

Clase 11 Evaluación de supuestos pruebas paramétricas

Diplomado en Análisis de datos con R para la acuicultura.

Dr. José A. Gallardo y Dra. María Angélica Rueda.
jose.gallardo@pucv.cl | Pontificia Universidad Católica de
Valparaíso

22 October 2021

PLAN DE LA CLASE

1.- Introducción

- ▶ Supuestos de los análisis paramétricos.
- ▶ Consecuencias de la violación de los supuestos.
- ▶ Métodos gráficos y análisis de residuos para evaluar supuestos.
- ▶ Pruebas de hipótesis para evaluar supuestos.

2.- Práctica con R y Rstudio cloud

- ▶ Evaluar supuestos de las pruebas paramétricas.
- ▶ Elaborar un reporte dinámico en formato pdf.

SUPUESTOS: INDEPENDENCIA

Independencia

Cada observación de la muestra no debe estar relacionada con otra observación de la muestra.

*Si se viola este supuesto la prueba paramétrica **NO** es válida.*

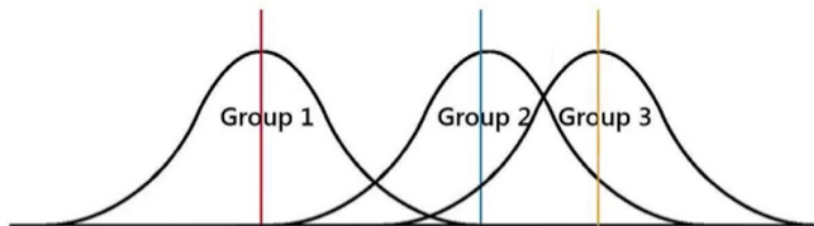
Ejemplo violación del supuesto

- Muestreo de peces de una misma familia.
- Diversidad de especies en una misma muestra de plancton.
- Medidas repetidas en un mismo individuo (antes y después de un tratamiento).

SUPUESTOS: HOMOGENEIDAD DE VARIANZAS

Homocedasticidad

En el caso de comparación de dos o más muestras éstas deben provenir de poblaciones con la misma varianza.



Alguna heterogeneidad es permitida, particularmente con $n > 30$.

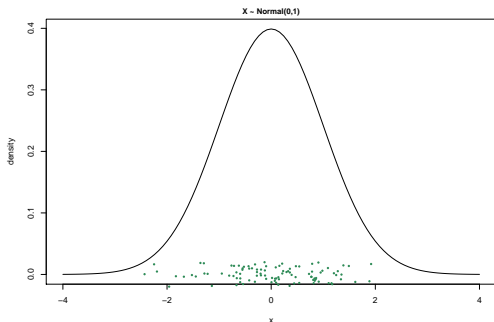
SUPUESTOS: NORMALIDAD

Normalidad

Los datos de muestreo se obtienen de una población que tiene distribución normal.

Ejemplos de violación del supuesto

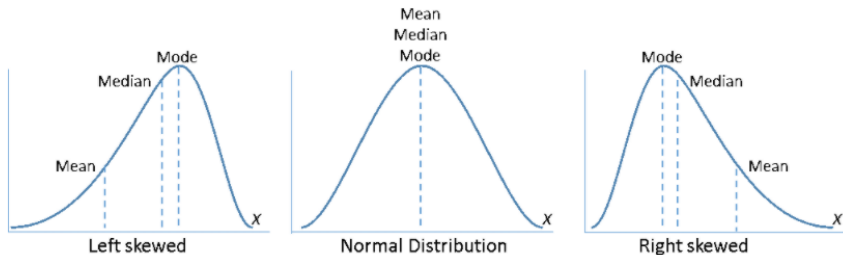
- La distribución no es simétrica.
- La variable no es continua.
- Tiene *límites* a la izquierda o derecha como los porcentajes.



VIOLACIÓN DEL SUPUESTO DE NORMALIDAD

¿Cuál es el problema?

Cambia la probabilidad de rechazar la hipótesis nula.



VIOLACIÓN DEL SUPUESTO DE NORMALIDAD 2

¿Cómo afecta una población no normal a la probabilidad de rechazar?

n	Cola izq.	Cola der.	$\alpha_{Empírica}$
5	0,20	0,26	0,46
10	0,24	0,28	0,52
20	0,23	0,26	0,49
30	0,24	0,27	0,51
50	0,24	0,26	0,50
100	0,24	0,26	0,50

*En la práctica aproximadamente normal es suficiente, particularmente con **$n > 30$** .*

ANÁLISIS DE RESIDUALES

¿Qué son los residuos?

Residuo = valor observado - valor predicho

$$e = y - \hat{y}$$

Residuos en ANOVA

$$\sum_{i=1}^n (y - \hat{y})^2$$

Note que la suma de residuos representa la variabilidad no explicada por el modelo.

¿Para qué sirven?

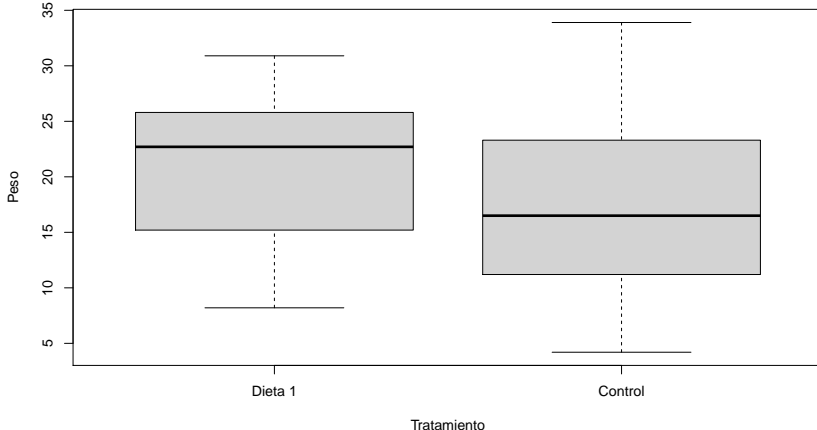
Para someter a prueba los supuestos de muchos análisis paramétricos como **ANOVA**, **ANCOVA** o **REGRESIÓN**.

EVALUACIÓN DE SUPUESTOS

Regla de oro

Primero evalúe independencia, luego homogeneidad de varianzas y finalmente normalidad.

Estudio de caso



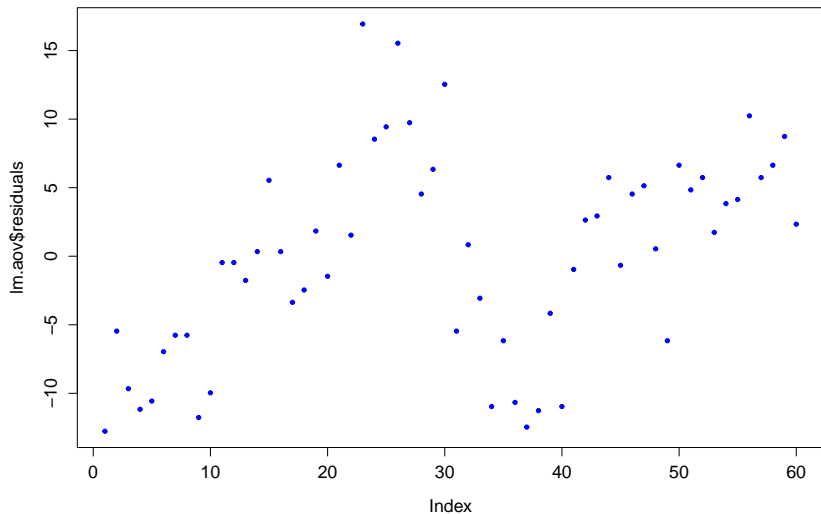
ANOVA EN LA FORMA DE MODELO LINEAL

```
lm.aov <- lm(Peso ~ Tratamiento, data = my_data)
anova(lm.aov) %>% kable(digits = 3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Tratamiento	1	205.350	205.35	3.668	0.06
Residuals	58	3246.859	55.98	NA	NA

INDEPENDENCIA: ANÁLISIS DE RESIDUALES

```
plot(lm.aov$residuals, pch=20, col = "blue",  
     cex.lab=1.25, cex.axis=1.25)
```



INDEPENDENCIA: PRUEBA DE DURBIN-WATSON

Hipótesis

H_0 : Son independientes o no existe autocorrelación.

H_A : No son independientes y existe autocorrelación.

```
dwtest(Peso ~ Tratamiento, data = my_data,  
       alternative = c("two.sided"),  
       iterations = 15) # library(lmtest)
```

```
##
```

```
## Durbin-Watson test
```

```
##
```

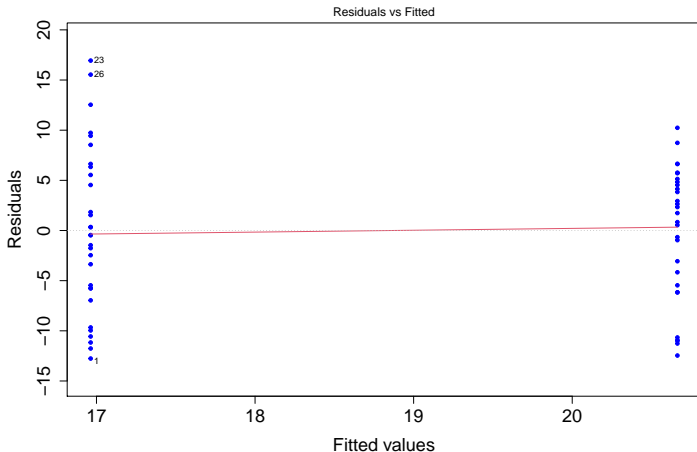
```
## data:  Peso ~ Tratamiento
```

```
## DW = 0.61428, p-value = 1.166e-10
```

```
## alternative hypothesis: true autocorrelation is not 0
```

HOMOGENEIDAD DE VARIANZAS: ANÁLISIS DE RESIDUALES

```
plot(lm.aov, 1, pch=20, col = "blue",  
     cex.lab=1.5, cex.axis=1.5, sub = "")
```



HOMOGENEIDAD DE VARIANZAS: PRUEBA DE LEVENE

$$H_0: \sigma_1^2 = \sigma_2^2$$

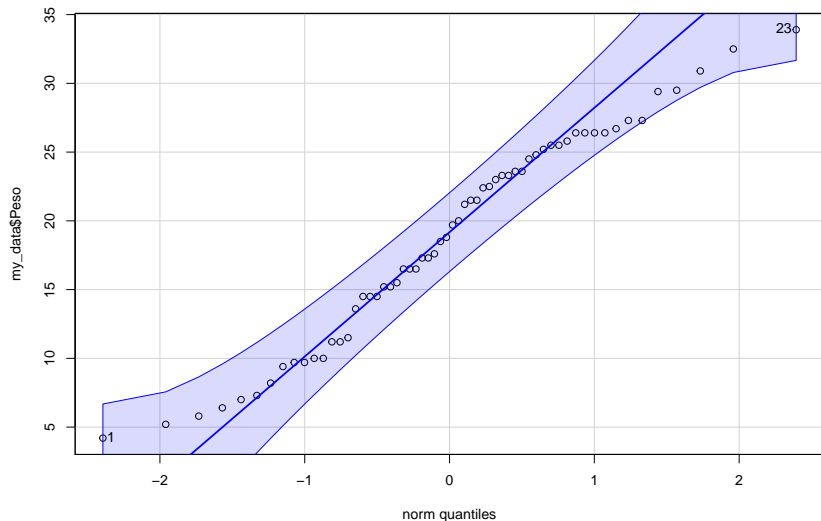
$$H_A: \sigma_1^2 \neq \sigma_2^2$$

```
lv <- leveneTest(Peso ~ Tratamiento, data = my_data,  
                  center = "median") # library(car)  
lv %>% kable(digits = 3)
```

	Df	F value	Pr(>F)
group	1	1.214	0.275
	58	NA	NA

NORMALIDAD: GRÁFICO DE CUANTILES

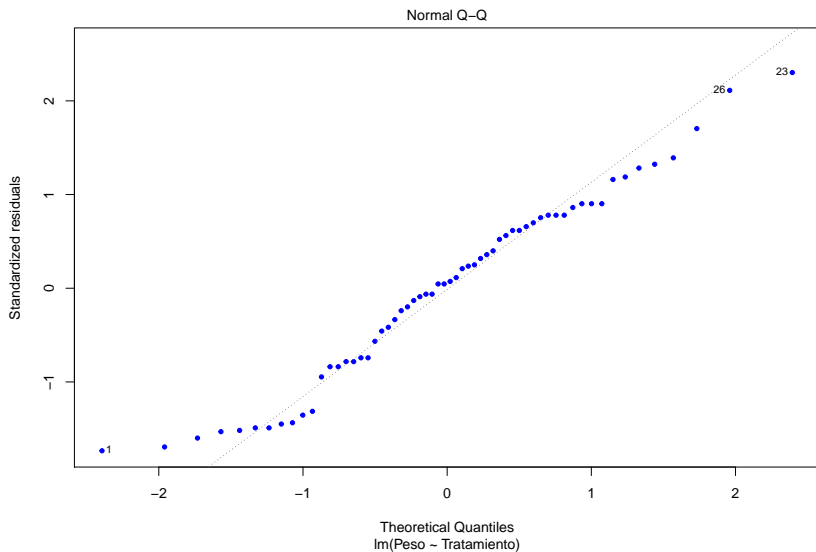
```
qqPlot(my_data$Peso) # library(car)
```



```
## [1] 23 1
```

NORMALIDAD: ANÁLISIS DE RESIDUALES

```
plot(lm.aov, 2, pch=20, col = "blue")
```



NORMALIDAD: PRUEBA DE SHAPIRO-WILKS

H₀: La distribución es normal.

H_A: La distribución no es normal.

```
aov_residuals <- residuals(object = lm.aov)
shapiro.test(x= aov_residuals)
```

```
##
```

```
##  Shapiro-Wilk normality test
```

```
##
```

```
## data:  aov_residuals
```

```
## W = 0.96949, p-value = 0.1378
```

PRÁCTICA ANÁLISIS DE DATOS

- ▶ Guía de trabajo práctico disponible en drive y Rstudio.cloud.

Clase_11

- ▶ El trabajo práctico se realiza en Rstudio.cloud.

Clase 11 - Evaluación de supuestos

RESUMEN DE LA CLASE

- ▶ **Teoría**

- ▶ Supuestos de los análisis paramétricos.
- ▶ Consecuencias de la violación de los supuestos.
- ▶ Interpretación de métodos gráficos, análisis de residuos y pruebas de hipótesis para evaluar supuestos.

- ▶ **Evaluación de supuestos**

- ▶ Independencia.
- ▶ Homocedasticidad.
- ▶ Normalidad.