Taller de Consultoria - TP3

Gonzalo Barrera Borla

23/09/2019

Setup

Ejercicio 1

> Los datos siguientes corresponden a un experimento realizado por Charles Darwin en 1876. En cada maceta se plantan dos brotes de maíz, uno producido por fertilización cruzada, y el otro por auto-fertilización. El objetivo era mostrar las ventajas de la fertilización cruzada. Los datos son las altura finales de las plantas después de un período de tiempo. ¿Alguno de los dos tipos de maíz es demostrablemente mejor?. Si es así, ¿cómo se puede describir la diferencia?.

Sea Ω_0 la familia de distribuciones tal que

$$\Omega_0 = \{F : F \text{ es absolutamente continua con única mediana en } 0\}$$

Sea Y_i la altura del brote de maíz producido por fertilización cruzada de la maceta $i \in \{1, \dots, n\}, n = 15$, y X_i la altura del producido por auto-fertilización. Supondremos además, razonablemente, que $Y_i \overset{iid}{\sim} F_Y(t) = F(t - \theta_Y)$, $F \in \Omega_0$ (es decir que F_Y es absolutamente continua con única mediana en θ_Y). Análogamente, $X_i \overset{iid}{\sim} F_X(t) = F(t - \theta_X)$. Como *"el objetivo era mostrar las ventajas de la fertilización cruzada"*, una forma razonable de plantear este test será:

$$H_0 := \theta_Y = \theta_X \quad \text{vs.} \quad H_1 := \theta_Y > \theta_X$$

Como los brotes están naturalmente apareados en sus respectivas macetas, es razonable realizar un test de muestras apareadas. Para determinar qué clase de test realizar, nos resta contestar: ¿Es normal la distribución de las alturas de ambos tipos de brotes? Si la respuesta es "sí", podemos usar un "test t", mientras que si no será conveniente recurrir a alguna alternativa no paramétrica, como el test de Wilcoxon de rangos signados, que sólo requiere que la distribución bajo la hipótesis nula sea simétrica.

Para testear la normalidad de los datos, utilizamos el test de Shapiro univariado. Para la diferencia $D_i = Y_i - X_i$, el p-valor es de 0.093, que dependiendo del nivel de significaión utilizado puede alcanzar o no para rechazar la normalidad. Si realizamos dos tests por separado, para la muestra $(X_1, \dots, X_n)$ de brotes autofertilizados obtenemos un p-valor 0.38 con lo cual es razonable asumir su normalidad, pero al aplicar el test a la muestra de fertilización cruzada, obtenemos un p-valor de 9.7×10^{-4} que nos lleva a rechazar su normalidad. Resulta difícil suponer que por casualidad las diferencias de altura resulten normales si cada muestra no tiene una distribución normal subyacente, así que por seguridad convendrá recurrir a un test no paramétrico.

Nótese que bajo $H_0 := \theta_Y = \theta_X \Rightarrow F_X = F_Y$, de manera que la distribución de $Y_i - X_i$ será simétrica, y podemos utilizar un test de Wilcoxon para datos apareados con confianza.

Este test arroja un p-valor de 0.021, de manera que con el tradicional criterio de significación $\alpha = 0.05$, podemos rechazar la hipótesis nula y concluir que la diferencia entre las medianas de los brotes de fertilización cruzada y los autofertilizados es positiva. Aprovechando el hecho de que el estadístico T^+ es un estimador de Hodges-Lehmann, podemos encontrar también el intervalo de confianza de nivel 95% correspondiente al test, que resulta ser $[1, \infty)$.

Por último, y a modo ilustrativo, incluimos en la siguiente tabla los p-valores e intervalos de confianza de nivel 95% correspondientes a cuatro alternativas de test razonables en este problema:

Muestras	Test	p-valor	Lím. inf. IC
apareadas	Wilcoxon	0.0206	1.0000
2 muestras	Wilcoxon	0.0013	2.0000
apareadas	T de Student	0.0251	0.4634
2 muestras	T de Student	0.0118	0.7666

Ejercicio 2

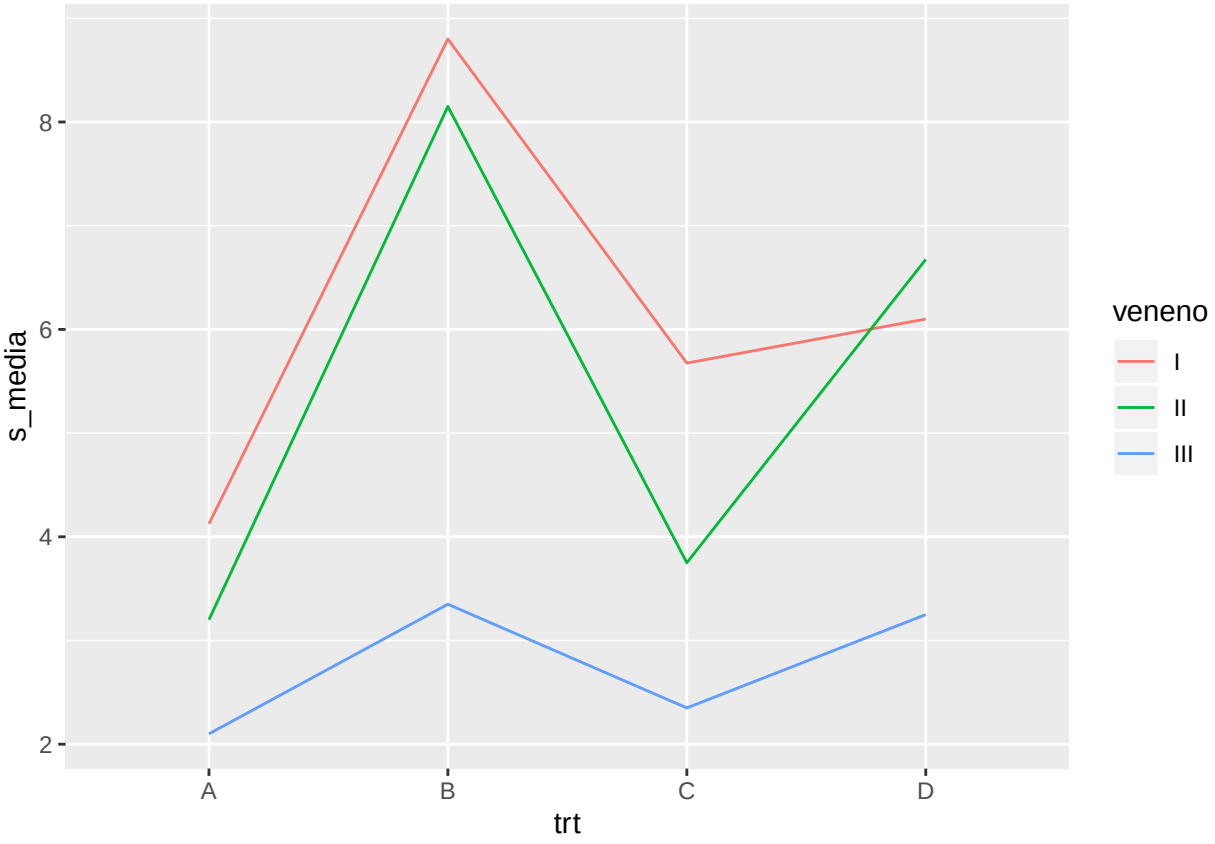
```
##
## Call:
## lm(formula = sobrevida ~ veneno + trt, data = juntos2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5167 -0.9625 -0.1490  0.6177  4.9833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5229     0.5592   8.088 4.22e-10 ***
## venenoII     -0.7313     0.5592  -1.308  0.19813
## venenoIII    -3.4125     0.5592  -6.102 2.83e-07 ***
## trtB          3.6250     0.6458   5.614 1.43e-06 ***
## trtC          0.7833     0.6458   1.213  0.23189
## trtD          2.2000     0.6458   3.407  0.00146 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.582 on 42 degrees of freedom
## Multiple R-squared:  0.6503, Adjusted R-squared:  0.6087
## F-statistic: 15.62 on 5 and 42 DF,  p-value: 1.123e-08
```

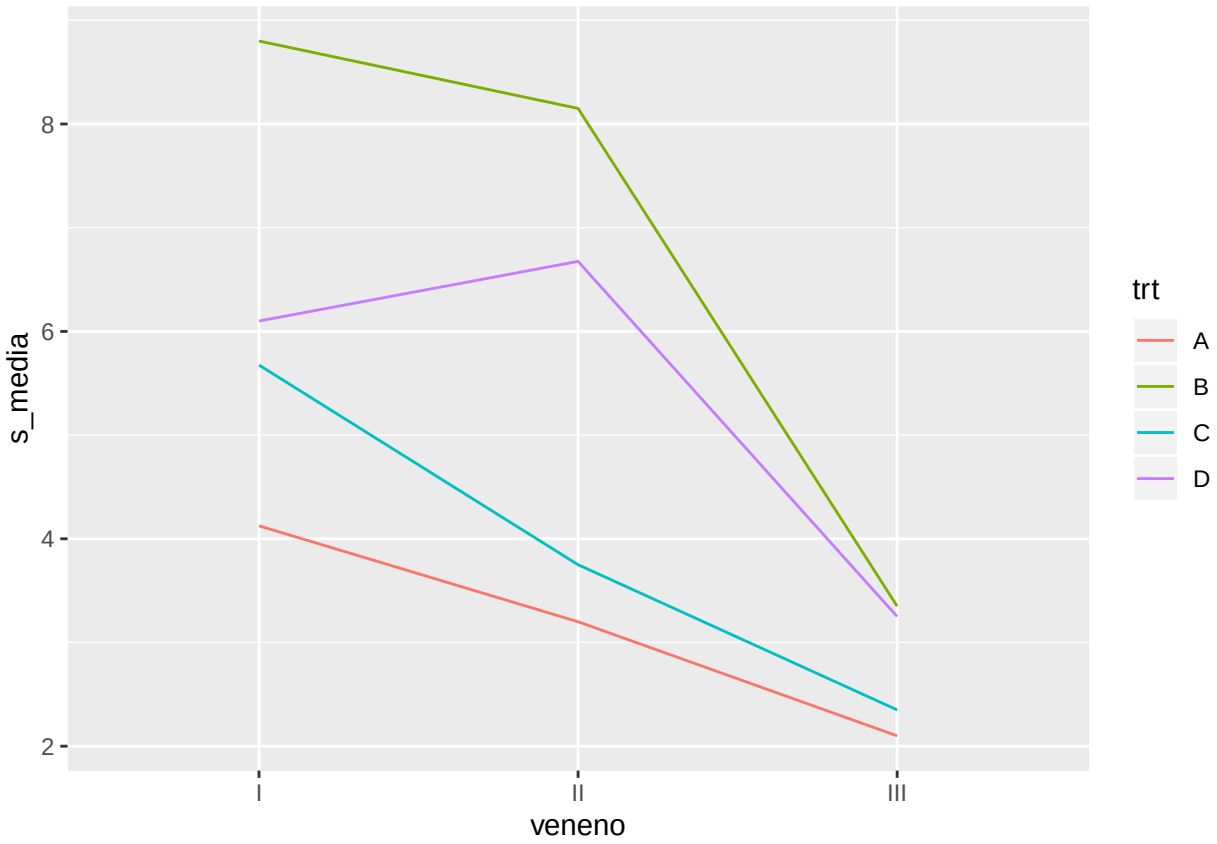
```
## Analysis of Variance Table
##
## Response: sobrevida
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## veneno     2 103.301  51.651  20.643 5.704e-07 ***
## trt        3  92.121  30.707  12.273 6.697e-06 ***
## Residuals 42 105.086   2.502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## lm(formula = sobrevida ~ veneno * trt, data = juntos2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2500 -0.4875  0.0500  0.4313  4.2500
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.1250     0.7457   5.532 2.94e-06 ***
## venenoII        -0.9250     1.0546  -0.877   0.3862
```

```
## venenoIII       -2.0250     1.0546  -1.920   0.0628 .
## trtB             4.6750     1.0546   4.433 8.37e-05 ***
## trtC             1.5500     1.0546   1.470   0.1503
## trtD             1.9750     1.0546   1.873   0.0692 .
## venenoII:trtB    0.2750     1.4914   0.184   0.8547
## venenoIII:trtB  -3.4250     1.4914  -2.297   0.0276 *
## venenoII:trtC   -1.0000     1.4914  -0.671   0.5068
## venenoIII:trtC  -1.3000     1.4914  -0.872   0.3892
## venenoII:trtD    1.5000     1.4914   1.006   0.3212
## venenoIII:trtD  -0.8250     1.4914  -0.553   0.5836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.491 on 36 degrees of freedom
## Multiple R-squared:  0.7335, Adjusted R-squared:  0.6521
## F-statistic:  9.01 on 11 and 36 DF,  p-value: 1.986e-07
```

```
## Analysis of Variance Table
##
## Response: sobrevida
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## veneno      2 103.301  51.651 23.2217 3.331e-07 ***
## trt         3  92.121  30.707 13.8056 3.777e-06 ***
## veneno:trt  6  25.014   4.169  1.8743    0.1123
## Residuals  36  80.072   2.224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```






```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sobrevida ~ veneno + trt, data = juntos2)
##
## $veneno
##             diff       lwr       upr     p adj
## II-I   -0.73125 -2.089936  0.627436 0.3989657
## III-I  -3.41250 -4.771186 -2.053814 0.0000008
## III-II -2.68125 -4.039936 -1.322564 0.0000606
##
## $trt
##           diff        lwr        upr     p adj
## B-A  3.6250000  1.8976135  5.3523865 0.0000083
## C-A  0.7833333 -0.9440532  2.5107198 0.6221729
## D-A  2.2000000  0.4726135  3.9273865 0.0076661
## C-B -2.8416667 -4.5690532 -1.1142802 0.0004090
## D-B -1.4250000 -3.1523865  0.3023865 0.1380432
## D-C  1.4166667 -0.3107198  3.1440532 0.1416151
```

```
## List of 5
##  $ statistic: Named num 9
##   ..- attr(*, "names")= chr "Friedman chi-squared"
##  $ parameter: Named num 3
##   ..- attr(*, "names")= chr "df"
##  $ p.value  : num 0.0293
##  $ method   : chr "Friedman rank sum test"
##  $ data.name: chr "sobrevida and trt and veneno"
```

```
##  - attr(*, "class")= chr "htest"
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = sobrevida ~ veneno + trt, data = juntos2)
##
## $veneno
##           diff       lwr       upr     p adj
## II-I   -0.73125 -2.089936  0.627436 0.3989657
## III-I  -3.41250 -4.771186 -2.053814 0.0000008
## III-II -2.68125 -4.039936 -1.322564 0.0000606
##
## $trt
##           diff        lwr        upr     p adj
## B-A  3.6250000  1.8976135  5.3523865 0.0000083
## C-A  0.7833333 -0.9440532  2.5107198 0.6221729
## D-A  2.2000000  0.4726135  3.9273865 0.0076661
## C-B -2.8416667 -4.5690532 -1.1142802 0.0004090
## D-B -1.4250000 -3.1523865  0.3023865 0.1380432
## D-C  1.4166667 -0.3107198  3.1440532 0.1416151
```

```
##
## Call:
## lm(formula = prop ~ dosis * substancia, data = df3)
##
## Residuals:
##          1          2          3          4          5          6
## -0.0176411 -0.0015121  0.0380040 -0.0188508  0.0045625 -0.0011250
##          7          8          9
## -0.0043125 -0.0006875  0.0015625
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.514e-02  1.480e-02   1.698    0.150
## dosis              6.855e-05  1.319e-04   0.520    0.625
## substanciaB       -1.720e-02  2.006e-02  -0.858    0.430
## dosis:substanciaB -1.755e-05  1.420e-04  -0.124    0.906
##
## Residual standard error: 0.02077 on 5 degrees of freedom
## Multiple R-squared:  0.2958, Adjusted R-squared:  -0.1268
## F-statistic: 0.6999 on 3 and 5 DF,  p-value: 0.5913
```

```
##
## Call:
## lm(formula = prop ~ dosis * substancia0, data = df3)
##
## Residuals:
##          1          2          3          4          5          6
## -0.0025000  0.0066929  0.0287863  0.0009882  0.0025000 -0.0084735
##          7          8          9
## -0.0307054  0.0007924  0.0019191
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         4.325e-02  2.088e-02   2.071    0.107
## dosis              -2.033e-05  6.888e-05  -0.295    0.783
## substancia0B       -2.502e-02  2.811e-02  -0.890    0.424
## substancia0O       -3.325e-02  2.596e-02  -1.280    0.270
## dosis:substancia0B  2.425e-05  1.342e-04   0.181    0.865
## dosis:substancia0O         NA         NA      NA       NA
##
## Residual standard error: 0.02183 on 4 degrees of freedom
## Multiple R-squared:  0.3777, Adjusted R-squared:  -0.2446
## F-statistic: 0.6069 on 4 and 4 DF,  p-value: 0.6798
```

```
##
## Call:
## glm(formula = prop ~ substancia * dosis, family = binomial, data = df3,
##     weights = muestreadas)
##
## Deviance Residuals:
##       1         2         3         4         5         6         7
## -1.99628   0.36885   2.98958  -1.66756   0.40807  -0.27430  -0.50882
##       8         9
##  0.11125   0.05572
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.911459   0.287501 -13.605   <2e-16 ***
## substanciaB       -0.647289   0.455109  -1.422    0.155
## dosis              0.003815   0.002367   1.612    0.107
## substanciaB:dosis -0.001374   0.002619  -0.525    0.600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25.619  on 8  degrees of freedom
## Residual deviance: 16.356  on 5  degrees of freedom
## AIC: 54.15
##
## Number of Fisher Scoring iterations: 5
```