ANOVA de una vía

Sebastián Sánchez Sandí

2025-08-01

Preparacion

Los datos para trabajar en este laboratorio se encuentran en el archivo *manzanas.csv*.

```
base = read.csv("manzanas.csv", sep = ';')
str(base)
```

```
## 'data.frame':    40 obs. of  3 variables:
##  $ juez : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ trat : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ color: int  5 3 2 4 3 2 5 4 1 3 ...
```

Cambiamos el tipo de la columna trat de entero a factor para poder utilizar las herramientas para el análisis.

```
base$trat = factor(base$trat)
levels(base$trat) = c("control", "tapar", "bolsa", "limon")
```

Análisis Gráfico

Comenzamos el análisis calculando la media de los datos bajo cada tratamiento. Recordamos que $E[Y|trat] = \mu + \tau_{trat}$.

```
m = with(data = base, tapply(color, trat, mean))
m
```

```
## control   tapar   bolsa   limon
##     5.4     3.2     2.8     1.8
```

Esta es una estimación de la media de los datos bajo cada uno de los tratamientos. Ahora vamos a calcular la varianza bajo cada tratamiento.

```
v = with(data = base, tapply(color, trat, var))
v
```

```
##   control     tapar     bolsa     limon
## 0.4888889 1.7333333 1.0666667 0.6222222
```

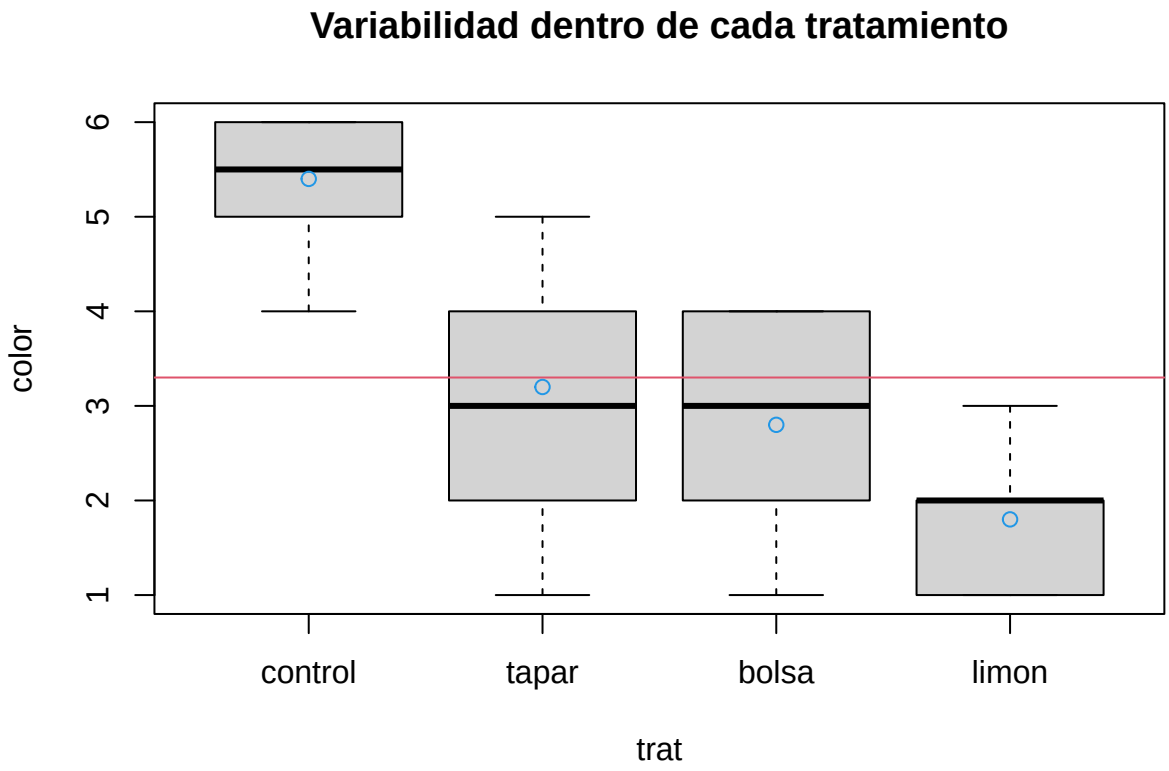
Finalmente obtenemos la media general.

```r
mediaGeneral = mean(base$color)
mediaGeneral
```

```
## [1] 3.3
```

Para analizar de forma gráfica la variabilidad dentro de cada grupo construimos un diagrama de cajas donde cada caja es un tratamiento.

```r
boxplot(color~trat, data = base)
abline(h = mediaGeneral, col = 2)
points(1:4, m, col = 4)
title("Variabilidad dentro de cada tratamiento")
```



Calculamos los efectos muestrales. Sabemos que $\mu_i = \mu + \tau_i$ entonces despejando el efecto obtenemos que $\tau_i = \mu_i - \mu$. Cabe destacar que se están trabajando con estimaciones por lo que la notación correcta sería $\hat{\tau}_i = \bar{y}_i - \bar{y}$.

```r
tau = m - mediaGeneral
tau
```

```
## control   tapar   bolsa   limon
##     2.1    -0.1    -0.5    -1.5
```

Retomando el gráfico anterior podemos ver que la distancia entre las medias (los puntos celestes) y la media (línea roja) son exactamente los efectos calculados. Esto quiere decir que el efecto de un tratamiento es que tanto mueven de forma postiva o negativa la media del tratamiento con respecto a la media general.

Teóricamente estos valores sumados deben dar como resultado cero por la carácteristica que tienen los efectos de suma nula.

```
round(sum(tau), 2)
```

```
## [1] 0
```

Finalmente calculamos una estimación de la varianza del error. Esta se calcula como $CMRes = \frac{SCRes}{n-k}$ donde $SCRes = \sum_{j=1}^{k}(r_j - 1)s_j^2$, n = número de observaciones, k = número de tratamientos, r = réplicas por tratamiento y s_j^2 = varianza del j-ésimo tratamiento. En el caso de que sea un experimento balanceado, cada tratamiento tenga la misma cantidad de réplicas, se puede demostrar que el CMRes es igual al promedio simple de las varianzas de los tratamientos.

```
table(base$trat)
```

```
##
## control   tapar   bolsa   limon
##      10      10      10      10
```

Como estamos ante un experimento balanceado, basta con promediar las varianzas para obtener el CMRes.

```
CMRes = mean(v)
CMRes
```

```
## [1] 0.9777778
```

Análisis de Varianza

Para el apartado de análisis R nos permite construir el modelo de dos formas: utilizando lm o aov. Si utilizamos lm podemos obtener los coeficientes del modelo y si utilizamos aov podemos obtner la tabla de efectos. Por preferencia personal vamos a utilizar lm.

```
mod = lm(color~trat, data = base)

anova(mod)
```

```
## Analysis of Variance Table
##
## Response: color
##           Df Sum Sq Mean Sq F value    Pr(>F)
## trat       3   69.2 23.0667  23.591 1.278e-08 ***
## Residuals 36   35.2  0.9778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Con esto obtenemos la tabla del ANOVA. Los grados de libertad de tratamiento son 3 ya que tenemos 4 tratamientos y los grados de libertad de los residuales son 36 ya que tenemos 40 observaciones y 4 tratamientos. La suma de cuadrados de tratamientos nace de $SCRes = \sum_{j=1}^{k} r_j \hat{\tau_j}^2$ y la suma de cuadrados residual

como se discutió anteriormente. El estadístico F se obtiene de dividir el CMTrat entre el CMRes y este se evalua en la función de distrubución F para obtener el valor p. Este valor se compara contra el nivel de signficancia (usualmente 0.05).

Ahora, como el objetivo general del estudio es seleccionar el tratamiento que mejor mantenga el color de la manzana planteamos las siguientes hipótesis.

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ con H_a : al menos una media es diferente.
- $H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$ con H_a : al menos un efecto es distinto a cero.

Ambas son equivalentes. Volviendo a la tabla anterior podemos ver que el valor p es menor a 0.05. Como la probabilidad de cometer error tipo 1 es menor a 0.05, entonces rechazamos la hipótesis nula. Se espera que al menos un tratamiento tenga un efecto distinto a 0.

Estimación de Parámetros

Cuando construimos el modelo tenemos dos opciones: modelo de suma nula o modelo de tratamiento de referencia. Vamos a analizar primero el modelo de tratamiento de referencia.

```
mod$coefficients
```

```
## (Intercept)   trattapar   tratbolsa   tratlimon
##         5.4        -2.2        -2.6        -3.6
```

En este caso el intercepto representa la media bajo el primer tratamiento, R toma esta media como referencia. El resto de valores son las diferencias entre la media bajo otro tratamiento y bajo el tratamiento de referencia $\delta_j = \mu_j - \mu_1$.

```
#model.matrix(mod)
contrasts(base$trat)
```

```
##         tapar bolsa limon
## control     0     0     0
## tapar       1     0     0
## bolsa       0     1     0
## limon       0     0     1
```

Esta es la matriz de estructura del modelo. Esta contiene las variables dummies que nos indican a cual tratamiento pertenecen.

Ahora pasamos al modelo de suma nula.

```
options(contrasts = c("contr.sum", "contr.poly"))
contrasts(base$trat)
```

```
##         [,1] [,2] [,3]
## control    1    0    0
## tapar      0    1    0
## bolsa      0    0    1
## limon     -1   -1   -1
```

Como podemos ver las variables dummies ahora cambian ya que los coeficientes ahora representan los efectos de los tratamientos.