DATA101 Assignment 2, Semester 2 2025

Due Date: 1st August 2025

Student ID Number: 83972706

Total Marks: 20

Instructions:

- Complete all questions in R Markdown format
- If using this .Rmd file for your solution, please enter your student ID number in the space provided above
- Submit your completed assignment as a PDF through LEARN by 11:59pm on the due date
- Show all your R code and explain your reasoning where required
- Code that produces errors will receive minimal marks

For some reason the histograms code is not showing with ECHO = TRUE but all others are good to go

Question One: Blood pressure medication testing (4 marks)

A pharmacy quality control department tested 6 tablets of Lisinopril to determine the concentration of the active ingredient (in mg). The data is shown below.

9.87 10.23 9.94 10.18 9.76 10.02

For parts a-c please show either your working or the R code you used to answer these questions.

a. What is the mean and median for this dataset? (1 mark)

```
# Define the data vector
data <- c(9.87, 10.23, 9.94, 10.18, 9.76, 10.02)

# Calculate mean and median
data_mean <- mean(data)
data_median <- median(data)

# Output values
cat("The mean for this data is", data_mean, "and the median is", data_median)
```

```
## The mean for this data is 10 and the median is 9.98
```

b. What is the sample variance and sample standard deviation for this dataset? (1 mark)

```
## The standard deviation for this data is 0.1809972 and the sample variance is 0.03276
```

c. What type of data is being considered here? (1 mark)

A) Binary categorical data
B) Multiple categorical data
C) Continuous numerical data
D) Discrete numerical data
E) Ordinal categorical data

C: Continuous numerical data

Justify your answer.

Since the data is just numerical (No yes/no or alphabetical values) it is just continuous numerical

d. Lisinopril tablets are supposed to contain 10 mg active ingredient. If the test had been changed to record whether the tablet met the quality standard (within ±0.5 mg of target) or not, what type of data would you be dealing with? (1 mark)

A) Binary categorical data
B) Multiple categorical data
C) Continuous numerical data
D) Discrete numerical data
E) Ordinal categorical data

A: Binary categorical data

Justify your answer.

Since the value would be either yes/no or true/false this makes it Binary categorical data

Question Two: Ordinal versus nominal (3 marks)

a. What is the difference between an ordinal and nominal categorical variable?

Ordinal data is data that is in some sort of order; for example data that is in acsending order whereas nominal data are values that are not ordered. For example something such as [red, blue green] is nominal where [1, 2, 5, 9] is ordinal

b. Include in your answer any difference in the statistics you are able to calculate from each of these variable types.

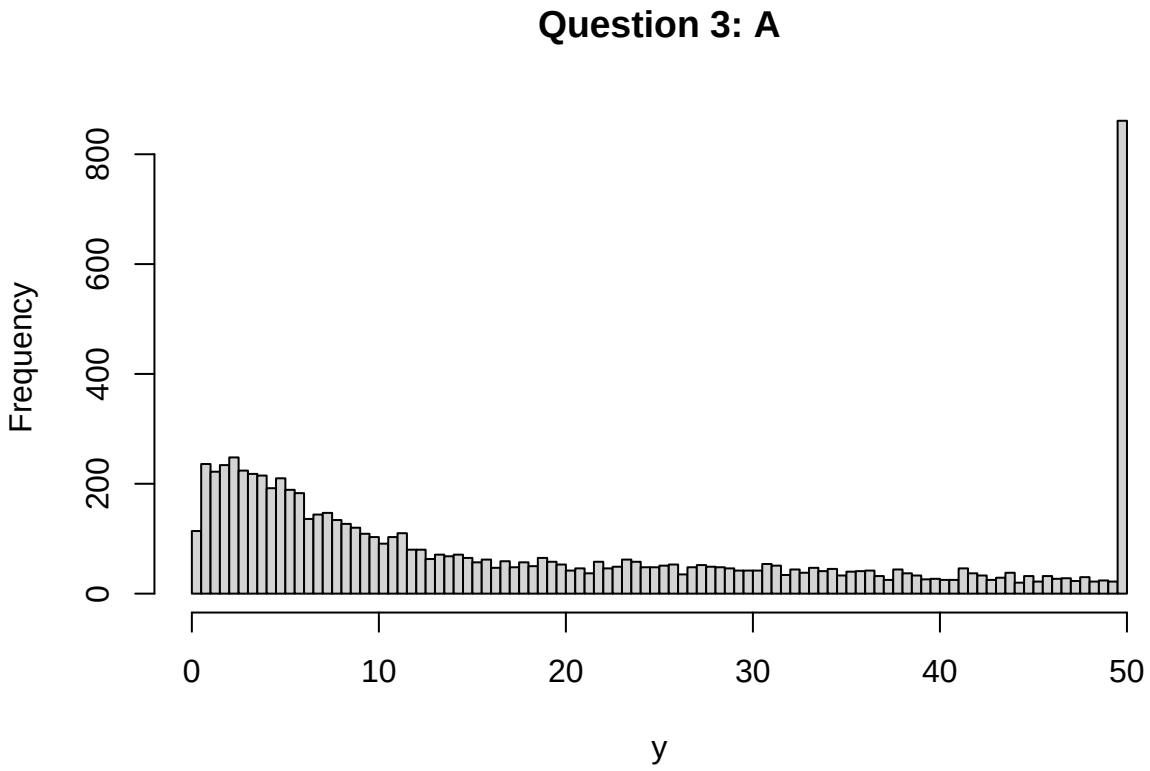
Ordinal data allows you to calculate things such as the median or mode where nominal values since they have no order does not allow you to calculate the median and only allows for calculations such as mode and frequency counts

Question Three: Analyze a distribution (7 marks)

You first need to run the code shown below. This will generate a random sample of 8000 observations, which you will use to answer the questions below.

To ensure the results can be reproduced, you will need to set the seed using the `set.seed` function, with your student id number as the input. If you decide to write your assignment answers in this R markdown document, remove the `eval=FALSE` you will see at the start of the chunk below.

a. Construct a histogram of y with 100 breaks in R. (1 mark)



b. Describe this distribution. Topics to consider in your answer include:

- Number of modes
- Symmetry/skewness
- Shape
- Type of variable, quantitative (continuous or discrete) or categorical
- Range of possible values.

The type of variable in this histogram is quantitative and the frequency ranges from about 25 to 800. The mode is at 50 and it's frequency is 800. The shape is positively skewed (Or right hand skewed). The only obvious outlier is at 50 and is considerably larger than every other frequency.

(3 marks)

c. Find the following quantiles of the sample y: 1%, 5%, 10%, 15%, ..., 85%, 90%, 95%, 99%. (1 mark)

d. For this distribution, what do you expect the order (from lowest to highest) of mean, median and mode to be? Justify your answer. (2 marks)

median mean mode

Since by the historgram we can see that the mode is so large that is obviously going to be the highest. The data around the middle is quite low so the median is going to be quite low. Then the mean is going to be heavily effected by the high mode and pretty high values at the start.

Question Four: Emergency department data (6 marks)

For this question, you will work with a dataset containing information about emergency department patient flow: `emergency_department_data.csv`. The dataset is avalaible on LEARN for you. To answer parts of this question, you will need to import this dataset into R.

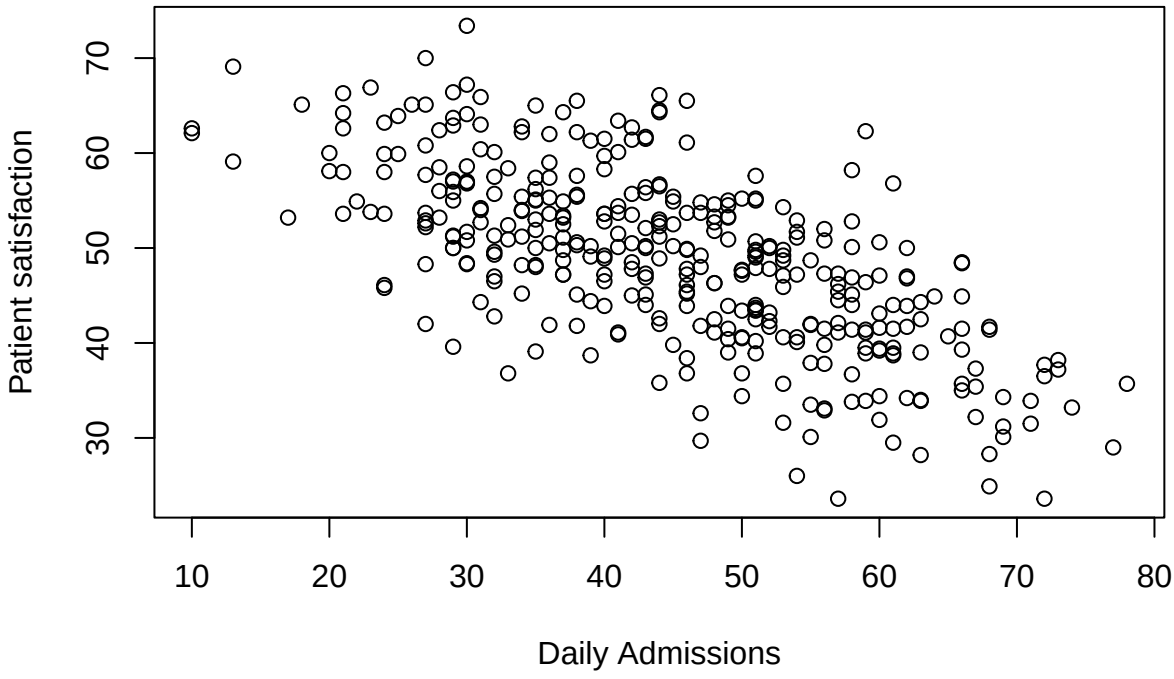
a. What does Pearson correlation, r, measure? (1 mark)

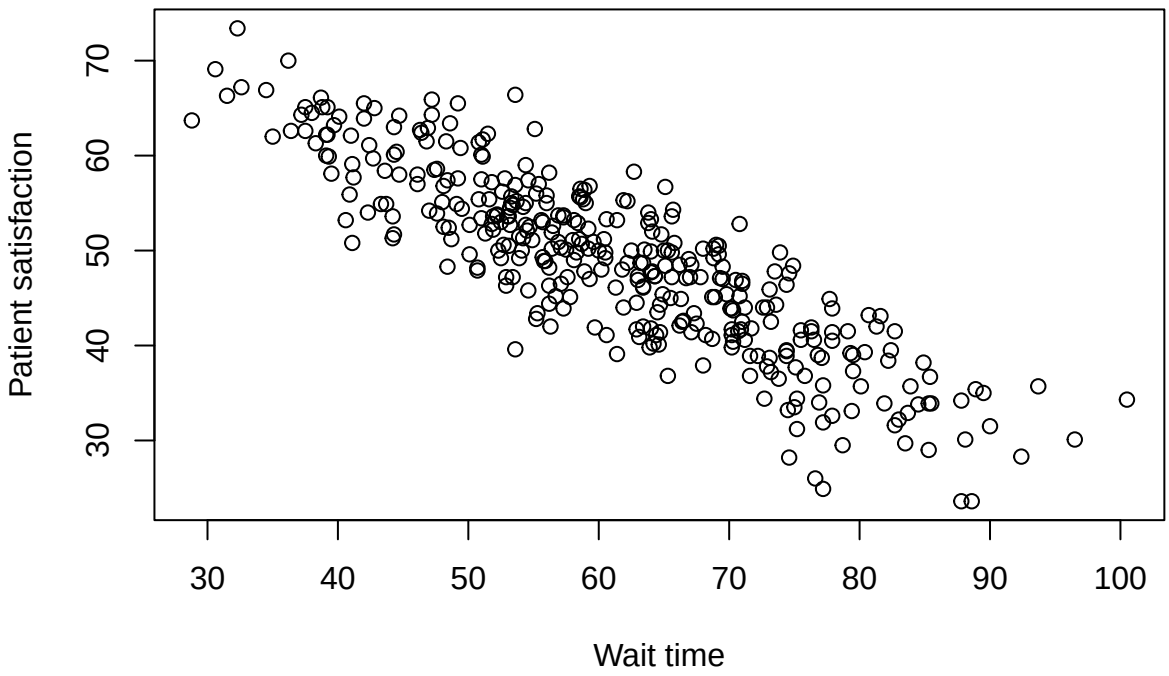
Direction and strength of a linear association between two quantitative variables

b. Compute the Pearson correlations between daily_admissions and the variables wait_time and patient_satisfaction. (1 mark)

c. Create scatter plots with:

- `daily_admissions` (as the explanatory variable) and `patient_satisfaction` (as the response variable)
- `wait_time` (as the explanatory variable) and `patient_satisfaction` (as the response variable)





Comment on the features of these plots. (2 marks)

Daily admissions vs Patient satisfaction

As Daily Admissions increases, overall patient satisfaction tends to decrease. From 20 daily admissions onward, there is quite consistent amounts of patient satisfaction ranging around 40-55 however this ceases from 66 onward. Overall the pattern appears to be decreasing linear but only very slightly thanks to the consistent values from 40-55. However, as admissions rise, the lowest satisfaction values become noticeably lower.

Wait time vs patient satisfaction

As wait time increases patient satisfaction very clearly decreases. The shape of the graph is decreasing linearly

d. Based on the plots above, do you believe Pearson correlation was an appropriate measure of association to use for:

- `daily_admissions` and `patient_satisfaction`
- `wait_time` and `patient_satisfaction`

Pearson correlation is appropriate for both daily_admissions and patient_satisfaction and wait_time and patient_satisfaction, because the plots for both pairs suggest roughly linear relationships. The second graph is easier to interpret and shows a clear linear trend, so Pearson correlation is especially suitable there.

Justify your answer. (2 marks)