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

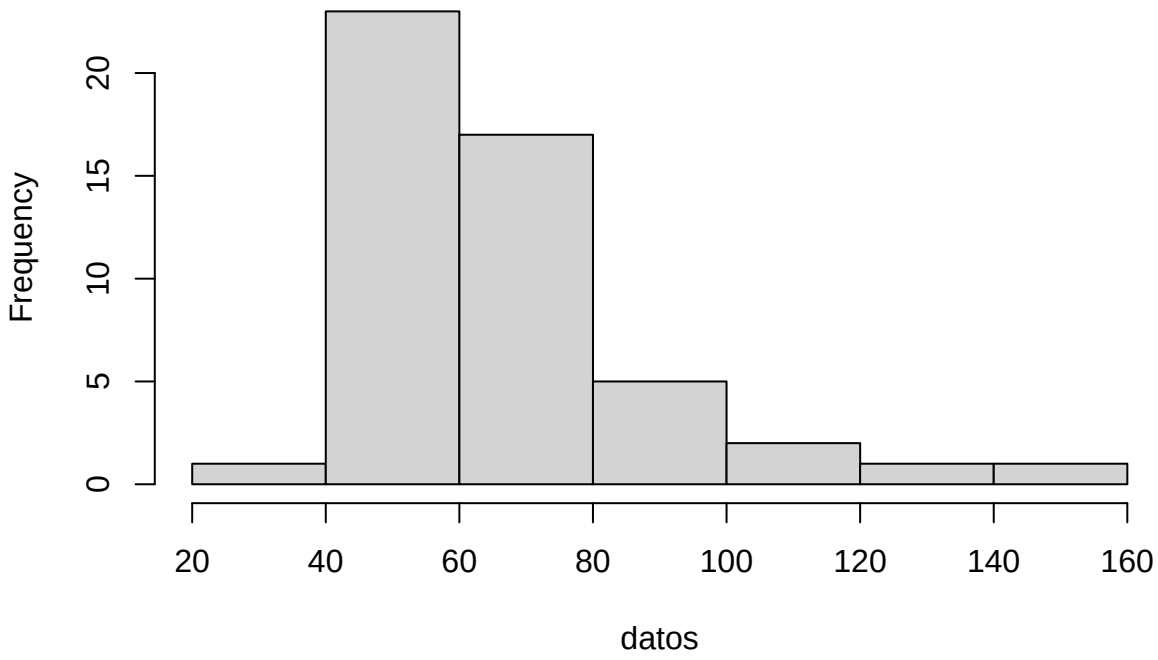
El ingreso mediano anual de las familias que viven en Estados Unidos es de $56 200 (The New York Times Almanac, 2008). Se presentan los ingresos anuales en miles de dólares para una muestra de 50 familias que viven en Chicago, Illinois. Utilice los datos de la muestra para ver si se puede concluir que 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)
```




```r
#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:

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

```
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: $0.0.02165407 \leq 0.05$

```
#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

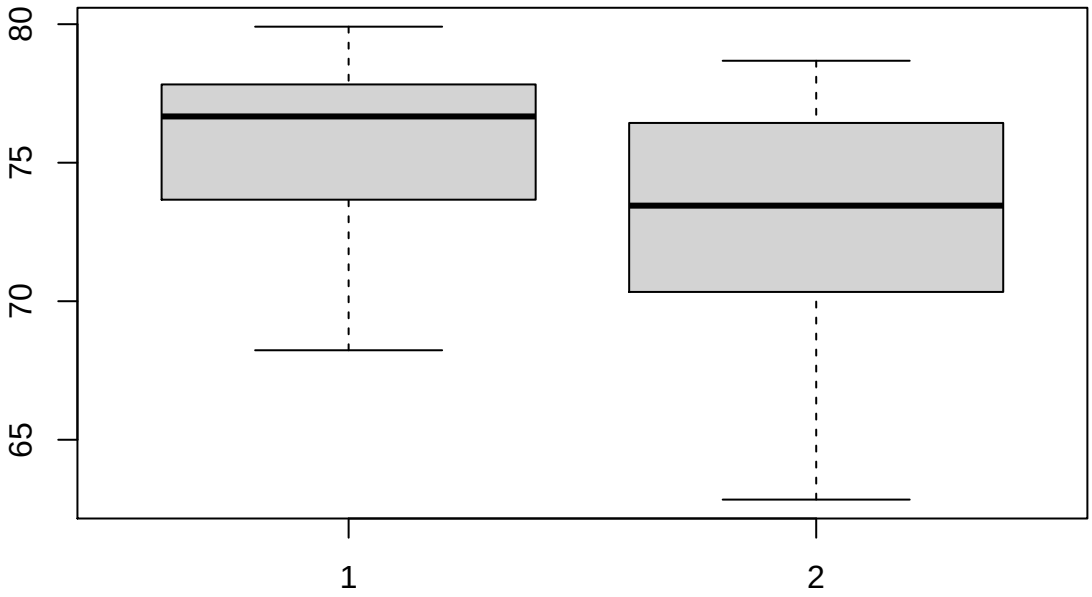
Los porcentajes de llegadas puntuales (Percent on Time) de vuelos en 2006 y 2007 fueron recabados aleatoriamente de 11 aeropuertos (Airport). Los datos se muestran en la parte superior de la página siguiente (página web de Research and Innovative Technology Administration, 29 de agosto de 2008). Utilice $\alpha = 0.05$ 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:

```
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
```

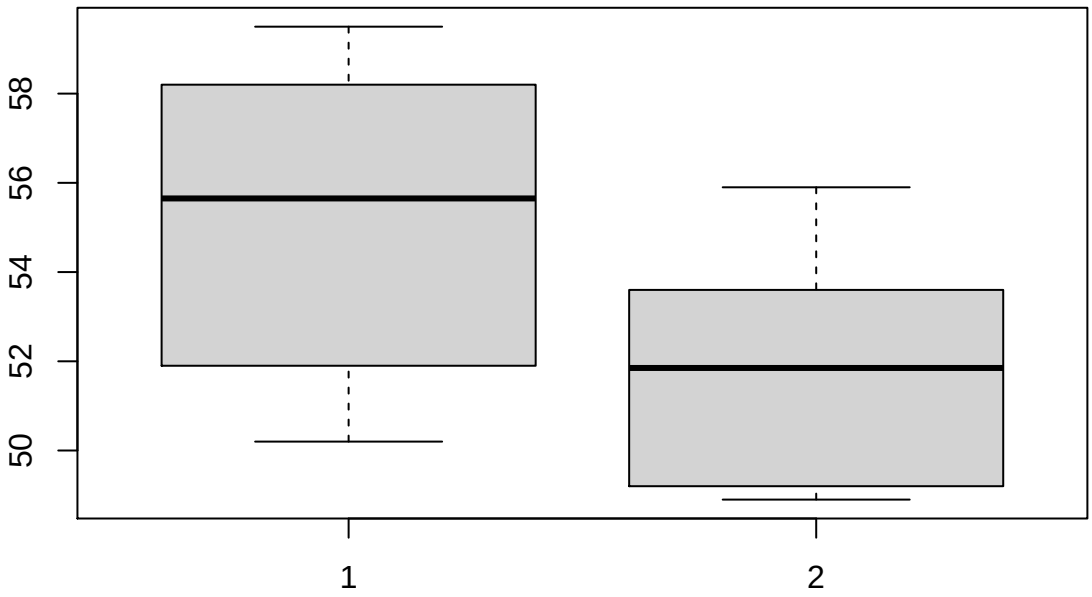
$\therefore$ Por la prueba del valor-p podemos concluir que $0.009766 < 0.05$; existe evidencia suficiente con una significancia de 0.05 para afirmar que las medianas del porcentaje de llegadas a tiempo para los dos años difieren. $\square$

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. □

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
```

$\therefore$: 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. $\square$