1014SCG - Statistics

Himath Ratnayake

s5209861

himath.ratnayake@griffithuni.edu.au

Signature: 

Part 1

When comparing different modes of transports, a general stereotype is usually thought of, that public transport such as busses or trams are of a lower cost than private or other methods, often due to price of fuel and maintenance for a car being a greater burden on a University student when compared to using public transport modes, which are often also offered at a discounted concession rates for students.

Research Question: *Thus, is there a dependency between the cost of transport and frequency in usage of certain transport modes?*

Table 1: Variables Used In Test

Variable Name	Type	Range
Mode	Categorical (Nominal)	*3 Sub-categories:* Public, Private, Other
Cost	Categorical (Ordinal)	*3 Sub-categories:* Low, Medium, High
Frequency of Transport Usage	Numerical (Discrete)	-

Statistical Hypotheses for a *Chi-Squared Test of Independence*

- **Null Hypothesis (H_0):** There is no dependency between cost and the usage frequency of transport modes
- **Alternate Hypothesis (H_A):** There is an unspecified dependency between cost and the usage frequency of transport modes

Part 2

The chi squared test of independence analyses the frequency of two categorical variables to determine whether there is a statistically significant association between them – in this report, mode of transport, and weekly cost on transport services are explored.
The cost variable which was originally numerical was split into three levels named “low”, “medium”, and “high” to transform it into a categorical variable using the ‘cut’ command in R software (see Appendix)

The rationale for the boundaries between low, medium, and high, was based on the range of cost values given in the data set. The minimum value for cost was $35.85 and went up to a $87.44 maximum within the given sample, giving a range of 51.59. This range was divided by 3 to give the following boundaries exhibited in the code (Table 2).

Price Category	Cost Range ($)
Low	35.85 – 53.04
Medium	53.05 – 70.24
High	70.25-87.45

Table 2: Price boundaries chosen for cost

With two categorical variables in cost and transport mode, as well a numerical variable of transport mode frequency (the number of people that use that mode of transport for the given price range), a contingency table can be created to summarise to the data required for the test of independence.

Table 3: Summary Table (Values Found Through R)				
OBSERVED	**Frequency of Transport Mode**			**Total**
Cost	**Public**	**Private**	**Other**	
Low	18	39	26	**83**
Medium	37	26	45	**108**
High	17	7	1	**25**
Total	**72**	**72**	**72**	**216**

It can be seen there is an equal number of people surveyed for each mode of transport (72), but the frequency of usage differs between cost levels. Visually this is represented in Fig.1.

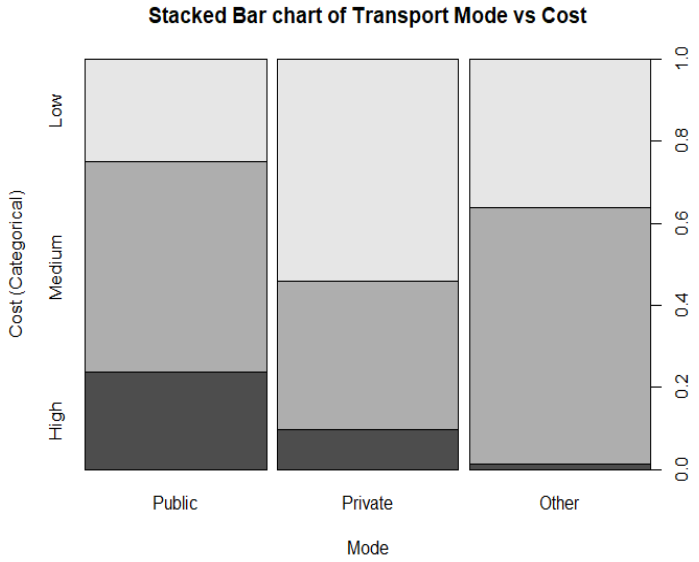

Figure 1: Stacked bar chart of different modes of transport vs cost

A percent-stacked bar plot was chosen to display how common each price range was for each mode of transport. Overall, the cost level of “medium” is most common, with half of all responders (108/216) answering within that price range for the survey. Within the medium price range, the “other” option was most popular. “Low” is the next most common price range, with 83/216 responses being in this range. Private was the most popular mode of transport in the low range, while public was the least popular. Responses in the “high” range were much less common when compared to the low and medium (Fig.1), with only 25/216 responders falling into this criterion. For high cost, public transport was the most popular option, while the “Other” transport method had only one responder who fit into the "Other, High" criterion. This value could be classed as an outlier, as it significantly deviates from the expected value of 8.33, potentially changing the chi-squared test statistic value in Part 3 by a statistically significant amount.

A boxplot histogram was plotted using the “PackHV” library in R, which illustrates cost per week (x axis) against the density of each cost value (y axis). Fig.2 shows the distribution of cost values – in this data set, there is central tendency around the mean of $57.12; the most frequent values that appear are between the range of 55 and 60 dollars per week. This is also highlighted further by the density curve plotted, wherein the peak of the curve also represents the mean value of the data set.

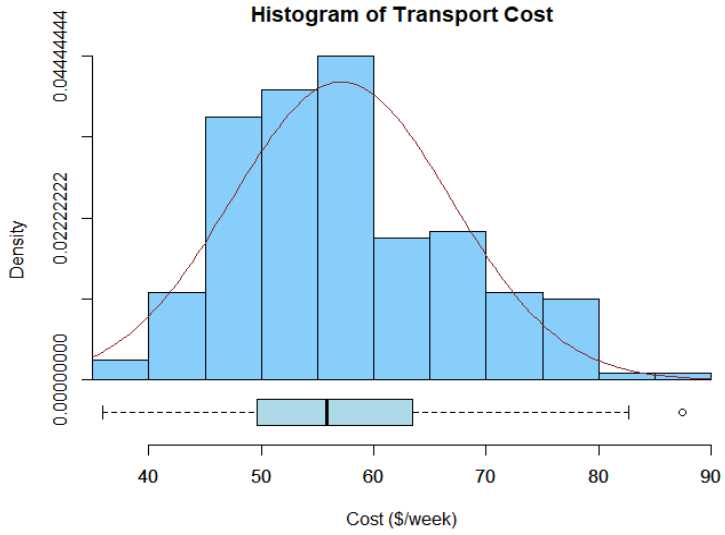

Figure 2: Histogram-Boxplot and density curve of numerical transport cost

The center and spread can also be shown by the boxplot. The line within the box represents the median cost value of $55.79 – another measure of central tendency. As the mean is greater than the median, the distribution is positively skewed. The skewness is almost moderate with a value of 0.55 (see appendix 4). The interquartile range of the data is between $49.71 and $63.30, meaning 50% of the data values lie within this range. The whiskers that extend from the box account for the bottom and top 25% of data, and are of noticeably greater range than the interquartile range, which should be expected, as the density curve also shows that the region of highest density is within the interquartile range.

Within the context of the scenario, this means that it’s expected that most individuals pay between $50-60 dollars a week on their transport needs, with a lower to medium cost per week being more common than a high one, as seen by the left-skewed distribution and central tendency of the histogram. Furthermore, there is an equal amount of responses for each transport mode, with low cost being most common in private methods while medium costs are more prevalent in public methods. “Other” modes were also the least likely to have high transport costs.

Note: see appendix for R code used for R outputs

Part 3

Table 4	Description	Notes
Significance Level	0.05	The standard significance level for an independence hypothesis test is 0.05
Degrees of Freedom	4	Contingency table has 3 columns and 3 rows. Therefore, (3-1)*(3-1) = 4
Decision Rule	If the TSV>9.488 reject the null hypothesis ($p < 0.05$)	The decision rule value was evaluated based upon the chi-squared table given a significance level of 0.05 and 4 degrees of freedom.
Test Statistic (Chi-Squared Formula)	$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$	**Where:** • O = Observed Value • E = Expected Value

Expected Values: The expected values for this distribution were calculated from the original contingency table above, with the equation:

$$\frac{Row\ Total\ \times Column\ Total}{Table\ Total}$$

Table 5			
EXPECTED	**Frequency of Transport Mode**		
Cost	**Public**	**Private**	**Other**
Low	$\frac{72 \times 83}{216} = 27.67$	27.67	27.67
Medium	$\frac{72 \times 108}{216} = 36$	36	36
High	$\frac{72 \times 25}{216} = 8.33$	8.33	8.33

Test Statistic Value:

$$X^2 = \frac{(18-27.67)^2}{27.67} + \frac{(39-27.67)^2}{27.67} + \frac{(26-48.67)^2}{27.67} + \frac{(37-36)^2}{36} + \frac{(26-36)^2}{36} + \frac{(45-36)^2}{36} + \frac{(17-8.33)^2}{8.33} + \frac{(7-8.33)^2}{8.33} + \frac{(1-8.33)^2}{8.33} = \mathbf{28.86}$$

Once evaluated, a final TSV of 28.86 is found. This value was cross-checked with R using the code seen at right and found to be correct.

```
> chisq.test(CatPrice, mode) # Run a chi squared test

        Pearson's Chi-squared test

data:  CatPrice and mode
X-squared = 28.856, df = 4, p-value = 8.362e-06
```

Conclusion:

As the final TSV of 28.86 is greater than the critical value of 9.488 at 5% significance, we reject the null hypothesis in favour of the alternate; that is to say, there is evidence of some unspecified dependency between the cost and the usage frequency of different transport mode types per week ($p < 0.05$), thus addressing the research question. These results are in line with official statistics presented by the Australian Trade and Investment commission; they also show that University students spend less than the average Australian on transport costs, such as by taking low to medium cost "Other" transport (like carpooling), or public transport like trains due to lower cost of living (Studyinaustralia.gov.au, 2020).

The results from this study could assist in further ventures analysing transport use in wider demographics, extending beyond students as a point of comparison to see if the trends between transport mode and cost continue to hold with a larger sample size/population, so that the commuting habits of Australians can be better understood.

References

Studyinaustralia.gov.au. 2020. *Education And Living Costs In Australia*. [online] Available at: <https://www.studyinaustralia.gov.au/english/live-in-australia/living-costs> [Accessed 26 August 2020].

Appendix – R Code

```
attach(TransportCopy)
plot(TransportCopy$mode, TransportCopy$CatPrice,
     main = "Stacked Bar chart of Transport Mode vs Categorical Cost",
     xlab = "Mode",
     ylab = "Cost (Categorical)")
```

Appendix 1: R code used to create percentage stacked bar chart of transport mode vs cost (Fig.1)

```
# Make numerical cost variable into a categorical one
CostLevels <- cut(Cost, breaks = c(35.84,53.04,70.24,87.45)) # Add category boundaries
levels(CostLevels) <- c("Low", "Medium", "High") # Name each level

# make a copy of data
TransportCopy <- Transport # Copy integers in Transport to TransportCopy
TransportCopy$CatPrice <- CostLevels # Create new column called CatPrice
#This will now have the price for transport as a categorical variable
```

Appendix 1: R code used to convert cost from a numerical to categorical variable (Table 2)

```
x = Cost
# Create A Histogram with box plot)
hist_boxplot(x, main = "Histogram of Transport Cost",
             xlab = "Cost ($/week)",
             freq = FALSE, # Frequency must be changed to density for a density curve
             col = "lightskyblue",
)
# Create a density curve above histogram with box plot
s = sd(x); m = mean(x); curve(dnorm(x, mean = m, sd = s),
                              col = "brown4",
                              add = TRUE)
```

Appendix 3: R code used to create the histogram and density plot (Fig.2)

```
Minimum             35.850000
Maximum             87.440000
1. Quartile         49.707500
3. Quartile         63.302500
Mean                57.118704
Median              55.785000
Sum              12337.640000
SE Mean              0.663460
LCL Mean            55.810985
UCL Mean            58.426422
Variance            95.078686
Stdev                9.750830
Skewness             0.549016
```

Appendix 4: Statistics summary of numerical transport cost