Report Title

A Statistical Analysis of Campus Coffee Shop Data

Submitted To

Peter Neal

University of Nottingham

Report Prepared By

Charlie Davies

University of Nottingham

Date

15 Dec 2022

Introduction

A coffee shop on campus has commissioned an analysis of four sets of collected data to assist with planning for the next academic year. This report aims to collate the findings of an investigation into the datasets which attempts to provide an answer to the coffee shop owner's questions and presumptions. The coffee shop's opening hours are 08:00 – 18:00 from Monday to Friday and 10:00 – 16:00 on weekends. The following data has been provided:

- The number of customers served in 5-minute intervals on a Tuesday
- The number of customers served in 5-minute intervals between 08:30 – 09:00 from Monday to Friday for 10 weeks
- The number of customers served each Saturday and Sunday for 10 weeks
- The total income per day for 10 weeks along with several descriptive covariates

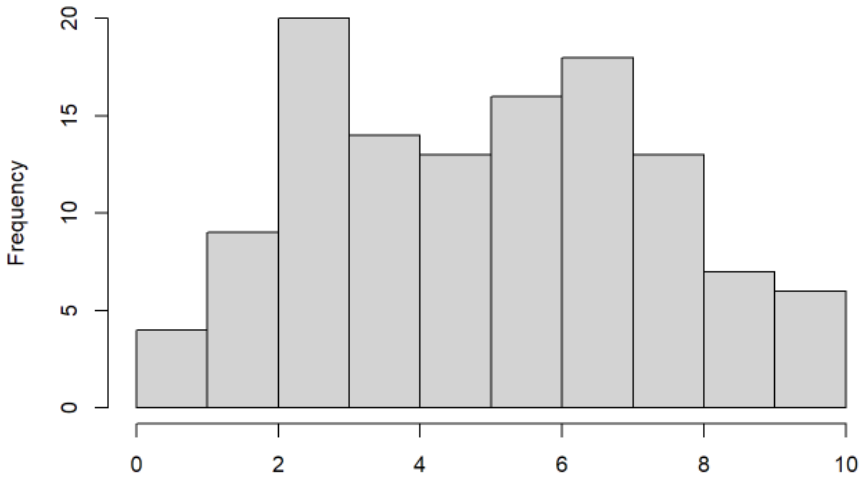
Please note the accompanying R code used for this report is uploaded as **coffee.html** and **coffee.Rmd**.

Question 1

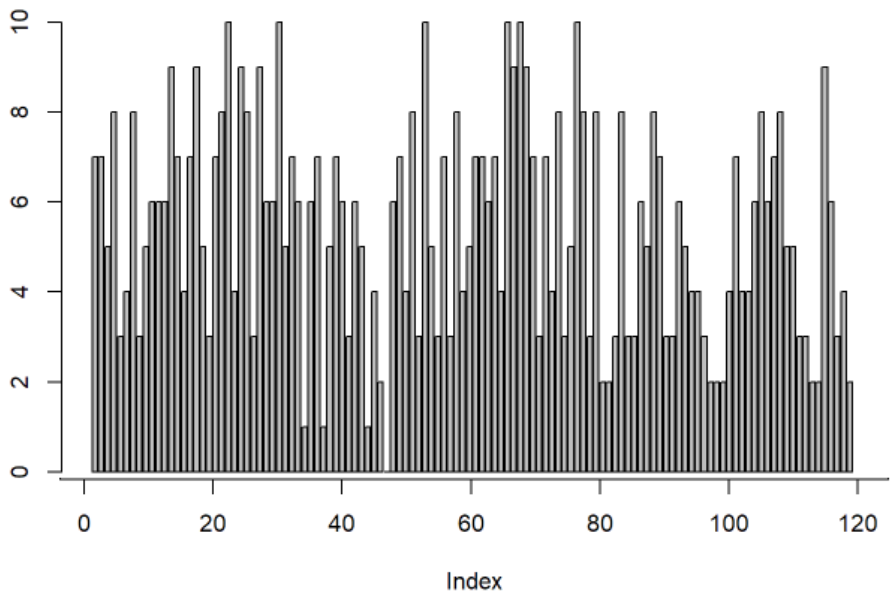
First of all, generating a summary of the data produces the following results.

```
     Time           Number.served
 Length:120         Min.   : 0.000
 Class :character   1st Qu.: 3.000
 Mode  :character   Median : 5.500
                    Mean   : 5.425
                    3rd Qu.: 7.000
                    Max.   :10.000
```

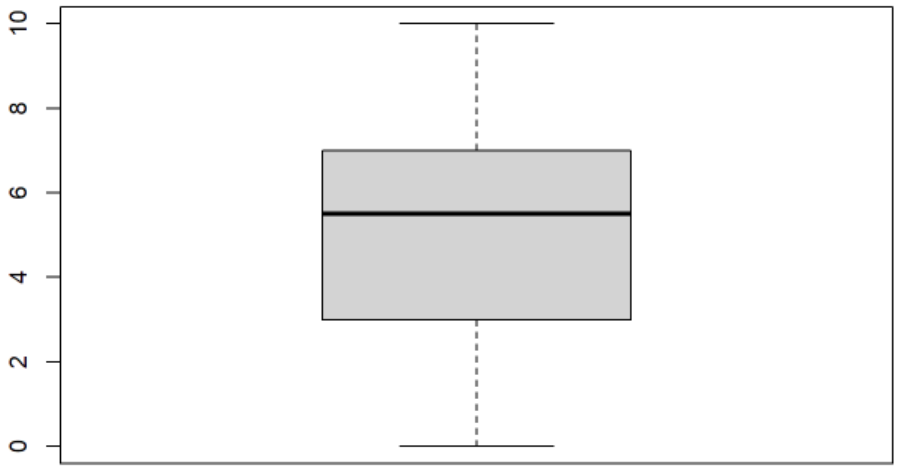
The mean number of customer served in each 5-minute window on this given Tuesday is 5.425. The minimum and maximum customers served are 0 and 10 respectively. 75% of the 5-minute periods see 7 or less customers served. There are 120 observations which means there are 10 hours of sales present. The below histogram asserts that the most frequent number of customers served in a 5-minute window is 2.



The next plot describes the number of customers served in each of the 120 5-minute periods in order. The observations align with the owner's assertion that the busiest periods are between 08:00 – 10:00 (0 – 24) and 12:00 – 14:00 (48 - 72). The quietest time is also clearly after 16:00 (84 – 120) as stated.



The following boxplot visually represents the quartiles in the data. As stated previously, 75% of the periods see 7 or less customers served and 3 or less customers are served 25% of the time.



Question 2

2A

The maximum likelihood estimate function produces a probabilistic estimate of the parameters of a selected model. In this scenario, the owner has suggested that the data follows a Poisson distribution which is defined as follows:

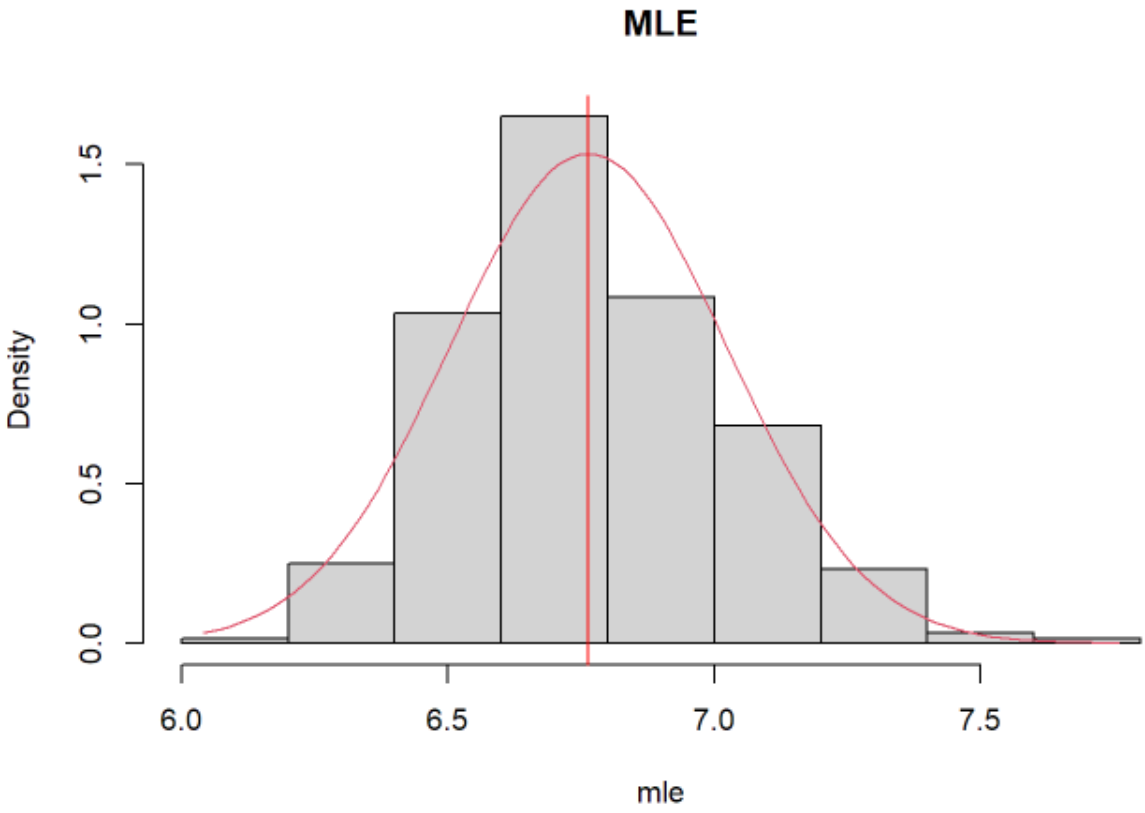
$$P(X = x) = \frac{\lambda^x}{x!} exp(-\lambda) \text{ denoted as } X \sim Po(\lambda)$$

Lambda (λ) is the sample mean and the function is performed on each data point (X). The output is a plotted distribution which will or will not match the Poisson shape. The maximum likelihood estimation function can be used to estimate the value for lambda (λ) to a confidence calculated probabilistically. The function is defined as follows:

$$\prod_{i=1}^{X} P(X_i = x_i)$$

$$\prod_{i=1}^{X} \left\{ \frac{\lambda^{x_i}}{x_i!} exp(-\lambda) \right\}$$

Using these formulas with the given data produces an estimated likelihood value of lambda (λ) between 6.5 and 7. This aligns with the calculated mean of 6.76 as shown on the following plot. This means that the observed sample mean is likely the most appropriate value to use for lambda (λ) for the given data.



2B

Should an extra staff member enable the coffee shop to serve 2 more customers per 5-minute period, the amount of time the new staff member will be busy can be estimated using the following.

The current maximum potential for serving customers in a given 5-minute period is 10. With the new staff member, this becomes 12. For this scenario, the mean of the data is considered as lambda (λ) and passed into a probability Poisson function. The probability output of this function is then multiplied by the length of the data given. This produces the following result.

$$0.02127501 * 300 = 6.38$$

This means that, within a 10-week window, the new member of staff is likely to be busy 6 times.

2C

If the aforementioned member of staff costs £20 per hour and the average profit per customer is £2.10, some simple calculations can be done to assess whether it would be worth paying for them to work for each 08:30 – 09:00 period in the week.

The maximum potential for serving customers for the new member of staff is 2 per 5-minute window. The average profit per customer is £2.10. The cost of the new member of staff is £20.

Consider the following:

- Maximum serving potential of the new member of staff per 5-minute period (q) = 2
- Average profit per customer (d) = 2.10
- Cost per hour for the new member of staff (c) = 20
- 5-minute intervals in one hour (t) = 12
- Probability of the new member of staff being busy (p) = 0.02127501

Maximum earning potential:

$$max = q * d * c = £50.40$$

Likely profit without considering hourly wage:

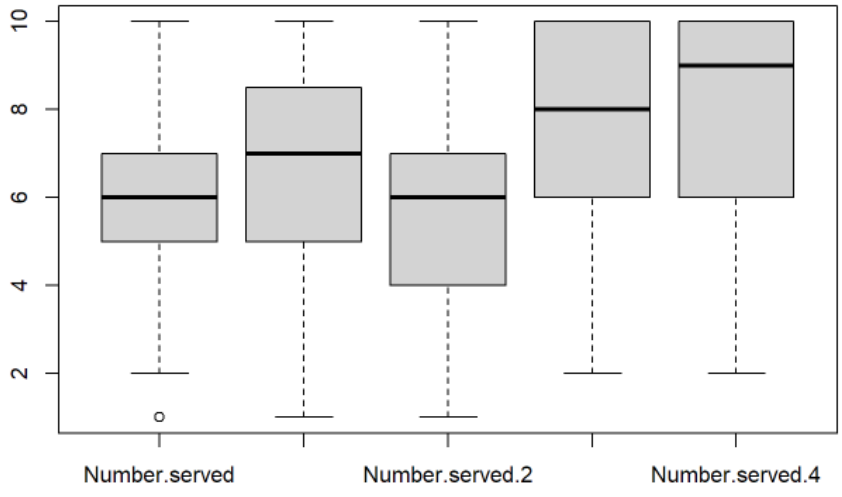
$$\max * p = £1.07$$

Since the likely profit to be made, based on the likelihood of the new member of staff being busy within a 10-week period, is less than £20 it would not be worth hiring them between 08:30 – 09:00 every day.

Question 3

3A

By looking the given data, assessing whether the busy period changes each early morning can be assessed with the following plots.

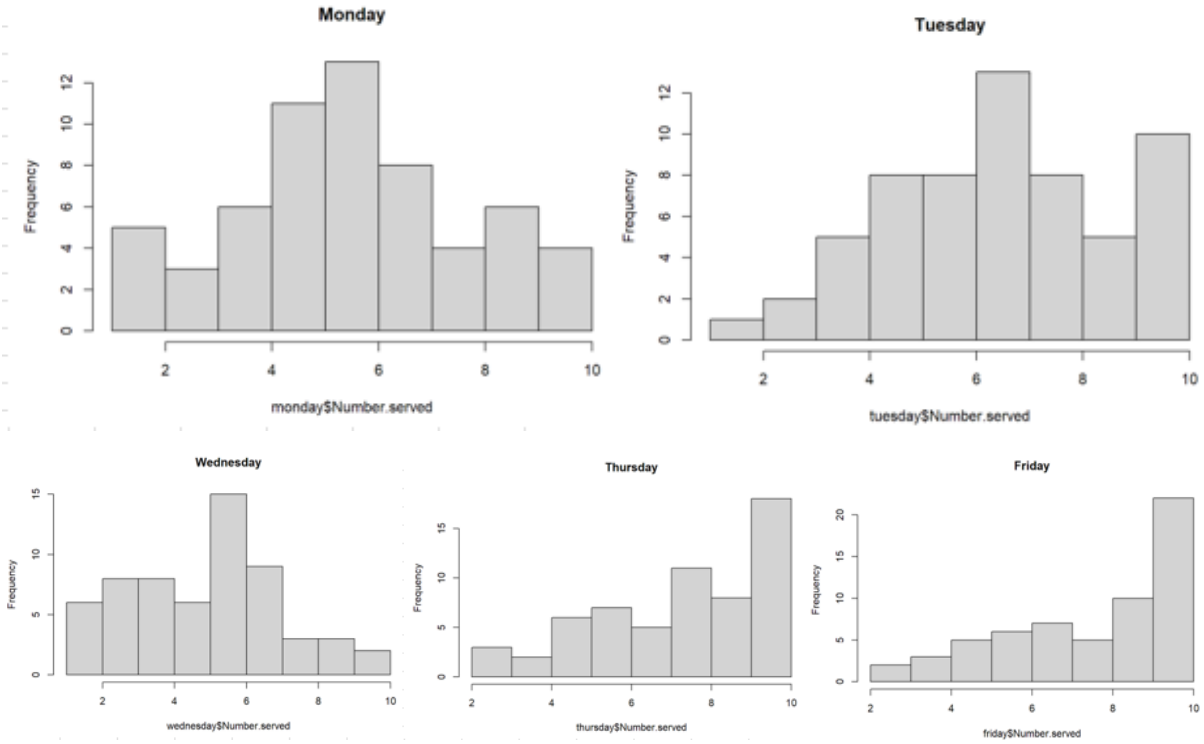


Each boxplot describes the number of customers served in each busy period between Monday to Friday in order. It is clear that the number of customers averages noticeably higher on Thursdays and Fridays.

The quietest day is most likely to be Wednesday. Monday has the least variance in the middle 50% of the data from the median and so the number of customers served can more reliably be assumed compared to, for example Thursday and Friday where the variance is greater. This could mean that it would be harder to assess the average number of staff required on those days if the upper third quartile didn't match the observed maximum.

3B

Observing the given data for each early morning period across 10 weeks, it can be inferred that Thursdays and Fridays conflict with the conclusions drawn previously regarding hiring an extra member of staff. Where previously it was determined that hiring an extra member of staff was not likely to generate profit, analysis of the data suggests that hiring a new member of staff for two out of five days may be appropriate. This is due to the highest frequency of customers equalling the current maximum limit for the staff, along with the third quartile matching the maximum of the data. It suggests it is likely that there's demand past the maximum for these days.



As seen above, Thursdays and Fridays, the number of customers being served equalling the maximum is far higher. Tuesday could also benefit from another member of staff, but it is not likely that Monday or Wednesday would. This would require further analysis of the data.

Question 4

4A

To assess the owner's assumption that the mean number of customers on Saturdays and Sundays are 250 and 230 respectively, the data for each day and be used with a one sample t-test to determine the range a 95% confidence for which the unknown population mean sits. It is assumed the owner is referencing this unknown population mean and not the sample mean.

```
##
##  One Sample t-test
##
## data:  saturday
## t = -1.377, df = 9, p-value = 0.2018
## alternative hypothesis: true mean is not equal to 250
## 95 percent confidence interval:
##  214.5868 258.6132
## sample estimates:
## mean of x
##     236.6

##
##  One Sample t-test
##
## data:  sunday
## t = -3.7831, df = 9, p-value = 0.004328
## alternative hypothesis: true mean is not equal to 230
## 95 percent confidence interval:
##  169.4371 214.7629
## sample estimates:
## mean of x
##     192.1
```

According to the results of the first test, the population mean for Saturday is 95% likely to sit within the range 214.6 – 258.6. Therefore, the null hypothesis is accepted as the population mean is equal to 250. The owner's assumption is highly likely to be correct. However, the results for Sunday indicate that the owner's assumption is highly likely to be incorrect. Since the p-value is below 0.05, the null hypothesis that the population mean is equal to 230 is rejected. The results suggest it is far more likely to sit within 169.4 – 214.8.

4B

To assess the difference in the data, and whether the owner's assumption that the mean difference between the two sets is at least 20, a paired t-test can be used to test this null hypothesis.

```
##
##  Paired t-test
##
## data:  saturday and sunday
## t = 4.3202, df = 9, p-value = 0.999
## alternative hypothesis: true mean difference is less than 20
## 95 percent confidence interval:
##      -Inf 54.89572
## sample estimates:
## mean difference
##            44.5
```

Since the p-value is more than 0.05, the above results suggest that the difference in means at least 20, and therefore the null hypothesis and the owner's assumption is accepted. The test is 95% confident that the difference in means is 44.5.

Question 5

5A

The owner has requested to know the dependence of income per day on the following covariates:

- Day of the week
- Mean number of staff working throughout the day
- Number of meal options
- Daily temperature
- Term time (yes or no)

The best model is found after applying a step function in R which assesses the affect of each covariate on the dependent variable using differences in Akaike Information Criterion (AIC) score. The function cycles through each covariate and determines its individual effect on the dependent variable, which in this case is income per day. The AIC function is defined as the following.

$$2p - 2\log\,(L)$$

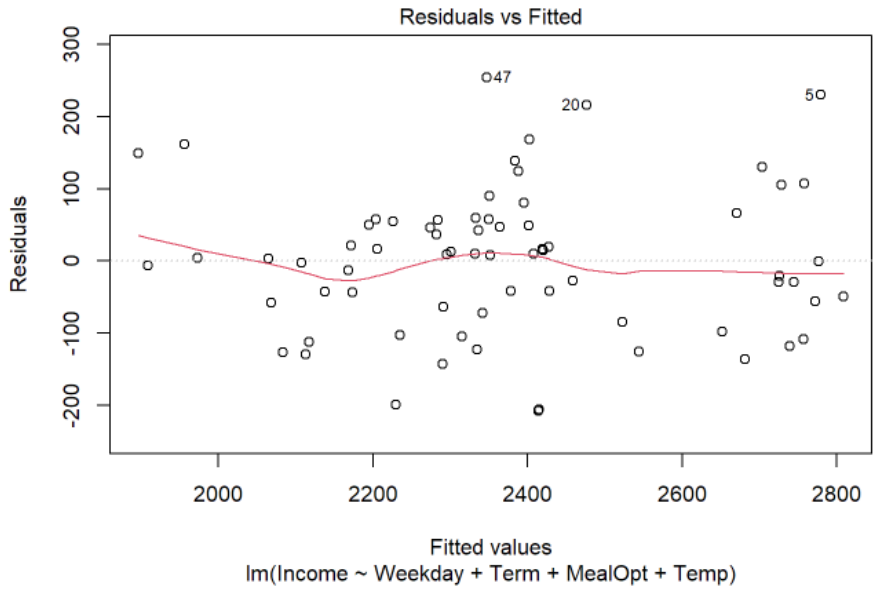
p holds the number of parameters in the model and L is the maximum value of the likelihood function of the model.

```
## Start:  AIC=666.69
## Income ~ Weekday + NoStaff + Term + MealOpt + Temp
##
##           Df Sum of Sq     RSS    AIC
## - NoStaff  1       267  699929 664.72
## <none>                  699662 666.69
## - MealOpt  1     39243  738905 668.51
## - Weekday  6    365012 1064673 684.08
## - Temp     1    328013 1027675 691.60
## - Term     1    915747 1615409 723.26
##
## Step:  AIC=664.72
## Income ~ Weekday + Term + MealOpt + Temp
##
##           Df Sum of Sq     RSS    AIC
## <none>                  699929 664.72
## - MealOpt  1     39320  739248 666.54
## - Temp     1    337022 1036951 690.23
## - Weekday  6    765199 1465127 704.43
## - Term     1    916998 1616927 721.33
```

The above results define show that removing 'NoStaff' produces the lowest AIC. Removing any of the other covariates increases the AIC value and is therefore not useful for increasing the accuracy of the model. It is seen within the results that 'Term' and 'Temp' have the largest impact on the AIC. Keeping these in the model appears to be important for predicting income.

5B

The model fits quite well since the observed the residuals following the zero line closely. The line is mostly narrowly sitting in the negatives meaning the model predicts a lot of values slightly higher than they should be.



Data Independence: Using the Durbin-Watson test, an observed value of rounded 0 for autocorrelation and 2 Durbin-Watson statistic indicates there is no correlation in the residuals. Therefore, the data is independent.

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.03435636      2.017775   0.962
##  Alternative hypothesis: rho != 0
```

Data Linearity: Since there is no pattern in the previous residual plot, there is no linear relationship between predictors and the outcome variable.

5C

To estimate income for a given day using the following covariates, the predict model in R applies to trained model to the unseen data.

- Daily temperature: 19.2°C
- Average members of staff: 5.7
- Meal options: 4
- Day: Thursday

The model estimate is given as: £2601.91

Conclusion

To conclude this report, it is worth noting that the data suggests there may be a need to hire a new member of staff for two days per week and it would be worth some further analysis to determine the exact potential benefits to this. The majority of the owner's assumptions were correct but it would be interesting to analyse data for the full year. It is worth noting the shortcomings of the data and that there is a limit to the predictions made on very small samples. If all data spanned the entire previous year the results may look very different since the report relies on probabilistic assumptions about unseen data. More data would also be beneficial for improving the prediction quality of a machine learning model. Perhaps the owner could provide further measured data to improve this. In terms of weaknesses of the report itself, a deeper analysis of the residuals and output plots form the best model could prove beneficial given more time. Although, the first most useful improvement would still be increasing the data fed into the model.