Ejercicios clases 11

Test de Kruskal-Wallis

Joaquin Cavieres G.

Ejercicio 1

En un estudio se comparan las cosechas de paltas 3 regiones distintas. ¿Existen diferencias significativas dependiendo de las condiciones climáticas de cada zona?

```
data = data.frame(zonas = c(rep("zona1", 18), rep("zona2", 18), rep("zona3", 18)),
            n_paltas =  c(1, 2, 3, 4, 5, 6, 7, 8, 9, 16, 27, 28, 29, 30, 51, 52, 53,
                          342, 40, 41, 42, 43, 44, 45, 46, 47, 48, 67, 88, 89, 90,
                          91,92, 93, 94, 293, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28,
                          25, 36, 37, 58, 59, 60, 71, 72))
head(data)
```

```
##   zonas n_paltas
## 1 zona1        1
## 2 zona1        2
## 3 zona1        3
## 4 zona1        4
## 5 zona1        5
## 6 zona1        6
```

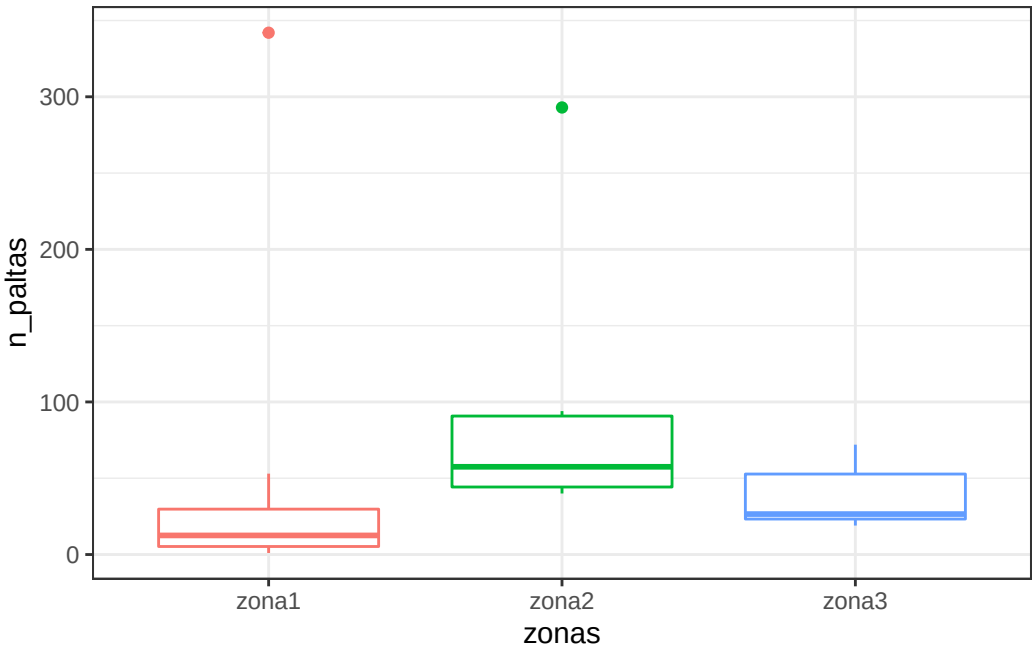
```
# Vemos la mediana para cada zona
aggregate(n_paltas ~ zonas, data = data, FUN = median)
```

```
##   zonas n_paltas
## 1 zona1     12.5
## 2 zona2     57.5
## 3 zona3     26.5
```

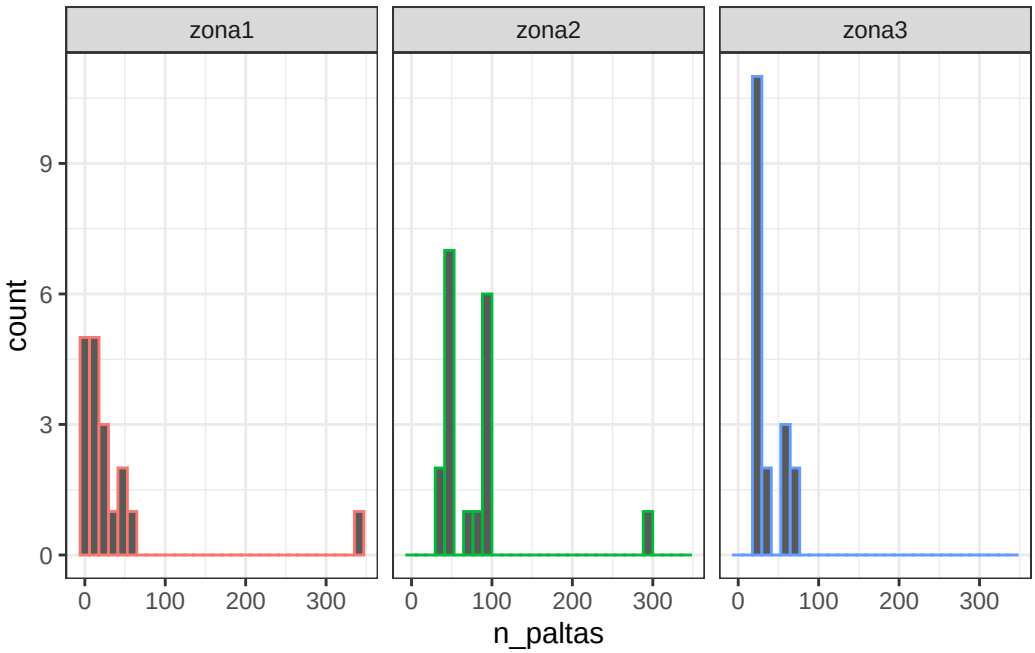
```
# Vemos la mediana desviacion estandar de cada zona
aggregate(n_paltas ~ zonas, data = data, FUN = sd)
```

```
##   zonas n_paltas
## 1 zona1 78.10637
## 2 zona2 58.41750
## 3 zona3 18.59097
```

```r
# Visualización
library(ggplot2)
ggplot(data = data, mapping = aes(x = zonas, y = n_paltas, colour = zonas)) +
  geom_boxplot() +
  theme_bw() +
  theme(legend.position = "none")
```




```r
ggplot(data = data, mapping = aes(x = n_paltas, colour = zonas)) +
  geom_histogram() +
  theme_bw() +
  facet_grid(. ~ zonas) +
  theme(legend.position = "none")
```



- El historgrama nos muestra que las muestras no se distribyen Normal, por tanto no podemos aplicar un test de ANOVA.
- Las tres muestras presentan asimetría hacia la derecha
- El Test de Kruskal-Wallis es la opción más adecuada para este caso particular (sin comprobar por el momento homogeneidad de la varianza)
- Otra alternativa serían las técnicas de resampling.

Verificamos si existe homogeneidad de la varianza entre los grupos

```
library(car)
leveneTest(n_paltas ~ zonas, data = data, center = "median")
```

```
## Levene's Test for Homogeneity of Variance (center = "median")
##       Df F value Pr(>F)
## group  2  0.7929  0.458
##       51
```

No hay evidencias en contra de la homogeneidad de varianzas ya que $Pr(> F)$0,458

Nota: Test de Levene - Se caracteriza, además de por poder comparar 2 o más poblaciones, por permitir elegir entre diferentes estadísticos de centralidad :mediana (por defecto), media, media truncada. Esto es importante a la hora de contrastarla homocedasticidad dependiendo de si los grupos se distribuyen de forma normal o no.

Test de Kruskal-Wallis

```
kruskal.test(n_paltas ~ zonas, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  n_paltas by zonas
## Kruskal-Wallis chi-squared = 19.964, df = 2, p-value = 4.623e-05
```

Se encuentra significancia en la diferencia de al menos 2 grupos.

¿ Cuales son los grupos difieren entre sí?

Análisis post-hoc

Utilizamos el método de Holm por que es más flexible que el de Bonferroni(se deben cumplir muchos supuestos pero es más exacto)

```
pairwise.wilcox.test(x = data$n_paltas, g = data$zonas, p.adjust.method = "holm" )
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  data$n_paltas and data$zonas
##
##       zona1   zona2
## zona2 0.00029 -
## zona3 0.04795 0.00058
##
## P value adjustment method: holm
```

En este caso particular todas las posibles combinaciones para las 'zonas' mediante el método de Holm nos otorga un valor menor al p-value.

Ejercicio 2

Thomson y Short (1969) han evaluado la eficiencia mucociliar a partir de la velocidad de eliminación del polvo en sujetos normales, sujetos con enfermedad obstructiva de las vías respiratorias y sujetos con asbestosis. La siguiente tabla se basa en un subconjunto de los datos de Thomson-Short. Los rangos conjuntos (r_{ij}) de las observaciones se dan entre paréntesis después de que los valores de los datos y las sumas de los rangos de tratamiento (R_1, R_2 y R_3) se proporcionan en la parte inferior de las columnas (Hollander, M (2003))

Sujetos con:		
Sujetos normales	Enfermedad respiratoria	Asbestosis
2.9 (8)	3.8 (13)	2.8 (7)
3.0 (9)	2.7 (6)	3.4 (11)
2.5 (4)	4.0 (14)	3.7 (12)
2.6 (5)	2.4 (3)	2.2 (2)
3.2 (10)		2.0 (1)
$R_1 = 36$	$R_2 = 36$	$R_3 = 33$

Para comenzar, y a modo de ejemplo, hacemos:

```
#install.packages("NSM3")
library(NSM3)
cKW(0.0502, c(5, 4, 5), "Exact")
```

```
## Group sizes: 5 4 5
## For the given alpha=0.0502, the upper cutoff value is Kruskal-Wallis H=5.64285714,
## with true alpha level=0.0502
```

```
# Group sizes: 5 4 5
# For the given alpha=0.0502, the upper cutoff value is Kruskal-Wallis H=5.64285714,
# with true alpha level=0.0502
```

```
# Data
normal = c(2.9, 3.0, 2.5, 2.6, 3.2)      # Sujetos normales
oadisease = c(3.8, 2.7, 4.0, 2.4)         # Sujetos con enfermedad respiratoria
asbestosis = c(2.8, 3.4, 3.7, 2.2, 2.0)  # Sujetos con asbestosis
```

Nosotros estamos interesados en utilizar Kruskal-Wallis el test de para probar si existen diferencias en la mediana de los tiempos medios de eliminación mucociliar para las tres poblaciones de sujetos. Para este ejemplo consideramos el nivel de significancia $\alpha = ,0502$

En estos datos específicos tenemos n1 = n3 = 5, n2 = 4 y N = 14. Combinando estos hechos con las sumas de rango de tratamiento calculamos:

$$H = \frac{12}{14(14+1)}(\frac{36^2}{5} + \frac{36^2}{4} + \frac{33^2}{5}) - 3(14+1) = 0{,}771$$

```r
kruskal.test(list(normal, oadisease, asbestosis))
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  list(normal, oadisease, asbestosis)
## Kruskal-Wallis chi-squared = 0.77143, df = 2, p-value = 0.68
```

```r
# data:  list(normal, oadisease, asbestosis)
# Kruskal-Wallis chi-squared = 0.7714, df = 2, p-value = 0.68
```

Como este valor de H es menor que el valor crítico 5.643, no rechazamos H_0, con un nivel de confianza de $\alpha = 0{,}052$.

Ejercicio 3

Se cuenta con la siguiente información relacionada a pacientes sanos y otros quienes estan siendo sometidos a tratamientos bajo distintos fármacos. Los datos corresponden a algún índice que nos indca si los pacientes se encuentran dentro de los rangos normales de salud.

a) ¿Cual debería ser la hipótesis nula para (H_0) este caso?:

```
sin_trat = c(4.302, 4.017, 4.049, 4.176)
con_trat1 = c(2.201, 3.190, 3.250, 3.276, 3.292, 3.267)
con_trat2 = c(3.397, 3.552, 3.630, 3.578, 3.612)
con_trat3 = c(2.699, 2.929, 2.785, 2.176, 2.845, 2.913)
```

Unimos los datos disponibles con las mediciones para los pacientes

```
all = c(sin_trat, con_trat1, con_trat2, con_trat3)
trat = c(rep("1", 4), rep("2", 6), rep("3", 5), rep("4", 6))
mydata = data.frame(Y=all, X=as.factor(trat))
mydata
```

```
##        Y X
## 1  4.302 1
## 2  4.017 1
## 3  4.049 1
## 4  4.176 1
## 5  2.201 2
## 6  3.190 2
## 7  3.250 2
## 8  3.276 2
## 9  3.292 2
## 10 3.267 2
## 11 3.397 3
## 12 3.552 3
## 13 3.630 3
## 14 3.578 3
## 15 3.612 3
## 16 2.699 4
## 17 2.929 4
## 18 2.785 4
## 19 2.176 4
## 20 2.845 4
## 21 2.913 4
```

Aplicamos el test de Kruskall-Wallis

```
kruskal.test(Y ~ X, data=mydata)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Y by X
## Kruskal-Wallis chi-squared = 17.359, df = 3, p-value = 0.0005961
```

¿Cual debería ser la respuesta dado H_0 planteado anteriormente?

Ejercicio 4

Test de Jonckheere-Terpstra

Supongamos que queremos comparar distintos tratamientos aplicados a distintos grupos.

a) Podemos hacer este ejercicio en `R` mediante la función `jonckheere.test()` de la librería `clinfun`

```
## install.packages( "clinfun" )
library( "clinfun" )
trt1 = c(13.0, 24.1, 11.7, 16.3, 15.5, 24.5)
trt2 = c(42.0, 18.0, 14.0, 36.0, 11.6, 19.0)
trt3 = c(15.6, 23.8, 24.4, 24.0, 21.0, 21.1)
trt4 = c(35.3, 22.5, 16.9, 25.0, 23.1, 26.0)
n1 = length(trt1)
n2 = length(trt2)
n3 = length(trt3)
n4 = length(trt4)
```

Unimos el conjunto de datos y aplicamos la función `jonckheere.test()`

```
## install.packages( "clinfun" )
library( "clinfun" )
all = c(trt1, trt2, trt3, trt4)
N = length(all)
g = rep(c(1, 2, 3, 4), each=6)
jonckheere.test(all, g, alternative="increasing")
```

```
##
##  Jonckheere-Terpstra test
##
## data:
## JT = 145, p-value = 0.02991
## alternative hypothesis: increasing
```

Podemos aproximar con un N grande:

```
mu = (N^2 - (n1^2 + n2^2 + n3^2 + n4^2))/4          # Por formula
# n1=n2=n3=n4=6
variance = ((N^2)*(2*N + 3) - 4*(6^2)*(2*6 + 3))/72 # Por formula
(pA <- pnorm(145, 108, sqrt(378), lower.tail=FALSE))
```

```
## [1] 0.0285154
```