

PRÁCTICA 4: Análisis de la varianza/estadística no paramétrica

4.1 Análisis de la varianza

DESARROLLO DE LA PRÁCTICA

El software R dispone de varios programas para realizar un análisis de la varianza a un conjunto de datos (aov, anova, lm,...), teniendo en cuenta los tres diseños que hemos estudiado, usaremos solo dos programas de análisis de la varianza. Se presupone que la variable a analizar tiene una distribución normal con dispersión común en todos los tratamientos.

Primer supuesto: Análisis de la varianza de un factor (diseño completamente aleatorizado).

Ejemplo: El conjunto de datos a analizar es “madera.RData”, que describe el grado de desgaste de las muestras de enchapados de madera sintéticos para cinco tipos de madera.

Para realizar dicho análisis, debemos ejecutar una serie de comandos:

```
#load("madera.RData")
load(url("https://ddv.stic.ull.es/users/cpgonzal/public/Data/R/madera.RData"))
str(madera)
attach(madera)
mod1F<-lm(formula(desgaste~tipo))
anova1F<-anova(object=mod1F)
print(anova1F)
```

que nos lee el fichero de datos, vemos la estructura del fichero e indicamos el modelo, mediante la orden print obtenemos la siguiente tabla ANOVA

Analysis of Variance Table						
Response: desgaste						
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
tipo	4	0.6170	0.154250	7.404	0.001683	**
Residuals	15	0.3125	0.020833			

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

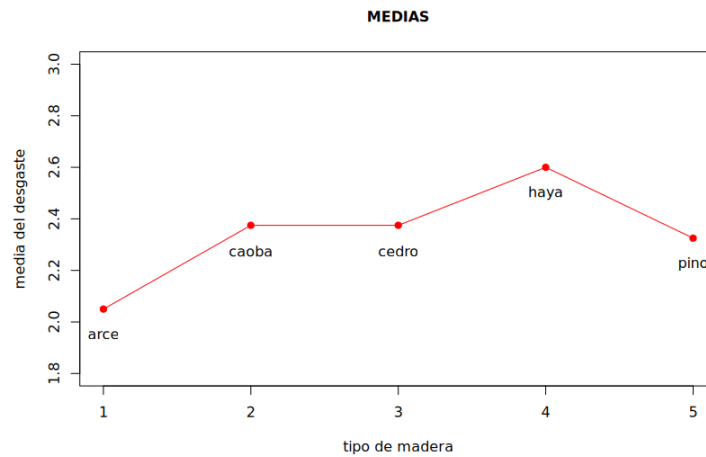
que nos indica un efecto significativo (estadísticamente) del desgaste según el tipo de madera de enchapado con los valores usuales (0.1, 0.05 o 0.01) de α . Para visualizar la situación que tenemos en los datos, podemos calcular las medias y desviaciones típicas de la variable respuesta en cada nivel del factor con

```
tapply(desgaste, tipo, mean)
tapply(desgaste, tipo, sd)
```

arce	caoba	cedro	haya	pino
2.050	2.375	2.375	2.600	2.325
arce	caoba	cedro	haya	pino
0.129	0.0957	0.1708	0.1414	0.1708

y la representación gráfica, de las medias de la variable respuesta, es

```
plot(mod1F,type='o',pch=16, xlab="tipo",ylab="media del
desgaste",main="MEDIAS",cex.main=1.0)
```



Dado que tenemos un diseño balanceado en los tratamientos, podemos realizar todos los procedimientos de comparaciones múltiples estudiados. Para ello debemos utilizar un subprograma (ANOVA_CMP) realizado con dicho propósito. Para los datos de madera se proporciona los siguientes métodos (Scheffe, Sidack y Ryan)

INTERVALOS DE CONFIANZA PAREADOS SEGÚN SCHEFFE:

	i	j	med(i)	med(j)	dif_med	Ampl	LI_Scheffe	LS_Scheffe	RES
1	arce	caoba	2.050	2.375	-0.325	0.35681	-0.681813	0.031813	NO Sig
2	arce	cedro	2.050	2.375	-0.325	0.35681	-0.681813	0.031813	NO Sig
3	arce	haya	2.050	2.600	-0.550	0.35681	-0.906813	-0.193187	SIG
4	arce	pino	2.050	2.325	-0.275	0.35681	-0.631813	0.081813	NO Sig
5	caoba	cedro	2.375	2.375	0.000	0.35681	-0.356813	0.356813	NO Sig
6	caoba	haya	2.375	2.600	-0.225	0.35681	-0.581813	0.131813	NO Sig
7	caoba	pino	2.375	2.325	0.050	0.35681	-0.306813	0.406813	NO Sig
8	cedro	haya	2.375	2.600	-0.225	0.35681	-0.581813	0.131813	NO Sig
9	cedro	pino	2.375	2.325	0.050	0.35681	-0.306813	0.406813	NO Sig
10	haya	pino	2.600	2.325	0.275	0.35681	-0.081813	0.631813	NO Sig

INTERVALOS DE CONFIANZA PAREADOS SEGÚN SIDACK:

	i	j	med(i)	med(j)	dif_med	Ampl	LI_Sidack	LS_Sidack	RES
1	arce	caoba	2.050	2.375	-0.325	0.33424	-0.659236	0.0092361	NO Sig
2	arce	cedro	2.050	2.375	-0.325	0.33424	-0.659236	0.0092361	NO Sig
3	arce	haya	2.050	2.600	-0.550	0.33424	-0.884236	-0.2157639	SIG
4	arce	pino	2.050	2.325	-0.275	0.33424	-0.609236	0.0592361	NO Sig
5	caoba	cedro	2.375	2.375	0.000	0.33424	-0.334236	0.3342361	NO Sig
6	caoba	haya	2.375	2.600	-0.225	0.33424	-0.559236	0.1092361	NO Sig
7	caoba	pino	2.375	2.325	0.050	0.33424	-0.284236	0.3842361	NO Sig
8	cedro	haya	2.375	2.600	-0.225	0.33424	-0.559236	0.1092361	NO Sig
9	cedro	pino	2.375	2.325	0.050	0.33424	-0.284236	0.3842361	NO Sig
10	haya	pino	2.600	2.325	0.275	0.33424	-0.059236	0.6092361	NO Sig

INTERVALOS DE CONFIANZA PAREADOS SEGÚN RYAN:

	i	j	med(i)	med(j)	dif_med	Amp1	RES
1	arce	pino	2.050	2.600	0.550	0.31516	SIG
2	arce	haya	2.050	2.375	0.325	0.30594	SIG
3	caoba	pino	2.325	2.600	0.275	0.30594	NO Sig
4	arce	cedro	2.050	2.375	0.325	0.29175	SIG
5	caoba	haya	2.325	2.375	0.050	0.29175	NO Sig
6	cedro	pino	2.375	2.600	0.225	0.29175	NO Sig
7	arce	caoba	2.050	2.325	0.275	0.26483	SIG
8	caoba	cedro	2.325	2.375	0.050	0.26483	NO Sig
9	cedro	haya	2.375	2.375	0.000	0.26483	NO Sig
10	haya	pino	2.375	2.600	0.225	0.26483	NO Sig

También la línea de comando

lybrary(stats)

TukeyHSD(aov1F)

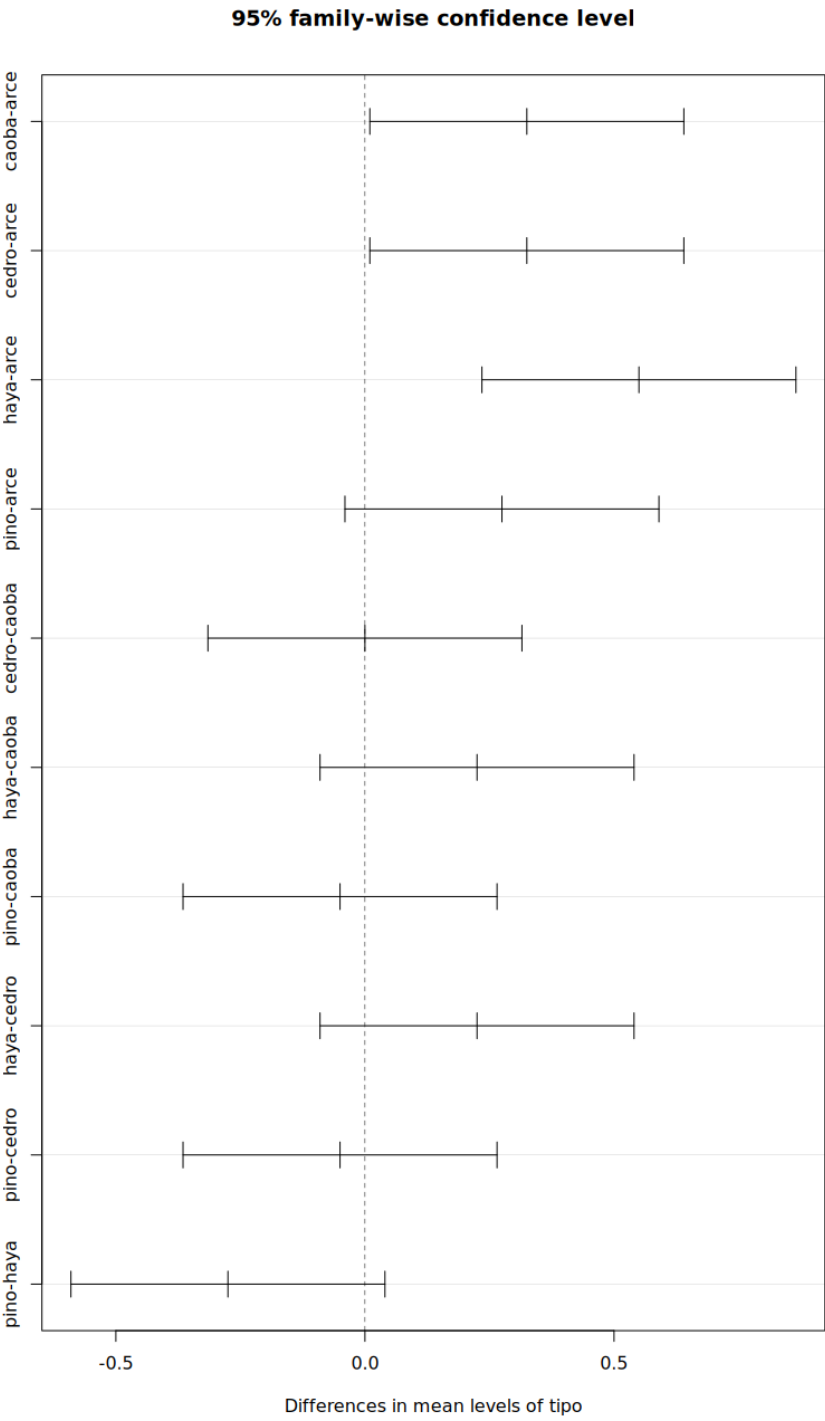
produce el siguiente cuadro (cuando se utiliza aov)

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = desgaste ~ tipo, data = madera)

\$tipo

	diff	lwr	upr	p adj
caoba-arce	0.325	0.009840027	0.64015997	0.0417456
cedro-arce	0.325	0.009840027	0.64015997	0.0417456
haya-arce	0.550	0.234840027	0.86515997	0.0006152
pino-arce	0.275	-0.040159973	0.59015997	0.1021412
cedro-caoba	0.000	-0.315159973	0.31515997	1.0000000
haya-caoba	0.225	-0.090159973	0.54015997	0.2304525
pino-caoba	-0.050	-0.365159973	0.26515997	0.9871310
haya-cedro	0.225	-0.090159973	0.54015997	0.2304525
pino-cedro	-0.050	-0.365159973	0.26515997	0.9871310
pino-haya	-0.275	-0.590159973	0.04015997	0.1021412



Segundo supuesto: Diseño de bloques completos aleatorizados.

Ejemplo: Cinco métodos de riego (cuenca, inundación, aspersión-a, aspersión-b, goteo) (basin, flood, spray, sprinkler, trickle) se implementan en cada uno de los ocho lugares (por ejemplo, partes de un campo que pueden tener las condiciones del suelo y luz similares), y se registra el peso de los frutos de cada uno de las cuarenta observaciones. Los datos están en “*frutas.RData*”.

Los comandos necesarios para analizar estos datos son:

```
#load("frutas.RData")
load(url("https://ddv.stic.ull.es/users/cpgonzal/public/Data/R/frutas.RData"))
str(frutas)
attach(frutas)
mod1F1B<-lm(formula(pesofruta~riego+bloque),data= frutas)
anova1F1B<-anova(object=mod1F1B)
print(anova1F1B) # Para obtener la tabla ANOVA completa
```

Los datos básicos de la tabla anova de este diseño es:

Analysis of Variance Table					
Response: pesofruta					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
riego	4	44026	11006	3.2734	0.02539 *
bloque	7	401308	57330	17.0503	1.452e-08 ***
Residuals	28	94147	3362		

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

Debemos proceder a analizar, mediante comparaciones múltiples los efectos del riego (factor) y de los lugares (bloques) que dejamos para analizar en el aula.

Tercer supuesto: Diseño factorial balanceado

Ejemplo: Se proporciona un ejemplo en el que hay dos variables de tratamiento (factores) –F1 cinco variedades de semilla (variedad) y F2 tres métodos de crecimiento (método), por lo que hay 15 condiciones experimentales (tratamientos). En este ejemplo, seis repeticiones (macetas) se realizan en cada condición, y el rendimiento (variable dependiente) se registró para cada uno de las 90 (5x3x6) observaciones. El conjunto de datos recogidos están en el data frame *semillas.RData*. Las instrucciones para realizar el análisis pueden ser

```
#load("semillas.RData")
load(url("https://ddv.stic.ull.es/users/cpgonzal/public/Data/R/semillas.RData"))
attach(semillas)
mod2F<-lm(formula(rendimiento~metodo*variedad),data= semillas)
anova2F<-anova(object=mod2F)
print(anova2F)
```

La tabla anova para este diseño es

Analysis of Variance Table

Response: rendimiento

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
metodo	2	953.16	476.58	24.2531	7.525e-09 ***
variedad	4	11.38	2.85	0.1448	0.96476
metodo:variedad	8	374.49	46.81	2.3822	0.02409 *
Residuals	75	1473.77	19.65		

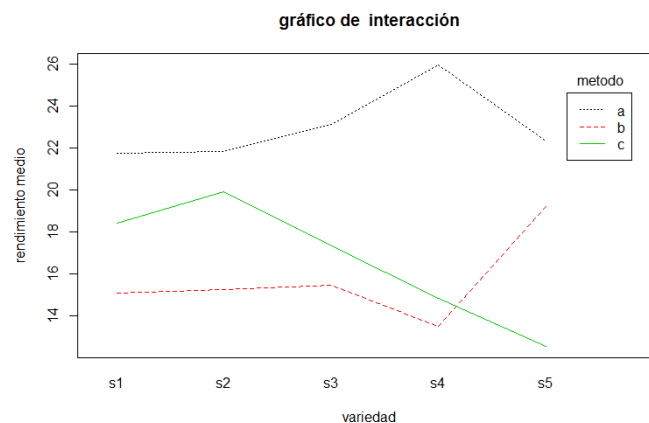
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

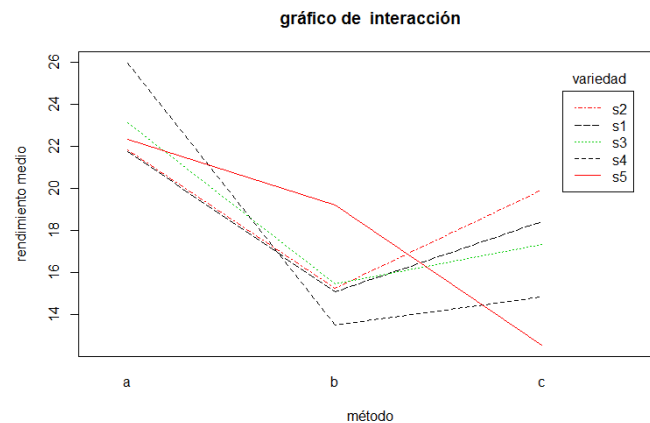
Tenemos al 5% de nivel de error que el efecto de interacción es significativo, no procede analizar cada efecto (Factor) por separado. Una idea de lo que esta ocurriendo la podemos tener observando las medias para cada tratamiento ensayado:

tapply(rendimiento, INDEX=list(metodo,variedad), FUN=mean)

	s1	s2	s3	s4	s5
a	21.76667	21.85000	23.13333	25.96667	22.33333
b	15.08333	15.23333	15.45000	13.50000	19.21667
c	18.41667	19.91667	17.31667	14.83333	12.55000

Aunque, mejor será una representación gráfica de las mismas, en los dos gráficos siguientes:





Podemos realizar todas las comparaciones pareadas, con un ajuste global según Bonferroni, con el siguiente comando:

```
pairwise.t.test(rendimiento, metodo:variedad, p.adjust.method="bonferroni")
```

cuya salida es

Pairwise comparisons using t tests with pooled SD

data: rendimiento and metodo:variedad

	a:s1	a:s2	a:s3	a:s4	a:s5	b:s1	b:s2	b:s3	b:s4	b:s5	c:s1	c:s2	c:s3	c:s4
a:s2	1.00000	-	-	-	-	-	-	-	-	-	-	-	-	-
a:s3	1.00000	1.00000	-	-	-	-	-	-	-	-	-	-	-	-
a:s4	1.00000	1.00000	1.00000	-	-	-	-	-	-	-	-	-	-	-
a:s5	1.00000	1.00000	1.00000	1.00000	-	-	-	-	-	-	-	-	-	-
b:s1	1.00000	1.00000	0.24971	0.00633	0.62193	-	-	-	-	-	-	-	-	-
b:s2	1.00000	1.00000	0.29768	0.00780	0.73293	1.00000	-	-	-	-	-	-	-	-
b:s3	1.00000	1.00000	0.38226	0.01054	0.92548	1.00000	1.00000	-	-	-	-	-	-	-
b:s4	0.19304	0.17464	0.03464	0.00063	0.09652	1.00000	1.00000	1.00000	-	-	-	-	-	-
b:s5	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	-	-	-	-	-
c:s1	1.00000	1.00000	1.00000	0.44487	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	-	-	-	-
c:s2	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	-	-	-
c:s3	1.00000	1.00000	1.00000	0.12116	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	-	-
c:s4	0.87734	0.80217	0.18547	0.00444	0.47070	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	-
c:s5	0.05946	0.05343	0.00961	0.00015	0.02842	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000	0.54652	1.00000	1.00000

P value adjustment method: bonferroni

EJERCICIOS:

1.- Para el fichero de datos “HIPER200.RData”, se solicita:

- ¿La tensión arterial sistólica final es la misma según la toma de sal?
- ¿La tensión arterial sistólica final es la misma según la actividad física?
- ¿La tensión arterial sistólica inicial es la misma según su clasificación del peso?
- ¿La tensión arterial sistólica inicial es la misma según su tiempo sin beber?

Si fuese el caso, indicar que grupos tienen medias diferentes.

2.- Para las variables incluidas en el data.frame iris (cargar en R como data(“iris”)) analizados por Pearson (1936). “The data set consists of 50 samples from each of three species of *Iris* ([Iris setosa](#), [Iris virginica](#) and [Iris versicolor](#)). Four [features](#) were measured from each sample: the length and the width of the [sepals](#) and [petals](#), in centimetres”.



Iris Setosa



Iris Versicolor



Iris virginica

Se pide realizar un análisis de la varianza para cada una de las variables incluidas en dichos datos.
(Fuente de fotografías wikipedia)

4.2 Estadística No Paramétrica

DESARROLLO DE LA PRÁCTICA

El software R dispone de varios procedimientos para realizar un test de adherencia de unos datos a una determinada distribución, en particular la normalidad de los datos. También realiza el contraste de independencia o de homogeneidad asociado a una tabla de contingencia. De otro lado se tiene otros programas para realizar un análisis estadístico de carácter no paramétrico a un conjunto de datos. Se presupone en este caso que la(s) variable(s) a analizar tiene una distribución desconocida pero de carácter continuo, al menos. Tomaremos de nuevo el conjunto de datos HIPER200.RData, por ser un conjunto no numeroso de datos y poder aplicarse todos los procedimientos.

Primer supuesto: Contraste de adherencia a una distribución determinada.

Cargamos el fichero de datos HIPER200.RData, para ejecutar dicho análisis, debemos ejecutar una serie de comandos:

```
#load("HIPER200.RData")
load(url("https://ddv.stic.ull.es/users/cpgonzal/public/Data/R/ HIPER200.RData "))
attach(HIPER200)
ks.test(TAsist0,"pnorm", 145,24) #test de Kolmogorov-Smirnov de 1 o 2 muestras
library(nortest)
lillie.test(TAsist0) # test de Lilliefors
pearson.test(TAsist0, n.classes=8) # test de bondad de ajuste segun Chi cuadrado
shapiro.test(TAsist0) # test de normalidad de Shapiro-Wilk
```

para realizar el contraste de Kolmogorov-Smirnov a la variable tensión arterial sistólica inicial (sist_ini) para una distribución normal indicando sus parámetros. Cargamos en memoria otro paquete (nortest) para realizar el test de Lilliefors sobre la misma variable anterior para el caso de la normalidad de los datos. El test de bondad de ajuste de Pearson y el de Shapiro y Wilk.

```
One-sample kolmogorov-Smirnov test

data:  TAsist0
D = 0.10252, p-value = 0.02988
alternative hypothesis: two-sided

One-sample kolmogorov-Smirnov test

data:  TAsist0
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided

Lilliefors (kolmogorov-Smirnov) normality test

data:  TAsist0
D = 0.099678, p-value = 5.113e-05

Pearson chi-square normality test

data:  TAsist0
P = 11.84, p-value = 0.03705
```

```
Shapiro-wilk normality test

data:  TAsist0
W = 0.96312, p-value = 4.367e-05
```

Segundo supuesto: Contraste de independencia y homogeneidad

Debemos seleccionar dos variables cualitativas y obtener una tabla de contingencia de 2xk o kx2

```
tab1<-table(genero,sal) #variables Genero y Sal
tab1
```

Genero	Sal		
	poca	normal	mucha
masculino	24	64	18
femenino	41	53	0

Aplicamos el test de independencia de la chi-cuadrado (test asintótico)

```
chisq.test(genero,sal,correct=FALSE)
```

Pearson's Chi-squared test				
data: tab3				
X-squared = 22.843, df = 2, p-value = 1.096e-05				
	X^2	df	P(> X^2)	
Likelihood Ratio	29.767	2	3.4372e-07	
Pearson	22.843	2	1.0960e-05	

```
library(vcd)
assocstats(tab1)
```

	X^2	df	P(> X^2)	
Likelihood Ratio	29.767	2	3.4372e-07	
Pearson	22.843	2	1.0960e-05	

Supongamos ahora que queremos dividir los individuos en 2 grupos (los de 40 años o menos, y los de más de 40 años). Creamos una variable dicotómica

```
grupo_edad<-ifelse(edad<=40,1,2) #recodificación de la edad en dos clases
grupo_edad <-factor(grupo_edad, labels=c("<=40",">40"))
```

Queremos comparar si hay relación entre esta variable ed y la variable concentración HTA (conc_hta). Por tanto, obtenemos la tabla de contingencia y vemos el resultado del contraste

```
table(grupo_edad,conc_hta) #variables ed y conc_hta
chisq.test(grupo_edad,conc_hta, correct=FALSE)
```

conc_hta					
ed	normotenso	bordeline	hipertenso		
<=40	64	12	3		
>40	68	8	45		
		X^2	df	P(> X^2)	
Likelihood Ratio	36.139	2	1.4209e-08		
Pearson	30.182	2	2.7926e-07		

Del mismo modo, también podemos hacer un contraste de homogeneidad para la igualdad de varias proporciones independientes desde una perspectiva asintótica. Por ejemplo, para comparar la proporción de mujeres de acuerdo a la cantidad de sal que consumen:

```
# test de homogeneidad
x<-c(41,53,0) # datos de genero="femenino" y Sal
n<-c(65,117,18)
prop.test(x,n, conf.level=0.95,correct=FALSE)
```

Tercer supuesto: Contrastar que la mediana de la distribución toma un valor determinado.

Para realizar el test ordinario de los signos, debemos indicar la variable y el valor de la mediana a contrastar

```
library(BSDA)
SIGN.test(TAsist0, md=150)
```

que nos proporciona

```
One-sample Sign-Test

data:  TAsist0
s = 66, p-value = 0.0002592
alternative hypothesis: true median is not equal to 150
95 percent confidence interval:
 140 148
sample estimates:
median of x
      140
```

Si utilizamos las instrucciones del cálculo de la función de distribución aplicado a la binomial, podemos determinar que la región crítica con $\alpha = 0.05$ es

$$H_1 : M_E(x) \neq 150 \text{ y } C = \{S \leq 77 \text{ o } S \geq 105\} \text{ pues } \sum_{k=0}^{77} \binom{182}{k} 0.5^{182} \leq 0.022527 \text{ y } \sum_{k=105}^{182} \binom{182}{k} 0.5^{182} \leq 0.022527$$

Mediante Wilcoxon de rangos con signo

```
wilcox.test(TAsist0, md=150, paired=FALSE, correct=FALSE, conf.int=TRUE)
```

```
wilcoxon signed rank test

data:  TAsist0
V = 20100, p-value < 2.2e-16
alternative hypothesis: true location is not equal to 0
95 percent confidence interval:
 139.9999 146.9999
sample estimates:
(pseudo)median
      143.5
```

Cuarto supuesto: Contrastar que la mediana de las diferencias pareadas es nula (muestras dependientes)

Utilizaremos las variables *TAsist0* y *TAsist1*, que representa la medida de HTA Sistólica a los individuos al comienzo del cuestionario y al finalizar este, un tiempo nunca superior a una hora.

```
SIGN.test(TAsist1,TAsist0, md=0)
```

```

Dependent-samples Sign-Test

data:  TAsist1 and TAsist0
S = 11, p-value < 2.2e-16
alternative hypothesis: true median difference is not equal to 0
95 percent confidence interval:
 -12 -10
sample estimates:
median of x-y
 -10
    
```

Si usamos de nuevo el cálculo de probabilidades de la Binomial con RSudio, podemos calcular

$$H_1 : M_D(x) \neq 0 \text{ y } C = \{S \leq 74 \text{ o } S \geq 103\} \text{ pues } \sum_{k=0}^{74} \binom{177}{k} 0.5^{177} \leq 0.017514 \text{ y } \sum_{k=103}^{177} \binom{177}{k} 0.5^{177} \leq 0.017514$$

Si tomamos las diferencias medida final menos medida inicial nuestro estadístico vale 11.

Utilizando el estadístico de rangos con signo de Wilcoxon

```
wilcox.test(TAsist1,TAsist0i, paired=TRUE, correct=FALSE)
```

tenemos

```

wilcoxon signed rank test

data:  TAsist1 and TAsist0
V = 650.5, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
    
```

Quinto supuesto: Contrastar que las medianas de las distribuciones son iguales en muestras independientes

Para este caso tomamos la variable TAsist0. Deseamos comparar la igual de medianas por genero, supuesto que ambas distribuciones son iguales salvo un desplazamiento. Para el test de Mann-Whitney-Wilcoxon

```
wilcox.test(TAsist0~genero, paired=FALSE, correct=FALSE, conf.int=TRUE)
```

```

wilcoxon rank sum test

data:  TAsist0 by genero
W = 4473, p-value = 0.2117
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -10.000001  1.999992
    
```

resultando un test no significativo.

Sexto supuesto: Contrastar que las medianas de las distribuciones son iguales en tres o más muestras independientes.

Tomamos la variable TAsist1 estando los grupos formados por la cantidad de Sal que se toma en las comidas. Para el test de Kruskal-Wallis los resultados son

```
kruskal.test(TAsist1~sal)
```

```

kruskal-wallis rank sum test

data:  TAsist1 by sal
kruskal-wallis chi-squared = 11.831, df = 2, p-value = 0.002698
    
```

EJERCICIOS:

Sin asumir una distribución específica de las variables implicadas se pide resolver:

1.- Para el fichero de datos “HIPER200.SAV”, se solicita:

- ¿La tensión arterial sistólica final es la misma según la clasificación de peso?
- ¿La tensión arterial sistólica final es la misma según la actividad física?
- ¿La tensión arterial sistólica inicial es la misma según su clasificación del peso?
- ¿La tensión arterial sistólica inicial es la misma según su tiempo sin beber?
- ¿La tensión arterial sistólica inicial y la tensión arterial sistólica final es la misma para las personas de menos de 30 años? ¿Y para los de más de 60 años?

2. En EEUU, en el curso de 2 años, se ha producido un notable incremento de incidentes en los que la policía ha disparado y abatido a personas. Los datos siguientes (obtenidos del Washington Post) corresponden a la distribución de un total de 1,642 personas de los diversos grupos de población y atendiendo a si fueron abatidos (es decir, recibieron un disparo fatal) portando armas o se encontraban desarmados.

		Situación de la persona abatida		Total
		Armados	Desarmados	
Raza	Blanca	403	54	457
	Hispana	282	25	307
	Negra	825	53	878
	Total	1510	132	1642

- Analizar si existe relación entre la raza y la posibilidad de recibir un disparo estando desarmado. Interpretar el estadístico del test de independencia (qué grupo es el que destaca sobre el resto) para $\alpha=0.05$. Acotar el p-valor e interpretarlo.
- Calcular un intervalo de confianza del 95% en la proporción de personas desarmadas.