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

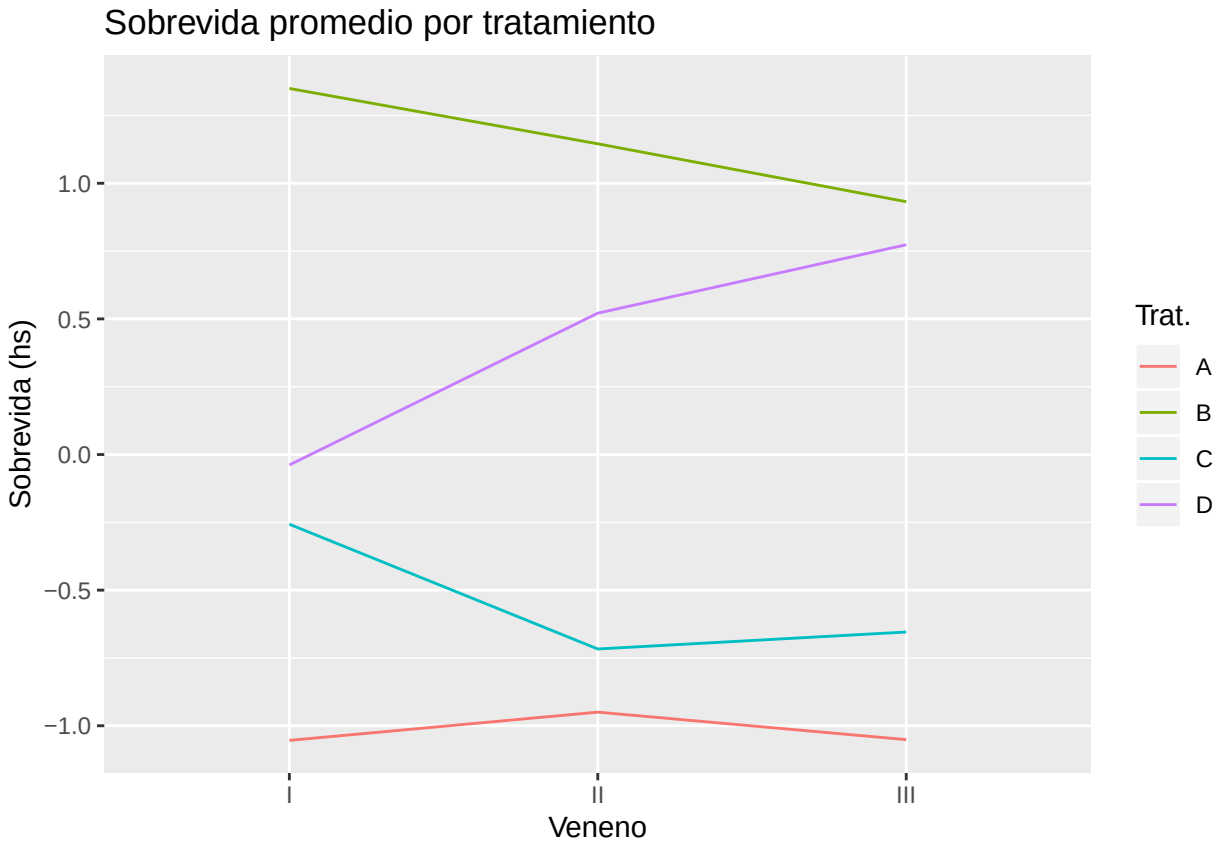
Es decir, que en todos los casos hubiésemos rechazado la hipótesis nula, pero la estimación de la diferencia en las medianas/medias hubiese sido bastante distinta.

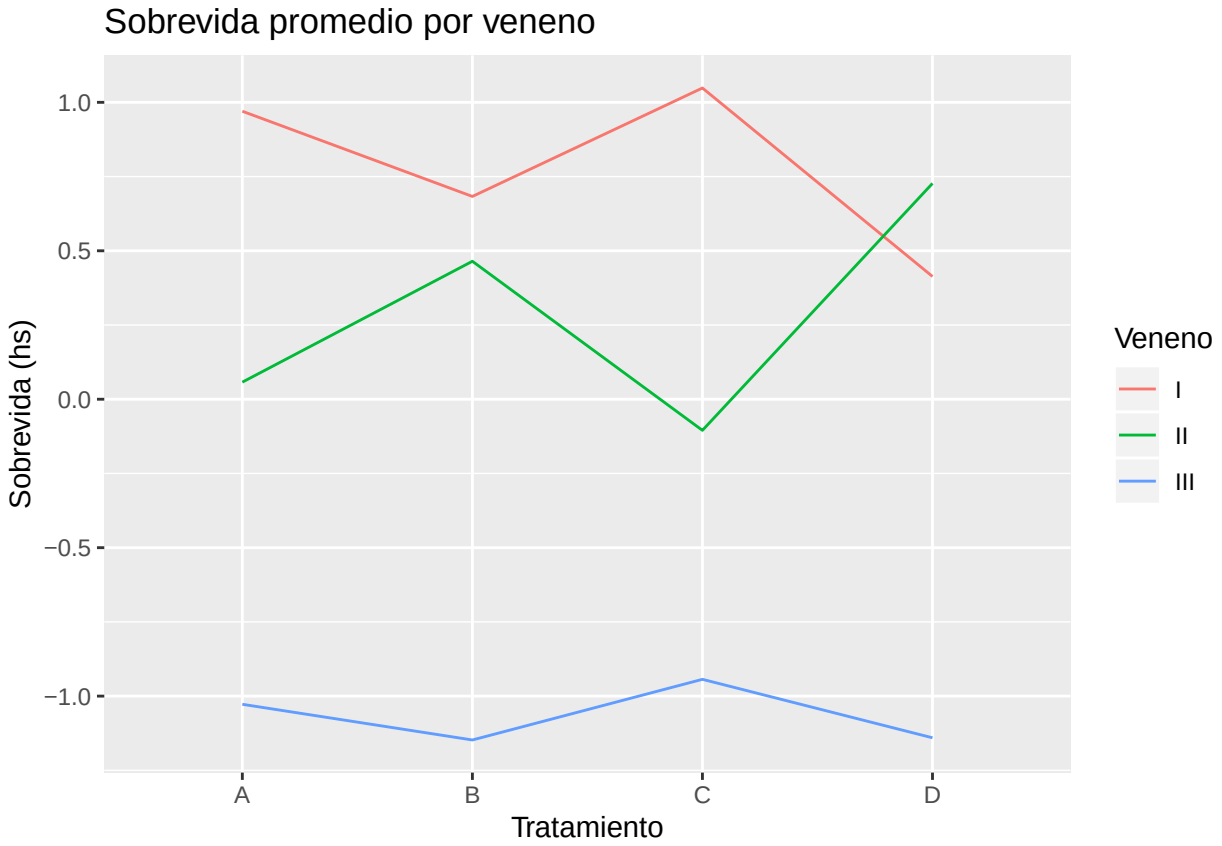
Ejercicio 2

> Se dan a continuación los tiempos de sobrevida (en unidades de 10 horas) de animales, sometidos a 3 tipos de veneno, y 4 tratamientos antitóxicos. Cada combinación veneno-tratamiento se prueba con 4 animales. Describir la influencia de ambos factores en la sobrevida. ¿Hay algún tratamiento demostrablemente mejor?

Para hacer más intuitiva la información, cambiamos las unidades de `sobrevida` de modo que esté en horas. Sea y_{ijk} la sobrevida del k-ésimo animal, sometida al tratamiento j para el veneno i, en general planteamos $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, donde $i \in \{I, II, II\}$, $j \in \{A, B, C, D\}$, $k \in \{1, 2, 3, 4\}$ y $\epsilon_{ijk} \sim N(0, \sigma^2)$.

Para darnos una idea de la relación entre venenos y tratamientos, comenzamos graficando los perfiles:





A primera vista, no se observa paralelismo en ninguno de los dos gráficos Sin embargo, salvo un ligero entrecruzamiento en los perfiles de los venenos II y III para los tratamientos C y D, los perfiles no se cruzan en ningún otro punto, por lo que tal vez sea razonable ignorar las interacciones y considerar un modelo de efectos únicamente aditivos.

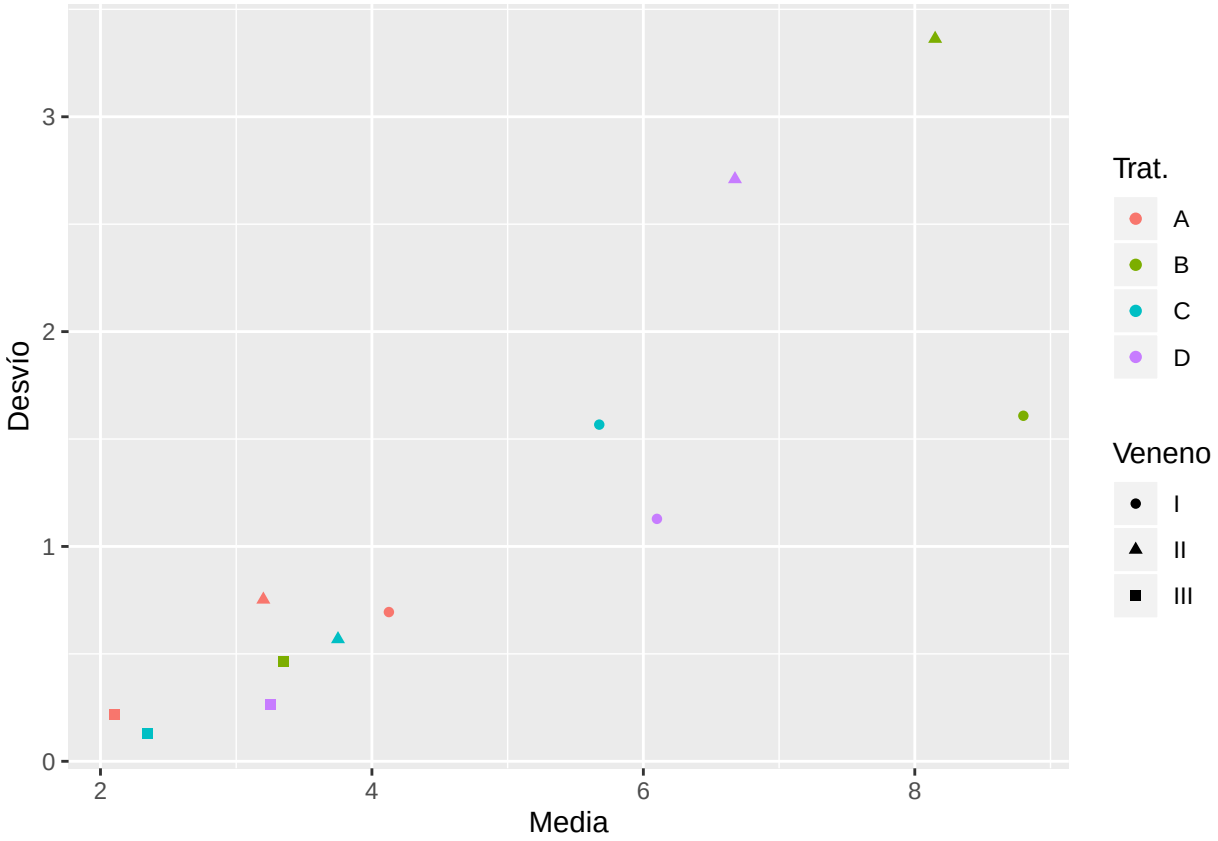
Por otra parte, debemos corroborar el fortísimo supuesto de *homocedasticidad* de los ϵ_{ijk} que hicimos inicialmente. Seber [1977, p. 195] siguiendo a Scheffé menciona que cuando se cuenta con un diseño balanceado como aquí, la desigualdad en las varianzas para cada combinación de bloque y tratamiento no hace demasiado "daño", siempre y cuando la razón entre el mayor y el menor desvío estándar es menor a 2. Lamentablemente, nuestros datos son "bochornosos" en tal sentido:

Table 2: La tasa entre los desvios de II:B y III:C es mayor a 26

veneno	trt	mean	sd
II	B	8.150	3.363
II	D	6.675	2.710
I	B	8.800	1.608
I	C	5.675	1.567
I	D	6.100	1.128
II	A	3.200	0.753
I	A	4.125	0.695
II	C	3.750	0.569
III	B	3.350	0.465
III	D	3.250	0.265
III	A	2.100	0.216
III	C	2.350	0.129

Antes de proseguir, nos convendrá entonces considerar una transformación estabilizadora de la varianza que al menos morigere esta situación.

Si graficamos el desvío en función de la media de cada grupo, observamos una estructura razonablemente lineal en la media:

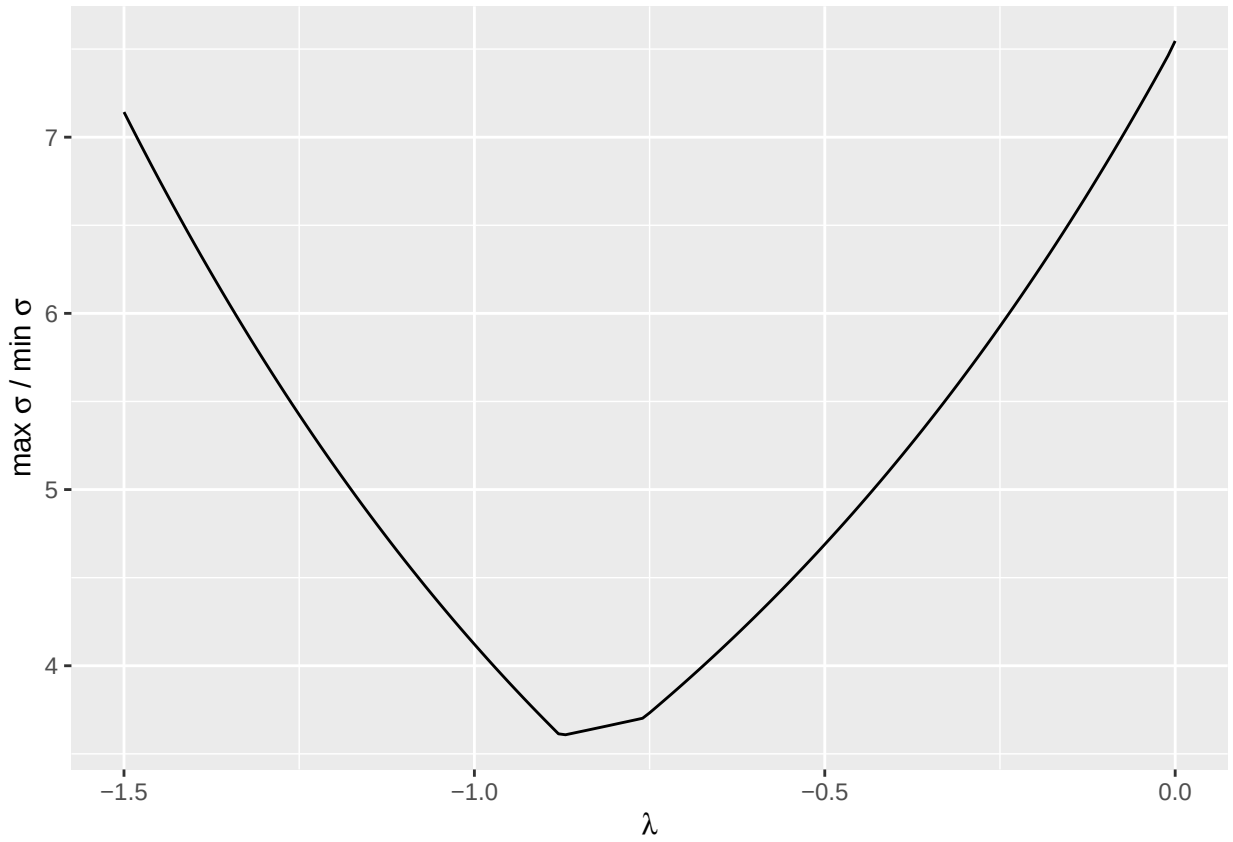


En este caso, donde $\sigma \approx k \times \mu$, la transformación sugerida es el logaritmo. Esto es razonable, ya que al hablar por ejemplo del logaritmo base 2 de la *sobrevida*, un tratamiento que duplica la supervivencia, aumentará $\log_2(y)$ en una unidad. Aplicando esta transformación, el la razón entre el máximo y mínimo desvíos anteriormente mencionada mejora bastante, pero sigue estando por encima de 7, muy lejos del corte de 2 que propone Seber.

Haremos un intento más, y en lugar de considerar la transformación logarítmica, buscaremos la transformación "óptima" según este criterio de optimalidad, de entre la familia de transformaciones Box-Cox de un parámetro, que incluye el logaritmo natural como un caso especial cuando $\lambda = 0$:

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

Si graficamos la relación entre el mayor y menor desvío estándar por bloque y tratamiento para distintos λ, observamos que el mínimo se encuentra alrededor de $\lambda = -0.87$, que lleva la tasa de desvíos a ≈ 3.6:



Reducir la tasa de desvíos a 3.6 sigue sin ser suficiente para el criterio originalmente autoimpuesto, pero tampoco está demasiado lejos, y es una enorme mejoría respecto a la tasa original de más de 26, así que aplicaremos la transformación de Box-Cox y calificaremos momentáneamente de "tolerable" la heterocedasticidad restante. Lo que sí, para ayudar a la comprensión, en lugar de tomar el óptimo $\lambda^\star = -0.87$, tomaremos el mucho más interpretable $\lambda = -1$, que consigue una tasa de desvíos similar (4.12), pero provee una transformación más clara: $g(y) = 1 - \frac{1}{y}$, que cuando $y \in (1, +\infty) \Rightarrow g(y) \in (0, 1)$.

Aplicamos la transformación, y volvemos a graficar los perfiles.

```
## # A tibble: 11 x 2
##        l     t
##    <dbl> <dbl>
##  1 -1     4.12
##  2 -0.89  3.65
##  3 -0.88  3.61
##  4 -0.87  3.61
##  5 -0.86  3.62
##  6 -0.85  3.63
##  7 -0.84  3.63
##  8 -0.83  3.64
##  9 -0.82  3.65
## 10 -0.81  3.66
## 11  0     7.55
```

Como la variable respuesta y indica la *sobrevida* de los animales, un procedimiento razonable sería transformarla por el logaritmo natural, de manera que

Usando el comando `anova` sobre la salida de `lm`, obtenemos fácilmente una tabla con los datos necesarios para testear esta hipótesis. Por un lado, los coeficientes significativos al 5% del modelo multiplicativo ajustado son

term	estimate	p.value
(Intercept)	4.125	0.0000
trtB	4.675	0.0001
venenoIII:trtB	-3.425	0.0276
venenoIII	-2.025	0.0628
trtD	1.975	0.0692

No hay entrecruzamientos escandalosos, pareciera que B>D>C>A y I>II>III

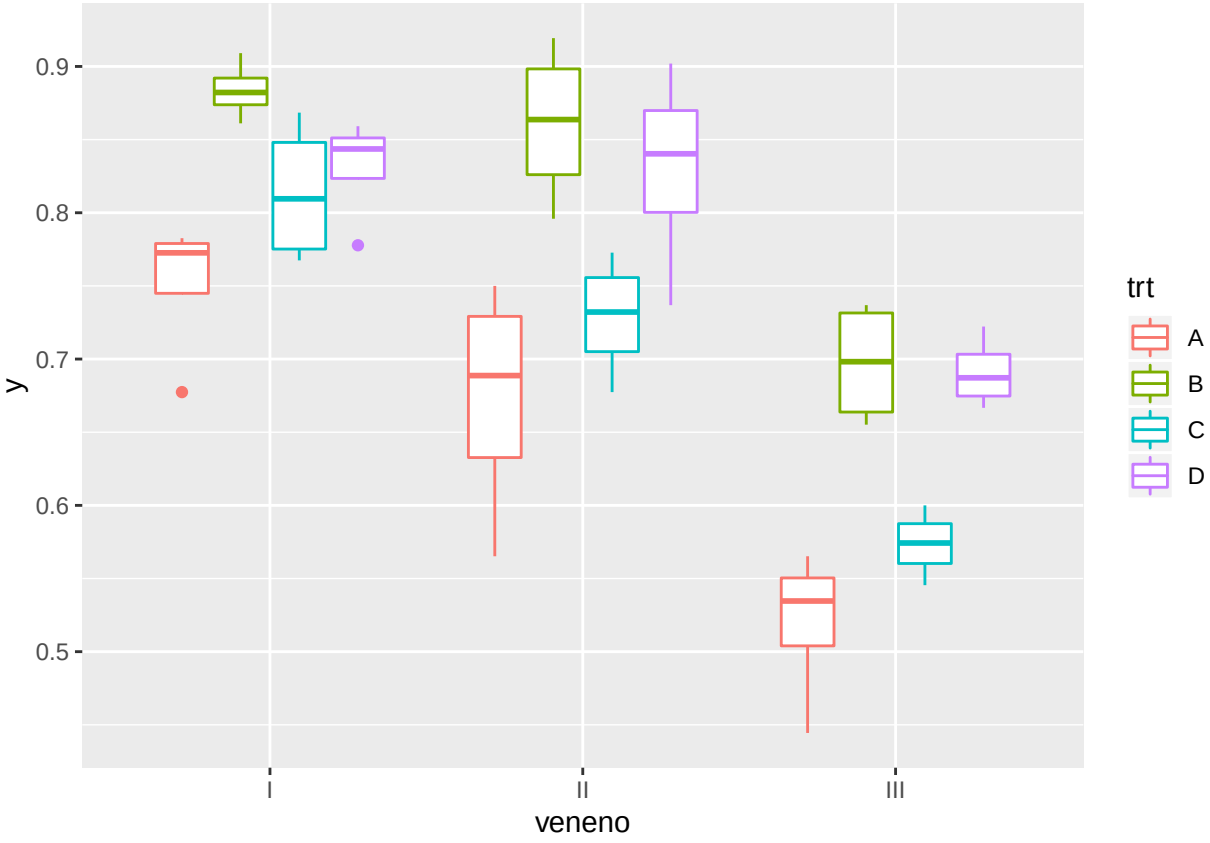
con un pequenio cruce entre I:D y II:D

```
## # A tibble: 4 x 6
##   term          df sumsq meansq statistic      p.value
##   <chr>      <int> <dbl>  <dbl>     <dbl>        <dbl>
## 1 veneno         2 103.   51.7      23.2   0.000000333
## 2 trt            3  92.1  30.7      13.8   0.00000378
## 3 veneno:trt     6  25.0   4.17      1.87  0.112
## 4 Residuals     36  80.1   2.22     NA    NA
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    4.52      0.559      8.09 4.22e-10
## 2 venenoIII     -3.41      0.559     -6.10 2.83e- 7
## 3 trtB           3.62      0.646      5.61 1.43e- 6
## 4 trtD           2.20      0.646      3.41 1.46e- 3
## 5 venenoII      -0.731     0.559     -1.31 1.98e- 1
## 6 trtC           0.783     0.646      1.21 2.32e- 1
```

```
##
## Call:
## lm(formula = y ~ veneno + trt, data = df2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118153 -0.027568 -0.002116  0.037619  0.082757
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.73023    0.01744  41.883  < 2e-16 ***
## venenoII    -0.04686    0.01744  -2.688  0.01026 *
## venenoIII   -0.19964    0.01744 -11.451 1.69e-14 ***
## trtB         0.16574    0.02013   8.233 2.66e-10 ***
## trtC         0.05721    0.02013   2.842  0.00689 **
## trtD         0.13583    0.02013   6.747 3.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04931 on 42 degrees of freedom
## Multiple R-squared:  0.8441, Adjusted R-squared:  0.8255
## F-statistic: 45.47 on 5 and 42 DF,  p-value: 6.974e-16
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq  Mean Sq F value    Pr(>F)
## veneno     2 0.34877 0.174386  71.708 2.865e-14 ***
## trt        3 0.20414 0.068048  27.982 4.192e-10 ***
## Residuals 42 0.10214 0.002432
## ---
## 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 = df2)
##
## $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
```