End-Semester

Nilanjana Dey

2024-02-17

# Healthcare dataset is taken.   
#The set consists of 10000 rows and 15 columns.  
#The columns are as follows:Name,Age,Gender,Blood Type,Medical Condition,Date of Admission,Doctor,Hospital,Insurance Provider,Billing Amount,Room Number,Admission Type,Discharge Date,Medication,Test Results  
  
library(readr)

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

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

## Rows: 10000 Columns: 15  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (10): Name, Gender, Blood Type, Medical Condition, Doctor, Hospital, In...  
## dbl (3): Age, Billing Amount, Room Number  
## date (2): Date of Admission, Discharge Date  
##   
## 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.

dim(healthcare)

## [1] 10000 15

head(healthcare)

## # A tibble: 6 × 15  
## Name Age Gender `Blood Type` `Medical Condition` `Date of Admission` Doctor  
## <chr> <dbl> <chr> <chr> <chr> <date> <chr>   
## 1 Tiff… 81 Female O- Diabetes 2022-11-17 Patri…  
## 2 Rube… 35 Male O+ Asthma 2023-06-01 Diane…  
## 3 Chad… 61 Male B- Obesity 2019-01-09 Paul …  
## 4 Anto… 49 Male B- Asthma 2020-05-02 Brian…  
## 5 Mrs.… 51 Male O- Arthritis 2021-07-09 Dusti…  
## 6 Patr… 41 Male AB+ Arthritis 2020-08-20 Robin…  
## # ℹ 8 more variables: Hospital <chr>, `Insurance Provider` <chr>,  
## # `Billing Amount` <dbl>, `Room Number` <dbl>, `Admission Type` <chr>,  
## # `Discharge Date` <date>, Medication <chr>, `Test Results` <chr>

library(readr)  
Crop<- read\_csv("C:/Users/NILANJANA/Downloads/Crop\_recommendation.csv")

## Rows: 2200 Columns: 8  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): label  
## dbl (7): N, P, K, temperature, humidity, ph, rainfall  
##   
## 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.

# the libraries are called.  
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

library(dplyr)

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

##   
## 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

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

library(ggpubr)

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

## Loading required package: ggplot2

## Warning: package 'ggplot2' 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 '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 ──  
## ✔ 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

Question 1:

Dynamic Solutions Ltd., with its head office situated in Kolkata and a branch in Mumbai, aims to introduce a new work plan. To gauge worker attitudes towards this plan, a survey was conducted. Out of a sample of 500 workers in Kolkata, 62% expressed support for the new plan. Conversely, out of a sample of 400 workers in Mumbai, 41% were against it. The Personnel Director seeks to determine if there is a significant difference in the attitudes of workers between the two locations at a 5% significance level.

We perform the 2 sample proportion test here.

H0: p1=p2

vs

H1: p1!=p2

where, p1=proportion of workers from Kolkata who are supporting the new plan

p2=proportion of workers from Mumbai who are against the new plan.

sample taken are given as follows:

n1=500  
n2=400  
x1=0.62\*n1# Out of a sample of 500 workers in Kolkata, 62% expressed support for the new plan  
x2=0.41\*n2# out of a sample of 400 workers in Mumbai, 41% were against it

x1

## [1] 310

x2

## [1] 164

x=c(x1,x2)  
n=c(500,400)  
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 = 38.474, df = 1, p-value = 5.549e-10  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.1434599 0.2765401  
## sample estimates:  
## prop 1 prop 2   
## 0.62 0.41

#Interpretation: We reject H0 because p-value<0.05. This implies that there is difference in the attitude of workers from Kolkata and Mumbai.

QUESTION 2:

Global Solutions Inc. Satisfaction Analysis Question The HR department of Global Solutions Inc Corporation conducted a satisfaction survey among employees to assess job satisfaction levels. The results revealed that 772 employees expressed dissatisfaction with their current roles, 1611 reported moderate satisfaction, and 737 indicated high satisfaction and the remaining employees did not provide satisfaction ratings. However, Global Solutions Inc Corporation’s HR policies anticipate a distribution of job satisfaction levels in the ratio of 1:2:1 for dissatisfied, moderately satisfied, and highly satisfied employees respectively. Considering this, analyze whether the observed satisfaction levels align with the expected ratios. Perform a hypothesis test using an appropriate statistical method to determine if there are any notable deviations exist. Based on the test results, draw conclusions about whether further inquiries or actions are necessary to address any notable deviations effectively.

Here, we perform the chis-square test of goodness of fit.

H0:The ratios of satisfaction level are equal to the observed frequencies)

vs

H1: at least one inequality in H0 (The given ratios are not equal)

obs=c(772,1611,737)# these are the observed frequencies given.  
expected\_prob=c(1/4,2/4,1/4)# expected probabilities (1:2:1 -> p1:p2:p3)  
# Expected values  
expected <- sum(obs) \* expected\_prob  
chisq.test(obs,p=expected\_prob)# we perform the chi-square test

##   
## Chi-squared test for given probabilities  
##   
## data: obs  
## X-squared = 4.1199, df = 2, p-value = 0.1275

expected # we can see that there is not much significant difference between observed and expected values

## [1] 780 1560 780

#Hence the chi-square test statistic value is not so high...that's why we get the p-value>0.05 which indicates the acceptance of H0. That is, the given probabilities is equivalent to the observed frequencies.

Question 3:

Statistical Analysis of Healthcare Data a) Using the Health Care dataset, examine whether individuals admitted under emergency conditions have an average bill of 25,500, based on a suitable sample size, with a significance level of 1%. b) Investigate whether there is equal variance in billing amounts among individuals with Diabetes and Hypertension. Utilize a suitable statistical test for validation and draw meaningful inferences from the results.

Part a:

We perform the one sample mean test where population variance is known.

H0: mu=25000(given)

vs

H1: mu!=25000

Our target variable is Billing Amount

# First we filter the dataset according to the emergency conditon of the patients  
pop1=filter(healthcare,`Admission Type`=="Emergency")  
dim(pop1)# we get 3367 observations

## [1] 3367 15

# Now we draw sample of 1000 observations from the "Billing amount" column without replacement.   
sample1=sample(pop1$`Billing Amount`,1000,replace=FALSE)

# we perform the z test from the sample values.  
z.test(sample1,alternative="two.sided",mu=25000,sigma.x=sd(pop1$`Billing Amount`),conf.level = 0.9)

##   
## One-sample z-Test  
##   
## data: sample1  
## z = -0.54091, p-value = 0.5886  
## alternative hypothesis: true mean is not equal to 25000  
## 90 percent confidence interval:  
## 24018.08 25495.93  
## sample estimates:  
## mean of x   
## 24757

#Interpretation: The p-value>0.05 indicates the acceptance of H0. This indicates that the mean billing amount is equal to 25000 for emergency patients.

Part b:

We perform the test for equality of 2 population variance.

H0: var1=var2

H1: var1!=var2

var1=sample variance of patients having Diabetes var2=sample variance of patients having Hypertension

# First we filter the dataset according to "Diabetes" and "Hypertension"  
pop2=filter(healthcare,`Medical Condition`=="Diabetes")  
  
pop3=filter(healthcare,`Medical Condition`=="Hypertension")  
  
dim(pop2) #there are 1628 observations for Diabetes patients

## [1] 1623 15

dim(pop3) #there are 1688 observations for Hypertension patients

## [1] 1688 15

# We draw samples of 500 observations from both the populations without replacement  
sample2=sample(pop2$`Billing Amount`,500,replace=FALSE)  
sample3=sample(pop3$`Billing Amount`,500,replace=FALSE)

# Now we perform the F test to check the equality of variances.  
var.test(sample2,sample3,alternative="two.sided",conf.level = 0.95)

##   
## F test to compare two variances  
##   
## data: sample2 and sample3  
## F = 0.98258, num df = 499, denom df = 499, p-value = 0.8444  
## alternative hypothesis: true ratio of variances is not equal to 1  
## 95 percent confidence interval:  
## 0.8242684 1.1712861  
## sample estimates:  
## ratio of variances   
## 0.9825752

#Interpretation: The p-vaue>0.05 indicates the acceptance of H0,i.e,the variances of billing amounts from both the populations are equal. We can conclude that it doesnt matter if a patient is Diabetic or having Hypertension, there will be no difference in variability between the billing amounts of both the categories.

Question 4:

Exploring the Impact of Nitrogen Content in Soil and Crop Types on Soil pH Test the significance of the ratio of Nitrogen (N) content in soil and the type of crop label on the average pH level. If the pH level is influenced by the ratio of nitrogen content in soil and the type of crop, conduct appropriate tests to determine which specific ratios of nitrogen content in soil and crop types result in differences.

We perform the two-way ANOVA test :

There are two factors: A:ratio of Nitrogen B:Crop Type

Observations: Average pH level.

H0A: A is insignificant vs H1A: A is significant

H0B: B is insignificant vs H1B: B is significant

# We check the levels of Factor A  
Crop$N=as.factor(Crop$N)  
levels(Crop$N)

## [1] "0" "1" "2" "3" "4" "5" "6" "7" "8" "9" "10" "11"   
## [13] "12" "13" "14" "15" "16" "17" "18" "19" "20" "21" "22" "23"   
## [25] "24" "25" "26" "27" "28" "29" "30" "31" "32" "33" "34" "35"   
## [37] "36" "37" "38" "39" "40" "41" "42" "43" "44" "45" "46" "47"   
## [49] "48" "49" "50" "51" "52" "53" "54" "55" "56" "57" "58" "59"   
## [61] "60" "61" "62" "63" "64" "65" "66" "67" "68" "69" "70" "71"   
## [73] "72" "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" "83"   
## [85] "84" "85" "86" "87" "88" "89" "90" "91" "92" "93" "94" "95"   
## [97] "96" "97" "98" "99" "100" "101" "102" "103" "104" "105" "106" "107"  
## [109] "108" "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119"  
## [121] "120" "121" "122" "123" "125" "126" "127" "129" "130" "131" "132" "133"  
## [133] "134" "135" "136" "139" "140"

## We check the levels of Factor B  
  
Crop$label=as.factor(Crop$label)  
levels(Crop$label)

## [1] "apple" "banana" "blackgram" "chickpea" "coconut"   
## [6] "coffee" "cotton" "grapes" "jute" "kidneybeans"  
## [11] "lentil" "maize" "mango" "mothbeans" "mungbean"   
## [16] "muskmelon" "orange" "papaya" "pigeonpeas" "pomegranate"  
## [21] "rice" "watermelon"

# We first perform the two-way ANOVA test with interaction effect of both Factor A and FActor B for better Precision.  
two\_way=aov(ph~label\*N,data=Crop)  
two\_way

## Call:  
## aov(formula = ph ~ label \* N, data = Crop)  
##   
## Terms:  
## label N label:N Residuals  
## Sum of Squares 484.4772 47.4703 258.0412 527.1674  
## Deg. of Freedom 21 136 670 1372  
##   
## Residual standard error: 0.6198651  
## 2186 out of 3014 effects not estimable  
## Estimated effects may be unbalanced

#Interpretation: There is no interaction effect between the 2 factors. So, we again perform the ANOVA test without the interaction effect.

# Without interaction effect  
  
two\_way2=aov(ph~label+N,data=Crop)  
summary(two\_way2)

## Df Sum Sq Mean Sq F value Pr(>F)   
## label 21 484.5 23.070 59.996 <2e-16 \*\*\*  
## N 136 47.5 0.349 0.908 0.766   
## Residuals 2042 785.2 0.385   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#Interpretation: Here we can see that Factor B-Crop type is significantly affecting the average pH level Which is evident from its p-value<0.05. Hence we reject H0B and conclude that Crop type is a significant factor.Hence we perform the Tukey-HSD test to check which Crop types are affecting the ph levels in the soil.

#TukeyHSD(two\_way2, conf.level=.95)  
##Under $label, we take the combinations whose p adj<0.05. Those levels are affectin the pH of the soil.