PRINCIPIOS DE DISEÑOS EXPERIMENTALES

Lab. No.2 - ANALISIS DE VARIANCIA DE UNA VIA

Manzanas

Las manzanas tienen un compuesto llamado polifenol oxidasa que hace que se oscurezcan rápidamente en contacto con el aire una vez cortadas. Para evitar el pardeamiento se probaron tres tratamientos: 1) tapar, 2) poner en bolsa plástica cerrada, y 3) aplicar jugo de limón. Se incluyó además un control sin aplicar nada (4). Una vez aplicados los tratamientos el resultado fue evaluado por 10 jueces que calificaron el color en una escala de 1 a 6 donde 1 es el color normal de la fruta y 6 es el más oscuro. El objetivo final es seleccionar el tratamiento que mantenga mejor el color original para una empresa que se encarga de banquetes para altos ejecutivos.

En un primer análisis sólo se va a investigar si hay alguna diferencia en el color promedio resultante con los cuatro tratamientos.

Ejercicios

1. Lea el archivo `manzanas.csv` en R.

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

---

2. Defina correctamente el factor y ponga las etiquetas correspondientes a cada tratamiento. Salve la base en un archivo llamado `manzanas.Rdata` para ser utilizado en futuros laboratorios.

```
base$trat
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4
## [36] 4 4 4 4 4
```

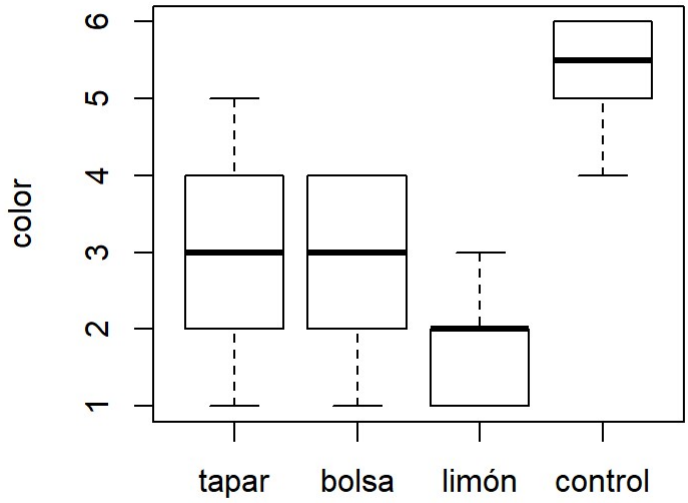
```
base$trat=factor(base$trat)
levels(base$trat)=c("tapar","bolsa","limón","control")
base$trat
```

```
##  [1] tapar   tapar   tapar   tapar   tapar   tapar   tapar   tapar
##  [9] tapar   tapar   bolsa   bolsa   bolsa   bolsa   bolsa   bolsa
## [17] bolsa   bolsa   bolsa   bolsa   limón   limón   limón   limón
## [25] limón   limón   limón   limón   limón   limón   control control
## [33] control control control control control control control control
## Levels: tapar bolsa limón control
```

```
save(base,file="manzanas.Rdata")
attach(base)
```

- Haga un boxplot para analizar los efectos de cada tratamiento sobre la respuesta.

```
boxplot(color~trat,ylab="color")
```



Se nota que los puntajes de color son mucho más bajos en los tres tratamientos que en el control. Cuando se aplicó limón estos puntajes tienen a ser más bajos que cuando se cubrió de alguna forma. También se nota que los dos tratamientos en que se cubrió producen resultados muy similares.

- Obtenga una tabla con las medias de la respuesta por tratamiento, llame a este objeto `m`.

```
m=tapply(color,trat,mean)
round(m,2)
```

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

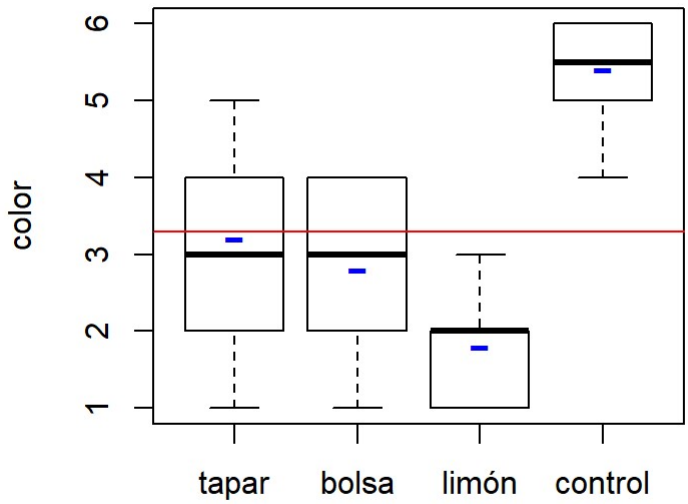
- Obtenga una tabla con las variancias de la respuesta por tratamiento, llame a este objeto `v`.

```
  v=tapply(color,trat,var)
  round(v,2)
```

```
##   tapar   bolsa   limón control
##    1.73    1.07    0.62    0.49
```

- Obtenga la media general de la respuesta y agréguela al boxplot usando:
  `abline(h=media, col=2)`
- Agregue las medias de los tratamientos al boxplot usando:
  `points(1:4,m,col=4,pch="-",cex=2)` .

```
  media=mean(color)
  boxplot(color~trat,ylab="color")
  abline(h=media, col=2)
  points(1:4,m,col=4,pch="-",cex=2)
```



- Obtenga los efectos muestrales de cada tratamiento a partir de la tabla de medias y compare estos resultados con lo que ve en el gráfico. Recuerde que el efecto del tratamiento j-ésimo se define como: $\tau_j = \mu_j - \mu$. Cada efecto se puede estimar como: $\hat{\tau}_j = \bar{y}_j - \bar{y}$, donde $\bar{y}$ representa la media general de la respuesta y $\bar{y}_j$ la media de la respuesta en el j-éstimo tratamiento.

```
  ef=m-media
  ef
```

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

Estos números coinciden con el gráfico puesto que los valores negativos concuerdan con aquellas medias que están por debajo de la media general y el valor positivo del control concuerda con el gráfico en que su media está por encima de la media general.

- Explique el significado de cada uno de los valores obtenidos para los efectos muestrales.

El control tiene una media que está 2.1 puntos sobre la media general por lo que se dice que el control tiene el efecto de subir la media 2.1 puntos. El limón produce una media 1.5 puntos por debajo de la media general, es decir, tiene el efecto de bajar la media 1.5 puntos. Similarmente, los dos tratamientos en que se cubre tienen un leve efecto sobre la media ya que la bajan muy poco.

- Obtenga la suma de los efectos anteriores.

```
  sum(ef)
```

```
## [1] 1.110223e-15
```

Aunque no da exactamente cero, esto se debe a un asunto de computacional pero la suma de los efectos debe ser siempre cero por su misma construcción.

- Obtenga una estimación de la variancia del error a partir de la tabla de variancias. La estimación debe ser la media ponderada de las variancias en los tratamientos, las cuales se ponderan con los grados de libertad, sin embargo, en este caso se tiene el mismo número de réplicas en todos los tratamientos, por lo que basta hacer un promedio simple de las variancias.

```
  n=table(trat)
  n
```

```
## trat
##   tapar   bolsa   limón control
##      10      10      10      10
```

```
  v1=sum((n-1)*v)/(sum(n)-4)
  round(v1,2)
```

```
## [1] 0.98
```

```
  v2=mean(v)
  round(v2,2)
```

```
## [1] 0.98
```

- Justifique por qué la variancia del error es igual a la variancia de la respuesta en cada tratamiento.

Por la definición de error se tiene que:

$$\epsilon_{ij} = y_{ij} - \mu_j \Rightarrow V(\epsilon|X_j) = V(y|X_j, \mu_j) = V(y|X_j)$$

3. Ajuste un modelo lineal. Se puede usar tanto la función `aov` como la función `lm`, la diferencia principal es que con `lm` se pueden obtener los coeficientes del modelo, mientras que con `aov` se puede obtener la tabla de efectos. En todo caso si se usa `lm`,por ejemplo `mod=lm(y~x)`, luego se puede obtener `mod1=aov(mod)` de la misma forma que haciendo `mod1=aov(y~x)`.

- Obtenga los resultados del análisis de variancia mediante `anova(mod)` o `anova(mod1)`. Si se usó la función `aov` da lo mismo usar `summary(mod1)` o `anova(mod1)`.

```
mod=lm(color~trat)
mod1=aov(color~trat)
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
```

```
summary(mod1)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## trat         3   69.2  23.067   23.59 1.28e-08 ***
## Residuals   36   35.2   0.978
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod1)
```

```
## 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
```

- Observe la línea de residuales para obtener el cuadrado medio residual y compárelo con la estimación de la variancia del error obtenida en el punto anterior.

```
v3=anova(mod)[2,3]
round(v3,2)
```

```
## [1] 0.98
```

Se obtiene el mismo valor que se tenía al promediar las variancias, es decir que el cuadrado medio residual es una medida de la variabilidad de color dentro de cada tratamiento.

- Observe los grados de libertad residuales y justifique por qué se obtiene ese número.

Hay 36 grados de libertad en los residuales ya que se cuenta con 40 datos pero se usaron 4 grados de libertad para calcular los promedios de los 4 tratamientos y a partir de ahí obtener los residuales dentro de cada tratamiento. Entonces quedan 40-4=36 grados de libertad.

- Observe la línea del tratamiento y obtenga la suma de cuadrados de tratamiento.

```
anova(mod)[1,2]
```

```
## [1] 69.2
```

- Haga la suma de los cuadrados de los efectos obtenidos anteriormente. Observe que estos cuadrados deben multiplicarse por el número de réplicas para obtener exactamente la suma de cuadrados de tratamiento. Justifique por qué esto debe ser así.

```
sum(10*ef^2)
```

```
## [1] 69.2
```

En la descomposición de la suma de cuadrados total se tiene una parte que va del promedio del tratamiento al promedio general, esa cantidad es la misma para todos los valores de un mismo tratamiento por lo que debe repetirse tantas veces como datos haya en ese tratamiento. De ahí viene que esa distancia o efecto deba multiplicarse por r_j , el número de réplicas en el j-ésimo tratamiento.

- Compare la variabilidad de los promedios con la variabilidad residual para determinar si hay alguna evidencia de diferencias entre las medias de la respuesta.

```
cmtrat=anova(mod)[1,3]
cmres=anova(mod)[2,3]
f=cmtrat/cmres
round(f,2)
```

```
## [1] 23.59
```

```
p=pf(f,3,36,lower.tail = F)
round(p,4)
```

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

Se observa que la variabilidad entre las medias de los tratamientos es 23.6 veces la de los residuales, lo cual es una cantidad enorme. Esto va a favor de pensar que las medias están alejadas unas de otras.

- Establezca adecuadamente la hipótesis que está poniendo a prueba y dé una conclusión.

$H_O : \mu_1 = \mu_2 = \mu_3 = \mu_4 \qquad o \qquad \tau_1 = \tau_2 = \tau_3 = \tau_4 = 0$

$H_1 : al\ menos\ una\ de\ las\ medias\ es\ diferente$

Puesto que la probabilidad asociada al error tipo I es casi cero, se puede esperar que si se rechaza la hipótesis nula, la probabilidad de estar cometiendo un error sea ínfima, entonces se toma la decisión de rechazar esa hipótesis tranquilamente. Por lo tanto, se puede esperar que no todas las medias sean iguales (se sospecha que la del control sea la que se comporta diferente).

---

4. Obtenga las estimaciones de los parámetros del modelo. Esto se logra con el ajuste que se hizo con `lm` mediante `summary(mod)` o `mod$coef`.

```
mod$coef
```

```
## (Intercept)   tratbolsa   tratlimón tratcontrol
##         3.2        -0.4        -1.4         2.2
```

- Escriba el modelo que está usando R (use notación matemática con letras griegas).

$$\mu_j = \mu + \delta_j$$

Aquí se usa la restricción: $\delta_1 = 0$ **y la definición** $\delta_j = \mu_j - \mu_1$ **.**

- ¿Qué significa el intercepto en este modelo?

```
m
```

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

Puesto que se usa el modelo con el tratamiento *tapar* como referencia, el intercepto coincide con la media de ese tratamiento que es justamente 3.2.

- ¿Qué representa cada uno de los coeficientes del modelo?

Los otros coeficientes representan la distancia que hay de la media de cada uno de los otros tratamientos con respecto a la media del tratamiento *tapar*.

- Obtenga la matriz de estructura.

```
model.matrix(mod)
```

```
##    (Intercept) tratbolsa tratlimón tratcontrol
## 1            1         0         0           0
## 2            1         0         0           0
## 3            1         0         0           0
## 4            1         0         0           0
## 5            1         0         0           0
## 6            1         0         0           0
## 7            1         0         0           0
## 8            1         0         0           0
## 9            1         0         0           0
## 10           1         0         0           0
## 11           1         1         0           0
## 12           1         1         0           0
## 13           1         1         0           0
## 14           1         1         0           0
## 15           1         1         0           0
## 16           1         1         0           0
## 17           1         1         0           0
## 18           1         1         0           0
## 19           1         1         0           0
## 20           1         1         0           0
## 21           1         0         1           0
## 22           1         0         1           0
## 23           1         0         1           0
## 24           1         0         1           0
## 25           1         0         1           0
## 26           1         0         1           0
## 27           1         0         1           0
## 28           1         0         1           0
## 29           1         0         1           0
## 30           1         0         1           0
## 31           1         0         0           1
## 32           1         0         0           1
## 33           1         0         0           1
## 34           1         0         0           1
## 35           1         0         0           1
## 36           1         0         0           1
## 37           1         0         0           1
## 38           1         0         0           1
## 39           1         0         0           1
## 40           1         0         0           1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$trat
## [1] "contr.treatment"
```

En esta matriz de estructura se tienen variables auxiliares con 0 y 1 solamente.

- A partir de los coeficientes obtenidos, obtenga los efectos muestrales y compárelos con los obtenidos en el punto 1.

En esta forma no es inmediato obtener los efectos. Por la forma en que están definidos los coeficientes debe restarse la media general y sumarse el intercepto a cada coeficiente para obtener el efecto respectivo.

```
  mod$coef[2:4]+mod$coef[1]-media
```

```
##   tratbolsa   tratlimón tratcontrol
##        -0.5        -1.5         2.1
```

```
  mod$coef[1]-media
```

```
## (Intercept)
##        -0.1
```

- Obtenga los efectos directamente con `model.tables(mod)` (sólo funciona si el modelo fue hecho con la función `aov` .

```
  model.tables(mod1)
```

```
## Tables of effects
##
##  trat
## trat
##   tapar   bolsa   limón control
##    -0.1    -0.5    -1.5     2.1
```

---

5. El modelo se puede parametrizar de dos formas:

- **Tratamiento referencia**: se asume que el coeficiente de uno de los tratamientos es cero. Esta forma es la que R usa por default.
- **Suma nula**: se asume que la suma de los coeficientes de todos los tratamientos es cero. En tal caso se estima un coeficiente menos que la cantidad de niveles del factor ya que el restante sale por diferencia:

$$\sum_{j=1}^{k} \tau_j = 0 \Rightarrow \tau_1 = -\sum_{j=2}^{k} \tau_j$$

- Cambie al modelo de **suma nula** usando la siguiente instrucción:
  `options(contrasts=c("contr.sum","contr.poly"))` .

  Para volver al modelo de **tratamiento referencia** se usa:
  `options(contrasts=c("contr.treatment","contr.poly"))` .

- Verifique la codificación con `contrasts(trat)` .

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

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

- Repita los pasos del punto 4. Compare los resultados.

Estimaciones:

```
mod2=lm(color~trat)
mod2$coef
```

```
## (Intercept)       trat1       trat2       trat3
##         3.3        -0.1        -0.5        -1.5
```

Modelo: el modelo se escribe $\mu_j = \mu + \tau_j$ y la restricción en los parámetros cambia ya que ahora se debe establecer que: $\tau_4 = -(\tau_1 + \tau_2 + \tau_3)$.

Intercepto: ahora el intercepto representa la media general que es 3.3.

Coeficientes: los otros coeficientes representan el efecto que tiene cada tratamiento, es decir la diferencia entre la media de un tratamiento respecto a la media general. Aparecen sólo 3 efectos puesto que el cuarto se obtiene a partir de la restricción.

Matriz de estructura: ahora las variables auxiliares contienen un -1 en el tratamiento de referencia que en este caso es el cuarto.

```
model.matrix(mod2)
```

```
##    (Intercept) trat1 trat2 trat3
## 1            1     1     0     0
## 2            1     1     0     0
## 3            1     1     0     0
## 4            1     1     0     0
## 5            1     1     0     0
## 6            1     1     0     0
## 7            1     1     0     0
## 8            1     1     0     0
## 9            1     1     0     0
## 10           1     1     0     0
## 11           1     0     1     0
## 12           1     0     1     0
## 13           1     0     1     0
## 14           1     0     1     0
## 15           1     0     1     0
## 16           1     0     1     0
## 17           1     0     1     0
## 18           1     0     1     0
## 19           1     0     1     0
## 20           1     0     1     0
## 21           1     0     0     1
## 22           1     0     0     1
## 23           1     0     0     1
## 24           1     0     0     1
## 25           1     0     0     1
## 26           1     0     0     1
## 27           1     0     0     1
## 28           1     0     0     1
## 29           1     0     0     1
## 30           1     0     0     1
## 31           1    -1    -1    -1
## 32           1    -1    -1    -1
## 33           1    -1    -1    -1
## 34           1    -1    -1    -1
## 35           1    -1    -1    -1
## 36           1    -1    -1    -1
## 37           1    -1    -1    -1
## 38           1    -1    -1    -1
## 39           1    -1    -1    -1
## 40           1    -1    -1    -1
## attr(,"assign")
## [1] 0 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$trat
## [1] "contr.sum"
```

Efectos: los coeficientes son directamente los efectos, salvo el cuarto que se tiene que obtener a partir de los otros.

```
mod2$coef[2:4]
```

```
## trat1 trat2 trat3
##  -0.1  -0.5  -1.5
```

```
  -sum(mod2$coef[2:4])
```

```
## [1] 2.1
```

Efectos directamente:

```
  model.tables(aov(mod2))
```

```
## Tables of effects
##
##  trat
## trat
##   tapar   bolsa   limón control
##    -0.1    -0.5    -1.5     2.1
```