

Taller Entregable parte 2

Oscar Stiven Muñoz Ramirez

2024-05-14

1. Jaime un nuevo empleado de una compañía desea seleccionar la ruta de autobús que lo lleve más rápido a su trabajo. Desde su casa al trabajo hay cuatro rutas que le sirven a Jaime. Para tomar con mayor seguridad su decisión, evaluó cinco veces el recorrido en cada ruta, y registró el tiempo en minutos que demora cada una desde su casa hasta su trabajo. Los resultados de las observaciones fueron:

```
#Pasar a un data frame
Problema1 <- data.frame(Rutas = c(rep("Ruta 1",5),rep("Ruta 2",5),rep("Ruta 3",5),rep("Ruta 4",5)),
                        Tiempo = c(18,21,20,22,19,28,25,23,26,24,20,24,23,25,22,26,34,30,35,30))
Problema1
```

##	Rutas	Tiempo
## 1	Ruta 1	18
## 2	Ruta 1	21
## 3	Ruta 1	20
## 4	Ruta 1	22
## 5	Ruta 1	19
## 6	Ruta 2	28
## 7	Ruta 2	25
## 8	Ruta 2	23
## 9	Ruta 2	26
## 10	Ruta 2	24
## 11	Ruta 3	20
## 12	Ruta 3	24
## 13	Ruta 3	23
## 14	Ruta 3	25
## 15	Ruta 3	22
## 16	Ruta 4	26
## 17	Ruta 4	34
## 18	Ruta 4	30
## 19	Ruta 4	35
## 20	Ruta 4	30

a. Identifique: el factor de estudio y la variable de respuesta. Explique porque los considera así.

- Factor de estudio: El factor de estudio es la variable que el investigador manipula o controla en un experimento para estudiar su efecto sobre otra variable, es decir, la variable de respuesta. En este

caso, Jaime está interesado en determinar cuál de las cuatro rutas de autobús lo llevará más rápido a su trabajo. Las rutas de autobús son el factor que Jaime está estudiando y analizando para tomar su decisión

- Variable de respuesta: La variable de respuesta es la variable que se observa y mide en respuesta al factor de estudio. En este caso, Jaime está registrando el tiempo en minutos que le toma recorrer cada una de las cuatro rutas de autobús desde su casa hasta su trabajo. Por lo tanto, el tiempo de recorrido es la variable que responde a la elección de la ruta de autobús y es la variable de respuesta en este contexto.

b. A un nivel de significancia del 5%, se puede asegurar que hay diferencia en el tiempo promedio de recorrido entre las rutas? Plantee, desarrolle e interprete adecuadamente la hipótesis.

- H0: No hay diferencia en el tiempo promedio de recorrido entre las rutas
- H1: Hay diferencia en el tiempo promedio de recorrido entre las rutas

Se realiza un análisis de varianza para determinar si hay diferencias en el tiempo promedio de recorrido entre las rutas.

```
modelo1 <- aov(Tiempo ~ Rutas, data = Problema1)
summary(modelo1)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Rutas      3  328.2  109.38    19.11 1.54e-05 ***
## Residuals 16   91.6    5.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Datos obtenidos:

- Valor de P: 1.54e-05
- Valor de F: 19.11

Interpretación:

Dado que el valor de P es menor que el nivel de significancia de 0.05, se rechaza la hipótesis nula de que no hay diferencia en el tiempo promedio de recorrido entre las rutas. Por lo tanto, se concluye que hay diferencias significativas en el tiempo promedio de recorrido entre las rutas de autobús.

c. Cual o cuales rutas de bus considera que debe tomar Jaime para desplazarse a su lugar de trabajo? Plantee, desarrolle e interprete adecuadamente las hipótesis.

Para determinar cuál o cuáles rutas de autobús Jaime debe tomar para desplazarse a su lugar de trabajo, se realizará un análisis de comparaciones múltiples utilizando la prueba de Tukey.

- H0: No hay diferencia en el tiempo promedio de recorrido entre las rutas
- H1: Hay diferencia en el tiempo promedio de recorrido entre las rutas

```
TukeyHSD(modelo1)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = Tiempo ~ Rutas, data = Problema1)
##
## $Rutas
##          diff          lwr          upr          p adj
## Ruta 2-Ruta 1  5.2  0.8704914  9.529509  0.0160811
## Ruta 3-Ruta 1  2.8 -1.5295086  7.129509  0.2873235
## Ruta 4-Ruta 1 11.0  6.6704914 15.329509  0.0000102
## Ruta 3-Ruta 2 -2.4 -6.7295086  1.929509  0.4136046
## Ruta 4-Ruta 2  5.8  1.4704914 10.129509  0.0072015
## Ruta 4-Ruta 3  8.2  3.8704914 12.529509  0.0002999
```

Interpretación:

- $\mu_2 > \mu_1$

- Ruta 2: La diferencia en el tiempo promedio de recorrido entre la Ruta 2 y la Ruta 1 es de 5.2 minutos, con un intervalo de confianza del 95% entre 0.87 y 9.53 minutos. El valor de P es 0.016, lo que indica que hay una diferencia significativa en el tiempo promedio de recorrido entre estas dos rutas. Por lo tanto, se recomienda que Jaime tome la Ruta 2 en lugar de la Ruta 1 para desplazarse a su lugar de trabajo.

- $\mu_3 = \mu_1$

- Ruta 3: La diferencia en el tiempo promedio de recorrido entre la Ruta 3 y la Ruta 1 es de 2.8 minutos, con un intervalo de confianza del 95% entre -1.53 y 7.13 minutos. El valor de P es 0.287, lo que indica que no hay una diferencia significativa en el tiempo promedio de recorrido entre estas dos rutas. Por lo tanto, no se puede concluir que Jaime deba tomar la Ruta 3 en lugar de la Ruta 1 para desplazarse a su lugar de trabajo.

- $\mu_4 > \mu_1$

- Ruta 4: La diferencia en el tiempo promedio de recorrido entre la Ruta 4 y la Ruta 1 es de 11 minutos, con un intervalo de confianza del 95% entre 6.67 y 15.33 minutos. El valor de P es 0.00001, lo que indica que hay una diferencia significativa en el tiempo promedio de recorrido entre estas dos rutas. Por lo tanto, se recomienda que Jaime tome la Ruta 4 en lugar de la Ruta 1 para desplazarse a su lugar de trabajo.

- $\mu_3 = \mu_2$

- Ruta 3 vs. Ruta 2: La diferencia en el tiempo promedio de recorrido entre la Ruta 3 y la Ruta 2 es de -2.4 minutos, con un intervalo de confianza del 95% entre -6.73 y 1.93 minutos. El valor de P es 0.414, lo que indica que no hay una diferencia significativa en el tiempo promedio de recorrido entre estas dos rutas. Por lo tanto, no se puede concluir que Jaime deba tomar la Ruta 3 en lugar de la Ruta 2 para desplazarse a su lugar de trabajo.

- $\mu_4 > \mu_2$

- Ruta 4 vs. Ruta 2: La diferencia en el tiempo promedio de recorrido entre la Ruta 4 y la Ruta 2 es de 5.8 minutos, con un intervalo de confianza del 95% entre 1.47 y 10.13 minutos. El valor de P es 0.007, lo que indica que hay una diferencia significativa en el tiempo promedio de recorrido entre estas dos rutas. Por lo tanto, se recomienda que Jaime tome la Ruta 4 en lugar de la Ruta 2 para desplazarse a su lugar de trabajo.

- $\mu_4 > \mu_3$

- Ruta 4 vs. Ruta 3: La diferencia en el tiempo promedio de recorrido entre la Ruta 4 y la Ruta 3 es de 8.2 minutos, con un intervalo de confianza del 95% entre 3.87 y 12.53 minutos. El valor de P es 0.0003, lo que indica que hay una diferencia significativa en el tiempo promedio de recorrido entre estas dos rutas. Por lo tanto, se recomienda que Jaime tome la Ruta 4 en lugar de la Ruta 3 para desplazarse a su lugar de trabajo.

En resumen, se recomienda que Jaime tome la Ruta 2 o la Ruta 4 en lugar de la Ruta 1 para desplazarse a su lugar de trabajo, ya que hay diferencias significativas en el tiempo promedio de recorrido entre estas rutas. Además, se recomienda que Jaime tome la Ruta 4 en lugar de la Ruta 3 para desplazarse a su lugar de trabajo, ya que también hay diferencias significativas en el tiempo promedio de recorrido entre estas rutas.

d. Desarrolle las pruebas de supuestos del modelo. Plantee, desarrolle e interprete adecuadamente las hipótesis.

- Variable de resultado cualitativo: En este caso, la variable de resultado es el tiempo en minutos que Jaime demora en recorrer cada una de las cuatro rutas de autobús desde su casa hasta su trabajo. Esta variable es cuantitativa y continua.

Distribución normal de los residuos

- H_0 : Los residuos del modelo siguen una distribución normal

```
shapiro.test(residuals(modelo1))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(modelo1)
## W = 0.98309, p-value = 0.9675

#Kolmogorov-Smornov
ks.test(residuals(modelo1), "pnorm", mean=mean(residuals(modelo1)), sd=sd(residuals(modelo1)))

## Warning in ks.test.default(residuals(modelo1), "pnorm", mean =
## mean(residuals(modelo1)), : ties should not be present for the
## Kolmogorov-Smirnov test

##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  residuals(modelo1)
## D = 0.092352, p-value = 0.9956
## alternative hypothesis: two-sided
```

Datos obtenidos:

1. Prueba de Shapiro-Wilk:

- Valor de P: 0.9675

2. Prueba de Kolmogorov-Smirnov:

- Valor de P: 0.9956

Interpretación:

Dado que el valor de P es mayor que el nivel de significancia de 0.05 en ambas pruebas, no se rechaza la hipótesis nula de que los residuos del modelo siguen una distribución normal. Por lo tanto, se cumple el supuesto de normalidad para el modelo de análisis de varianza.

Homogeneidad de varianzas

- H_0 : Las varianzas de los residuos son iguales para todas las rutas ($\text{var}(e_i) = 0$)
- H_1 : Las varianzas de los residuos no son iguales para todas las rutas ($\text{var}(e_i) \neq 0$)

```
bartlett.test(residuals(modelo1) ~ Problema1$Rutas)
```

```
##  
## Bartlett test of homogeneity of variances  
##  
## data: residuals(modelo1) by Problema1$Rutas  
## Bartlett's K-squared = 3.1932, df = 3, p-value = 0.3628
```

Datos obtenidos:

- Valor de P: $0.3628 > 0.05 \rightarrow$ No Se rechaza H_0

Interpretación:

Dado que el valor de P es mayor que el nivel de significancia de 0.05, no se rechaza la hipótesis nula de que las varianzas de los residuos son iguales para todas las rutas. Por lo tanto, se cumple el supuesto de homogeneidad de varianzas para el modelo de análisis de varianza.

2. Una compañía desea comprar una máquina para realizar un trabajo específico, decide realizar un experimento y evaluar cuatro tipos de máquinas sometiéndolos a un trabajo que simula el trabajo a realizar, de tal manera que cada hora de trabajo representa seis meses de tiempo real. En la tabla siguiente registra el tiempo en que la maquina presenta alguna falla.

```
#Pasar a un data frame
Problema2 <- data.frame(Maquina = c(rep("Maquina 1",5),rep("Maquina 2",6),
                                   rep("Maquina 3",6),rep("Maquina 4",5)),
                        Tiempo = c(6.4,7.8,5.3,7.4,8.4,8.7,7.4,9.4,10.1,9.2,
                                   9.8,11.1,10.3,9.7,10.3,9.2,8.8,9.9,12.8,12.1,
                                   10.8,11.3))

Problema2
```

##	Maquina	Tiempo
## 1	Maquina 1	6.4
## 2	Maquina 1	7.8
## 3	Maquina 1	5.3
## 4	Maquina 1	7.4
## 5	Maquina 1	8.4
## 6	Maquina 2	8.7
## 7	Maquina 2	7.4
## 8	Maquina 2	9.4
## 9	Maquina 2	10.1
## 10	Maquina 2	9.2
## 11	Maquina 2	9.8
## 12	Maquina 3	11.1
## 13	Maquina 3	10.3
## 14	Maquina 3	9.7
## 15	Maquina 3	10.3
## 16	Maquina 3	9.2
## 17	Maquina 3	8.8
## 18	Maquina 4	9.9
## 19	Maquina 4	12.8
## 20	Maquina 4	12.1
## 21	Maquina 4	10.8
## 22	Maquina 4	11.3

A un nivel de significancia del 5%, se puede asegurar que hay diferencia en el tiempo promedio de falla de las maquinas? Plantee las hipótesis a probar.

- $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$
- H_1 : Hay diferencia en el tiempo promedio de falla de las máquinas (al menos una media es diferente)

Se realiza un análisis de varianza para determinar si hay diferencias en el tiempo promedio de falla de las máquinas.

```
modelo2 <- aov(Tiempo ~ Maquina, data = Problema2)
summary(modelo2)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## Maquina        3   49.0  16.335    15.31 3.38e-05 ***
## Residuals     18   19.2   1.067
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Datos obtenidos:

- Valor de P: $3.38e-05 < 0.05 \rightarrow$ Rechazamos H_0

Interpretación:

Dado que el valor de P es menor que el nivel de significancia de 0.05, se rechaza la hipótesis nula de que no hay diferencia en el tiempo promedio de falla de las máquinas. Por lo tanto, se concluye que hay diferencias significativas en el tiempo promedio de falla de las máquinas (al menos una media es diferente).

b. Que maquina le recomienda a la compañía para que realice el trabajo, justifique y explique claramente su respuesta.

Para determinar qué máquina se recomienda a la compañía para realizar el trabajo, se realizará un análisis de comparaciones múltiples utilizando la prueba de Tukey.

```
TukeyHSD(modelo2)
```

```
##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = Tiempo ~ Maquina, data = Problema2)
##
## $Maquina
##              diff          lwr          upr          p adj
## Maquina 2-Maquina 1 2.04   0.2724700 3.807530 0.0205352
## Maquina 3-Maquina 1 2.84   1.0724700 4.607530 0.0013191
## Maquina 4-Maquina 1 4.32   2.4738751 6.166125 0.0000181
## Maquina 3-Maquina 2 0.80  -0.8852738 2.485274 0.5497622
## Maquina 4-Maquina 2 2.28   0.5124700 4.047530 0.0091105
## Maquina 4-Maquina 3 1.48  -0.2875300 3.247530 0.1199678
```

- $H_1 : \mu_2 > \mu_1$

- valor p= 0.0205352
- Maquina 2: La diferencia en el tiempo promedio de falla entre la Maquina 2 y la Maquina 1 es de 2.04 horas, con un intervalo de confianza del 95% entre 0.2724700 y 3.807530 horas. El valor de P es 0.0205, lo que indica que hay una diferencia significativa en el tiempo promedio de falla entre estas dos máquinas. Por lo tanto, se recomienda que la compañía utilice la Maquina 2 para realizar el trabajo en lugar de la Maquina 1.

- $H_2 : \mu_3 > \mu_1$

- valor $p=0.0013191$
- Maquina 3: La diferencia en el tiempo promedio de falla entre la Maquina 3 y la Maquina 1 es de 2.84 horas, con un intervalo de confianza del 95% entre 1.0724700 y 4.607530 horas. El valor de P es 0.0013, lo que indica que hay una diferencia significativa en el tiempo promedio de falla entre estas dos máquinas. Por lo tanto, se recomienda que la compañía utilice la Maquina 3 para realizar el trabajo en lugar de la Maquina 1.

- $H_3 : \mu_4 > \mu_1$

- valor $p= 0.0000181$
- Maquina 4: La diferencia en el tiempo promedio de falla entre la Maquina 4 y la Maquina 1 es de 4.32 horas, con un intervalo de confianza del 95% entre 2.4738751 y 6.166125 horas. El valor de P es 0.0000181, lo que indica que hay una diferencia significativa en el tiempo promedio de falla entre estas dos máquinas. Por lo tanto, se recomienda que la compañía utilice la Maquina 4 para realizar el trabajo en lugar de la Maquina 1.

- $H_4 : \mu_3 = \mu_2$

- valor $p= 0.549762$
- Maquina 3 vs. Maquina 2: La diferencia en el tiempo promedio de falla entre la Maquina 3 y la Maquina 2 es de 0.8 horas, con un intervalo de confianza del 95% entre -0.8852738 y 2.485274 horas. El valor de P es 0.5498, lo que indica que no hay una diferencia significativa en el tiempo promedio de falla entre estas dos máquinas. Por lo tanto, no se puede concluir que la Maquina 3 sea mejor que la Maquina 2 para realizar el trabajo.

- $H_5 : \mu_4 > \mu_2$

- valor $p= 0.0091105$
- Maquina 4 vs. Maquina 2: La diferencia en el tiempo promedio de falla entre la Maquina 4 y la Maquina 2 es de 2.28 horas, con un intervalo de confianza del 95% entre 0.5124700 y 4.047530 horas. El valor de P es 0.0091, lo que indica que hay una diferencia significativa en el tiempo promedio de falla entre estas dos máquinas. Por lo tanto, se recomienda que la compañía utilice la Maquina 4 para realizar el trabajo en lugar de la Maquina 2.

- $H_6 : \mu_4 = \mu_3$

- valor $p= 0.1199678$
- Maquina 4 vs. Maquina 3: La diferencia en el tiempo promedio de falla entre la Maquina 4 y la Maquina 3 es de 1.48 horas, con un intervalo de confianza del 95% entre -0.2875300 y 3.247530 horas. El valor de P es 0.1199, lo que indica que no hay una diferencia significativa en el tiempo promedio de falla entre estas dos máquinas. Por lo tanto, no se puede concluir que la Maquina 4 sea mejor que la Maquina 3 para realizar el trabajo.

En resumen, se recomienda que la compañía utilice la Maquina 2, la Maquina 3 o la Maquina 4 para realizar el trabajo en lugar de la Maquina 1, ya que hay diferencias significativas en el tiempo promedio de falla entre estas máquinas. Además, se recomienda que la compañía utilice la Maquina 4 en lugar de la Maquina 2 o la Maquina 3 para realizar el trabajo, ya que también hay diferencias significativas en el tiempo promedio de falla entre estas máquinas.