PRINCIPIOS DE DISEÑO EXPERIMENTAL: ANÁLISIS DE VARIANZA DE UNA VÍA

ISADORE NABI

I. FUNDAMENTOS TEÓRICOS GENERALES DEL DISEÑO DE EXPERIMENTOS

I.I. Conceptos Fundamentales: Diseño Experimental, Factor, Niveles, Tratamiento, Unidad Experimental y Variables Confusoras

Incluye el diseño de todos los ejercicios de recolección de información donde hay variación y donde el experimentador puede tener completo control o no. El

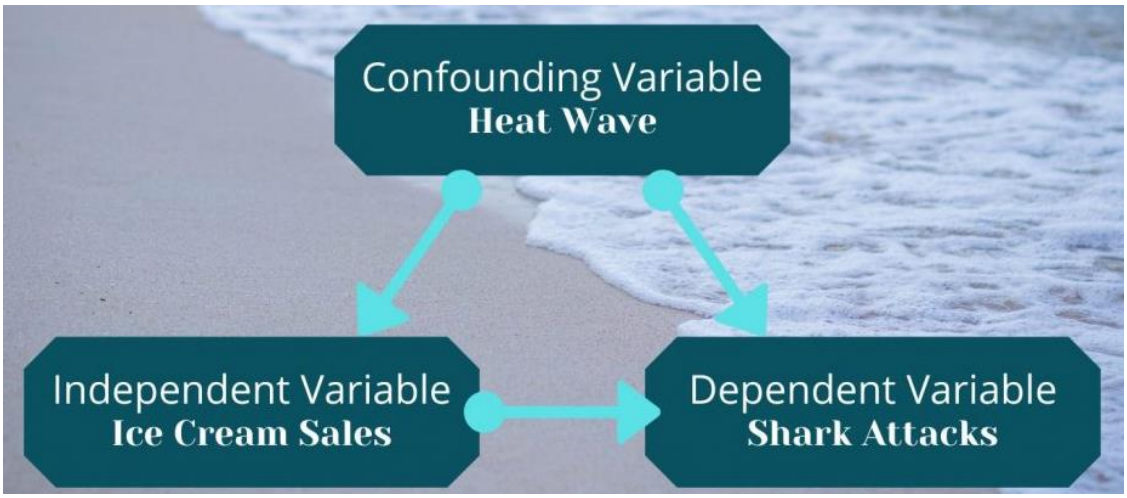
experimentador está interesado en el efecto de algún proceso o intervención llamado *tratamiento* sobre algunos objetos llamados *unidades experimentales*.

En este contexto, un *factor* es una variable categórica (o categorizada) con varios niveles para la cual se quiere investigar el efecto que tiene un cambio en ella sobre una variable respuesta. Los factores tienen "niveles" de aplicación, donde los niveles implican una cantidad o magnitud; *i.e*, si se les aplica dosis de 5mg, 10mg y 15mg de un medicamento a las unidades experimentales, estas cantidades son los niveles del factor. También se usa el término "nivel" para variables categóricas; *i.e*, en un experimento donde se prueban tres tipos de droga, cada tipo de droga administrada es un nivel del factor (A, B, o C).

En experimentos, un tratamiento es algo que los investigadores administran a las unidades experimentales. Cuando un experimento tiene un solo factor los tratamientos son los niveles del factor, en cambio, si el experimento tiene dos o más factores, cada tratamiento es la combinación de los niveles de los factores. Las unidades experimentales pueden ser personas, ratas, muestras de material, pedazos de tierra o paquetes de salchichas con las que el investigador trabaja y a las que se les aplica los tratamientos bajo estudio.

Una *variable confusora* está relacionada tanto con la pertenencia al grupo como con la respuesta. Su presencia hace difícil establecer que la respuesta sea consecuencia directamente de la pertenencia al grupo. Así, una variable confusora es una variable que influencia tanto a los predictores como a la respuesta y que conduce a asociación espuria, *i.e.*, es fuente de asociación espuria.

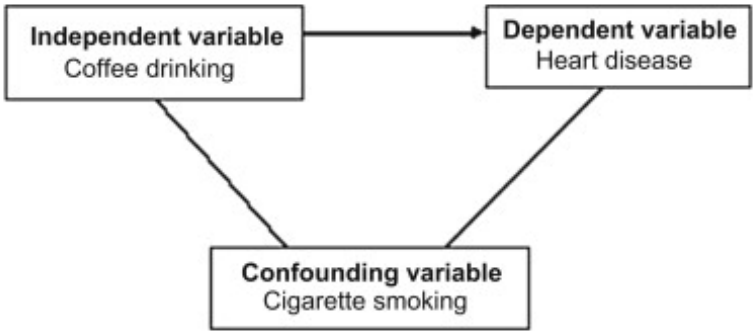
```
knitr::include_graphics("IMG1.png")
```



#*Figura 1: Variables Confusoras (Confounding Variables)*

#*Fuente:* (Helmenstine, 2020).

```
knitr::include_graphics("IMG2.png")
```



#*Figura 2: Variables Confusoras (Confounding Variables)*

#*Fuente:* (Science Direct, 2022).

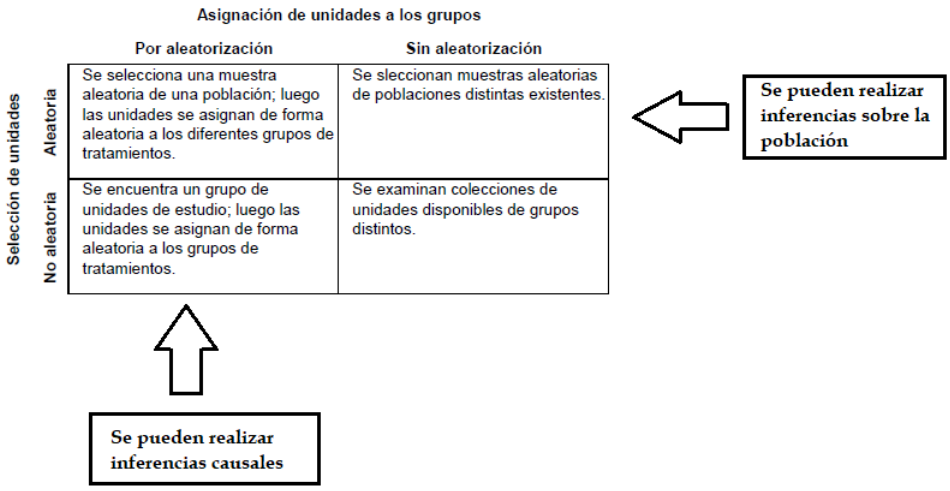
I.II. Consideraciones en el Diseño de Experimentos

Aunque no se requiere tener igual cantidad de unidades experimentales bajo cada tratamiento, la potencia de las pruebas de hipótesis aumenta conforme las muestras tengan tamaños tan parecidos como sea posible. El desempeño de las pruebas será mejor conforme las unidades experimentales sean más similares en

todos los aspectos excepto en el tratamiento aplicado; sin embargo, para obtener resultados más generalizables se pueden incluir factores de ruido controlables o medibles que pueden ser considerados en el modelo. En un diseño totalmente experimental, cada unidad experimental debe ser asignada aleatoriamente a cada uno de los tratamientos bajo estudio. En la práctica no siempre es posible hacer una asignación aleatoria de las unidades a los tratamientos. Esto limita el alcance del diseño. Los factores de ruido que no se pueden controlar se deben medir para tomarse en cuenta en el momento del análisis de tal forma que no oscurezcan los resultados.

I.III. Aleatorización

```
knitr::include_graphics("IMG3.png")
```



#Figura 3: Sobre la Aleatorización en el Diseño de Experimentos

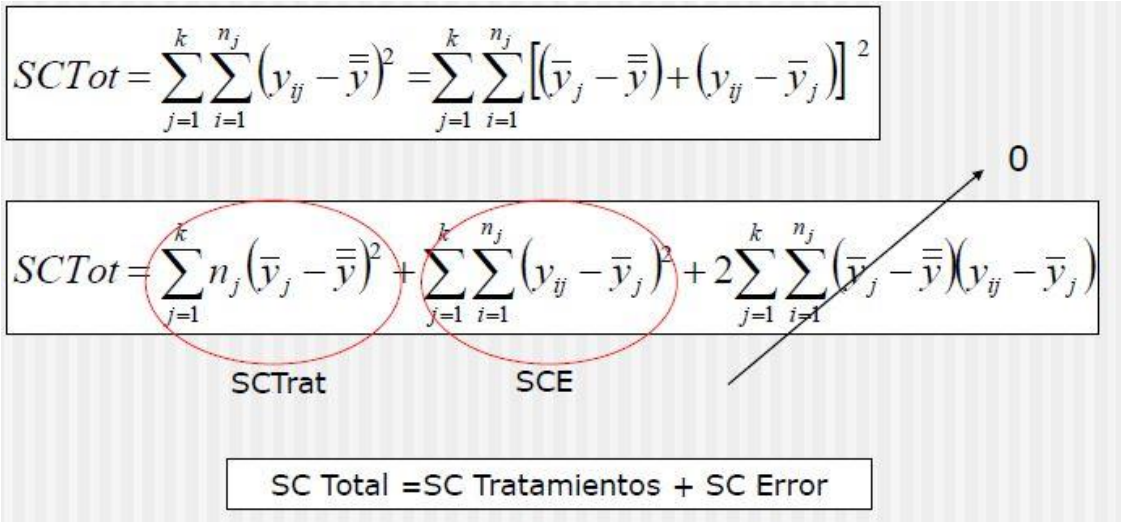
I.IV. Error Experimental

Las diferencias entre unidades experimentales que están bajo un mismo tratamiento se atribuyen solamente a un error experimental aleatorio o a la presencia de un factor de ruido no considerado en el experimento que esté

causando variaciones. Aun cuando se incluyan todos los factores que puedan estar influyendo la respuesta, se espera que exista variación entre unidades experimentales que están bajo un mismo tratamiento. El error experimental es inherente a todo proceso.

I.V. Descomposición de Suma de Cuadrados Totales

```
knitr::include_graphics("IMG4.JPG")
```



#Figura 4: Descomposición de la Suma de Cuadrados Totales del Tratamiento

I.VI. Distribución de Probabilidad de las Sumas de Cuadrados y Teorema de Cochran

En William Cochran (Cochran, 1934) estableció un teorema para justificar formalmente los resultados relativos a las distribuciones de probabilidad que se usan para realizar Análisis de Varianza (ANOVA).

Como señala (Cochran, 1934, pág. 178), muchas de las aplicaciones más utilizadas de la teoría de la estadística, como por ejemplo los métodos de análisis de varianza y covarianza, la prueba general de regresión múltiple y la prueba de un coeficiente de regresión, dependen esencialmente de la distribución conjunta de varias formas

cuadráticas. en un sistema normal univariante. El objetivo de la investigación de Cochran fue probar los principales resultados relevantes sobre esta distribución.

Adicionalmente, como aplicación de estos resultados, profundiza en la teoría involucrada en el método de análisis de covarianza. El sistema probabilístico normal con media nula y varianza unitaria usado por Cochran es uno tal que las regresoras son independientes e idénticamente distribuidas, el cual es válido, según el autor, para todo tipo de formas cuadráticas[1].

El teorema de Cochran establece que las formas cuadráticas Q_i son independientes y cada una de ellas tiene una distribución chi-cuadrado con r_i grados de libertad, en donde los grados de libertad son calculados como el rango de la matriz en la que se expresa cada forma cuadrática. De manera menos formal, tales grados de libertad son el número de combinaciones lineales incluidas en la suma de cuadrados que definen Q_i, siempre que estas combinaciones lineales sean linealmente independientes (Cochran, 1934, pág. 180).

Así, la suma de cuadrados totales sigue una distribución χ^2 con $n-1$ grados de libertad, la suma de cuadrados del tratamiento sigue una distribución χ^2 con $k-1$ grados de libertad y, finalmente, la suma de cuadrados del error sigue una distribución χ^2 con $n-k$ grados de libertad, donde k es el número de coeficientes de regresión a estimar (que representan los efectos y la media global -equivalente al intercepto de un modelo econométrico-).

I.VII. Contraste de Hipótesis

En este contexto, el contraste de hipótesis consiste en verificar si las medias obtenidas bajo todos los tratamientos son iguales o si, por el contrario, al menos

---

[11] Una forma cuadrática es un polinomio con términos todos de grado dos ("forma" es otro nombre para un polinomio homogéneo). Por ejemplo, $4x^2 + 2xy - 3y^2$ es una forma cuadrática en las variables x e y). A estas formas cuadráticas Cochran las denota como Q_i (Cochran, 1934, pág. 178).

uno de los tratamientos produce un promedio diferente a los demás. En tal caso se dice que el factor tiene un efecto sobre la media global de la variable respuesta. Se busca determinar si al menos un par de tratamientos producen medias estadísticamente diferentes entre sí.

II. CASO DE APLICACIÓN: ANÁLISIS DEL PROCESO DE OXIDACIÓN DE LAS MANZANAS

II.I. Efectos Individuales y Efectos Agregados en los Tratamientos

Las manzanas tienen un compuesto llamado *polifenol oxidasa*, el cual hace que se oscurezcan rápidamente en contacto con el aire una vez cortadas, que es como se presenta visualmente el proceso su oxidación. Para evitar el pardeamiento, se probaron tres tratamientos y el tratamiento de control:

1. Taparla.
2. Ponerla en bolsa plástica cerrada.
3. Aplicarle jugo de limón.
4. Tratamiento de control (sin aplicar nada).

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 interesa investigar si existe alguna diferencia en el color promedio resultante con los cuatro tratamientos. Para ello, es recomendable definir correctamente el factor y aplicar las etiquetas correspondientes a cada tratamiento; este archivo así creado puede guardarse en el disco duro del equipo para ser utilizado en futuras ocasiones.

```
base=read.csv("manzanas.csv",sep=";")
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 4 4
## [39] 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   tapar
## [10] tapar   bolsa   bolsa   bolsa   bolsa   bolsa   bolsa   bolsa   bolsa
## [19] bolsa   bolsa   limón   limón   limón   limón   limón   limón   limón
## [28] limón   limón   limón   control control control control control control
## [37] control control control control
## Levels: tapar bolsa limón control
```

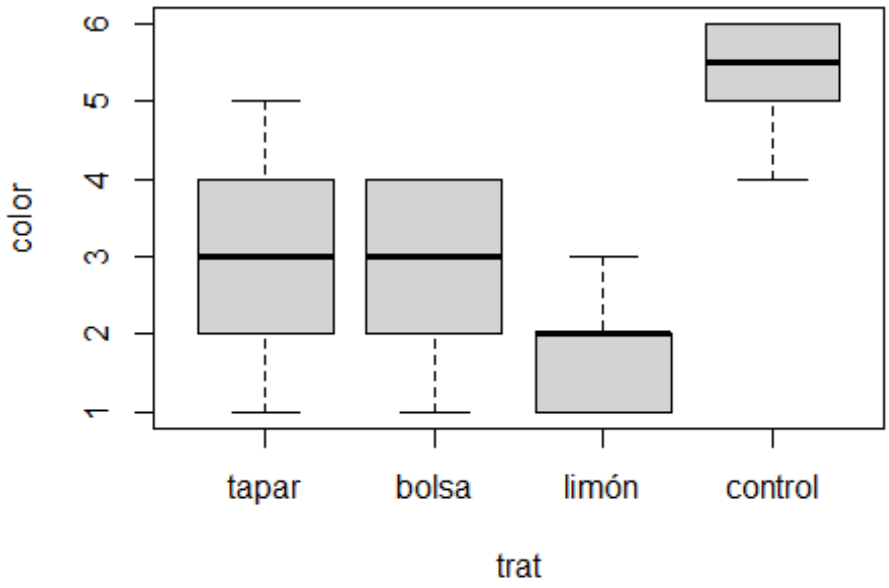
```
save(base,file="manzanas.Rdata")
attach(base)
```

Con la finalidad de realizar un estudio preliminar sobre los efectos de cada tratamiento en la variable de respuesta, puede construirse un boxplot[2], donde la línea negra en cada caja representa la mediana de la distribución.

---

[2] Conocido en la literatura en español como *gráfico de caja*, es aquel gráfico que incluye, en forma de cajas, la información relativa a el valor mínimo, el primer cuartil (percentil 25), el valor mediano, el tercer cuartil (percentil 75) y el valor máximo. La construcción estos obedece al siguiente algoritmo: 1. Dibujar una caja desde el primer hasta el tercer cuartil, 2. dibujar una línea vertical en la mediana, 3. dibujar "bigotes" desde los cuartiles hasta el mínimo y valor máximo. Los escenarios en que los boxplot son de utilidad son: 1. Para visualizar la distribución de valores en un conjunto de datos; un diagrama de caja nos permite visualizar rápidamente la distribución de valores en un conjunto de datos y ver dónde se encuentran los valores de resumen de cinco números. 2. Para comparar dos o más distribuciones; los diagramas de caja, uno al lado del otro, permiten visualizar las diferencias entre dos o más distribuciones y comparar los valores medianos y la dispersión de valores entre distribuciones. 3. Para identificar valores atípicos; en los diagramas de caja, los valores atípicos suelen estar representados por pequeños círculos que se extienden más allá de cualquiera de los bigotes. Una observación se define como un valor atípico si cumple con uno de los siguientes

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



Se observa que los puntajes de color son mucho más bajos en los tres tratamientos que en el del grupo de control (en el que no se hace nada). Cuando se aplicó limón estos puntajes tienden a ser más bajos que cuando se cubrió de alguna forma. También se observa que los dos tratamientos en que se cubrió la manzana producen resultados muy similares.

Complementariamente, puede construirse una tabla con las medias de la respuesta por tratamiento y llamar al objeto así construido m.

---

criterios: a. una observación es menor que Q1 – 1.5*(rango intercuartílico), b. una observación es mayor que Q3 + 1,5* (rango intercuartílico). Al crear un diagrama de caja, es posible visualizar rápidamente si una distribución tiene valores atípicos o no (Statology, 2021). Adicionalmente, como se señala en la discusión (Cross Validated, 2014), no es posible deducir la varianza de los boxplot sin asumir muchos supuestos restrictivos, en caso contrario, el investigador será conducido a conclusiones equivocadas.

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

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

Además, puede construirse una tabla que contenga las varianzas de la respuesta por tratamiento y llamar a este objeto v.

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

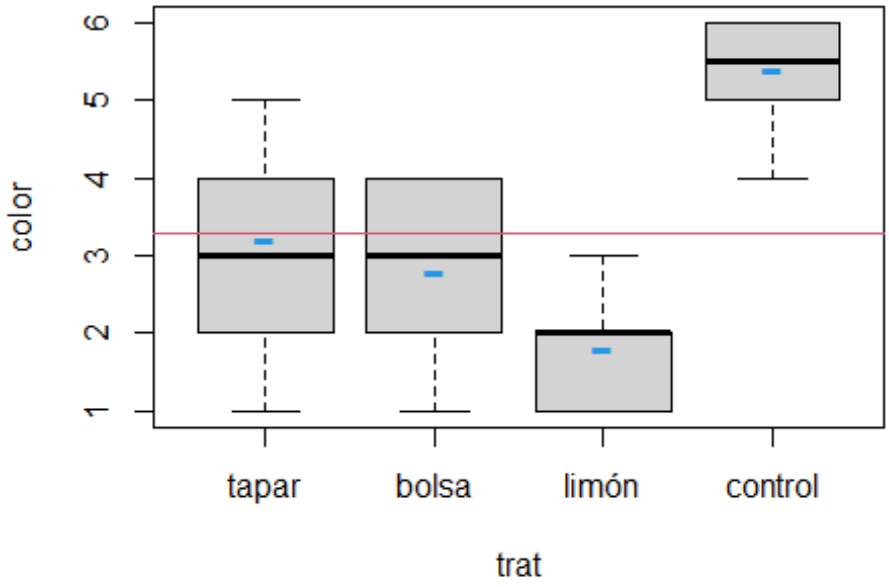
```
##   tapar   bolsa   limón control
##    1.73    1.07    0.62    0.49
```

Es posible también obtener la media general de la respuesta y agregarla al boxplot utilizando la sintaxis abline(h=media, col=2), así como también agregarle las medias de los cuatro tratamientos (el de control es *placebo*) mediante la sintaxis points(1:4,m,col=4,pch="-",cex=2).

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

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

```
abline(h=media, col=2)
points(1:4,m,col=4,pch="-",cex=2)
```



Además del análisis exploratorio anterior, es de especial interés determinar los efectos muestrales de cada tratamiento a partir de la tabla de medias y comparar tales resultados con lo que se observa gráficamente. Recuérdese 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.

```
#EFECTOS
ef=m-media
ef

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

#SUMA DE EFECTOS
sum(ef) #Aunque no da exactamente cero por motivos de cálculo computacional, la suma
```

```
# de los efectos se aproxima a cero empíricamente y debe ser cero teóricamente por su misma construcción
```

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

Los resultados numéricos coinciden con los gráficos, 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.

El significado de cada uno de los valores obtenidos para los efectos muestrales es el siguiente: 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 ambos bajan la media muy poco.

Es recomendable estimar la varianza del error (CME) a partir de la tabla de variancias antes construida. La estimación debe ser la media ponderada de las varianzas 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
```

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

```
## [1] 0.98

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

## [1] 0.98
```

¿Por qué la varianza del error es igual a la varianza de la respuesta en cada tratamiento?

$$e_{ij} = y_{ij} - \mu_j => V(e/x_j) = V(y/x_j) - \mu_j = V(y/x_j)$$

En este caso, μ_j aparece como una constante que no va a modificar las varianzas de los tratamientos y, a causa del supuesto de homogeneidad de varianzas, esta varianza del error sería el mejor estimador.

II.II. Ajuste de un Modelo Lineal

En este caso, para realizar una regresión lineal puede usarse tanto la función 'aov' como la función 'lm'. La principal diferencia entre ambas radica en 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)'. Además, es también de especial interés obtener los resultados del análisis de varianza de una vía mediante 'anova(mod)' o 'anova(mod1)'. Si se usó la función 'aov', es indiferente 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
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = color ~ trat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -0.8    0.2    0.6    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.2000     0.3127  10.234 3.34e-12 ***
## tratbolsa    -0.4000     0.4422  -0.905  0.37173
## tratlimón    -1.4000     0.4422  -3.166  0.00314 **
## tratcontrol   2.2000     0.4422   4.975 1.62e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9888 on 36 degrees of freedom
## Multiple R-squared:  0.6628, Adjusted R-squared:  0.6347
## F-statistic: 23.59 on 3 and 36 DF,  p-value: 1.278e-08
```

```
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
```

En los cuadros informativos generados antes, puede observarse la fila correspondiente a los residuales y obtener de ahí el cuadrado medio residual (CMR) para compararlo con la estimación de la varianza del error obtenida en el punto anterior y verificar que sean equivalentes.

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

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

Como puede observarse, se obtiene el mismo valor que se tenía al promediar las varianzas, es decir, que el cuadrado medio residual es una medida de la variabilidad del color que presentan las manzanas dentro de cada tratamiento.

También se observó antes que existen 36 grados de libertad disponibles para la estimación de los residuales. Esto es así puesto a que se cuenta con 40 datos, 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.

II.III. Descomposición de Suma de Cuadrados Totales

Es posible calcular la suma de cuadrados de *trat*.

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

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

También es posible construir la suma de cuadrados de los efectos obtenidos anteriormente. Al hacerlo, se observará que estos deben multiplicarse por el número de réplicas para obtener exactamente la suma de cuadrados de tratamiento, tal como se estableció antes teóricamente.

```
sum(10*ef^2) #recuérdese la expresión matemática de SCtrat expuesta en la sección A4
```

```
## [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. Dicha cantidad es la misma para todos los valores de un mismo tratamiento, por lo que debe repetirse tantas veces como datos existan para ese tratamiento. De lo anterior se deriva que esa distancia o efecto deba multiplicarse por el número de réplicas en el j-ésimo tratamiento, como se adelantó en la sección B1.

II.IV. Contraste de Hipótesis Sobre las Medias de los Tratamientos

Adicionalmente, es de interés comparar la variabilidad de los promedios con la variabilidad residual para determinar si existe alguna evidencia de diferencias relevantes 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, factor de proporcionalidad extremadamente grande. Esto constituye fuerza de evidencia a favor de que las medias están alejadas unas de otras.

Lo anterior puede investigarse con mayor rigor estableciendo formalmente la hipótesis que se está poniendo a prueba, llevando posteriormente a cabo el contraste de hipótesis y finalmente concluyendo con base en los resultados de este. Así, la hipótesis nula global se establece de la siguiente forma:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$$

o bien,

$$\tau_1 = \tau_2 = \tau_3 = \tau_4$$

Mientras que H_1 establece que al menos una de las medias es diferente.

Puesto que la probabilidad asociada a la probabilidad de cometer error tipo I es casi cero (valor p), se puede esperar que, si se rechaza la hipótesis nula, la probabilidad de estar cometiendo error sea ínfima, por lo que se toma la decisión de rechazar esa hipótesis. Por lo tanto, se puede esperar que no todas las medias

sean iguales (se sospecha que la media del control sea la que se comporta diferente).

Así, pueden obtenerse 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
```

Complementariamente, puede escribirse el modelo que está usando R como.

$$\mu_j = \mu_1 + \delta_j$$

En la ecuación anterior se verifica que $\delta_1 = 0$ (puesto que se toma el tratamiento 1, tapar, como tratamiento de referencia), donde δ_j es el j-ésimo efecto medio que cada tratamiento j-ésimo $\mu_j$tiene con relación al efecto medio de referencia μ_1. A causa de la definición anterior, $\delta_j = \mu_j - \mu_1$. ¿Qué significa el intercepto en estos modelos?

```
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 (*i.e.*, μ_1), el intercepto de cada modelo coincide con la media de ese tratamiento, que es 3.2. Así, los otros coeficientes de regresión obtenidos representan la distancia que hay de la media de cada uno de los otros tratamientos con respecto a la media del tratamiento tapar (tratamiento de referencia).

Adicionalmente, puede estimarse la matriz de diseño[3].

```
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
```

[3] Los diferentes valores que se asignan a cada factor estudiado en un diseño experimental se llaman *niveles*. Una combinación de niveles de todos los factores estudiados se llama *tratamiento* o *punto de diseño*. Un aspecto fundamental del diseño de experimentos es decidir cuáles pruebas o tratamientos se van a realizar y cuántas repeticiones de cada uno se requieren, de manera que se obtenga la máxima información al mínimo costo posible. El arreglo formado por los diferentes tratamientos que serán corridos, incluyendo las repeticiones, recibe el nombre de *matriz de diseño* o simplemente *diseño* (Gutiérrez Pulido & de la Vara Salazar, 2012, pág. 6).

```
## 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 diseño se tienen variables auxiliares con 0 y 1 solamente.

A partir de los coeficientes obtenidos, es posible estimar los efectos muestrales y compararlos con los obtenidos mediante 'mod$coef'.

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

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

Como se observa, mediante 'mod$coef' no es inmediata la obtención de los efectos. Por la forma en que están definidos los coeficientes de regresión, debe restarse la media general y sumarse el intercepto a cada coeficiente para obtener el efecto respectivo.

Pueden obtenerse de forma directa los efectos de los predictores mediante la sintaxis 'model.tables(mod)' (esto 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
```

II.V. Parametrizaciones Posibles del Modelo de Regresión

1. Tratamiento referencia. Se asume que el coeficiente de uno de los tratamientos es cero. Esta forma es la que R usa por default.

2. 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 se obtiene por diferencia:

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

Puede estudiarse el modelo **suma nula** usando la siguiente instrucción: 'options(contrasts=c("contr.sum","contr.poly"))'. Adicionalmente, para volver al modelo tratamiento referencia se puede utilizar la sintaxis 'options(contrasts=c("contr.treatment","contr.poly"))'.

La codificación antes descrita puede verificarse mediante la sintaxis 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

#Esto significa que para "tapar" se usa el coeficiente 1, para "bolsa" el coeficiente 3, para "limón" el coeficiente 3 y para control se restan todos los coeficientes, con la finalidad de poder estimar el efecto
```

II.VI. Estimaciones mediante el Modelo de Regresión

Los pasos antes realizados pueden repetirse con un nuevo modelo, con la finalidad de realizar por separado las estimaciones de los cálculos antes hechos y evitar con ello confusiones.

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

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

El modelo se escribe

$$\mu_j = \mu + \tau_j$$

y la restricción en los parámetros cambia, ya que ahora se debe establecer que:

$$\tau_4 = -(\tau_1 + \tau_2 + \tau_3)$$

Como se observa, ahora el intercepto representa la media global de la variable de respuesta, que es 3.3. Complementariamente, 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.

II.VII. Matriz de Diseño para el Modelo de Suma Nula

Ahora las variables auxiliares son denotadas por un -1 en el tratamiento de referencia, que en este caso es el cuarto (el tratamiento de observación o placebo, en el que no se interviene).

```
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"
```

Merece la pena mencionar que los coeficientes son directamente los efectos, salvo el cuarto coeficiente de regresión 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
```

Finalmente, también es posible estimar los efectos directos que cada tipo de tratamiento tiene sobre las manzanas.

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

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

III. REFERENCIAS

Cochran, W. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society, 30*(2), 178-191.

Cross Validated. (2014, Marzo 27). *Deduce variance from boxplot*. Retrieved from Questions: https://stats.stackexchange.com/questions/91536/deduce-variance-from-boxplot

Gutiérrez Pulido, H., & de la Vara Salazar, R. (2012). *Análisis y diseño de experimentos* (Tercera ed.). México, D.F.: McGraw Hill.

Helmenstine, A. (2020, Septiembre 15). *What Is a Confounding Variable? Definition and Examples*. Retrieved from Science Notes: https://sciencenotes.org/what-is-a-confounding-variable-definition-and-examples/

Science Direct. (2022, Julio 21). *Confounding Variable*. Retrieved from Topics: https://www.sciencedirect.com/topics/nursing-and-health-professions/confounding-variable

Statology. (2021, Octubre 4). *When Should You Use a Box Plot? (3 Scenarios)*. Retrieved from statology.org: https://www.statology.org/when-to-use-box-plot/