# Lab #7 – Statistical Testing with Multiple Groups

Most recently we’ve learned about Chi-Square testing for association and ANOVA, both of which allow for comparisons across multiple groups. This lab will involve using these new tools, along with some of the additional topics we’ve covered, such as log-transformations and outliers.

### Question #1

Fill in the blanks for the following sentence with either “categorical” or “quantitative”.

Chi-Square testing for association involves comparing a \_\_\_\_\_\_\_\_\_\_ outcome variable across multiple groups defined by a \_\_\_\_\_\_\_\_\_\_ explanatory variable. ANOVA involves evaluating a \_\_\_\_\_\_\_\_\_\_ outcome variable for multiple groups defined by a \_\_\_\_\_\_\_\_\_\_ explanatory variable.

## PART #1

In part 1 of this lab you will analyze a dataset recorded by a GPS app that a driver uses to track the daily commute to his workplace in the Toronto area. The data are available on p-web as “Commute Tracker”, the available variables include:

* **Date** of travel
* **Month** of the year
* **StartTime**: when getting into the car
* **DayOfWeek**: the day name
* **GoingTo**: direction of travel
* **Distance** travelled in kilometers
* **MaxSpeed**: fastest speed recorded (all trips are on the 407 highway for some portion)
* **AvgSpeed**: the average speed for the entire trip
* **AvgMovingSpeed**: the average speed recorded only while the car is moving
* **TotalTime**: duration of the entire trip, in minutes
* **MovingTime**: duration when the car was considered to be moving (i.e. not counting traffic delays, accidents, or time while the car is stationary)
* **Take407All**: is Yes if the 407 toll highway was taken for the entire trip. I try to avoid taking the 407, taking slower back routes to save costs. But some days I'm running late, or just lazy, and take it all the way.
* **Comments**: comments from the driver about that day’s travel

### Question #2

Construct an appropriate plot illustrating the distribution of the variable “TotalTime”, do you see evidence of skew or outliers? Include your plot and comments in your write-up

### Question #3

In class, we discussed various arguments for and against including outliers in a statistical analysis. There are two observations with commute durations (TotalTime) that are substantially longer than the rest of the data. Should these outliers be excluded when analyzing these data? Write 1-2 sentences explaining your decision and rationale.

### Question #4

Create a new variable “LogTime” by applying a log-transformation to the variable “TotalTime”. Construct a plot(s) comparing the distributions of “LogTime” and “TotalTime” using “GoingTo” as a by/grouping variable. Does the trip appear to take longer going to or from work based upon these data? Include your plot(s) and comments in your write-up.

### Question #5

Construct a 95% confidence interval to estimate how much longer/shorter the trip home takes. **Provide both relative and absolute assessments**. Also, provide a one-sentence summary of the interval tells you.

|  |  |  |
| --- | --- | --- |
| Comparison | 95% Confidence Interval | Interpretation |
| Absolute |  |  |
| Relative |  |  |

### Question #6

To determine which other factors are associated with the trip’s total time, fill out the table below. I’ve partially provided the first row as an example:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (to be related with “TotalTime”) | Statistical Approach | Null Hypothesis | P-value and conclusion |
| “GoingTo” | Two-sample t-test |  | p = 0.435, we conclude there is insufficient evidence to indicate a longer/shorter average commute time for either direct of the trip |
| “DayOfWeek” |  |  |  |
| “MaxSpeed” |  |  |  |
| “Month” |  |  |  |
| “Take407All” |  |  |  |

### Question #7

In Question #6 you performed several hypothesis tests. How many are statistically significant at the level? How many are statistically significant after applying the Bonferonni Correction? Why might you want to apply the Bonferonni Correction here? Provide the number of statistically significant findings and briefly explain.

### Question #8

In addition to relationships with “TotalTime”, the driver is also interested in which variables are associated with whether they use the 407 toll-way for the entire trip. What type of variable is “Take407All”? (be specific) Are you concerned with outliers or skew when analyzing this variable? Can you apply a log transformation to “Take407All”? State the variable’s type, and briefly explain your answer to each of these questions.

### Question #9

To determine which other factors are associated with taking toll highway 407, fill out the table below. I’ve partially provided the first row as an example:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (to be related with “Take407All”) | Statistical Approach | Null Hypothesis | P-value and conclusion |
| “GoingTo” | Difference in proportions test (two-sample z-test) |  |  |
| “DayOfWeek” |  |  |  |
| “MaxSpeed” |  |  |  |
| “Month” |  |  |  |

### Question #10-A

For this question I’d like you devise a way to evaluate how “StartTime” is related to “TotalTime”, “LogTime” (created in Question #4), and “Take407All”. To do this, you should consider a few key details, including the direction of the trip, and whether it makes sense to use a quantitative start time variable or derive a new categorical version of “StartTime”.

For Question #10-part A, describe how you plan to accommodate direction of the trip in 1-2 sentences.

### Question #10-B

For Question #10-part B, describe what you plan to use as your new start time variable. For example, you might consider a quantitative variable: minutes past the earliest start time, or a categorical variable which divides the start time into categories.

### Question #10-C

For Question #10-part C, fill out the table below.

NOTE: DEPENDING ON YOUR ANSWER TO #10-A YOU LIKELY NEED TO MODIFY THIS TABLE

|  |  |  |  |
| --- | --- | --- | --- |
| Variable (to be related with your new start time variable) | Statistical Approach | Null Hypothesis | P-value and conclusion |
| “LogTime” |  |  |  |
| “TotalTime” |  |  |  |
| “Take407All” |  |  |  |

### Question #10-D

Is the commentary provided in the description of “Take407All” in the data dictionary substantiated by your findings in Question #10-C? Briefly explain why or why not.