ANOVA (Análisis de la Varianza)

El análisis de la varianza es un procedimiento estadístico que pretende contrastar si existen diferencias en el comportamiento de una variable, Y , entre los distintos grupos determinados por un cierto factor. (Tenemos 1 vble. cuantitativa y 1 factor o vble. cualitativa)
p.ej. : Los salarios entre las comunidad autónomas

Viene a ser una generalización del test t de student de igualdad de (dos) medias a un número cualquiera de grupos.
El problema que afronta se puede plantear de formas (equivalentes) alternativas:

-El comportamiento de la variables Y es diferente (o no) entre los grupos (que forma el factor)
-La media de Y es diferente (o no) en los distintos grupos
-El factor afecta, produce efecto (o no) al comportamiento de la variable Y
-El factor discrimina , marca diferencias (o no) el comportamiento de Y

El contraste (ANOVA) se lleva a cabo asumiendo algunas hipótesis ("MODELO de ANOVA") pero, es un método bastante **robusto** ya que (en el fondo) se basa en una idea general de cumplimiento universal : Cuando mezclamos dos (o más) conjuntos de datos muy **diferentes** entre sí la varianza total **crece más** que si mezclamos dos (o más) conjuntos de datos parecidos.
Veamos un ejemplo trivial.

	X1	X2	X1 U X2
	6	4	
	4	6	
media	5	5	5
var	1	1	1
SCI	2	2	
		SCT	**4**

SCE S.C.entre grupos= 0

	X1	X2	X1 U X2
	2	8	
	2	8	
media	2	8	5
var	0	0	9
SCI	0	0	
		SCT	**36**

SCE S.C.entre grupos= 36

$$\text{Suma de cuadrados externa (entre-grupos)} = SCE = \sum_{\forall grupos} n_g(\bar{X}_g - \bar{X})^2$$

SCT= SCE+SCI

En el primer caso la SCT es 4 (La SCI=2+2=4 y La SCE=0) Los dos conjuntos son idénticos

En el segundo caso SCT es 36 (La SCI=0 pero la SCE= 36 Los dos conjuntos son muy diferentes

EJEMPLO PARA LA COMPRESIÓN DE LA SITUACIÓN:

Supongamos una población de las notas $y_{i.j}$ de un universo de 9 alumnos de tres grupos distintos , así:

grupo 1	grupo 2	grupo 3
5	5	5
5	5	5
5	5	5

evidentemente en este caso la media global es 5 y la de cada grupo también $\mathbf{y_{i.j}} = \boldsymbol{\mu}$ cada valor es igual a la media general. NO HAY DIFERENCIAS ENTRE GRUPOS NI DENTRO DE LOS GRUPOS

Supongamos que aplicamos un método de enseñanza (factor) que afecta : subiendo las notas del grupo 1 en 1 punto , las del grupo dos en 2 puntos y no modificando las del grupo 3 . Así.

grupo 1	grupo 2	grupo 3
5+1=6	5+2=7	5
5+1=6	5+2=7	5
5+1=6	5+2=7	5

ahora la nota de un alumno sería $\mathbf{y_{i.j}} = \boldsymbol{\mu} + \boldsymbol{\alpha_i}$

en el que los α_i son 1, 2, 0 los EFECTOS QUE PRODUCEN EL FACTOR (MÉTODO) EN CADA NIVEL . PARECE CLARO QUE EL FACTOR INFLUYE EN ESTABLECER DIFERENCIAS ENTRE GRUPOS; PERO NO DENTRO

EJEMPLO PARA LA COMPRESIÓN DE LA SITUACIÓN II

Lo más habitual es que haya alumnos que rindan más que otros (por diversas razones aleatorias o que en principio no dependan de un factor) ,son por tanto, comportamientos aleatorios individuales que denominamos $\varepsilon_{i,j}$ implantando algunos en el ejemplo ,sería:

grupo 1	grupo 2	grupo 3
5+1-1=5	5+2+2=9	5+0+3=8
5+1-2=4	5+2+0=7	5+0+4=9
5+1+0=6	5+2+1=8	5+0+0=5

en el que los efectos aleatorios $\varepsilon_{i,j}$ serían -1,-2,0,2,0,1,3,4,0 que fomentan la variablidad **dentro** de los grupos INTRA-GRUPOS

Entonces para cada valor tendremos el modelo $\boxed{y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}}$

TENEMOS DOS TIPOS DE VARIABILIDAD : LA ENTRE GRUPOS (DEBIDA AL FACTOR) Y LA INTRA GRUPOS (DEBIDA A LA ALEATORIEDAD....) .

PARA PODER AFIRMAR QUE EL FACTOR PRODUCE EFECTOS. LA VARIABILIDAD ENTRE LOS GRUPOS HA DE SER *SIGNIFICATIVAMENTE* GRANDE RESPECTO A LA INTRA GRUPOS

MODELO E HIPÓTESIS

EL MODELO SERÍA EL SIGUIENTE :	$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

donde :

Y_{ij} = observación j-ésima del nivel i μ = media general

α_i = efecto el i-ésimo nivel del factor ε_{ij} = error aleatorio independiente $N(0,\sigma)$

PANTEAMOS LA SIGUIENTE HIPÓTESIS NULA

$H_O = \alpha_1 = \alpha_2 = \alpha_3 = \ldots\ldots = \alpha_K$

o bien si consideramos $\mu_i = \mu + \alpha_i$

entonces $H_o = \mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_K$

en definitiva se quiere comprobar la **no influencia** del factor α

DE OTRA FORMA : SI TODAS LAS MUESTRAS PROCEDEN DE LA MISMA POBLACIÓN

CONDICIONES GENERALES DE APLICACIÓN.

A- **INDEPENDENCIA DE LOS ERRORES**. Los errores experimentales han de ser independientes . Se consigue si los sujetos son asignados aleatoriamente. Es decir , se consigue esta condición si las elementos de los diversos grupos han sido elegidos ***por muestreo aleatorio*** .

B- **NORMALIDAD** . Se supone que los errores experimentales se distribuyen normalmente.Lo que supone que cada una de las puntuaciones $y_{i.i}$ se distribuirá normalmente . Para comprobarlo se puede aplicar un test de ajuste a la distribución Normal como el de **Kolmogov-Smirnov**

C- **HOMOGENEIDAD DE VARIANZAS (HOMOSCEDASTICIDAD).** La varianza de los subgrupos ha de ser homogénea $\sigma^2_1 = \sigma^2_2 = \ldots\ldots = \sigma^2_K$ ya que están debidas al error. Se comprobarán mediante los test de : Razón de varianzas (máx/mín) , C de Cochran , Barlett-Box, Sea Y una variable aleatoria sobre la que se han tomado N muestras ;de manera que obtenemos k muestras correspondientes a las k categorías(niveles) del factor. Si el tamaño de la muestra para cada categoría es el mismo (n) para todas, estaremos ante un modelo **BALANCEADO** en el que N = nk   Y sigue una $N(\mu_i, \sigma)$ para i = 1,2,3,,,,,,,,k.

	NIVELES DEL FACTOR				
	1	2	....	....	k
1	$Y_{1.1}$	$Y_{2.1}$	....		$Y_{k.1}$
2	$Y_{1.2}$	$Y_{2.2}$	....		$Y_{k.2}$.
j	$Y_{1.j}$	$Y_{2.j}$.	.... Y_{ij}		$Y_{k.j}$
n	$Y_{1.n}$	$Y_{2.n}$	....		$Y_{k.n}$

i = 1,2,3,....k   j = 1,2,3,....n (en el caso de balanceado)

media muestral correspondiente al nivel i del factor $= 1/n \sum_{j=1}^{n} Y_{ij} = \bar{y}_{i.}$

media general $= 1/N \sum_{i=1}^{k} \sum_{j=1}^{n} Y_{ij} = \bar{y}$

DESCOMPOSICIÓN DE LA VARIABILIDAD TOTAL

que quedaría establecida de la siguiente forma:

$$\sum\sum (Y_{ij} - \bar{\mathbf{y}})^2 = \sum\sum (Y_{ij} - \bar{\mathbf{y}}_{i.})^2 + n\sum_{i=1}^{k} (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}})^2$$

$$S.C.T. = S.C.I. + S.C.E.$$

Donde : S.C.T. = SUMA DE CUADRADOS TOTAL

S.C.I. = SUMA DE CUADRADOS INTRA-GRUPOS (within-groups)

S.C.E. = SUMA DE CUADRADOS ENTRE-GRUPOS (between-groups)

En el caso de NO ser BALANCEADO , no habría una solo n sino k distintos valores.

la descomposición quedaría entonces así:

$$\sum\sum (Y_{ij} - \bar{\mathbf{y}})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} (Y_{ij} - \bar{\mathbf{y}}_{i.})^2 + \sum_{i=1}^{k} n_i (\bar{\mathbf{y}}_{i.} - \bar{\mathbf{y}})^2$$

$$S.C.T. = S.C.I. \qquad S.C.E.$$

S.C.T. será la suma de las desviciones de cada Y_{ij} respecto a la media general $\bar{y}$

S.C.I. será la variación total de las observaciones alrededor de cada una de las medias muestrales

S.C.E. será la variación de las medias muestrales de cada grupo alrededor de la media general

SI DIVIDIMOS LA DESCOMPOSICIÓN DE LA VARIABILIDAD POR LA VARIANZA

$$\boxed{\frac{S.C.T.}{\sigma^2} = \frac{S.C.I.}{\sigma^2} + \frac{S.C.E.}{\sigma^2}}$$

Dado que las observaciones de las k muestras son independientes , cada uno de los tres sumatorios desde i = 1 hasta i = k es la suma de k variables aleatorias que tienen distribuciones χ^2 de tal manera que

$S.C.T. / \sigma^2 \longrightarrow \chi^2$ con (N-1) grados de libertad

$S.C.I. / \sigma^2 \longrightarrow \chi^2$ con k(n-1) grados de libertad

$S.C.E. / \sigma^2 \longrightarrow \chi^2$ con (k-1) grados de libertad

en el caso de diseño NO BALANCEADO $S.C.I. / \sigma^2 \longrightarrow \chi^2$ con $(n_1-1)+(n_2-1)\ldots(n_k-1)$ grados de libertad

Bajo la hipótesis nula; la variabilidad entre grupos no deberá superar significativamente a la variabilidad intra grupo luego el cociente **S.C.E / S.C.I. no deberá ser significativamente grande** . Esa ***Significabilidad*** nos la dará una distribución de probabilidad asociada. Así:

$$\frac{S.C.E. \longrightarrow}{S.C.I. \longrightarrow} \quad \frac{\chi^2_{(k-1)}}{\chi^2_{k(n-1)}} \text{ ó bien } \chi^2_{(N-k)}$$

si multiplicamos el conciente por N-k / k-1

$$\frac{S.C.E.\cdot(N-k)}{S.C.I\cdot(k-1)} \begin{matrix}\longrightarrow \\ = \\ \longrightarrow\end{matrix} \frac{\chi^2_{(k-1)}/(k-1)}{\chi^2_{(N-k)}/(N-k)} \longrightarrow \boxed{F_{(k-1),(N-k)} \Rightarrow \mathbf{F}}$$

Así para un nivel de significación α

Si $\mathbf{F} > F_{\alpha}$ se rechaza la H_0 medias de los grupos no son iguales .

Si $\mathbf{F} < F_{\alpha}$ no se puede rechazar la H_0 ; luego medias de los grupos son iguales

no influye el factor

no hay diferencias significativa entre los niveles

TABLA " ANOVA " DE UN FACTOR .

FUENTE DE VARIACIÓN	G.L	SUMA DE CUADRADOS	$F_{(k-1)(N-k)}$
ENTRE-GRUPOS	k-1	S .C. E.	$F = \dfrac{S.C.E(N-k)}{S.C.I.(k-1)}$
INTRA-GRUPOS	N-1 $(n_1-1)+(n_2-1)+.(n_k-1)$ no balanceado	S. C. I.	
TOTAL		S. C. T.	

Ejemplo (simple de las notas de antes)

grupo 1	grupo 2	grupo 3
5+1-1=5	5+2+2=9	5+0+3=8
5+1-2=4	5+2+0=7	5+0+4=9
5+1+0=6	5+2+1=8	5+0+0=5

nota<-c(5,4,6,9,7,8,8,9,5)
grupo<-as.factor(c("1","1","1","2","2","2","3","3","3"))
notas<-as.data.frame(cbind(grupo,nota))

anova<-aov(nota~grupo,data=notas)
summary(anova)

Console Terminal × Jobs ×

```
~/
> anova<-aov(nota~grupo,data=notas)
> summary(anova)
            Df Sum Sq Mean Sq F value Pr(>F)
grupo        1  3.068   3.068   0.877   0.38
Residuals    7 24.487   3.498
>
>
```

En este caso no habría diferencias significativas en las notas entres los diferentes grupos.
No podríamos descartar que la nota media es la misma en todos los grupos.
No podemos considerar que el factor grupo tiene efecto en las notas.

COMPARACIONES MULTIPLES ENTRE MEDIAS.

Si se ha **rechazado la hipótesis nula** de igualdad de medias esto supone : EXISTE AL MENOS UNA DIFERENCIA. . PERO NO SABEMOS CUÁNTAS,NI ENTRE QUE MEDIAS.

Para ello se establecen pruebas de comparación múltiple (a posteriori) ;

Hay muchas . Entre ellas destacan la LSD (Mínimas diferencias significativas), HSD de Tukey (Diferencia "honestas" significativas) , SNK, y Scheffé.

La más "conservadora " (con las hipótesis nulas de medias iguales) es la de Scheffé.
Y La que tiende a detectar más "liberalmente" las diferencias entre medias es la LSD.
HSD y SNK son intermedias.

HSD está incorporada a los paquetes básicos de R. En cambio LSD y Scheffe requieren del paquete agricolae

Ejemplo ANOVA un factor

Se trata de comprobar si el factor día de la semana influye en el tiempo de elaboración de un componente.

Estaremos ante un ANOVA de un factor (día de la semana) con cinco niveles y no balanceado pues los valores por día son distintos

```
#los datos están en esta dirección:
data<-read.table("http://www.uv.es/lejarza/eaa/tareas/t3r/avo.txt” ,header=T, sep="", )
names(data) #Comprobamos que hay dos variables al y día
```

DATOS DE LOS TIEMPOS , EN MINUTOS, QUE SE TARDA EN ELABORAR UN COMPONENTE ASÍ COMO EL DÍA QUE SE REALIZÓ LA PRUEBA

MINUTOS	DIA	MINUTOS	DIA	MINUTOS	DIA	MINUTOS	DIA	MINUTOS	DIA
215	MARTES	209	VIERNES	220	JUEVES	204	MIERCOLES	201	MARTES
215	VIERNES	221	MARTES	219	MIERCOLES	212	MIERCOLES	220	MARTES
219	MIERCOLES	208	VIERNES	219	JUEVES	219	MIERCOLES	222	LUNES
204	MARTES	214	VIERNES	207	JUEVES	204	JUEVES	215	MIERCOLES
209	MARTES	212	MARTES	219	MIERCOLES	203	MIERCOLES	205	VIERNES
216	MIERCOLES	208	JUEVES	213	JUEVES	221	LUNES	207	MARTES
207	MARTES	223	JUEVES	216	JUEVES	204	MIERCOLES	206	MIERCOLES
206	MARTES	209	LUNES	225	LUNES	224	JUEVES	216	MARTES
220	LUNES	204	MIERCOLES	208	JUEVES	221	JUEVES	211	MARTES
228	LUNES	202	MIERCOLES	224	VIERNES	206	VIERNES	217	MIERCOLES
220	MARTES	205	LUNES	220	MIERCOLES	200	JUEVES	212	VIERNES
202	MIERCOLES	223	MIERCOLES	222	VIERNES	211	MIERCOLES	216	MARTES
210	MARTES	206	LUNES	228	LUNES	212	MARTES	223	MIERCOLES
200	MARTES	200	MARTES	203	VIERNES	208	JUEVES	216	VIERNES
205	MIERCOLES	208	VIERNES	200	VIERNES	223	JUEVES	206	MIERCOLES
205	LUNES	204	JUEVES	218	LUNES	228	LUNES	203	VIERNES
205	MARTES	210	JUEVES	203	VIERNES	209	LUNES	200	MIERCOLES
221	MARTES	218	LUNES	210	JUEVES	211	JUEVES	219	MIERCOLES
213	LUNES	222	LUNES	210	VIERNES	204	JUEVES	210	MIERCOLES
200	MARTES	225	LUNES	203	VIERNES	222	JUEVES	215	MARTES

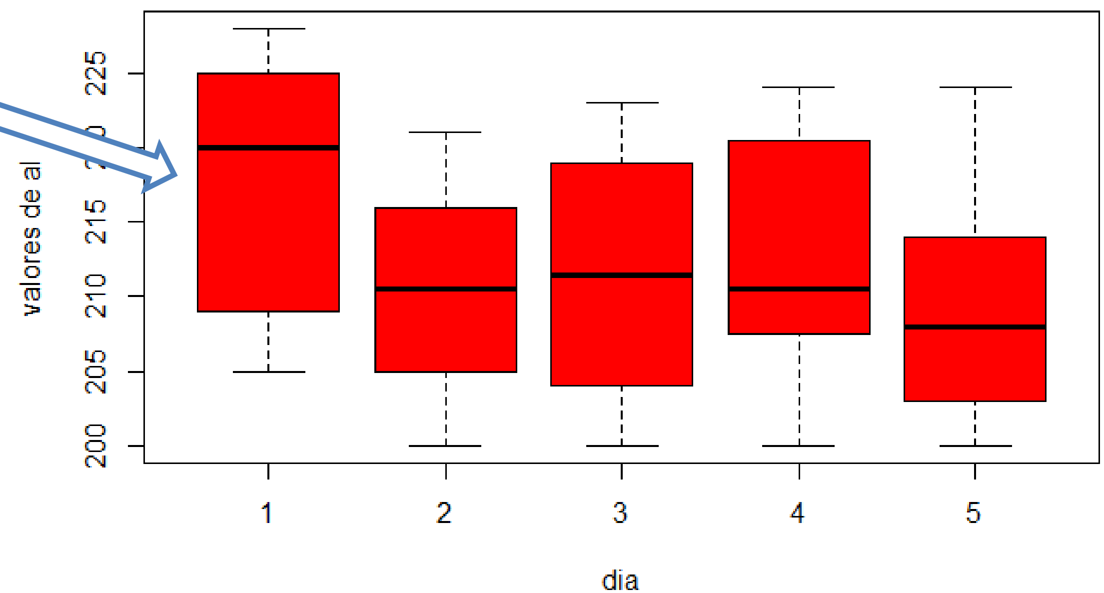
```
> summary(data)
       al              dia      
 Min.   :200.0   Min.   :1.00  
 1st Qu.:205.0   1st Qu.:2.00  
 Median :211.0   Median :3.00  
 Mean   :212.2   Mean   :2.98  
 3rd Qu.:219.0   3rd Qu.:4.00  
 Max.   :228.0   Max.   :5.00  
> medal<-mean(data$al);sdal<-sd(data$al)
> medal;sdal
[1] 212.24
[1] 7.85361
>
```

Análisis exploratorio

#gráfico conjunto: Gráfico de cajas de al en función del día

boxplot(al~dia, ylab="valores de al", xlab="dia",col="red" ,data=data)

Parece que los lunes….



#recodificamos el factor “dia” como factor (si no la función aov no funciona)
data$Factordia <-factor(data$dia,
labels=c("lunes","martes","miercoles","jueves","viernes"))
fix(data) # editamos los datos para comprobarlo

#medias y varianzas de al por día
tapply(data$al,data$Factordia,mean)
tapply(data$al, data$Factordia, var)

```
~/
> #medias y varianzas de al por día
> tapply(data$al,data$Factordia,mean)
   lunes    martes miercoles    jueves   viernes 
217.7647  210.3636  211.5833  212.7500  209.4706 
> tapply(data$al, data$Factordia, var)
   lunes    martes miercoles    jueves   viernes 
69.44118  49.86147  58.25362  58.09211  47.63971 
>
>
```

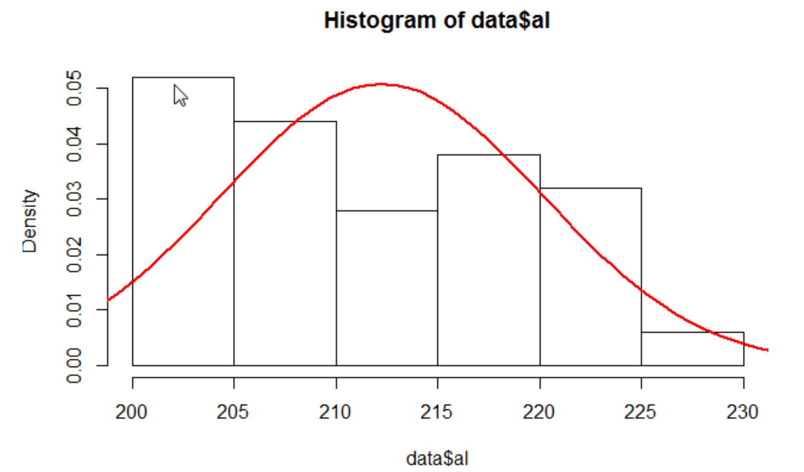
Normalidad

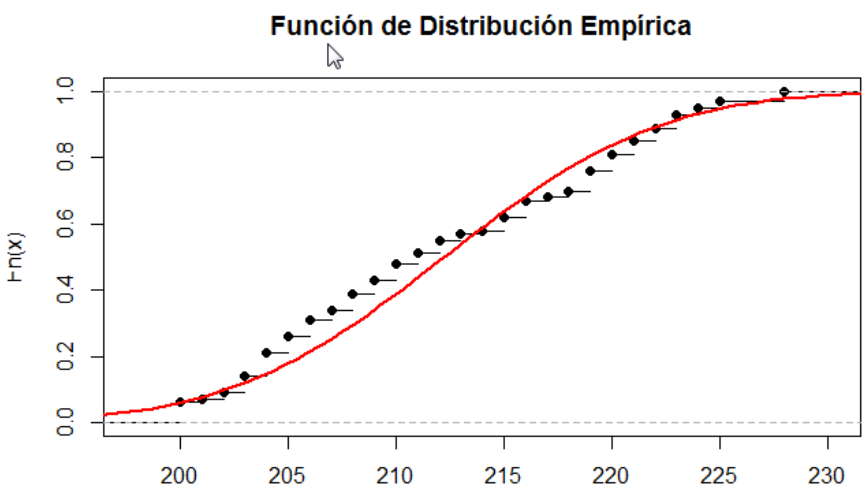
Cumplimiento de hipótesis del modelo

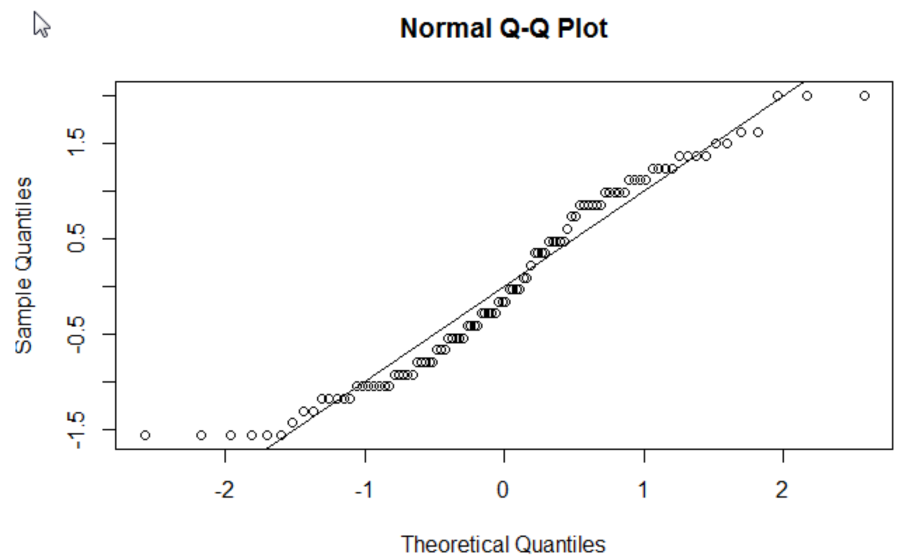
```
#Comparación con normal. Visualización en gráficos
hist(data$al,freq = FALSE)
curve(dnorm(x,medal,sdal), 185, 240, add=T, col="red",lwd=2 )

plot(ecdf(data$al), main="Función de Distribución Empírica")
curve(pnorm(x,medal,sdal),185,240,add=T,col="red",lwd=2)

qqnorm((data$al-medal)/sdal) # Grafico Q-Q
abline(0,1) # se añade la recta y=x
```







Normalidad

Cumplimiento de hipótesis del modelo

Contraste de Kolmogorv-Smirnov

ks.test(data$al,pnorm,mean=212.24,sd=7.85361)

```
>
>
> ks.test(data$al,pnorm,mean=212.24,sd=7.85361)

        One-sample Kolmogorov-Smirnov test

data:  data$al
D = 0.10531, p-value = 0.2173
alternative hypothesis: two-sided
```

Homoscedasticidad

#test de levenne de igualdad de varianzas (Homoscedasticidad) se requiere paquete car

library(car)

leveneTest(data$al, data$Factordia)

#test de Bartlett de Igualdad de varianzas

bartlett.test(al ~ Factordia, data=data)

```
> leveneTest(data$al, data$Factordia)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  4  0.4568 0.7672
      95
>
>
> bartlett.test(al ~ Factordia, data=data)

        Bartlett test of homogeneity of variances

data:  al by Factordia
Bartlett's K-squared = 0.74463, df = 4, p-value = 0.9457
```

anova

```
> Anova <- aov(al ~ Factordia, data=data)
>
> summary(Anova)
            Df Sum Sq Mean Sq F value Pr(>F)
Factordia    4    742  185.57   3.287 0.0144 *
Residuals   95   5364   56.46
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

Al 5 % de significación descartamos la Hip. Nula (medias iguales) . // al 1 % no podríamos.
Existen diferencias significativas (al 5 %) en el tiempo de fabricación entre los distintos días
El día de fabricación influye significativamente (al 5%) en el tiempo de fabricación

Comparaciones múltiples

```
> TukeyHSD(Anova)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = al ~ Factordia, data = data)

$Factordia
                        diff        lwr        upr     p adj
martes-lunes      -7.4010695 -14.148799 -0.6533403 0.0241638
miercoles-lunes   -6.1813725 -12.805413  0.4426681 0.0793852
jueves-lunes      -5.0147059 -11.907932  1.8785206 0.2630799
viernes-lunes     -8.2941176 -15.461353 -1.1268821 0.0148644
miercoles-martes   1.2196970  -4.948001  7.3873949 0.9817005
jueves-martes      2.3863636  -4.069576  8.8423031 0.8418536
viernes-martes    -0.8930481  -7.640777  5.8546811 0.9960246
jueves-miercoles   1.1666667  -5.159882  7.4932154 0.9859016
viernes-miercoles -2.1127451  -8.736786  4.5112955 0.9010641
viernes-jueves    -3.2794118 -10.172638  3.6138147 0.6776695
```

Según el test de HSD de Tukey
Hay diferencias "honestamente" significativas (5 %) entre el tiempo de fabricación de los lunes y el de los martes y, también entre los de los lunes y los viernes

¿Otros test?

```
> library(agricolae)
> lsd<-LSD.test(Anova,"Factordia")
> lsd
$statistics
   MSerror Df   Mean       CV
  56.46282 95 212.24 3.540414

$parameters
        test p.ajusted    name.t ntr alpha
  Fisher-LSD      none Factordia   5  0.05

$means
                al      std  r      LCL      UCL Min Max    Q25   Q50    Q75
jueves    212.7500 7.621818 20 209.4143 216.0857 200 224 207.75 210.5 220.25
lunes     217.7647 8.333137 17 214.1467 221.3827 205 228 209.00 220.0 225.00
martes    210.3636 7.061266 22 207.1832 213.5441 200 221 205.25 210.5 215.75
miercoles 211.5833 7.632406 24 208.5383 214.6284 200 223 204.00 211.5 219.00
viernes   209.4706 6.902152 17 205.8526 213.0886 200 224 203.00 208.0 214.00

$comparison
NULL

$groups
                al groups
lunes     217.7647      a
jueves    212.7500      b
miercoles 211.5833      b
martes    210.3636      b
viernes   209.4706      b
```

casi tampoco se solapa

```
> scheffe<-scheffe.test(Anova,"Factordia")
> scheffe
$statistics
   MSerror Df        F   Mean       CV
  56.46282 95 2.467494 212.24 3.540414

$parameters
     test    name.t ntr alpha
  Scheffe Factordia   5  0.05

$means
                 al      std  r Min Max    Q25   Q50    Q75
jueves     212.7500 7.621818 20 200 224 207.75 210.5 220.25
lunes      217.7647 8.333137 17 205 228 209.00 220.0 225.00
martes     210.3636 7.061266 22 200 221 205.25 210.5 215.75
miercoles  211.5833 7.632406 24 200 223 204.00 211.5 219.00
viernes    209.4706 6.902152 17 200 224 203.00 208.0 214.00

$comparison
NULL

$groups
                 al groups
lunes      217.7647      a
jueves     212.7500     ab
miercoles  211.5833     ab
martes     210.3636     ab
viernes    209.4706      b
```

Basándonos en el criterio de Scheffé, que es el que exige más evidencia para detectar diferencias:
El tiempo de fabricación es significativamente superior el lunes y significativamente inferior el viernes . Siendo el resto de los días intermedio.

Dados estos resultados que parecen indicar claramente que el lunes tiene un comportamiento bien diferente (mayor tiempo de fabricación) podríamos plantearnos qué pasaría si sólo consideráramos el resto de los días (eliminando los datos de los lunes). Caso de no haber diferencias significativas la conclusión resultaría meridiana.

```
> # hacemos el ANOVA sin el lunes que parece tener un comportamiento "especial"
>
> datos2<-subset(data, data$Factordia != "lunes")
> fix(datos2)
>
> Anova2<-aov(al~Factordia,data=datos2)
> summary(Anova2)
            Df Sum Sq Mean Sq F value Pr(>F)
Factordia    3    117   39.04   0.725   0.54
Residuals   79   4253   53.83
>
>
```

Ahora NO PODEMOS DESCARTAR la hipótesis nula. No podemos descartar que el tiempo medio de fabricación es el mismo los martes, miércoles jueves y viernes. Como, sin embargo, cuando considerábamos el lunes sí lo hicimos, la conclusión …

Interacción de factores : Anova de dos factores o más

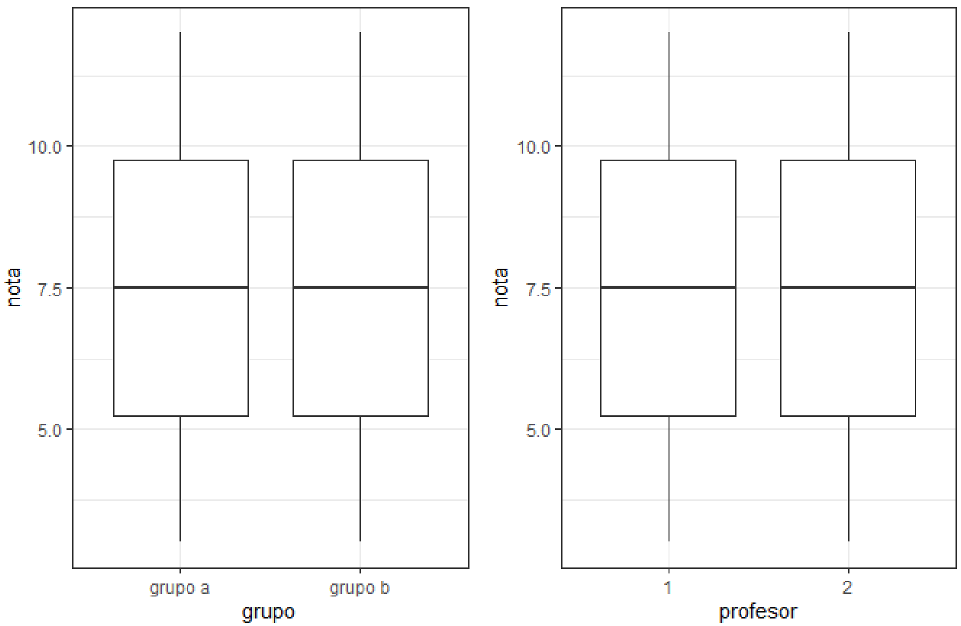
Si queremos investigar el posible efecto de 2 o más factores sobre una variable ,el uso de una anova (de un factor) para cada factor nos puede llevar a conclusiones erróneas debido al fenómeno de la interacción: Podría ocurrir que los factores por separado no causarán efecto en la variable pero que la interacción de ambos (o más) factores sí ,en según que combinaciones de niveles de ambos factores

X	FactorA	FactorB
3	1	1
4	1	1
5	1	1
6	1	1
7	1	1
11	1	2
10	1	2
9	1	2
8	1	2
12	1	2
9	2	1
8	2	1
10	2	1
12	2	1
11	2	1
7	2	2
5	2	2
3	2	2
4	2	2
6	2	2

		Factor B		MEDIA
	X	B1	B2	
		3	11	
		4	10	
	A1	5	9	7,5
		6	8	
Factor A		7	12	
		9	7	
		8	5	
	A2	10	3	7,5
		12	4	
		11	6	
MEDIA		7,5	7,5	7,5

MEDIAS EN INTERACCIÓN	B1	B2
A1	5	10
A2	10	5

		gl		F	
SCE_GRUPOS A	0	1	(grupos-1)	0	No hay diferencias según A
SCE_GRUPOS B	0	1	(grupos-1)	0	No hay diferencias según B
SCE INTRACCIÓN AxB	125	1	(grupos-f-1)	50	Sí hay diferencias si se combina A y B
SCI (residuo)	40	16	(n-grupos)		
SCT	165	19	(n-1)		



Introducimos los datos anteriores

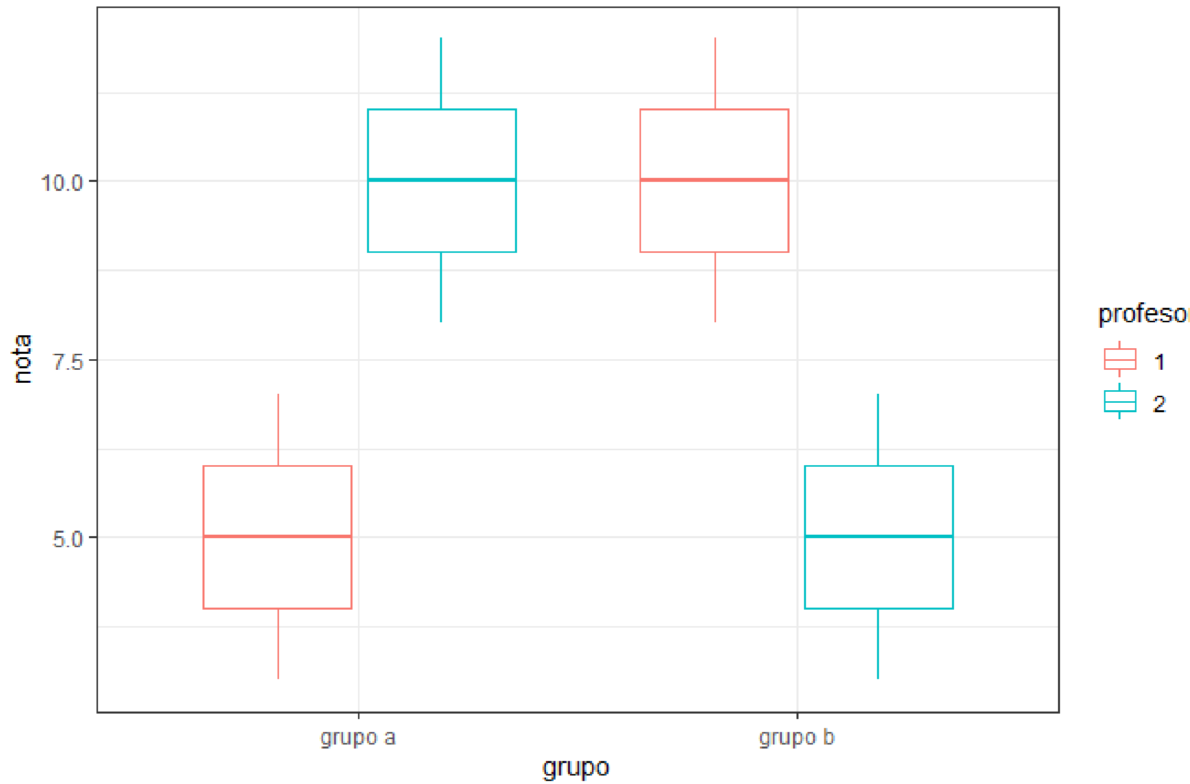
```
nota <- c(3,4,5,6,7,11,10,9,8,12,9,8,10,12,11,7,
          5,3,4,6)
grupo <- c(rep(c("grupo a", "grupo b"), c(10,10)))
profesor <- rep(c(1,2), each = 5, times = 2)
datos <- data.frame(grupo = grupo,
   profesor = as.factor( profesor), nota= nota)
```

Anovas   por separado

```
> summary(aov(nota~grupo,data=datos))
            Df Sum Sq Mean Sq F value Pr(>F)
grupo        1      0   0.000       0      1
Residuals   18    165   9.167
> summary(aov(nota~profesor,data=datos))
            Df Sum Sq Mean Sq F value Pr(>F)
profesor     1      0   0.000       0      1
Residuals   18    165   9.167
```

```
> anova2F<-aov(nota~profesor*grupo,data=datos)
> summary(anova2F)
               Df Sum Sq Mean Sq F value   Pr(>F)    
profesor        1      0     0.0       0        1    
grupo           1      0     0.0       0        1    
profesor:grupo  1    125   125.0      50 2.65e-06 ***
Residuals      16     40     2.5                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
>
>
```

Anova de (los) dos factores



Ejemplo: con los datos de las base datos bankdata estudiar si la edad y el lugar de residencia afecta al saldo medio ponderado de los clientes.

```
> anova<-aov(saldomedio~region, data=bankdata)
> summary(anova)
             Df    Sum Sq  Mean Sq F value Pr(>F)
region        2 9.333e+06  4666404   0.404  0.668
Residuals   597 6.891e+09 11543135
> anovab<-aov(saldomedio~edad,data=bankdata)
> summary(anovab)
             Df    Sum Sq  Mean Sq F value Pr(>F)
edad          2 1.028e+07  5140026   0.445  0.641
Residuals   597 6.890e+09 11541548
> summary(anovac)
Error in summary(anovac) : object 'anovac' not found
> anovac<-aov(saldomedio~edad*region, data=bankdata)
> summary(anovac)
             Df    Sum Sq  Mean Sq F value Pr(>F)
edad          2 1.028e+07  5140026   0.446  0.641
region        2 9.533e+06  4766559   0.413  0.662
edad:region   4 6.372e+07 15930433   1.381  0.239
Residuals   591 6.817e+09 11534771
> 
```