Statistics Group Assignment

Daniel Alkurdi Sammi Guan Stephanie Yang

Veda Rena William Huynh

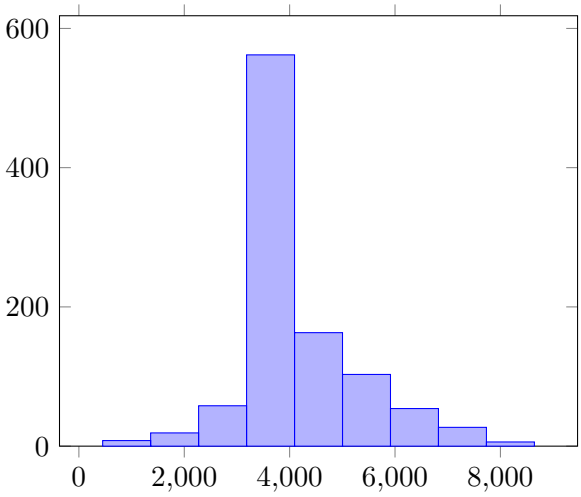
Business Statistics 26134

Contents

1 Everstain Investment Group

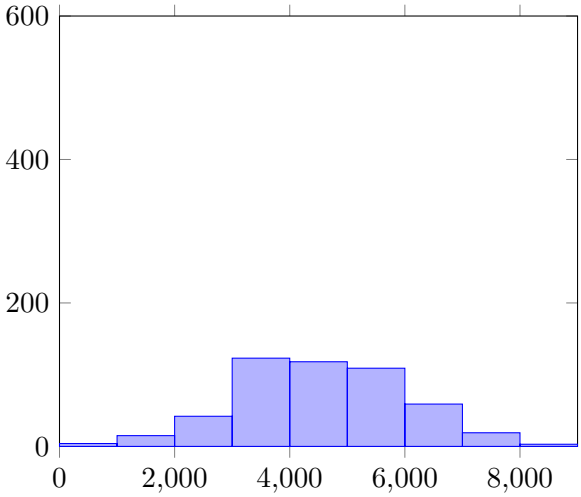
1.1 Frequency Distribution of Monthly Payments for all Employees

bin	Frequency
(0, 1000]	4
(1000, 2000]	15
(2000, 3000]	42
(3000, 4000]	490
(4000, 5000]	258
(5000, 6000]	109
(6000, 7000]	59
(7000, 8000]	19
(8000, 9000]	3



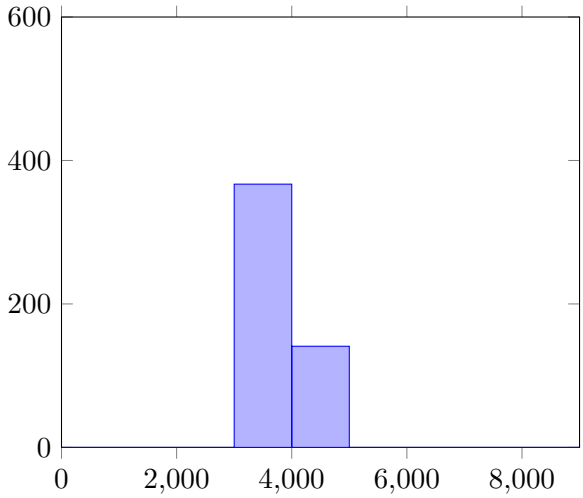
1.2 Frequency Distribution of Monthly Payment of Males

bin	Frequency
(0, 1000]	4
(1000, 2000]	15
(2000, 3000]	42
(3000, 4000]	123
(4000, 5000]	118
(5000, 6000]	109
(6000, 7000]	59
(7000, 8000]	18
(8000, 9000]	3



1.3 Frequency Distribution of Monthly Payment of Females

bin	Frequency
(0, 1000]	0
(1000, 2000]	0
(2000, 3000]	0
(3000, 4000]	367
(4000, 5000]	140
(5000, 6000]	0
(6000, 7000]	0
(7000, 8000]	0
(8000, 9000]	0



1.4 Descriptive Statistics Summary

The population is defined as the total number of employees in the company, with the subsets of the population (Gender: Male and Female) being samples A and B respectively.

Company ($) (N=1000)	
Mean μ	4244.59
Median $\widetilde{x}$	3975.79
Variance σ^2	1135892.69
Standard Deviation σ	1065.78

Male ($) (n=492)		**Female ($) (n=508)**	
Mean $\bar{x}_A$	4565.70	Mean $\bar{x}_B$	3933.59
Variance s_A^2	2089960	Variance s_B^2	19440
Standard Deviation s_A	1445.67	Standard Deviation s_B	139.43

1.5 Gender Pay Gap Calculation

The Australian Workplace Gender Equality Agency (WGEA) provides information on calculating the gender pay gap.[1]

Gender Pay Gap

$$GPG = 100 \times \frac{\bar{x}_A - \bar{x}_B}{\bar{x}_A} \tag{1}$$

$$GPG = 100 \times \frac{4565.70 - 3933.59}{4565.70} \tag{2}$$

$$GPG = 13.84\% \tag{3}$$

The Gender Pay Gap (GPG) is expressed as a percentage of male earnings. The above results show a 13.84% difference in the monthly salary of males and females. This GPG calculation is non-adjusted, meaning specific job functions of employees are not considered.

1.6 Summary of findings

There is a difference in monthly wages between males and females within Everstain Investment Group. All females within the company earn between $3000 and $5000. While male salaries are more normally distributed. Noting there is an approximately 14% difference in male and female salaries, it is important to note that job functions are not considered in these results. We advise Everstain Investment Group to collect more data in order to compare gender wages within job categories. Two possibilities exist with job functions considered:

1. Females are paid less than males
2. Males tend to have more job functions within the company

Understanding which of the two scenarios is occurring may help Everstain Investment Group to improve reducing their gender pay gap.

[1] Cassells, R., Duncan, A.S. and Ong, R., 2017. Gender equity insights 2017: Inside Australia's gender pay gap (No. GE02). Bankwest Curtin Economics Centre (BCEC), Curtin Business School.

2 Warhammer

2.1 Contingency Tables

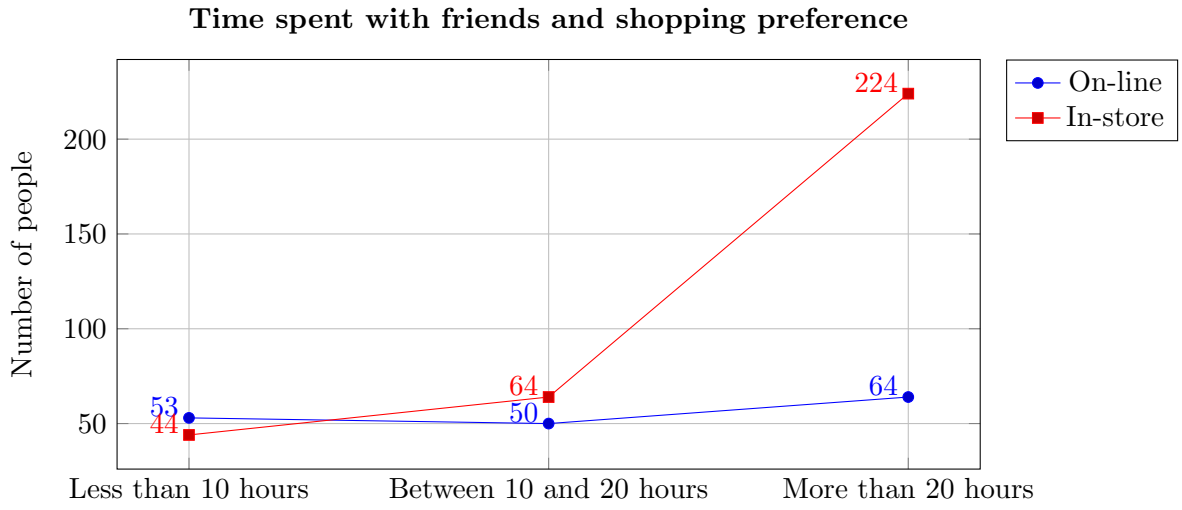
2.1.1 Counts

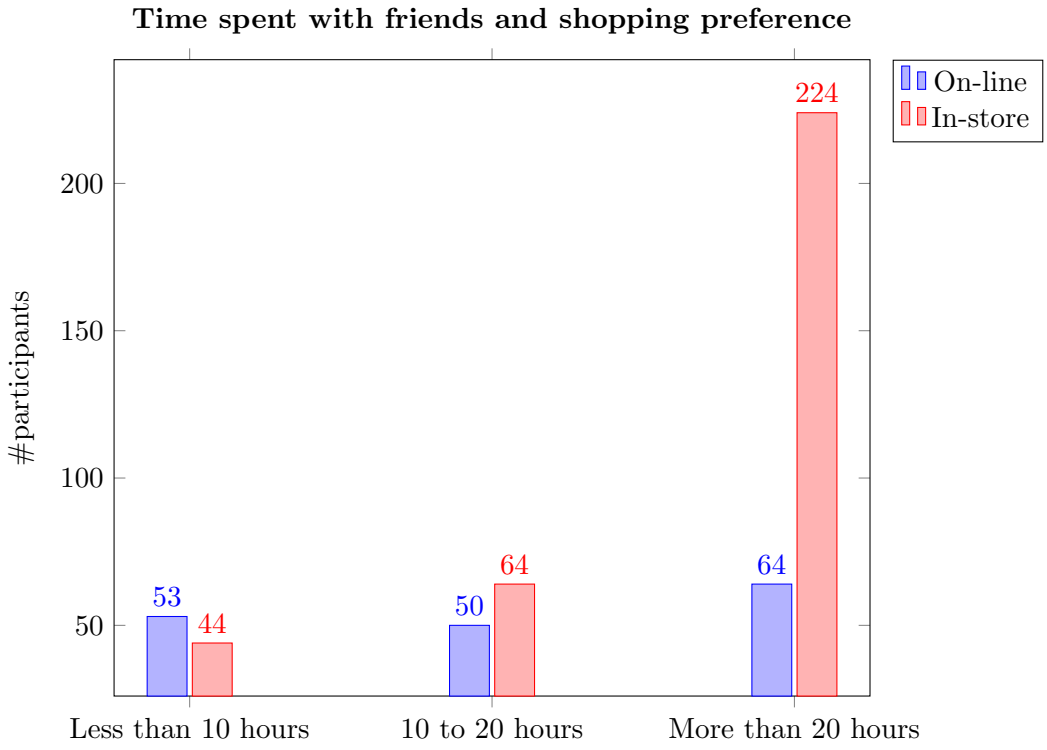
Contingency Table: Count				
	Time <10 Hours	10 Hours <Time <20 Hours	Time >20 Hours	**Total**
Online Shopping	53	50	64	167
In-Store Shopping	44	64	224	333
Total	97	115	288	500

2.1.2 Relative Frequency

Contingency Table: Relative Frequency				
	Time <10 Hours	10 Hours <Time <20 Hours	Time >20 Hous	**Total**
Online Shopping	0.106	0.100	0.128	0.334
In-Store Shopping	0.088	0.128	0.448	0.664
Total	0.194	0.228	0.576	1

2.2 Graphs





2.3 Is time spent with friends and shopping behaviour independent?

The procured results suggest that as people spend more time with their friends they tend to spend more time shopping In-store. This correlates with the assumption that people enjoy participating in activities with their friends. Intuitively, the results suggest that time spent with friends and their shopping behaviour may not be independent. Further statistical analysis (correlation tests) may provide Warhammer with a better suggestion as to whether these two variables are independent.

2.4 Summary

From the analysis we have provided, we believe there is a positive relationship between the time that people spend with their friends and the propensity to;

1. Partake in shopping activity
2. Shop in-store rather than on-line

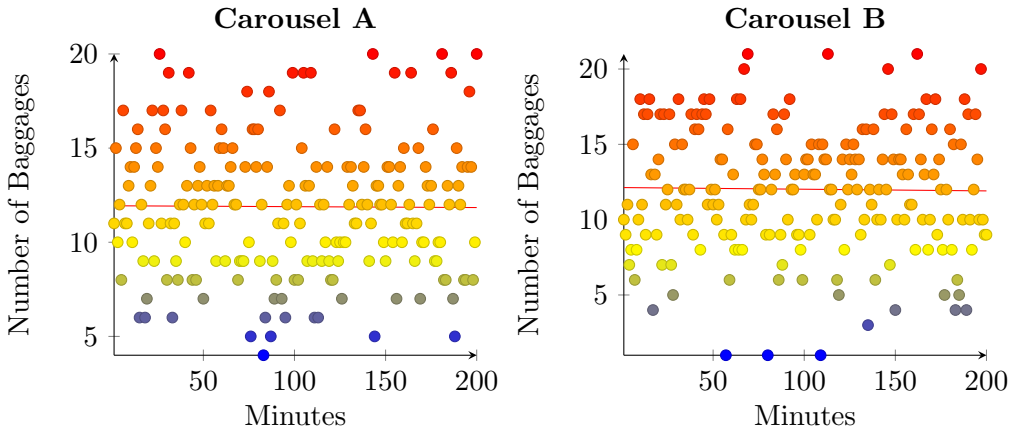
As such, we suggest to Warhammer to consider placing more weight towards off-line marketing such as in-store advertisements. However, it is important to note that there is a proportion of customers who do undertake their shopping on-line. Furthermore, with the virtually global use of the internet, on-line advertising is an important factor in the marketing of the brand. A study has found that on-line advertising has shown to have a positive impact on buyers' purchase decision. However, it notes that heterogeneous advertising has the most beneficial impact on a firm's economic activity.[2]

[2]Bayer, E., Srinivasan, S., Riedl, E.J. and Skiera, B., 2020. The impact of online display advertising and paid search advertising relative to offline advertising on firm performance and firm value. International Journal of Research in Marketing.

3 Istanbul International Airport

3.1 Data

3.1.1 Visualisation



3.1.2 Descriptive Statistics Summary

Carousel A (n=**200**)	
Mean $\bar{x}_a$	11.89
Median	12
Mode	12
Standard Deviation	3.57
Sample Variance	12.74
Kurtosis	-0.411
Skewness	0.246
Range	16

Carousel B (n=**200**)	
Mean $\bar{x}_b$	12.02
Median	12
Mode	10
Standard Deviation	4.10
Sample Variance	16.82
Kurtosis	-0.217
Skewness	-0.190
Range	20

From the above results, we can see that the mean of Carousel B is marginally higher than that of Carousel A. Furthermore the number of bags that go through Carousel B have a higher variance than that of Carousel A.

3.2 Computations

3.2.1 Average time to wait to see the next bag

If we assume that the time which elapses between two events follows the exponential distribution with a mean of μ units of time. Additionally assuming that these times are independent, meaning that the time between events is not affected by the times between previous events. If these assumptions hold, then the number of events per unit time follows a Poisson distribution with mean $\lambda = 1/\mu$.

$$E(X) = \frac{1}{\lambda} \tag{4}$$

$$E(X) = \frac{1}{11.89} = 0.0833 \tag{5}$$

$$0.0833 \times 60 = 5 \tag{6}$$

These calculations show that it will take on average, 5 seconds for the next bag to arrive in each Carousel.

3.2.2 Probability Calculations

Probability that there are less than 7 bags in one minute.
Let x be the number of bags collected per minute. With λ being 12

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{7}$$

$$P(X \leq 6; 12) = \frac{12^0 e^{-12}}{0!} + \frac{12^1 e^{-12}}{1!} + ... + \frac{12^6 e^{-12}}{6!} \tag{8}$$

$$P(X \leq 6; 12) = 0.049 \tag{9}$$

Probability that there are less than 10 bags in one minute.
Let x be the number of bags collected per minute. With λ being 12

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{10}$$

$$P(X \leq 9; 12) = \frac{12^0 e^{-12}}{0!} + \frac{12^1 e^{-12}}{1!} + ... + \frac{12^9 e^{-12}}{9!} \tag{11}$$

$$P(X \leq 9; 12) = 0.252 \tag{12}$$

The results of these probability calculations make sense in that the approximate average number of bags to arrive in one minute is 12. As such for less than 7 bags to arrive in one minute, a low probability result correlates with the mean. Similarly, an increase in the probability, when calculating the probability of less than 10 bags arriving in 1 minute, also correlates with the means of the data.

3.3 Assumptions within Statistic Calculations

The above calculations for the data provided assume that the data is Poisson distributed. Based on the descriptive statistics calculated in 3.1.2 there is no over-dispersion within the mean and standard deviatino of the data, however the variance does not equal the mean which is a property of Poisson distributions.

$$\lambda_a \approx 12 \tag{13}$$

$$\lambda_b \approx 12 \tag{14}$$

$$Var\ (X_a) \neq \lambda_a \tag{15}$$

$$Var\ (X_b) \neq \lambda_b \tag{16}$$

Therefore a χ^2 goodness of fit test is required to ascertain if the data follows a Poisson distribution.

4 MyOriental

4.1 Approaches to summarise the dataset

Statistical inference is important to analyse data and refers to the process of drawing conclusions from a model estimation.

One approach of statistical inference is interval estimation for a single population. We can draw conclusions on population parameters based on a sample. Although a sample is often a good representation of the population, it can however have discrepancies. Due to sampling error, we cannot say with absolute certainty that if a=b, as this requires us to gather data on the entire population. If we were to keep re-sampling, all sample statistics should form an interval, otherwise known as a confidence interval, where the actual population parameters would fall, allowing room for error. This approach is appropriate as confidence intervals consider the sample size and the possible variations in the population to determine an approximation of the range where the real answer lies.

Another approach is hypothesis testing for a single population or for two populations. For a single population, the basis of a hypothesis test is to decide if a sample is typical or atypical compared to a population. We can use hypothesis testing to quantify the distance between a sample statistic and the hypothesised population parameters, however the test results vary depending on the accuracy of the sample. Population parameters can be assumed by using a sample to either reject the null and accept the alternative, or not reject the null and maintain it. This approach is appropriate as it evaluates two statements about a population to determine which statement is supported by the sample data.

For comparing two populations, hypothesis testing gives an assessment of statistical significance, but rather than looking at the estimate of the difference, we use the difference to determine whether it may be important or not. This approach is appropriate as comparing two population means is very common in studies, and also evaluates statements about the populations.

4.2 Why a sample mean can serve as an estimate for the mean daily revenue of that city but the accuracy of the estimate is not the same across cities

Although a sample mean can serve as an estimate for the mean daily revenue, sampling error can be expected, as the sample statistic is only a subset of

the population. Thus the accuracy of a sample statistic is dependent on the variability of sampling error or sampling distribution. The larger the standard error, the more variation the sampling distribution has and the less accurate the sample becomes as a representation for the population. Therefore, the accuracy of the estimate may vary between cities.

4.3 Parametric Analysis of Variance (ANOVA)

The ANOVA test will allow us to test if the means of more than two samples (multiple cities) are equal.

H_0: Sample means are equal
H_A: Sample means are not equal

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Darwin	441	482426.2	1093.937	41883.78
Perth	186	200754.9	1079.327	165565
Melbourne	352	844438.4	2398.973	656694.8
Sydney	417	1413643	3390.032	607248

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1.37E+09	3	4.55E+08	1190.569	0	2.611295
Within Groups	5.32E+08	1392	382308.5			
Total	1.9E+09	1395				

As $F > F - crit$ we reject the null hypothesis H_0 and accept H_A, the sample means of the daily revenue of the 4 cities are not equal.

As for the sample means for Sydney and Melbourne:

H_0: Sample means are equal
H_A: Sample means are not equal

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	352	844438.4	2398.973	656694.8
Column 2	417	1413643	3390.032	607248

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1.87E+08	1	1.87E+08	297.6435	1.32E-56	3.853611
Within Groups	4.83E+08	767	629876.2			
Total	6.71E+08	768				
Total	1.9E+09	1395				

As $F > F - crit$ we reject the null hypothesis H_0 and accept H_A, the sample means of the daily revenue of the 2 cities are not equal.

From these two ANOVA tests we can see that the means across the cities are not equal. This means that the average daily revenue of MyOriental is not equal in each of the different cities.