Question	Mark	Out of
A1		8
A2		12
A3		10
A4		6
A5		9
B1		8
B2		7
B3		10
C1		6
C2		12
C3		6
C4		6
TOTAL		100

Note you will need to use Formula Sheet in this unit to answer the questions.

Instructions

Answer each question in the space provided. You can write in pen or pencil. Marks are indicated next to each question. The total mark for the exam is 100.

Part A (45 marks in total)

Question A.1 (1+1+1+1+2+1+1=8 marks)

Consider the following set of numbers: -25, 2, 3, 8, 10, 14, 18, 21, 32. For each of the questions below, state your answer, showing working if necessary.

(a) What is the median? 10

The given data set has already been ranked and the number of values is 9 which is odd, we can apply $x_{(n+1)/2}$ where $n = 9$ to get the median

(b) What is the 1st quartile? 2.5

The formula to be used here will be

$$Q_k = x_p + \frac{q}{4}(x_{p+1} - x_p)$$
$$p = floor((k(n+1))/4)$$
$$q = (k(n+1)) \ mod \ 4$$

where $n = 9$, $k = 1$. So

$$p = floor((1 \times (9+1))/4) = 2$$
$$q = (1 \times (9+1)) \ mod \ 4 = 2$$
$$Q_1 = x_2 + \frac{2}{4}(x_3 - x_2) = 2 + \frac{2}{4} \times (3-2) = 2.5$$

(c) What is the 3rd quartile? 19.5

Using the formula in Q(b), let $k = 3$, $n = 9$

$$p = floor((3 \times (9+1))/4) = 7$$
$$q = (3 \times (9+1)) \ mod \ 4 = 2$$
$$Q_3 = x_7 + \frac{2}{4}(x_8 - x_7) = 18 + \frac{2}{4} \times (21-18) = 19.5$$

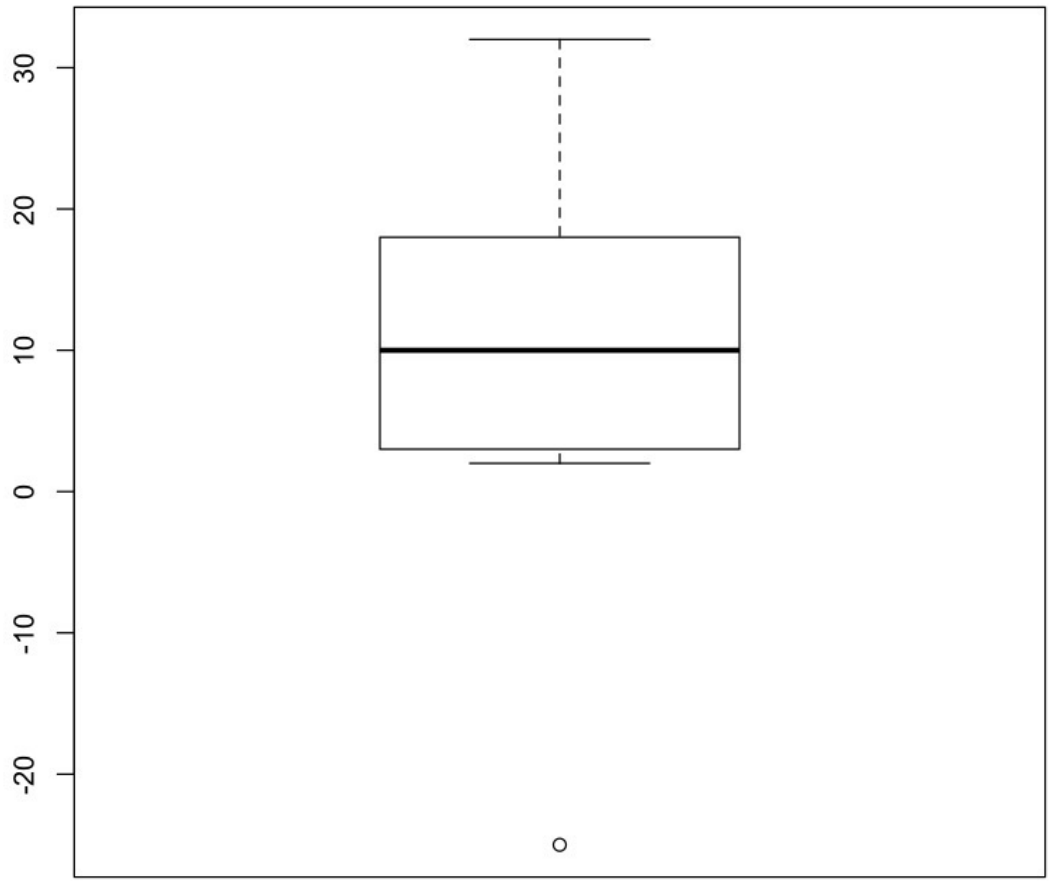
Marks // 3

(d) What is the interquartile range. 17

$\text{IQR} = Q_3 - Q_1 = 19.5 - 2.5 = 17$

(e) Hence sketch a box-plot. Lay it out horizontally below. Be sure to mark the values of the various parts.

upper and lower hinges at Q3 and Q1, median at 10, upper and lower whiskers at 2 and 32, one outlier at -25



Marks // **3**

(f) You are told the mean of the numbers is 9.222 and the mean of their square is 309.666. What is the sample standard deviation?

$\sqrt{9/8 * (309.666 - 9.222^2)} = 15.896$

Sample standard deviation will be calculated through $S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}^2)}$, so here, $n = 9$, $\sum_{i=1}^{9} x_i^2 = 9 \times 309.666$, $\sum_{i=1}^{9} x_i = 9 \times 9.222$. Then

$$\begin{aligned}
&\sum_{i=1}^{9}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{9} x_i^2 - 2\bar{x}\sum_{i=1}^{9} x_i + \sum_{i=1}^{9}\bar{x}^2 \\
&= 9 \times 309.666 - 2 \times 9.222 \times 9 \times 9.222 + 9 \times 9.222^2 \\
&= 9 \times (309.666 - 2 \times 9.222^2 + 9.222^2) = 9 \times (309.666 - 9.222^2)
\end{aligned}$$

$S_x^2 = \frac{1}{8}\sum_{i=1}^{9}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \frac{1}{8} \times (9 \times (309.666 - 9.222^2)) = \frac{9}{8} \times (309.666 - 9.222^2)$.
So $S_x = \sqrt{9/8 * (309.666 - 9.222^2)} = 15.896$

(g) If you only knew the mean and sample standard deviation of the sample, what does Chebyshev's inequality tell you?

That the number of points outside $(9.222 - k * 15.896, 9.222 + k * 15.896)$ is less than $9/k^2$.

Chebyshev's Inequality: $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$, where μ is the mean and σ is the value of standard deviation.
This inequality is used to describe at least $1 - 1/k^2$ of the distribution's values are within k standard deviations of the mean, or equivalently, no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean.
For example, 36 students' average mark in an exam is 80, the standard deviation is 10, then we may know, the number of students whose mark is less than 50 ($k = 3$, 3 standard deviation away from the mean) is 4 ($= 36 \times \frac{1}{3^2}$).

Marks // 2

Question A.2 (4+2+2+4=12 marks)

Throughout this question, show your working and leave your answer in a clear from. Of those reporting to a medical clinic, 2% have medical condition Z. It is assumed that this figure of 2% is also the base rate across the population. There is a test for condition Z such that, for those patients who have condition Z, 85% will test positive; and for those patients who do not have condition Z, 25% will test positive.

(a) If a patient tests positive, what is the probability that the patient has condition Z?

$p(Z, P) = 1/50 * 17/20;$
$p(not\ Z, P) = 49/50 * 1/4$
$p(P) = p(Z, P) + p(not\ Z, P) = 262/1000$
$p(Z|P) = p(Z, P)/p(P) = 1/50 * 17/20/(262/1000) = 17/262 = 0.065$

Let Z represent a patient has condition Z, P represent tests positive, so $not\ Z$ means do not have condition Z, then we will calculate the probability as $p(Z|P)$. According to the Bayes theorem,

$$p(Z|P) = \frac{p(Z,P)}{p(P)} = \frac{p(P|Z)p(Z)}{\sum_{x \in X} p(P|x)p(x)}$$

where $x = Z$ or $x = not\ Z$. So $\sum_{x \in X} p(P|x)p(x) = p(P|Z)p(Z) + p(P|not\ Z)p(not\ Z)$.
From the question, we know $p(Z) = 0.02$, $p(not\ Z) = 0.98$, $p(P|Z) = 0.85$, $p(P|not\ Z) = 0.25$. Apply these values into the formula:

$$p(Z|P) = \frac{p(P|Z)p(Z)}{p(P|Z)p(Z) + p(P|not\ Z)p(not\ Z)} = \frac{0.85 \times 0.02}{0.85 \times 0.02 + 0.25 \times 0.98} \approx 0.065$$

After some consideration, it is decided that the test gives too many false positives, and it is decided to modify the test as follows. The new test is simply to administer the original test twice, where it is assumed that these two tests give results that are independent of one another. A patient will be considered to have tested positive on the new test precisely in those cases where both tests on the original test return a positive result.

(b) If a patient has condition Z, what is the probability that the patient will test positive on the new test?

$p(P|Z) = 17/20$ and are independent, so simply multiple,
$p(PP|Z) = 17/20 * 17/20 = 0.7225$

Marks // 6

(c) If a patient does not have condition Z, what is the probability that the patient will test positive on the new test?

As above but using $p(P|not\ Z) = 1/4$,
$p(PP|not\ Z) = 1/16$

(d) If a patient returns a positive result on this new test, what is the probability that the patient has condition Z?

Build it up as done for part (a):
$p(Z, PP) = 1/50 * 17/20 * 17/20; p(not\ Z, PP) = 49/50 * 1/16$
$p(PP) = 1514/20000$
$p(Z|PP) = 1/50 * 17/20 * 17/20/(1514/20000) = 289/1514 = 0.191$

$p(Z, PP) = p(Z)p(PP|Z) = 1/50 * 17/20 * 17/20$

$p(not\ Z, PP) = p(notZ)p(PP|not\ Z) = 49/50 * 1/16$

$p(PP) = p(Z)p(PP|Z) + p(not\ Z)p(PP|not\ Z) = 1/50 * 17/20 * 17/20 + 49/50 * 1/16 = 1514/20000$

$$p(Z|PP) = \frac{p(Z)p(PP|Z)}{p(PP)} = \frac{1/50 * 17/20 * 17/20}{1514/20000} = \frac{289}{1514} \approx 0.191$$

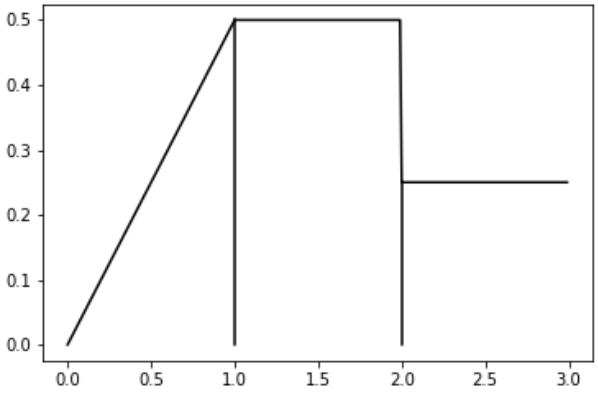
Marks	// **6**

Question A.3 (2+3+3+2=10 marks)

Consider the probability density function given at the right, defined by

$$p(x) = \begin{cases} \frac{1}{2}x & : \quad 0 \leq x \leq 1 \\ 0.5 & : \quad 1 \leq x \leq 2 \\ 0.25 & : \quad 2 \leq x \leq 3 \\ 0 & : \quad \text{otherwise} \end{cases}$$

Consider the cumulative density function $P(x)$ corresponding to $p(x)$, and the quantile function $Q(p)$.



(a) What is $P(0.5)$ and $Q(0.375)$?

Do not need a correct quantile function, can read off graph.
For $P(0.5)$, $P(1) = 0.25$, and geometrically, $P(0.5)$ is a quarter of that, so $P(0.5) = 0.0625$.
If you want to do it the long way you can integrate:

$$\begin{aligned} P(x = 0.5) &= \int_{-\infty}^{x=0.5} p(y)dy \\ &= \int_{-\infty}^{0} p(y)dy + \int_{0}^{0.5} p(y)dy \\ &= 0 + \int_{0}^{0.5} \frac{1}{2} y dy \\ &= \left[\frac{1}{2} \cdot \frac{1}{2} y^2\right]_0^{0.5} = \frac{1}{16} = 0.0625 \end{aligned}$$

For $Q(0.375)$:
We want $Q(p = 0.375)$ and p is an area under $p(x)$. We note by geometry the area between $0 \leq x < 1$ is 0.25, so then we just need the x value in the second rectangle that increases the area to 0.375. The area is $0.375 - 0.25 = 0.125$. In $1 \leq x < 2$, the height of the curve is 0.5, so $0.125 = 0.5(x - 1) \rightarrow 0.25 = x - 1 \rightarrow x = 1.25$.

(b) Derive the function for $P(x)$.

Have to do the intergral for the first part, for $0 < x < 1$, and the rest we can just write down

Marks // **2**

geometrically, since it will be linear (since we just integrate constants).

$$P(x) = \begin{cases} 0 & : \quad x \leq 0 \\ \frac{1}{4}x^2 & : \quad 0 < x \leq 1 \\ 0.25 + 0.5(x-1) = 0.5x - 0.25 & : \quad 1 \leq x \leq 2 \\ 0.75 + 0.25(x-2) = 0.25x + 0.25 & : \quad 2 < x < 3 \\ 1 & : \quad x \geq 3 \end{cases}$$

Here are the steps if you want to do all the integrals:

$$P(x) = \begin{cases} 0 & , \quad x < 0 \\ 0 + \int_0^x \frac{y}{2} dy & , \quad 0 \leq x < 1 \\ 0.25 + \int_1^x 0.5 dy & , \quad 1 \leq x < 2 \\ 0.75 + \int_2^x 0.25 dy & , \quad 2 \leq x < 3 \\ 1 & , \quad x \geq 3 \end{cases} = \begin{cases} 0 & , \quad x < 0 \\ \left[\frac{1}{2}\frac{y^2}{2}\right]_0^x & , \quad 0 \leq x < 1 \\ 0.25 + [0.5y]_1^x & , \quad 1 \leq x < 2 \\ 0.75 + [0.25y]_2^x & , \quad 2 \leq x < 3 \\ 1 + 0 & , \quad x \geq 3 \end{cases} = \begin{cases} 0 & , \quad x < 0 \\ \frac{x^2}{4} & , \quad 0 \leq x < 1 \\ 0.5x - 0.25 & , \quad 1 \leq x < 2 \\ 0.25x + 0.25 & , \quad 2 \leq x < 3 \\ 1 & , \quad x \geq 3 \end{cases}$$

(c) Hence give the quantile function $Q(p)$ corresponding to $p(x)$.

This comes from re-arranging part (b).
$p = 0$, $x < 0$ can ignore since next interval covers $p = 0$.
$p = \frac{x^2}{4} \rightarrow x = \sqrt{4p} = 2\sqrt{p}$, as $0 \leq x < 1$, $0 \leq p < \frac{1}{4}$.
$p = 0.5x - 0.25 \rightarrow x = 2(p + 0.25)$, as $1 \leq x < 2$, $\frac{1}{4} \leq p < \frac{3}{4}$.
$p = 0.25x + 0.25 \rightarrow x = 4(p - 0.25)$, as $2 \leq x < 3$, $\frac{3}{4} \leq p < 1$.
So

$$Q(p) = \begin{cases} 2\sqrt{p} & : \quad p \leq 0.25 \\ 1 + 2(p - 0.25) & : \quad 0.25 \leq p \leq 0.75 \\ 2 + 4(p - 0.75) & : \quad p \geq 0.75 \end{cases}$$

(d) Hence, or otherwise, write pseudo-code for an algorithm that will generate a sample from this distribution.

Use an inverse sampler with $Q(p)$ above. I.e.
Define $Q(p)$ as in (c)
Sample x from p(x) as follows
1. sample u uniformly from $[0, 1]$
2. take $x = Q(u)$.

Marks // **8**

Question A.4 (2+2+2=6 marks)

If $\mathbb{E}[X] = 1$ and $\mathbb{E}\left[X^2\right] = 4$, $\mathbb{E}[Y] = 0$ and $\mathbb{E}\left[Y^2\right] = 1$, and X and Y are independent, then:

(a) Calculate $\mathbb{E}\left[2X^2 + (X+1)^2\right]$.

$= 2\mathbb{E}\left[X^2\right] + \mathbb{E}\left[X^2 + 2X + 1\right] = 2\mathbb{E}\left[X^2\right] + \mathbb{E}\left[X^2\right] + 2\mathbb{E}[X] + \mathbb{E}[1] = 2*4+4+2*1+1 = 15$

(b) Calculate $\mathbb{E}\left[(X+1)(Y+1)^2\right]$.

$= \mathbb{E}[(X+1)]\,\mathbb{E}\left[(Y+1)^2\right] = \mathbb{E}[(X+1)]\,\mathbb{E}\left[Y^2 + 2Y + 1\right]$
$= (\mathbb{E}[X] + \mathbb{E}[1])(\mathbb{E}\left[Y^2\right] + 2\mathbb{E}[Y] + \mathbb{E}[1]) = (1+1)(1+2*0+1) = 4$

(c) Calculate $\mathbb{V}[(X+1)(Y+1)]$.

(According to $\mathbb{V}[X] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$)

$= \mathbb{E}\left[(X+1)^2(Y+1)^2\right] - \mathbb{E}[(X+1)(Y+1)]^2 = (4+2*1+1)(1+2*0+1) - (1+1)^2(1)^2 = 10$

Marks // **6**

Question A.5 (3+3+3=9 marks)

Consider the probability density function given by a mixture of two Gaussians with identical standard deviation σ, as

$$p(x|\rho, \mu_1, \mu_2, \sigma) = \rho N(x|\mu_1, \sigma) + (1-\rho)N(x|\mu_2, \sigma)$$

where $N(\cdot|\cdot)$ is the probability debsity function of a Gaussian. Thus the expected value of function $f(x)$ under this distribution is given by

$$\mathbb{E}_{\rho,\mu_1,\mu_2,\sigma}\left[f(x)\right] = \rho\mathbb{E}_{N(\mu_1,\sigma)}\left[f(x)\right] + (1-\rho)\mathbb{E}_{N(\mu_2,\sigma)}\left[f(x)\right]$$

where the two expected values on the right hand side are done using Gaussian distributions.

(a) What is the mean of x for the mixture of two Gaussians?

According to the second formula provided: $\mathbb{E}_{1,2}\left[f(x)\right] = \rho\mathbb{E}_1\left[f(x)\right] + (1-\rho)\mathbb{E}_2\left[f(x)\right]$. For the mean we use $f(x) = x$ and note
$\mathbb{E}_1\left[x\right] = \mu_1$ and $\mathbb{E}_2\left[x\right] = \mu_2$.
So $\mathbb{E}_{1,2}\left[x\right] = \rho\mu_1 + (1-\rho)\mu_2$

(b) What is the mean of x^2 for the mixture of two Gaussians?

Now $f(x) = x^2$
from the formula sheet $\mathbb{V}\left[g(x)\right] = \mathbb{E}\left[g(x)^2\right] - \mathbb{E}\left[g(x)\right]^2$,
for $g(x) = x \to \mathbb{V}\left[x\right] = \mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 \to \mathbb{E}\left[x^2\right] = \mathbb{V}\left[x\right] + \mathbb{E}\left[x\right]^2 = \sigma^2 + \mu^2$.
So,

$$\begin{aligned}\mathbb{E}_{1,2}\left[x^2\right] &= \rho\mathbb{E}_1\left[x^2\right] + (1-\rho)\mathbb{E}_2\left[x^2\right] \\ &= \rho(\sigma^2 + \mu_1^2) + (1-\rho)(\sigma^2 + \mu_2^2) \\ &= \sigma^2 + \rho\mu_1^2 + (1-\rho)\mu_2^2\end{aligned}$$

Marks // 6

(c) What is the variance for the mixture of two Gaussians?

Recall $\mathbb{V}[x] = \mathbb{E}\left[x^2\right] - \mathbb{E}[x]^2$, so

$$\begin{aligned}\mathbb{V}_{1,2}[x] &= \mathbb{E}_{1,2}\left[x^2\right] - \mathbb{E}_{1,2}[x]^2 \\ &= \sigma^2 + \rho\mu_1^2 + (1-\rho)\mu_2^2 - (\rho\mu_1 + (1-\rho)\mu_2)^2 \\ &= \sigma^2 + (\rho - \rho^2)(\mu_1 - \mu_2)^2\end{aligned}$$

Marks // **3**

Part B (25 marks in total)

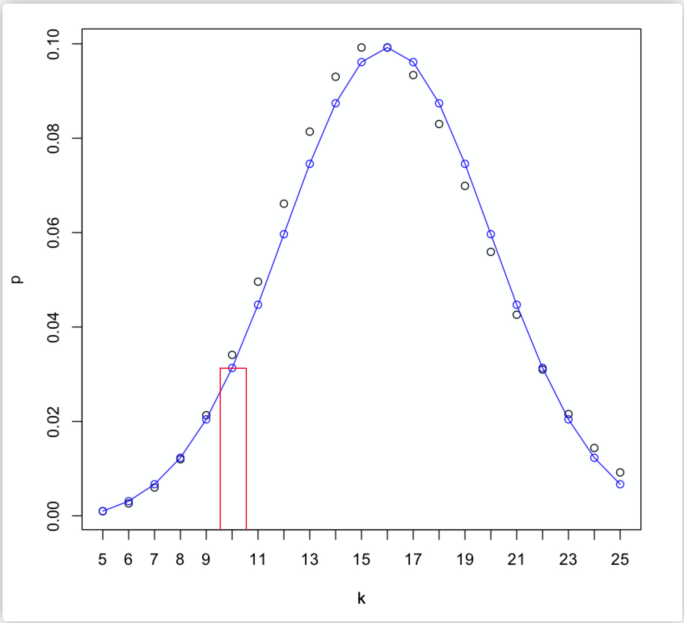
Question B.1 (3+2+3=8 marks)

You have data x distributed as Poisson with rate $\lambda = 16$, so $x \sim \text{Pois}(16)$.

(a) Show how to use the central limit theorem to get an approximate value for $p(10 \leq x \leq 20)$. Compute the approximate value, noting that the Z tables are only accurate to 2 decimal places.

So approximately $x \sim N(16, 16)$.

As shown in the figure, black dots are plotted using the Poisson distribution while the bule line plots the approximate Normal distribution. Since CLT is applied, when we computing the approximate value, we will standardise the normal distribution to $N(0, 1)$ by using the formula $z = \frac{x-\mu}{\sigma}$. As the original distribution is a discrete one, value 10 will be in the middle of 9.5 and 10.5, 20 will be in the middle of 19.5 and 20.5. Therefore $p(10 \leq x \leq 20) \approx p(9.5 \leq x \leq 20.5) = p(-6.5/4 \leq Z \leq 4.5/4)$ for Z a standard normal. Thus the answer from the Z tables, which only give a few decimal places, $0.86 - 0.05 = 0.81$.



You won't be penalised if use $p(-6.0/4 \leq Z \leq 4/4)$ in the exam.

(b) You have a sample of 10 values from this distribution, and compute its mean $\overline{x}$. What is an approximate distribution for $\overline{x}$?

So approximately $\overline{x} \sim N(16, 16/10)$.

Since it is a Poisson distribution, $\mathbb{E}[x] = \lambda = 16$, $\mathbb{V}[x] = \lambda = 16$. Using CLT in the formula sheet, mean(μ) of the distribution is $\lambda = 16$, and variance (σ^2) is $\lambda = 16$ as well. Then the sample mean's approximate distribution: $\overline{x} \sim N(\mu, \frac{1}{n}\sigma^2)$ where $n = 10$.

(c) What are 95% confidence intervals for the mean $\overline{x}$, according to this approximation?

So approximately $\overline{x} \sim N(16, 16/10)$. Using $Z = 1.96$ for confidence intervals, we get $[16 -$

Marks // 5

$1.96\sqrt{16/10}, 16 + 1.96\sqrt{16/10}]$ which is $[16 - 2.48, 16 + 2.48]$.

Using the approximation distribution for $\overline{x}$ in (b): $\overline{x} \sim N(16, 16/10)$, as 95% is required, we know $\alpha = 0.05 \rightarrow \alpha/2 = 0.025$. Looking up the z-table in the formula sheet for z-values greater than zero, we need to find the z-value $z_{0.025}$, for which the probability $p = 1-\alpha/2 = 1-0.025 = 0.975 \rightarrow z_{0.025} = 1.96$. So we get $[16 - 1.96\sqrt{16/10}, 16 + 1.96\sqrt{16/10}]$.

Marks // 3

Question B.2 (2+5=7 marks)

While IQ is considered to have a mean of 100 and standard deviation of 15. You expect students in your masters class will have a higher mean.

(a) Given a sample of size 10, compute a one-sided 95% confidence interval in the form $(-\infty, I]$ for where the measured mean should lie.

The mean has distribution $N(100, 15^2/10)$. For the one-sided case, $Z = 1.64$. So the confidence interval is $(-\infty, 100 + 1.64 * 15/\sqrt{10}]$ which is $(-\infty, 107.78]$

From the question, we know it is a normal distribution where $\mu = 100$, $\sigma = 15 \rightarrow \sigma^2 = 15^2$. With the given sample whose size is 10 and CLT, then we have the mean as $N(100, 15^2/10)$. For the one-sided case, we should find $z_{0.5}$ by using the z-table for which the probability $p = 1 - \alpha = 1 - 0.05 = 0.95$. So here $z = 1.64$, the confidence interval will be $(-\infty, 100 + 1.64 * 15/\sqrt{10}]$.

(b) You get data from 10 students with the form $[104, 120, 100, 112, 133, 138, 111, 118, 114, 118]$. Note that the mean of the sample is 116.8 and the mean of the squares of the sample is 13765.8. Test the null hypothesis that the students' IQ has mean 100. Without assuming you know the standard deviation, give the test statistic and the p-value for this data. Note the tables of statistics given at the back of the exam will not allow you to lookup the p-value precisely.

Null hypothesis $H_0 : \ \mu_0 = 100$
Alternative hypothesis $H_A : \ \mu_0 \neq 100$
sample standard deviation is $S = \sqrt{(13765.8 - 116.8^2) * 10/9} = 11.7$;
since don't know variance use Student-t; by null hypothesis, $t_{\alpha,9} = (116.8-100)*\sqrt{10}/S = 4.54$ which is the test statistic; p-value is about 0.0005 (not much accuracy though, its a bit more); thus we reject the null hypothesis at the 0.1% level (could choose the 5% level i.e. $\alpha = 0.05$, it is up to you to choose the decision threshold. In practice we don't go higher than $\alpha = 0.05$ and α should ideally be chosen before calculating p. If we set the threshold lower we can be more confident in our decision if the p-value is below this threshold) and conclude that the student's IQ does not have mean of 100.

The sample standard deviation can be calculated quickly by using the formula

$$S = \sqrt{\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2\right)}$$

.

Marks // 7

Question B.3 (2+2+4+2=10 marks)

You obtain paired data (X, Y) with values $\vec{x} = [4.59, 4.60, 6.32, 4.85, 3.27, 5.92, 1.92, 6.90, 4.82, 5.39]$ and $\vec{y} = [2.89, 2.46, 3.28, 2.34, 2.11, 3.56, 1.77, 3.29, 2.46, 2.60]$. The various sample means (using the above data) are:

$$\begin{aligned} \overline{x} &= 4.859 \\ \overline{y} &= 2.677 \\ \overline{x^2} &= 25.516 \\ \overline{y^2} &= 7.460 \\ \overline{xy} &= 13.670 \end{aligned}$$

(a) What is the correlation co-efficient between X and Y? What does this tell you about X and Y?

$r^2_{xy} = \dfrac{SS^2_{XY}}{SS_{XX}SS_{YY}} = \dfrac{(n(\overline{xy} - \overline{x}.\overline{y}))^2}{n(\overline{x^2} - \overline{x}^2)n(\overline{y^2} - \overline{y}^2)} = 0.784$, so X and Y are highly correlated.

(b) Fit a simple linear model to this data in the form

$$\hat{Y} = \beta_0 + \beta_1 X$$

What are your estimates for β_0 and β_1?

$\hat{\beta}_1 = \dfrac{SS_{XY}}{SS_{XX}} = \dfrac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - \overline{x}^2} = 0.348$
$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} = 0.988$
NB. data was generated with $Y = 0.3 * X + 1.3 + \sqrt{(}0.3)Z$.

Marks // **4**

(c) What are the standard errors for β_0 and β_1?

From the formula sheet, we have

$$\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{\overline{X^2}}{\overline{X^2}-\overline{X}^2}}}(\hat{\beta}_0-\beta_0)\sim Student-t(n-2)$$

$$\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{1}{\overline{X^2}-\overline{X}^2}}}(\hat{\beta}_1-\beta_1)\sim Student-t(n-2)$$

within which, the square-root terms are the corresponding standard errors (i.e., error in estimating β_0, β_1).

Standard error of $\hat{\beta}_1$ is $\sqrt{\frac{RSS}{n(n-2)}\frac{1}{\overline{X^2}-\overline{X}^2}}$;

Standard error of $\hat{\beta}_0$ is $\sqrt{\frac{RSS}{n(n-2)}\frac{\overline{X^2}}{\overline{X^2}-\overline{X}^2}}$.

And

$$\begin{aligned}RSS(\beta_0,\beta_1)&=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1x_i)^2\\&=\sum_{i=1}^{n}y_i^2-\beta_0y_i-\beta_1x_iy_i-\beta_0y_i+\beta_0^2+\beta_0\beta_1x_i-\beta_1x_iy_i+\beta_0\beta_1x_i+\beta_1^2x_i^2\\&=\sum_{i=1}^{n}y_i^2-2\beta_0y_i-2\beta_1x_iy_i+\beta_0^2+\beta_1^2x_i^2+2\beta_0\beta_1x_i\\&=n\left(\overline{y^2}-2\beta_0\overline{y}-2\beta_1\overline{xy}+\beta_0^2+\beta_1^2\overline{x^2}+2\beta_0\beta_1\overline{x}\right)\end{aligned}$$

So $RSS(\hat{\beta}_0,\hat{\beta}_1)=0.631$
and $s.e.(\hat{\beta}_0)=0.325$
and $s.e.(\hat{\beta}_1)=0.064$.

(d) Test the hypothesis the $\beta_1=0$. What is your test statistic and its p-value? What is the outcome of the test?

The formula $\frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{1}{\overline{X^2}-\overline{X}^2}}}(\hat{\beta}_1-\beta_1)\sim Student-t(n-2)$ will be used to test the hypothesis:
$H_0:\ \beta_1=0$
$H_1:\ \beta_1\neq 0$
This is a linear regression problem, so t-test will be used, $n=10$, so $n-2=8$

$$\begin{aligned}t_{\alpha,n-2}&=\frac{1}{s.e.(\hat{\beta}_1)}(\hat{\beta}_1-\beta_1)\\&=\frac{1}{0.064}(0.348-0)=5.406\end{aligned}$$

p-value is slightly more than 0.0005, so we reject null hypothesis.

Marks // 6

Part C (30 marks in total)

Question C.1 (2+2+2=6 marks)

You have a data set supplied as real-valued pairs (X, Y) and you wish to regress X onto Y. You have 2 models:

A: a 4 degree polynomial

$$\hat{y} = \sum_{i=0}^{4} a_i x^i$$

B: a 20 degree polynomial

$$\hat{y} = \sum_{i=0}^{20} a_i x^i$$

These questions can be answered based on the contents in Alexandria and One in ten rule.

(a) Describe how the bias of models A and B differ.

A has higher bias with 5 versus 21 parameters; generally, A should not give as good a fit to the training data

(b) Describe how the variance of models A and B differ.

B has higher variance because it has more parameters generally, B give as good a closer giit to the training data but it can overfit, and when it does the variance will be higher

(c) If you had 100 data points in your sample, which of ther two models would you recommend? Justify your answer.

by the rule of thumb, with 100 data points, you should fit about 100/10=10 parameters, so safer to go with model A

Marks // 6

Question C.2 (5+3+2+2=12 marks)

(a) You wish to build a naïve Bayes classifier regressing Booleans A, B and C onto the Boolean X. Someone has already counted the data for you to create frequency tables below:

	A=0	A=1		B=0	B=1		C=0	C=1
X=0	10	40		30	20		15	35
X=1	30	20		5	45		40	10

Construct probability tables as needed to specify the estimated naïve Bayes classifier for the task. Then give the formula for the classifier and describe how it would be used.

First $p(X{=}0) = p(X{=}1) = 50/100 = 0.5$. Then the tables for $p(A|X)$, $p(B|X)$ and $p(C|X)$ respectively are created by normalising the above tables along the rows.

	A=0	A=1		B=0	B=1		C=0	C=1
X=0	10/50=0.2	40/50=0.8		30/50=0.6	20/50=0.4		15/50=0.3	35/50=0.7
X=1	30/50=0.6	20/50=0.4		5/50=0.1	45/50=0.9		40/50=0.8	10/50=0.2

Now using the unnormalised NBC formula we have $p(X|A,B,C) \propto p(X)p(A|X)p(B|X)p(C|X)$. To classify a vector $(A=a, B=b, C=c)$ as being in class $X=1$ or $X=0$, we look up $p(X=1|A=a,B=b,C=c)$ and $p(X=0|A=a,B=b,C=c)$ and set the class label to the value of X that gives the highest probability.

Marks // 5

(b) Consider the probabilities $p(A{=}0|X{=}0)$ and $p(B{=}0|X{=}1)$. Compute their standard errors, making any assumptions as needed? What can you say about the resulting estimates?

The RVs A and B are binary so treat as Bernoulli with parameter Θ. Now for Bernoulli RVs, the parameter estimate is $\hat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} y_i$ where y_i correspond to n observations (for A or B) that can be 0 or 1. Now standard error is the defined to be the standard deviation of the sampling distribution of a statistic. From the formula above we see that $\hat{\Theta}$ is equal to the sample mean for the Bernoulli RV. So by applying the the sample mean version of the CLT we see that the variance of the sampling distribution of the mean in this case is $\sigma^2/n = \Theta(1{-}\Theta)/n$ where $\Theta(1-\Theta)$ is the variance for a Bernoulli RV. Now standard deviation (aka standard error for the reason noted above) is the square root of the variance. So we have:

$$s.e.(\hat{\Theta}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\hat{\Theta}(1-\hat{\Theta})}{n}} = \sqrt{\frac{\hat{\Theta}(1-\hat{\Theta})}{50}},\ n = 50 \text{ for } X = 0 \text{ or } X = 1.$$

And so:

$$s.e.(p(A = 0|X = 0)) = \sqrt{\frac{0.2 \times 0.8}{50}} = 0.056 \approx 6\% \text{ error;}$$

$$s.e.(p(B = 0|X = 1)) = \sqrt{\frac{0.1 \times 0.9}{50}} = 0.042 \approx 4\% \text{ error.}$$

So both estimates are quite poor.

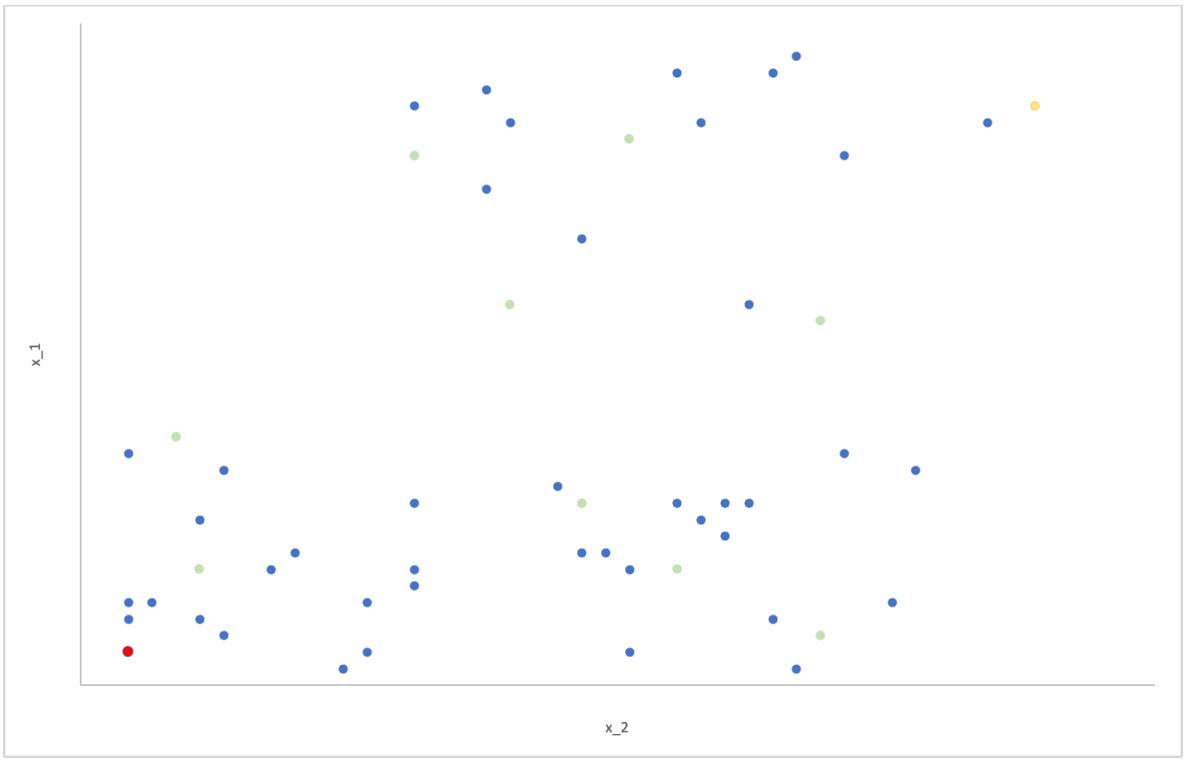
(c) Which would be better, the naïve Bayesian classifier or the logistic regression classifier for this data set? Justify your answer.

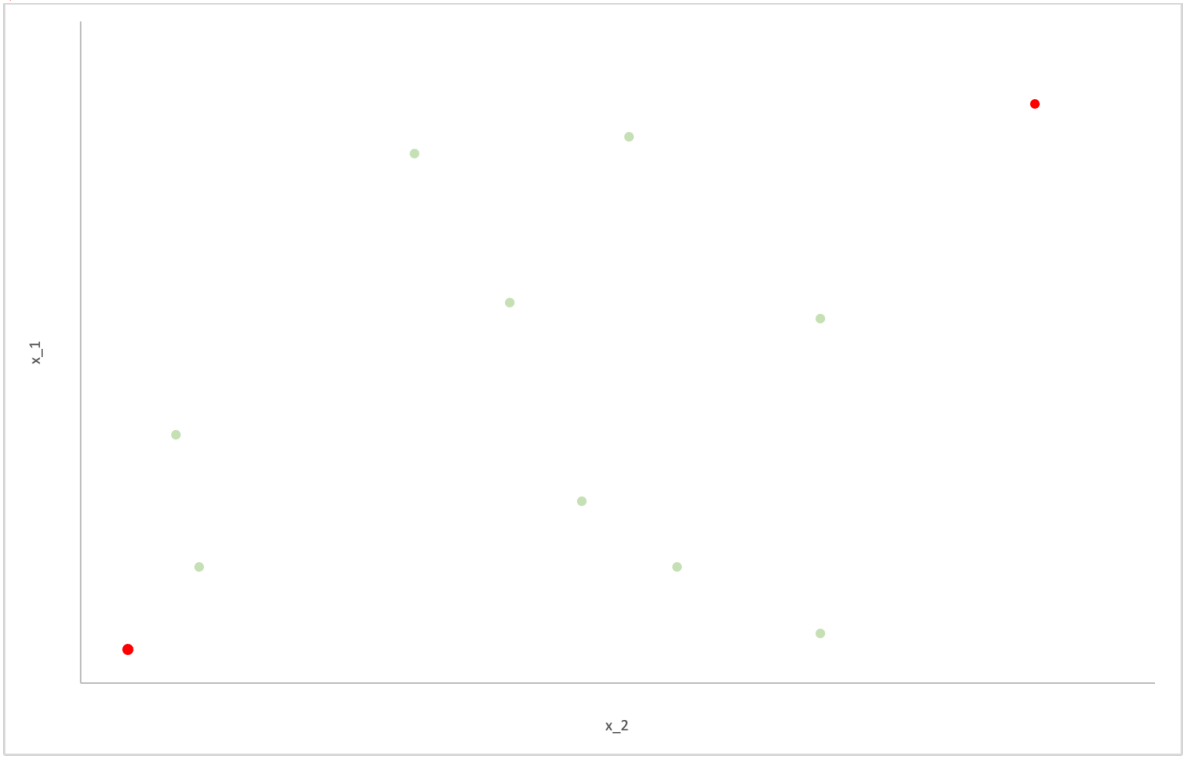
Its not really clear, since they have the same bias. But at a pinch since there is so few data, perhaps the NBC works better because it can be more robust with little data.

(d) The first step of the k-means algorithm is to initialise the centroids. Describe a way this could be done, and why it is OK to use it.

Many solutions. Pick a random point as first centroid. For the next $k-1$ centroids: select 10 random points and select the point furtherest away from the current batch of centroids.
It is OK to use this approach because there are many different possibilities to initialise and without any prior knowledge a random initialisation is a good starting point. Moreover, keeping the centroids away from each other initially will also probably lead to a better solution because if we start with them too close together we may end up with more than one centroid per actual cluster as we saw in the lecture.

Marks // 7


x_1
x_2

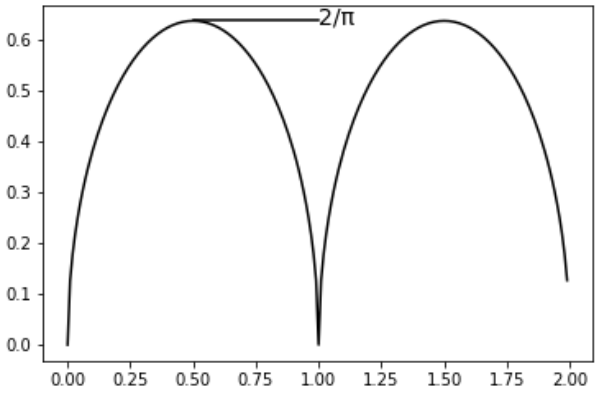

x_1
x_2

Question C.3 (6=6 marks)

Consider the probability density function given below, defined by

$$p(x) = \begin{cases} \frac{2}{\pi}\sqrt{1-(2x-1)^2} & : \quad 0 \le x \le 1 \\ \frac{2}{\pi}\sqrt{1-(2x-3)^2} & : \quad 1 \le x \le 2 \\ 0 & : \quad \text{otherwise} \end{cases}$$

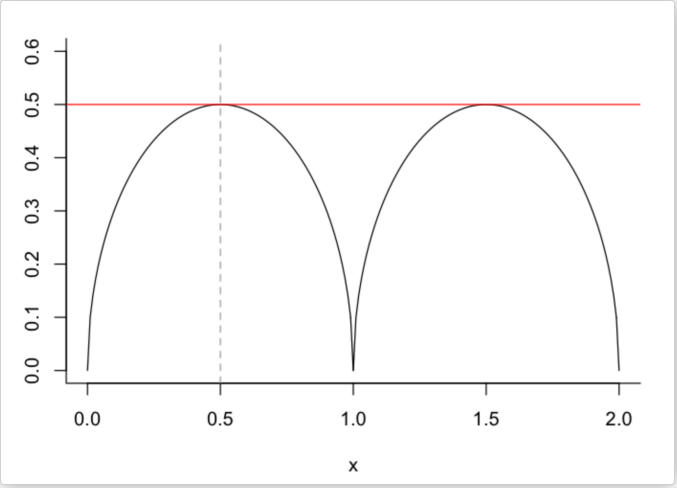
This is two semi-circles side-by-side of radius 1/2, then scaled by $4/\pi$ to get a PDF.



(a) Devise pseudo-code for a rejection sampler for this distribution. Note the maximum value is marked at $\frac{2}{\pi}$.

First we Choose $p_{prop}(x) = \frac{1}{2}$ on $x \in [0, 2]$ and let $q(x) = p(x)$. Now we need to find C by aligning the maxima of $p_{prop}(x)$ and $Cq(x)$ to minimise the number of samples that will be rejected:

$\frac{1}{2} = Cq(x)$ at $x = \frac{1}{2}$, this is
$\frac{1}{2} = C \cdot \frac{2}{\pi} \rightarrow C = \frac{\pi}{4}$.



Now here is the pseudo-code:
Choose $p_{prop}(x) = \frac{1}{2}$ on $x \in [0, 2]$, $C = \frac{\pi}{4}$ and $q(x) = p(x)$, sample x from $p(x)$ as follows:
1. Sample x from $p_{prop}(x)$
2. Sample U as uniform in $[0, 1]$
3. Reject x if $U > \frac{Cq(x)}{p_{prop}(x)} = \frac{\frac{\pi}{4}q(x)}{\frac{1}{2}} = \frac{\pi}{2}q(x)$ and return to step 1
4. Otherwise, accept x.

Marks // 6

Question C.4 (5+1=6 marks)

You wish to build a decision tree to predict a three-valued variable X. The first two features to test are Booleans A and B. Someone has already counted the data for you to create frequency tables below:

	A=0	A=1		B=0	B=1
X=0	10	40		30	20
X=1	30	20		5	45
X=2	30	20		45	5

(a) Compute and report the quality measure for the attributes A and B using the information gain metric.

Get conditional probabilities by normalising in columns:

	$p(X\|A)$ A=0	A=1		$p(X\|B)$ B=0	B=1
X=0	$10/70 = 1/7$	$40/80 = 1/2$		$30/80 = 3/8$	$20/70 = 2/7$
X=1	$30/70 = 3/7$	$20/80 = 1/4$		$5/80 = 1/16$	$45/70 = 9/14$
X=2	$30/70 = 3/7$	$20/80 = 1/4$		$45/80 = 9/16$	$5/70 = 1/14$

Then we need to find $H(X|A)$ and $H(X|B)$ as they determine quality of A and B in terms of information gain defined as $H(X) - H(X|Y)$ where Y can be A or B.
From formula sheet:

$$H(X|Y) = \sum_y p(Y = y) H(X|Y = y)$$

and

$$H(X|Y = y) = \mathbb{E}\left[\log_2 \frac{1}{p(X|Y = y)}\right] = \sum_i p(X = x_i|Y = y) \log_2 \frac{1}{p(X = x_i|Y = y)}$$

So we need

$$p(A = 0) = \frac{70}{150} = \frac{7}{15}$$
$$p(A = 1) = \frac{80}{150} = \frac{8}{15}$$
$$p(B = 0) = \frac{80}{150} = \frac{8}{15}$$
$$p(B = 1) = \frac{70}{150} = \frac{7}{15}$$

and conditional probabilities given in table above.
Plug conditional probs into RHS of $H(X|A=0)$, $H(X|A=1)$, $H(X|B=0)$, $H(X|B=1)$.
Once you have these, plug the values into RHS of $H(X|A)$ and $H(X|B)$.
This gives $H(X|A) = 1.48$ and $H(X|B) = 1.23$.
Important Note!!!!!
If your calculator cannot compute $\log_2(x)$ directly, then use:

$$\log_2(x) = \frac{\log_{10}(x)}{\log_{10}(2)}$$

or

$$\log_2(x) = \frac{\log_e(x)}{\log_e(2)}$$

(b) Hence say which attribute is recommended to use at the root of the tree?

In (a), $H(X|A) > H(X|B)$, so B should be split at the root of the tree, because entropy of B is lower which means the value of Information Gain is larger.
Actually don't need to solve (a) to conclude this since it can be seen in the table above that B is more biased to specific values than A, and so B is a purer prediction of X for the training data.

Marks // 6

Blank page for additional answers if needed.

Blank page for additional answers if needed.