PRINCIPIOS DE DISEÑO EXPERIMENTAL: COMPARACIONES MÚLTIPLES

ISADORE NABI

I. ALGUNOS ASPECTOS TEÓRICOS GENERALES 2

I.I. Análisis de Varianza (ANOVA) **2**

I.I. I. Generalidades 2

I.I. II. Descomposición de la Varianza como Suma de Cuadrados 5

I.II. Modelo General ANOVA **7**

I.III. Contrastes (Diseño de Experimentos por Bloques) **9**

I.IV. Varianza de un Contraste **12**

I.V. Comparaciones Múltiples **13**

I.IV. I. Intervalos LSD (Least Significance Method) de Fisher 15

I.IV. II. Ajuste de Bonferroni 15

I.IV. III. Ajuste de Holm–Bonferroni 16

I.IV. IV. Ajuste por Diferencia Honestamente Significativa (HSD) de Tukey o Método de Tukey-Kramer 16

I.IV. V. Método de Scheffé 18

I.IV. VI. Corrección de Dunnett 19

I.IV. VI. Tasa de Falso Descubrimiento (FDR) de Benjamini & Hochberg (BH) 19

II. PRIMER CASO DE APLICACIÓN: DETERMINACIÓN DE LA LOCALIDAD DE PRODUCCIÓN DE LAS UVAS MÁS DULCES *23*

II.I. Desarrollo en RStudio **23**

II.II. Conclusiones **32**

III. SEGUNDO CASO DE APLICACIÓN: EVALUACIÓN DE MÉTODOS DE PREVENCIÓN DE LA OXIDACIÓN DE LAS MANZANAS *33*

III.I. Desarrollo en RStudio **33**

III.II. Conclusiones **46**

IV. TERCER CASO DE APLICACIÓN: ESTIMACIÓN NO-PARAMÉTRICA DE LA CANTIDAD PROMEDIO DE MOSCAS SEGÚN TIPO DE VEGETACIÓN *46*

IV.I. Desarrollo en RStudio **46**

IV.II. Conclusiones **54**

V. REFERENCIAS *54*

I. ALGUNOS ASPECTOS TEÓRICOS GENERALES

I.I. Análisis de Varianza (ANOVA)

I.I. I. Generalidades

Como señala (Amat Rodrigo, 2016), la técnica de análisis de varianza (ANOVA) también conocida como análisis factorial y desarrollada por Fisher en 1930, constituye la herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua[1]. Es, por tanto, la prueba estadística a emplear cuando se desea comparar las medias de dos o más grupos. Esta técnica puede generalizarse también para estudiar los posibles efectos de los factores sobre la varianza de una variable.

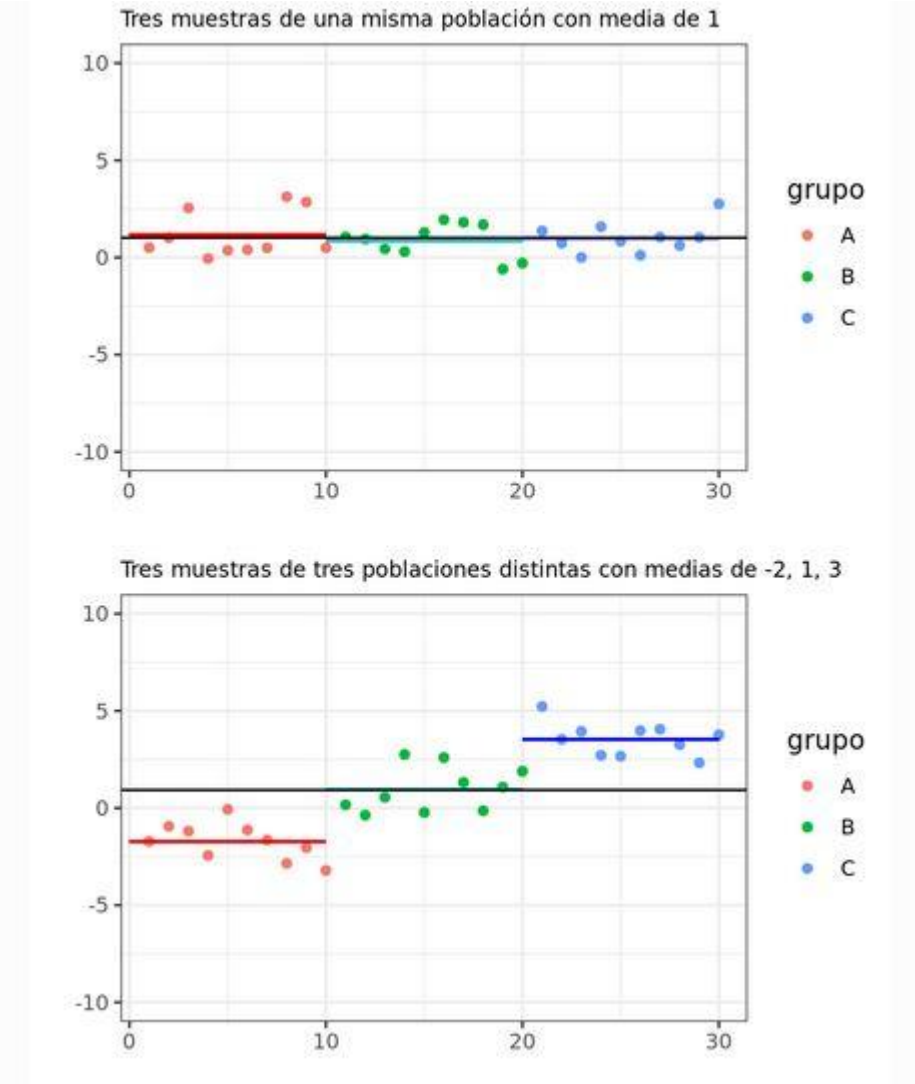
La hipótesis nula de la que parten los diferentes tipos de ANOVA es que la media de la variable estudiada es la misma en los diferentes grupos, en contraposición a la hipótesis alternativa de que al menos dos medias difieren de forma significativa. ANOVA permite comparar múltiples medias, pero lo hace mediante el estudio de las varianzas.

El funcionamiento básico de un ANOVA consiste en calcular la media de cada uno de los grupos para a continuación comparar la varianza de estas medias (varianza explicada por la variable grupo, intervarianza) frente a la varianza promedio dentro de los grupos (la no explicada por la variable grupo, intravarianza). Bajo la hipótesis nula de que las observaciones de los distintos grupos proceden en su totalidad de la misma población (*i.e.*, tienen la misma media y varianza), la varianza ponderada entre grupos será la misma que la varianza promedio dentro de los grupos.

[1] Que es la variable dependiente o variable que se busca explicar en el estudio/investigación.

Conforme las medias de los grupos estén más alejadas las unas de las otras, la varianza entre medias se incrementará y dejará de ser igual a la varianza promedio dentro de los grupos.

```
knitr::include_graphics("AMAT1.JPG")
```



#*Figura 1. Comparación de Medias Entre Grupos*

#*Fuente:* (Amat Rodrigo, 2016)

La línea negra es la media para todas las observaciones o media general.

El estadístico estudiado en el ANOVA, conocido como F_{ratio}, es la razón entre la varianza de las medias de los grupos y el promedio de la varianza dentro de los

grupos. Este estadístico sigue una distribución conocida como "F de Fisher-Snedecor". Si se cumple la hipótesis nula, el estadístico F adquiere el valor de 1 ya que la intervarianza será igual a la intravarianza. Cuanto más difieran las medias de los grupos mayor será la varianza entre medias en comparación al promedio de la varianza dentro de los grupos, obteniéndose valores de F superiores a 1 y por lo tanto menor la probabilidad de que la distribución adquiera valores tan extremos (menor el p-value).

En concreto, si S_1^2 es la la varianza de una muestra de tamaño N_1 extraída de una población normal de varianza σ_1^2 y S_2^2 es la la varianza de una muestra de tamaño N_2 extraída de una población normal de varianza σ_2^2, y ambas muestras son independientes, el cociente $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ se distribuye como una variable F de Snedecor con (N_1 y N_2) grados de libertad. En el caso del ANOVA, dado que dos de las condiciones son la normalidad de los grupos y la homocedasticidad de varianza ($\sigma_1^2 = \sigma_2^2$), el valor F se puede obtener dividiendo las dos varianzas calculadas a partir de las muestras (intervarianza y intravarianza).

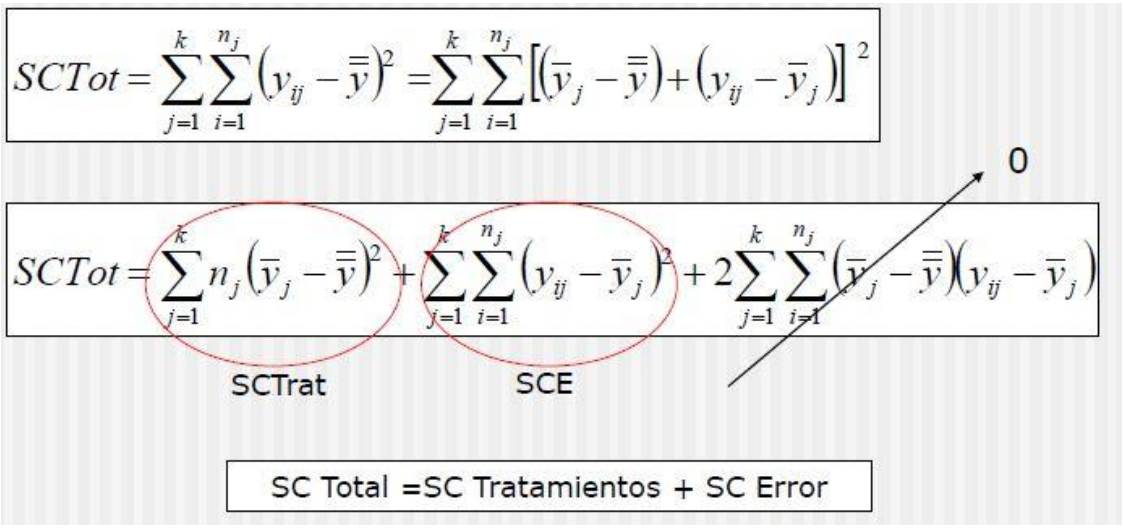
Existen diferentes tipos de ANOVA dependiendo de si se trata de datos independientes (ANOVA entre sujetos), si son pareados (ANOVA de mediciones repetidas), si se comparan la variable cuantitativa dependiente contra los niveles de una única variable explicativa o factor (ANOVA de una vía) o frente a dos factores (ANOVA de dos vías). Este último puede ser a su vez aditivo o de interacción (*i.e.*, los factores son independientes o no lo son, respectivamente). Cada uno de estos tipos de ANOVA tiene una serie de requerimientos.

El ANOVA de una vía, ANOVA con un factor o modelo factorial de un solo factor es el tipo de análisis que se emplea cuando los datos no están pareados y se quiere estudiar si existen diferencias significativas entre las medias de una variable aleatoria continua en los diferentes niveles de otra variable cualitativa o factor. Es una extensión de las pruebas t independientes para más de dos grupos.

I.I. II. Descomposición de la Varianza como Suma de Cuadrados

La diferencia entre medias se detecta a través del estudio de la varianza entre grupos y dentro de grupos. Para lograrlo, el ANOVA requiere de una descomposición de la varianza basada en la siguiente idea:

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




```
#Figura 2: Descomposición de la Suma de Cuadrados Totales del Tratamiento
```

Donde *SC Total* es la variabilidad total, *SC Tratamientos* es la variabilidad debida a los diferentes niveles del factor (*i.e.*, variabilidad explicada por el factor o varianza entre niveles), mientras que *SC Error* es la variabilidad residual (*i.e.*, variabilidad no-explicada por el factor o varianza dentro de los niveles). Precisamente para calcular estas fuentes de variabilidad es que se recurre a la suma de cuadrados y su descomposición, tal como se muestra a continuación:

1. *Suma de Cuadrados Total o Total Sum of Squares (SC Total o TSS)*. Mide la variabilidad total de los datos y se define como la suma de los cuadrados de las diferencias de cada observación respecto a la media general de todas las observaciones. Los grados de libertad de la suma de cuadrados totales es igual al número total de observaciones menos uno ($N - 1$).

2. *Suma de Cuadrados del Factor o Sum of Squares due to Treatment (SC Tratamientos o SST)*. Mide la variabilidad en los datos asociada al efecto del factor sobre la media (la diferencia de las medias entre los diferentes niveles o grupos). Se obtiene como la suma de los cuadrados de las desviaciones de la media de cada proveedor respecto de la media general, ponderando cada diferencia al cuadrado por el número de observaciones de cada grupo. Los grados de libertad correspondientes son igual al número niveles del factor menos uno (k-1).

3. *Suma de Cuadrados Residual/Error o Sum of Squares of Errors (SC Error o SSE)*. Mide la variabilidad dentro de cada nivel, es decir, la variabilidad que no es debida a variable cualitativa o factor. Se calcula como la suma de los cuadrados de las desviaciones de cada observación respecto a la media del nivel al que pertenece. Los grados de libertad asignados a la suma de cuadrados residual equivale la diferencia entre los grados de libertad totales y los grados de libertad del factor, o lo que es lo mismo $(N - k)$. En estadística se emplea el termino error o residual ya que se considera que esta es la variabilidad que muestran los datos debido a los errores de medida. Desde el punto de vista biológico tiene más sentido llamarlo Suma de cuadrados dentro de grupos ya que se sabe que la variabilidad observada no solo se debe a errores de medida, si no a los muchos factores que no se controlan y que afectan a los procesos biológicos.

$$TSS = SST + SSE$$

Una vez descompuesta la suma de cuadrados se puede obtener la descomposición de la varianza dividiendo la Suma de Cuadrados entre los respectivos grados de libertad. De forma estricta, al cociente entre la suma de cuadrados y sus correspondientes grados de libertad se le denomina *cuadrados medios* o *mean sum of squares* y pueden ser empleado como estimador de la varianza; ANOVA se define

como análisis de varianza, pero en un sentido puramente matemático, se trata de un análisis de la Suma de Cuadrados Medios.

1. $S_T^2 = TSS/(N-1)$ = Cuadrados Medios Totales = Cuasivarianza Total (varianza muestral total).

2. $S_t^2 = SST/(k-1)$ = Cuadrados Medios del Factor = Intervarianza (varianza entre las medias de los distintos niveles).

3. $S_E^2 = SSE/(N-k)$=Cuadrados Medios del Error = Intravarianza (varianza dentro de los niveles, conocida como varianza residual o de error).

Una vez descompuesta la estimación de la varianza, se obtiene el estadístico F_{ratio} dividiendo la intervarianza entre la intravarianza, es decir:

F_ratio = (Cuadrados Medios del Factor) / (Cuadrados Medios del Error) = S_t^2/S_E^2 = intervarianza/intravarianza $\sim F_{k-1,N-k}$.

Dado que por definición el estadístico F_{ratio} sigue una distribución F Fisher-Snedecor con $k-1$ y $N-k$ grados de libertad, se puede conocer la probabilidad de obtener valores iguales o más extremos que los observados.

I.II. Modelo General ANOVA

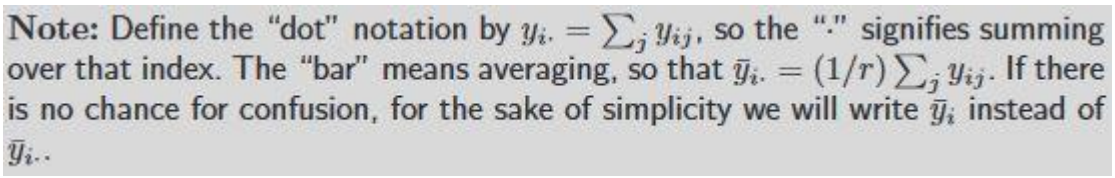
$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},$$

donde $i = 1, \ldots, t$ y $j = 1, \ldots, r$.

Como señala (Casella, Statistical Design, 2008, pág. 2), en la expresión anterior, Y_{ij} es la variable de respuesta, μ es la media general del modelo, τ_i es el efecto del tratamiento i-ésimo y ϵ_{ij} es el término de error, a menudo asumido como $N(0, \sigma^2)$, independiente e idénticamente distribuido (iid); la obra citada no debe confundirse con su presentación de clases en la Universidad de Florida citada más abajo.

El modelo antes presentado, tiene la característica de estar sobreparametrizado[2] o, lo que es lo mismo, el modelo es no-identificable, es decir, es un modelo para el cual la función sobre la cual se aplican los valores de los parámetros (estimados a partir del conjunto de datos) para obtener una predicción sobre la respuesta no es de carácter inyectivo (que establece una relación uno-a-uno entre los elementos de dos conjuntos) (Cross Validated, 2012); el modelo antes expuesto está sobreparametrizado por cuanto para cada valor i-ésimo se deben estimar no únicamente un parámetro τ_i sino que $(\mu + \tau_i)$ (Oxford Reference, 2022). Por ello, es común imponer al modelo la restricción $\sum_i \tau_i = 0$, lo que permite que el modelo sea identificable en cuanto garantiza la ortogonalidad entre los efectos (*i.e.*, su independencia lineal).

```
knitr::include_graphics("F01.JPG")
```

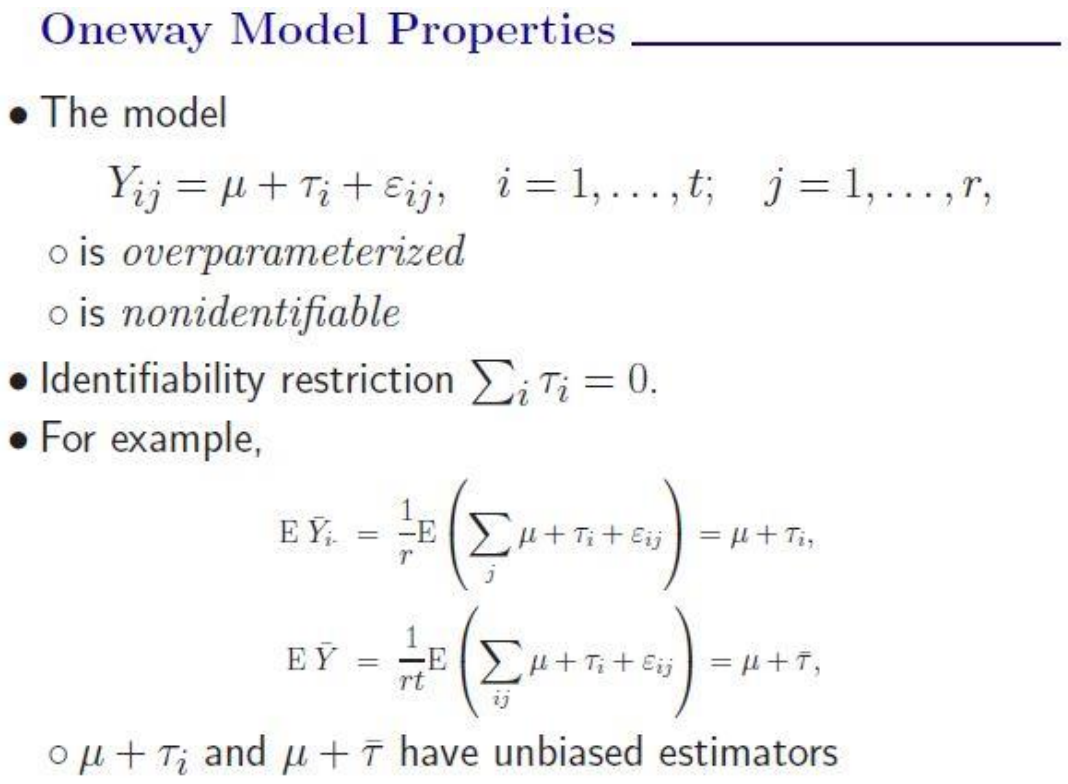


#*Figura 2: Aclaraciones Sobre la Notación*

#*Fuente:* (Casella, Statistical Design. Principles, Recommendations, and Opinions, 2022, pág. 3).

```
knitr::include_graphics("F0.JPG")
```

[2] Este término se utiliza más en Psicometría, Bioestadística y Aprendizaje Estadístico, en Econometría se utiliza sobreespecificado.



#*Figura 3: Resumen de las Propiedades del Modelo de Una Vía*
#*Fuente:* (Casella, Statistical Design. Principles, Recommendations, and Opinions, 2022, pág. 6).

En las identidades anteriores, $E(\bar{Y}_{i\cdot} - \bar{Y}_{i'\cdot}) = \tau_i - \tau_{i'}$, por lo que las diferencias entre tratamientos son siempre estimables. Además, es posible asumir siempre que $\bar{\tau} = 0$ sin perder generalidad (Casella, Statistical Design, 2008, pág. 3), por lo que μ y τ_i son también estimables. Finalmente, sin perder generalidad (loc. cit), puede asumirse que $\mu^* + \bar{\tau_i}$ y $\tau_i^* = \tau_i - \bar{\tau}$ y usar el modelo $Y_{ij} = \mu^* + \tau_i^* + \epsilon_{ij}$ con $\sum_i \tau_i^* = 0$.

I.III. Contrastes (Diseño de Experimentos por Bloques)

Como señala (Casella, Statistical Design, 2008, pág. 11), normalmente el objetivo de un experimento estadístico es comprender el efecto de un tratamiento o comparar y contrastar los efectos del tratamiento. Con un buen diseño, el experimentador se esfuerza por optimizar la varianza que utiliza para hacer comparaciones, pero también requiere entender cómo hacer estas comparaciones de interés. Por ello, se

utilizan los contrastes y, de ser posible, aquellos con la característica de ortogonalidad para comparar los efectos del tratamiento.

Formalmente, un contraste es una combinación lineal L de los promedios de los diferentes tratamientos (los parámetros a estimar del conjunto de datos) y sus respectivos pesos, bajo la condición de que la suma de dichos pesos sea 0. Idealmente, los contrastes deben ser ortogonales, *i.e.*, generados mediante el producto escalar de los diferentes pares de combinaciones lineales entre parámetros μ_i[3] y los pesos c_i igual a cero. La ortogonalidad de los contrastes permite que su comparación sea estadísticamente relevante en cuanto se están comparando dos contrastes verdaderamente diferentes (donde su diferencia pasa porque no existan efectos "duplicados" en consideración, es decir, efectos que se expresen como combinación lineal de otro efecto y su respectivo peso).

```
knitr::include_graphics("F4.JPG")
```

[3] Haciendo un cambio de notación, μ_i equivale a lo que Casella representa como θ_i, los cuales no deben confundirse con μ, que en la notación de Casella representa la media global (esta media global, en la notación aquí utilizada, se expresa como $\bar{\bar{y}}$).

Orthogonal and Uncorrelated

A Oneway Model

$$Y_{ij} = \theta_i + \varepsilon_{ij}, \quad i = 1, \ldots, t; \quad j = 1, \ldots, r_i,$$

$\sum_{i=1}^{t} a_i\theta_i$	$\sum_i a_i = 0$	Contrast
$\sum_{i=1}^{t} a_i\theta_i$ and $\sum_{i=1}^{t} b_i\theta_i$	$\sum_{i=1}^{t} a_i b_i = 0$	Orthogonal Contrasts
$\sum_{i=1}^{t} a_i\bar{y}_i$ and $\sum_{i=1}^{t} b_i\bar{y}_i$	$\sum_{i=1}^{t} a_i b_i = 0$	Orthogonal Contrasts
$\sum_{i=1}^{t} a_i\bar{y}_i$ and $\sum_{i=1}^{t} b_i\bar{y}_i$	$\sum_{i=1}^{t} a_i b_i/r_i = 0$	Uncorrelated Contrasts

#*Figura 4: Contraste, Contrastes Ortogonales, Contrastes No-Correlacionados*

#*Fuente:* (Casella, Statistical Design. Principles, Recommendations, and Opinions, 2022, pág. 5)

knitr::include_graphics("F1.JPG")

$$L = \sum_{i=1}^{k} c_i\mu_i \quad \text{donde} \quad \sum_{i=1}^{k} c_i = 0$$

#*Figura 5: Definición Formal de Contraste (Diseño de Experimentos por Bloques)*

#*Fuente:* (Fallas, 2012)*Fallas, Jorge. ANÁLISIS DE VARIANZA. 2012. p. 26.*

De lo anterior surge la interrogante, ¿en qué condiciones un estadístico de prueba[4] o coeficiente se convierte en una magnitud de efecto?

[4] Que es la denominación adecuada en lugar de llamarlos "parámetros", puesto que estos últimos son técnicamente los verdaderos valores poblacionales.

Según (Kelley & Preacher, 2012, pág. 140), para comprender tal diferencia debe comprenderse primero qué es el tamaño del efecto. Defínase el tamaño del efecto como el reflejo cuantitativo de la magnitud de algún fenómeno que es usada (tal magnitud) para abordar una pregunta de interés. Como ellos señalan, la definición es más que una combinación de "efecto" y "tamaño" porque depende explícitamente de la pregunta de investigación que se aborde. Lo anterior significa que lo que separa a un tamaño de efecto de un estadístico de prueba (o estimador) es la orientación de su uso, si responde una pregunta de investigación en específico entonces el estadístico de prueba se convierte en un "tamaño de efecto" y si sólo es parte de un proceso global de predicción entonces es un estadístico de prueba a secas, *i.e.*, su distinción/identificación de cuándo un estadístico (o parámetro) se convierte en un tamaño de efecto, es una cuestión, en última instancia, gnoseológica, no matemática.

Lo anterior simplemente significa que, dependiendo del tipo de pregunta que se desee responder el investigador, un estadístico (o parámetro) será un tamaño de efecto o simplemente un estadístico (o parámetro) sin más.

I.IV. Varianza de un Contraste

Un contraste se puede estimar fácilmente con la misma combinación lineal sobre los promedios muestrales, en donde nuevamente se lleva a cabo un cambio de notación, específicamente μ_i por $\bar{y}_i$, para denotar que $\bar{y}_i$ es una estimación del verdadero valor de dicha media, mientras que su verdadero valor (el poblacional) es μ_i.

```
knitr::include_graphics("F2.JPG")
```

$$\hat{L} = \sum_{i=1}^{k} c_i \bar{y}_i$$

#*Figura 6: Estimación de un Contraste*

Para construir intervalos de confianza y hacer pruebas de hipótesis sobre los contrastes es preciso conocer su varianza y error estándar.

```
knitr::include_graphics("F3.JPG")
```

$$Var(\hat{L}) = \sum_{i=1}^{k} c_i^2 Var[\bar{y}_i] = \sum_{i=1}^{k} c_i^2 \frac{\sigma^2}{n_i} = CME \times \sum_{i=1}^{k} \frac{c_i^2}{n_i}$$

#*Figura 7: Varianza de la Estimación de un Contraste*

En la expresión anterior, n_i representa el tamaño del bloque i-ésimo.

I.V. Comparaciones Múltiples

Como se señala en (Scientific European Federation of Osteophats, 2022, pág. 1), una vez que se ha determinado que existen diferencias entre las medias, las pruebas de rango post hoc permiten determinar qué medias difieren. La prueba de rango post hoc identifica subconjuntos homogéneos de medias que no se diferencian entre sí.

Retomando lo señalado en (Glen, 2022), una prueba ANOVA puede decirle al investigador si las diferencias entre medias son significativas (o no) en general, pero no le dirá exactamente dónde (de entre las medias comparadas) se encuentran esas

diferencias. Una vez que se haya ejecutado un ANOVA y se hayan encontrado resultados significativos, es posible realizar alguna prueba de comparaciones múltiples para averiguar qué medias de grupos específicos (comparados entre sí) son diferentes como las que se presentarán en breve.

Para determinar cuáles contrastes son diferentes de cero se debe evitar hacer pruebas simples de forma separada ya que se estaría aumentando el error tipo I. Cuando se tiene un modelo de efectos fijos, hay métodos estadísticos para hacer todas las comparaciones y a la vez mantener el error tipo I total en un nivel deseado α (usualmente 0.05). Si la hipótesis de igualdad de medias no se rechaza, este paso no tiene sentido (puesto que implicaría que todas son estadísticamente iguales y no tiene sentido compararlas).

Como señala (Amat Rodrigo, 2016), si un análisis de varianza resulta significativo, implica que al menos dos de las medias comparadas son significativamente distintas entre sí, pero no se indica cuáles. Para identificarlas hay que comparar dos a dos las medias de todos los grupos introducidos en el análisis mediante una prueba t u otra prueba que compare dos grupos, a esto se le conoce como análisis post-hoc.

Debido a la inflación del error de tipo I, cuantas más comparaciones se hagan más aumenta la probabilidad de encontrar diferencias significativas (para $\alpha = 0.05$, de cada 100 comparaciones se esperan 5 significativas solo por azar). Los niveles de significancia pueden ser ajustados en función del número de comparaciones (a esto se le conoce como *corrección de significancia*). Si no se hace ningún tipo de corrección se aumenta la posibilidad de falsos positivos (error tipo I), pero si se es muy estricto con las correcciones se pueden considerar como no-significativas diferencias que realmente podrían serlo (error tipo II). La necesidad de corrección o no, y de qué tipo, se ha de estudiar con detenimiento en cada caso. Los principales métodos de comparación post-hoc (algunas con corrección y otros no) se presentan a continuación.

I.IV. I. Intervalos LSD (Least Significance Method) de Fisher

Los intervalos LSD son básicamente un conjunto de pruebas t individuales con la única diferencia de que en lugar de calcular una desviación estándar agrupada[5] (DSA) empleando solo los dos grupos comparados, calcula la DSA a partir de todos los grupos. Cuanto más se alejen los intervalos de dos grupos más diferentes son sus medias, siendo significativa dicha diferencia si los intervalos no se solapan. Es importante comprender que los intervalos LSD se emplean para comparar las medias, pero no se pueden interpretar como el intervalo de confianza para cada una de las medias. El método LSD no conlleva ningún tipo de corrección de significancia, es por esto que su uso parece estar desaconsejado para determinar significancia, aunque sí para identificar que grupos tienen las medias más distantes. En R se pueden obtener los intervalos LSD y su representación gráfica mediante la función 'LSD.test()'.

I.IV. II. Ajuste de Bonferroni

Se pueden construir intervalos de confianza simultáneamente sobre varios contrastes que se hayan definido antes de llevar a cabo el análisis. El método de Bonferroni usa la distribución t con los grados de libertad del error y ajustando el nivel de confianza según el número de contrastes (r) de tal forma que los intervalos se hacen más amplios que si se hicieran con un nivel $(1-\alpha)$; la confianza global será $(1-\alpha)$. Este es posiblemente el ajuste de significancia más extendido a pesar de que no está recomendado para la mayoría de las situaciones que se dan en el ámbito de la biomedicina. Con esta corrección se asegura que la probabilidad de obtener al

[5] La desviación estándar agrupada es un método para estimar una única desviación estándar para representar todas las muestras o grupos independientes en su estudio cuando se supone que provienen de poblaciones con una desviación estándar común. La desviación estándar agrupada es la dispersión promedio de todos los puntos de datos sobre la media de su grupo (no la media general). Es un promedio ponderado de la desviación estándar de cada grupo. La ponderación otorga a los grupos más grandes un efecto proporcionalmente mayor en la estimación general. Las desviaciones estándar agrupadas se utilizan en pruebas t de 2 muestras, ANOVA, gráficos de control y análisis de capacidad. (Minitab, 2022).

menos un falso positivo entre todas las comparaciones (denominado también "family-wise error rate") es $\leq \alpha$. Permite por lo tanto contrastar una hipótesis nula general (la de que todas las hipótesis nulas testadas son verdaderas) de forma simultánea, cosa que raramente es de interés en las investigaciones. Se considera un método excesivamente conservativo sobre todo a medida que se incrementa el número de comparaciones. Se desaconseja su utilización excepto en situaciones muy concretas. Alternativamente a este ajuste, se recomienda usar la tasa de control de falsos descubrimientos, al menos en lo que respecta a los estudios sobre salud[6]. En R se puede realizar este ajuste mediante la función 'pairwise.t.test()' indicando en los argumentos 'p.adj = "bonferroni"'.

I.IV. III. Ajuste de Holm–Bonferroni

Con este método, el valor de significancia α se corrige secuencialmente haciéndolo menos conservativo que el de Bonferroni. Aun así, parece que tampoco es indicado si se realizan más de seis comparaciones. El proceso consiste en realizar un t-test para todas las comparaciones y ordenarlas de menor a mayor p-value. El nivel de significancia para la primera comparación (la que tiene menor p-value) se corrige dividiendo α entre el número total de comparaciones, si no resulta significativo se detiene el proceso, si sí que lo es, se corrige el nivel de significancia de la siguiente comparación (segundo menor p-value) dividiendo entre el número de comparaciones menos uno. El proceso se repite hasta detenerse cuando la comparación ya no sea significativa.

I.IV. IV. Ajuste por Diferencia Honestamente Significativa (HSD) de Tukey o Método de Tukey-Kramer

Es el ajuste recomendado cuando el número de grupos a comparar es mayor de seis y el diseño es equilibrado (mismo número de observaciones por grupo). En el caso de modelos no equilibrados el método HSD es conservativo (en este contexto,

[6] Véase https://pubmed.ncbi.nlm.nih.gov/24831050/.

conservador implica que construye intervalos de confianza más amplios), requiere diferencias grandes para que resulte significativo. Solo aplicable si se trata de datos no pareados. El Tukey's test es muy similar a un t-test, excepto que corrige el experiment wise error rate. Esto lo consigue empleando un estadístico que sigue una distribución llamada studentized range distribution en lugar de una distribución t. El estadístico se define como $q_{calculado} = \frac{\bar{x}_{max} - \bar{x}_{min}}{S\sqrt{2/n}}$, en donde $\bar{x}_{max}$ es la mayor media entre las dos en comparación, $\bar{x}_{min}$ es la menor entre las dos medias en comparación y $S\sqrt{2/n}$ es el error estándar de la suma de medias, lo cual también se puede expresar como $q_{ij} = \frac{\bar{y}_i - \bar{y}_j}{ee_{ij}}$. Para cada par de grupos, se obtiene el valor q_{ij} y se compara con el esperado acorde a una distribución de rango studentizado con los correspondientes grados de libertad. Si la probabilidad es menor al nivel de significancia α preestablecido, se considera significativa la diferencia de medias. Al igual que con los intervalos LSD, es posible calcular intervalos HSD para estudiar su solapamiento. En R, las funciones 'TukeyHSD()' y 'plot(TukeyHSD)' permiten calcular los valores p corregidos por Tukey y representar los intervalos. Cuando sólo se quieren hacer contrastes para comparar pares de medias, el método de Tukey produce intervalos más angostos que los de Bonferroni o Scheffé. Al utilizar el CME en la fórmula del error estándar se está utilizando la información de todas las varianzas. El CME es una ponderación de todas las varianzas.

El uso del CME presupone que hay igualdad de varianzas. Si éste no fuera el caso es mejor usar las varianzas estimadas de cada grupo por separado. Se conoce como Tukey-Kramer cuando las muestras no tienen el mismo número de datos y la hipótesis nula es rechazada cuando los intervalos de confianza no contienen al elemento nulo 0 (Scientific European Federation of Osteophats, 2022, pág. 1). La prueba compara todos los posibles pares de medias (Glen, 2022). Como se señala en la última fuente citada, los supuestos para la validez de las inferencias desprendidas de esta prueba son:

1. Las observaciones son independientes dentro de los grupos y entre ellos.
2. Los grupos de cada media en la prueba se distribuyen normalmente.
3. Existe la misma varianza dentro del grupo entre los grupos asociados con cada media en la prueba (homogeneidad de varianza).

I.IV. V. Método de Scheffé

Como se señala en (Wikipedia, 2021), es un método para ajustar los niveles de significancia en un análisis de regresión lineal para tener en cuenta las comparaciones múltiples. Es particularmente útil en el análisis de varianza (un caso especial de análisis de regresión) y en la construcción de bandas de confianza simultáneas para regresiones que involucran funciones básicas (un espacio funcional puede verse como un espacio vectorial de dimensión infinita cuyos vectores de base son funciones, no vectores. Esto significa que cada función en el espacio funcional puede representarse como una combinación lineal de las funciones de base, al igual que en el álgebra lineal con los vectores). El método de Scheffé es un procedimiento de comparación múltiple de un solo paso que se aplica al conjunto de estimaciones de todos los posibles contrastes entre las medias de nivel de factor, no solo las diferencias por pares consideradas por el método de Tukey-Kramer. Funciona con principios similares al procedimiento Working-Hotelling para estimar respuestas medias en regresión, que se aplica al conjunto de todos los niveles de factores posibles. Si solo se realiza un número fijo de comparaciones por pares, el método Tukey-Kramer dará como resultado un intervalo de confianza más preciso; Scheffé propone un método más conservador, es decir, que tiende a construir intervalos más amplios. La ventaja de este método es que permite construir los contrastes a posteriori. Este método no depende del número de contrastes que se prueben ya que el error de estimación se construye con una distribución F que depende solamente de los grados de libertad en el ANDEVA realizado y del nivel de significancia escogido. La confianza global seguirá siendo $(1-\alpha)$. En el caso

general de que muchos o todos los contrastes puedan ser de interés, el método Scheffé es más apropiado y proporcionará intervalos de confianza más estrechos en el caso de un gran número de comparaciones. En R, esto puede calcularse mediante la sintaxis 'scheffeCI()'. La interpretación de estos intervalos es similar a la de los de Bonferroni. Sólo hay que observar que todos son más anchos debido al factor que multiplica el error estándar para obtener el error de estimación.

I.IV. VI. Corrección de Dunnett

Es el equivalente a la prueba Tukey-Kramer (HSD) recomendado cuando en lugar de comparar todos los grupos entre sí ($[(k-1)k]/2$ comparaciones) solo se quieren comparar frente a un grupo control (k−1 comparaciones). Se emplea con frecuencia en experimentos médicos. Su sintaxis en RStudio es 'dunnettCI'.

I.IV. VI. Tasa de Falso Descubrimiento (FDR) de Benjamini & Hochberg (BH)

Al realizar múltiples comparaciones es importante controlar la inflación del error de tipo I. Sin embargo, correcciones como las de Bonferroni o similares conllevan una serie de problemas. La primera es que el método se desarrolló para contrastar la hipótesis nula universal de que los dos grupos son iguales para todas las variables testadas, no para aplicarlo de forma individual a cada prueba. A modo de ejemplo, supóngase que un investigador quiere determinar si un nuevo método de enseñanza es efectivo empleando para ello estudiantes de 20 colegios. En cada colegio se selecciona de forma aleatoria un grupo control, un grupo que se somete al nuevo método y se realiza una prueba estadística entre ambos considerando un nivel de significancia $\alpha=0.05$.

La corrección de Bonferroni implica comparar el valor p obtenido en las 20 pruebas estadísticas frente a $0.05/20 = 0.0025$. Si alguno de los valores p es significativo, la conclusión de Bonferroni es que la hipótesis nula de que el nuevo sistema de enseñanza no es efectivo en todos los grupos (colegios) queda rechazada, por lo que se puede afirmar que el método de enseñanza es efectivo para alguno de los 20

grupos, pero no cuáles ni cuántos. Este tipo de información no es de interés en la gran mayoría de estudios, ya que lo que se desea conocer es qué grupos difieren.

El segundo problema de la corrección de Bonferroni es que una misma comparación será interpretada de forma distinta dependiendo del número de pruebas que se hagan. Supóngase que un investigador realiza 20 contrastes de hipótesis y que todos ellos resultan en un valor p de 0.001. Aplicando la corrección de Bonferroni si el límite de significancia para una prueba individual es de $\alpha = 0.05$, el nivel de significancia corregido resulta ser 0.05/20 = 0.0025, por lo que el investigador concluye que todos los test son significativos. Un segundo investigador reproduce los mismos análisis en otro laboratorio y llega a los mismos resultados, pero para confirmarlos todavía más, realiza 80 test estadísticos adicionales con lo que su nivel de significancia corregido pasa a ser de 0.05/100 = 0.0005. Ahora, ninguna de las pruebas se puede considerar significativa, por lo que debido a aumentar el número de contrastes las conclusiones son totalmente contrarias. Viendo los problemas que implica, ¿para qué sirve entonces la corrección de Bonferroni? Que su aplicación en las disciplinas biomédicas no sea adecuada no descarta que pueda serlo en otras áreas. Imagínese, por ejemplo, una factoría que genera bombillas en lotes de 1000 unidades y que testar cada una de ellas antes de repartirlas no es práctico. Una alternativa consiste en comprobar únicamente una muestra de cada lote, rechazando cualquier lote que tenga más de x bombillas defectuosas en la muestra. Por supuesto, la decisión puede ser errónea para un determinado lote, pero según la teoría de Neyman-Pearson, se puede encontrar el valor x para el que se minimiza la ratio/razón de error. Ahora bien, la probabilidad de encontrar x bombillas defectuosas en la muestra depende del tamaño que tenga la muestra, o, en otras palabras, del número de pruebas que se hagan por lote. Si se incrementa el tamaño también lo hace la probabilidad de rechazar el lote, es aquí donde la corrección de Bonferroni recalcula el valor de x que mantiene minimizado la ratio de errores.

Los métodos descritos anteriormente se centran en corregir la inflación del error de tipo I ("false positive rate"), es decir, la probabilidad de rechazar la hipótesis nula siendo esta cierta. Esta aproximación es útil cuando se emplea un número limitado de comparaciones. Para escenarios de pruebas múltiples de gran escala ("large-scale multiple testing"), como los estudios genómicos en los que se realizan miles de test de forma simultánea, el resultado de estos métodos es demasiado conservativo e impide que se detecten diferencias reales. Una alternativa es controlar el "false discovery rate".

La tasa de falsos descubrimientos (FDR, por su nombre en inglés) se define como (todas las definiciones son equivalentes):

a. La proporción esperada de pruebas en los que la hipótesis nula es cierta, de entre todos los test que se han considerado significativos.

b. FDR es la probabilidad de que una hipótesis nula sea cierta habiendo sido rechazada por la prueba estadística.

c. De entre todos los test considerados significativos, el FDR es la proporción esperada de esas pruebas para los que la hipótesis nula es verdadera.

d. Es la proporción de pruebas aparentemente significativas que realmente no lo son.

e. La proporción esperada de falsos positivos de entre todas las pruebas consideradas como significativos.

El objetivo de controlar la tasa de falsos descubrimientos es establecer un límite de significancia para un conjunto de prueba tal que, de entre todos los test considerados como significativos, la proporción de hipótesis nulas verdaderas (falsos positivos) no supere un determinado valor. Otra ventaja añadida es su fácil interpretación, por ejemplo, si un estudio publica resultados estadísticamente significativos para un FDR del 10%, el lector de dicha investigación tiene la seguridad de que como

máximo un 10% de los resultados considerados como significativos son realmente falsos positivos.

Cuando un investigador emplea un nivel de significancia α, por ejemplo, de 0.05, suele esperar cierta seguridad de que solo una pequeña fracción de las pruebas significativas se correspondan con hipótesis nulas verdaderas (falsos positivos). Sin embargo, esto no tiene por qué ser así. La razón por la que una FDR baja no tiene por qué traducirse en una probabilidad baja de hipótesis nulas verdaderas entre las pruebas significativas se debe a que la FDR depende de la frecuencia con la que la hipótesis nula contrastada es realmente verdadera. Un caso extremo sería el planteado en el ejemplo 1, en el que todas las hipótesis nulas son realmente ciertas por lo que el 100% de las pruebas que en promedio resultan significativas son falsos positivos. Por lo tanto, la FDR depende de la cantidad de hipótesis nulas que sean ciertas de entre todos los contrastes[7].

Los análisis de tipo exploratorio en los que el investigador trata de identificar resultados significativos sin apenas conocimiento previo se caracterizan por una proporción alta de hipótesis nulas falsas. Los análisis que se hacen para confirmar hipótesis, en los que el diseño se ha orientado en base a un conocimiento previo, suelen tener una proporción de hipótesis nulas verdaderas alta. Idealmente, si se conociera de antemano la proporción de hipótesis nulas verdaderas de entre todos los contrastes se podría ajustar con precisión el límite significancia adecuado a cada escenario, sin embargo, esto no ocurre en la realidad. La primera aproximación para controlar el FDR fue descrita por Benjamini y Hochberg en 1995.

[7] Implícitamente aquí se pone de manifiesto que la definición ontológica de probabilidades no es frecuentista.

II. PRIMER CASO DE APLICACIÓN: DETERMINACIÓN DE LA LOCALIDAD DE PRODUCCIÓN DE LAS UVAS MÁS DULCES

II.I. Desarrollo en RStudio

En Costa Rica se cultivan dos especies de uva: roja y blanca. Como medida de dulzor se utiliza la *escala Brix*. Se tomaron medidas del diámetro mayor y de los grados brix de una muestra de 100 frutas de ambas especies en 3 sitios: La Garita de Alajuela, La Guácima y San Vito de Java. El objetivo principal era determinar si en alguna de las localidades se tienden a producir uvas más dulces.

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

## 'data.frame': 100 obs. of 4 variables:
## $ localidad: chr "Garita" "Guacima" "Sanvito" "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 ...

base$localidad=as.factor(base$localidad)
attach(base)
```

Si sólo se están comparando los promedios entre las tres localidades, ¿es necesario tomar en cuenta que las uvas son de dos tipos? Si bien no se quieren comparar las uvas según la especie, si las uvas de una especie tienen un nivel de grados Brix muy diferente al de la otra especie, sería muy importante incluir este factor en el modelo, puesto que de no hacerlo se tendría más *ruido* (el ruido estadístico es una variabilidad inexplicable dentro de una muestra de datos) traducido en una mayor variancia residual. Esto podría producir que en el experimento no se logren detectar diferencias que sí existen. En este caso no se incluye la especie porque así se está planteando el ejercicio, pero eso no significa que no deba incluirse en un escenario

de aplicación rigurosamente científico y no meramente didáctico; por el contrario, dicho factor debería incluirse en el análisis en un contexto como el especificado.

La hipótesis básica para verificar que en efecto las tres localidades no producen el mismo promedio de dulzor es

$$\mu_1 = \mu_2 = \mu_3$$

. Esta hipótesis puede verificarse como se muestra a continuación.

```
mod1=aov(brix~localidad)
anova(mod1)

##                    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
```

Así, estableciéndose como referencia un nivel de significancia de 0.05, dado que la probabilidad asociada a la prueba de igualdad de los tres promedios es de 0.004, se rechaza esta hipótesis y se concluye que no se puede esperar que los promedios de grados Brix en las tres localidades sea el mismo. Entonces, vale la pena investigar más para decidir en cuál o cuáles de ellas se puede esperar que el dulzor sea mayor.

Dado que el objetivo es comparar todas las localidades entre sí, se trata de un problema de comparación de todos los pares de promedios. Sin embargo, es importante que la hipótesis alternativa incluya la dirección apropiada que vaya a favor de los datos, es decir, si por ejemplo la media observada del primer tratamiento $\bar{y}_1$ es mayor que la media observada del segundo tratamiento $\bar{y}_2$, entonces la

hipótesis alternativa debe escribirse $H_1: \mu_1 > \mu_2$. Así, deben escribirse todos los pares de hipótesis se deban probar.

```
m=tapply(brix,localidad,mean)
m
```

```
##    Garita   Guacima   Sanvito
## 16.95526 16.10400 15.97027
```

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 > \mu_2$$

$$H_0: \mu_1 = \mu_3$$

$$H_0: \mu_1 > \mu_3$$

$$H_0: \mu_2 = \mu_3$$

$$H_0: \mu_2 > \mu_3$$

Adicionalmente, puede estimarse el Cuadrado Medio Residual.

```
cmres=anova(mod1)[2,3]
round(cmres,2)
```

```
## [1] 1.81
```

Es recomendable estimar los estadísticos de interés para realizar cada prueba, es decir, calcular las diferencias $\bar{y}_i - \bar{y}_j$, asegurándose que la hipotesis alternativa sea $\bar{y}_i > \bar{y}_j$.

```
d12=m[1]-m[2]
d13=m[1]-m[3]
d23=m[2]-m[3]
d=c(d12,d13,d23)
```

```
names(d)=c("Ga-Gu","Ga-SV","Gu-SV")
round(d,3)
```

```
##                Ga-Gu                Ga-SV                Gu-SV
## 0.851 0.985 0.134
```

Es de interés la estimación del error estándar del estadístico $\bar{y}_i - \bar{y}_j$ en cada caso usando.

$$V(\bar{y}_i - \bar{y}_j) = \sigma_\epsilon^2 \left(\frac{1}{n_i} + \frac{1}{n_j}\right)$$

```
n=table(localidad)
ee12=sqrt(cmres*(1/n[1]+1/n[2]))
ee13=sqrt(cmres*(1/n[1]+1/n[3]))
ee23=sqrt(cmres*(1/n[2]+1/n[3]))
ee=c(ee12,ee13,ee23)
names(ee)=names(d)
round(ee,3)
```

```
##                Ga-Gu                Ga-SV                Gu-SV
## 0.346 0.311 0.348
```

Complementariamente, es de interés calcular el valor estandarizado del estadístico haciendo mediante

$$q_{ij} = \frac{\bar{y}_i - \bar{y}_j}{ee_{ij}},$$

que es precisamente la distribución de rango studentizada antes mencionada.

```
q=d/ee
names(q)=names(d)
round(q,3)
```

```
##          Ga-Gu          Ga-SV          Gu-SV
## 2.457 3.170 0.384
```

Es posible determinar la probabilidad des-acumulada (la acumulada se define como la probabilidad de obtener un valor menor o igual que un valor x de referencia) de obtener un valor igual o mayor al estadístico usando la distribución del rango studentizado de Tukey. Para ello debe indicarse ptukey(q*sqrt(2),k,df,lower.tail = F), donde k es el número de grupos y df son los grados de libertad de los residuales. Además se obtiene directamente la cola derecha con el argumento lower.tail=F.

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

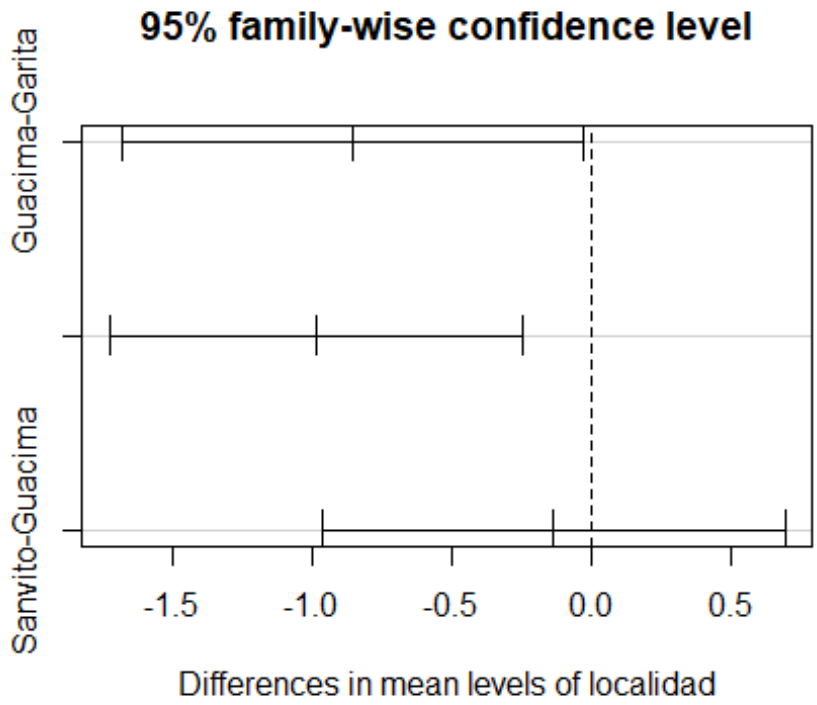
```
##          Ga-Gu          Ga-SV          Gu-SV
## 0.041 0.006 0.922
```

Para realizar comparaciones entre pares de medias con los resultados de la prueba de Tukey, deben comprobarse las probabilidades anteriores y calcular el intervalo de confianza correspondiente.

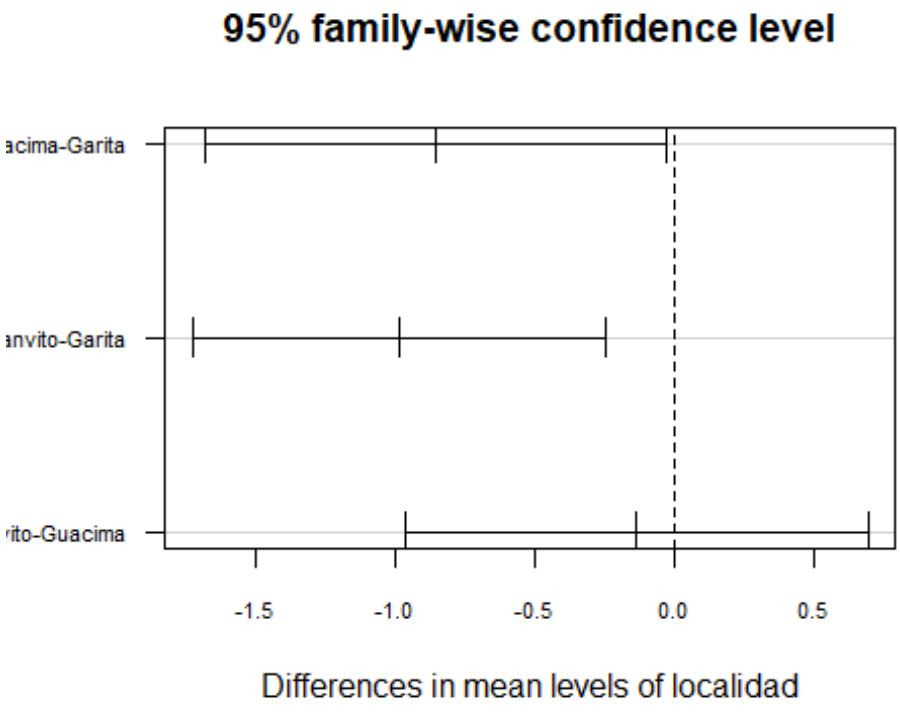
```
mod1<-aov(brix~localidad,base)
Tukey<-TukeyHSD(mod1,"localidad")
Tukey
```

```
##                  Tukey multiple comparisons of means
##                    95% family-wise confidence level
##
##  Fit: aov(formula = brix ~ localidad, data = base)
##
##                                                $localidad
##                          diff        lwr         upr     p adj
##  Guacima-Garita -0.8512632 -1.6757890 -0.02673733 0.0413810
```

```
##      Sanvito-Garita          -0.9849929    -1.7244854     -0.24550041      0.0057301
## Sanvito-Guacima -0.1337297 -0.9626653  0.69520583 0.9220088

plot(Tukey,cex.axis=1)
```




```
plot(Tukey,cex.axis=0.7,las=1)
```



La conclusión relativa a las pruebas de hipótesis generadas dados los resultados del contraste de rango de Tukey es que todas las parejas de medias de dulzor de uvas según localidad de producción puestas en comparación, únicamente la pareja Sanvito-Guácima posee una diferencia de medias estadísticamente no-significativa, puesto que el cero pertenece al intervalo de confianza generado. Esto se verifica también porque el valor p ajustado para comparaciones múltiples (que aparece en la tabla generada por la sintaxis 'aov()' como "p adj").

Esto parece indicar que las uvas producidas en la localidad de Garita son las que son diferentes, puesto que Garita es la localidad que no aparece en la pareja de medias para cuya diferencia se falla en rechazar H_0, mientras que es la localidad común en las parejas de medias para las cuales se rechaza la hipótesis nula de igualdad de medias. Lo anterior está directamente vinculado al hecho de que la media de dulzor en Garita es mayor que la de Guácima y San Vito, sin embargo, entre estas últimas dos no se ha demostrado que exista una diferencia estadísticamente significativa.

Puede obtenerse una cota inferior para la diferencia de las medias sólo en los casos en que se encontró una diferencia significativa.

Como sólo interesa conocer en cuánto es mayor un promedio que el otro, se debe obtener el valor que acumula un 95% de área en la distribución t con los grados de libertad residuales (gl), en caso se desee un nivel de confianza de 0.95. Este valor se puede ajustar por Bonferroni (con las limitaciones antes señaladas) para obtener un nivel de confianza global para los dos intervalos construidos (correspondientes a las dos diferencias de medias que sí son significativas). Esto se hace con 'qt(1-0.05/k,gl)', donde k es el número de intervalos. Luego se calcula LIM con:

$$LIM = \left(\bar{y}_i - \bar{y}_j\right) - t_{1-\alpha} ee_{ij}$$

```
t=qt(1-0.05/2,97)
lim=d[1:2]-t*ee[1:2]
names(lim)=names(d)[1:2]
round(lim,2)
```

```
##                                Ga-Gu                                Ga-SV
##  0.16  0.37
```

Adicionalmente, puede encontrarse el intervalo de confianza, pero ahora para los tres pares con Bonferroni, Scheffé y Dunnett para poder compararlos con los de Tukey.

```
localidad2=as.factor(localidad)


#Bonferroni
pairwise.t.test(brix, localidad2, p.adj = "bonf")
```

```
##
##         Pairwise comparisons using t tests with pooled SD
##
```

```
##                 data:                 brix                 and                 localidad2
##
##                                                                    Garita          Guacima
##                    Guacima                                         0.0473                -
##                    Sanvito                                         0.0061             1.0000
##
## P value adjustment method: bonferroni

library(asbio)

## Loading required package: tcltk

bonfCI(brix, localidad2, conf.level = 0.95)

##
##                 95%                 Bonferroni                 confidence                 intervals
##
##                                        Diff      Lower      Upper    Decision    Adj. p-value
##   muGarita-muGuacima     0.85126   0.00734   1.69518   Reject   H0              0.04731
##   muGarita-muSanvito     0.98499   0.22811   1.74188   Reject   H0              0.006113
## muGuacima-muSanvito 0.13373 -0.7147 0.98216    FTR H0          1

#Scheffé
scheffeCI(brix, localidad2, conf.level = 0.95)

##
##                 95%                 Scheffe                 confidence                 intervals
##
##                                        Diff      Lower      Upper    Decision    Adj. P-value
##   muGarita-muGuacima     0.85126   -0.00992  1.71244        FTR   H0              0.053442
##   muGarita-muSanvito     0.98499     0.21263   1.75736   Reject   H0              0.008379
## muGuacima-muSanvito 0.13373 -0.73206 0.99952    FTR H0     0.928978
```

```
#Dunnett (recomendado solo para comparar contra un grupo control)
dunnettCI(brix, localidad2, conf.level = 0.95, control = "Garita")

##
##                    95% Dunnett confidence intervals
##
##                                  Diff      Lower      Upper Decision
##  muGuacima-muGarita  -0.851263 -1.631058 -0.071469 Reject H0
## muSanvito-muGarita -0.984993 -1.684367 -0.285619 Reject H0
```

II.II. Conclusiones

1) Dunnet es mejor cuando es para un grupo control comparado con todos.

2) Si se quiere comparar todos los pares entre sí, Tukey genera intervalos más angostos.

3) Complementario al uso del contraste de rango de Tukey, puede utilizarse el ajuste de Bonferroni siempre que los contrastes se hayan definido antes de llevar a cabo el análisis, proporcionando de forma adicional la ventaja de permitir comparar más de dos contrastes (combinaciones lineales de los efectos del modelo y sus pesos -cuya suma es nula-).

4) Scheffé permite generar intervalos a posteriori (su diferencia más importante respecto a Bonferroni).

5) Cuando el número de comparaciones que se haga sea mayor al número de tratamientos o en general sea grande, Scheffé podría ser preferido a Bonferroni, ya que el α de comparación se vuelve muy pequeño.

Los resultados de las pruebas realizadas indican que donde hay una mayor diferencia es entre Garita y San Vito, ya que se puede esperar que Garita tenga un nivel promedio al menos de 0.37 grados Brix más que San Vito. Además, se observa que en lo que aventaja Garita a Guácima es en 0.16 grados Brix, aunque también

podría ser superior (por el hecho de fallar en rechazar la H_0) y ello pone en duda la confiabilidad estadística sobre la superioridad en grados brix de la uva de Garita frente a la de Guácima, ya que no hay una fuerte evidencia de que esté marcadamente por encima. Sin embargo, en última instancia cuánto debería ser una diferencia en grados Brix entre dos promedios para considerarse uno superior al otro y qué tanta superioridad implica esa diferencia es un criterio cuya naturaleza técnica es no-matemático y no-estadístico (más allá que estas herramientas sean complementos analíticos importantes), por lo que debe ser determinado por un experto en dicha tarea.

III. SEGUNDO CASO DE APLICACIÓN: EVALUACIÓN DE MÉTODOS DE PREVENCIÓN DE LA OXIDACIÓN DE LAS MANZANAS

III.I. Desarrollo en RStudio

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. Se busca determinar¨lo siguiente:

a. Verificar si los tratamientos definidos controlan el pardeamiento produciendo una menor puntuación promedio que el caso en que no se aplica nada (control).

b. Verificar si aplicar ácido controla mejor que cubrir (tapar o poner en bolsa).

c. Identificar cuál de las dos formas de cubrir es mejor.

El objetivo final es seleccionar el tratamiento que mantenga mejor el color original para una empresa que se encarga de banquetes para altos ejecutivos. Para ello, primero se debe probar que hay alguna diferencia en el color promedio resultante con los cuatro tratamientos. Por la naturaleza de la variable respuesta, el tratamiento que produzca un valor promedio más bajo se considera el mejor. Una vez que esta hipótesis básica de igualdad de promedios se rechaza, interesa hacer varias comparaciones para cumplir con los objetivos de la investigación:

1. En primer lugar, es necesario verificar si la respuesta promedio en realidad es menor en los tratamientos aplicados que en el caso control donde no se aplicó nada.

2. En segundo lugar, deben compararse los tratamientos donde se ha cubierto contra aquél en que se usó ácido y verificar si el promedio del segundo en realidad es menor que el promedio de los primeros.

3. Finalmente debe verificarse cuál de los dos promedios en que se cubrió es menor.

```
load("manzanas.Rdata")
attach(base)
```

Para realizar las estimaciones pertinentes, debe especificarse el modelo de **suma nula** (se le llama así por lo antes expuesto en relación a la suma de los pesos) en la sintaxis a utilizar, para lo cual debe especificarse antes de usar la sintaxis lm (usada para realizar las estimaciones de los coeficientes) la línea de código options(contrasts=c("contr.sum","contr.poly")).

```
options(contrasts=c("contr.sum","contr.poly"))
mod2=lm(color~trat)
mod2$coef
```

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

Para verificar cuál es el tratamiento que se está tomando de referencia (*i.e.*, para cuál de los tratamientos se adopta el valor de -1 en todas las variables binarias) debe utilizarse la sintaxis contrasts(trat). Las variables binarias son en este contexto aquellas que indican 1 para el tratamiento estudiado y 0 para los otros dos tratamientos que no son estudiados, mientras que el tratamiento de referencia, como se dijo, adoptará siempre el valor de -1.

```
contrasts(trat)

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

El tratamiento de control es el tratamiento de referencia.

Tras lo anterior, es recomendable plantear las hipótesis necesarias para cumplir los objetivos del problema especificado.

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

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

$$H_0: \tau_4 = \frac{1}{3}(\tau_1 + \tau_2 + \tau_3)$$

$$H_1: \tau_4 > \frac{1}{3}(\tau_1 + \tau_2 + \tau_3)$$

$$H_0: \frac{1}{2}(\tau_1 + \tau_2) = \tau_3$$

$$H_1: \frac{1}{2}(\tau_1 + \tau_2) > \tau_3$$

$$H_0: \tau_1 = \tau_2$$

$$H_1: \tau_1 > \tau_2$$

¿Por qué es equivalente plantear tales hipótesis en términos de los efectos a plantearlas en términos de los promedios de cada tratamiento? Esto es así debido a que es posible definir matemáticamente un efecto mediante $\tau_j = \mu_j - \mu$, y al usar estos τ_j en lugar de los promedios se tendría μ a ambos lados de la expresión, con lo cual se cancelan y queda establecida la expresión en términos de μ_j. De ahí su equivalencia matemática.

Es recomendable a continuación definir los contrastes ortogonales necesarios para realizar las pruebas de hipótesis pertinentes.

$$L_1: -\frac{1}{3}\tau_1 - \frac{1}{3}\tau_2 - \frac{1}{3}\tau_3 + \tau_4$$

$$L_2: \frac{1}{2}\tau_1 + \frac{1}{2}\tau_2 - \tau_3$$

$$L_3: \tau_1 - \tau_2$$

Debido a que se está usando el modelo con suma nula, se tiene la restricción $\tau_4 = -(\tau_1 + \tau_2 + \tau_3)$. Por ello, debe realizarse la sustitución de τ_4 en L_1 para que todo quede en términos de los otros 3 efectos (dado que τ_4 es el tratamiento de control).

$$L_1: -\frac{4}{3}\tau_1 - \frac{4}{3}\tau_2 - \frac{4}{3}\tau_3$$

Así, puede construirse una matriz para los coeficientes de los contrastes, para el cual debe incluirse el intercepto (que representa la media global del modelo).

```
c1=c(0,-4/3,-4/3,-4/3)
c2=c(0,0.5,0.5,-1)
c3=c(0,1,-1,0)
cont=cbind(c1,c2,c3)
cont

##             c1   c2 c3
## [1,]  0.000000  0.0  0
## [2,] -1.333333  0.5  1
## [3,] -1.333333  0.5 -1
## [4,] -1.333333 -1.0  0
```

Recuérdese que dos contrastes, cuyos coeficientes están en los vectores c_1 y c_2 (respectivamente), son ortogonales si el producto escalar de ambos vectores es cero, es decir, $c_1^T c_2 = 0$. Por ello, es requerido verificar que los tres vectores son ortogonales. Para esto deben usarse los coeficientes originales.

```
ck1=c(-1/3,-1/3,-1/3,1)
ck2=c(1/2,1/2,-1,0)
ck3=c(1,-1,0,0)
t(ck1)%*%ck2

##      [,1]
## [1,]    0

t(ck1)%*%ck3

##      [,1]
## [1,]    0

t(ck2)%*%ck3
```

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

Se verifica su independencia lineal, por lo que es válido realizar inferencias estadísticas a través de dichos coeficientes y, por tanto, puede procederse con la estimación de los contrastes. Así, si el vector de coeficientes del contraste se denota por c entonces:

$$\hat{L} = c_0\hat{\beta}_0 + c_1\hat{\tau}_1 + c_2\hat{\tau}_2 + c_3\hat{\tau}_3 = c^T\hat{\beta}$$

Sobre la ecuación anterior debe decirse lo siguiente:

1. Como se señala en (Naval Posgraduate School, 2022), las "restricciones" establecidas en el entorno R del usuario del programa determinan cómo se manejan las variables categóricas en los modelos. El esquema más común en el análisis de regresión se llama "contrastes de tratamiento": con los contrastes de tratamiento, al primer nivel de la variable categórica se le asigna el valor 0, y luego otros niveles miden el cambio desde el primer nivel.

2. Como señala (Casella, Statistical Design, 2008, pág. 13), en un análisis de varianza equilibrado unidireccional, el uso de contrastes ortogonales tiene la ventaja de dividir completamente la suma de cuadrados del tratamiento en componentes aditivos no superpuestos que representan la variación debida a cada contraste.

3. Como se señala en (Wikipedia, 2021), un contraste se define como la suma de la media de cada grupo multiplicada por un coeficiente para cada grupo (es decir, un número con signo, c). En la ecuación antes presentada, $\hat{L}$ es la suma ponderada de las medias del grupo, los coeficientes c_j representan los pesos asignados de las medias (estos deben sumar 0 para los contrastes ortogonales) y j representa las medias grupales. Los coeficientes pueden ser positivos o negativos, y fracciones o números enteros, según la comparación de interés.

Los contrastes lineales son muy útiles y se pueden usar para probar hipótesis complejas cuando se usan junto con ANOVA o regresión múltiple. En esencia, cada contraste define y prueba un patrón particular de diferencias entre las medias.

```
L=t(cont)%*%mod2$coef
round(L,3)

##      [,1]
## c1    2.8
## c2    1.2
## c3    0.4
```

Los resultados anteriores pueden verificarse empíricamente haciendo ahora el contraste basado en las medias estimadas.

```
m[4]-mean(m[1:3])

## control
##     2.8

mean(m[1:2])-m[3]

## limón
##   1.2

m[1]-m[2]

## tapar
##   0.4
```

Adicionalmente, debe encontrarse el error estándar de cada contraste mediante la ecuación:

$$V(\hat{L}) = c^T V(\hat{\beta}) c$$

```
ee=sqrt(diag(t(cont)%*%vcov(mod2)%*%cont))
round(ee,3)

##                 c1                 c2                 c3
## 0.361 0.383 0.442
```

El valor estandarizado del contraste puede determinarse mediante:

$$t = \frac{\hat{L}_j}{ee_j}$$

```
t=L/ee
round(t,3)

##                                    [,1]
##                 c1                7.755
##                 c2                3.133
## c3 0.905
```

Es posible calcular la probabilidad de obtener un valor igual o mayor al estadístico usando la distribución t. Esto se puede hacer sin importar la cantidad de comparaciones que se hagan simultáneamente debido a que los contrastes son ortogonales. Para ello debe utilizarse la sintaxis 'pt(t,36,lower.tail=F)', que requiere especificarle los grados de libertad de los residuales, que en este caso son $N - k = 40 - 4 = 36$, puesto que son 40 observaciones y cuatro coeficientes a estimar (incluyendo el intercepto).

```
p=pt(t,36,lower.tail=F)
round(p,3)

##                                    [,1]
##                 c1                0.000
##                 c2                0.002
## c3 0.186
```

Las primeras dos probabilidades son menores a 0.05 por lo que se rechazan la primera y la segunda hipótesis, es decir, se encontró que la media de color es mayor cuando no se aplica ningún tratamiento que en los otros casos, además que usar limón ácido controla mejor que cubriendo. Se compararon los dos tratamientos en que se cubre y no se demostró que usar bolsa sea mejor que tapar.

Adicionalmente, es posible realizar la misma prueba utilizando en lugar de la distribución t la distribución F mediante las sumas de cuadrados de los contrates. Por ello, es necesario determinar tales sumas para cada uno de los contrastes. Para ello, deben utilizarse los coeficientes de los contrastes como se definieron originalmente. Se tiene que cada suma de cuadrados se obtiene mediante:

$$SCCont = \frac{r\hat{L}^2}{\sum c_j^2}$$

```
ck1=sum(c(-1/3,-1/3,-1/3,1)^2)
ck2=sum(c(1/2,1/2,-1,0)^2)
ck3=sum(c(1,-1,0,0)^2)
ck=c(ck1,ck2,ck3)
table(trat)

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

sccont=(10*L^2)/ck
sccont

## [,1]
## c1 58.8
## c2 9.6
## c3 0.8
```

Puede verificarse que la suma de las tres SCCont es igual a la Suma de Cuadrados de Tratamiento (que son los cuadrados totales de los tratamientos).

```
sum(sccont)

## [1] 69.2

anova(mod2)[1,2]

## [1] 69.2
```

De forma complementaria, es de interés estimar el valor F asociado a la prueba F para cada contraste, recordando que la suma de cuadrados de cada contraste tiene un grado de libertad cuando estos contrastes son ortogonales. De esta forma, la F se construye como:

$$F = \frac{SCCont}{CMRes}$$

```
cmres=anova(mod2)[2,3]
f=sccont/cmres
p1=pf(f,1,36,lower.tail=F)
round(p1,3)

##                                  [,1]
## c1                              0.000
## c2                              0.003
## c3 0.372
```

Debido a que con la prueba F se comparan las medias y tiene como hipótesis alternativa la diferencia entre ellas, esta prueba es equivalente a una prueba de dos colas. Puesto que interesa hacer la prueba con una sola dirección, los valores de las probabilidades obtenidos anteriormente deben dividirse por 2 para que la prueba

sea de una cola. Por lo anterior, pueden obtenerse tales probabilidades y compararlas con las obtenidas en el punto anterior.

```
round(p1/2,3)

##                [,1]
## c1            0.000
## c2            0.002
## c3 0.186
```

Como puede observarse, se llega a las mismas conclusiones.

También es posible estimar una cota inferior para la diferencia de las medias sólo en los casos en que se encontró una diferencia significativa (c_1 y c_2).

```
t=qt(0.95,36)
lim=L[1:2]-t*ee[1:2]
names(lim)=names(L)[1:2]
round(lim,2)

## [1] 2.19 0.55
```

El promedio de color cuando no se aplica nada es al menos 2.2 puntos mayor que en los otros tres casos en conjunto. Sabiendo que la escala va de 1 a 6, tener 2 puntos en promedio más es una cantidad importante, por lo que se nota que aplicar alguno de estos tratamientos ayuda a mejorar el color. Por otra parte, aplicar limón da un color promedio al menos 0.55 puntos menor que cubrir. Esta diferencia no es tan grande como para afirmar que definitivamente el limón esté produciendo una mejora con respecto a cubrir. Aquí el experto debe dar su valoración de cuánto es una diferencia que para él o ella sea relevante.

Además, pueden realizarse dos nuevas comparaciones adicionales bajo el escenario en la que el investigador quisiera solamente saber si cubrir da mejores resultados que no hacer nada y también si poner limón da mejores resultados que no hacer

nada. En ambos casos se quiere cuantificar en cuanto es esta mejoría, en caso de existir.

Para ello, deben establecerse las hipótesis, escribir los contrastes asociados y verificar si cumplen o no la condición de ortogonalidad.

$$H_0: \tau_4 = \frac{1}{2}(\tau_1 + \tau_2)$$

$$H_1: \tau_4 > \frac{1}{2}(\tau_1 + \tau_2)$$

$$H_0: \tau_4 = \tau_3$$

$$H_1: \tau_4 > \tau_3$$

$$L_1: -\frac{1}{2}\tau_1 - \frac{1}{2}\tau_2 + \tau_4$$

$$L_1: -\tau_3 + \tau_4$$

```
c1=c(0,-3/2,-3/2,-1)
c2=c(0,-1,-1,-2)
cont=cbind(c1,c2)
cont
```

```
##                    c1   c2
##      [1,]          0.0    0
##      [2,]         -1.5   -1
##      [3,]         -1.5   -1
## [4,] -1.0 -2
```

```
ck1=c(-1/2,-1/2,0,1)
ck2=c(0,0,-1,1)
t(ck1)%*%ck2
```

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

Como se verifica, el producto de los dos vectores de coeficientes de los contrastes es distinto de cero, por lo que estos contrastes no son ortogonales.

Pueden estimarse los coeficientes de los contrastes, los errores estándar y el valor estandarizado del contraste.

```
L=t(cont)%*%mod2$coef
L
```

```
##                                                                  [,1]
##                                c1                                2.4
## c2  3.6
```

```
ee=sqrt(diag(t(cont)%*%vcov(mod2)%*%cont))
round(ee,2)
```

```
##                                      c1                          c2
## 0.38 0.44
```

```
t=L/ee
round(t,2)
```

```
##                                                                  [,1]
##                                      c1                        6.27
## c2 8.14
```

Finalmente, debe encontrarse la probabilidad de obtener un valor igual o mayor al estadístico usando la distribución t. Como se trata de dos pruebas simultáneas con contrastes no ortogonales, debe hacer alguna corrección pertinente, como por ejemplo la corrección de Bonferroni. Esta consiste en dividir el nivel de significancia

(α) por el número de pruebas que se realizan (k), entonces la probabilidad asociada a cada prueba se compara contra α/k.

```
p=pt(t,36,lower.tail=F)
round(p,3)

##                                                                    [,1]
##                    c1                                                0
## c2    0

k=2; 0.05/2

## [1] 0.025
```

III.II. Conclusiones

En ambos casos la probabilidad obtenida es menor a 0.025, por lo que se rechaza la hipótesis nula y se concluye que con un nivel de significancia global de 0.05 el promedio de color cuando se cubre es menor que cuando no se hace nada, y de la misma forma es menor cuando se usa limón que cuando no se hace nada.

IV. TERCER CASO DE APLICACIÓN: ESTIMACIÓN NO-PARAMÉTRICA DE LA CANTIDAD PROMEDIO DE MOSCAS SEGÚN TIPO DE VEGETACIÓN

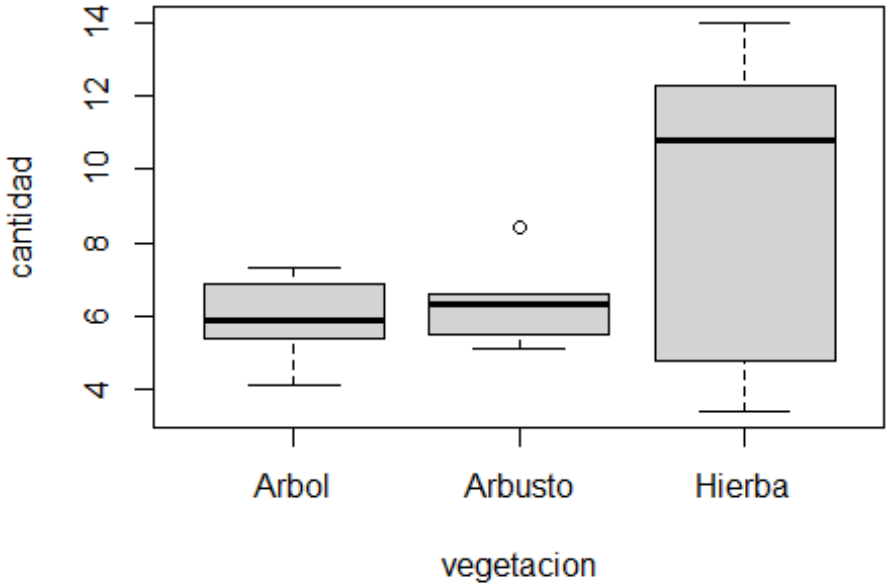
IV.I. Desarrollo en RStudio

Como se señala en (Heidel, 2022), la prueba de Kruskal-Wallis es el equivalente no paramétrico de un ANOVA (análisis de varianza). Kruskal-Wallis se usa cuando los investigadores comparan tres o más grupos independientes en un resultado continuo, pero se viola la suposición de homogeneidad de varianza entre los grupos en el análisis ANOVA. La prueba de Kruskal-Wallis es resistente a las violaciones de esta suposición estadística. Los investigadores deberán informar las medianas y

los rangos intercuartílicos en lugar de las medias y las desviaciones estándar cuando utilicen la prueba de Kruskal-Wallis.

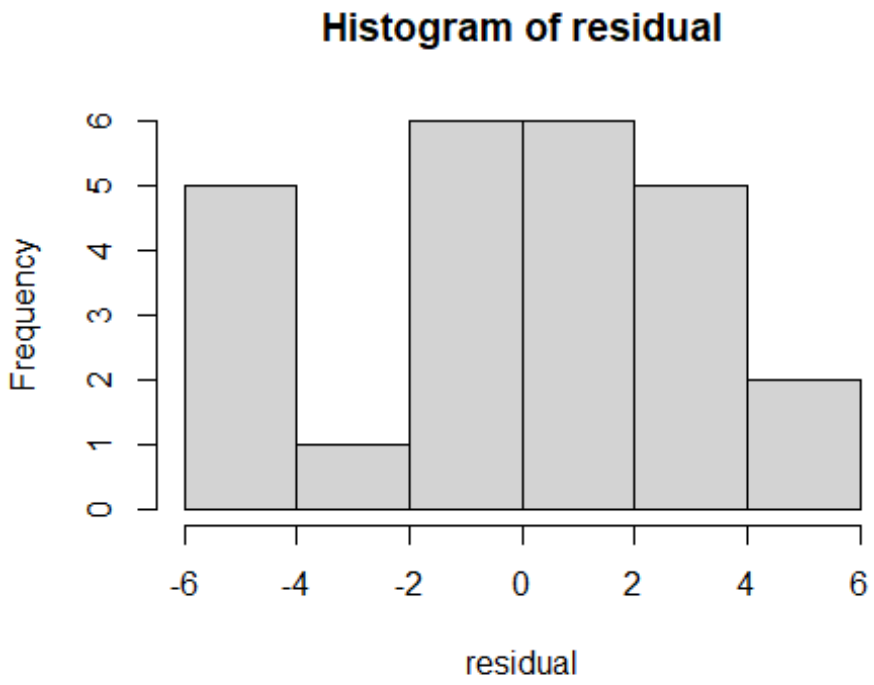
Así, supóngase que se quiere estimar la cantidad promedio de moscas a encontrar en un lugar según su tipo de vegetación (arbusto, árbol o hierba) y para ello se cuenta con la base de datos moscas.csv que contiene un total de 25 observaciones de la cantidad de moscas existentes en cada bloque de una determinada superficie, desglosadas en 15 observaciones para hierba, 5 para arbusto y 5 para árbol. Este se trata de un caso de diseño de bloques incompleto (la totalidad de los tratamientos -tipos de vegetación- no están en cada bloque, aunque los bloques son de tamaño homogéneo).

```
moscas<-read.csv("moscas.csv")
str(moscas)

## 'data.frame': 25 obs. of 2 variables:
## $ cantidad : num 3.4 3.9 4.2 4.7 4.8 5.8 10.5 10.8 10.9 11.8 ...
## $ vegetacion: chr "Hierba" "Hierba" "Hierba" "Hierba" ...

attach(moscas)
boxplot(cantidad~vegetacion)
```

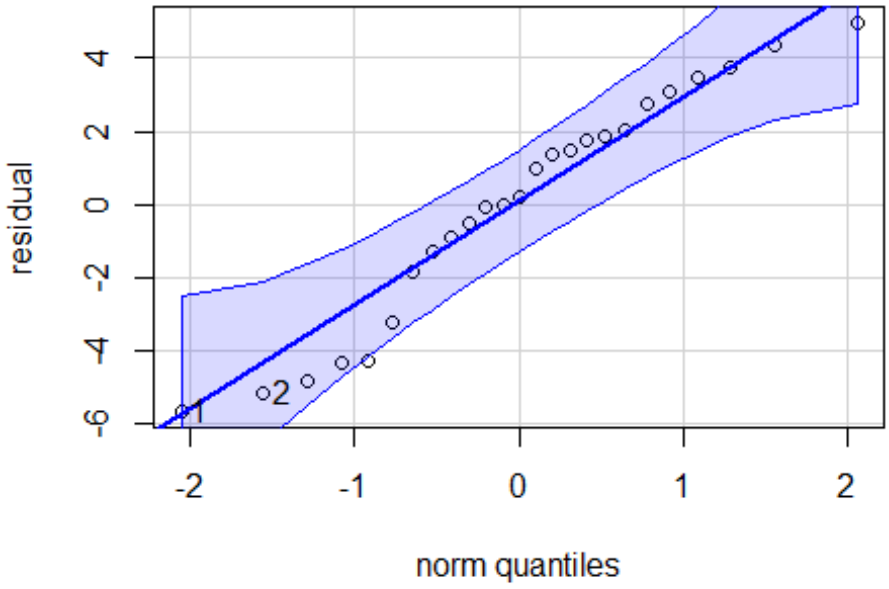



```
mod1<-lm(cantidad~vegetacion)
anova(mod1)

## Analysis of Variance Table
##
## Response: cantidad
## Df Sum Sq Mean Sq F value Pr(>F)
## vegetacion 2 50.642 25.321 2.347 0.1192
## Residuals 22 237.352 10.789

#Prueba de homogeneidad de variancias
bartlett.test(cantidad~vegetacion) #supone normalidad

##
## Bartlett test of homogeneity of variances
##
```

```
## data:  cantidad by vegetacion
## Bartlett's K-squared = 8.8329, df = 2, p-value = 0.01208

library(car)

## Loading required package: carData

leveneTest(cantidad~vegetacion,center=median,data=moscas) #sin norm

## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  2  3.6442 0.04295 *
##       22
##       ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Prueba de normalidad
residual<-mod1$res
hist(residual)
```

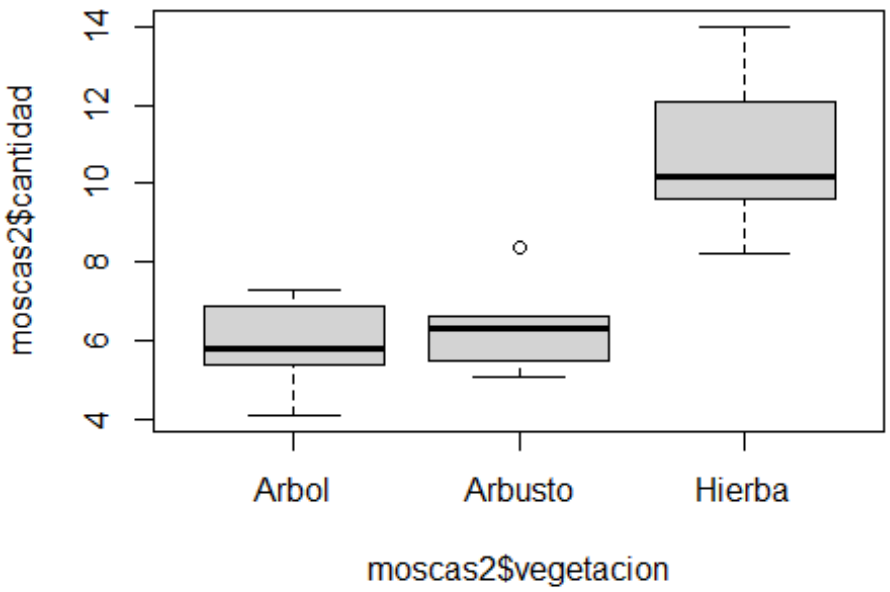



```
shapiro.test(mod1$res)

##
##                    Shapiro-Wilk                    normality                    test
##
##                          data:                                          mod1$res
## W = 0.948, p-value = 0.2259

qqPlot(residual)
```




```
## [1] 1 2

oneway.test(cantidad~vegetacion) #Prueba de Welch asume normalidad, varianzas diferentes.

##
##      One-way analysis of means (not assuming equal variances)
##
## data:  cantidad and vegetacion
## F = 3.3791, num df = 2.000, denom df = 12.461, p-value = 0.06722

kruskal.test(cantidad~vegetacion)  #Prueba no parametrica

##
##      Kruskal-Wallis rank sum test
##
```

```
## data: cantidad by vegetacion
## Kruskal-Wallis chi-squared = 1.5557, df = 2, p-value = 0.4594

#Comparaciones múltiples (no paramétrica)
#Como ejercicio se cambian algunos datos, para probar las comparaciones no-paramétricas.

moscas2<-read.csv("moscas2.csv")
boxplot(moscas2$cantidad~moscas2$vegetacion)
```




```
#install.packages("FSA", dependencies = T)
library(FSA)

## Registered S3 methods overwritten by 'FSA':
##   method       from
##   confint.boot car
##   hist.boot    car
```

```
## ## FSA v0.9.3. See citation('FSA') if used in publication.
## ## Run fishR() for related website and fishR('IFAR') for related book.

##
## Attaching package: 'FSA'

## The following object is masked from 'package:car':
##
##    bootCase

library(dunn.test)
kruskal.test(moscas2$cantidad~moscas2$vegetacion)  #Prueba no paramétrica

##
##                 Kruskal-Wallis rank sum test
##
##         data: moscas2$cantidad by moscas2$vegetacion
## Kruskal-Wallis chi-squared = 8.72, df = 2, p-value = 0.01278

tapply(moscas2$cantidad,moscas2$vegetacion,mean)

##                 Arbol              Arbusto              Hierba
##    5.90    6.38   10.82

dunnTest(moscas2$cantidad~moscas2$vegetacion)

## Warning: moscas2$vegetacion was coerced to a factor.

## Dunn (1964) Kruskal-Wallis multiple comparison

##   p-values adjusted with the Holm method.

##                  Comparison          Z      P.unadj        P.adj
##     1        Arbol - Arbusto -0.2828427 0.777297411   0.77729741
```

```
## 2 Arbol - Hierba -2.6870058 0.007209571 0.02162871
## 3 Arbusto - Hierba -2.4041631 0.016209541 0.03241908
```

IV.II. Conclusiones

Acá no se da un intervalo, solo indica cuales son diferentes. Considerando que la relación árbol-arbusto obtuvo un valor p ajustado de 0.77, la relación árbol-hierba uno de 0.02 y la relación arbusto-hierva de 0.03, se concluye que sólo los promedios de moscas que hay en la vegetación de tipo árbol y de tipo hierba son significativamente diferentes.

V. REFERENCIAS

Amat Rodrigo, J. (Enero de 2016). *ANOVA análisis de varianza para comparar múltiples medias*. Obtenido de Ciencia de Datos, Estadística, Machine Learning y Programación: https://www.cienciadedatos.net/documentos/19_anova.html

Casella, G. (2008). *Statistical Design.* New York: Springer.

Casella, G. (22 de Julio de 2022). *Statistical Design. Principles, Recommendations, and Opinions.* Obtenido de University of Florida: https://archived.stat.ufl.edu/casella/STA6209/SDSC09.pdf

Cross Validated. (5 de Enero de 2012). *What is model identifiability?* Obtenido de StackExchange: https://stats.stackexchange.com/questions/20608/what-is-model-identifiability

Fallas, J. (2012). *ANÁLISIS DE VARIANZA. Comparando tres o más medias.* Obtenido de Universidad para la Cooperación Internacional: https://www.ucipfg.com/Repositorio/MGAP/MGAP-05/BLOQUE-ACADEMICO/Unidad-2/complementarias/analisis_de_varianza_2012.pdf

Glen, S. (22 de Julio de 2022). *Tukey Test / Tukey Procedure / Honest Significant Difference*. Obtenido de StatisticsHowTo: https://www.statisticshowto.com/tukey-test-honest-significant-difference/

Heidel, E. (22 de Julio de 2022). *Kruskal-Wallis and Homogeneity of Variance*. Obtenido de Scalestatistics: https://www.scalestatistics.com/kruskal-wallis-and-homogeneity-of-variance.html

Kelley, K., & Preacher, K. J. (2012). On Effect Size. *Psychological Methods, 17*(2), 137-152. Obtenido de https://pubmed.ncbi.nlm.nih.gov/22545595/

Minitab. (22 de Julio de 2022). *What is the pooled standard deviation?* Obtenido de Support: https://support.minitab.com/en-us/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/what-is-the-pooled-standard-deviation/

Naval Posgraduate School. (22 de Julio de 2022). *Setting and Keeping Contrasts*. Obtenido de R: http://faculty.nps.edu/sebuttre/home/r/contrasts.html

Oxford Reference. (22 de Julio de 2022). *overparameterized model*. Obtenido de Overview: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803100258143

Scientific European Federation of Osteophats. (22 de Julio de 2022). *PRUEBAS POST HOC*. Obtenido de Content: https://www.scientific-european-federation-osteopaths.org/wp-content/uploads/2019/01/PRUEBAS-POST-HOC.pdf

Wikipedia. (10 de Junio de 2021). *Contrast (statistics)*. Obtenido de Multiple comparisons: https://en.wikipedia.org/wiki/Contrast_(statistics)

Wikipedia. (26 de Diciembre de 2021). *Método de Scheffé*. Obtenido de Contraste de hipótesis: https://es.wikipedia.org/wiki/M%C3%A9todo_de_Scheff%C3%A9