


Facultad de Ciencias
UNAM

Estadística II

Tarea-Examen, versión B

Pruebas no paramétricas

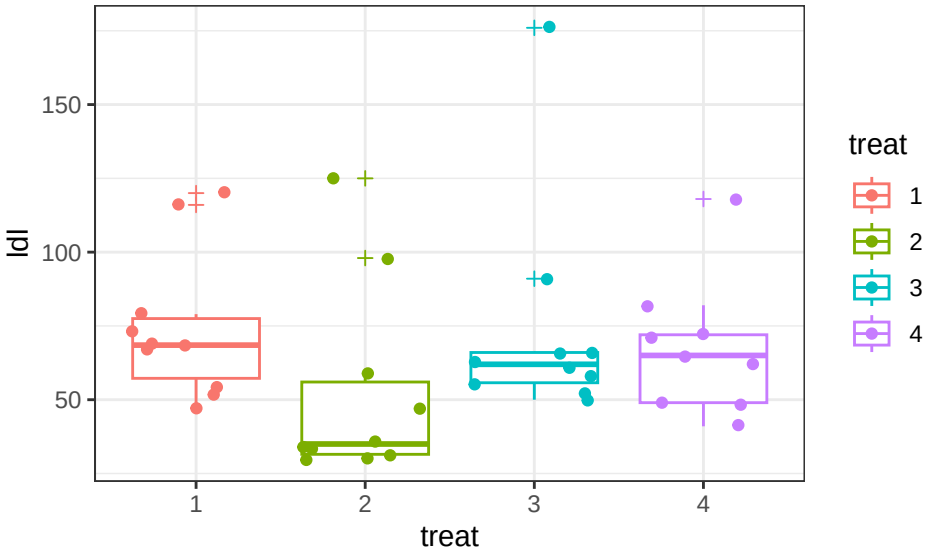
Enríquez Hernández Leobardo
Huitrón Zambrano Victor Manuel
Suárez López David

4 de junio de 2024

Índice

Ejercicio 1

En esta sección trabajaremos con datos de colesterol malo (ldl) y cuatro tratamientos (treat: 1,2,3,4) que se probaron para reducir los niveles. Se tienen 39 observaciones aleatorias e independientes en el conjunto de datos `quail` de la biblioteca `Rfit`. A continuación se presenta el BoxPlot, en donde podemos observar que la mediana de ldl para el segundo tratamiento es menor que en los demás casos, y que hay algunos valores que son outliers (los puntos asociados al símbolo de +). No parece que se pudiera asumir que los datos tengan un comportamiento normal para todos los grupos, por ejemplo en el caso del segundo grupo de tratamiento, no hay simetría entre los intercuantiles Q1 y Q3, con una media más cercana al cuantil Q1.



En la prueba de normalidad `Lilliefors (Kolmogorov-Smirnov) normality test` para cada grupo, tenemos que no se rechaza la hipótesis nula de normalidad, para los grupos 1 y 4, mientras que se rechaza para los grupos 2 y 3. El p-value para el grupo 1 es de 0.145818, para el 4 de 0.3653111, para el 2 de 0.0178636 y para el 3 de 0.0001785. (Chunk normalidad, linea de código 40)

Por otra parte, en las pruebas de homocedasticidad entre grupos, no se rechaza la hipótesis nula de misma varianza, pues el p-value para la prueba de Bartlett es de 0.4556299, para la prueba Fligner de 0.9807608 y para la prueba Levene de 0.9798342. (Chunk homocedasticidad, linea de código 53)

Si no asumimos ninguna distribución en particular pero sí que es la misma, la prueba Kruskal–Wallis se basa en una transformación de rangos. En este caso no hay problema de varianza heterogénea, por lo que no será necesario esto para comparar grupos. Sin embargo, con esta prueba tenemos que la mediana o la distribución no es la misma en todos los grupos, porque se rechaza la hipótesis nula (p-value de 0.0612452), con un nivel de significancia $\alpha = 0{,}1$. El resultado es similar si usamos distribución asintótoca con distribución Chisquare. (Chunk kruskaltest, linea de código 80)

```
##
##  Kruskal-Wallis test
##
## data:  ldl by treat
## H = 7.1879, k = 4.00000, U = 0.41111, N = 39.00000, p-value = 0.06125
```

```
##
##  Kruskal-Wallis test
##
## data:  ldl by treat
## chi-squared = 7.1879, df = 3, p-value = 0.06614
```

Además, en las pruebas simultáneas de Dunn, para pares de grupos, se tienen p-values mayores a 0.1 en casi

todos los casos, por lo que la dstribución es igual entre los grupos comparados por pares, a excepción de los grupos 1 y 2. (Chunk kwAllPairsDunnTest1, linea de código 100)

```
##            z value Pr(>|z|)
## 2 - 1 == 0   2.521 0.070292 .
## 3 - 1 == 0   0.598 1.000000
## 4 - 1 == 0   0.598 1.000000
## 3 - 2 == 0   1.922 0.272823
## 4 - 2 == 0   1.855 0.272823
## 4 - 3 == 0   0.016 1.000000
```

Comparando un grupo contra todos los demás (ManyOne) usando Dunn test, tenemos que si tomamos el grupo de referencia (Grupo 1), para responder si el grupo 1 es el mejor, hay un p-value menor a 0.1 para la comparación con el grupo 2, lo que rechaza que sean iguales estos grupos. Si cambiamos al grupo de referencia, se rechaza la hipótesis de que éstos dos grupos (1 Y 2) sean iguales en todos los casos. (Chunk kwManyOneDunnTest1, linea de código 108)

```
##            z value Pr(>|z|)
## 2 - 1 == 0  -2.521 0.031777 *
## 3 - 1 == 0  -0.598 0.881845
## 4 - 1 == 0  -0.598 0.881873
```

```
## [1] "1" "2" "3" "4"
```

Haciendo una prueba con dirección, podemos concluir que el grupo 2 es mejor que el grupo 1, pues se rechaza la hipótesis nula, donde la hipótesis alternativa es que el grupo 2 es mejor que el grupo 1, este es el único caso. (Chunk kwManyOneDunnTestLess, linea de código 125)

```
##            z value   Pr(<z)
## 2 - 1 >= 0  -2.521 0.015955 *
## 3 - 1 >= 0  -0.598 0.499306
## 4 - 1 >= 0  -0.598 0.499261
```

Por otra parte, si hacemos esta misma prueba tomando como referencia al grupo 2, no hay suficiente evidencia para decir que el tratamiento 2 es mejor o reduce más los niveles de colesterol en comparación con el resto de tratamientos. (Chunk kwManyOneDunnTestLess2, linea de código 136)

```
##            z value  Pr(<z)
## 1 - 2 >= 0   2.521 0.99986
## 3 - 2 >= 0   1.922 0.99825
## 4 - 2 >= 0   1.855 0.99774
```

Análogamente, esta prueba se hizo también para los grupos 3 y 4, con lo que tampoco hay suficiente evidencia para afirmar que los tratamientos 3 y 4 pudieran ser mejores que los demás.

Podemos concluir que el tratamiento 2 es mejor que el 1, sin embargo no es posible afirmar que sea mejor que los tratamientos 3 y 4. En las pruebas de los demás tratamientos, no es posible afirmar que los tratamientos 1,3 y 4 sean mejores que algún otro.

Ejercicio 2

1) Prueba parametrica

```
x<-c(1.53,1.68,1.88,1.55,3.06,1.3,0.5,1.62,2.48)
y<-c(0.578,1.06,1.29,1.06,3.14,1.29,0.647,0.59,2.05)
w<-x-y
```

Para esta prueba usamos el modelo de regresion:

$$w_i : \beta_0 + \epsilon_i$$

Para saber si el tratamiento tuvo exito neseistamos la siguiente prueba de hipotesis:

$$H_0 : \beta_0 \leq 0 vs H_a : \beta_0 > 0$$

```
K=matrix(c(1), ncol=1, nrow=1, byrow=TRUE)
m=c(0)
summary(glht(fit, linfct=K, rhs=m, alternative="greater"))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = w ~ 1)
##
## Linear Hypotheses:
##        Estimate Std. Error t value  Pr(>t)
## 1 <= 0   0.4328     0.1427   3.032 0.00813 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Se rechaza H_0 por lo que podemos decir que el tratamiento tuvo exito.

2)Prueba no parametrica

prueba de hipotesis:

$$H_0 : \theta \leq 0 vs H_a : \theta > 0$$

```
wilcox.test(x,y,paired = TRUE, alternative = c("greater"), exact = TRUE, correct = FALSE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  x and y
## V = 40, p-value = 0.01953
## alternative hypothesis: true location shift is greater than 0
```

En esta prueba tambien se rechaza H_0, lo que nos indica que el tratamiento tuvo exito.

Ejercicio 3

```
x <- c(3.9, 7.9, 4.1, 8.8, 9.4, 0.46, 5.3, 8.92, 5.5, 4.6, 47.9, 23.2, 34.2, 29.1, 6.0,
       45.1, 13.1, 3.1, 17.1, 47.8)
y <- c(2.3, 2.4, 1.8, 2.3, 1.7, 0.1, 2.2, 0.22, 2.4, 1.0, 2.8, 2.9, 2.5, 2.6, 1.2, 2.1,
       3.4, 1.3, 1.7, 1.6)
```

I)

Vamos a obtener el coeficiente de correlacion de Pearson.

```
cor.test(x,y,method = "pearson", alternative="two.sided")
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 1.6827, df = 18, p-value = 0.1097
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08823671  0.69741789
## sample estimates:
##       cor
## 0.3686798
```

Entonces el coeficiente de correlacion de Pearson es de 0.3686798 y ademas el p-value es mayor a .05 por lo que no se rechaza H_0, es decir, no estamos rechazando que x y y sean independientes. Pero para realizar este coeficiente estamos suponiendo que se cumple la normalidad bivariada, entonces vamos a hacer pruebas para ver si efectivamente se cumple este supuesto.

```
mvn(data=cbind(x,y), mvnTest = "hz")
```

```
## $multivariateNormality
##            Test        HZ    p value MVN
## 1 Henze-Zirkler 0.8589563 0.02203508  NO
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        1.6558    0.0002    NO
## 2 Anderson-Darling     y        0.3796    0.3704    YES
##
## $Descriptives
##    n   Mean    Std.Dev Median  Min  Max  25th   75th       Skew   Kurtosis
## x 20 16.274 15.8947633   8.86 0.46 47.9 5.125 24.675  0.9733591 -0.6340077
## y 20  1.926  0.8506617   2.15 0.10  3.4 1.525  2.425 -0.5606597 -0.3801318
```

```
mvn(data=cbind(x,y), mvnTest = "mardia")
```

```
## $multivariateNormality
##              Test          Statistic           p value Result
## 1 Mardia Skewness   5.80081251091552 0.214525741744589    YES
## 2 Mardia Kurtosis -0.629418159366215 0.529075332506437    YES
## 3             MVN               <NA>              <NA>    YES
##
## $univariateNormality
##               Test  Variable Statistic   p value Normality
## 1 Anderson-Darling     x        1.6558    0.0002    NO
```

```
## 2 Anderson-Darling     y        0.3796    0.3704    YES
##
## $Descriptives
##    n   Mean    Std.Dev Median  Min  Max  25th   75th       Skew   Kurtosis
## x 20 16.274 15.8947633   8.86 0.46 47.9 5.125 24.675  0.9733591 -0.6340077
## y 20  1.926  0.8506617   2.15 0.10  3.4 1.525  2.425 -0.5606597 -0.3801318
```

En ambas pruebas podemos ver que marginalmente rechazamos que x se distribuya como una normal, por lo que se rechaza H_0, es decir, hay evidencia que nos dice que no se cumple el supuesto de normalidad bivariada. Y por lo tanto el coeficiente que obtuvimos solo se puede usar como una estadistica que nos habla de la asociacion monotona de las variables y en este caso la prueba de hipotesis asociada al coeficiente de correlacion de Pearson no tiene validez.

II)

Ahora vamos a obtener el coeficiente τ_b de Kendall

```
cor.test(x,y,method = "kendall", alternative="two.sided")
```

```
##
##  Kendall's rank correlation tau
##
## data:  x and y
## z = 1.9172, p-value = 0.05521
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.3130073
```

En este caso el coeficiente τ_b de Kendall es de 0.3130073

III)

Ahora vamos a calcular el coeficiente ρ_s de Spearman

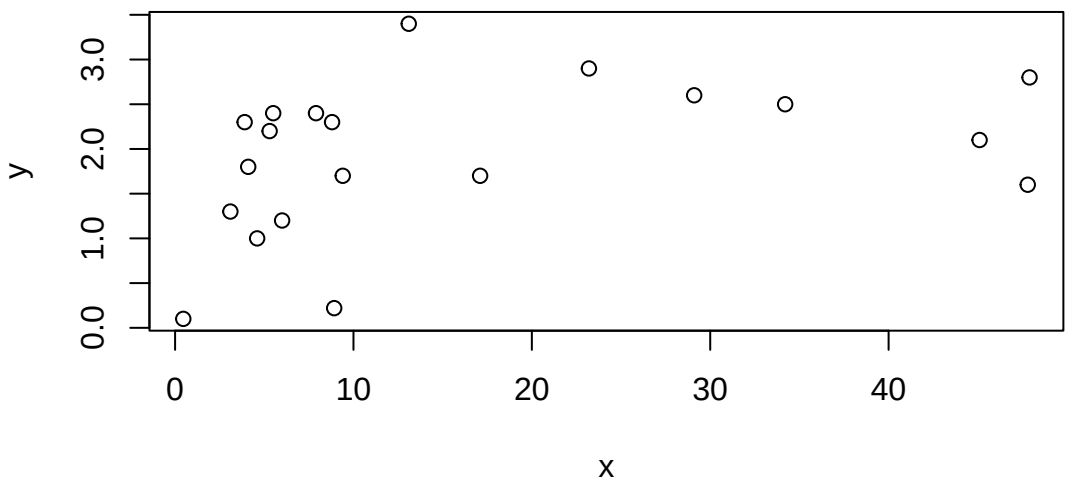
```
cor.test(x,y,method = "spearman", alternative="two.sided")
```

```
##
##  Spearman's rank correlation rho
##
## data:  x and y
## S = 710.3, p-value = 0.0384
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.4659393
```

Entonces tenemos que el coeficiente ρ_s de Spearman es de 0.4659393

Vamos a realizar un diagrama de dispersion para ver como se comportan las variables

```r
plot(x,y)
```



Podemos ver en el diagrama de dispersion que no hay algun patron o comportamiento en especifico que nos pueda indicar que las variables sean independientes e incluso podemos ver que no parece haber una relacion completamente monotona positiva entre las variables

Ejercicio 4

Primero vamos a transformar nuestras variables a tipo factor de una manera que nos sea conveniente para poder analizar si a mayor nivel de estudio menor es el impacto de las fakenews

```
datos$NivEdu=factor(datos$NivEdu, levels = c("Primaria", "Secundaria",
                                             "Bachillerato", "Profesional"))
datos$FakeNews=factor(datos$FakeNews,levels = c("Muy Poco", "Poco", "Regular", "Mucho"))
str(datos)
```

```
## 'data.frame':    1000 obs. of  2 variables:
##  $ NivEdu  : Factor w/ 4 levels "Primaria","Secundaria",..: 2 2 4 3 3 4 3 1 1 2 ...
##  $ FakeNews: Factor w/ 4 levels "Muy Poco","Poco",..: 1 1 1 2 1 3 2 4 4 2 ...
```

Y ahora procedemos a calcular los coeficientes τ_b de Kendall y ρ_s de Spearman

```
cor.test(rank(datos$NivEdu),rank(datos$FakeNews), method = "kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  rank(datos$NivEdu) and rank(datos$FakeNews)
## z = -9.4456, p-value < 0.00000000000000022
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##        tau
## -0.2491923
```

```
cor.test(rank(datos$NivEdu),rank(datos$FakeNews), method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  rank(datos$NivEdu) and rank(datos$FakeNews)
## S = 215969482, p-value < 0.00000000000000022
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2958182
```

En ambos casos podemos ver que el p-value es de 2.2e-16 por lo que rechazamos H_0, es decir, hay evidencia de que hay una relacion monotona entre las variables. En el caso del coeficiente τ_b de Kendall tiene un valor de -0.2491923, mientras que en el coeficiente ρ_s de Spearman tenemos un valor de -0.2958182 por lo que en ambos caso tendriamos una relacion monotona negativa. Entonces podriamos decir que a mayor nivel de estudio menor es el impacto de las fakenews

Ejercicio 5

A partir de la información de 1475 pacientes que sufrieron un paro cardiaco, de los cuales 733 recibieron el medicamento Sulphinpyrazone usada para disminuir la muerte cardiaca y 742 un placebo, durante 2 años, se quiere saber si el medicamento funciona o no. A continuación se muestra el número de pacientes (Frec) de acuerdo a si tomó o no el medicamento y su condición de vivo o muerto después del paro cardiaco.

Frecuencia	Tratamiento	Vivo
692	Sulphinpyrazone	Si
41	Sulphinpyrazone	No
682	Placebo	Si
60	Placebo	No

Con esta información haremos una prueba de hipótesis para indicar si la condición de muerte después de un paro cardíaco es diferente de acuerdo a si se recibió o no el tratamiento con Sulphinpyrazone, considerando un novel de significancia estadística de $\alpha = 0{,}1$.

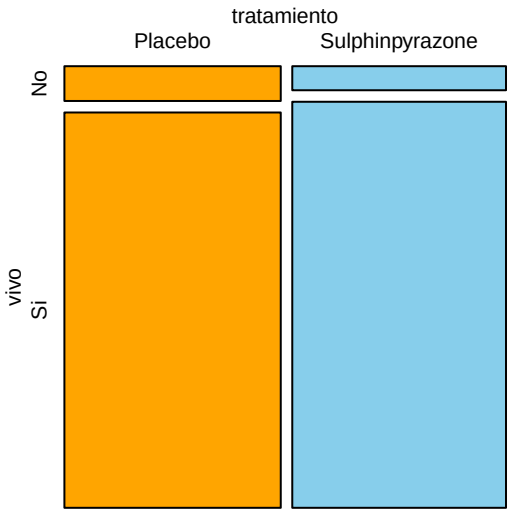
En primer lugar creamos una tabla de contingencia, y los valores esperados para cada celda bajo el supuesto de independencia entre las variables de tratamiento y vivo en la hipótesus nula H_0. Estos últimos valores se muestran a continuación.

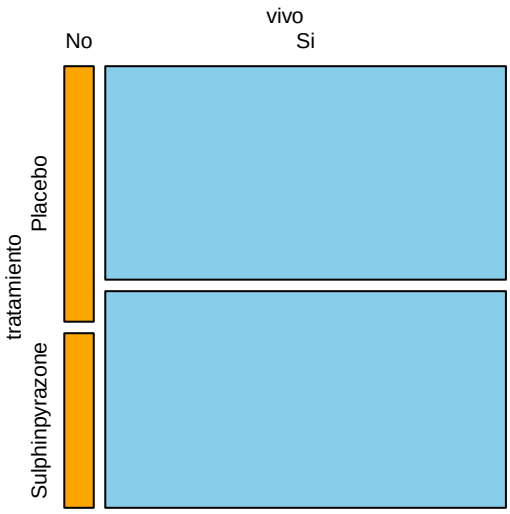
```
##            No       Si
## [1,] 50.80814 691.1919
## [2,] 50.19186 682.8081
```

A continuación mostramos la prueba de independencia con la función `loglm` de la biblioteca `MASS`, donde tenemos la prueba del cociente de verosimilitudes generalizadas y la chi cuadrada (Pearson). En ambos casos la hipótesis nula H_0 es que las variables tratamiento y vivo son independientes y como el p-value es menor considerando $\alpha = 0{,}1$, se rechaza H_0. Por lo tanto estas variables no son independientes.

```
## Call:
## loglm(formula = frec ~ tratamiento + vivo, data = data)
##
## Statistics:
##                       X^2 df   P(> X^2)
## Likelihood Ratio 3.613508  1 0.05731210
## Pearson          3.592256  1 0.05804938
```

En la suguiente gráfica de mosaico, mostramos las distribuciones por cada una de las clases. En esta caso podemos ver que en efecto el medicamento hace una diferencia en estar vivo o no estarlo. Con esto podríamos concluir de manera exploratoria que no son independientes el tratamiento de estar o no vivo.


tratamiento
Placebo
Sulphinpyrazone
vivo
No
Si


vivo
No
Si
tratamiento
Placebo
Sulphinpyrazone

Ejercicio 6

Tenemos la prueba de hipotesis:

H_0:provienen de la distribucion exp(2)

vs

H_a:no provienen de esa distribucion

```
observados<-c(0.0023, 0.0150, 0.0298, 0.0337, 0.0729, 0.0943, 0.0950, 0.1080,
              0.1180, 0.1300, 0.1500, 0.1592, 0.1617, 0.2016,0.2083, 0.2316,
              0.2403, 0.2863, 0.3427, 0.3766, 0.4384, 0.4715, 0.4895, 0.5544,
              0.5575, 0.5910, 0.5960, 0.6224,0.6517, 0.6602, 0.7197, 0.7317,
              0.7687, 0.8212, 0.9439, 1.1242, 1.2681, 1.2885, 2.3626, 2.6055)
```

Usamos la prueba de bondad de ajuste ji-cuadrada

```
gofTest(observados, test = "chisq", distribution = "exp", param.list = list(rate = 2),
            cut.points=c(0,0.3,0.7,1.1,Inf) )
```

```
##
## Results of Goodness-of-Fit Test
## -------------------------------
##
## Test Method:                     Chi-square GOF
##
## Hypothesized Distribution:       Exponential(rate = 2)
##
## Data:                            observados
##
## Sample Size:                     40
##
## Test Statistic:                  Chi-square = 0.1078527
##
## Test Statistic Parameter:        df = 3
##
## P-value:                         0.9908787
##
## Alternative Hypothesis:          True cdf does not equal the
##                                  Exponential(rate = 2)
##                                  Distribution.
```

No se rechaza H_0, por lo cual no hay evidencia suficiente para rechazar que proviene de la distribución exp(2).

Ejercicio 7

Tenemos la prueba de hipotesis:

H_0:provienen de la distribucion exponencial

vs

H_a:no provienen de esa distribucion

Usamos la prueba Kolmogorov–Smirnov con correcion Lilliefors

```
set.seed(123)
library(KScorrect)
Test=LcKS(observados, cdf = "pexp", nreps = 10000)
Test$D.obs
```

```
## [1] 0.07152941
```

```
Test$p.value
```

```
## [1] 0.9435056
```

No se rechaza H_0, por lo tanto no hay suficiente evidencia para rechazar que provenga de la distribución exponencial.