

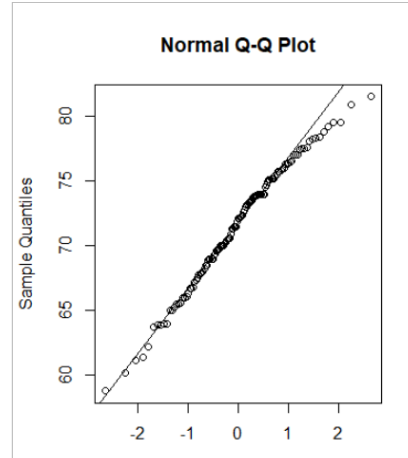
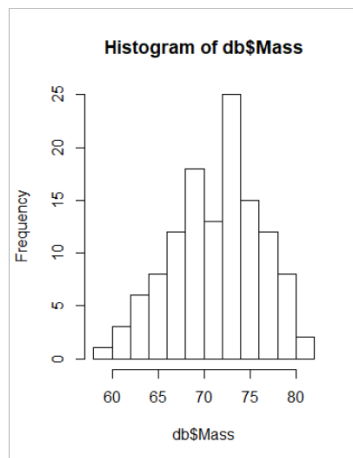
Lunes 28	Martes 29	Miércoles 30	Jueves 31	Viernes 1
	<ul style="list-style-type: none"> • Conceptos básicos • T-test 	<ul style="list-style-type: none"> • One-way ANOVA • Two-way ANOVA 	<ul style="list-style-type: none"> • LM Simples • LM múltiples sin interacción 	
Lunes 4	Martes 5	Miércoles 6	Jueves 7	Viernes 8
		<ul style="list-style-type: none"> • LM múltiples con interacción 	<ul style="list-style-type: none"> • Resolución de práctica • GLMs 	

Repaso de Conceptos básicos y T-test

- **Asunciones:** Todos los análisis estadísticos asumen ciertas características de los datos. Se deben comprobar antes de llevar a cabo el modelo.

Asunciones no cumplidas → Conclusiones erróneas

Normalidad



```
>shapiro.test(db$Mass)
```

Shapiro-wilk normality test

data: db\$Mass
W = 0.98599, p-value = 0.2366

Homogeneidad de varianza (Homocedasticidad)

```
> leveneTest(db$Mass~db$KnownSex)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  1  4.4591 0.03677 *
      121

---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

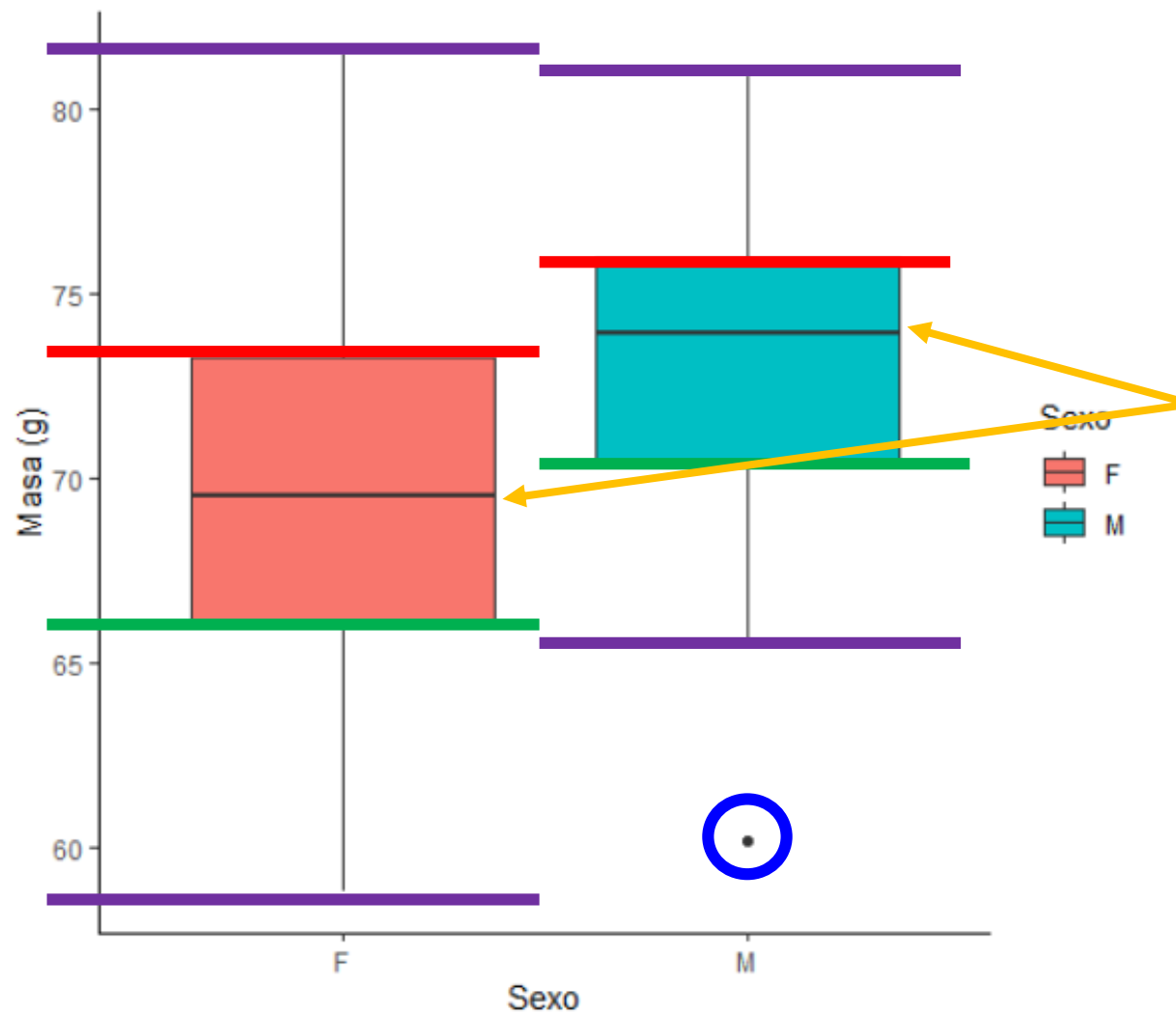
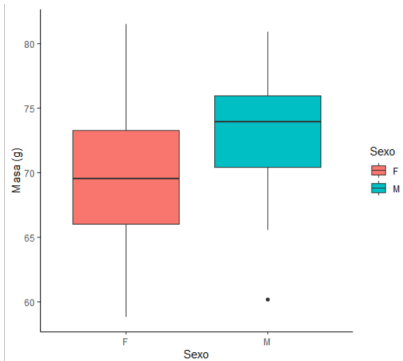
- **Relación entre variables:** Efecto que quieres estudiar

Y ~ X → Y depende de X = X afecta a Y

DEPENDIENTE ~ INDEPENDIENTE

RESPUESTA ~ EXPLICATIVA

- **Boxplot:**



Tercer cuartil (Q3): 75% de los datos es inferior a ese valor

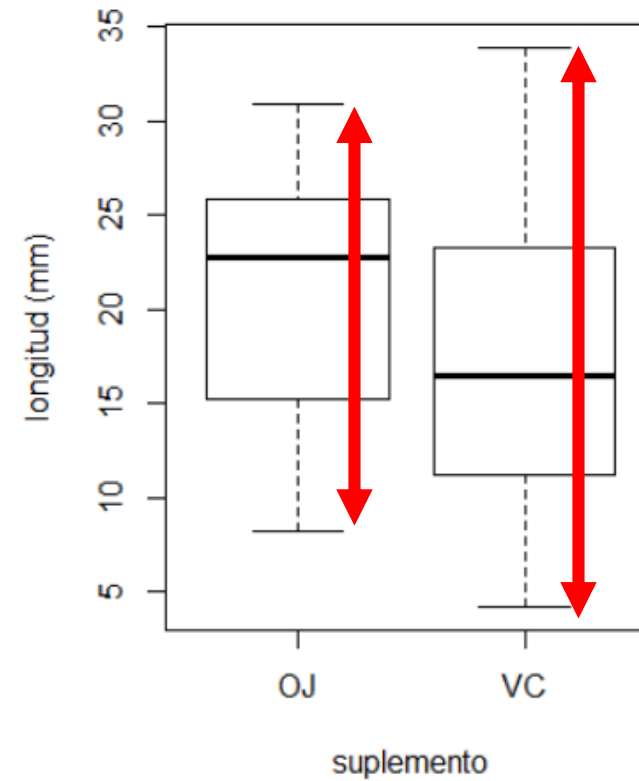
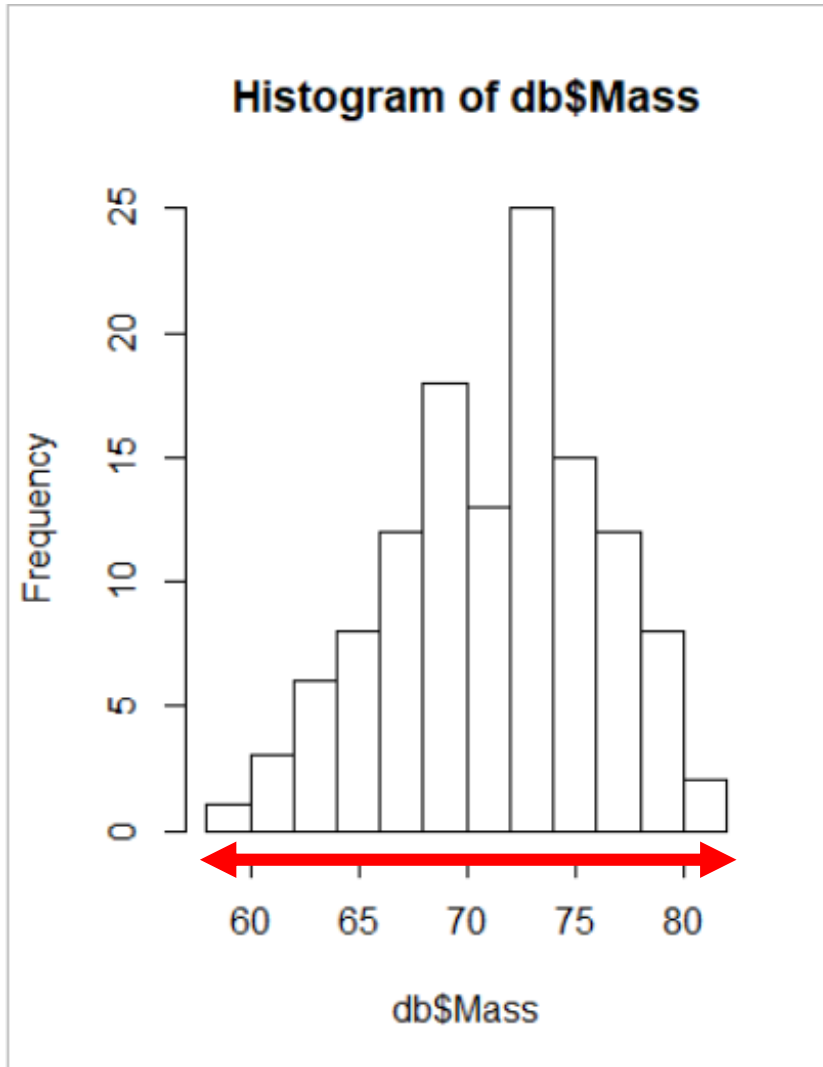
Mediana (Q2): 50% de los datos es inferior a ese valor; valor “en medio” de la base de datos

Primer cuartil (Q1): 25% de los datos es inferior a ese valor

Min. Y Máx: valores “máximo y mínimo”

Outlier: valores atípicos

- **Varianza:** Medida de dispersión de nuestros datos.



- **P-valor:** *“Probabilidad de haber obtenido el resultado obtenido siendo H_0 cierta”*

Indica significancia del modelo:

$P\text{-valor} < 0.05 \rightarrow$ Significativo, aceptamos H_a

$P\text{-valor} > 0.05 \rightarrow$ No significativo, rechazamos H_a

T-test

- Comparar **dos** grupos
 - H0= las medias de los dos grupos son iguales
 - Ha= las medias de los dos grupos son distintas
- `>t.test(Y ~ X)`

Estadística aplicada en R

Análisis de Varianza (ANOVA)



-Marzo 2022-

Carlota Solano

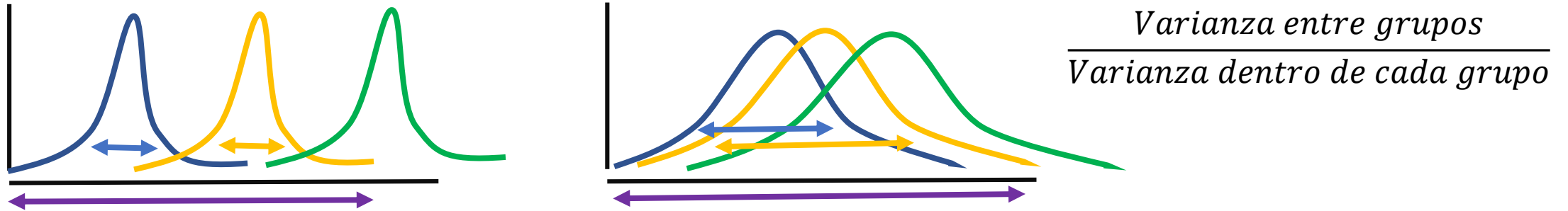
carlota.solano.udina@upm.es



3.1. Análisis de Varianza (One-way ANOVA)

1. ¿Qué es?

Es un análisis factorial de la varianza de grupos, comparando la varianza dentro y entre grupos.



2. ¿Cuándo se puede utilizar?

- Cuando quieres comparar una variable continua entre más de dos niveles.
- Cuando quieres estudiar el efecto de un factor (con tres o más niveles) sobre la media de una variable continua.

3. ¿Qué tipo de datos se necesitan?

Variable respuesta (dep.; y) → Numérica continua

Variables explicativas (indep.; x) → Una variable categórica con tres o más niveles

3.1. Análisis de Varianza (One-way ANOVA)

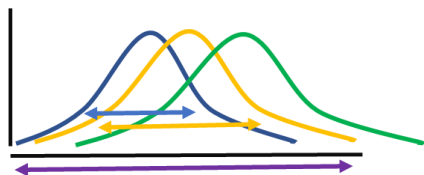
4. ¿Qué asunciones tiene?

- **Independencia** de las observaciones
- Distribución **normal** de la variable a estudiar en cada uno de los grupos.
- Igualdad de **varianza** en los grupos a comparar (Homocedasticidad) → Se asume durante el muestreo, pero puede ser comprobada posteriormente.
- Diseños equilibrados → Mismo número muestral en los distintos grupos.

5. Matemáticamente, ¿cuál es la hipótesis?

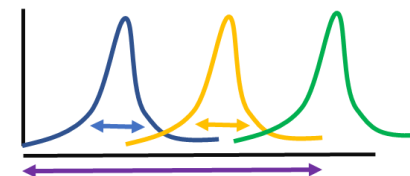
H0: La varianza dentro de los grupos es mayor que entre grupos

La media de los grupos no difiere → $\mu_1 = \mu_2 = \dots = \mu_i$



Ha: La varianza dentro de los grupos es menor que entre grupos

La media de al menos dos de los grupos difiere → $\mu_1 \neq \mu_2 \dots = \mu_i$



6. ¿Cómo se corre en R?

```
> aov(respuesta ~ explicativa) %>% summary() #<- con paquete tidyr
```

ó

R

```
> holi<-aov(...)  
> summary(holi)
```

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

```
> aov(ChickWeight$weight~ChickWeight$Diet)
```

Call:

```
aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

Terms:

	ChickWeight\$Diet	Residuals
Sum of Squares	155862.7	2758693.3
Deg. of Freedom	3	574

Sum of Squares (Suma de cuadrados): para el tratamiento, i.e. la dieta (entre dietas), y para los residuos, i.e. el error (dentro de cada dieta).

Residual standard error: 69.32594

Estimated effects may be unbalanced

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%summary()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ChickWeight\$Diet	3	155863	51954	10.81	6.43e-07 ***
Residuals	574	2758693	4806		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

```
> aov(ChickWeight$weight~ChickWeight$Diet)
```

Call:

```
aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

Terms:

	ChickWeight\$Diet	Residuals
Sum of Squares	155862.7	2758693.3
Deg. of Freedom	3	574

Residual standard error: 69.32594

Estimated effects may be unbalanced

Sum of Squares (Suma de cuadrados): para el tratamiento, i.e. la dieta (entre dietas), y para los residuos, i.e. el error (dentro de cada dieta).

Grados de libertad: *cantidad de valores independientes que tenemos para calcular el modelo.*- Para el tratamiento: n-1 - Para los residuos: k-n

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%summary()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ChickWeight\$Diet	3	155863	51954	10.81	6.43e-07 ***
Residuals	574	2758693	4806		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

```
> aov(ChickWeight$weight~ChickWeight$Diet)
```

Call:

```
aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

Terms:

	ChickWeight\$Diet	Residuals
Sum of Squares	155862.7	2758693.3
Deg. of Freedom	3	574

Sum of Squares (Suma de cuadrados): para el tratamiento, i.e. la dieta (entre dietas), y para los residuos, i.e. el error (dentro de cada dieta).

Residual standard error: 69.32594

Estimated effects may be unbalanced

Grados de libertad: *cantidad de valores independientes que tenemos para calcular el modelo.*- Para el tratamiento: n-1 - Para los residuos: k-n

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%summary()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ChickWeight\$Diet	3	155863	51954	10.81	6.43e-07 ***
Residuals	574	2758693	4806		

F value: $\frac{MS_{inter}}{MS_{intra}} = \frac{51954}{4806} = 10.81$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

```
> aov(ChickWeight$weight~ChickWeight$Diet)
```

Call:

```
aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

Terms:

	ChickWeight\$Diet	Residuals
Sum of Squares	155862.7	2758693.3
Deg. of Freedom	3	574

Sum of Squares (Suma de cuadrados): para el tratamiento, i.e. la dieta (entre dietas), y para los residuos, i.e. el error (dentro de cada dieta).

Residual standard error: 69.32594
Estimated effects may be unbalanced

Grados de libertad: *cantidad de valores independientes que tenemos para calcular el modelo.*- Para el tratamiento: n-1 - Para los residuos: k-n

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%summary()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ChickWeight\$Diet	3	155863	51954	10.81	6.43e-07 ***
Residuals	574	2758693	4806		

F value: $\frac{MS_{inter}}{MS_{intra}} = \frac{51954}{4806} = 10.81$

P-valor y significancia (*)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

```
> aov(ChickWeight$weight~ChickWeight$Diet)
```

Call:

```
aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

Terms:

	ChickWeight\$Diet	Residuals
Sum of Squares	155862.7	2758693.3
Deg. of Freedom	3	574

Sum of Squares (Suma de cuadrados): para el tratamiento, i.e. la dieta (entre dietas), y para los residuos, i.e. el error (dentro de cada dieta).

Residual standard error: 69.32594

Estimated effects may be unbalanced

Grados de libertad: *cantidad de valores independientes que tenemos para calcular el modelo.*- Para el tratamiento: n-1 - Para los residuos: k-n

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%summary()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ChickWeight\$Diet	3	155863	51954	10.81	6.43e-07 ***
Residuals	574	2758693	4806		

F value: $\frac{MS_{inter}}{MS_{intra}} = \frac{51954}{4806} = 10.81$

P-valor y significancia (*)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

La dieta afecta significativamente a la masa de los pollitos, pero... ¿qué dietas son las que más engordan a los pollitos y cuales las que menos? → **Post-Hoc test**

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

La dieta afecta significativamente a la masa de los pollitos, pero... ¿qué dietas son las que más engordan a los pollitos y cuales las que menos? → **Post-Hoc test**

Post-Hoc test: Identifica subconjuntos de medias que no difieren entre sí → *Qué grupos específicos difieren entre sí.*

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%TukeyHSD()  
Tukey multiple comparisons of means  
95% family-wise confidence level
```

```
> anova<-aov(...)  
> TukeyHSD(anova)
```



```
Fit: aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

```
$`ChickWeight$Diet`  
      diff      lwr      upr      p adj  
2-1 19.971212 -0.2998092 40.24223 0.0552271  
3-1 40.304545 20.0335241 60.57557 0.0000025  
4-1 32.617257 12.2353820 52.99913 0.0002501  
3-2 20.333333 -2.7268370 43.39350 0.1058474  
4-2 12.646045 -10.5116315 35.80372 0.4954239  
4-3 -7.687288 -30.8449649 15.47039 0.8277810
```

→ Comparaciones pareadas de variable explicativa (diet)

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

La dieta afecta significativamente a la masa de los pollitos, pero... ¿qué dietas son las que más engordan a los pollitos y cuales las que menos? → **Post-Hoc test**

Post-Hoc test: Identifica subconjuntos de medias que no difieren entre sí → *Qué grupos específicos difieren entre sí.*

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%TukeyHSD()  
Tukey multiple comparisons of means  
95% family-wise confidence level
```

```
> anova<-aov(...)  
> TukeyHSD(anova)
```



```
Fit: aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

```
$`ChickWeight$Diet`  
      diff      lwr      upr      p adj  
2-1 19.971212 -0.2998092 40.24223 0.0552271  
3-1 40.304545 20.0335241 60.57557 0.0000025  
4-1 32.617257 12.2353820 52.99913 0.0002501  
3-2 20.333333 -2.7268370 43.39350 0.1058474  
4-2 12.646045 -10.5116315 35.80372 0.4954239  
4-3 -7.687288 -30.8449649 15.47039 0.8277810
```

→ Comparaciones pareadas de variable explicativa (diet)

→ P-valor

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

La dieta afecta significativamente a la masa de los pollitos, pero... ¿qué dietas son las que más engordan a los pollitos y cuales las que menos? → **Post-Hoc test**

Post-Hoc test: Identifica subconjuntos de medias que no difieren entre sí → *Qué grupos específicos difieren entre sí.*

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%TukeyHSD()  
Tukey multiple comparisons of means  
95% family-wise confidence level
```

```
> anova<-aov(...)  
> TukeyHSD(anova)
```



```
Fit: aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

\$`ChickWeight\$Diet`	diff	lwr	upr	p adj	
2-1	19.971212	-0.2998092	40.24223	0.0552271	→ Comparaciones pareadas de variable explicativa (diet)
3-1	40.304545	20.0335241	60.57557	0.0000025	→ P-valor
4-1	32.617257	12.2353820	52.99913	0.0002501	→ Intervalo de confianza (95%CI): menor (lwr) y mayor (upr) valor
3-2	20.333333	-2.7268370	43.39350	0.1058474	
4-2	12.646045	-10.5116315	35.80372	0.4954239	
4-3	-7.687288	-30.8449649	15.47039	0.8277810	

3.1. Análisis de Varianza (One-way ANOVA)

7. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo afectan distintas dietas a la masa de pollitos?

Unos pollitos son alimentados con cuatro dietas distintas (diet: 1, 2, 3 , ó 4), y se mide su masa al final del experimento (weight).

La dieta afecta significativamente a la masa de los pollitos, pero... ¿qué dietas son las que más engordan a los pollitos y cuales las que menos? → **Post-Hoc test**

Post-Hoc test: Identifica subconjuntos de medias que no difieren entre sí → *Qué grupos específicos difieren entre sí.*

```
> aov(ChickWeight$weight~ChickWeight$Diet)%>%TukeyHSD()  
Tukey multiple comparisons of means  
95% family-wise confidence level
```

```
> anova<-aov(...)  
> TukeyHSD(anova)
```



```
Fit: aov(formula = ChickWeight$weight ~ ChickWeight$Diet)
```

```
$`ChickWeight$Diet`  
      diff      lwr      upr      p adj  
2-1 19.971212 -0.2998092 40.24223 0.0552271  
3-1 40.304545 20.0335241 60.57557 0.0000025  
4-1 32.617257 12.2353820 52.99913 0.0002501  
3-2 20.333333 -2.7268370 43.39350 0.1058474  
4-2 12.646045 -10.5116315 35.80372 0.4954239  
4-3 -7.687288 -30.8449649 15.47039 0.8277810
```

→ Comparaciones pareadas de variable explicativa (diet)

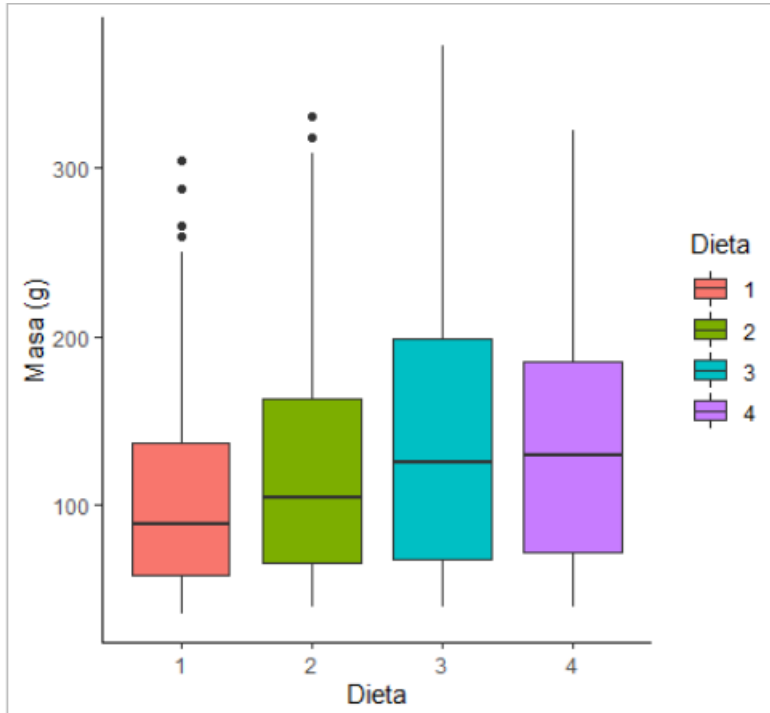
→ P-valor

→ Intervalo de confianza (95%CI): menor (lwr) y mayor (upr) valor

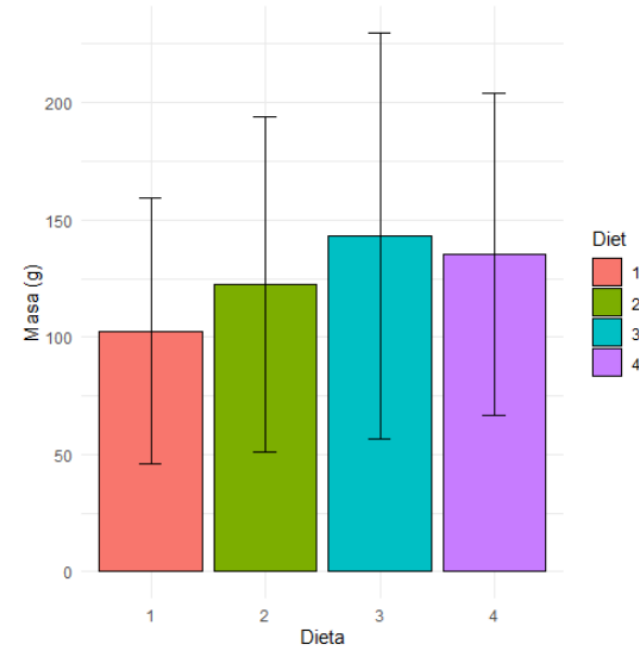
→ Diferencia entre los valores de la variable respuesta (weight) entre comparaciones

3.1. Análisis de Varianza (One-way ANOVA)

8. ¿Cómo se puede representar?



```
>ggplot(ChickWeight,  
aes(x=Diet,y=weight,fill=Diet))+  
  geom_boxplot()+  
  theme_classic()+  
  labs(x="Dieta",y="Masa (g)",fill="Dieta")
```



```
>ggplot(CW,aes(x=Diet,y=weight,fill=Diet))+  
  geom_bar(stat="identity", color="black",  
           position=position_dodge()) +  
  geom_errorbar(aes(ymin=weight-sd, ymax=weight+sd),  
                width=.2,position=position_dodge(.9))+  
  theme(legend.position="top")+  
  theme_minimal()+  
  labs( x="Dieta", y = "Masa (g)")
```

3.2. Análisis de Varianza (Two-way ANOVA)

1. ¿Cómo difiere de One-way ANOVA?

Las variables explicativas son dos o más factores con dos o más niveles. Es decir, en vez de un tratamiento, puede haber una combinación de tratamientos afectando a la variable respuesta.

2. ¿Qué tipo de datos se necesitan?

Variable respuesta (dep.; y) → Numérica continua

Variables explicativas (indep.; x) → Variables categóricas con dos o más niveles

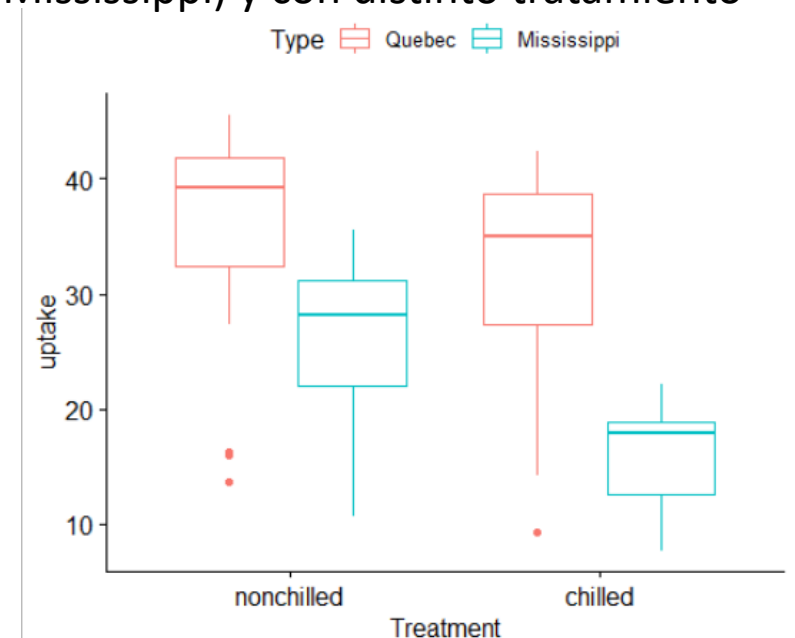
```
> anova<-aov(respuesta ~ expl1* expl2)
> summary(anova)
> TukeyHSD(anova)
```



3. ¿Cómo se interpreta el resultado de R?

E.g.: ¿Cómo varía el consumo de CO₂ en distintos tipos de plantas (Quebec & Mississippi) y con distinto tratamiento (chilled & nonchilled)?

```
> aov(CO2$uptake~CO2$Type*CO2$Treatment)%>%summary()
              Df Sum Sq Mean Sq F value    Pr(>F)
CO2$Type         1   3366      3366  52.509 2.38e-10 ***
CO2$Treatment     1    988       988  15.416 0.000182 ***
CO2$Type:CO2$Treatment 1    226       226   3.522 0.064213 .
Residuals       80   5128        64
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



3.2. Análisis de Varianza (Two-way ANOVA)

1. ¿Cómo difiere de One-way ANOVA?

Las variables explicativas son dos o más factores con dos o más niveles. Es decir, en vez de un tratamiento, puede haber una combinación de tratamientos afectando a la variable respuesta.

2. ¿Qué tipo de datos se necesitan?

Variable respuesta (dep.; y) → Numérica continua

Variables explicativas (indep.; x) → Variables categóricas con dos o más niveles

```
> anova<-aov(respuesta ~ expl1* expl2)
> summary(anova)
> TukeyHSD(anova)
```



3. ¿Cómo se interpreta el resultado de R?

```
> aov(CO2$uptake~CO2$Type*CO2$Treatment)%>%TukeyHSD()
Tukey multiple comparisons of means
95% family-wise confidence level
```

```
Fit: aov(formula = CO2$uptake ~ CO2$Type * CO2$Treatment)
```

```
$`CO2$Type`
```

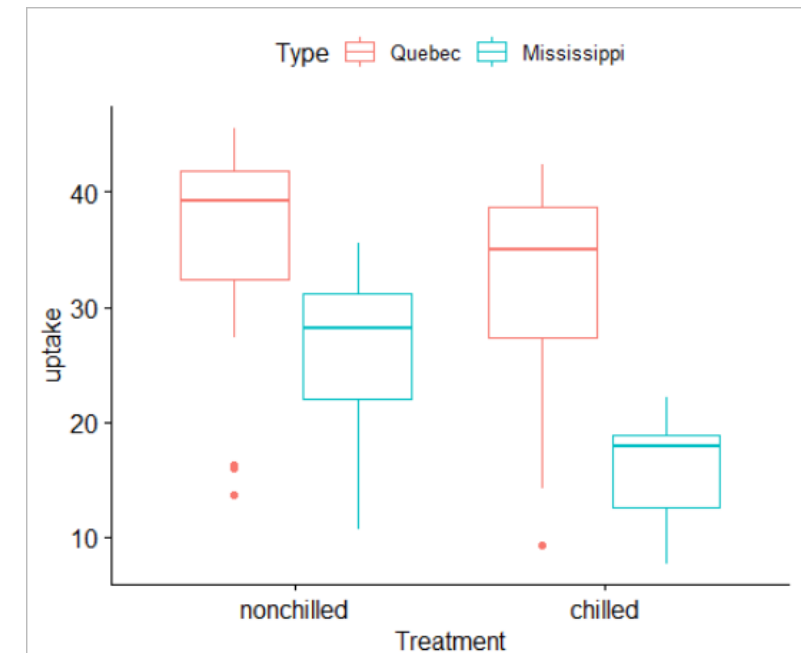
	diff	lwr	upr	p adj
Mississippi-Quebec	-12.65952	-16.13624	-9.182808	0

```
$`CO2$Treatment`
```

	diff	lwr	upr	p adj
chilled-nonchilled	-6.859524	-10.33624	-3.382808	0.0001817

```
$`CO2$Type:CO2$Treatment`
```

	diff	lwr	upr	p adj
Mississippi:nonchilled-Quebec:nonchilled	-9.380952	-15.8636917	-2.898213	0.0015893
Quebec:chilled-Quebec:nonchilled	-3.580952	-10.0636917	2.901787	0.4727714
Mississippi:chilled-Quebec:nonchilled	-19.519048	-26.0017869	-13.036308	0.0000000
Quebec:chilled-Mississippi:nonchilled	5.800000	-0.6827393	12.282739	0.0959830
Mississippi:chilled-Mississippi:nonchilled	-10.138095	-16.6208345	-3.655356	0.0005553
Mississippi:chilled-Quebec:chilled	-15.938095	-22.4208345	-9.455356	0.0000000



3.3. ANOVA ejercicio

Ejercicios: 3. Ejer_ANOVA