PRACTICE QS.

Nilanjana Dey

2024-02-16

library(readr)

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

BankChurners<- read\_csv("C:/Users/NILANJANA/Downloads/BankChurners.csv.zip")

## Rows: 10127 Columns: 23  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (6): Attrition\_Flag, Gender, Education\_Level, Marital\_Status, Income\_Ca...  
## dbl (17): CLIENTNUM, Customer\_Age, Dependent\_count, Months\_on\_book, Total\_Re...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

View(BankChurners)  
BankChurners

## # A tibble: 10,127 × 23  
## CLIENTNUM Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level  
## <dbl> <chr> <dbl> <chr> <dbl> <chr>   
## 1 768805383 Existing Custo… 45 M 3 High School   
## 2 818770008 Existing Custo… 49 F 5 Graduate   
## 3 713982108 Existing Custo… 51 M 3 Graduate   
## 4 769911858 Existing Custo… 40 F 4 High School   
## 5 709106358 Existing Custo… 40 M 3 Uneducated   
## 6 713061558 Existing Custo… 44 M 2 Graduate   
## 7 810347208 Existing Custo… 51 M 4 Unknown   
## 8 818906208 Existing Custo… 32 M 0 High School   
## 9 710930508 Existing Custo… 37 M 3 Uneducated   
## 10 719661558 Existing Custo… 48 M 2 Graduate   
## # ℹ 10,117 more rows  
## # ℹ 17 more variables: Marital\_Status <chr>, Income\_Category <chr>,  
## # Card\_Category <chr>, Months\_on\_book <dbl>, Total\_Relationship\_Count <dbl>,  
## # Months\_Inactive\_12\_mon <dbl>, Contacts\_Count\_12\_mon <dbl>,  
## # Credit\_Limit <dbl>, Total\_Revolving\_Bal <dbl>, Avg\_Open\_To\_Buy <dbl>,  
## # Total\_Amt\_Chng\_Q4\_Q1 <dbl>, Total\_Trans\_Amt <dbl>, Total\_Trans\_Ct <dbl>,  
## # Total\_Ct\_Chng\_Q4\_Q1 <dbl>, Avg\_Utilization\_Ratio <dbl>, …

# 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 '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 ✔ stringr 1.5.0  
## ✔ forcats 1.0.0 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── 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)  
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

BankChurners$Education\_Level=as.factor(BankChurners$Education\_Level)  
levels(BankChurners$Education\_Level)

## [1] "College" "Doctorate" "Graduate" "High School"   
## [5] "Post-Graduate" "Uneducated" "Unknown"

BankChurners$Marital\_Status=as.factor(BankChurners$Marital\_Status)  
levels(BankChurners$Marital\_Status)

## [1] "Divorced" "Married" "Single" "Unknown"

two\_way=aov(Credit\_Limit~Education\_Level\*Marital\_Status,data = BankChurners)  
summary(two\_way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Education\_Level 6 2.030e+08 33839553 0.411 0.872   
## Marital\_Status 3 2.848e+09 949193434 11.523 1.54e-07 \*\*\*  
## Education\_Level:Marital\_Status 18 1.518e+09 84356244 1.024 0.428   
## Residuals 10099 8.319e+11 82374286   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

two\_way=aov(Credit\_Limit~Education\_Level+Marital\_Status,data = BankChurners)  
summary(two\_way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Education\_Level 6 2.030e+08 33839553 0.411 0.872   
## Marital\_Status 3 2.848e+09 949193434 11.522 1.55e-07 \*\*\*  
## Residuals 10117 8.334e+11 82377813   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(two\_way, conf.level=.95)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Credit\_Limit ~ Education\_Level + Marital\_Status, data = BankChurners)  
##   
## $Education\_Level  
## diff lwr upr p adj  
## Doctorate-College -271.27715 -1786.3900 1243.8357 0.9984512  
## Graduate-College -118.43520 -1086.0042 849.1338 0.9998263  
## High School-College -78.71258 -1109.7520 952.3269 0.9999892  
## Post-Graduate-College 178.02433 -1269.5524 1625.6011 0.9998214  
## Uneducated-College 214.97288 -875.4055 1305.3513 0.9973404  
## Unknown-College -192.73718 -1278.4518 892.9774 0.9985244  
## Graduate-Doctorate 152.84195 -1195.2721 1500.9560 0.9998890  
## High School-Doctorate 192.56457 -1201.8041 1586.9333 0.9996497  
## Post-Graduate-Doctorate 449.30149 -1276.0102 2174.6131 0.9879255  
## Uneducated-Doctorate 486.25003 -952.5504 1925.0505 0.9548898  
## Unknown-Doctorate 78.53997 -1356.7293 1513.8093 0.9999985  
## High School-Graduate 39.72262 -725.0564 804.5017 0.9999989  
## Post-Graduate-Graduate 296.45954 -975.2807 1568.1997 0.9933194  
## Uneducated-Graduate 333.40808 -509.6635 1176.4796 0.9068067  
## Unknown-Graduate -74.30198 -911.3329 762.7290 0.9999737  
## Post-Graduate-High School 256.73692 -1063.9355 1577.4094 0.9975413  
## Uneducated-High School 293.68546 -621.5316 1208.9025 0.9649366  
## Unknown-High School -114.02460 -1023.6802 795.6310 0.9998004  
## Uneducated-Post-Graduate 36.94855 -1330.5522 1404.4493 1.0000000  
## Unknown-Post-Graduate -370.76152 -1734.5465 993.0235 0.9848729  
## Unknown-Uneducated -407.71006 -1384.1096 568.6894 0.8820750  
##   
## $Marital\_Status  
## diff lwr upr p adj  
## Married-Divorced -1265.79766 -2184.0209 -347.5744 0.0022550  
## Single-Divorced -342.96454 -1273.0338 587.1047 0.7791447  
## Unknown-Divorced 98.72425 -1106.7723 1304.2208 0.9967082  
## Single-Married 922.83312 418.8786 1426.7876 0.0000153  
## Unknown-Married 1364.52191 446.8274 2282.2164 0.0007724  
## Unknown-Single 441.68879 -487.8585 1371.2361 0.6135192

pop1=filter(BankChurners,Attrition\_Flag == 'Existing Customer')  
pop1

## # A tibble: 8,500 × 23  
## CLIENTNUM Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level  
## <dbl> <chr> <dbl> <chr> <dbl> <fct>   
## 1 768805383 Existing Custo… 45 M 3 High School   
## 2 818770008 Existing Custo… 49 F 5 Graduate   
## 3 713982108 Existing Custo… 51 M 3 Graduate   
## 4 769911858 Existing Custo… 40 F 4 High School   
## 5 709106358 Existing Custo… 40 M 3 Uneducated   
## 6 713061558 Existing Custo… 44 M 2 Graduate   
## 7 810347208 Existing Custo… 51 M 4 Unknown   
## 8 818906208 Existing Custo… 32 M 0 High School   
## 9 710930508 Existing Custo… 37 M 3 Uneducated   
## 10 719661558 Existing Custo… 48 M 2 Graduate   
## # ℹ 8,490 more rows  
## # ℹ 17 more variables: Marital\_Status <fct>, Income\_Category <chr>,  
## # Card\_Category <chr>, Months\_on\_book <dbl>, Total\_Relationship\_Count <dbl>,  
## # Months\_Inactive\_12\_mon <dbl>, Contacts\_Count\_12\_mon <dbl>,  
## # Credit\_Limit <dbl>, Total\_Revolving\_Bal <dbl>, Avg\_Open\_To\_Buy <dbl>,  
## # Total\_Amt\_Chng\_Q4\_Q1 <dbl>, Total\_Trans\_Amt <dbl>, Total\_Trans\_Ct <dbl>,  
## # Total\_Ct\_Chng\_Q4\_Q1 <dbl>, Avg\_Utilization\_Ratio <dbl>, …

sample1=sample(pop1$Total\_Trans\_Amt,2000,replace=TRUE)

pop2=filter(BankChurners,Attrition\_Flag == 'Attrited Customer')  
pop2

## # A tibble: 1,627 × 23  
## CLIENTNUM Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level  
## <dbl> <chr> <dbl> <chr> <dbl> <fct>   
## 1 708508758 Attrited Custo… 62 F 0 Graduate   
## 2 708300483 Attrited Custo… 66 F 0 Doctorate   
## 3 779471883 Attrited Custo… 54 F 1 Graduate   
## 4 714374133 Attrited Custo… 56 M 2 Graduate   
## 5 712030833 Attrited Custo… 48 M 2 Graduate   
## 6 711013983 Attrited Custo… 55 F 4 Unknown   
## 7 711887583 Attrited Custo… 47 M 2 Unknown   
## 8 720201033 Attrited Custo… 53 M 2 Graduate   
## 9 789322833 Attrited Custo… 48 F 5 High School   
## 10 767712558 Attrited Custo… 59 M 1 College   
## # ℹ 1,617 more rows  
## # ℹ 17 more variables: Marital\_Status <fct>, Income\_Category <chr>,  
## # Card\_Category <chr>, Months\_on\_book <dbl>, Total\_Relationship\_Count <dbl>,  
## # Months\_Inactive\_12\_mon <dbl>, Contacts\_Count\_12\_mon <dbl>,  
## # Credit\_Limit <dbl>, Total\_Revolving\_Bal <dbl>, Avg\_Open\_To\_Buy <dbl>,  
## # Total\_Amt\_Chng\_Q4\_Q1 <dbl>, Total\_Trans\_Amt <dbl>, Total\_Trans\_Ct <dbl>,  
## # Total\_Ct\_Chng\_Q4\_Q1 <dbl>, Avg\_Utilization\_Ratio <dbl>, …

sample2=sample(pop2$Total\_Trans\_Amt,500,replace=TRUE)

z.test(x=sample1,y=sample2,  
alternative = "two.sided",  
mu = 0,  
sigma.x = sd(pop2$Total\_Trans\_Amt),  
sigma.y = sd(pop2$Total\_Trans\_Amt),  
conf.level = 0.95  
)

##   
## Two-sample z-Test  
##   
## data: sample1 and sample2  
## z = 13.047, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 1279.556 1731.960  
## sample estimates:  
## mean of x mean of y   
## 4543.632 3037.874

one\_way=aov(Credit\_Limit~Card\_Category,data = BankChurners)  
summary(one\_way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Card\_Category 3 2.237e+11 7.458e+10 1232 <2e-16 \*\*\*  
## Residuals 10123 6.127e+11 6.053e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

one\_way=aov(Credit\_Limit~Gender,data = BankChurners)  
summary(one\_way)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Gender 1 1.481e+11 1.481e+11 2179 <2e-16 \*\*\*  
## Residuals 10125 6.883e+11 6.798e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pop3=filter(BankChurners,Card\_Category=='Blue')  
pop3

## # A tibble: 9,436 × 23  
## CLIENTNUM Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level  
## <dbl> <chr> <dbl> <chr> <dbl> <fct>   
## 1 768805383 Existing Custo… 45 M 3 High School   
## 2 818770008 Existing Custo… 49 F 5 Graduate   
## 3 713982108 Existing Custo… 51 M 3 Graduate   
## 4 769911858 Existing Custo… 40 F 4 High School   
## 5 709106358 Existing Custo… 40 M 3 Uneducated   
## 6 713061558 Existing Custo… 44 M 2 Graduate   
## 7 710930508 Existing Custo… 37 M 3 Uneducated   
## 8 719661558 Existing Custo… 48 M 2 Graduate   
## 9 708790833 Existing Custo… 42 M 5 Uneducated   
## 10 710821833 Existing Custo… 65 M 1 Unknown   
## # ℹ 9,426 more rows  
## # ℹ 17 more variables: Marital\_Status <fct>, Income\_Category <chr>,  
## # Card\_Category <chr>, Months\_on\_book <dbl>, Total\_Relationship\_Count <dbl>,  
## # Months\_Inactive\_12\_mon <dbl>, Contacts\_Count\_12\_mon <dbl>,  
## # Credit\_Limit <dbl>, Total\_Revolving\_Bal <dbl>, Avg\_Open\_To\_Buy <dbl>,  
## # Total\_Amt\_Chng\_Q4\_Q1 <dbl>, Total\_Trans\_Amt <dbl>, Total\_Trans\_Ct <dbl>,  
## # Total\_Ct\_Chng\_Q4\_Q1 <dbl>, Avg\_Utilization\_Ratio <dbl>, …

sample3=sample(pop3$Gender,3000,replace=TRUE)

s=as.data.frame(sample3)  
count(s,sample3)

## sample3 n  
## 1 F 1669  
## 2 M 1331

n=3000  
x=1638  
prop.test(x, n, p = 0.6, alternative = "two.sided",  
 correct = TRUE)

##   
## 1-sample proportions test with continuity correction  
##   
## data: x out of n, null probability 0.6  
## X-squared = 36.225, df = 1, p-value = 1.758e-09  
## alternative hypothesis: true p is not equal to 0.6  
## 95 percent confidence interval:  
## 0.5279694 0.5639118  
## sample estimates:  
## p   
## 0.546

pop4=filter(BankChurners,Gender=='F')  
pop4

## # A tibble: 5,358 × 23  
## CLIENTNUM Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level  
## <dbl> <chr> <dbl> <chr> <dbl> <fct>   
## 1 818770008 Existing Custo… 49 F 5 Graduate   
## 2 769911858 Existing Custo… 40 F 4 High School   
## 3 712396908 Existing Custo… 57 F 2 Graduate   
## 4 709327383 Existing Custo… 45 F 2 Graduate   
## 5 708508758 Attrited Custo… 62 F 0 Graduate   
## 6 811604133 Existing Custo… 47 F 4 Unknown   
## 7 771071958 Existing Custo… 41 F 3 Graduate   
## 8 718813833 Existing Custo… 44 F 3 Uneducated   
## 9 788658483 Existing Custo… 53 F 2 College   
## 10 715318008 Existing Custo… 55 F 1 College   
## # ℹ 5,348 more rows  
## # ℹ 17 more variables: Marital\_Status <fct>, Income\_Category <chr>,  
## # Card\_Category <chr>, Months\_on\_book <dbl>, Total\_Relationship\_Count <dbl>,  
## # Months\_Inactive\_12\_mon <dbl>, Contacts\_Count\_12\_mon <dbl>,  
## # Credit\_Limit <dbl>, Total\_Revolving\_Bal <dbl>, Avg\_Open\_To\_Buy <dbl>,  
## # Total\_Amt\_Chng\_Q4\_Q1 <dbl>, Total\_Trans\_Amt <dbl>, Total\_Trans\_Ct <dbl>,  
## # Total\_Ct\_Chng\_Q4\_Q1 <dbl>, Avg\_Utilization\_Ratio <dbl>, …

pop5=filter(BankChurners,Gender=='M')  
pop5

## # A tibble: 4,769 × 23  
## CLIENTNUM Attrition\_Flag Customer\_Age Gender Dependent\_count Education\_Level  
## <dbl> <chr> <dbl> <chr> <dbl> <fct>   
## 1 768805383 Existing Custo… 45 M 3 High School   
## 2 713982108 Existing Custo… 51 M 3 Graduate   
## 3 709106358 Existing Custo… 40 M 3 Uneducated   
## 4 713061558 Existing Custo… 44 M 2 Graduate   
## 5 810347208 Existing Custo… 51 M 4 Unknown   
## 6 818906208 Existing Custo… 32 M 0 High School   
## 7 710930508 Existing Custo… 37 M 3 Uneducated   
## 8 719661558 Existing Custo… 48 M 2 Graduate   
## 9 708790833 Existing Custo… 42 M 5 Uneducated   
## 10 710821833 Existing Custo… 65 M 1 Unknown   
## # ℹ 4,759 more rows  
## # ℹ 17 more variables: Marital\_Status <fct>, Income\_Category <chr>,  
## # Card\_Category <chr>, Months\_on\_book <dbl>, Total\_Relationship\_Count <dbl>,  
## # Months\_Inactive\_12\_mon <dbl>, Contacts\_Count\_12\_mon <dbl>,  
## # Credit\_Limit <dbl>, Total\_Revolving\_Bal <dbl>, Avg\_Open\_To\_Buy <dbl>,  
## # Total\_Amt\_Chng\_Q4\_Q1 <dbl>, Total\_Trans\_Amt <dbl>, Total\_Trans\_Ct <dbl>,  
## # Total\_Ct\_Chng\_Q4\_Q1 <dbl>, Avg\_Utilization\_Ratio <dbl>, …

sample4=sample(pop4$Credit\_Limit,2000,replace=FALSE)

sample5=sample(pop5$Credit\_Limit,1000,replace=FALSE)

var.test(sample4,sample5, alternative = "two.sided", conf.level = 0.9)

##   
## F test to compare two variances  
##   
## data: sample4 and sample5  
## F = 0.22756, num df = 1999, denom df = 999, p-value < 2.2e-16  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 90 percent confidence interval:  
## 0.2077741 0.2488328  
## sample estimates:  
## ratio of variances   
## 0.2275574

sample6=sample(pop3$Total\_Revolving\_Bal,350,replace=FALSE)

z.test(x=sample6,alternative="two.sided",mu=1325,sigma.x=sd(pop3$Total\_Revolving\_Bal),conf.level=0.95)

##   
## One-sample z-Test  
##   
## data: sample6  
## z = -4.4196, p-value = 9.889e-06  
## alternative hypothesis: true mean is not equal to 1325  
## 95 percent confidence interval:  
## 1046.703 1217.703  
## sample estimates:  
## mean of x   
## 1132.203