Lab_10

Jyosna Philip

2024-02-09

1) Movie data:

A)Test whether the genre and main production company of a movie is having any significant effect on the the IMDB score.

B)If the score is affected due to the genre or production company, perform appropriate test to find out for which genre or which production company it is differing. [use multiple pairwise comparison test]

2) For the given nike data, test whether the average price of nike products that are in stock and out of stock are significantly differing or not.[Note: Check the assumptions first]

3)Take a sample of 250 from the population and test whether the variance of price of nike products is 45. Validate using an appropriate test.

4)Test whether the proportion of nike products that are in white, black, Navy are significantly different or not.

Ans 1

AIM : To test whether the genre and main production company of a movie is having any significant effect on the the IMDB score.

```
library(readxl)
Movie<- read_excel("C:/Users/jyosn/Downloads/Best Movie by Year
Netflix.xlsx")
```

About the dataset.

This dataset contains the best movies on IMDB from 1954 to 2021. It has 49 rows and 5 attributes.

```
head(Movie)

## # A tibble: 6 × 6
##   index TITLE                RELEASE_YEAR SCORE MAIN_GENRE MAIN_PRODUCTION
##   <dbl> <chr>                       <dbl> <dbl> <chr>      <chr>
## 1     0 White Christmas              1954   7.5 romance    US
## 2     1 The Guns of Navarone         1961   7.5 war        US
## 3     2 My Fair Lady                 1964   7.8 drama      US
## 4     3 Bonnie and Clyde             1967   7.7 drama      US
## 5     4 Dirty Harry                  1971   7.7 thriller   US
## 6     5 The Exorcist                 1973   8.1 horror     US
```

We carry out TWO-WAY ANOVA TEST as we have to check whether two attributes (with more than 2 levels) have an effect on the IMDB score.

```
#extract variables
genre<-Movie$MAIN_GENRE
production<- Movie$MAIN_PRODUCTION

genre<-as.factor(genre)
production<-as.factor(production)
levels(genre)

##  [1] "action"      "comedy"      "crime"       "documentary" "drama"
##  [6] "fantasy"     "horror"      "romance"     "scifi"       "thriller"
## [11] "war"         "western"

levels(production)

## [1] "DE" "FR" "GB" "HK" "IN" "JP" "US"
```

Therefore the two attributes have: Genre: it has 12 levels production: it has 7 levels

We try this with both ‘With interaction’ and ‘without interaction’.

WITHOUT INTERACTION

Define hypothesis

Ha0: Genre has no significant effect on the IMDB Score. Ha1: Genre has significant effect on the IMDB Score.

Hb0: main production company has no significant effect on the IMDB Score. Hb1: main production company has significant effect on the IMDB Score.

```
result<-aov(Movie$SCORE~genre+production,data=Movie)
result

## Call:
##    aov(formula = Movie$SCORE ~ genre + production, data = Movie)
##
## Terms:
##                     genre production Residuals
## Sum of Squares   5.502341   2.186973 11.530686
## Deg. of Freedom        11          6        31
##
## Residual standard error: 0.6098833
## Estimated effects may be unbalanced

summary(result)

##             Df Sum Sq Mean Sq F value Pr(>F)
## genre       11  5.502  0.5002   1.345  0.248
## production   6  2.187  0.3645   0.980  0.455
## Residuals   31 11.531  0.3720
```

Since p-values for both genre and production is greater than 0.05, we accept Ha0 and Hb0. Hence, Genre has no significant effect on the IMDB Score. Also Production company has no significant effect on IMDB Score.

WITH INTERACTION EFFECT

Defining the null and alternative hypothesis:
Ha0: There is no significant difference in the score of movies with difference in genre.
Ha1: There is a significant difference in the score of movies with difference in genre.

Hb0: There is no significant difference in the score of movies with difference in production company.

Hb1: There is a significant difference in the score of movies with difference in production company.

Hc0: There is no interaction effect

Hc1: There is an interaction effect

```
result1<-aov(Movie$SCORE~genre*production,data=Movie)
summary(result1)
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## genre             11  5.502  0.5002   1.289  0.283
## production         6  2.187  0.3645   0.939  0.483
## genre:production   4  1.054  0.2635   0.679  0.612
## Residuals         27 10.477  0.3880
```

All p-values obtained in this test is greater than 0.05. Hence we fail to reject Ha0,Hb0,Hc0. There is no significant effect on the score of movies due to their genre or production company. there is no interaction effect between genre and production company.

<u>Answer 2</u>

AIM: Test whether the average price of nike products that are in stock and out of stock are significantly differing or not.

```
library(readxl)

nike<- read_excel("C:/Users/jyosn/Downloads/nike_data_2022_09.xlsx")
head(nike)
```

```
## # A tibble: 6 × 18
##   index url       name  sub_title brand  model color price currency
availability
##   <dbl> <chr>     <chr> <chr>     <chr>  <dbl> <chr> <dbl> <chr>    <chr>
## 1     0 https://… Nike… Men's Lo… Nike  1.42e7 Navy     40   USD
InStock
```

```
## 2      1 https://… Club… Women's … Nike  1.38e7 Blac…  90   USD
InStock
## 3      2 https://… Nike… Men's Ov… Nike  1.30e7 Blac… 140   USD
OutOfStock
## 4      3 https://… Nike… Big Kids… Nike  1.38e7 Blac…  23.0 USD
OutOfStock
## 5      4 https://… Pari… Big Kids… Nike  1.33e7 Dark…  70   USD
InStock
## 6      5 https://… NFL … Men's Ga… Nike  1.41e7 White 130   USD
InStock
## # ℹ 8 more variables: description <chr>, raw_description <chr>,
## #   avg_rating <dbl>, review_count <dbl>, images <chr>, available_sizes
<chr>,
## #   uniq_id <chr>, scraped_at <chr>
```

```
availability<-as.factor(nike$availability)
levels(availability)
```

```
## [1] "InStock"    "OutOfStock"
```

Since availability attribute has only two levels, and we need to find the average mean of both types, we carry out the z test for equality of two means .

HYPOTHESIS

H0:average price of nike products that are in stock and out of stock are NOT significantly different. mu1=mu2

H1:average price of nike products that are in stock and out of stock are significantly different. mu1!=mu2

```
stock<-subset(nike,availability==c('InStock'))
not_instock<-subset(nike,availability==c('OutOfStock'))

stock_sample<-sample(stock$price,100,replace=TRUE)
outOfStock_sample<-sample(not_instock$price,100,replace=TRUE)

mean_stock<-mean(stock_sample)
mean_notInStock<-mean(outOfStock_sample)
sd_stock<-sd(stock_sample)
sd_notInStock<-sd(outOfStock_sample)

mean_stock
```

```
## [1] 66.138
```

```
mean_notInStock
```

```
## [1] 49.905
```

```
sd_stock
```

```
## [1] 44.63617
```

```
sd_notInStock

## [1] 32.8894

library(BSDA)

## Loading required package: lattice

##
## Attaching package: 'BSDA'

## The following object is masked _by_ '.GlobalEnv':
##
##     Movie

## The following object is masked from 'package:datasets':
##
##     Orange

z.test(x=stock_sample,y=outOfStock_sample,alternative='two.sided',mu=0,sigma.
x=sd_stock,sigma.y=sd_notInStock,conf.level = 0.95)

##
##  Two-sample z-Test
##
## data:  stock_sample and outOfStock_sample
## z = 2.9278, p-value = 0.003414
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.366065 27.099935
## sample estimates:
## mean of x mean of y
##    66.138    49.905
```

Since the p-value is less than 0.05, we reject H0. Hence the price of instock items and items not in stock are significantly different.

RIGHT-SIDED TEST

H0:average price of nike products that are in stock and out of stock are NOT significantly different. mu1=mu2

H1:average price of nike products that are in stock is significantly greater than that of out of stock . mu1<mu2

```
z.test(x=stock_sample,y=outOfStock_sample,alternative='greater',mu=0,sigma.x=
sd_stock,sigma.y=sd_notInStock,conf.level = 0.95)

##
##  Two-sample z-Test
##
## data:  stock_sample and outOfStock_sample
## z = 2.9278, p-value = 0.001707
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  7.113181       NA
## sample estimates:
## mean of x mean of y 
##    66.138    49.905
```

since p-value is less than 0.05, we reject H0. Hence, average price of nike products that are in stock is significantly greater than that of out of stock . mu1<mu2

Conclusion A z-test was carried out to check whether the average price of products instock significantly differed from that which were out of stock. Results show that the average price of products in stock are signifcantly greater than that which are out of stock.

Answer3:

```
sample1<-sample(nike,size=250,replace=TRUE)
```

We need to check whether the variance in price is 45. We need to carry out chi square test for this.

Importing required library

```
library(EnvStats)
```

```
## 
## Attaching package: 'EnvStats'
```

```
## The following objects are masked from 'package:stats':
## 
##     predict, predict.lm
```

Here our target variable is price

CHI SQUARE TEST FOR SINGLE VARIANCE

H0: The variance of price of nike products is 45

H1: The variance of price of nike products is not 45

Setting significance level α=0.05

Two tailed test

```
varTest(sample1$price,alternative = "two.sided",sigma.squared =45)
```

```
## 
## Results of Hypothesis Test
## --------------------------
## 
## Null Hypothesis:                 variance = 45
```

```
##
## Alternative Hypothesis:          True variance is not equal to 45
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 1591.032
##
## Data:                            sample1$price
##
## Test Statistic:                  Chi-Squared = 3924.545
##
## Test Statistic Parameter:        df = 111
##
## P-value:                         0
##
## 95% Confidence Interval:         LCL = 1243.268
##                                  UCL = 2109.088
```

We have obtained a p-value less tha 0.05 i.e., 0. Hence we reject H0 and conclude that the variance of nike products is not 45

Right sided test

H0: The variance of price of nike products is 45

H1: The variance of price of nike products is greater than 45

```
varTest(sample1$price,alternative = "greater",sigma.squared =45)
```

```
##
## Results of Hypothesis Test
## --------------------------
##
## Null Hypothesis:                 variance = 45
##
## Alternative Hypothesis:          True variance is greater than 45
##
## Test Name:                       Chi-Squared Test on Variance
##
## Estimated Parameter(s):          variance = 1591.032
##
## Data:                            sample1$price
##
## Test Statistic:                  Chi-Squared = 3924.545
##
## Test Statistic Parameter:        df = 111
##
## P-value:                         0
##
## 95% Confidence Interval:         LCL = 1292.943
##                                  UCL =      Inf
```

We obtain a p-value of 0. Hence we reject H0 and conclude that the variance of nike product price is greater than 45.

Conclusion A chi-square test for single variance was used to check whether the variance of nike products price was 45. Followed by the rejection of null hypothesis in two tailed test, a right tatiled test was conducted to confirm that the nike product price variance is greater than 45.

Ans 4

# AIM: To check whether the proportion of nike products in white,black and navy are same or not.

For this we need to use prop.test

```
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Setting up hypothesis

H0: The proportion of black, navy and white products are same.

H1: The proportion of black, navy and white products are not same.

Setting the significance level α=0.05

```
nike_df<-as.data.frame((nike))
colors<- subset(nike_df, nike_df$color %in%
c("White","Black","Navy","Black/Black","White/White","Navy/Navy"),color)

count(colors,color)

##         color  n
## 1       Black 15
## 2 Black/Black  1
## 3        Navy  3
## 4       White  9

count(nike_df)

##     n
## 1 112
```

```r
x=c(16,3,9) #number of black, navy, and white products respectively

n=c(112,112,112) #total number of products
prop.test(x, n, alternative = "two.sided", conf.level = 0.95)

##
##  3-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## X-squared = 9.8961, df = 2, p-value = 0.007097
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3
## 0.14285714 0.02678571 0.08035714
```

We got a p-value less than 0.05. Hence we reject H0 and conclude that atleast on among the proportion of black, navy and white nike products are significantly different.

```r
#checking proportion of black and white products
#Ha0:The proportion of black, and white products are same.
#Ha1: The proportion of black, and white products are not same.
x=c(16,9) #number of black,and white products respectively

n=c(112,112) #total number of products
prop.test(x, n, alternative = "two.sided", conf.level = 0.95)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 1.6209, df = 1, p-value = 0.203
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02849268  0.15349268
## sample estimates:
##     prop 1     prop 2
## 0.14285714 0.08035714
```

The p-value is greater than 0.05 hence we fail to reject Ha0 and conclude that the proprotion of white and black products are not signifcantly different.

```r
#checking proportion of white and navy products
#Hb0:The proportion of white, and navy products are same.
#Hb1: The proportion of white, and navy products are not same.
x=c(9,3) #number of white,and navy products respectively

n=c(112,112) #total number of products
prop.test(x, n, alternative = "two.sided", conf.level = 0.95)

##
##  2-sample test for equality of proportions with continuity correction
```

```
##
## data:  x out of n
## X-squared = 2.2013, df = 1, p-value = 0.1379
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.01391293  0.12105579
## sample estimates:
##     prop 1     prop 2
## 0.08035714 0.02678571
```

The p-value is greater than 0.05 hence we fail to reject Hb0 and conclude that the proportion of white and navy products are not significantly different.

```
#checking proportion of black and navy products
#Hc0:The proportion of black, and navy products are same.
#Hc1: The proportion of black, and navy products are not same.
x=c(16,3) #number of black,and navy products respectively

n=c(112,112) #total number of products
prop.test(x, n, alternative = "two.sided", conf.level = 0.95)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 8.2814, df = 1, p-value = 0.004005
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.03577092 0.19637194
## sample estimates:
##     prop 1     prop 2
## 0.14285714 0.02678571
```

The p-value is less than 0.05 hence we reject Hc0 and conclude that the proportion of navy and black products are significantly different.

Right tailed test for proportion of navy and black products

```
#Hd0:The proportion of black, and navy products are same.
#Hd1: The proportion of black products is greater than that of navy products.

prop.test(x, n, alternative = "greater", conf.level = 0.95)

##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## X-squared = 8.2814, df = 1, p-value = 0.002003
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04724564 1.00000000
## sample estimates:
##     prop 1     prop 2
## 0.14285714 0.02678571
```

Our p-value is less than 0.05 and hence we can reject Hd0. therefore the proportion of Black products is significantly greater than that of navy products.

Conclusion We did 3 sample proportion test to check whether the proportion of black, white and navy products are the same. our intitial test resulted in rejection of H0. therefore at least one pair of color had significantly different proportions. Further tests proved that that the proportion of black products were significantly greater than that of navy products. On the other hand, we also found that proportion of black and white products and the proportion of navy and white products are not significantly different.