Tarea dos de Diseño de Experimentos

Sofía Cuartas García (C.C. 1.000.874.937) – scuartasg@unal.edu.co
Simón Cuartas Rendón (C.C. 1.037.670.103) – scuartasr@unal.edu.co

Septiembre 7, 2022

Punto uno

- ***Enunciado.*** *Se desea investigar el efecto del pH en el crecimiento de cierto microorganismo en un medio específico. Para ello se realiza un experimento, teniendo como punto de partida la misma cantidad de microorganismos. Se hacen cuatro repeticiones y se obtienen los siguientes resultados:*

Table 1: Crecimiento porcentual promedio según el nivel de pH

Nivel de pH	Crecimiento porcentual promedio
1	80
2	105
3	75

¿Estos datos son evidencia suficiente para afirmar que los niveles de pH donde se logra menor y mayor crecimiento son el tres y el dos respectivamente? Explique y liste las consideraciones que se deben tener en cuenta para que las conclusiones obtenidas sean válidas.

Consideraciones básicas del experimento

A continuación se explicarán los elementos básicos relacionados al diseño de este experimento.

- **Unidad experimental.** Cada una de las poblaciones de microorganismos consideradas para hacer la medición de su población.
- **Factor.** Nivel de pH.
- **Niveles.** Esta variable fue categorizada para estudiar el efecto que tienen tres valores de pH sobre la variable de respuesta: $pH = 1, 2, 3$.
- **Variable respuesta.** Crecimiento de cierta especie de microorganismo.
- **Garantía de la aleatorización del experimento.** Se puede asignar de forma aleatoria el nivel de pH que tendrá el medio en el que habita cada grupo con ayuda de R, definiendo los grupos 'uno', 'dos' y 'tres', se ejecuta un comando del tipo `sample(1:3, size = 3, replace = FALSE)`, y el primer elemento se le asigna al primer grupo y de forma análoga para los otros dos. Luego, se estudian los microorganismos de forma aleatoria ejecutando un comando similar para los n microorganismos que hayan sido considerados para la realización del experimento.

¿Se puede concluir sobre el crecimiento de los microorganismos?

En primer lugar es importante tener presente que no se puede asumir un acercamiento determinístico a este problema para considerar si alguno de los tres tratamientos es mayor o menor que los otros a partir del valor de la media muestral reportada en la tabla 1, ya que esto implica pasar por alto la variabilidad que pudo ser observada en cada una de las repeticiones de este experimento para los tres tratamientos considerados.

Entonces, una primera aproximación para poder hacer juicios basados en la evidencia estadística es mediante un análisis de varianza con la tabla ANOVA para este experimento de un solo factor, la cual permite determinar si la media de cada uno de los tratamientos considerados es igual o no desde un punto de vista estadístico. Así pues, si Y_{ij} es el crecimiento promedio porcentual de los microorganismos de acuerdo con el i-ésimo nivel de pH en la j-ésima repetición, entonces tal modelo es:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim NID(0, \sigma^2), \quad i = 1, 2, 3, \quad j = 1, 2, 3, 4$$

donde μ es la media global, por lo que es un parámetro común a todos los tratamientos, τ_i es un parámetro único del i-ésimo tratamiento y ε_{ij} es el error aleatorio relacionado a cada repetición para cada tratamiento.

Sin embargo, como el mismo nombre lo dice, este método requiere de la varianza para poder llevar a cabo el análisis, pero en este caso solo conocemos el valor de las medias para cada tratamiento (los $\bar{y}_{i.},\ i = 1, 2, 3$), lo cual impide estudiar si quiera si existe al menos una media que sea diferente (mayor o menor) a las otras. Esto quiere decir que, deberíamos conocer los siguientes valores: $y_{11}, \ldots, y_{14}, y_{21}, \ldots, y_{24}, y_{31}, \ldots, y_{41}, \ldots, y_{44}$, así como los residuales para poder calcular la suma de los cuadrados totales y la suma de los cuadrados asociados a los tratamientos, necesarios para poder obtener los valores que se reportan en una tabla ANOVA.

Y a propósito de los residuales, un asunto muy importante es que deben ser estudiados los supuestos del modelo ANOVA, y es que los errores aleatorios son independientes e idénticamente distribuidos siguiendo una normal con media nula y homocedásticos. Para poder chequear estos supuestos es necesario conocer los residuales del experimento, teniendo en cuenta que hacer conclusiones a partir de metodologías estadísticas sin validar los supuestos que las sustentan es incorrecto y puede derivar a conclusiones inadecuadas.

Asimismo, y ahora con el objetivo de determinar si algún tratamiento en particular tiene resultados en promedio mayores o menores a los de algún otro tratamiento (o varios de ellos), se podría considerar cualquiera de los testes que comparan las medias de los tratamientos de un experimento, tales como el test de Tukey o el de Duncan, pero semejante al caso de la ANOVA, es necesario conocer los valores obtenidos en cada una de las cuatro repeticiones para los cuatro tratamientos considerados, de tal forma que se puedan calcular las varianzas para obtener los estadísticos de prueba de estos testes.

En conclusión, no se tienen elementos suficientes para sugerir que alguno de los tratamientos considerados derivan en un crecimiento porcentual promedio significativamente mayor, menor o igual que los otros dos desde el punto de vista estadístico.

Punto dos. Problema uno.

- **Selección del punto.** La suma de los últimos dígitos de las cédulas de los dos estudiantes que presentamos este trabajo es diez, cuyo módulo cinco es cero, por lo que nos corresponde el problema uno.
- ***Enunciado.*** Se hace un estudio sobre la efectividad de tres marcas de espray para matar moscas. Para ello, cada espray se aplica a un grupo de 100 moscas, y se cuenta el número de moscas muertas, expresado en porcentajes. Se hacen seis réplicas y los resultados obtenidos se muestran en la tabla 2.

Table 2: Número de moscas muertas en cada una de las seis repeticiones para las tres diferentes marcas de *espray* consideradas.

Marca de espray	Nro. de moscas muertas
1	72 65 67 75 62 73
2	55 59 68 70 53 50
3	64 74 61 58 51 69

I. Identificar los elementos básicos del modelo

Identificar la unidad experimental, el factor, los niveles del factor y la variable de respuestas.

- **Unidad experimental.** Es cada una de los dieciocho grupos de cien moscas considerados, a los que se les aplicó uno de los tres *esprays* diferentes.
- **Factor.** Se está controlando la marca del *espray* que está siendo usado. Este es un factor controlable.
- **Niveles del factor.** Son cada una de las tres marcas distintas de *espray* para matar moscas que fueron considerados. De acuerdo con la tabla 2, estas marcas fueron denominadas como marca *'uno'*, marca *'dos'* y marca *'tres'*.
- **Variable respuesta.** Está dada por la cantidad de moscas que mueren en el grupo original de cien moscas luego de haber aplicado el *espray*.

II. Escribir el modelo estadístico asociado al respectivo diseño, indicando cuál es la variable respuesta y los supuestos del modelo.

De acuerdo a lo descrito en el numeral anterior, se está considerando un experimento de un solo factor, de tal suerte que los datos pueden ser estudiados siguiendo el modelo estadístico lineal de **análisis de varianza** o **de un solo factor**, que es conocido más popularmente como ***modelo ANOVA***. Así, si se considera a Y_{ij} como el número de moscas que mueren dentro de un grupo original de cien moscas luego de aplicarle el *espray* de la i-ésima marca en la j-ésima repetición, se tiene entonces que el modelo de análisis de varianza viene dato por:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim NID(0, \sigma^2), \quad i = 1, 2, 3, \quad j = 1, 2, 3, 4, 5, 6$$

donde μ es la media global, por lo que es un parámetro común a los tres tratamientos, τ_i es un parámetro único del i-ésimo tratamiento y ε_{ij} es el error aleatorio relacionado a cada repetición para cada tratamiento.

Nótese que se está suponiendo que los errores aleatorios del modelo ANOVA son independientes entre sí y se distribuyen idénticamente como una normal de media nula con varianza constante.

III. Verificar los supuestos del modelo:

Normalidad, homogeneidad de varianza (prueba de Bartlett y la prueba de Levene modificada) e independencia.

Primero se va a comenzar estudiando que se cumpla el supuesto de **normalidad**, por lo que prueba de hipótesis que se tiene para validar este supuesto es:

$$\begin{cases} H_0 : \varepsilon_{ij} \sim Normal \\ H_1 : \varepsilon_{ij} \nsim Normal \end{cases}$$

Para poder dirimir esta prueba de hipótesis con ayuda de R, se van a comenzar diligenciando los datos en un vector, para así obtener la tabla ANOVA que permita conseguir los residuales y así poder aplicarles el test correspondiente, que para este caso van a ser considerados dos: uno gráfico (gráfico cuantil-cuantil) y uno analítico (prueba de Shapiro Wilk, considerando que se tiene una muestra que puede tomarse como pequeña), para el cual se va a considerar un nivel de significancia de $\alpha = 0.05$.

```
# Vector con los valores obtenidos
y <- c(72,65,67,75,62,73, 55,59,68,70,53,50,64,74,61,58,51,69)

# Vector con el espray asociado a cada valor del vector anterior
factor <- c(rep(1,6),rep(2,6),rep(3,6))

# Los valores del vector anterior son nominales
factor <- as.factor(factor)
```

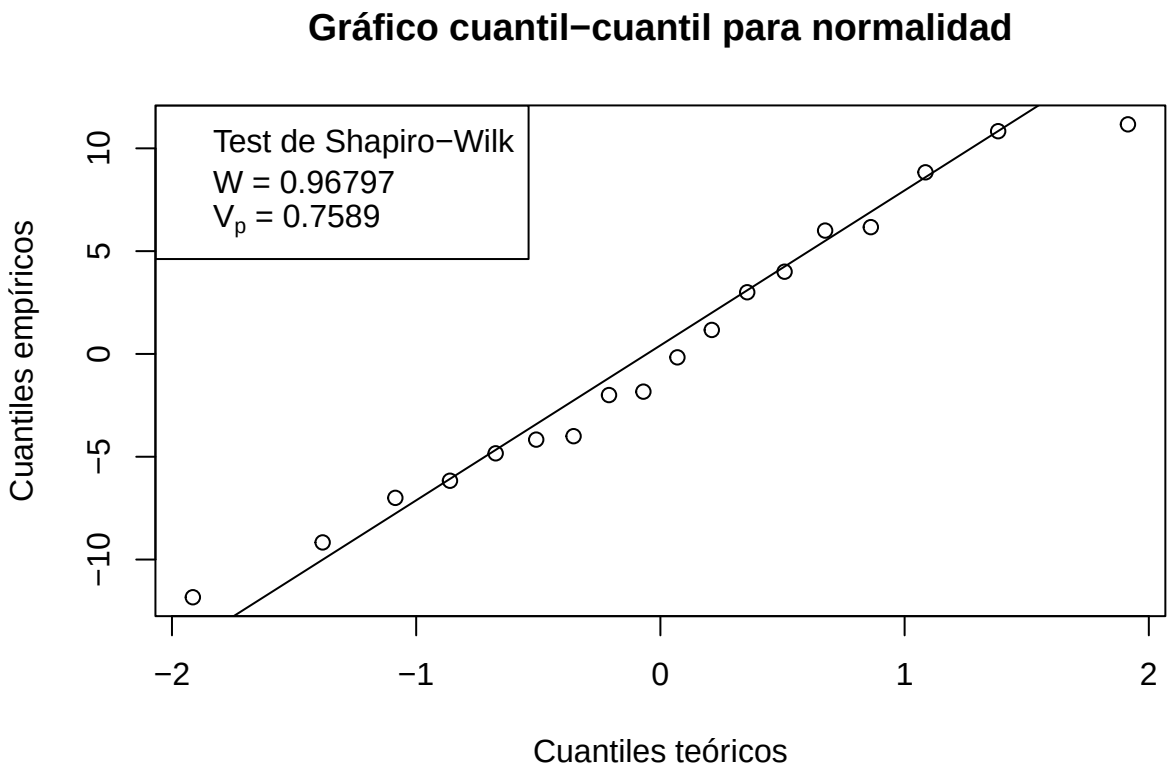
```
# Tabla ANOVA
anv <- aov(y~factor)

# Test de Shapiro-Wilk
shapiro.test(anv$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anv$residuals
## W = 0.96797, p-value = 0.7589
```

```
leyenda <- c('Test de Shapiro-Wilk',
             expression(paste('W', ' = ', '0.96797')),
             expression(paste('V'[p], ' = ', '0.7589')))

# Gráfico cuantil-cuantil
qqnorm(residuals(anv),
       main = 'Gráfico cuantil-cuantil para normalidad',
       xlab = 'Cuantiles teóricos',
       ylab = 'Cuantiles empíricos')
legend('topleft', leyenda)
qqline(residuals(anv))
```



Así pues, de la gráfico cuantil-cuantil se evidencia que los cuantiles de la muestra se apegan de forma considerable a los cuantiles teóricos, por lo que resulta razonable considerar que los errores se distribuyen siguiendo una distribución normal. Ahora bien, considerando la prueba analítica de Shapiro-Wilk, considerando que se tiene un valor p $V_p = 0.7598 > 0.05$, no se encuentra evidencia muestral suficiente en contra del supuesto de que los errores se distribuyen normal con una significancia de $\alpha = 0.05$.

A continuación, se usa la prueba de Bartlet para verificar el supuesto de homocedasticidad, que tiene por hipótesis lo siguiente:

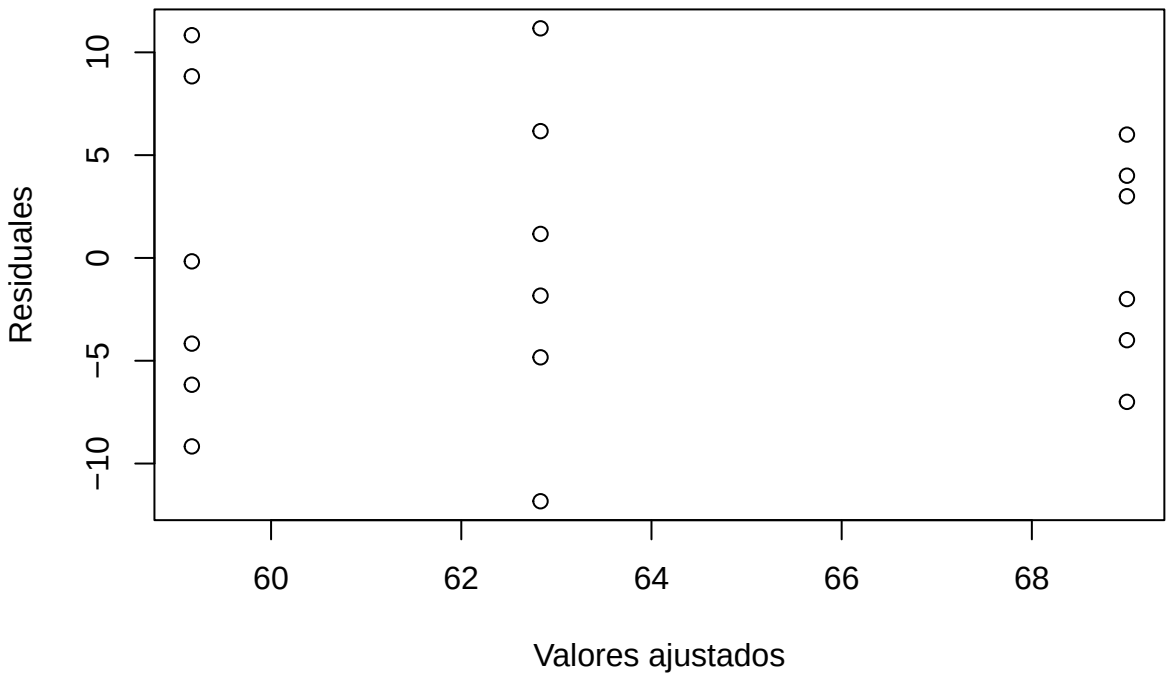
$$\begin{cases} H_0 : Var(\varepsilon_{ij}) = \sigma^2 \ \forall (i,j), \ i = 1,2,3, \ j = 1,2,3,4,5,6 \\ H_1 : Var(\varepsilon_{ij}) \ no \ es \ constante \end{cases}$$

Así pues, con ayuda de R, se tiene que:

```
bartlett.test(y ~ factor)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  y by factor
## Bartlett's K-squared = 1.1889, df = 2, p-value = 0.5519
```

```
plot(anv$fitted.values,anv$residuals,
     xlab="Valores ajustados",
     ylab = "Residuales")
```



Como el valor p es $V_p 0.5519 > 0.05$, no se rechaza la hipótesis nula, por tanto no se tiene evidencia muestral suficiente para decir que los errores no tienen varianza constante, con un nivel de significancia de $\alpha = 0.5$. Además, en el gráfico de valores ajustados contra residuales no se observa ningún patrón que el indique que no hay varianza constante.

Luego, se pasa a verificar el cumplimiento del supuesto de independencia entre los errores, gráficamente y usando el test Durbin-Watson el cual tiene como hipótesis:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

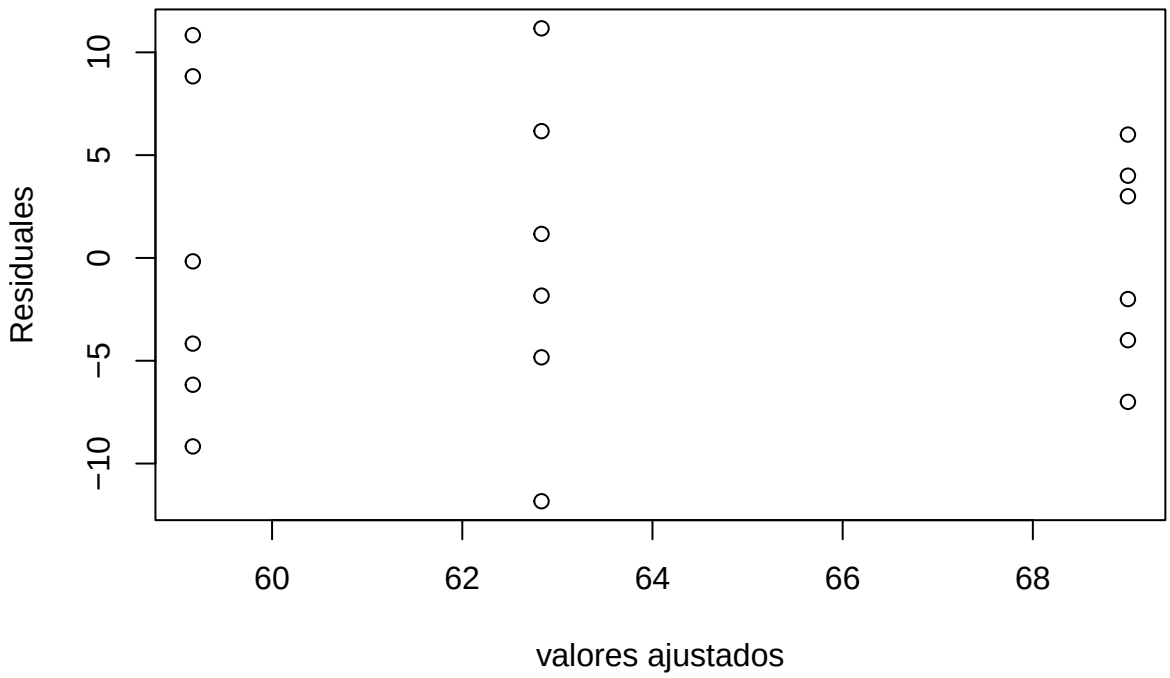
,

donde $\varepsilon_t = \rho\ \varepsilon_{t-1} + a_t\ con\ a_t \sim NID(0, \sigma_a^2)$

```
require(car)
durbinWatsonTest(anv)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.05414747       2.04919   0.666
##  Alternative hypothesis: rho != 0
```

```
plot(anv$fitted.values,anv$residuals, xlab="valores ajustados", ylab = "Residuales")
```



Como el valor p es $V_p = 0.65 > 0.05$, no hay evidencias suficientes para rechazar H_0, por lo tanto, los errores no tienen correlación de orden uno; por tanto no hay evidencias suficientes para rechazar el supuesto de independencia entre los errores.

IV. Tabla ANOVA.

Construir la tabla ANOVA y probar la significancia de los efectos de los tratamientos. Use $\alpha = 0.05$***. A partir de las expresiones para SSTrat, SSE, etc, verifique los resultados de la tabla ANOVA reportada por el R.***

Se construye la tabla ANOVA como sigue:

```
n <- 6                                                  # Repeticiones
a <- 3                                                  # Tratamientos
N <- n * a                                              # Cantidad total de observaciones
yi. <- c(sum(y[1:6]), sum(y[7:12]), sum(y[13:18]))      # Sumatoria para cada tratamiento
y.. <- sum(y)                                           # Sumatoria general
sstrat <- sum(yi.^2) / n - (y.. ^ 2) / N
sst <- sum(y ^ 2) - ((y.. ^ 2) / N)
sse <- sst - sstrat
mstrat <- sstrat / (a - 1)
mse <- sse / (N - a)
estf <- mstrat / mse                                    # Cálculo del estadístico F
pval <- pf(estf, (a - 1), (N - a), lower.tail = F)      # Cálculo del valor p
```

Y se puede construir una tabla como la que devolvería R como sigue:

```
tabla<-data.frame("Fuente de variacion" = c("Marca de espray", "residuales"),
                  "g.l" = c(a-1,N-a),
                  "Suma Cuad." = c(sstrat, sse),
                  "Media Cuad." = c(mstrat,mse),
                  "EstF" = c(estf, NA),
                  "P- Value" = c(pval,NA)
                  )
tabla
```

```
##   Fuente.de.variacion g.l Suma.Cuad. Media.Cuad.     EstF   P..Value
## 1     Marca de espray   2   296.3333   148.16667 2.793255 0.09307091
## 2          residuales  15   795.6667    53.04444       NA         NA
```

Se quiere probar si los efectos de cada uno de los tratamientos, τ_i son cero, lo que equivale a la prueba de hipótesis:

$$\begin{cases} H_0 : \tau_1 = \tau_2 = \tau_3 = 0 \\ H_1 : \exists\ \tau_i \neq 0,\ \ i = 1, 2, 3 \end{cases}$$

Con el p-valor obtenido en los anteriores cálculos y el nivel de significancia que fijó en $\alpha = 0.05$, se tiene que $0.09 > 0.05$, por lo que no se rechaza la hipótesis nula y se concluye que no hay evidencia suficiente para decir que la marca de *espray* influye en el número de moscas muertas.

V. Agrupaciones entre medias de los tratamientos.

Según sea el caso, realice agrupaciones entre medias de los tratamientos. Usar el método de Duncan. Plantee un contraste para comparar la media del primer tratamiento con las otras medias y haga la prueba de significancia de dicho contraste,use $\alpha = 0.05$***.***

Comenzando pues con el test Duncan para agrupaciones entre medias, se obtiene que:

```
require(agricolae)
dt <- duncan.test(anv, "factor")
dt$groups
```

```
##          y groups
## 1 69.00000      a
## 3 62.83333     ab
## 2 59.16667      b
```

Tratamiento	Número de moscas muertas	Grupos
1	69	a
3	62.8	ab
2	59.2	b

Parece que el tratamiento 1 y 2 se diferencian lo suficientemente bien entre sí, mientras que para el tratamiento 3 no queda muy claro a qué grupo pertenece, por lo que es pertinente hacer una comparación por contrastes que se define a continuación:

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

$$\begin{cases} H_0 : \mu_1 - \mu_3 = 0 \\ H_1 : \mu_1 - \mu_3 \neq 0 \end{cases}$$

```
require(multcomp)
contraste <- rbind("Marca espray 1 vs Marca espray 2"= c(1,-1,0),
                   "Marca espray 1 vs Marca espray 3"= c(1,0,-1))
columnas <- c("1","2","3")
filas <- c("Marca espray 1 vs Marca espray 2","marca espray 1 vs Marca espray 3")
dimnames(contraste) <- list(filas,columnas)
```

```
compar<-glht(anv,linfct = mcp(factor= contraste))
```

	Estimate Std.	Error	t value	V_p
Marca espray 1 vs Marca espray 2 == 0	9.833	4.205	2.339	0.0605
Marca espray 1 vs Marca espray 3 == 0	6.167	4.205	1.467	0.2729

Siguiendo el criterio donde si $|T_{cal}| > t_{\frac{\alpha}{2}}(N-a)$ se rechaza la hipótesis nula usando un nivel de significancia de α= 0.05, para la comparación de la marca 1 y la marca 2 se tiene que: $|2.339| > 2.13145$ por lo que se rechaza H_0, y se dice que la media del tratamiento 1 es diferente de la media del tratamiento 2.

Para la comparación de la marca espray 1 y la marca espray 3, se obtiene que: $|1.467| < 2.13145$ en este caso *no* se rechaza la hipótesis nula, y se dice que la media de los tratamientos 1 y 3 son iguales.

Con estos resultados y teniendo en cuenta los resultados arrojados en el test de Duncan se podría decir que existen dos grupos: el primero que contiene solamente al tratamiento 2, es decir la marca de espray 2, y un segundo que contiene a los tratamientos 1 y 3, es decir las marcas de espray 1 y 3.