Lab-10

Nilanjana Dey

2024-02-12

library(readxl)

## Warning: package 'readxl' was built under R version 4.2.3

nike <- read\_excel("C:/Users/NILANJANA/Downloads/nike\_data\_2022\_09.xlsx")  
View(nike)

library(readxl)  
netflix <- read\_excel("C:/Users/NILANJANA/Downloads/Best Movie by Year Netflix.xlsx")  
View(netflix)  
netflix

## # A tibble: 49 × 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   
## 7 6 Monty Python and the Hol… 1975 8.2 comedy GB   
## 8 7 Taxi Driver 1976 8.3 crime US   
## 9 8 Life of Brian 1979 8 comedy GB   
## 10 9 The Blue Lagoon 1980 5.8 romance US   
## # ℹ 39 more rows

# the libraries are installed  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.3

library(ggpubr)

## Warning: package 'ggpubr' was built under R version 4.2.3

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.1 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ lubridate 1.9.2 ✔ tibble 3.2.1  
## ✔ purrr 1.0.1 ✔ tidyr 1.3.0  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(broom)

## Warning: package 'broom' was built under R version 4.2.3

library(AICcmodavg)

## Warning: package 'AICcmodavg' was built under R version 4.2.3

library(dplyr)

# we try to identify the levels of "MAIN\_GENRE" factor  
netflix$MAIN\_GENRE=as.factor(netflix$MAIN\_GENRE)  
levels(netflix$MAIN\_GENRE)

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

# we try to identify the levels of "MAIN\_PRODUCTION" factor  
netflix$MAIN\_PRODUCTION=as.factor(netflix$MAIN\_PRODUCTION)  
levels(netflix$MAIN\_PRODUCTION)

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

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.

We will first check for difference with interaction effect between factors “MAIN\_PRODUCTION” and “MAIN\_GENRE”.

H0A: There is no significant difference in the SCORE with 7 levels of ‘MAIN\_PRODUCTION’ factor μ1=μ2=μ3=μ4=μ5=μ6=μ7

H0B: There is no significant difference in the SCORE with 12 levels of ‘MAIN\_GENRE’ factor

H0AB: The interaction effect of “MAIN\_PRODUCTION” and “MAIN\_SCORE” is insignificant.

vs

H1A: atleast one inequality in H0A H1B: atleast one inequality in H0B H1AB: atleast one inequality in H0A

two.way <- aov(SCORE ~ MAIN\_PRODUCTION\*MAIN\_GENRE, data = netflix)  
summary(two.way)

## Df Sum Sq Mean Sq F value Pr(>F)  
## MAIN\_PRODUCTION 6 3.670 0.6117 1.577 0.192  
## MAIN\_GENRE 11 4.019 0.3654 0.942 0.518  
## MAIN\_PRODUCTION:MAIN\_GENRE 4 1.054 0.2635 0.679 0.612  
## Residuals 27 10.477 0.3880

#Interpretation: We can see that the two factors and their interaction effect is insignificant which is evident from the p-value.In every case we accept the H0.

#Without interaction effect  
two.way <- aov(SCORE ~ MAIN\_PRODUCTION+MAIN\_GENRE, data = netflix)  
summary(two.way)

## Df Sum Sq Mean Sq F value Pr(>F)  
## MAIN\_PRODUCTION 6 3.670 0.6117 1.645 0.168  
## MAIN\_GENRE 11 4.019 0.3654 0.982 0.483  
## Residuals 31 11.531 0.3720

#Interpretation: Here also the two factors are concluded to be insignificant as p-value>0.05. Hence we don't perform the Tukey-HSD test.

1. 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]

population\_1=filter(nike,availability=='InStock')

sample\_1=sample(population\_1$price,20,replace=FALSE)  
sample\_1

## [1] 115.00 90.00 55.97 110.00 50.00 28.97 34.97 130.00 40.00 100.00  
## [11] 160.00 33.97 58.00 37.97 92.97 26.97 40.00 34.97 13.97 18.97

population\_2=filter(nike,availability=='OutOfStock')  
population\_2

## # A tibble: 41 × 18  
## index url name sub\_title brand model color price currency availability  
## <dbl> <chr> <chr> <chr> <chr> <dbl> <chr> <dbl> <chr> <chr>   
## 1 2 https:/… Nike… Men's Ov… Nike 1.30e7 Blac… 140 USD OutOfStock   
## 2 3 https:/… Nike… Big Kids… Nike 1.38e7 Blac… 23.0 USD OutOfStock   
## 3 6 https:/… Kris… Nike NBA… Nike 1.18e7 Rush… 105. USD OutOfStock   
## 4 10 https:/… Nike… Little K… Nike 1.23e7 Aege… 46.0 USD OutOfStock   
## 5 14 https:/… Nike… Men's Sh… Nike 1.40e7 Black 45 USD OutOfStock   
## 6 16 https:/… Nike… Women's … Nike 1.29e7 Grey… 41.0 USD OutOfStock   
## 7 18 https:/… Nike… Hat Nike 1.34e7 Grey… 22.0 USD OutOfStock   
## 8 20 https:/… Denv… Men's NB… Nike 1.24e7 White 22.0 USD OutOfStock   
## 9 21 https:/… Hurl… Boys' Bo… Nike 1.30e7 <NA> 40 USD OutOfStock   
## 10 26 https:/… Nike Little K… Nike 1.30e7 Black 18 USD OutOfStock   
## # ℹ 31 more rows  
## # 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>

sample\_2=sample(population\_1$price,10,replace=FALSE)  
sample\_2

## [1] 63.97 15.97 119.97 24.97 38.97 92.97 18.97 160.00 165.00 95.00

#Two-tailed H0:MEAN\_1=MEAN\_2 VS H1:MEAN\_1!=MEAN\_2

library(BSDA)

## Warning: package 'BSDA' was built under R version 4.2.3

## Loading required package: lattice

##   
## Attaching package: 'BSDA'

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

z.test(x=sample\_1,y=sample\_2,  
alternative = "two.sided",  
mu = 0,  
sigma.x = sd(population\_1$price),  
sigma.y = sd(population\_2$price),  
conf.level = 0.95  
)

##   
## Two-sample z-Test  
##   
## data: sample\_1 and sample\_2  
## z = -1.1048, p-value = 0.2692  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -44.229 12.341  
## sample estimates:  
## mean of x mean of y   
## 63.635 79.579

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

# our target variable is color  
target\_var=nike$color

# a sample of 50 observations are drawn  
sample1=sample(target\_var,50,replace=TRUE)  
sample1

## [1] "Multi-Color/Lavender Mist"   
## [2] "Black"   
## [3] "Black/White/Metallic Gold"   
## [4] "Rush Blue/College Navy/White"   
## [5] "College Navy/White/University Red/White"  
## [6] "Hyper Pink"   
## [7] "White"   
## [8] "Grey Heather/Valor Blue"   
## [9] "Black"   
## [10] "Atmosphere"   
## [11] "Aegean Storm"   
## [12] "Bleached Coral"   
## [13] "Sail/Pearl White"   
## [14] "Seafoam/Brushed Silver"   
## [15] "Royal"   
## [16] "Navy"   
## [17] NA   
## [18] "White/Black"   
## [19] "Navy"   
## [20] "Black"   
## [21] "Black"   
## [22] "Midnight Navy/White/Midnight Navy"   
## [23] "White"   
## [24] "Black/Gold"   
## [25] "Midnight Navy"   
## [26] "Black/Smoke Grey"   
## [27] "Black/White/White"   
## [28] "Bleached Coral"   
## [29] "Multi-Color"   
## [30] "Green"   
## [31] "White/Black"   
## [32] "Atmosphere"   
## [33] "University Red"   
## [34] "Purple"   
## [35] "Court Blue/College Navy"   
## [36] "Grey Heather/White/Pro Green"   
## [37] "Black"   
## [38] "Scarlet/Black"   
## [39] "Black"   
## [40] "Black"   
## [41] "Black"   
## [42] "Black/White"   
## [43] "Team Gold"   
## [44] "Black Heather/Black/Heather/Black"   
## [45] "Medium Olive/Black/Particle Grey"   
## [46] "Black"   
## [47] "Navy"   
## [48] "Black/Iron Grey/White/Black"   
## [49] "Black"   
## [50] "White/Black"

# a dataframe is created from the samples  
s=as.data.frame(sample1)  
s

## sample1  
## 1 Multi-Color/Lavender Mist  
## 2 Black  
## 3 Black/White/Metallic Gold  
## 4 Rush Blue/College Navy/White  
## 5 College Navy/White/University Red/White  
## 6 Hyper Pink  
## 7 White  
## 8 Grey Heather/Valor Blue  
## 9 Black  
## 10 Atmosphere  
## 11 Aegean Storm  
## 12 Bleached Coral  
## 13 Sail/Pearl White  
## 14 Seafoam/Brushed Silver  
## 15 Royal  
## 16 Navy  
## 17 <NA>  
## 18 White/Black  
## 19 Navy  
## 20 Black  
## 21 Black  
## 22 Midnight Navy/White/Midnight Navy  
## 23 White  
## 24 Black/Gold  
## 25 Midnight Navy  
## 26 Black/Smoke Grey  
## 27 Black/White/White  
## 28 Bleached Coral  
## 29 Multi-Color  
## 30 Green  
## 31 White/Black  
## 32 Atmosphere  
## 33 University Red  
## 34 Purple  
## 35 Court Blue/College Navy  
## 36 Grey Heather/White/Pro Green  
## 37 Black  
## 38 Scarlet/Black  
## 39 Black  
## 40 Black  
## 41 Black  
## 42 Black/White  
## 43 Team Gold  
## 44 Black Heather/Black/Heather/Black  
## 45 Medium Olive/Black/Particle Grey  
## 46 Black  
## 47 Navy  
## 48 Black/Iron Grey/White/Black  
## 49 Black  
## 50 White/Black

count(s,sample1)

## sample1 n  
## 1 Aegean Storm 1  
## 2 Atmosphere 2  
## 3 Black 10  
## 4 Black Heather/Black/Heather/Black 1  
## 5 Black/Gold 1  
## 6 Black/Iron Grey/White/Black 1  
## 7 Black/Smoke Grey 1  
## 8 Black/White 1  
## 9 Black/White/Metallic Gold 1  
## 10 Black/White/White 1  
## 11 Bleached Coral 2  
## 12 College Navy/White/University Red/White 1  
## 13 Court Blue/College Navy 1  
## 14 Green 1  
## 15 Grey Heather/Valor Blue 1  
## 16 Grey Heather/White/Pro Green 1  
## 17 Hyper Pink 1  
## 18 Medium Olive/Black/Particle Grey 1  
## 19 Midnight Navy 1  
## 20 Midnight Navy/White/Midnight Navy 1  
## 21 Multi-Color 1  
## 22 Multi-Color/Lavender Mist 1  
## 23 Navy 3  
## 24 Purple 1  
## 25 Royal 1  
## 26 Rush Blue/College Navy/White 1  
## 27 Sail/Pearl White 1  
## 28 Scarlet/Black 1  
## 29 Seafoam/Brushed Silver 1  
## 30 Team Gold 1  
## 31 University Red 1  
## 32 White 2  
## 33 White/Black 3  
## 34 <NA> 1

H0 : the proportion of nike products that are in white, black, Navy are same.

vs

H1 : the proportion of nike products that are in white, black, Navy are significantly different.

x=c(3,1,4)  
n=c(50,50,50)  
prop.test(x, n, alternative = "two.sided", conf.level = 0.95)

## Warning in prop.test(x, n, alternative = "two.sided", conf.level = 0.95):  
## Chi-squared approximation may be incorrect

##   
## 3-sample test for equality of proportions without continuity correction  
##   
## data: x out of n  
## X-squared = 1.8486, df = 2, p-value = 0.3968  
## alternative hypothesis: two.sided  
## sample estimates:  
## prop 1 prop 2 prop 3   
## 0.06 0.02 0.08

#Interpretation: The proportion of nike poducts in white,black and navy blue is sme which is indicated by the acceptance of H0 from the p-value.

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.

A sample of 25 observations are taken and replicated 250 times.Now we conduct the one-sample variance test where the test sample follows chi-square test.

#replicating the sample   
  
samp\_dist=replicate(250,mean(sample(nike$price,25,replace=TRUE)))

H0 : sigma.squared = 45 vs H1 : sigma.squared != 45

library(EnvStats)

## Warning: package 'EnvStats' was built under R version 4.2.3

##   
## Attaching package: 'EnvStats'

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

varTest(samp\_dist, alternative = "two.sided", conf.level = 0.95, 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 = 55.95251  
##   
## Data: samp\_dist  
##   
## Test Statistic: Chi-Squared = 309.6039  
##   
## Test Statistic Parameter: df = 249  
##   
## P-value: 0.01072532  
##   
## 95% Confidence Interval: LCL = 47.29171  
## UCL = 67.24490

#Interpretation: We reject H0 which implies that the varinace is not 45.