Comparaciones Multples no Ortogonales

Sebastián Sánchez Sandí

2025-08-01

Preparación

Los datos se encuentran en el archivo *uvas.csv*.

```
base = read.csv("uvas.csv")
str(base)
```

```
## 'data.frame':    100 obs. of  4 variables:
##  $ localidad: chr  "Garita" "Guacima" "San Vito" "Garita" ...
##  $ especie  : chr  "blanca" "roja" "roja" "roja" ...
##  $ diam     : num  1.2 0.9 0.8 0.7 1.6 1 1.2 1 1.1 1.3 ...
##  $ brix     : num  17 16.8 17.9 18.1 17.9 15 18.1 16.2 17 16 ...
```

Como lo que se quiere probar es si el dulzor varia dependiendo de la localidad, entonces el factor de diseño es la localidad.

```
base$localidad = factor(base$localidad)
```

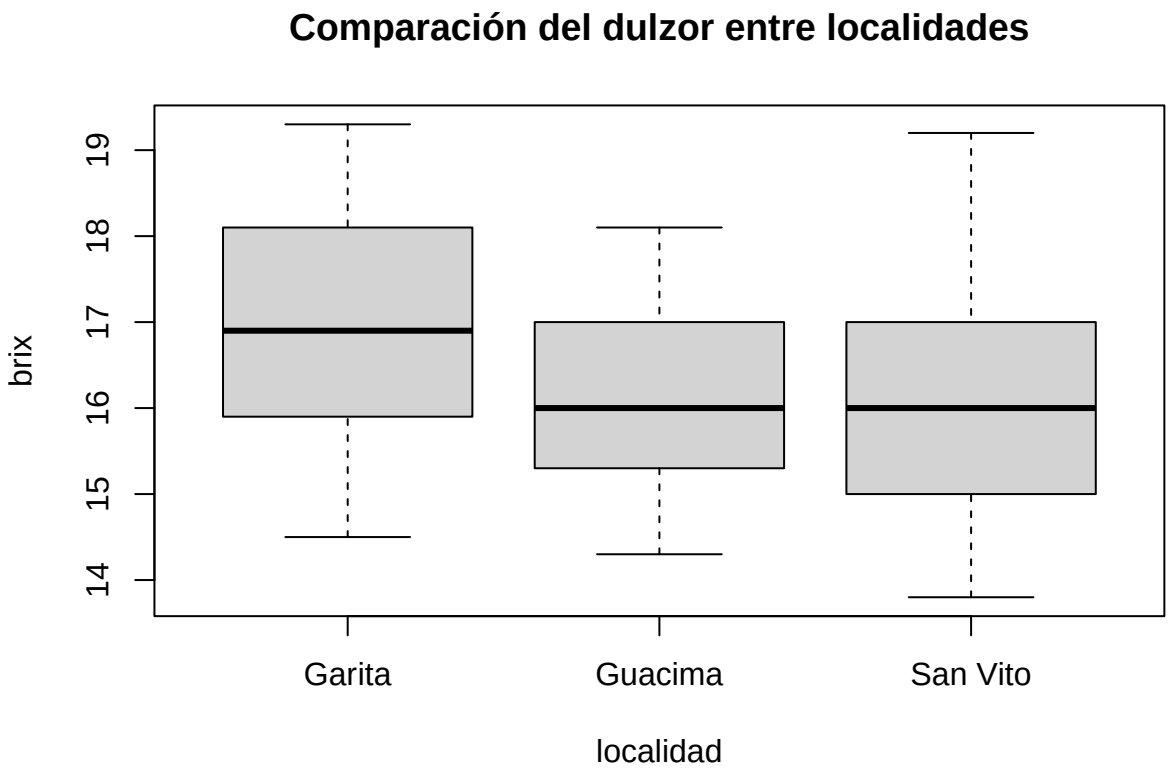
Hipótesis Básica

Para probar que las 3 localidades no producen el mismo dulzor planteamos las siguientes hipótesis:

- $H_0 = \mu_1 = \mu_2 = \mu_3$ con $H_a =$ al menos una media es distinta.

Ahora construimos un boxplot para visualizar la variabilidad de cada tratamiento.

```
boxplot(brix~localidad, data = base)
title("Comparación del dulzor entre localidades")
```



Podemos ver que parece que las medias entre la Guacima y San Vito son parecidas y la Garita está un poco por encima, pero al existir variabilidad no podemos afirmar esto. Estimamos las medias de cada tratamiento para visualizar diferencias.

```
m = with(data = base, tapply(brix, localidad, mean))
m
```

```
##   Garita  Guacima San Vito
## 16.95526 16.10400 15.97027
```

Las medias parecen respaldar la idea inicial que Garita produce en promedio uvas más dulces. Ponemos a prueba esta hipótesis con un análisis de varianza.

```
mod = lm(brix~localidad, data = base)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: brix
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## localidad  2  20.691 10.3454  5.7173 0.004496 **
## Residuals 97 175.521  1.8095
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como la probabilidad de cometer error tipo 1 es menor a 0.05, entonces rechazamos la hipótesis nula. Se espera que al menos una localidad produzca en promedio uvas más dulces.

Comparaciones de Promedios

Como en este caso se deben comparar todas las medias entre sí utilizamos tukey. Primero planteamos todas las hipótesis necesarias.

- $H_0 : \mu_1 = \mu_2$ con $H_a : \mu_1 \neq \mu_2$
- $H_0 : \mu_1 = \mu_3$ con $H_a : \mu_1 \neq \mu 3$
- $H_0 : \mu_2 = \mu_3$ con $H_a : \mu_2 \neq \mu_3$

Para verificar que estas hipótesis no son ortogonales realizamos el siguiente procedimiento. Si suponemos que el vector de coeficientes es $(\mu_1, \mu_2, \mu_3)^T$ entonces podemos multiplicar las hipótesis y revisar la ortogonalidad.

```
h1 = c(1, -1, 0)
h2 = c(1, 0, -1)
h3 = c(0, 1, -1)

h1 %*% h2
```

```
##      [,1]
## [1,]    1
```

```
h2 %*% h3
```

```
##      [,1]
## [1,]    1
```

```
h1 %*% h3
```

```
##      [,1]
## [1,]   -1
```

Como podemos ver no son ortogonales, por lo que requieren corrección de bonferroni. Con esto en mente calculamos el CMRes.

```
CMRes = anova(mod)[2, 3]
CMRes
```

```
## [1] 1.809493
```

Ahora calculamos los estadísticos de interés que serían todas las diferencias entre las medias en valor absoluto.

```
d1 = abs(m[1] - m[2])
d2 = abs(m[1] - m[3])
d3 = abs(m[2] - m[3])
diff = c(d1, d2, d3)
names(diff) = c("Ga-Gu", "Ga-Sv", "Gu-Sv")
diff
```

```
##     Ga-Gu     Ga-Sv     Gu-Sv
## 0.8512632 0.9849929 0.1337297
```

Ahora calculamos el error estándar de los contrastes.

```
r = table(base$localidad)
se1 = sqrt(CMRes * (1 / r[1] + 1 / r[2]))
se2 = sqrt(CMRes * (1 / r[1] + 1 / r[3]))
se3 = sqrt(CMRes * (1 / r[2] + 1 / r[3]))
se = c(se1, se2, se3)
se
```

```
##    Garita    Garita   Guacima
## 0.3464072 0.3106823 0.3482599
```

Con todo esto ya podemos calcular el estadístico estandarizado para obtener su probablidad con la distribución de Tukey.

```
q = diff / se
q
```

```
##     Ga-Gu     Ga-Sv     Gu-Sv
## 2.4574058 3.1704188 0.3839941
```

Con esto y como las hipótesis no son ortogonales, entonces utilizamos la distribución de Tukey para hacer esta corrección. Los grados de libertad de esta distribución son los grados de libertad del tratamiento y del residuo.

```
ptukey(q * sqrt(2), 3, 97, lower.tail = F)
```

```
##       Ga-Gu       Ga-Sv       Gu-Sv
## 0.041380986 0.005730055 0.922008844
```

Como la prueba ya hizo la correción entonces podemos comparar todo contra 0.05. En los casos de la primera y segunda hipótesis se rechazan ya que son menores a 0.05 y la tercera no se rechaza. En otras palabras, la media de dulzor de las uvas de Garita es distinta a las de San Vito y a la Guacima, pero San Vito y la Guacima no tienen diferencia.

Cabe destacar que todo este análisis puede hacerse de manera automática de la siguiente forma.

```
TukeyHSD(aov(mod))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mod)
##
## $localidad
##                         diff        lwr         upr     p adj
## Guacima-Garita   -0.8512632 -1.6757890 -0.02673733 0.0413810
## San Vito-Garita  -0.9849929 -1.7244854 -0.24550041 0.0057301
## San Vito-Guacima -0.1337297 -0.9626653  0.69520583 0.9220088
```

Límites para las diferencias

Como solamente debemos hacer 2 intrervalos tenemos que recurrir a la corrección de Bonferroni. Entonces tenemos que d = 2.

```
t = qt(1 - 0.05 / (2 * 2), 97)
lim = cbind(diff[1:2] - t * se[1:2], diff[1:2] + t * se[1:2])
lim
```

```
##             [,1]     [,2]
## Ga-Gu 0.06258823 1.639938
## Ga-Sv 0.27765400 1.692332
```

Para realizar la conclusión es importante definir que puede ser una diferencia relevante ya que en este caso la diferencia entre el promedio de dulzor de las uvas entre Garita y la Guacima puede llegar a ser tan baja como 0.06. Lo cual puede ser insignificante para el investigador. Finalmente, todas estas conclusiones se hacen con un 95%, no cada una por separado.