



Tema 6 - Unidad 2: Análisis Estadístico

2.1 Diferencia Estadística
2.2 Correlación

Interacción y Visualización de la Información
Pablo.Bermejo@uclm.es

Análisis Estadístico

1



Introducción

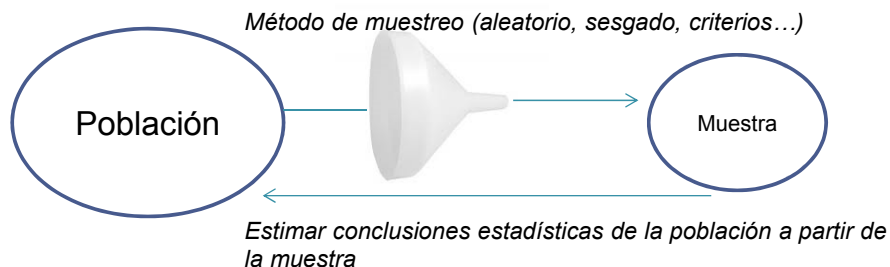
- La **estadística descriptiva** se refiere a las distintas formas de organizar y presentar la información:
 - Gráficos
 - Tablas
 - Resumen textual
- El **análisis estadístico** nos permite inferir conclusiones, presumiblemente correctas, acerca de la/s muestra/s que hemos obtenido de una población:
 - Comparación de métodos
 - Correlación
- La **estadística predictiva** se utiliza para crear modelos, a partir de nuestra muestra, capaces de predecir futuras mediciones.

Análisis Estadístico

2

Introducción

- Cualquier tipo de estadística se ejecuta sobre una **Muestra**, que es un subconjunto aleatorio de observaciones de la **Población**.



- Llamamos **variable** a un conjunto de observaciones (sean el total de la población o el subconjunto de la muestra)
- La cardinalidad del conjunto de casos u observaciones que contiene la variable se conoce como **tamaño muestral**.

Introducción

- **Parámetro:** valor calculado sobre la población entera
- **Estadístico:** valor calculado sobre la muestra
- Normalmente siempre trabajamos con estadísticos, esperando que las conclusiones sean fieles a la población. En este caso el estadístico también se conoce como **estimador**.

2.1 Diferencia estadística

- El tipo de conclusión estadística que más nos suele interesar es comparar dos muestras, obtenidas de la misma población con diferentes métodos, y saber si sus medias son estadísticamente iguales o diferentes.
- Esto es un proceso que pertenece al **contraste de hipótesis**.
- Tenemos 2 hipótesis:
 - H_0 : $\mu_1 = \mu_2$ las medias son iguales (**hipótesis nula**)
 - H_1 : $\mu_1 \neq \mu_2$ las medias son distintas (**hipótesis alternativa**), se dice que 'se rechaza la hipótesis nula'.

2.1 Diferencia estadística

- Cuando rechazamos la hipótesis nula sin tener que rechazarla, cometemos un **Error Tipo I** → Falso Positivo: afirmamos algo y nos equivocamos
- Cuando aceptamos la hipótesis nula sin tener que aceptarla, cometemos un **Error de Tipo II** → Falso negativo: negamos algo y nos equivocamos
- El **nivel de confianza** α es la probabilidad de cometer un Error de Tipo I. Así cuanto menor es, el test es más restrictivo y es más difícil encontrar diferencias, pero estas son más fuertes.
 - $\alpha = 0.05$ → 5% probabilidad de encontrar una diferencia estadística donde no la hay.
 - $\alpha = 0.01$ → 1% probabilidad de encontrar una diferencia estadística donde no la hay.
- Cada test genera un valor **p-value**, en esta asignatura no nos interesa cómo.
- Si $p\text{-value} < \alpha$, rechazamos la hipótesis nula. Es decir, aceptamos H_1 : $\mu_1 \neq \mu_2$



2.1.1 Definiciones



2.1.1 Diferencia estadística - definiciones

Definiciones:

- Supongamos que calculamos el volumen prostático de pacientes, divididos en 2 muestras X1 y X2 según diferentes criterios:
 - X1: pacientes entre 30 y 50 años
 - X2: pacientes mayores de 50 años
 - Lo que cambia es el método de muestreo, no la variable medida: → **test no pareado**
- Supongamos que calculamos el volumen prostático de una muestra de pacientes, en 2 momentos distintos de su vida
 - X1: antes del tratamiento
 - X2: después del tratamiento
 - Lo que cambia es el método de obtener el estadístico, pero la muestra poblacional es la misma: → **test pareado**

2.1.1 Diferencia estadística - definiciones

Test no pareado:

X1	X2
4.5	3.9
3.4	5.2
5.6	4.5
3.8	4.7
4.7	4.1
5.2	4.4
6.2	5.6
4.2	6.2
5.3	5.6
4.5	
4.8	
2.7	
4.58	4,9

Test pareado:

X1	X2
4.5	4.8
3.4	3.6
5.6	5.9
3.8	5.1
4.7	4.5
5.2	5.2
6.2	6.1
4.2	4.2
5.3	5.0
4.5	4.6
4.8	4.4
2.7	2.8
4.58	4.68

2.1.1 Diferencia estadística - definiciones

- Normalmente, del estadístico calculado de una muestra nos interesa saber solo:
 - si es mayor: test de **una cola (derecha) one tail**
 - si es menor: test de **una cola (izquierda)**
- Si se encuentra diferencia con una cola, también se encontrará al hacer un test de 2 colas, ya que mira que sea mayor por cualquiera de los 2 lados (es menos restrictivo).

X1	X2
4.5	4.8
3.4	3.6
5.6	5.9
3.8	5.1
4.7	4.5
5.2	5.2
6.2	6.1
4.2	4.2
5.3	5.0
4.5	4.6
4.8	4.4
2.7	2.8
4.58	4.68

¿Es $4.68 > 4.58$?

Test pareado de una cola (derecha)

Si se acepta la hipótesis nula, entonces son estadísticamente iguales.


¿Es $4.68 \neq 4.58$? Comparar con 2 colas, que es la salida por defecto en los tests de SPSS.

Cuestión de ética científica: hay que decidir el tipo de tests antes de ver las 2 medias; es decir, al diseñar el experimento, ¿sabemos **seguro** que una va a ser algo mayor o menor?



2.1.1 Diferencia estadística - definiciones

- Dependiendo de la distribución de la muestra, tendremos que decidir el tipo de test.
 - Distribución normal: **paramétrico**
 - No normal: **no paramétrico**
- Suele asumirse que una distribución es normal a partir de un tamaño muestral >30 , pero lo correcto utilizar **tests de normalidad**:
 - Kolmogorov-Smirnov: SPSS
 - Shapiro Wilk: R



2.1.2 Tipos de tests

Usaremos:

- SPSS. Todas las imágenes usadas son de su GUI.
- R (desde el cliente Rstudio). Podéis cargar las bases de datos en .csv usando el comando `data <- read.csv("ruta")`

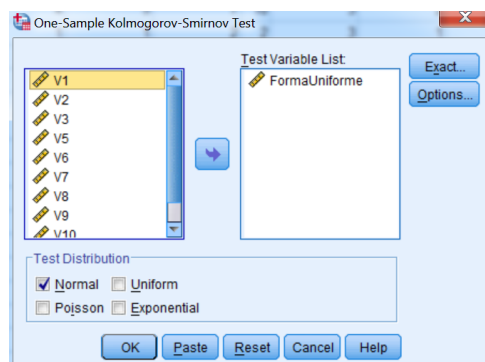
2.1.2 Diferencia estadística – tests

Tests de normalidad

- Utilizando una base de datos que describe tejidos de tumor obtenidos del pecho de pacientes:
- <http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29>
- Antes de cargarla, añade 11 nombres de variable en la primera línea para cuando lo cargues en R.
- La cuarta variable indica la uniformidad del tejido. ¿Sigue esta variable una distribución normal?
- Shapiro Wilk en R: `shapiro.test(data$V4)` siendo V4 el nombre dado a la cuarta variable.
- Kolmogorov-Smirnov en SPSS: **Analyze → non-parametric tests → Legacy Dialogs → 1-sample K-S**

El test de Shapiro es considerado más fiable.

2.1.2 Diferencia estadística – tests



Prueba de Kolmogorov-Smirnov para una muestra		
		Forma Uniforme
N		699
Parámetros normales ^{a, b}	Media	3,21
	Desviación típica	2,972
Diferencias más extremas	Absoluta	,276
	Positiva	,276
	Negativa	-,229
Z de Kolmogorov-Smirnov		7,302
Sig. asintót. (bilateral)		,000
a. La distribución de contraste es la Normal.		
b. Se han calculado a partir de los datos.		

P-value= 0

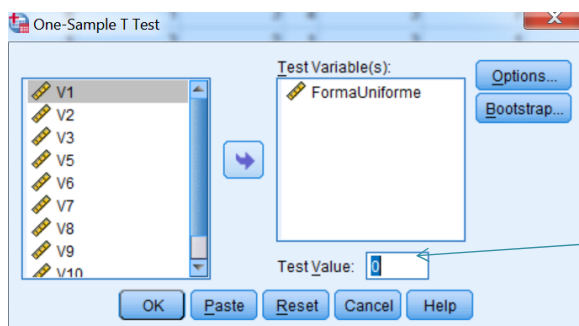
Nivel de confianza = 0.05 por defecto

P-value < α , se rechaza la hipótesis nula → no sigue distribución normal, al contrario de lo que nos podríamos haber esperado por tener >30 casos.

2.1.2 Diferencia estadística – tests

t-test de 1 muestra

- Cuando lo que nos interesa no es saber si nuestra muestra sigue una distribución normal, sino que la media de nuestra muestra es significativamente igual (o distinta) a otra media, entonces utilizamos t-test de 1 muestra.
- Ej: se dice que la población de los estudiantes del Grado en Informática asisten 5 horas diarias de clase de media. Cogemos una muestra de 10 estudiantes y obtenemos una media de 7. ¿Ha sido por casualidad? ¿He seleccionado a los que más trabajadores o es que la media de 5 horas era incorrecta?
- **En R:** `t.test(vector de datos, mean=la media a comparar)`
- **Analyze → compare means → One-sample T test...**



Media contra la que comparar

2.1.2 Diferencia estadística – tests

Test para comparar la media de 2 muestras paramétricas y dependientes

T-test pareado

- Abrid el archivo dietstudy.sav, que puede encontrarse en el subdirectorío Samples en la ruta de instalación de SPSS.
- Las variables wgt0 y wgt4 indican el peso de pacientes antes y después de un tratamiento, y queremos ver si el tratamiento ha funcionado.
- El test será pareado, porque cada medición se ha hecho sobre el mismo paciente.
- Primero, comprueba que cada variable sigue una distribución normal. Con el test de K-S.
- Si el p-value es mayor que 0.05, no se rechaza la hipótesis y por lo tanto son normales.
- **En R:** `t.test(data$wgt0,data$wgt4,paired=TRUE)`
- Luego realiza el test en **Analyze → Compare Means → Paired-Samples T Test...**

2.1.2 Diferencia estadística – tests

Estadísticos de muestras relacionadas

	Media	N	Desviación tip.	Error típ. de la media
Par 1 Weight	198,38	16	33,472	8,368
Final weight	190,31	16	33,508	8,377

Prueba de muestras relacionadas

		Diferencias relacionadas					t	gl	Sig. (bilateral)
		Media	Desviación tip.	Error tip. de la media	95% Intervalo de confianza para la diferencia				
					Inferior	Superior			
Par 1	Weight - Final weight	8,063	2,886	,722	6,525	9,600	11,175	15	,000

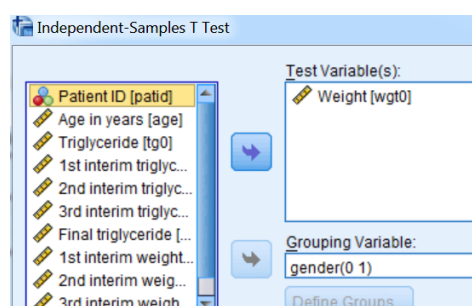
- $0.00 < 0.05 \rightarrow$ se rechaza la hipótesis nula: las medias son distintas
 \rightarrow el tratamiento ha funcionado
- Si se quisiera hacer de una cola, dividimos el p-value entre 2, pero como es 0 no hace falta.

2.1.2 Diferencia estadística – tests

- Test para comparar la media de 2 muestras paramétricas no pareadas

T-test no pareado

- Ahora imaginemos que queremos comparar la media de peso original entre sexos: las mediciones ya no son dependientes
- Hay que dividir las mediciones de la variable wgt0 según los valores de la variable categórica gender={male,female}
- **En R:** `t.test(data[data$gender==0,]$wgt0,data[data$gender==1,]$wgt0, paired==FALSE)`
- **Analyze \rightarrow Compare Means \rightarrow Independent-Samples T Test...**



¿son diferentes las medias?

2.1.2 Diferencia estadística – tests

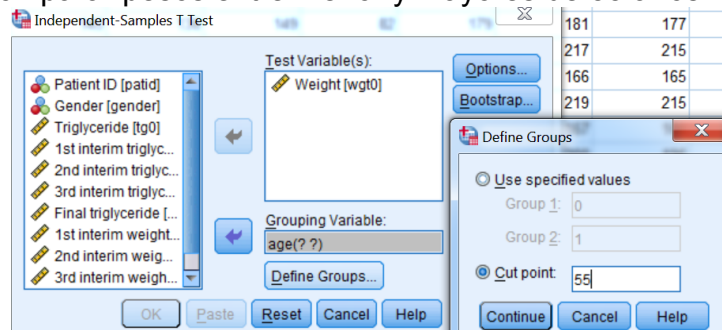
Estadísticos de grupo					
	Gender	N	Media	Desviación tip.	Error tip. de la media
Weight	Male	9	223,78	18,754	6,251
	Female	7	165,71	10,935	4,133

Prueba de muestras independientes										
		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error tip. de la diferencia	95% Intervalo de confianza para la diferencia	
Weight	Se han asumido varianzas iguales	1,521	,238	7,255	14	,000	58,063	8,003	40,898	75,229
	No se han asumido varianzas iguales			7,748	13,168	,000	58,063	7,494	41,895	74,232

- Ojo, aunque la variable original sea gaussiana, al dividirla es posible que las submuestras no lo sean, y además el revisor del artículo puede decirnos que un test paramétrico con tamaños muestrales tan pequeños no es significativo.

2.1.2 Diferencia estadística – tests

- Si la variable en función de la cual queremos dividir el peso fuera numérica, puede indicarse un punto de corte
- Ej: comparar pesos entre menor y mayores de 55 años.



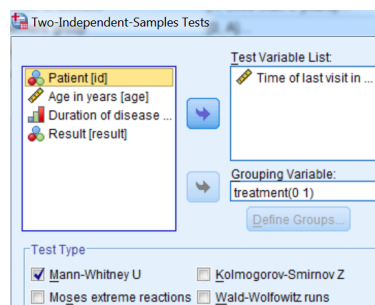
- En R:** `t.test(data[data$age<=55],$wgt0,data[data$age>55],$wgt0, paired==FALSE)`
- ¿Se obtiene diferencia estadística?

2.1.2 Diferencia estadística – tests

- [Test para comparar la media de 2 muestras no-paramétricas y no pareadas](#)

Mann-Whitney test (U test) = Wilcoxon Rank Sum test

- Abramos el archivo ulcer_recurrence.sav, también en el subdirectorio Samples en la ruta de instalación de SPSS.
- Comprobamos que la variable *time* (tiempo desde la última visita) no sigue una distribución normal.
- Comprobamos si el tiempo que un paciente tarda en volver a la consulta médica depende del tipo de tratamiento.
- **En R:** `wilcox.test(data[data$treatment=="A"],$time, data[data$treatment=="B"],$time,paired=FALSE)`
- **Analyze → non-parametric tests → Legacy Dialogs → 2 Independent samples**



¿Existe diferencia?

Análisis Estadístico

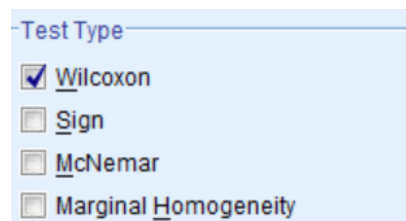
21

2.1.2 Diferencia estadística – tests

- [Test para comparar la media de 2 muestras no-paramétricas y pareadas](#)

Wilcoxon Signed-Rank test

- **En R:** `wilcox.test(x,y,paired=TRUE)`
- En **Analyze → non-parametric tests → Legacy Dialogs → 2 related samples**
- El más comúnmente utilizado es el de los rangos con signo de Wilcoxon,

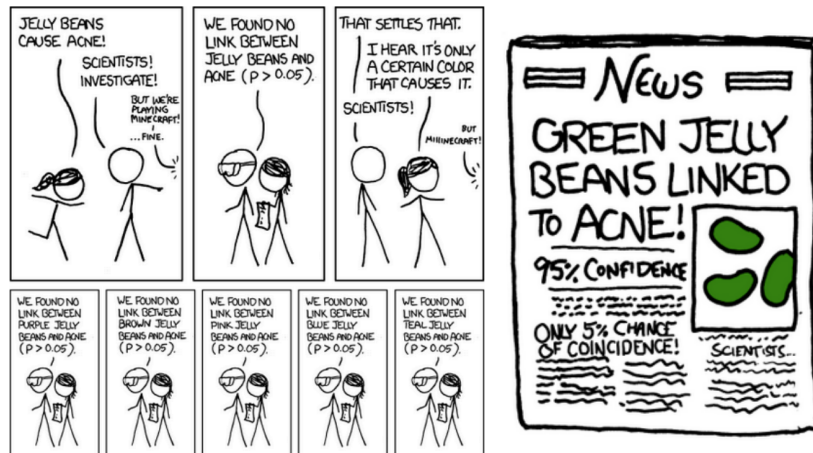


Análisis Estadístico

22

2.1.2 Diferencia estadística – tests

- ¿Qué ocurre si queremos comparar qué media es más alta con más de 2 series? A, B, C:
 - Si obtenemos $A > B$ y $B > C$, ¿Podemos decir que $A > C$?
 - Si obtenemos $A > B$ y $A > C$, ¿Podemos decir que A es la mayor?
- La probabilidad del Error de Tipo I, si es un 5%, quiere decir que cometeremos seguro un error tras 20 comparaciones; quizás antes!



<http://xkcd.com/882/>

Análisis Estadístico

23

2.1.2 Diferencia estadística – tests

- Hay que comparar todas las series a la vez.
 - En este caso, la hipótesis nula es $H_0: \mu_A = \mu_B = \mu_C$
 - Y la hipótesis alternativa es que al menos una igualdad se rompe.
 - Cuantas más series se comparan a la vez, más difícil es encontrar diferencia estadística.
- Hay que ejecutar 2 fases:
 - 1) Test estadístico para más de 2 muestras
 - 2) Si se rechaza H_0 , entonces se realiza un test post-hoc, para comprobar qué serie/s es/son distinta/s.
- De nuevo el test a ejecutar depende de la normalidad de los datos.

Análisis Estadístico

24

2.1.2 Diferencia estadística – tests

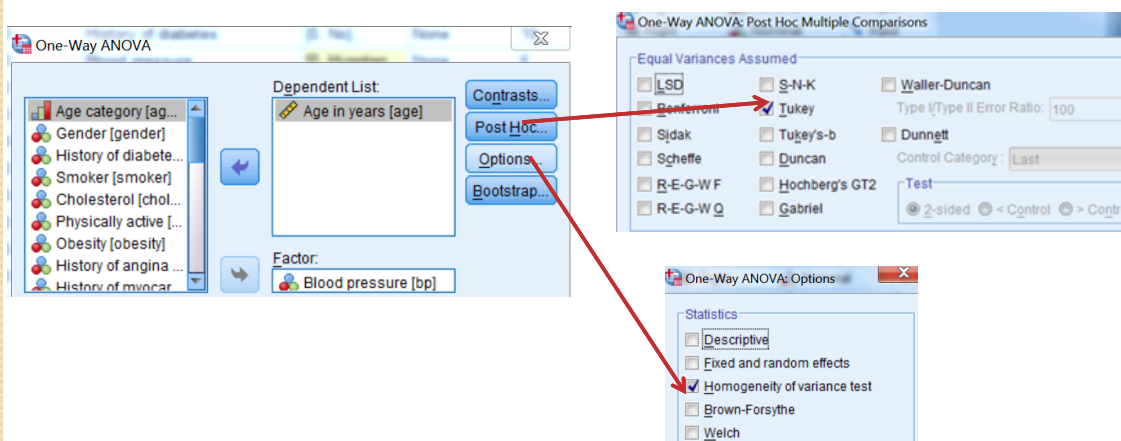
- Test para comparar 3 ó más muestras no pareadas, con distribución normal

ANOVA

- Además de normalidad, requiere varianzas homogéneas
- El mismo test ANOVA permite en sus opciones realizar un test de varianzas homogéneas (Levene's test).
- El factor por el que se divide la variable cualitativa para crear las series debe tener 3 ó más valores.
- SPSS no proporciona test para muestra pareadas normales

2.1.2 Diferencia estadística – tests

- Abrid el archivo de prueba de SPSS patient_los.sav, que contiene datos de pacientes ingresados ante la posibilidad de sufrir un infarto de miocardio.
- Queremos comprobar si la edad influye en el tipo de presión arterial (variables *age* y *bp*)
- **En R:** `summary(aov(data$age ~ data$bp))` (convertir *bp* a factor antes)
- **Analyze → Compare Means → One way ANOVA...**



2.1.2 Diferencia estadística – tests

Prueba de homogeneidad de varianzas

Age in years

Estadístico de Levene	gl1	gl2	Sig.
1,696	2	9997	,184

En R: library(lawstat)
levene.test(data\$age,data\$bp)

ANOVA

Age in years

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	2421,076	2	1210,538	15,126	,000
Intra-grupos	800059,339	9997	80,030		
Total	802480,414	9999			

El test post-hoc de Tukey nos divide las muestras que son estadísticamente Distintas, y une las que son iguales.

En R: TukeyHSD(objeto devuelto por el comando aov)

Age in years

Tukey B^{a, b}

Blood pressure	N	Subconjunto para alfa = 0.05	
		1	2
Normal	6134	62,03	
Hypertension	2659	62,46	
Hypotension	1207		63,56

2.1.2 Diferencia estadística – tests

Si seleccionamos el de Bonferroni, tenemos comparación a pares.

Nosotros elegimos el valor de control de la variable factor

Pruebas post hoc

Comparaciones múltiples

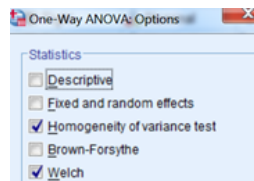
Age in years
Bonferroni

(I) Blood pressure	(J) Blood pressure	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
Hypotension	Normal	1,529 [*]	,282	,000	,85	2,20
	Hypertension	1,096 [*]	,310	,001	,35	1,84
Normal	Hypotension	-1,529 [*]	,282	,000	-2,20	-,85
	Hypertension	-,432	,208	,112	-,93	,07
Hypertension	Hypotension	-1,096 [*]	,310	,001	-1,84	-,35
	Normal	,432	,208	,112	-,07	,93

*. La diferencia de medias es significativa al nivel 0.05.

2.1.2 Diferencia estadística – tests

- Si el test de Levene devuelve un p-valor significativo, entonces los resultados de ANOVA no nos sirven porque no se cumple la homogeneidad de varianzas.
- Podemos pedir entonces, en el menú de opciones de ANOVA, la **Pruebas Robustas de Igualdad de Medias**, haciendo click en el test de Welch:



En R: `oneway.test(data$age~data$bp,var.equal=FALSE)`

- Los resultados se interpretan igual que con ANOVA.

Pruebas robustas de igualdad de las medias

Age in years

	Estadístico ^a	gl1	gl2	Sig.
Welch	14,504	2	3035,873	,000

Análisis Estadístico

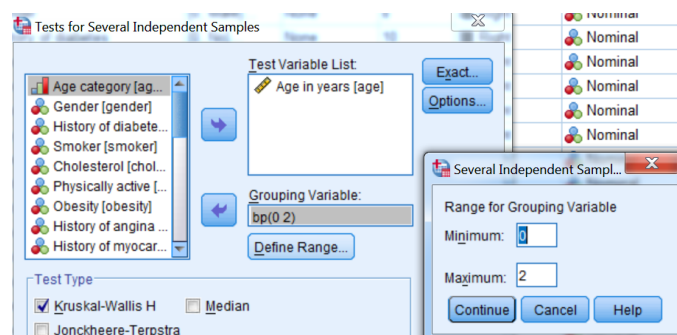
29

2.1.2 Diferencia estadística – tests

Test no paramétrico para comparar 3 ó más muestras no pareadas

Kruskal – Wallis

- Si no se cumplen las condiciones de ANOVA
 - Normalidad de la distribución y/o
 - Homogeneidad de varianzas
- Pasamos a un test no-paramétrico: Kruskal-Wallis (admite variables ordinales además de numéricas)
- La variable *age* no es gaussiana! (hacer test de normalidad)
- **En R:**
- **Analyze → non parametric tests → Legacy dialogs → k independent samples...**



30

2.1.2 Diferencia estadística – tests

- SPSS no ofrece tests post-hoc para tests no-paramétricos de k muestras.
- Pero sí **Prueba de Kruskal-Wallis** estimación.

Rangos			
	Blood pressure	N	Rango promedio
Age in years	Hypotension	1207	5373,27
	Normal	6134	4906,04
	Hypertension	2659	5049,20
	Total	10000	

Estadísticos de contraste^{a, b}

	Age in years
Chi-cuadrado	27,478
gl	2
Sig. asintót.	,000

- La edad cuando la presión es normal dista mucho de cuando es 'hypotension'.
- **En R:** `library(pgirmess); kruskalmc(data$age~data$bp, cont="two-tailed")`
Toma como control para la comparación el primer nivel de la variable categórica. Se pueden reordenar, por ejemplo al revés: `data$bp = factor(data$bp, levels(data$bp)[c(3,2,1)])`

2.1.2 Diferencia estadística – tests

- [Test para comparar 3 ó más muestras pareadas, con distribución no normal.](#)

Friedman

- El más común es el test de Friedman. Normalmente éste iría seguido de un post-hoc de Holm, pero ya hemos dicho que SPSS no ofrece directamente post-hocs para tests no paramétricos.
- **Analyze → Nonparametric tests → Legacy Dialogs → k related samples...**
- **En R:** `friedman.test(matriz)` donde la matriz tiene todos los valores en cada fila correspondientes a un mismo caso.
Y el post hoc: `pairwise.t.test(valores, grupo, p.adj="holm", paired=T)`

2.1.2 Diferencia estadística – tests

- Test para comparar frecuencias cruzadas de 2 variables nominales, con muestras no pareadas

Prueba Z para comparación de proporciones

- A partir de 2 variables categóricas, comparar las proporciones de frecuencia de aparición de cada par de valores cruzados de las 2 variables.
- Abriendo el archivo insurance_claims.sav, que contiene información sobre avisos al seguro
- Queremos comparar qué tipo de avisos (*claim_type*) son más frecuentes cuando el aviso es fraudulento (*fraudulent*).
- **Analyze → Tables → Custom Tables...**

2.1.2 Diferencia estadística – tests

Custom Tables

Table Titles Test Statistics Options

Compare column means (t-tests) Alpha: 0,05

Compare column proportions (z-tests) Alpha: 0,05

Variables:

Claim ID [claim...]
Date of incident...
Type of claim [c...]
Property uninha...
Cost of claim in...
Fraudulent clai...
Policy ID [policy...]
Date policy wen...
Amount of cove...
Deductible [ded...]
Size of hometo...
Gender [gender]
Date of birth [do...]
Level of educati...
Employment st...
Retired [retire]
Household inc...
Marital status f...

Colums

		Type of claim				
		Wind/Hail	Water ...	Fire/Smoke	Contaminati...	Theft/Vandali...
		Column N %	Column N %	Column N %	Column N %	Column N %
Fraudulent claim	No	nnn.n%	nnnn.n%	nnnn.n%	nnnn.n%	nnnn.n%
	Yes	nnnn.n%	nnnn.n%	nnnn.n%	nnnn.n%	nnnn.n%

Rows

Define

Summary Statistics

Summary Statistics... Categories and Totals...

Position: Columns Hide Category Position: Default

Source: Column Variables

Indicar que muestre los porcentajes de las columnas

2.1.2 Diferencia estadística – tests

		Type of claim				
		Wind/Hail	Water damage	Fire/Smoke	Contamination	Theft/Vandalism
		% del N de la columna	% del N de la columna	% del N de la columna	% del N de la columna	% del N de la columna
Fraudulent claim	No	91,4%	92,0%	88,5%	93,6%	86,4%
	Yes	8,6%	8,0%	11,5%	6,4%	13,6%

Comparaciones de proporciones de columnas^a

		Type of claim				
		Wind/Hail	Water damage	Fire/Smoke	Contamination	Theft/Vandalism
		(A)	(B)	(C)	(D)	(E)
Fraudulent claim	No	E	E		C E	
	Yes			D		A B D

Los resultados se basan en pruebas bilaterales con un nivel de significación 0.05. Para cada par significativo, la clave de la categoría con la proporción de columna menor aparece debajo de la categoría con mayor proporción de columna.

- Ojo, las comparaciones son entre columnas, no entre filas.
- *‘Cuando el cliente miente, casi siempre lo hace respecto a robos’*

2.1.2 Diferencia estadística – tests

- En R:

```
#crear tabla
cruces <- table(data$fraudulent,data$claim_type)
#prueba z
fraud <- c(cruces[2,1],cruces[2,2],cruces[2,3],cruces[2,4],cruces[2,5])
total <- c(sum(cruces[,1]),sum(cruces[,2]),sum(cruces[,3]),sum(cruces[,4]),sum(cruces[,5]))
prop.test(fraud,total)
```

Y si $p\text{-value} < \text{nuestro nivel de confianza}$, entonces lanzamos un test post-hoc
`pairwise.prop.test(t(tabla))`

2.1.2 Diferencia estadística – tests

- Si queremos encontrar diferencias entre los casos fraudulentos y verdaderos, invertimos la tabla, y obtenemos:

Comparaciones de proporciones de columnas^a

		Fraudulent claim	
		No	Yes
		(A)	(B)
Type of claim	Wind/Hail	B	
	Water damage	B	
	Fire/Smoke		
	Contamination	B	
	Theft/Vandalism		A

- ‘Cuando el tipo de aviso es por daño de viento, agua o contaminación, el porcentaje de los que mienten es menor que el de los que dicen la verdad.’
- ‘Cuando el tipo de aviso es por robo, mienten más que dicen la verdad’

2.1.2 Diferencia estadística – tests

- Se pueden anidar más variables categóricas

Custom Tables

Table Titles Test Statistics Options

Variables:

Columns

				Fraudulent claim	
				No	Yes
				Column N %	Column N %
Gender	Male	Type of claim	Wind/Hail	nnnn.n%	nnnn.n%
			Water ...	nnnn.n%	nnnn.n%
			Fire/Smoke	nnnn.n%	nnnn.n%
			Contaminati...	nnnn.n%	nnnn.n%
			Theft/Vandal...	nnnn.n%	nnnn.n%
	Female	Type of claim	Wind/Hail	nnnn.n%	nnnn.n%
			Water ...	nnnn.n%	nnnn.n%
			Fire/Smoke	nnnn.n%	nnnn.n%
			Contaminati...	nnnn.n%	nnnn.n%
			Theft/Vandal...	nnnn.n%	nnnn.n%

Comparaciones de proporciones de columnas^a

				Fraudulent claim	
				No	Yes
				(A)	(B)
Gender	Male	Type of claim	Wind/Hail		
			Water damage		
			Fire/Smoke		
			Contamination		
			Theft/Vandalism		A
	Female	Type of claim	Wind/Hail	B	
			Water damage		
			Fire/Smoke		
			Contamination	B	
			Theft/Vandalism		A

‘Cuando una mujer denuncia por contaminación o viento, suele ser cierto’
Esta conclusión no puede obtenerse para Los hombres

2.1.2 Diferencia estadística – tests

- Test para saber si las veces que una variable nominal obtiene un valor sigue una frecuencia determinada.
- Imagina que tiras una moneda 30 veces, y 20 te sale cara. Quieres comprobar si esa moneda está sesgada hacia cara, o ha sido fruto del azar y es justa; es decir, la probabilidad real de ser cara sigue siendo 50%.

- **Con R** es muy sencillo:

```
binom.test(x=20, n =30, p=0.5)
```

¡La moneda estaba trucada!

2.1.2 Diferencia estadística – tests

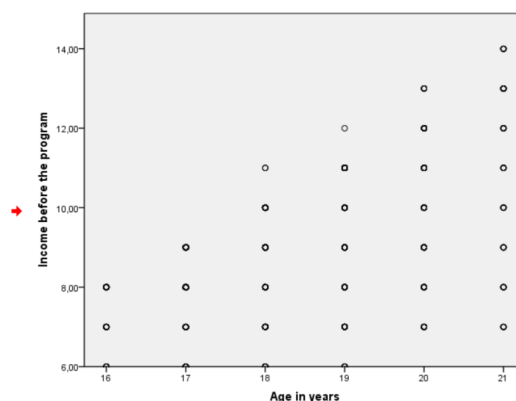
Cuadro Resumen

Quiero comprobar...	Test	Paramétrico
Distribución normal	Kolmogorov-Smirnov	No
Distribución normal con una media dada	T test de 1 muestra	Sí
Diferencia en la media de 2 muestras normales y pareadas	T test pareado	Sí
Diferencia en la media de una variable normal dividida en 2 según el valor de otra variable	T test no pareado	Sí
Diferencia en la media de 2 muestras pareadas	Wilcoxon Signed-Rank	No
Diferencia en la media de una variable dividida en 2 según el valor de otra variable	Mann-Whitney o Wilcoxon Rank Sum	No
Diferencia en la media de una variable normal dividida en K series según el valor de otra variable - varianzas homogéneas	ANOVA + post-hoc	Sí
Diferencia en la media de una variable normal dividida en K series según el valor de otra variable - varianzas NO homogéneas	Welch+ post-hoc	Sí
Diferencia en la media de una variable numérica u ordinal, dividida en K series según el valor de otra variable	Kruskal - Wallis	No
Diferencia en la media de K muestras	Friedman	No
Diferencia en proporciones de 2 variables categóricas cruzadas	Prueba Z	-

2.2 Correlación

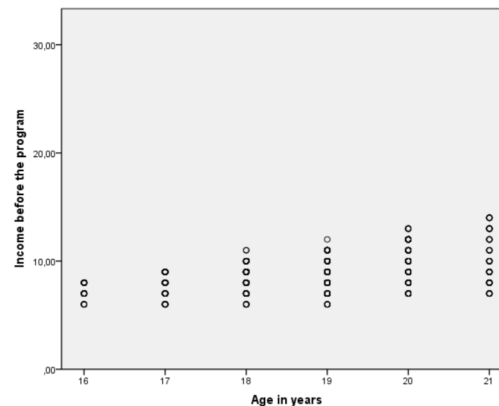
2.2 Correlación

- Decimos que dos variables A y B son independientes, si los valores que toma una no dependen de la otra.
- **Correlación:** grado de dependencia lineal entre dos variables
- Visualmente, es fácil intuir una dependencia lineal mediante diagramas de dispersión.
- Abrimos el archivo *workprog.sav*, y hacemos un gráfico de dispersión (scatter) con las variables *income* y *age*.



2.2 Correlación

- En ocasiones por culpa de la granularidad (zoom) del gráfico, no puede verse una correlación.
- Si repetimos el anterior indicando que el eje Y tenga una escala de 0 a 30:



- Se ve más claramente una relación lineal.
- Pero, ¿podemos fiarnos sabiendo que la apariencia puede *trucarse*?

2.2.1 Correlación – numéricas y ordinales

- Es preferible realizar tests de correlación, y así fiarnos de una métrica *objetiva*.
- Métricas más conocidas de correlación lineal
 - Correlación de **Pearson**: variables numéricas normales
 - Correlación **ro de Spearman** (ρ): variables numéricas no normales o con outliers, y variables ordinales.
 - Correlación **tau de Kendall** (τ): igual que el anterior, pero su coeficiente además puede interpretarse como la probabilidad de predecir correctamente una variable a partir de la otra.
- Las 3 métricas se encuentran en el rango $[-1, 1]$
 - 0: variables totalmente independientes (linealmente)
 - -1: correlación negativa perfecta. Cuando una aumenta la otra disminuye, y viceversa.
 - 1: correlación positiva perfecta. Cuando una aumenta la otra aumenta, y viceversa.

2.2.1 Correlación – numéricas y ordinales

- Usando *workprog.sav*
- **Analyze → Correlate → Bivariate...**

Pearson

Correlaciones		Age in years	Income before the program
Age in years	Correlación de Pearson	1	,526**
	Sig. (bilateral)		,000
	N	1000	1000
Income before the program	Correlación de Pearson	,526**	1
	Sig. (bilateral)	,000	
	N	1000	1000

** La correlación es significativa al nivel 0,01 (bilateral).

Correlación Lineal Positiva

Correlación Significativa
p-value<0.05

Si no fuera significativa, nos daría
Igual que el coeficiente de correlación fuese alto

Spearman

Correlaciones			Age in years	Income before the program
Rho de Spearman	Age in years	Coefficiente de correlación	1,000	,486**
		Sig. (bilateral)	.	,000
		N	1000	1000
	Income before the program	Coefficiente de correlación	,486**	1,000
		Sig. (bilateral)	,000	.
		N	1000	1000

** La correlación es significativa al nivel 0,01 (bilateral).

Estadístico

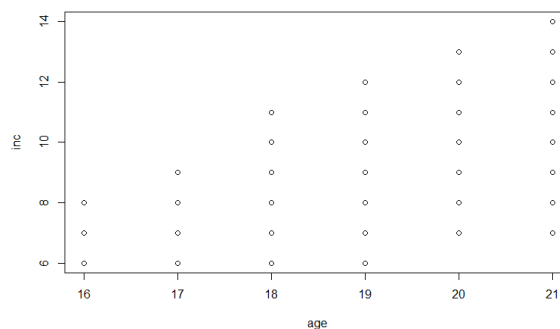
45

2.2.1 Correlación – numéricas y ordinales

- **En R:**

#dibujar gráfico de dispersión

```
plot(data$age,data$incbef)
```



#tests de correlación

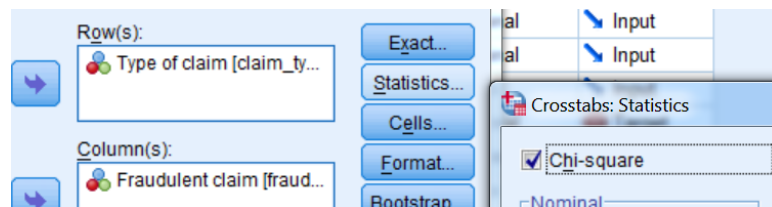
```
cor.test(data$age, data$incbef,method="pearson",conf.level=0.95)
```

```
cor.test(data$age, data$incbef,method="spearman",conf.level=0.95)
```

```
cor.test(data$age, data$incbef,method="kendall",conf.level=0.95)
```

2.2.2 Correlación - categóricas

- Cuando quiere encontrarse correlación entre variables categóricas, utiliza los siguientes tests para saber **si existe o no** dependencia:
 - **Chi-Cuadrado - χ^2 (test de independencia. H_0 : vars. independientes)**
 - **En R:** `chisq.test(table(data$fraudulent,data$claim_type))`
 - **Analyze → Descriptive Statistics → CrossTables...**



Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	29,996 ^a	4	,000
Razón de verosimilitudes	30,485	4	,000
Asociación lineal por lineal	14,672	1	,000
N de casos válidos	4415		

Análisis Estadístico

47

2.2.2 Correlación - categóricas

- **Fisher Exact**, se utiliza en lugar del Chi-cuadrado cuando:
 - para un cruce de categorías hay menos de 5 casos,
 - cuando las cantidades de casos son muy dispares entre celdas de la tabla creada.
- **En R:** `fisher.test(table(data$fraudulent,data$claim_type),conf.int=0.95workspace=1E8)`
(se indica el tamaño del workspace solo para tablas mayores de 2x2)
- En SPSS se hace por defecto cuando ejecutamos el Chi-cuadrado, pero solo si las tablas son de 2x2 (variables binomiales). Por eso en el test anterior no aparece.
- Calculemos el chi-cuadrado con las variables *gender* y *fraudulent*.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)	Sig. exacta (bilateral)	Sig. exacta (unilateral)
Chi-cuadrado de Pearson	2,820 ^a	1	,093		
Corrección por continuidad ^b	2,657	1	,103		
Razón de verosimilitudes	2,824	1	,093		
Estadístico exacto de Fisher				,095	,051
Asociación lineal por lineal	2,819	1	,093		
N de casos válidos	4415				

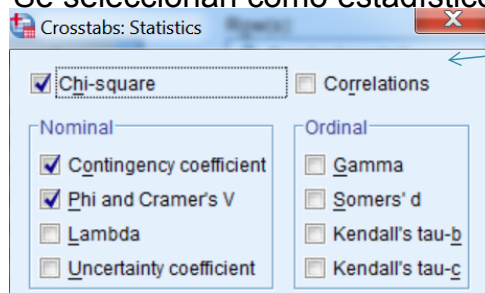
a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 228,09.

b. Calculado sólo para una tabla de 2x2.

48

2.2.2 Correlación - categóricas

- Para saber cómo de **fuerte** es esta **correlación**, podemos calcular los siguientes coeficientes,.
 - **Coeficiente de Contingencia – CC**: poco aconsejable ya que cuanto más valores tienen las variables, más grande es el coeficiente. . Rango [0,1]
 - **Coeficiente Phi - ϕ** : solo para variables binomiales (tablas 2x2). Rango [-1,1]. Cerca de +1 indica que el estado 1 de la variable se relaciona con el estado 1 de la otra. Cerca de -1 lo mismo para los estados 0.
 - **Cramer - V**: para variables con más de 2 estados. Rango [0,1]
- **Recuerda: fuerza, NO SENTIDO DE LA CORRELACIÓN. Ya que no puede existir un sentido en variables categóricas. Para diferencia de proporciones, usar prueba Z.**
- Se seleccionan como estadísticos a la vez que el chi-cuadrado.



(para forzar test de spearman como si fueran ordinales)

Y vemos que la correlación es muy floja

Análisis Estadístico

49

2.2.2 Correlación - categóricas

- **En R:**

```
tabla <- table(data$fraudulent,data$claim_type)
```

```
#chi-square independence test
```

```
chisq.test(tabla)
```

```
#fuerza de la correlación
```

```
library(vcd)
```

```
assocstats(tabla)
```

Análisis Estadístico

50

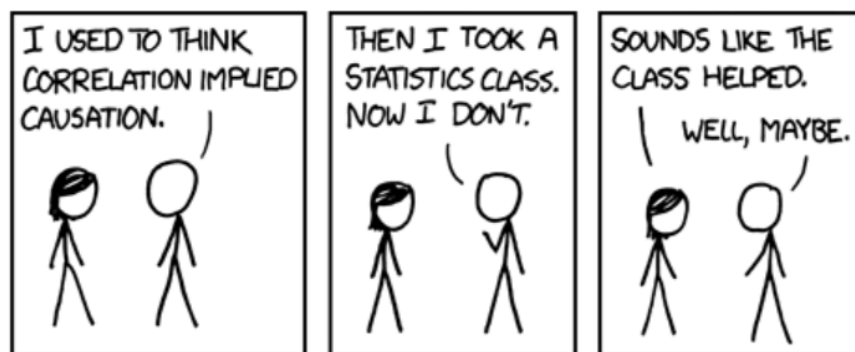
2.2.2 Correlación - categóricas

TABLA RESUMEN DE PRUEBAS DE CORRELACIÓN

Tipo variables	Test de correlación	Fuerza de la correlación
Numéricas normales	Pearson	-
Numéricas no normales	Spearman	-
Numéricas con outliers	Spearman	-
Ordinales	Spearman o tau de Kendall	-
Categóricas que crean tablas con celdas <5casos	Fisher's Exact	Variables binomiales: Phi Variables multinomiales: Cramer's V
Categóricas que crean tablas con celdas de cantidad de casos dispares	Fisher's Exact	
Otras categóricas	Chi-Cuadrado	

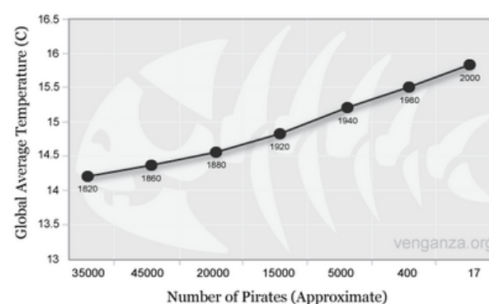
2.2.2 Correlación - categóricas

- Correlación no implica causalidad



Pirates & global warming:

Global Average Temperature Vs. Number of Pirates



2.2.3 Correlación, concordancia, homogeneidad

Correlación vs concordancia

- Muchos investigadores confunden estos términos.
- La correlación se mide entre variables distintas: peso vs sexo, para comprobar si el valor de una está relacionada con el de la otra.
- La concordancia se mide entre diferentes muestras de la misma variable, obtenidas con algún método distinto, y siendo una de ellas la de referencia.
- Ej: Sabemos que el aparato A es el más fiable para medir la tensión. Pero ha llegado el aparato B que es más barato. Queremos comprobar si las mediciones de B son fiables, es decir, concordantes con A.
 - **Coeficiente de concordancia W de Kendall:**
 - Rango [0-1]
 - **En R:** `library(irr); kendall(matriz nxm) # n sujetos, m mediciones`
 - SPSS no lo calcula
 - **Coeficiente de concordancia Kappa de Cohen (κ):**
 - Rango: [-1.1]: 0 ninguna concordancia. 1 total. -1, menos concordancia aún que por azar.
 - Numéricas o categóricas
 - **En R:** `library(vcd); kappa(datos en formato table)`
 - **Analyze → Descriptive Statistics → CrossTables... estadístico kappa**

2.2.3 Correlación, concordancia, homogeneidad

Test Z vs Test χ^2 de homogeneidad:

- Utilizamos la prueba Z para comparar proporciones de columnas al cruzar 2 variables categóricas: calidad de vida según tasa de mortalidad.
- La homogeneidad se calcula cuando queremos comparar el valor que toma 1 variable categórica o numérica respecto a varias poblaciones: sensación de temperatura según intervalo de edad.
- Se calcula igual que el test de independencia de Chi-cuadrado, siendo una variable utilizada para dividir en poblaciones diferenciadas. Y la otra es numérica y categórica y se quiere comprobar si cambia con la primera. Es cuestión de lenguaje, matemáticamente es lo mismo.



Conclusiones

- Para decidir qué tipo de prueba hacer, decide con qué encaja más tu objetivo:
 - **Diferencia estadística:** comprobar si 2 ó más muestras tienen una media distinta.
 - **Correlación:** comprobar si 2 variables son dependientes:
 - Diagrama de dispersión: dependencia lineal, curvi-línea o ninguna
 - Coeficientes: dependencia lineal
 - **Concordancia:** tenemos un medidor de referencia que sabemos que es correcto, y queremos saber si otro medidor es igual de fiable.