Universidad del Valle de Guatemala
Departamento de Matemática
Licenciatura en Matemática Aplicada

Estudiante: Rudik Roberto Rompich
E-mail: rom19857@uvg.edu.gt
Carné: 19857

MM2040 - Estadística 2 - Catedrático: Eugenio Aristondo
24 de mayo de 2021

Tarea 5

1. Capítulo 12

1.1. Problema 6

Ejericicio extra

Se resolvió en clase.

La American Bankers Association recoge datos sobre el uso de tarjetas de crédito o débito, cheques personales y efectivo para el pago de compras en tienda (The Wall Street Journal, 16 de diciembre de 2003). En 1999 los datos encontrados fueron los siguientes.

Compras en tienda	Porcentaje
Tarjeta de crédito	22
Tarjeta de débito	21
Cheque personal	18
Efectivo	39

En una muestra tomada en 2003 se encontró que de cada 220 compras en tienda, en 46 se usó tarjeta de crédito, en 67 tarjeta de débito, en 33 cheque personal y en 74 pago en efectivo.

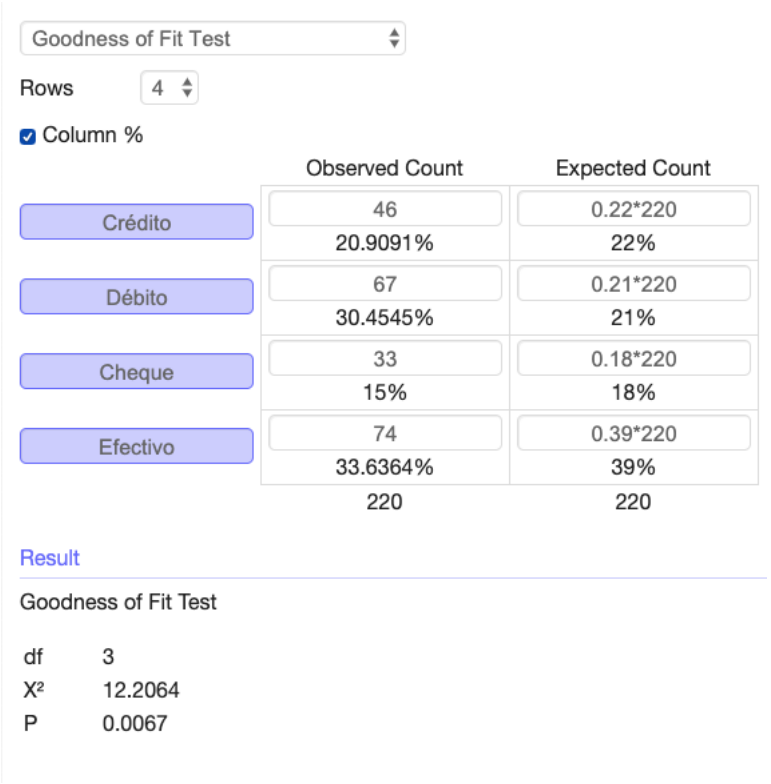
1. a) Con $\alpha = 0.01$, ¿se puede concluir que en este periodo de cuatro años, de 1999 a 2003, se ha generado un cambio en la manera en que los clientes pagan sus compras en las tiendas? ¿Cuál es el valor-p?

Solución. Se considera la prueba de hipótesis:

H_0 : la población tiene una distribución multinomial con la probabilidad específica de cada una de las k categorías.

H_a : la población no tiene una distribución multinomial con la probabilidad específica de cada una de las k categorías.

Se considera el análisis hecho en Geogebra:



a) El valor-p es 0.0067.

b) Por el método del valor-p: 0,0067 $<$ 0,01; por lo que se puede concluir que de 1900 a 2003 sí se ha generado un cambio en el que los clientes pagan sus compras.

□

2. b) 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 período de cuatro años?

Solución. Los porcentajes se observan en la imagen de arriba. El cambio principal es el pago con tarjeta de débito, que subió de 21 % a aproximadamente 30.45 % □

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

Solución.

$$20{,}9091\,\% + 30{,}4545\,\% = 51{,}36\,\%$$

□

1.2. Problema 12

Visa Card USA estudió la frecuencia con que los consumidores de diversos rangos de edad usan tarjetas plásticas (de crédito o de débito) para pagar sus compras (Associated Press, 16 de enero de 2006). A continuación se presentan los datos muestrales de 300 clientes divididos en cuatro grupos de edad.

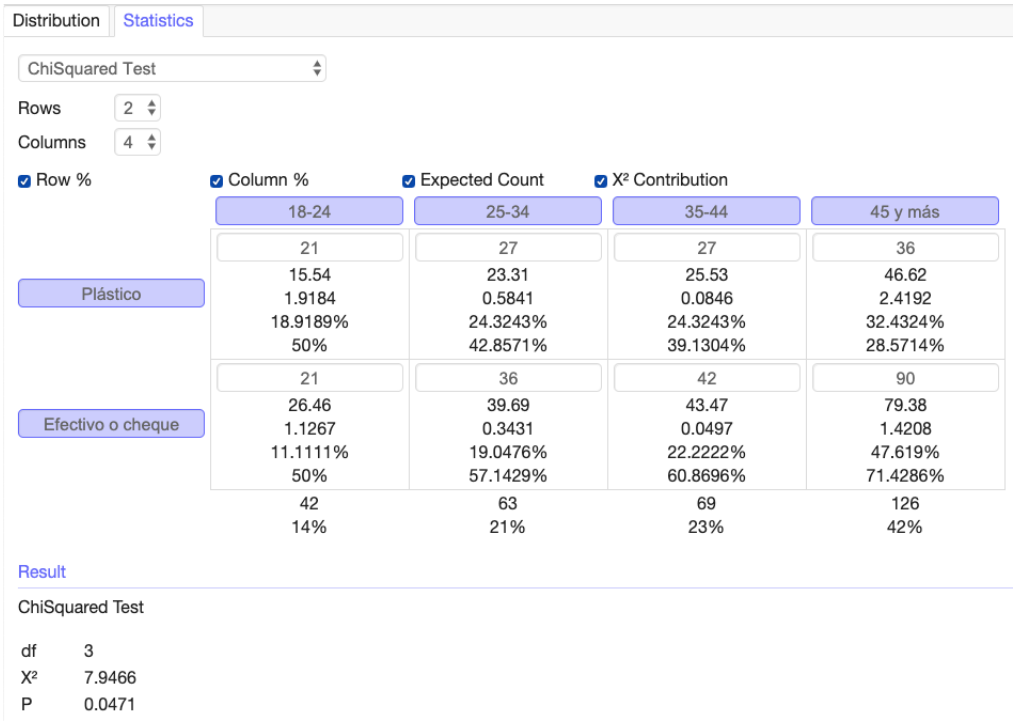
	Grupo de edad			
Forma de pago	**18–24**	**25–34**	**35–44**	**45 y más**
Plástico	21	27	27	36
Efectivo o cheque	21	36	42	90

1. a) Pruebe la independencia entre el método de pago y el grupo de edad. ¿Cuál es el valor-p? Usando $\alpha = 0.05$, ¿cuál es su conclusión?

   Solución. Se considera la prueba de hipótesis:

   H_0 : la variable de las columnas es independiente de la variable de las filas.
   H_a : la variable de las columnas no es independiente de la variable de las filas.

   Se considera el análisis hecho en Geogebra:



   a) El valor-p es 0.0471.

   b) Usando $\alpha = 0{,}05$, por la prueba del valor-p $0{,}0471 < 0{,}05$ por lo que H_0 se rechaza. Hay evidencia suficiente para afirmar que grupo de edad no es independiente de las formas de pago.

□

2. b) Si la forma de pago y el grupo de edad no son independientes, ¿qué observación puede formular acerca de la diferencia en el uso del plástico en los diversos grupos de edad?

   Solución. Las conclusiones son diversas, pero se puede observar que entre el grupo de personas es más joven, existe una preferencia en el plástico. □

3. c) ¿Qué consecuencias tiene este estudio para empresas como Visa, MasterCard y Discover?

   Solución. Las consecuencias son variadas, pero pareciera indicar que estas compañías deberían enfocar en el uso del plástico a los grupos más viejos. □

1.3. Problema 15

FlightStats, Inc. recolecta datos sobre el número de vuelos programados y realizados en los principales aeropuertos de Estados Unidos. Sus datos indican que 56 % de los vuelos programados en los aeropuertos de Newark, La Guardia y Kennedy se efectuaron durante una tormenta de nieve que duró tres días (The Wall Street Journal, 21 de febrero de 2006). Todas las aerolíneas afirman que operan siempre dentro de parámetros de seguridad establecidos: si las condiciones son muy malas, no vuelan. Los datos en la tabla superior de la siguiente página presentan una muestra de 400 vuelos programados durante tormentas de nieve.

	Aerolínea				
¿Voló?	**American**	**Continental**	**Delta**	**United**	**Total**
Sí	48	69	68	25	210
No	52	41	62	35	190

Use la prueba de independencia ji-cuadrada con un nivel de significancia de 0.05 para analizar estos datos. ¿Cuál es su conclusión? ¿Qué aerolínea elegiría para volar en condiciones de tormentas de nieve semejantes? Explique.

Solución. Primero se consideran las hipótesis:

H_0 : la variable de las columnas es independiente de la variable de las filas.

H_a : la variable de las columnas no es independiente de la variable de las filas.

Ahora bien, se considera el análisis hecho en Geogebra:

ChiSquared Test

Rows 2

Columns 4

Row % Column % Expected Count X² Contribution

	American	Continental	Delta	United
Sí	48 52.5 0.3857 22.8571% 48%	69 57.75 2.1916 32.8571% 62.7273%	68 68.25 0.0009 32.381% 52.3077%	25 31.5 1.3413 11.9048% 41.6667%
No	52 47.5 0.4263 27.3684% 52%	41 52.25 2.4222 21.5789% 37.2727%	62 61.75 0.001 32.6316% 47.6923%	35 28.5 1.4825 18.4211% 58.3333%
	100 25%	110 27.5%	130 32.5%	60 15%

Result

ChiSquared Test

df	3
X²	8.2515
P	0.0411

1. El valor-p es 0.0411.

2. Usando $\alpha = 0{,}05$, por la prueba del valor-p $0{,}0411 < 0{,}05$ por lo que H_0 se rechaza. Hay evidencia suficiente para afirmar que la aerolínea no es independiente de si voló o no voló durante una tormenta de nieve.

3. En un primera vistazo, elegiría entre Delta y Continental; sin embargo, me decidiría por Continental; ya que tiene una menor cantidad de vuelos que no volaron comparado a Delta.

□

1.4. Problema 20

A continuación se presenta el número de ocurrencias por periodo y su frecuencia observada. Use $\alpha = 0.05$ y la prueba de bondad de ajuste para determinar si estos datos se ajustan a una distribución de Poisson.

Número de ocurrencias	Frecuencia observada
0	39
1	30
2	30
3	18
4	3

Solución. Primero, se hacen los cálculos pertinentes con Excel, usando la distribución de Poisson definida como:

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}$$

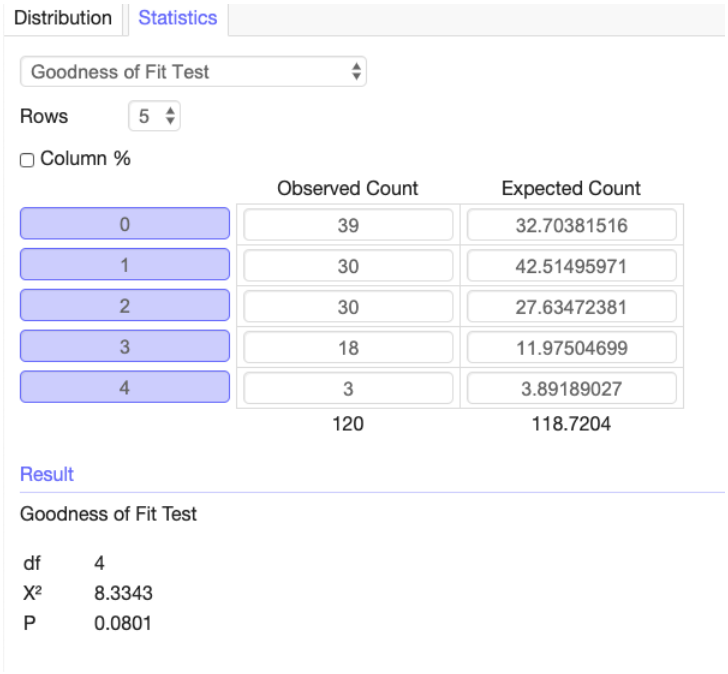
Por lo cual, se tiene:

Número de Ocurrencias	Frecuencia Observada	Llegadas	Propabilidad	Frecuencia Esperada
0	39	0	0.272531793	32.70381516
1	30	30	0.354291331	42.51495971
2	30	60	0.230289365	27.63472381
3	18	54	0.099792058	11.97504699
4	3	12	0.032432419	3.89189027
Datos	120			
Media	1.3			

Definimos las hipótesis:

H_0 : la población tiene una distribución de Poisson.
H_a : la población no tiene una distribución de Poisson.

Usando Geogebra:



1. El valor-p es 0.0801.
2. Por el método del valor-p: 0,0801 > 0,05; por lo que H_0 no se rechaza. Por lo tanto, podemos concluir que los datos sí se ajustan a una distribución de Poisson.

□

1.5. Problema 25

Use $\alpha = 0.01$ y realice una prueba de bondad de ajuste para comprobar si la siguiente muestra fue tomada de una distribución normal.

55	86	94	58	55	95	55	52	69	95	90	65	87	50	56
55	57	98	58	79	92	62	59	88	65					

Una vez realizada la prueba de bondad de ajuste, elabore un histograma con todos estos datos. ¿Este gráfico respalda la conclusión a la que se llegó con la prueba de bondad de ajuste? (Nota. $\overline{x} = 71$ y $s = 17$.)

Solución. Debido a la falta de instrucciones del problema, se propondrá distribuir la muestra en intervalos de 20 %, es decir que tendremos 5. Se propone utilizar Excel para esta tarea, con el comando

$$NORM.INV()$$

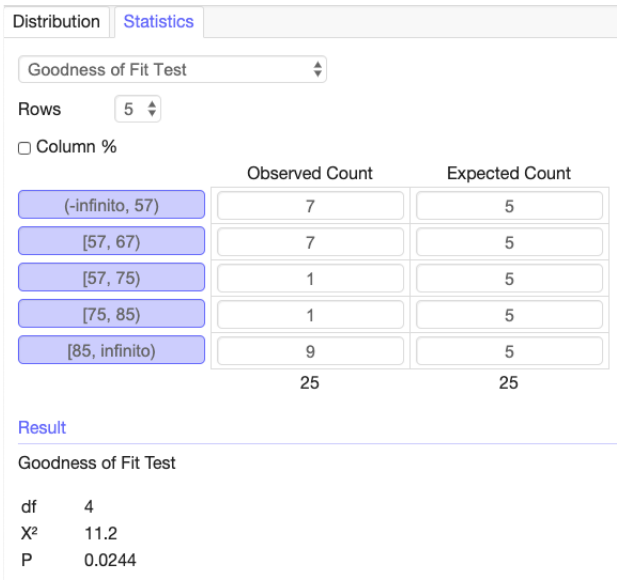
Por lo que tenemos:

Porcentaje	Separadores
0.2	57
0.4	67
0.6	75
0.8	85
Media	71
DV	16.99509733

Ahora bien, considerando el análisis de histograma de Excel, tenemos:

Intervalo	*Frecuencia observada*	*Frecuencia esperada*
(-infinito, 57)	7	5
[57,67)	7	5
[67-75)	1	5
[75,85)	1	5
[85,infinito)	9	5

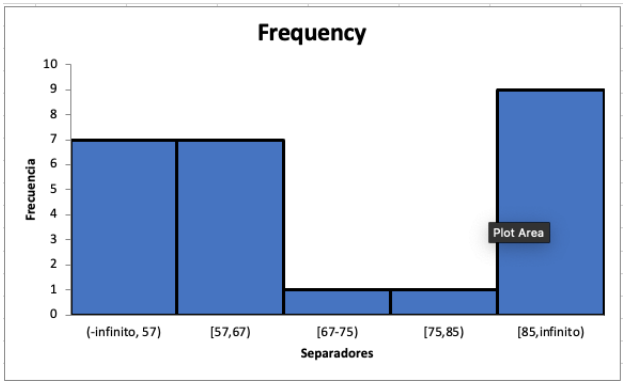
Usando Geogebra:



Se plantean las hipótesis:

$$H_0 : \text{la población tiene una distribución normal.}$$
$$H_a : \text{la población no tiene una distribución normal.}$$

1. El valor-p es 0.0244.
2. Por lo tanto, considerando la prueba del valor-p: 0.0244>0.01. Es decir que H_0 se acepta, entonces los datos tienen una distribución normal con una significancia de 0.01 .
3. El histograma:



El histograma no es de mucha ayuda, ya que a simple vista hace pensar que los datos no tienen un distribución normal; pero la conclusión de que los datos tienen una distribución normal es debido a la significancia que nos da el problema (α =0.01); ya que con una significancia de 0.05 sí hubiera sido posible rechazar la H_0.

□

Segunda parte

La segunda parte se hizo completamente en R-Markdown.

Paquetes

```
library(readxl)
library(gsubfn)
library(stats)
library(BSDA)
library(ggplot2)
library(cowplot)
```

Funciones Prueba de Signos

```
clasificar <- function(datos, mediana){
longitud <- length(datos)
positivos <- 0
negativos <- 0
iguales <- 0

for (i in datos){
  if (i< mediana){
    negativos<- negativos+ 1
 }
  else if (i> mediana){
    positivos<- positivos + 1}
  else{
    iguales <- iguales +1
  }
}
positivos <- positivos
muestras <- longitud-iguales
return(list(positivos,muestras))
}
```

Capítulo 19

Ejercicio 6

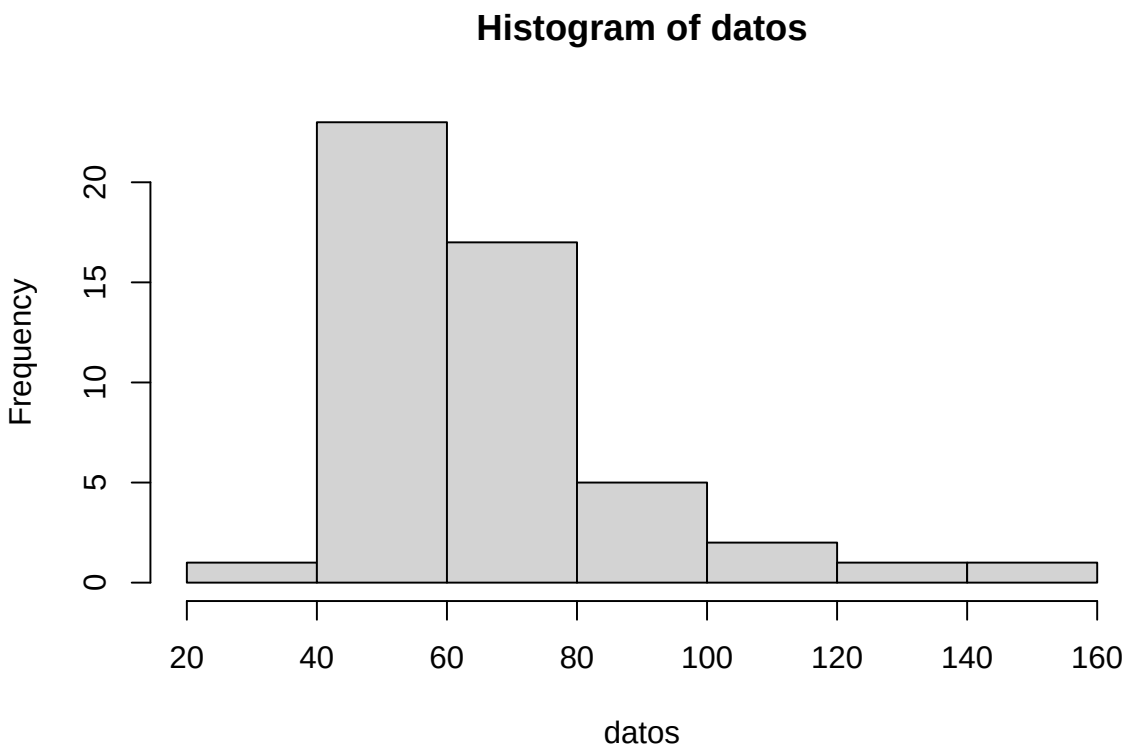
El ingreso mediano anual de las familias que viven en Estados Unidos es de $56 200 (The New York Times Almanac, 2008). Se presentan los ingresos anuales en miles de dólares para una muestra de 50 familias que viven en Chicago, Illinois. Utilice los datos de la muestra para ver si se puede concluir que las familias que viven en Chicago tienen un ingreso mediano anual de más de $56 200. Utilice $\alpha = 0.05$. ¿Cuál es su conclusión?

Solución.

El problema se plantea de la siguiente forma:

$$H_0 : \tilde{\mu} \leq \$56200$$
$$H_a : \tilde{\mu} > \$56200$$

```
#Base de datos
DB <-read_excel("/Users/rudiks/Desktop/Excel Files-3/Ch 19 Nonparametric/ChicagoIncome.xlsx")
datos <-DB$`Income ($1000)`
hist(datos)
```




```
#La mediana que queremos calcular
mediana <- 56.2
#La función para clasificar los datos
datos2 <- clasificar(datos, mediana)
positivos <- datos2[[1]]
muestras <- datos2[[2]]
#El test binomial, con alternative= "greater" porque nos piden calcular si la media es más grande.
binom.test(positivos,muestras, p=0.5, alternative="greater", conf.level=0.95)
```

```
##
##  Exact binomial test
##
## data:  positivos and muestras
## number of successes = 31, number of trials = 48, p-value = 0.02973
## alternative hypothesis: true probability of success is greater than 0.5
## 95 percent confidence interval:
##  0.5173383 1.0000000
## sample estimates:
## probability of success
##              0.6458333
```

Nos percatamos que el problema se pudo haber tratado con una distribución normal por el **Teorema del Límite Central**, sin embargo, se propuso usar la distribución binormal por motivos ilustrativos. Usando la distribución normal ($H_0 : p = 0.50$), se tiene:

$$\mu = 0.5n$$
$$\sigma = \sqrt{0.25n}$$

```r
media= 0.5*muestras
dv = sqrt(0.25*muestras)
#El test normal, con lower.tail= FALSE porque nos piden calcular si la media es más grande.
pnorm(positivos,media, dv, lower.tail = FALSE)
```

```
## [1] 0.02165407
```

- Binomial: $0.02973 \leq 0.05$
- Normal: $0.0.02165407 \leq 0.05$

```r
#Análogamente, el problema simplemente se pudo resolver con el SIGN.TEST.
SIGN.test(datos,md=mediana, alternative = "greater", conf.level = 0.95)
```

```
##
##  One-sample Sign-Test
##
## data:  datos
## s = 31, p-value = 0.02973
## alternative hypothesis: true median is greater than 56.2
## 95 percent confidence interval:
##  56.39491      Inf
## sample estimates:
## median of x
##       60.55
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level  L.E.pt U.E.pt
## Lower Achieved CI     0.9405 56.5000    Inf
## Interpolated CI       0.9500 56.3949    Inf
## Upper Achieved CI     0.9675 56.2000    Inf
```

Por lo que podemos concluir, basado en la prueba del valor-p que existe suficiente evidencia con una significancia de $\alpha = 0.05$ que las familias que viven en Chicago tienen un ingreso mediano anual de más de \$56 200. □

Ejercicio 14

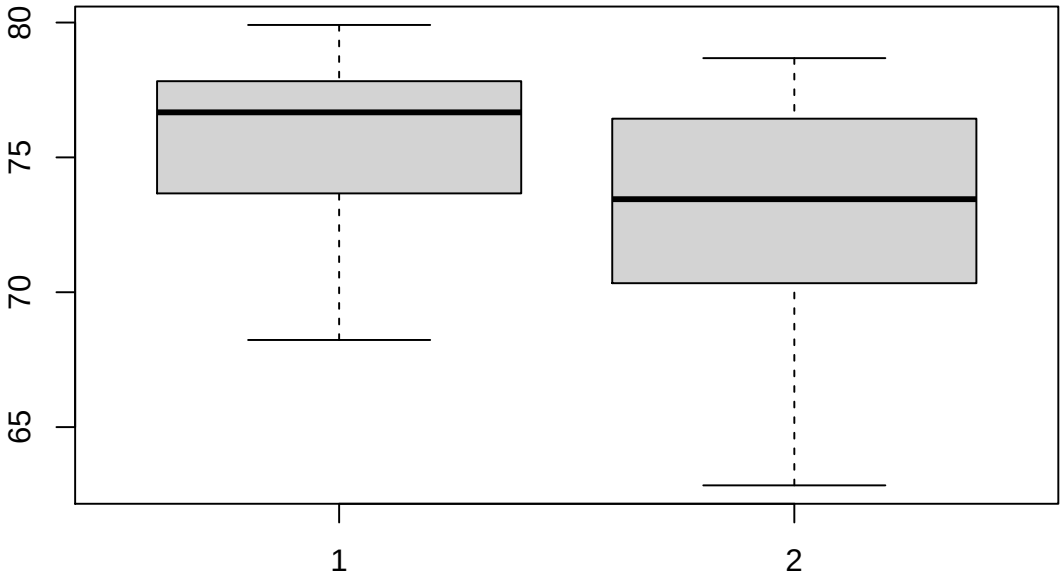
Los porcentajes de llegadas puntuales (Percent on Time) de vuelos en 2006 y 2007 fueron recabados aleatoriamente de 11 aeropuertos (Airport). Los datos se muestran en la parte superior de la página siguiente (página web de Research and Innovative Technology Administration, 29 de agosto de 2008). Utilice $\alpha = 0.05$ como nivel de significancia para probar la hipótesis de que no hay diferencia entre las medianas del porcentaje de llegadas a tiempo para los dos años. ¿Cuál es su conclusión?

Solución. El problema se plantea:

$$H_0 : \tilde{\mu}_{2006} = \tilde{\mu}_{2007}$$
$$H_a : \tilde{\mu}_{2006} \neq \tilde{\mu}_{2007}$$

Es decir, tenemos un problema de dos colas. Analizamos los datos presentandos:

```r
DB <- read_excel("/Users/rudiks/Desktop/Excel Files-3/Ch 19 Nonparametric/OnTime.xlsx")
datos2006 <- DB$'2006'
datos2007<- DB$'2007'
boxplot(datos2006,datos2007)
```



Se propone tratar el problema con el método de rangos con signo de Wilcoxon.

```
wilcox.test(datos2006,datos2007,paired = TRUE,  conf.int = 0.95)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  datos2006 and datos2007
## V = 61, p-value = 0.009766
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  0.965 4.320
## sample estimates:
## (pseudo)median
##         2.7475
```

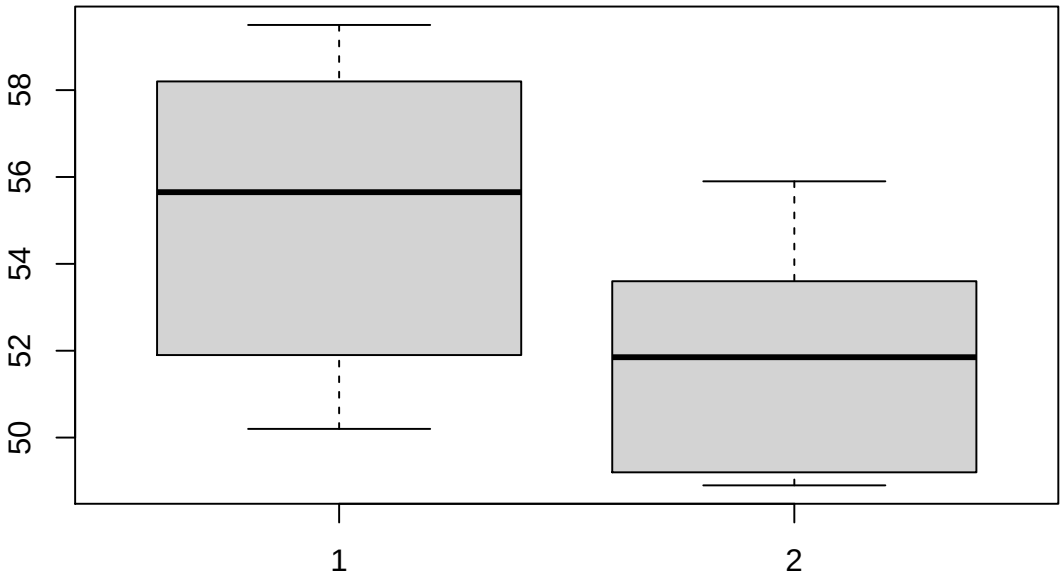
$\therefore$ Por la prueba del valor-p podemos concluir que $0.009766 < 0.05$; existe evidencia suficiente con una significancia de 0.05 para afirmar que las medianas del porcentaje de llegadas a tiempo para los dos años difieren. $\square$

Ejercicio 19

Las siguientes son muestras de los sueldos iniciales anuales de personas que ingresan a las carreras de contador público (Public Accountant) y de planificador financiero (Financial Planner). Los sueldos anuales se presentan en miles de dólares.

```
AcctPlanners <- read_xlsx("/Users/rudiks/Desktop/Excel Files-3/Ch 19 Nonparametric/AcctPlanners.xlsx")
public <- AcctPlanners$`Public Accountant`
financial <- AcctPlanners$`Financial Planner`

boxplot(public, financial)
```



a) Utilice 0.05 como nivel de significancia y la prueba de hipótesis de que no existe diferencia entre el sueldo inicial anual de los contadores públicos y los planificadores financieros. ¿Cuál es su conclusión?

Solución. Considerando:

$$H_0 : \tilde{\mu}_{contadores} = \tilde{\mu}_{planificadores}$$
$$H_a : \tilde{\mu}_{contadores} \neq \tilde{\mu}_{planificadores}$$

Se propone el test de Wilcox:

```
wilcox.test(public,financial,paired = TRUE, exact=FALSE, conf.int = 0.95)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  public and financial
## V = 55, p-value = 0.005889
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  1.749950 5.599947
## sample estimates:
## (pseudo)median
##       3.199958
```

$\therefore$ Por la prueba del valor-p podemos concluir que $0.005889 < 0.05$; existe evidencia suficiente con una significancia de 0.05 para afirmar que el sueldo de contadores públicos y planificadores financiero difieren. $\square$

b) ¿Cuáles son las medianas de los sueldos anuales de la muestra para las dos profesiones?

Contadores

```
median(public)
```

```
## [1] 55.65
```

Planificadores

```
median(financial)
```

```
## [1] 51.85
```

Ejercicio 29

La revista Condé Nast Traveler realiza un estudio anual entre sus lectores con el fin de calificar los 80 cruceros más importantes del mundo (Condé Nast Traveler, febrero de 2008). Con 100 como la calificación (Rating) más alta posible, se lista la siguiente puntuación global para una muestra de los barcos (Ship) de Holland America, Princess y Royal Caribbean. Utilice la prueba de Kruskal-Wallis con $\alpha = 0.05$ para determinar si en general las calificaciones entre las tres líneas de cruceros difieren significativamente. ¿Cuál es su conclusión?

Solución. Se comienza definiendo el problema:

$$H_0 : \tilde{\mu}_H = \tilde{\mu}_P = \tilde{\mu}_R$$
$$H_a : \tilde{\mu}_H \neq \tilde{\mu}_P \neq \tilde{\mu}_R$$

```
DB <- read_excel("/Users/rudiks/Desktop/Excel Files-3/Ch 19 Nonparametric/CruiseShips.xlsx")
holland <- DB$`Holland America`
princess <- DB$Princess
royal <- DB$`Royal Caribbean`

kruskal.test(list(holland, princess,royal))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(holland, princess, royal)
## Kruskal-Wallis chi-squared = 4.1925, df = 2, p-value = 0.1229
```

$\therefore$ Por la prueba del valor-p podemos concluir que 0.1229>0.05; existe evidencia suficiente con una significancia de 0.05 para afirmar que las medias de los cruceros NO difieren. □

Ejercicio 36

A continuación se presenta la clasificación de una muestra de golfistas (Golfer) profesionales respecto del driving distance como del putting. ¿Cuál es la correlación por rangos entre el driving distance y el putting para estos jugadores? Utilice 0.10 como nivel de significancia y pruebe la significancia de la correlación por rangos.

Solución. Se propone:

$$H_0 : \rho_s = 0$$
$$H_a : \rho_s \neq 0$$

```
DB<- read_xlsx("/Users/rudiks/Desktop/Excel Files-3/Ch 19 Nonparametric/ProGolfers.xlsx")
driving <- DB$`Driving Distance`
putting <- DB$Putting

cor.test(driving,putting, method="spearman", conf.level = 0.95)
```

```
##
##  Spearman's rank correlation rho
##
## data:  driving and putting
## S = 282, p-value = 0.02751
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.7090909
```

∴: el coeficiente: ρ=-0.7090909 Por otra parte, por la prueba del valor-p podemos concluir que 0.02751<0.1; existe evidencia suficiente con una significancia de 0.05 para afirmar que la correlación difiere. □