MATH 4773 Laboratory 1

Categorical Data Analysis using R

This is the first of two labs on Categorical response variables. The lab will function to help the student get up to speed with R. You will learn how to manipulate data in the form of one categorical response variable and write your own functions to perform more elaborate analysis. The data will thus be frequencies divided into levels of the category.

# Objectives

In this lab you will learn how to:

* Create tables from data
* Create appropriate investigative plots
* Create and interpret hypothesis tests
  + P values
  + Confidence Intervals
  + Test statistic
  + Distributional assumptions
* Create statistics by hand

### Tasks

We will use Rstudio for all of the course.

Use RMD and knit the final answers into an html file (this will be graded).

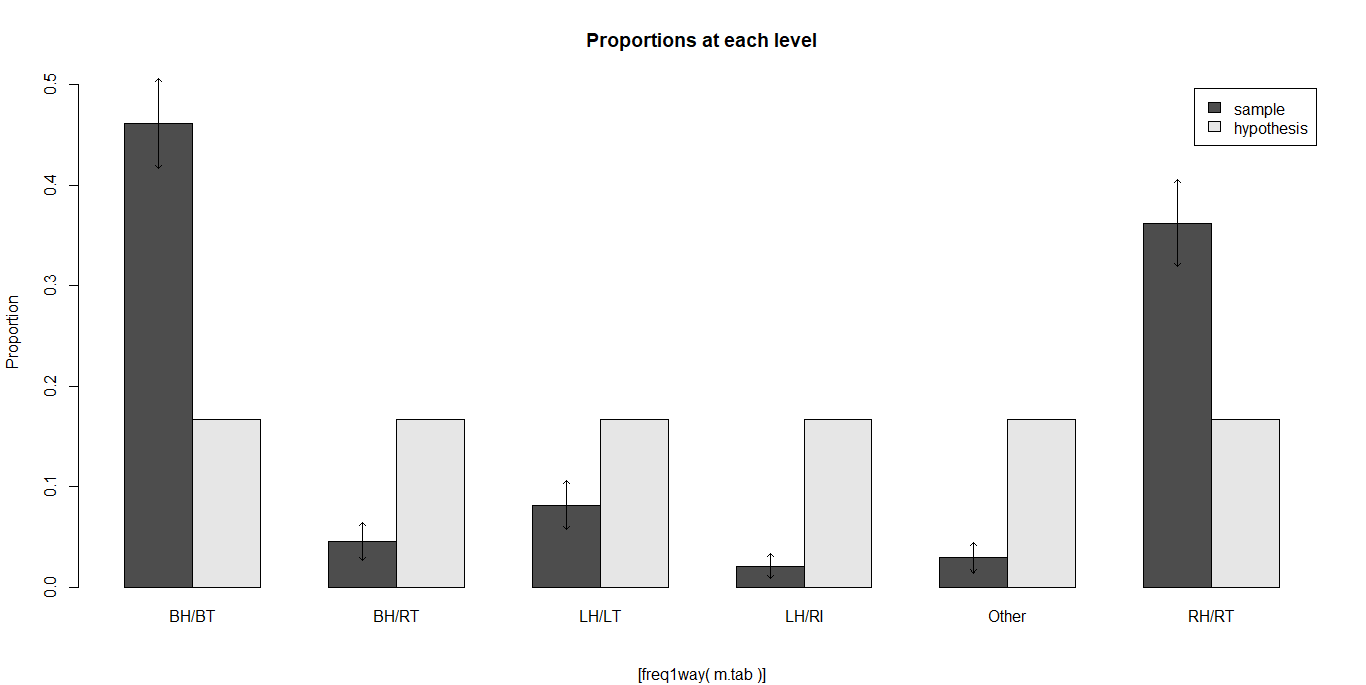
All R functions used will be placed in r chunks inside RMD. Formulae will be put into Latex. Save both rmd and html files and later place both in the dropbox when completed.

* Task 1
  + Download from CANVAS the zipped data files, “Dataxls”
  + Unzip the contents into a directory on your desktop (call it LAB1)
  + Download the file “lab1.r”
  + Place this file with the others in LAB1.
  + Start Rstudio
  + Open “lab1.r” from within Rstudio (this is an exemplar file).
  + Using hash commenting and at the top of Lab 1 place the task number eg #Task 1
  + Go to the “session” menu within Rstudio and “set working directory” to where the source files are located.
  + Copy and paste the working directory by issuing the command getwd(): under #Task 1
* Task 2
  + Make sure you install the readxl package.
  + Locate the file MOBILE.xls
  + Use read\_excel() to read the data into R, this function will already be available within the script lab1.r which you have opened in Rstudio.
  + Copy and paste the first six lines of the data using “head()” (use “courier new” font):
* Task 3
  + Using the code provided or your own and make a table of frequencies. Place here!
  + Make a function called mydata(filename) that will take any file in the same format as “MOBILE.xls”
    - Read it into R
    - Create a one column file of categorical data (use rep() )
    - Make a data frame
    - Release the data frame as command line output.
  + Issue the following in R and place results in this document after the code!

fff=mydata(“MOBILE.xls”)

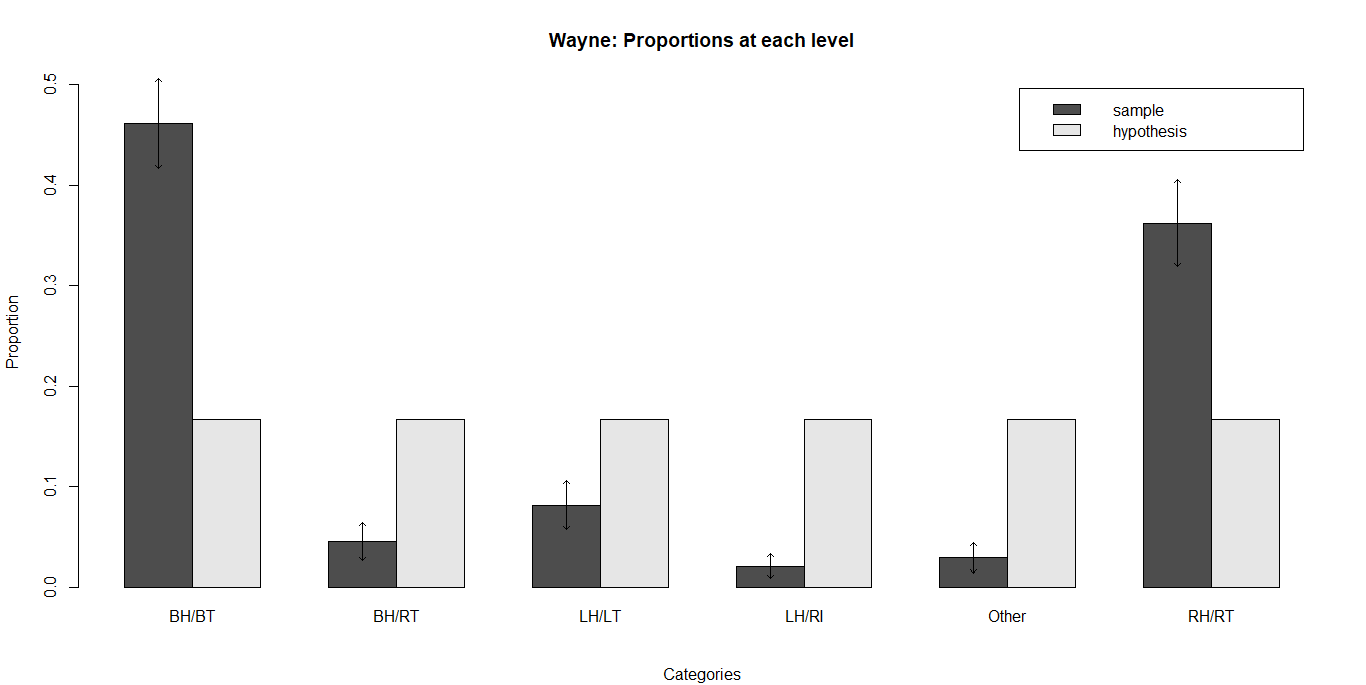
head(fff)

* + Explain the story behind the data by reading 9.11, page 451. Place here!
  + Use the table you created and make a barplot in R that has the following:
    - Title
    - Color filled bars
    - Labelled x and y axes
    - Bars with correct level names
    - Key for the colors
  + Place plot here!
* Task 4
  + Install and invoke the library “s20x”
  + We wish to test the NULL hypothesis that there is no difference in underlying text style categories.
  + What is the NULL hypothesis in terms of probabilities?
  + Using alpha = 0.10 conduct the test using the freq1way() function that is found in the “s20x” library. Paste the command line output here:
  + What is the P value for evidence against the NULL?
  + Interpret the P value.
  + Produce a plot of the chi square distribution that has the rejection region corresponding to on it. Place plot here!
* Task 5
  + If there is evidence against the NULL you need to examine which category or categories contribute most to the overall chi square statistic.
  + Create the chi square statistic for each category – use R as a calculator, Place here!
  + Which category gives most departure from the NULL?
  + Give interval estimates for the underlying proportions.
* Task 6
  + The following is the plot you should have made when using the freq1way()function.



* + - Alter the function freq1way()and call it myfreq1way()so that the plot will contain an x label “Categories” and the same title with your name on it eg. “Wayne: Proportions at each level”. Place the code snippet with the changes here:
    - Run your amended function and place the plot here.

Example:



* + - Please note that the above two questions do NOT presuppose that you understand all the code in the function.
* Task 7
  + What are the assumptions for the chi square test?
  + Does the above test satisfy the assumptions?