determine a formula for $P(X \leqslant x)$ for values of x such that $1 \leqslant x < 3$.

Show all working. Give numeric answers ***exactly, as fractions****, in all cases.*

Question 5. **10 marks**

A small electronics repair store has a single staff member working each day. The time this staff member takes to repair a device is exponentially distributed, with an average completion time of 24 minutes. The repair store is open for 8 hours each day, and at closing time, the staff member will stop working immediately. They will then continue any unfinished jobs at the start of the next working day. Customers bring devices to the store according to a Poisson process, with an average number of 15 devices arriving per 8-hour working day. Hence, the number of devices in the store can be modelled as M/M/1 queue.

Assume that the queue has been running for long enough that its steady state properties apply.

(a) **(1 mark)** Write down the service rate μ, the arrival rate λ, and the traffic intensity ρ, for this queue. Include relevant units.

(b) **(1 mark)** What is the average number of devices in the queue (including any which are currently being repaired)?

(c) **(2 marks)** On average, after a customer brings their device to the store, how long will it be before their device begins being repaired?

(d) **(3 marks)** If the store opens at 9AM, what is the probability that 2 or fewer customers bring a device to the store before noon?

(e) **(3 marks)** On average, how much time in one 8-hour work day will the staff member spend repairing a device?

Show all working. Give your answers to at least 3 decimal places.

Question 6. 8 marks

Suppose you had two candidate algorithms, algorithm A and algorithm B, designed to perform analysis on large data sets. You wish to determine which of the two algorithms, if any, has the fastest average running time on large data sets. Let μ_A and μ_B denote the true mean running time for algorithm A and B, respectively, and then let $\mu_d = \mu_A - \mu_B$. Assume that differences are normally distributed.

To determine the best algorithm, you conduct a paired two-sample t-test by running each algorithm on the same collection of $n = 11$ different data sets. Some results are summarised in the table below.

$\overline{x}_A$	s_A	$\overline{x}_B$	s_B	$\overline{x}_d$	s_d
67.45	25.27	56.82	26.81	10.64	7.09

The times are measured in minutes.

Some R console input and output is shown below. You will need to use some of these numbers in some parts of this question.

```
> pt(0.975, df = 10)
[1] 0.8237223
> pt(0.975, df = 11)
[1] 0.8247428
> pt(0.975, df = 19.93)
[1] 0.8293766
> qt(0.975, df = 10)
[1] 2.228139
> qt(0.975, df = 11)
[1] 2.200985
> qt(0.975, df = 19.93)
[1] 2.086433
```

(a) **(2 mark)** Calculate the standard error for the estimated mean difference.

(b) **(1 marks)** What is the degrees of freedom for this t-test?

(c) **(2 marks)** Calculate a 95% confidence interval for the difference in the two means.

(d) **(3 marks)** Write an appropriate conclusion on the basis of your confidence interval.

Show all working. Give your answers to at least 3 decimal places.

Question 7. 18 marks

Performance of computer hardware is often determined using software that applies a number of standardised tests, which returns a single unitless number called a *benchmark*, which can be used to compare different hardware models. A higher benchmark corresponds to better performing hardware. In this question, you will consider linear regression applied to GPU benchmarks.

The numerical variables you will consider are:

- `memSize`, the amount of memory available, in gigabytes (GB)
- `gpuClock`, the GPU clock speed, in megahertz (MHz)
- `unifiedShader`, the number of unified shaders
- `rop`, the number of render output units

You do not need to understand the technical aspects of these variables for this question, but you are given that each component may be relevant to the performance of a GPU.

Linear regression has been carried out in R using data from 46 GPUs. Partial output from the summary is shown on the next page, along with the residuals versus fitted plot and a normal Q-Q plot of the residuals.

(a) **(4 marks)** Are there any clear violations in the residuals versus fits plot or the Q-Q plot to be concerned with? Justify your answer clearly with references to both plots.

NOTE: Regardless of your answer to (a), for the remainder of this question, assume that there are no linear regression model violations.

(b) **(2 marks)** Does the R output suggest that the regression model fits the data well? Explain.

(c) **(3 marks)** What is the estimate of the coefficient of `memSize`? Give a sentence interpreting this estimated coefficient, including any relevant units.

(d) **(2 marks)** Let β_1 denote the true coefficient for the `memSize` variable and consider the hypotheses

$$H_0\colon \beta_1 = 0 \text{ versus } H_1\colon \beta_1 \neq 0.$$

Do you reject the null hypothesis at the $\alpha = 0.05$ significance level? Explain.

(e) **(2 marks)** Using the fact that $t_{41,0.975} = 2.020$, construct a 95% confidence interval for β_1.

(f) **(2 marks)** In the context of the problem, interpret the confidence interval you calculated in part (e).

(g) **(3 marks)** Suppose the following code was run, with associated output shown. The variable `model` contains the linear model created for this question.

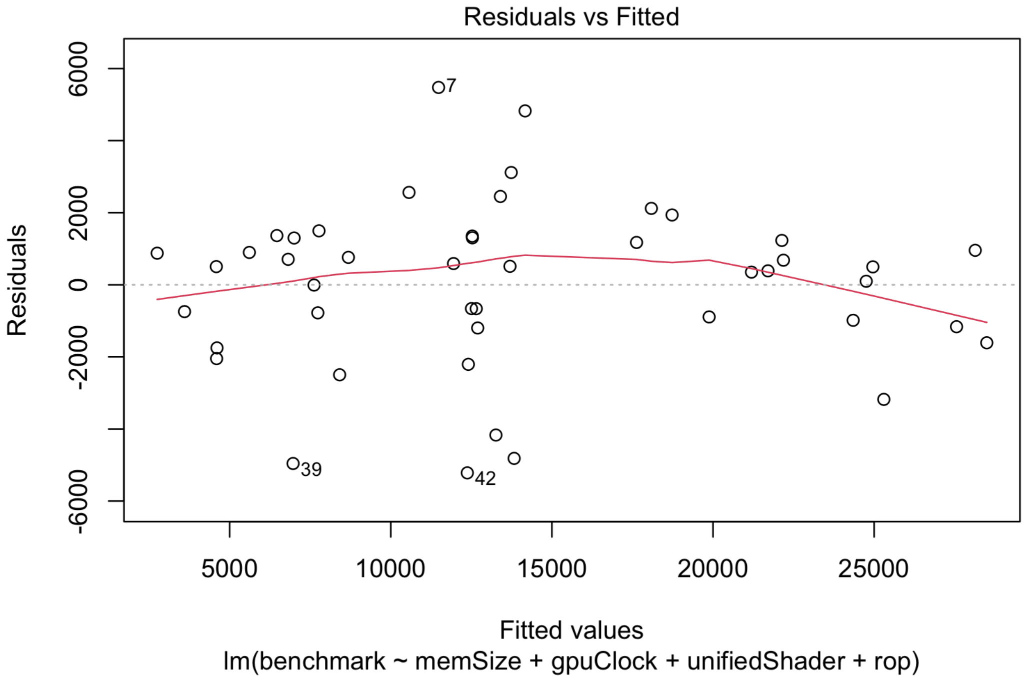
```
> df <- data.frame(memSize=10, gpuClock=1489, unifiedShader=2304, rop=64)
> predict(model, df, interval="prediction")
   fit     lwr     upr
1 12697.02 7816.93 17577.1
```

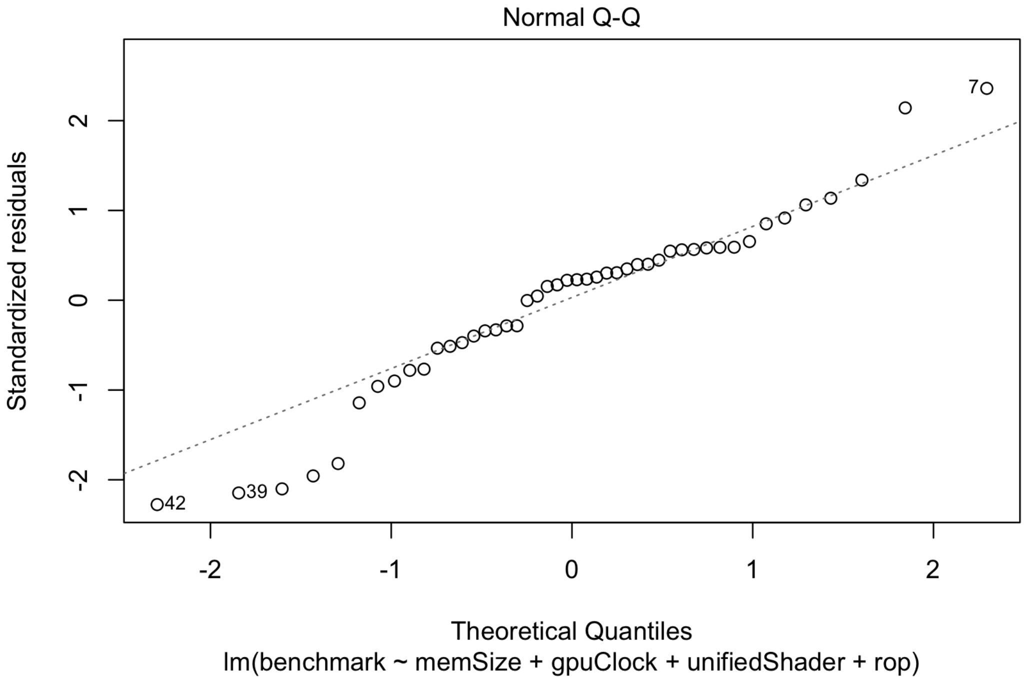
Provide a simple, in-context statement that interprets this output.

Show all working. Give your answers to at least 3 decimal places.

```
Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept)   -437.3939  1227.8856  -0.356   0.7235
memSize       -107.8043    62.0232  -1.738   0.0897 .
gpuClock         1.6601     0.8227   2.018   0.0502 .
unifiedShader    1.1944     0.2479   4.819 2.01e-05 ***
rop            140.4494    22.0894   6.358 1.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2376 on 41 degrees of freedom
Multiple R-squared:  0.9117,    Adjusted R-squared:  0.9031
F-statistic: 105.8 on 4 and 41 DF,  p-value: < 2.2e-16
```





Question 8 **4 marks**

From the following, which would be suitably described by a continuous random variable?

(a) The diameter of a tree trunk.

(b) The number of steps taken by a person in one day.

(c) The weight of a newborn baby.

(d) The length of an M/M/1 queue.

(e) The number of people enrolled at La Trobe University in one year.

(f) The distance a bird flies when looking for food.

There is at least one correct answer and there may be more than one. Select the correct answers.

For Question 8, you will be marked on the following basis:

- *If you select every correct answer and no incorrect answers, then you will get 4 marks.*
- *If you select no incorrect answers, but not every correct answer, you will get 2 marks.*
- *If you select at least one incorrect answer, and select more correct answers than incorrect answers, you will get 1 mark.*
- *Otherwise, you will get 0 marks.*

Question 9 — 4 marks

Suppose the following hypotheses were to be tested, for unknown means μ_1 and μ_2:

$$H_0\colon \mu = 0 \text{ versus } H_1\colon \mu_1 \neq 0.$$

To test the hypotheses, a t-test was performed in R, and the following output was obtained:

```
data:  x
t = 2.3438, df = 54, p-value = 0.0228
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.1733575 2.2244452
sample estimates:
mean of x
 1.198901
```

Assume a 5% level of significance. Which of the following statements are valid conclusions?

(a) Because $p < 0.05$, we know that the mean is not equal to zero.

(b) Because $p < 0.05$, we accept the null hypothesis.

(c) Because $p < 0.05$, we reject the null hypothesis.

(d) The evidence suggests that the mean is not zero.

(e) This is a two-sample test.

(f) To two decimal places, the 95% confidence interval is $(0.17, 2.22)$.

There is at least one correct answer and there may be more than one. Select the correct answers.

For Question 9, you will be marked on the following basis:

- *If you select every correct answer and no incorrect answers, then you will get 4 marks.*
- *If you select no incorrect answers, but not every correct answer, you will get 2 marks.*
- *If you select at least one incorrect answer, and select more correct answers than incorrect answers, you will get 1 mark.*
- *Otherwise, you will get 0 marks.*

Question 10 3 marks

A random variable X has its probability density function $f(x)$ given by

$$f_X(x) = \begin{cases} e^{x-3} & \text{if } x \leqslant 3 \\ 0 & \text{otherwise} \end{cases}$$

The following code was run:

```
> f <- function(x) { ifelse(x <= 3, exp(x-3), 0) }
> integrate(f, lower=-Inf, upper=2)$value
0.3678694
```

What information can be deduced from this output?

(a) $P(X \leqslant 2) \approx 0.368$

(b) $P(X \geqslant 2) \approx 0.368$

(c) $P(0 \leqslant X \leqslant 2) \approx 0.368$

(d) $E(X) \approx 0.368$

(e) $\mathrm{Var}(X) \approx 0.368$

(f) f is a probability density function

There is one correct answer. Select the correct answer.

Question 11 3 marks

Suppose that tasks arrive at a server at a rate of $\lambda = 4$ jobs per second. The server can process the tasks at a rate of $\mu = 7$ tasks per second. If the server is busy processing a task, any arriving tasks join a queue behind any other task already waiting to be processed. You may assume that the process can be modelled by an M/M/1 queue.

On average, how many seconds will a task spend in the system (this includes time waiting in the queue before processing starts and the time spent being processed)?

(a) 4/7

(b) 1/3

(c) 4/21

(d) 4/3

(e) 4

(f) 2/3

Question 12 **3 marks**

A probability mass function for a random variable Z is given to you below in table form.

z	1	3	4	5	11
$P(Z = z)$	0.168	0.209	0.12	0.213	0.29

The value of $P(Z < 5)$ is:

(a) 0.213

(b) 0.497

(c) 0.503

(d) 0.6

(e) 0.71

(f) 0.8

There is one correct answer. Select the correct answer.

Question 13 **3 marks**

Consider the following scenario: at a grocery store, on average, 1 in every 100 cans of baked beans has a dent in it. What is the appropriate calculation to perform to answer the following question: the store has ordered a delivery of 1250 cans of baked beans; what is the probability that at most 20 cans will have a dent?

(a) $P(X > 20)$, where $X \sim \text{Bin}(1250, 0.01)$

(b) $P(X < 20)$, where $X \sim \text{Bin}(1250, 0.01)$

(c) $P(X = 20)$, where $X \sim \text{NB}(1250, 0.01)$

(d) $P(X < 20)$, where $X \sim \text{Pois}(12.50)$

(e) $P(X < 21)$, where $X \sim \text{Bin}(1250, 0.01)$

(f) $P(X \leqslant 21)$, where $X \sim \text{Bin}(1250, 0.01)$

There is one correct answer. Select the correct answer.