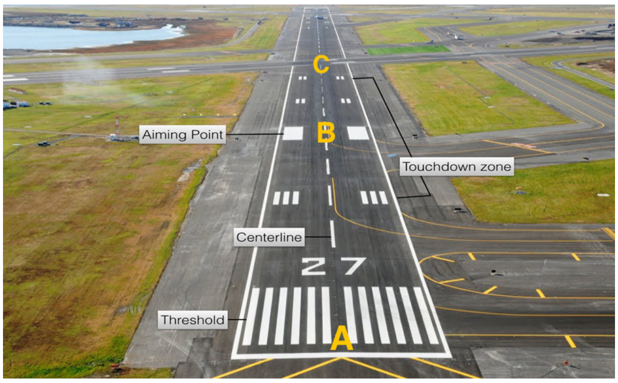
Assignment 1
Probability & Data (MATH 4043)

Instructions

- This assignment is worth 25% of your final grade and is due - 12pm Friday Week 6.
- Submission is online on the Learnonline website.
- Assignments will be marked and returned online via Learnonline.
- R output will be required to be produced for some of the solutions.
- Your answers should be typed and submitted as a PDF.
- The marks for each question are displayed next to the question.
- The assignment is worth a total of 80 marks.
- **A late submission will attract a penalty of 10% of maximum marks available per day, or part thereof, if the assignment is late. The cut-off time is 5pm each day.**

Questions



1. **(22 marks)** Ten small airplanes attempt independently of each other to land on a runway with the ground marked ABC (in order), where B is the “sweet” and aiming point for landing of the line AC. If the plane lands too early (A) then the plane is flying in too low, or if the plane lands too late (C) then the stopping distance becomes unsafe and must abort after position C. Each airplane can land somewhere on the runway with probability 0.9. Otherwise, the airplane ends up landing somewhere else or comes back around. If any one of the airplanes lands on the runway, then it is noticed that it will land in the AB section with probability 0.4, or in the BC section with probability 0.6.

a. **(4 marks)** Define the distribution for X and calculate the probability that any 6 out of the 10 airplanes manage to land somewhere on the runway.

b. **(5 marks)** What is the probability that any 6 out of the 10 airplanes manage to land somewhere on the AB section of the runway? The other ones could have either landed in the BC section or not on the runway at all.

c. **(4 marks)** Use Venn diagrams and probability axioms to prove that

$$(A_1 \cap A_2 \cap A_3^c) \cap (A_1 \cap A_2^c \cap A_3) = \emptyset$$

where A_1, A_2, A_3 are three sets.

d. **(4 marks)** *Flysafe* is one of the airplanes and always lands safely. The airport would like to analyse the scenario that a plane cannot land due to another airplane landing and FlySafe lands. Assume that most commonly six airplanes arrive close to one another, and it has been observed that usually one of the airplanes fails to land when *Flysafe* lands. What is the probability that *Flysafe* lands and one of the other five fails? *Hint: Use a result similar to part c.*

e. **(5 marks)** What is the probability that six airplanes manage to land somewhere on the runway, given that *Flysafe* is one of them?

Solution:

(a) Let the probability of landing on the runway, $p = 0.9$ and let X be the number of landings on the runway. Therefore, $X \sim Binomal(10, 0.9)$. For 6 out of 10 airplanes,

$$P(X = 6) = \binom{10}{6}(0.9)^6(0.1)^4 = 0.0117$$

(b) See that the probability of landing on runway in the AB section is $p = 0.9 \times 0.4 = 0.36$.
Let Y be the number of landings on the runway. Therefore, $Y \sim Binomal(10, 0.36)$.
Therefore, $P(Y = 6) = \binom{10}{6}(0.36)^6(0.64)^4 = 0.0767$.

(c) Let A_1, A_2, A_3 be three events. Use a Venn diagram to illustrate your argument. Essentially, the intersection between

$$A_3 \cap A_3^c = \emptyset$$

And

$$(A \cap A_3) \cap (A \cap A_3^c) = \emptyset.$$

Finally,

$$(A_1 \cap A_2 \cap A_3) \cap (A_1 \cap A_2^c \cap A_3^c) = \emptyset.$$

(d) Let A_1 be the event *Flysafe* lands on the runway, while A_i are the events of the other 5 airplanes landing on the runway. Thus

$$(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5 \cap A_6^c)$$

denotes a A_1 landing while A_6 fails. We must combine this with the other scenarios such as A_i not landing. Notice that

$$(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5 \cap A_6^c) \cap (A_1 \cap A_2^c \cap A_3 \cap A_4 \cap A_5 \cap A_6) = \emptyset$$

Thus, we have

$$P\left(\cup_{i=2}^{6} (A_1 \cap A_i^c \cap A_3 \cap A_4 \cap A_5 \cap A_6)\right) = 5 \times 0.9^5 0.1 = 0.295245.$$

(e) This is conditional on *Flysafe* landing on the runway. If part (c) was used then this is the result:

$$P(\cup_{i=2}^{6} (A_1 \cap A_i^c \cap A_3 \cap A_4 \cap A_5 \cap A_6) \mid A_1)$$
$$= \frac{P\left(\cup_{i=2}^{6} (A_1 \cap A_i^c \cap A_3 \cap A_4 \cap A_5 \cap A_6)\right)}{P(A_1)}$$
$$= \frac{0.295245}{0.9} = 0.32805.$$

However, if used part (a) by accident then I will accept it, so

$$P(six\ planes\ land \mid Flysafe) = \frac{0.0117}{0.9} = 0.13.$$

2. (10 marks) Consider an individual chosen at random from a city of interest, and let

$X = 0$, if the individual has neither lung cancer or tuberculosis disease;

$X = 1$, if the individual has lung cancer;
$X = 2$, if the individual has tuberculosis disease.

It is known from previous research that

$$P(X = 0) = 0.97, P(X = 1) = \ 0.01, P(X = 2) = 0.02.$$

These probabilities constitute the probability distribution of X prior to obtaining any information about the individual. We investigate further and take an X-ray of the individual. We can summarise the new information with the variable, Y. If the X-ray is positive then $Y = 1$, and if the X-ray is not "positive" then $Y = 0$.

Suppose the X-ray is positive, and we know

$$P(Y = 1 \mid X = 0) = 0.07,$$
$$P(Y = 1 \mid X = 1) = 0.95,$$
$$P(Y = 1 \mid X = 2) = 0.90.$$

It is critical to understand what the probabilities of no diseases, lung cancer and tuberculosis disease given the X-Ray is positive ($Y = 1$). Find the probabilities $P(X_i \mid Y = 1)$ for $i = 0, 1, 2$ and interpret each probability.

Solution: The aim in the question is to apply Bayes Rule and find $P(X_i \mid Y = 1)$, which can be written as

$$P(X_1 \mid Y = 1) = \frac{P(\, X = 1 \cap Y = 1)}{P(Y = 1)}$$

$$= \frac{P(Y = 1 \mid X = 1)P(X = 1)}{P(Y = 1 \mid X = 0)\ P(X = 0) + P(Y = 1 \mid X = 1)\ P(X = 1) + P(Y = 1 \mid X = 2)\ P(X = 2)}$$

= 0.0995

Repeat the same calculation for $X = 0, X = 2$ to obtain

$$P(X_0 \mid Y = 1) = 0.7117$$
$$P(X_2 \mid Y = 1) = 0.1886$$

Interpretation: There is a 9.95% chance that given the X-Ray is positive then the patient has lung cancer.
There is a 71.17% chance that given the X-Ray is positive then the patient is healthy/no disease is present which is what we want to see.
There is a 18.86% chance that given the X-Ray is positive then the patient has tuberculosis.

3. (12 marks) A fair coin is tossed 3 times. Let $X = (X_1, X_2)$, where X_1 counts the number of heads in the 3 tosses, and X_2 counts the number of tails in the 3 tosses.

a. **(2 marks)** Write down the sample space S.
b. **(3 marks)** Tabulate the joint probability mass function of $X_1$and X_2.

c. **(3 marks)** Calculate the marginal probability mass function of X_1 and X_2. What can you deduce about them?

d. **(4 marks)** Calculate $\rho(X_1, X_2)$, the correlation coefficient. Interpret this value and give an explanation for your answer.

Solution:

(a) The sample space is {HHH, TTT, HHT, THT, HTT, THH, HTH, TTH}.

(b) & (c)

X_2 X_1	0	1	2	3	$p_{X_2}(x_2)$
0	0	0	0	1/8	1/8
1	0	0	3/8	0	3/8
2	0	3/8	0	0	3/8
3	1/8	0	0	0	1/8
$p_{X_1}(x_1)$	1/8	3/8	3/8	1/8	

The marginal probability mass function has been calculated in the table and we deduce that they are the same. Also, we observe

$$p_{X_i}(x_i) = P(X_i = x_i) = \binom{3}{x_i}(0.5)^{x_i}(0.5)^{3-x_i}$$

which makes X_i a Binomial distribution.

(c) We can calculate the expected value and variance of X_1, X_2,

$$E(X_i) = 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = 1.5$$

$$Var(X_i) = \frac{24}{8} - \left(\frac{3}{2}\right)^2 = 0.75.$$

$$E(X_1X_2) = 2 \times \frac{3}{8} + 2 \times \frac{3}{8} = \frac{3}{2}.$$

Thus, the covariance can be calculated and is as follows,

$$Cov(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = \frac{3}{2} - \left(\frac{3}{2}\right)^2 = -0.75.$$

Thus, the correlation coefficient, $\rho(X_1, X_2) = \frac{Cov(X_1,X_2)}{\sigma_1\sigma_2} = -1$. This means that the number of heads and number of tails over 3 tosses are negatively correlated. This makes sense because if I toss 3 heads then there is no possibility for me to toss a tail, and vice versa, as well, if I toss 1 head and 2 tails then it's the same opposite 2 heads and 1 tails.

4. (18 marks) Computation Marketing: what is the expected revenue?

- Please download the sales data from Learnonline, ***sales.csv***.

You have started working at a sales company and they have asked you to investigate a particular product, its sales and the expected revenue where the cost of the product is $80. The data which has been given to you has three columns, “Date”, “Sales” and “Price ($)”.

a. **(2 marks)** Identify the most appropriate distribution that models the variable, *Sales*.
b. **(2 marks)** Use R to compute the mean of the *Price ($)* in the dataset over the time period.
c. **(3 marks)** Use R to calculate the revenue and attach the daily revenue to the right of the sales table in *Sales.csv*. Please provide a screenshot of the daily revenue in your solution.
d. **(2 marks)** Use R to calculate the mean and variance of the *revenue* ($) over the 61 days.
e. **(3 marks)** Use R to compute the probability density function for the number of *Sales* per day and use R to plot a histogram of the number of *Sales* and the *revenue*. Assume that the maximum sold per day is 20 units.
f. **(4 marks)** The company chooses to fix the *Price ($)* per day to be the expected (mean) of *Price ($)* over 61 days. The company states that the lowest revenue amount is $1021 before there is a loss of money. What is the probability that the *revenue* is less than $1021?
g. **(2 marks)** Summarise the results in parts a) – f) and give a conclusion in relation to sales and revenue when the company fixes the price.

Solution:

(a) The most appropriate distribution to model the variable, *Sales* is a Poisson distribution with the expected value/rate being $\lambda = 9.06 \ \approx 9$.

(b) The average (mean) *price* of a variable,

```
 > mean(sales$price)
[1] 145.8861
```

(c) Revenue is defined as Sales multiplied by Price, so use the various values in the sales column to produce a new column in R to represent the revenue.

```
> Revenue = Sales$sales * Sales$Price
```

(d) Use the commands > mean(revenue); var(revenue).

(e) From part (a), the probability density function can be represented in R, by the command

```
> dpois(0:20,9.06)
```

Using ggplot or plot command, create a probability plot, for example:

```
>plot(0:20, dpois(0:20,9.06),xlab=”Number of Sales”,ylab=”Probability”.
```

In addition, by using `>hist(sales$revenue), hist(sales$sales)` the histogram of revenue can be calculated.

(f) Let $\overline{price} = 145.8861$ from part (b). We are interested in

$$\begin{aligned} P(revenue < 1021) &= P(Sales * \overline{price} < 1021) \\ &= P(Sales < 6.998) \\ &= P(Sales < 7) \end{aligned}$$

By using

```
>ppois(7,9.06)
```

We have the probability of having less than 7 sales, thus

$$P(revenue < 1021) = 0.316917.$$

(g) We notice the distribution for the number of sales is Poisson, the histogram and probability density function reasonably align. The typical revenue made per day is

$1256.29. However, by analysing the histogram, we see that the distribution for revenue is slightly left-skewed and the average revenue is $1308. If they fix the price then we can use the sales distribution to calculate expected revenue and probability of making certain revenue targets. They need to be careful by how much they increase their price because the probability of the number of sales decreases.

5. (18 marks) *Multiple Choice, testing to see if you can count on luck.*

Suppose we conduct an experiment on taking a multiple choice test, let there be 20 questions, each with four possible answers, one of which is correct. We will test this over a large class of 50 students where each student guesses the answers randomly without reading the question. We place the condition that a student passes this test, if they score 8 or more correct answers. Let X be the number of correct answers for a randomly selected student.

a. **(2 marks)** What is the distribution for X?
b. **(3 marks)** What is the expected value for X and the standard deviation for X?
c. **(4 marks)** Use R to plot a probability density function (p.d.f) for X, produce a table for the p.d.f and calculate the probability of a student getting 8 or more correct answers.
d. **(4 marks)** Suppose we randomly select 10 students, what is the probability that at least three students pass just by guessing? From the 10 students, what is the expected number of students who will pass?
e. **(5 marks)** Use R to simulate the number of correct answers someone gets for 50 students to populate a sample for the class. Compute and compare the mean, variance, plot the distribution and probability of passing to the distribution of X.

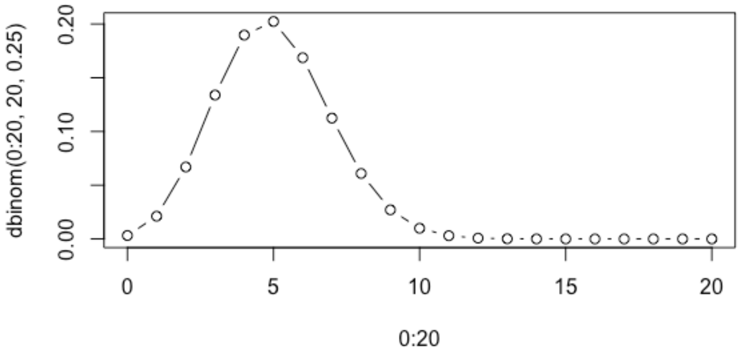
Solution:

(a) The most appropriate distribution for X is the Binomial Distribution, where X is the number of successes in an independent Bernoulli trial with probability p.

(b) The expected value and standard deviation for X is

$$\mu_X = 20 \times 0.25 = 5, \qquad \sigma_X = \sqrt{20 \times 0.25 \ \times 0.75} = 1.936.$$

```
(c) > dbinom(0:20,20,0.25)
     [1] 3.171212e-03 2.114141e-02 6.694781e-02 1.338956e-01 1.896855e-01
    2.023312e01
     [7] 1.686093e-01 1.124062e-01 6.088669e-02 2.706075e-02 9.922275e-03
    3.006750e03
    [13] 7.516875e-04 1.541923e-04 2.569872e-05 3.426496e-06 3.569266e-07
    2.799425e08
    [19] 1.555236e-09 5.456968e-11 9.094947e-13
```



```
> 1-pbinom(7,20,0.25)
[1] 0.1018119
```

(d) Let Y be the number of students who pass the exam. See that this follows a binomial distribution:

$Y \sim Binomial(10, 0.1018119)$. The expected value of Y is $E(Y) = 10 \times 0.1018119 = 1.018199$.

At least three students pass has the probability calculated by

$$P(Y \geq 3) =$$

```
> 1-pbinom(2,10,0.1018119)
[1] 0.07334512
```

(e)

```
> rbinom(50,20, 0.25)
[1] 3 5 6 4 8 6 3 8 9 4 3 5 6 4 6 2 5 7 4 8 6 5 8 4 1 6 4 7 3 2 3 3 6 4 8 6 5 7 4 0 5 5 8
[44] 5 8 7 3 6 4 3
Dist=rbinom(50,20,0.25)
> mean(rbinom(50,20,0.25))
5.38
> var(rbinom(50,20,0.25))
3.208571
```

The number that pass (TRUE) is the following:

```
> table(dist>=8)

FALSE  TRUE
   42     8
```

Therefore, we have a probability from our sample that $\frac{8}{50} = 0.16$.

By comparing the answers to part (a), we see that the mean is greater than part (a) but the variance is higher as well.