### Universidad de Granada

### Escuela Internacional de Posgrado

### Máster en Estadística Aplicada

### Materia: Diseño estadístico experimental y control de calidad. Aplicaciones en Biociencias e ingeniería.

### Alumno: Francisco Javier Marquez Rosales

# **ACTIVIDADES DE RECUPERACION:**

# **Tema 2. Análisis de la varianza de una vía**

# **Tema 3. Regresión lineal**

# **Tema 4. Diseño por bloques aleatorizado**

Septiembre, 2022

# **Tema 2. Análisis de la varianza de una vía**

## Planteamiento Actividad 2.1

Se desea comprobar si ciertos cambios en un proceso de fabricación aumentan la calidad del producto. Para ello se comparan los resultados con el método tradicional (A) frente a los obtenidos por los procedimientos que se desean probar (B y C). Los datos corresponden a una medida de calidad del proceso.

**Table

Description automatically generated**

### Tarea

Comprobar si existen diferencias entre los tres tratamientos. En caso de existir diferencias entre los tratamientos, determinar de cuál de ellos proviene. Estudiar la validez del modelo, es decir, que los residuos sean normales e independientes y la varianza constante.

### Solución

Para realizar el análisis se utilizará el software R. En primer lugar, examinamos visualmente la distribución de los datos en los tres métodos de fabricación basados en un gráfico de cajas.

Sintaxis:

### construcción de los datos

Resp21<- c(32,44,31,35,33,33,40,46,33,29,35,32,37,30,28,33,37,39)

Trat21<-c("A","A","A","A","A","A","B","B","B","B","B","B","C","C","C","C","C","C")

## el data frame

dat\_21<-data.frame(Resp21,Trat21)

trat\_21<-factor(Trat21)

#Determinacion de los factores

trat21f<-factor(trat\_21)

#analisis los datos

require(ggplot2)

ggplot(data = dat\_21, aes(x = Trat21, y = Resp21, color = Trat21)) +

geom\_boxplot() +

theme\_bw()

### Resultado

Chart, box and whisker chart

Description automatically generated

El gráfico nos muestra que la dispersión de los datos para el método de fabricación A es menor que para B y C al igual que, de forma leve, el valor del promedio de la medida de calidad. Lo anterior indica inicialmente que el método de fabricación A ofrece resultados más estables (menos dispersos) y en donde la media de calidad es un poco menor.

### Validación de los supuestos

### Supuesto 1: Independencia

Por la forma como está descrita la recolección de los datos en el ejercicio, asumimos que fueron recolectados en forma aleatoria.

### Supuesto 2: Distribucion normal

Debemos comprobar si la distribución que siguen los datos de fabricación de los tres métodos se pudiera considerar como ‘Distribución Normal’. Para ello generamos los respectivos gráficos QQ y la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es *H0: los datos siguen una distribución normal*.

En este caso la prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

par(mfrow = c(1,3))

qqnorm(dat\_21[dat\_21$Trat21 == "A","Resp21"], main = "A")

qqline(dat\_21[dat\_21$Trat21 == "A","Resp21"])

qqnorm(dat\_21[dat\_21$Trat21 == "B","Resp21"], main = "B")

qqline(dat\_21[dat\_21$Trat21 == "B","Resp21"])

qqnorm(dat\_21[dat\_21$Trat21 == "C","Resp21"], main = "C")

qqline(dat\_21[dat\_21$Trat21 == "C","Resp21"])

par(mfrow = c(1,1))

#test de normalidad (menos de 50 observaciones usamos el test Shapiro - Wilk)

hist(dat\_21$Resp21)

shapiro.test(dat\_21$Resp21)

### Resultado

Chart

Description automatically generated

Shapiro-Wilk normality test

data: dat\_21$Resp21

W = 0.92699, p-value = 0.1719

El gráfico nos muestra como la nube de puntos se ajusta a lo largo de la recta del modelo normal teórico, lo cual sugiere normalidad. El resultado del test Shapiro-Wilk ofrece un p-value del 0.17 que al hacer la prueba al 95% de confianza no tenemos evidencia para rechazar la hipótesis nula, por lo que aceptamos que los datos tienen una distribución normal.

### Supuesto 3: homocedasticidad o varianza constante entre grupos

Utilizaremos la prueba Barlett para evaluar la homocedasticidad (homogeneidad de varianza). Esta prueba no mantiene sensibilidad frente al supuesto de normalidad que acabamos de comprobar. La hipótesis nula de esta prueba es *H0: los datos presentan homogeneidad de varianza entre los grupos*.

#supuesto 3: homocedasticidad o varianza constante entre grupos

bartlett.test(Resp21~Trat21,dat\_21)

### Resultado

Bartlett test of homogeneity of variances

data: Resp21 by Trat21

Bartlett's K-squared = 0.61404, df = 2, p-value = 0.7356

El resultado del p-valor indica que no hay evidencias significativas de falta de homocedasticidad. De esta forma hemos comprobado los supuestos necesarios para ejecutar el ANOVA.

### Análisis de Varianza

Para comprobar si existen diferencias entre los tres tratamientos. Ahora aplicaremos el ANOVA. Este contrate plantea como hipótesis nula *Ho: las medias de los tratamientos son iguales*.

modelo21 <- lm(Resp21~trat21f)

ANOVA21 <- aov(modelo21)

RESUMEN\_ANOVA21 <- summary(ANOVA21)

RESUMEN\_ANOVA21

Resultado

Df Sum Sq Mean Sq F value Pr(>F)

trat21f 2 10.3 5.167 0.194 0.826

Residuals 15 400.2 26.678

Con base en este resultado, no tenemos suficiente evidencia para rechazar la hipótesis nula, por tanto concluimos que no existen diferencias significativas en los resultados de la calidad de los tres métodos de producción. Dado este resultado, no es válido estudiar el efecto particular de los tratamientos.

### Validez del modelo

Estudiamos la validez del modelo analizando los residuos. Evaluamos que los mismos sean normales e independientes y la varianza constante. Para ello usamos la opción ‘*plot*’ de R sobre el modelo obtenido con la siguiente sintaxis:

par(mfrow = c(2,2))

plot(modelo21)

par(mfrow = c(1,1))

Chart, diagram

Description automatically generated

La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residual vs fitted) y en el Gráfico NormalQ-Q (qqplot) los residuos se distribuyen muy cercanos a la linea de la normal, lo que indica un comportamiento normal. Estos valores nos permiten confirmar la validez del modelo.

## Planteamiento Actividad 2.2

Se comparan las emisiones de distintas empresas de C02, para ello se miden las emisiones a la atmósfera de cuatro empresas, obteniendo los siguientes resultados:

# **Table Description automatically generated**

### Tarea

Existen diferencias entre las emisiones de las diferentes empresas. En caso de existir, ¿de dónde proceden? Estudiar la validez del modelo.

### Solución

Para la solución de este problema usaremos el software R. Iniciamos haciendo la carga de los datos y la identificación de los factores que influyen en el análisis.

### construccion de los datos

Resp22<- c(32,45,68,29,41,37,78,54,23,33,35,11,30,23,41,42,31,37,22,45,61,34,38,39,28,29)

Trat22<-c("Em1","Em1","Em1","Em1","Em1","Em1","Em1","Em2","Em2","Em2","Em2","Em2","Em3",

"Em3","Em3","Em3","Em3","Em3","Em3","Em3","Em4","Em4","Em4","Em4","Em4","Em4")

## el data frame

dat\_22<-data.frame(Resp22,Trat22)

trat\_22<-factor(Trat22)

#Determinacion de los factores

trat22f<-factor(trat\_22)

Estudiamos la distribución de los datos, sus valores medios y su dispersión los cual nos permitirá saber si el análisis planteado es razonable. Para ello utilizamos el gráfico de cajas

#analisis los datos

require(ggplot2)

ggplot(data = dat\_22, aes(x = Trat22, y = Resp22, color = Trat22)) +

geom\_boxplot() +

theme\_bw()

### Resultado

Chart, box and whisker chart

Description automatically generated

Examinando el gráfico podemos observar inicialmente como la emisión de CO2 de la empresa 1 tiene una dispersión mayor con relación a las otras empresas, por ello tendremos que comprobar el supuesto de homocedasticidad. La emisión de Co2 producida por la empresa 2 presenta una ligera asimetría. De igual forma la media de la empresa 1 parece tener una diferencia con relación a las medias de las otras empresas.

Procedemos entonces a validar los supuestos.

### Validación de los supuestos

### Supuesto 1: Independencia

Por la forma como está descrita la recolección de los datos en el ejercicio, asumimos que fueron recolectados en forma aleatoria.

### Supuesto 2: Distribución normal

Debemos comprobar si la distribución de las emisiones de CO2 de las 4 empresas se pueden considerar como ‘Distribución Normal’. Para ello generamos los respectivos gráficos QQ y la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es *H0: los datos siguen una distribución normal*.

En este caso la prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

#supuesto 2: distribucion normal

par(mar=c(1,1,1,1)) #cambio de margenes para que ajusten los 4 graficos

par(mfrow = c(2,2))

qqnorm(dat\_22[dat\_22$Trat22 == "Em1","Resp22"], main = "Em1")

qqline(dat\_22[dat\_22$Trat22 == "Em1","Resp22"])

qqnorm(dat\_22[dat\_22$Trat22 == "Em2","Resp22"], main = "Em2")

qqline(dat\_22[dat\_22$Trat22 == "Em2","Resp22"])

qqnorm(dat\_22[dat\_22$Trat22 == "Em3","Resp22"], main = "Em3")

qqline(dat\_22[dat\_22$Trat22 == "Em3","Resp22"])

qqnorm(dat\_22[dat\_22$Trat22 == "Em4","Resp22"], main = "Em4")

qqline(dat\_22[dat\_22$Trat22 == "Em4","Resp22"])

par(mfrow = c(1,1))

#test de normalidad (menos de 50 observaciones usamos el test Shapiro - Wilk)

shapiro.test(dat\_22$Resp22)

### Resultado

Chart, scatter chart

Description automatically generated

Shapiro-Wilk normality test

data: dat\_22$Resp22

W = 0.9223, p-value = 0.05092

Los gráficos nos muestran como las nubes de puntos se ajustan a lo largo de la recta del modelo normal teórico, lo cual sugiere normalidad. El resultado del test Shapiro-Wilk ofrece un p-value del 0.05092 que indica que al hacer la prueba al 95% de confianza no tenemos evidencia para rechazar la hipótesis nula, por lo que aceptamos que los datos tienen una distribución normal. Este resultado está al borde del criterio por que hay que tomar en cuenta la sensibilidad de los datos al considerarlos normal.

### Supuesto 3: homocedasticidad o varianza constante entre grupos

Dado que se encuentra en el límite para aceptar que se distribuye de forma normal, la prueba de Fisher y la de Bartlett no son recomendables. En su lugar es mejor emplear una prueba basada en la mediana, por lo que emplearemos la prueba de Levene y la prueba de Fligner-Killeen. En ambas pruebas contrastamos la hipótesis nula *Ho: la varianza entre grupos es constante*.

require(car)

leveneTest(Resp22~Trat22,dat\_22,center = "median")

fligner.test(Resp22~Trat22,dat\_22)

### Resultado

Levene's Test for Homogeneity of Variance (center = "median")

Df F value Pr(>F)

group 3 0.5875 0.6296

Fligner-Killeen test of homogeneity of variances

data: Resp22 by Trat22

Fligner-Killeen:med chi-squared = 1.7094, df = 3, p-value = 0.6349

Como resultado en ambas pruebas obtenemos valores p-value mayores a 0.05, esto indica que no tenemos evidencia para rechazar la hipótesis nula, por lo que concluimos que se cumple el criterio de homocedasticidad o igualdad de la varianza entre empresas.

### Análisis de Varianza

Para comprobar si existen diferencias entre los tres tratamientos. Ahora aplicaremos el ANOVA. Este contrate plantea como hipótesis nula *Ho: las medias de las emisiones de CO2 de las empresas son iguales*.

modelo22 <- lm(Resp22~trat22f)

ANOVA22 <- aov(modelo22)

RESUMEN\_ANOVA22 <- summary(ANOVA22)

RESUMEN\_ANOVA22

### Resultado

Df Sum Sq Mean Sq F value Pr(>F)

trat22f 3 952 317.5 1.601 0.218

Residuals 22 4363 198.3

Con base en este resultado, no tenemos suficiente evidencia para rechazar la hipótesis nula, por tanto concluimos que no existen diferencias significativas en los resultados de emisión de Co2 de las cuatro empresas. Dado este resultado, no resulta útil estudiar el efecto individual de la emisión de CO2 de las empresas.

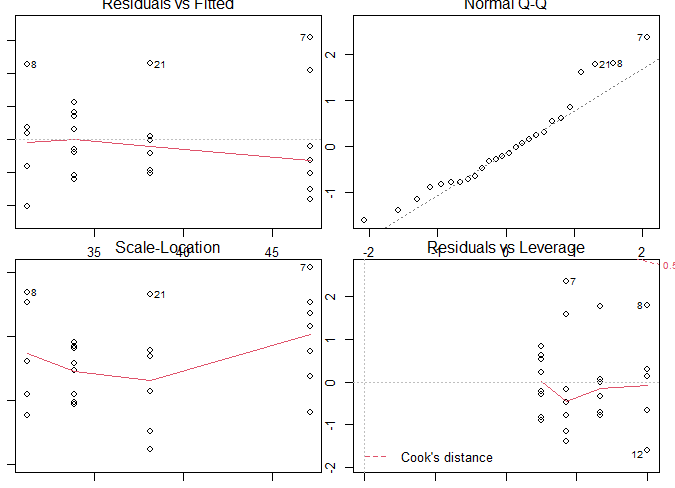
### Validez del modelo

Estudiamos la validez del modelo analizando los residuos. Evaluamos que los mismos sean normales e independientes y la varianza constante. Para ello usamos la opción ‘*plot*’ de R sobre el modelo obtenido con la siguiente sintaxis:

par(mfrow = c(2,2))

plot(modelo21)

par(mfrow = c(1,1))



La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residual vs fitted) y en el Gráfico NormalQ-Q (qqplot) los residuos se distribuyen muy cercanos a la linea de la normal, lo que indica un comportamiento normal. Estos valores nos permiten confirmar la validez del modelo.

## Planteamiento Actividad 2.3

Una empresa produce telas en diferentes telares. Les gustaría que los telares fuesen lo más homogéneos posibles para obtener tejidos de la misma calidad, aunque el ingeniero sospecha que por la variación en la fuerza puede producir diferencias significativas entre las telas producidas en diferentes telares. Para investigar esto, se selecciona al azar, cuatro telares y hace cuatro determinaciones fuerza en el tejido fabricado en cada telar. Obteniendo la siguiente tabla:

A screenshot of a calculator

Description automatically generated with low confidence

### Tarea

¿Está en lo cierto el investigador? Comprobar el modelo y determinar el origen de las posibles diferencias.

### Solución

Para realizar el análisis se utilizará el software R. En primer lugar, examinamos visualmente la distribución de los datos en los tres métodos de fabricación basados en un gráfico de cajas.

Sintaxis:

### construcción de los datos

### construccion de los datos

Resp23<- c(98,97,99,96,91,90,93,92,96,95,97,95,95,96,99,98)

Trat23<-c("T1","T1","T1","T1","T2","T2","T2","T2","T3","T3","T3","T3","T4","T4","T4","T4")

## el data frame.

dat\_23<-data.frame(Resp23,Trat23)

trat\_23<-factor(Trat23)

#Determinacion de los factores

trat23f<-factor(trat\_23)

#analisis visual de los datos

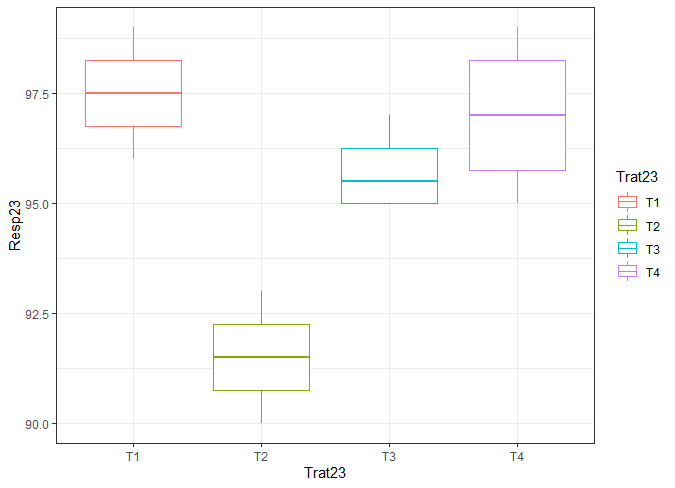
require(ggplot2)

ggplot(data = dat\_23, aes(x = Trat23, y = Resp23, color = Trat23)) +

geom\_boxplot() +

theme\_bw()

Resultado



El gráfico nos muestra que los valores medios de medida de tensión de los cuatro telares es diferente, en especial los telares 1 y 2. La dispersión de los datos para los cuales telares puede considerarse parecida. En ninguno de los casos se observa asimetría.

### Validación de los supuestos

### Supuesto 1: Independencia

Dado que los telares fuero seleccionados al azar para el regitro de las meiciones de fuerza, podemos considerar que los datos presentan la característica de independencia.

### Supuesto 2: Distribución normal

Debemos comprobar si la medida de fuerza de los 4 telares empresas se pueden considerar que tienen una ‘Distribución Normal’. Para ello generamos los respectivos gráficos QQ y la prueba Shapiro-Wilk, con un nivel de confianza del 95%, esta prueba es un contraste de hipótesis en donde la hipótesis nula es *H0: los datos siguen una distribución normal*.

En este caso la prueba Shapiro-Wilk es la recomendada por tener menos de 50 observaciones.

#supuesto 2: distribucion normal

par(mar=c(1,1,1,1)) #cambio de márgenes para que ajusten los 4 gráficos

par(mfrow = c(2,2))

qqnorm(dat\_23[dat\_23$Trat23 == "T1","Resp23"], main = "T1")

qqline(dat\_23[dat\_23$Trat23 == "T1","Resp23"])

qqnorm(dat\_23[dat\_23$Trat23 == "T2","Resp23"], main = "T2")

qqline(dat\_23[dat\_23$Trat23 == "T2","Resp23"])

qqnorm(dat\_23[dat\_23$Trat23 == "T3","Resp23"], main = "T3")

qqline(dat\_23[dat\_23$Trat23 == "T3","Resp23"])

qqnorm(dat\_23[dat\_23$Trat23 == "T4","Resp23"], main = "T4")

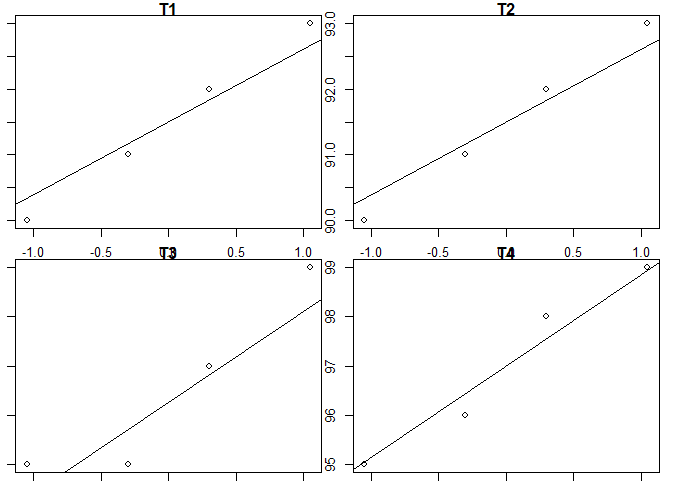
qqline(dat\_23[dat\_23$Trat23 == "T4","Resp23"])

par(mfrow = c(1,1))

#test de normalidad (menos de 50 observaciones usamos el test Shapiro - Wilk)

shapiro.test(dat\_23$Resp23)

### Resultado



Shapiro-Wilk normality test

data: dat\_23$Resp23

W = 0.93207, p-value = 0.2629

Los gráficos nos muestran como las nubes de puntos se ajustan a lo largo de la recta del modelo normal teórico, lo cual sugiere normalidad. El resultado del test Shapiro-Wilk ofrece un p-value del 0.26 lo que indica al 95% de confianza, que no tenemos evidencia para rechazar la hipótesis nula, por lo que aceptamos que los datos provienen de una distribución normal.

### Supuesto 3: homocedasticidad o varianza constante entre grupos

Utilizaremos la prueba Barlett para evaluar la homocedasticidad (homogeneidad de varianza). Esta prueba no mantiene sensibilidad frente al supuesto de normalidad que acabamos de comprobar. La hipótesis nula de esta prueba es *H0: los datos presentan homogeneidad de varianza entre los grupos*.

bartlett.test(Resp23~Trat23,dat\_23)

### Resultado

Bartlett test of homogeneity of variances

data: Resp23 by Trat23

Bartlett's K-squared = 1.1064, df = 3, p-value = 0.7755

El resultado del p-valor, p-value=0.73, indica que no hay evidencias significativas de falta de homocedasticidad. De esta forma hemos comprobado los supuestos necesarios para ejecutar el ANOVA.

### Análisis de Varianza

Para comprobar si existen diferencias significativas entre la tensión de fuerza de los cuatro telares. Ahora aplicaremos el ANOVA. Este contraste plantea como hipótesis nula *Ho: las medias de la tensión de fuerza de los cuatro telares son iguales*.

Df Sum Sq Mean Sq F value Pr(>F)

trat23f 3 89.19 29.729 15.68 0.000188 \*\*\*

Residuals 12 22.75 1.896

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

### Resultado

Df Sum Sq Mean Sq F value Pr(>F)

trat23f 3 89.19 29.729 15.68 0.000188 \*\*\*

Residuals 12 22.75 1.896

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Con base en los datos analizados y con un 95% de confianza, no podemos aceptar la hipótesis nula, por tanto concluimos que existen diferencias significativas en las medidas de fuerza ejercidas por los cuatro telares. Dado este resultado, procedemos a investigar el origen de las posibles diferencias, para ello utilizaremos prueba de comparación múltiple de Tukey, llamada prueba TukeyHSD.

#comparacion multiple

TukeyHSD(anova23)

plot(TukeyHSD(anova23))

Resultado

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = modelo23)

$trat23f

diff lwr upr p adj

T2-T1 -6.00 -8.890552 -3.109448 0.0002455

T3-T1 -1.75 -4.640552 1.140552 0.3209518

T4-T1 -0.50 -3.390552 2.390552 0.9542581

T3-T2 4.25 1.359448 7.140552 0.0044029

T4-T2 5.50 2.609448 8.390552 0.0005377

T4-T3 1.25 -1.640552 4.140552 0.5894146

Chart, box and whisker chart

Description automatically generated

Con base en los resultados de la prueba y del gráfico obtenido observamos como se tienen diferencias significativas en las siguientes combinaciones de telares:

T1 con T2, T2 con T3 y T2 con T4

Debido a este resultado, podemos afirmar que la diferencia identificada por el ANOVA proviene principalmente de que las medidas del telar 2 (T2), en promedio, fueron significativamente menores al resto de los telares.

### Validez del modelo

Estudiamos la validez del modelo analizando los residuos. Evaluamos que los mismos sean normales e independientes y la varianza constante. Para ello usamos la opción ‘*plot*’ de R sobre el modelo obtenido con la siguiente sintaxis:

par(mfrow = c(2,2))

plot(modelo23)

par(mfrow = c(1,1))

Chart

Description automatically generated

La representación gráfica de los residuos no muestra falta de homocedasticidad (gráfico Residual vs fitted) y en el Gráfico NormalQ-Q (qqplot) los residuos se distribuyen muy cercanos a la linea de la normal, lo que indica un comportamiento normal. Estos valores nos permiten confirmar la validez del modelo.