ANOVA\_Goodness\_Fit\_independence\_Statistics for Business and Economics

Miguel\_Gomez

30/11/2021

Ejercicios libro Estadística para negocios y economía Capítulo 12 Pruebas de bondad de ajuste e independencia

Ejercicio 6

la suma de todas las proporciones hipoteticas siempre es 100% k = 4 categor?as

PASO 1 Sistema de hipótesis: H0: pi1 = 0.22, pi2 = 0.21, pi3 = 0.18, pi4 = 0.39 frente a pi1 = 0.21, pi2 = 0.30, pi3 = 0.15, pi4 = 0.34 no representa ningun cambio en ninguna proporcion H1: por lo menos alguna proporción es diferente al valor hipotetizado Nivel de significancia es del 1%.

|  |
| --- |
| DESCRIPTION PROPORCIONES HIPOTÉTICAS |
| numero de d?as categor?a proporción (1999) |
| tarjetas de credito 1 0.22 = pi1^0 |
| Tarjeta de débito 2 0.21 = pi2^0 |
| Cheque personal 3 0.18 = pi3^0 |
| Efectivo 4 0.39 = pi4^0 |

n=220  
k\_1=4 #categorias  
p\_h = c(0.22, 0.21, 0.18, 0.39) # proporciones hipotéticas  
e\_1=p\_h\*n  
#calculo prop observadas  
f\_f = c(46, 67, 33, 74) #freciencias observadas  
p\_o= f\_f/n   
  
diferencia\_porcentajes=p\_o-p\_h

1. Con α . 0.01, ¿se puede concluir que en este periodo de cuatro años, de 1999 a 2003, se ha generado un cambio en la manera en que los clientes pagan sus compras en las tiendas? ¿Cuál es el valor-p?

alpha\_1=0.05  
# calculo valor p (PASO 3)  
  
# calculo del estad?stico de prueba  
est\_prueba\_1= sum( (f\_f - e\_1)^2 / e\_1)  
# mi estadistico de prueba da 12.20635

PASO 3 establecer la región crítica

percentil\_chi\_1 = qchisq(p = alpha\_1, df = k\_1 - 1, lower.tail = F)  
# 11.34487  
  
p\_value\_1 = pchisq(q = est\_prueba\_1, df = k\_1 - 1, lower.tail = FALSE)   
# el valor p es igual a 0.006708693

Respuestas

a1). Se puede concluir que hay un cambio en como los clientes pagan sus cuentas en las tiendas pues el estadístico de prueba se encuentra en la región de rechazo que niega la igualdad entre proporciones de frecuencias observadas (datos 2003) e hipotecas (datos 1999).

A2). el p valor es = 0.006708693

1. A partir de los datos muestrales de 2003, calcule el porcentaje de uso de cada método de pago. ¿Cuál parece haber sido el principal o los principales cambios ocurridos en este periodo de cuatro años?

|  |
| --- |
| DESCRIPTION PROPORCIONES OBSERVADAS (2003) |
| numero de d?as categor?a proporción (2003) doferemcia de porcentaje con 1999 |
| tarjetas de credito 1 0.21 = pi1^0 -0.01090909 |
| Tarjeta de débito 2 0.30 = pi2^0 0.09454545 |
| Cheque personal 3 0.15 = pi3^0 -0.03000000 |
| Efectivo 4 0.34 = pi4^0 -0.05363636 |

diferencia\_porcentajes=p\_o-p\_h #diferencia entre proporciones de 1999 y 2003

respuesta b

Las formas de pago que reportan un mayor cambio son las de tarjeta debito con un incremento de 9.45% y el efectivo con una disminución del 5.36% con respecto al 1999.

1. ¿Qué porcentaje de los pagos se efectuó con tarjeta (de crédito o de débito) en 2003?

Respuesta c El 30% de los pagos en el 2003 se efectúa con tarjeta débito, mientras el 21% de los pagos se efectuó con tarjeta de crédito.

1. El interés por fuentes alternativas de energía se incrementa a medida que aumenta el precio del petróleo. En un estudio de Financial Times/Harris Poll se entrevistó a ciudadanos en seis paises para evaluar sus actitudes hacia diversas formas alternas de energía (sitio web de Harris Interactive, 27 de febrero de 2008). Los datos de la siguiente tabla representan una parte de los hallazgos de la encuesta acerca de si las personas están a favor o en contra de la construcción de nuevas plantas de energía nuclear.

|  |
| --- |
| Respuesta: |
| 1 = muy favorable |
| 2 = Más a favor que en contra |
| 3 = En contra más que a favor |
| 4 = Muy en contra |
| carrera: |
| 1 = Gran Bretaña |
| 2 = Francia |
| 3 = Italia |
| 4 = España |
| 5 = Alemania |
| 6 = Estados Unidos |

#creando los vectores   
  
c1 = c(141, 348, 381, 217)  
c2 = c(161, 366, 334, 215)  
c3 = c(298, 309, 219, 219)  
c4 = c(133, 222, 311, 443)  
c5 = c(128, 272, 322, 389)  
c6 = c(204, 326, 316, 174)  
me = matrix(c(c1, c2, c3,c4, c5, c6),  
 nrow = 4,  
 ncol = 6)  
  
n\_2 = sum(me)   
chisq.test(me)

##   
## Pearson's Chi-squared test  
##   
## data: me  
## X-squared = 425.41, df = 15, p-value < 2.2e-16

nc\_2=6  
nf\_2=4  
alpha\_2=0.05

PASO 1 proponer o armar nuestro sistema de hipótesis Sistema de hipótesis: H0: la actitud hacia la construcción de nuevas plantas de energía nuclear es independiente del país frente a H1: la actitud hacia la construcción de nuevas plantas de energía nuclear NO es independiente del país PONIENDO EL NIVEL DE SIGNIFICANCIA

alpha = 0.05 #nivel de significancia en este caso esta del 95% si quiero 97% pongo 0.03  
  
perfil\_f = rowSums(me) # perfil fila: totales por fila  
#me da la suma de cada fila en este caso son 3 filas me da 3 valores en una fila  
#es decir me da la suma de cada fila ordenados en una fila  
  
perfil\_c = colSums(me) # perfil columna: totales por columna  
#me da la suma de las columnas en este caso son 4 me deben aparecer 4 valores en forma de fila

tabla\_f = matrix(data = rep(perfil\_f, nc\_2), nrow = nf\_2, ncol = nc\_2, byrow = FALSE)# aqui creamos una matriz con los valores   
#resultantes de la suma de cada fila repetidos los prolonga por columna hasta completar   
#el arreglo matricial   
tabla\_c = matrix(data = rep(perfil\_c, nf\_2), nrow = nf\_2, ncol = nc\_2, byrow = TRUE)  
#los resultados de la sumatoria de cada columna los da en una fila y simplemente se repite   
#hasta completar el arreglo en forma matricial  
  
  
e\_2 = tabla\_f \* tabla\_c / n\_2 # frecuencias esperadas  
#dadas en forma matricial

#PASO 2   
# calculo del estadistico de prueba  
est\_prueba\_2 = sum( (me - e\_2)^2 / e\_2 )  
#est\_prueba\_2 = 425.4063  
# PASO 3  
# establecer la regi?n critica  
percentil\_chi\_2 = qchisq(p = alpha\_2, df = (nf\_2-1)\*(nc\_2-1), lower.tail = FALSE)  
#24.99579  
# valor p  
valor\_p\_2 = pchisq(q = est\_prueba\_2, df = (nc\_2 - 1)\*(nf\_2 - 1), lower.tail = FALSE)  
# 3.133381e-81

se rechaza la hipótesis nula es decir las variables no son dependientes

1. ¿Qué tan grande fue la muestra en esta encuesta? ???????????????????????????
2. Realice una prueba de hipótesis para determinar si la actitud hacia la construcción de nuevas plantas de energía nuclear es independiente del país. ¿Cuál es su conclusión? Como el estadístico de prueba chi^2 = 425.4063 pertenece a la región critica (RC) pues 425.4063 > 24.99579, entonces SI se rechaza la hipótesis nula (H0).

Como se ha rechazo la hipótesis nula, entonces se concluye que hay suficiente evidencia en la muestra para establecer que la construcción de nuevas plantas de energía nuclear NO es independiente del país.

1. Utilizando el porcentaje de respuestas “muy a favor” y “más a favor que en contra”, ¿qué país tiene la actitud más favorable hacia la construcción de nuevas plantas de energía nuclear? ¿Cuál tiene la actitud menos favorable?

Teniendo en cuenta los segmentos “muy a favor” y “más a favor que en contra” Italia, mientras el menos favorable es España.

porcentaje\_me= (me/n\_2)\*100

1. La Encuesta de satisfacción de clientes de restaurantes de Consumer Reports se basa en más de 148 599 visitas a diferentes cadenas de restaurantes de servicio completo (sitio web de Consumer Reports). Una de las variables en el estudio es el precio de los alimentos, la cantidad promedio que paga una persona por la comida y la bebida, menos la propina. Suponga que un reportero del Sun Coast Times cree que sería de interés para sus lectores realizar un estudio similar en los restaurantes ubicados en la zona del Grand Strand en Myrtle Beach, Carolina del Sur. El reportero seleccionó una muestra de ocho restaurantes de mariscos (Seafood) ocho italianos (Italian) y ocho de carnes (Steakhouse). Los datos a continuación muestran los precios de la comida en dólares de los 24 negocios muestreados. Utilice α 0.05 para probar si hay una diferencia significativa entre el precio medio de la comida en los tres tipos de restaurantes.

#creando los vectores   
  
cost = c(12, 13, 15, 17, 18, 20, 17, 24, 16, 18, 17, 26, 23, 15, 19, 18,24, 19, 23, 25, 21, 22, 27, 31)  
tipo\_rest = c('italian','italian','italian','italian','italian','italian','italian','italian',  
 'seafood','seafood','seafood','seafood','seafood','seafood','seafood','seafood'  
 ,'steakhouse','steakhouse','steakhouse','steakhouse','steakhouse','steakhouse','steakhouse','steakhouse')  
  
restaurant = matrix(c(cost, tipo\_rest),nrow = 24,ncol = 2, dimnames = list(c(),c('cost','tipo\_rest')))

Probando los principales supuestos para poder aplicar ANOVA 1. Las distribuciones deben ser normales, se hace con cualquier test de normalidad que se ajuste a la muestra.

Normalidad H\_o: Las observaciones se distribuyen normal H\_1: Las observaciones No se distribuyen de forma normal.

tapply(cost, tipo\_rest, shapiro.test)

## $italian  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.96157, p-value = 0.8249  
##   
##   
## $seafood  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.88551, p-value = 0.2124  
##   
##   
## $steakhouse  
##   
## Shapiro-Wilk normality test  
##   
## data: X[[i]]  
## W = 0.96811, p-value = 0.8828

1. Homocedasticidad es igualdad de varianzas es decir que la varianza en cada una de las poblaciones es igual. Se hace mediante el test de bartlett

Homocedasticidad H\_o: La variabilidad del costo en cada tipo de restaurante es igual H\_1: La variabilidad del costo en cada tipo de restaurante NO es igual

tapply(cost, tipo\_rest, sd)

## italian seafood steakhouse   
## 3.854496 3.703280 3.741657

#H\_o: las varianzas son id?nticas en   
#todos los grupos  
bartlett.test(cost, tipo\_rest )

##   
## Bartlett test of homogeneity of variances  
##   
## data: cost and tipo\_rest  
## Bartlett's K-squared = 0.011433, df = 2, p-value = 0.9943

bartlett.test(cost ~ tipo\_rest, restaurant)#barlett test segmanetado por tipo\_rest

##   
## Bartlett test of homogeneity of variances  
##   
## data: cost by tipo\_rest  
## Bartlett's K-squared = 0.011433, df = 2, p-value = 0.9943

Al aplicar el test de bartlett obtuvimos un valor mayor a 0.05 es decir que NO rechazamos la hipotesis nula, por lo tanto las varianzas son iguales.

1. tercer supuesto independencia de las variables, se hace con el test de durming watson

Independencia de las observaciones H\_o: Las observaciones son independientes H\_1: Las observaciones NO son independientes

library(lmtest)

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

dwtest(cost ~ tipo\_rest)

##   
## Durbin-Watson test  
##   
## data: cost ~ tipo\_rest  
## DW = 1.5168, p-value = 0.04724  
## alternative hypothesis: true autocorrelation is greater than 0

p valor mayor que 0.05 por lo tanto no rechazo la hipotesis es decir las observaciones son independientes.

Aplicacion del ANOVA

med\_ita=mean(cost[tipo\_rest == "italian"])  
med\_seaf=mean(cost[tipo\_rest == "seafood"])  
med\_stk=mean(cost[tipo\_rest == "steakhouse"])  
med\_c=mean(cost)  
tapply(cost, tipo\_rest, mean)

## italian seafood steakhouse   
## 17 19 24

var\_x=((med\_ita-med\_c)^2+(med\_seaf-med\_c)^2+  
 (med\_stk-med\_c)^2)/(3-1)  
  
  
n=length(cost[tipo\_rest == "italian"])  
var=n\*var\_x  
  
var\_ita=var(cost[tipo\_rest == "italian"])  
var\_seaf=var(cost[tipo\_rest == "seafood"])  
var\_stk=var(cost[tipo\_rest == "steakhouse"])  
var\_dentro=(var\_ita+var\_seaf+var\_stk)/3  
prueba=tapply(cost, tipo\_rest, var)  
vardentro=sum(prueba)/3  
  
  
ANOVA1=aov(formula=cost ~ tipo\_rest)  
summary(ANOVA1)

## Df Sum Sq Mean Sq F value Pr(>F)   
## tipo\_rest 2 208 104.00 7.329 0.00385 \*\*  
## Residuals 21 298 14.19   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#las medias son diferentes porque el valor p es menor a 0.05  
SCTR=n\*((med\_ita-med\_c)^2+(med\_seaf-med\_c)^2+  
 (med\_stk-med\_c)^2)  
  
SCE=(n-1)\*(var\_ita+var\_seaf+var\_stk)

Gráfica de diferencia de medias

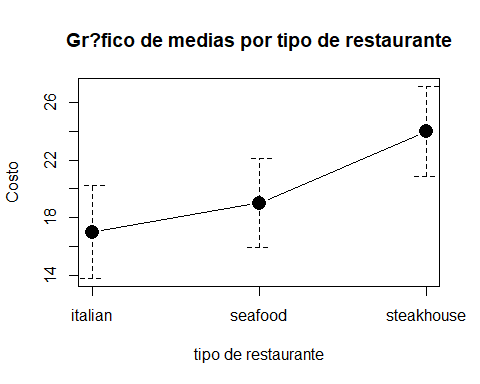
## grafica de diferencia de medias   
library(RcmdrMisc)

## Loading required package: car

## Loading required package: carData

## Loading required package: sandwich

windows()  
plotMeans(cost, tipo\_rest, error.bars = "conf.int",   
 xlab = "tipo de restaurante", main = "Gr?fico de medias por tipo de restaurante",   
 ylab = "Costo")



Aplicando prueba de Tukey

TukeyHSD(ANOVA1, "tipo\_rest")

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = cost ~ tipo\_rest)  
##   
## $tipo\_rest  
## diff lwr upr p adj  
## seafood-italian 2 -2.7475254 6.747525 0.5474395  
## steakhouse-italian 7 2.2524746 11.747525 0.0035046  
## steakhouse-seafood 5 0.2524746 9.747525 0.0378404

De acuerdo con los resultados de la prueba Tukey, se puede concluir que las medias entre los costos de seafood-italian no son diferentes, mientras que steakhouse -italian y steakhouse-seafood si difieren.