Hypothesis Tests Involving Categorical Vairables

Fang Yu | yufzhy@gmail.com

This repository hosts codes to perform and visualize statistic tests involving at least one categorical variable. Chi Square test is performed if it is between two categorical variables. Tukey HSD test is performed if it is between one categorical and one continuous variable. I did not include the test between two continuous variables as it is really easy to perform using all kinds of tests such as T test.

The running examples here are based a public dataset called cars93.

```
# Load the libraries ------------------------------------------------------
library(tidyverse)
library(magrittr)
library(ggalluvial)
library(gridExtra)
library(zoo)
library(scales)
#library(readxl)
library(MASS)
#library(xlsx)
library(modeest)
library(multcompView)
# setup working directory -------------------------------------------------
setwd("C:/Users/fang/Documents/FangRepo")
set.seed(99)
rm(list=ls())

# Functions ---------------------------------------------------

# find mode of a variable grouped by variable ID
findModes <- function(varID,var,varIDName,varModeName){
  a = table(varID,var)
  aColNames = colnames(a)
  #if there is empty string, do not use it
  if (aColNames[1] == "") { a = a[,2:ncol(a)]
  aColNames = aColNames[2:length(aColNames)] }
  colIdxMax = max.col(a)
  modes = aColNames[colIdxMax]
  b=cbind(rownames(a),modes)
  colnames(b) = c(varIDName,varModeName)
  return(b)
  #return(modes)
}

# I need to group the treatments that are not different from each other together.
```

```r
generate_label_df <- function(TUKEY, variable){

  # Extract labels and factor levels from Tukey post-hoc
  Tukey.levels <- TUKEY[[variable]][,4]
  Tukey.labels <- data.frame(multcompLetters(Tukey.levels)['Letters'])

  #I need to put the labels in the same order as in the boxplot :
  Tukey.labels$treatment=rownames(Tukey.labels)
  Tukey.labels=Tukey.labels[order(Tukey.labels$treatment) , ]
  return(Tukey.labels)
}

# Perform and visualize Chi square test and Tukey HSD test
test2CatVar <- function(data,xlabText,ylabText,xlabTickAngle=0,titleText,testType){
  #library(ggplot2)
  #library(gridExtra)
  if (testType=='ChiSquared'){
    FreqPlot <- ggplot(data, aes(x=x, fill=y, color=y, group=y)) +
      geom_histogram( position="dodge", stat = "count",alpha=0.5) +
      #geom_density(alpha=0.6) +
      theme(legend.position="top",legend.title = element_blank()) +
      theme(axis.text.x = element_text(angle = xlabTickAngle)) +
      ylab('Frequency') + xlab('')
    #FreqPlot
    #
    DensPlot <- ggplot(data, aes(x=x, fill=y, color=y, group=y)) +
      #geom_histogram( position="dodge", stat = "count",alpha=0.5) +
      geom_density(alpha=0.6) +
      theme(legend.position="none") +
      theme(axis.text.x = element_text(angle = xlabTickAngle)) +
      ylab('Density') + xlab('')
    #DensPlot
    #
    chisqTest = chisq.test(table(data$x,data$y))
    contributionChiSq = sort(rowSums((chisqTest$expected-chisqTest$observed)^2/chisqTest$expected))
    contributionDf = data.frame(contribution=contributionChiSq,category=names(contributionChiSq))
    contributionPlot<-ggplot(contributionDf, aes(x = reorder(category, -contribution), y = contribution
      geom_bar(stat="identity") +
      theme(axis.text.x = element_text(angle = xlabTickAngle)) +
      xlab(xlabText) + ylab('Contribution to Chi-Square') +
      labs(title =paste('Chi-Square Test: pValue = ',round(chisqTest$p.value,2)))
    #contributionPlot

    #
    grid.arrange(FreqPlot, DensPlot, contributionPlot, nrow = 3,  top = titleText)
  }
  #
  else if (testType=='TukeyHSD') {
    # What is the effect of the treatment x on y ?
    model=lm( data$y ~ data$x )
    ANOVA=aov(model)

    # Tukey test to study each pair of treatment :
```

```
    TUKEY <- TukeyHSD(x=ANOVA, 'data$x', conf.level=0.95)

    # Apply the function on my dataset
    LABELS <- generate_label_df(TUKEY , "data$x")
    color = grDevices::colors()[grep('gr(a|e)y', grDevices::colors(), invert = T)]


    # A panel of colors to draw each group with the same color :
    my_colors <- sample(color,length(LABELS))

    # Draw the basic boxplot
    a <- boxplot(data$y ~ data$x , ylim=c(min(data$y,na.rm = T) , 1.1*max(data$y,na.rm = T)), las = 2,
                 par(mar = c(18, 5, 4, 2)+ 0.1), col=my_colors[as.numeric(LABELS[,1])] ,
                 ylab=ylabText , main=titleText, xlab = xlabText)

    # I want to write the letter over each box. Over is how high I want to write it.
    over <- 0.1*max( a$stats[nrow(a$stats),] )

    #Add the labels
    text( c(1:length(unique(data$x))) , a$stats[nrow(a$stats),]+over , LABELS[,1]  , col=my_colors[as.nu
  }
  else {print('testType need to be either ChiSquared or TukeyHSD!')}

}
```

Examples using a public dataset cars93

This section shows examples of how to use functions in this repository to perform and visualizae the tests using a public dataset called cars93.

```
### Quick test using public dataset
data("Cars93")
```

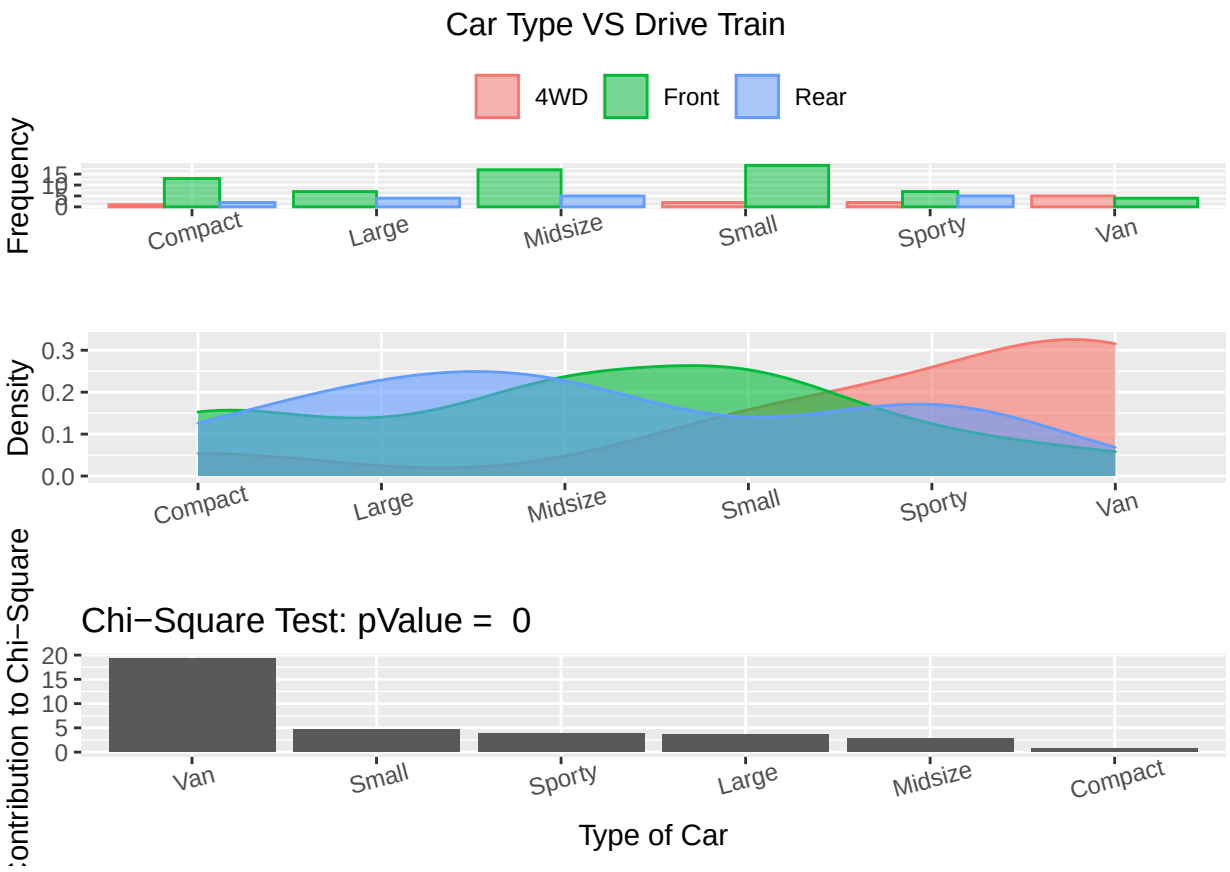
Chi Square test of two categorical variables

Do different types of cars have different drive trains? If so, whether the difference are significant or not?

```
# Chi Square test of two categorical variables
data1=data.frame(x=Cars93$Type,y=Cars93$DriveTrain)
test2CatVar(data = data1, xlabText='Type of Car', ylabText = 'Drive Train',xlabTickAngle = 15, titleTex
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```

```
## Warning in chisq.test(table(data$x, data$y)): Chi-squared approximation may be
## incorrect
```



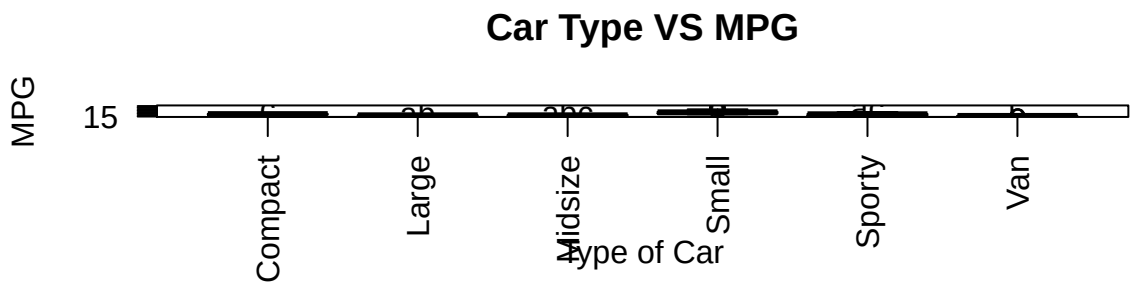
Tukey HSD test of one categorical variable and one continous variable.

Do different types of cars have different metrics of Mileage Per Hour (MPG) ? If so, whether the difference are significant or not?

```
# Tukey HSD test of one categorical variable and one continuous variable
data2=data.frame(x=Cars93$Type,y=Cars93$MPG.city)
test2CatVar(data = data2, xlabText='Type of Car', ylabText = 'MPG',xlabTickAngle = 15, titleText='Car Ty
```

```
## Warning in boxplot.default(split(mf[[response]], mf[-response], drop = drop, :
## NAs introduced by coercion
```

```
## Warning in text.default(c(1:length(unique(data$x))), a$stats[nrow(a$stats), :
## NAs introduced by coercion
```


Car Type VS MPG
MPG
15
Compact
Large
Midsize
Small
Sporty
Van
Type of Car