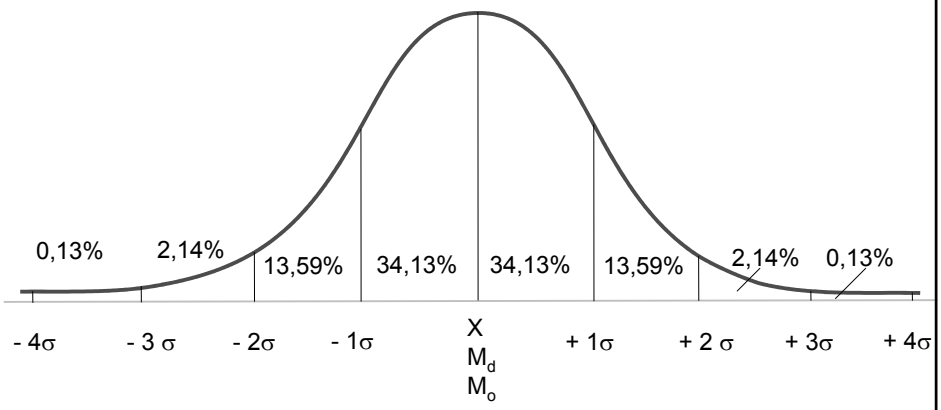
Propiedades de la distribución *«Normal»*



- Es simétrica y media, mediana y moda coinciden en el punto central
- Si añadimos el valor de la desviación típica y lo restamos a la media entre ambos extremos queda comprendido el 68.26% de los sujetos.
- Si sumamos o restamos 2 σ a la media el universo comprendido entre estos extremos es de 97.7% . Con 3 σ será el 99.9% del universo.
  - ➡ Con σ ±1,96 σ queda comprendido entre ambos valores el 95%

Estandarizar puntuaciones la *«z»*

- Para poder comparar puntuaciones de dos sujetos en distintas distribuciones o de un sujeto en distintas variables se utiliza la puntuación estandarizada basada en puntuarlo en unidades de desviación estándar.

$$Z = \frac{\text{Unidades que se desvía de su media}}{\textit{Desviación típica}}$$

➡ Así la puntuación n de un sujeto en una distribución de media «μ» y desviación «σ» será:

$$Z = \frac{n - \mu}{\sigma}$$

Ejemplo: Un sujeto que ha puntuado 95 en una distribución de media = 100 y desviación típica =15 tendría una puntuación estándar z de (100-95)/15 = - 0,67

- La puntuación z además de permitir comparar a sujetos en diferentes distribuciones tiene propiedades muy interesantes al tener de μ =0 y σ =1 y distribuirse de forma normal
  - ➡ Los sujetos que puntúan más de cero están por encima de la media y viceversa
  - ➡ Un sujeto que puntué 1 dejaría por debajo de él a más del 76,02%

Ejemplo: Tenemos dos series de notas correspondientes a dos asignaturas diferentes de un grupo de alumnos:

4 $\boxed{5}$ 5 6 7 8 7 7 6 6 7 5 9 10 8 6 7 8 3 8 7 μ=6,61 σ=1,65

1 $\boxed{5}$ 6 3 2 0 6 7 9 8 10 1 3 5 3 4 7 8 3 2 4 μ= 4,61 σ=2,81

Si quisiéramos saber si el segundo sujeto (comenzando por la derecha) que ha calificado con 5 en ambas asignaturas ha sacado una puntuación equivalente tendríamos que estandarizar ambas calificaciones:

$$z_1 = \frac{5-6{,}61}{1{,}65} = -1{,}0 \quad z_2 = \frac{5-4{,}61}{2{,}81} = 0{,}01$$

Como vemos ambas calificaciones no son equivalentes, pues mientras un 5 en la primera asignatura tiene por encima el 76% de las calificaciones de la clase. En la otra asignatura es una medida muy cercana a la media de la clase

Una vez estandarizadas todas las puntuaciones en ambas asignaturas éstas si son comparables y nos indican si los sujetos están por debajo o encima de la media (valores negativos o positivos) y que posición ocupan en la distribución:

-1,6 -1,0 -1,0 -0,4 0,2 0,8 0,2 0,2 -0,4 -0,4 0,2 -1,0 1,4 2,0 0,8 -0,4 0,2 0,8 -2,2 0,8 0,2

-1,3 0,1 0,5 -0,6 -0,9 -1,6 0,5 0,8 1,6 1,2 1,9 -1,3 -0,6 0,1 -0,6 -0,2 0,8 1,2 -0,6 -0,9 -0,2

Análisis de los residuos en T. Contingencia

- Utiliza las ideas de Ji-cuadrado para aplicarlas no al estudio de la tabla global, sino a cada una de las parejas de categorías de la misma.
  - ***«El residuo»*** Como en Ji-cuadrado es la diferencia entre frecuencias observadas y esperadas $R_{ij} = O_{ij} - E_{ij}$
  - ***«El residuo tipificado»*** Elimina los efectos que sobre el residuo tengan los valores marginales de ambas variables dividiendo los residuos por la raíz cuadrada de las frecuencias esperadas $SR_{ij} = \frac{R_{ij}}{\sqrt{E_{ij}}}$ por lo que $= \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$
  - ***«El residuo ajustado»*** Estandariza los valores de los residuos tipificados dividiendo por la varianza estimada

$$AR_{ij} = \frac{SR_{ij}}{\sqrt{V_{ij}}} \text{ donde } V_{ij} = \left(1 - \frac{O_{i.}}{n}\right) \times \left(1 - \frac{O_{.j}}{n}\right)$$

- Los residuos así ajustados de Haberman (1978). Tienen una distribución normal con μ= 0 y σ =1 por lo que si son mayores en valor absoluto a ±1,96 tienen un 95% de posibilidades de no deberse al azar y ser significativos

Un ejemplo de «Residuos ajustados»

- Partimos de un grupo de 220 sujetos, divididos en tres grupos de edad en torno al voto emitido a diferentes ámbitos políticos. La distribución de frecuencias observadas es:

	Jóvenes	Adultos	Mayores	Total
Izquierda	20	40	20	80
Centro	20	10	10	40
Derecha	20	40	40	100
Total	60	90	70	220

- Calculamos las esperadas para cada celda «ij» multiplicando el valor de los marginales de filas y columnas dividido por el total de casos

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n}$$

$$E_{11} = \frac{60 \times 80}{220} = 21{,}82... \quad E_{12} = \frac{90 \times 80}{220} = 32{,}72$$

...

$$E_{21} = \frac{60 \times 40}{220} = 10{,}91... \quad E_{22} = \frac{90 \times 40}{220} = 16{,}36$$

	Jóvenes	Adultos	Mayores	Total
Izquierda	21,82	32,73	25,45	80
Centro	10,91	16,36	12,73	40
Derecha	27,27	40,91	31,82	100
Total	60	90	70	220

- Se calculan a continuación los residuos restando a las frecuencias observadas las esperadas:

$$R_{ij} = O_{ij} - E_{ij}$$

$$R_{11} = 20 - 21{,}82 = -1{,}82...$$
$$R_{21} = 20 - 10{,}91 = 9{,}09...$$

	Jóvenes	Adultos	Mayores
Izquierda	-1,82	7,27	-5,45
Centro	9,09	-6,36	-2,73
Derecha	-7,27	-0,91	8,18

- Se calculan los residuos tipificados dividiendo los anteriores por la raíz cuadrada de las frecuencias esperadas

$$SR_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}} = \frac{R_{ij}}{\sqrt{E_{ij}}}$$

$$SR_{12} = \frac{-1{,}82}{\sqrt{21{,}82}} = -038925$$

	Jóvenes	Adultos	Mayores
Izquierda	-0,38925	1,27128	-1,08112
Centro	2,75241	-1,57313	-0,76447
Derecha	-1,39262	-0,14213	1,45048

- Se ajustan finalmente para normalizar su distribución mediante:

$$AR_{ij} = \frac{SR_{ij}}{\sqrt{V_{ij}}} \text{ de dónde } = \frac{SR_{ij}}{\sqrt{\left(1-\frac{O_{i..}}{n}\right)\times\left(1-\frac{O_{.j}}{n}\right)}}$$

$$AR_{11} = \frac{-0{,}5722}{\sqrt{\left(1-\frac{60}{220}\right)\times\left(1-\frac{80}{220}\right)}} = -0{,}6 \ldots AR_{12} \frac{2{,}0731}{\sqrt{\left(1-\frac{90}{220}\right)\times\left(1-\frac{80}{220}\right)}} = 2{,}1$$

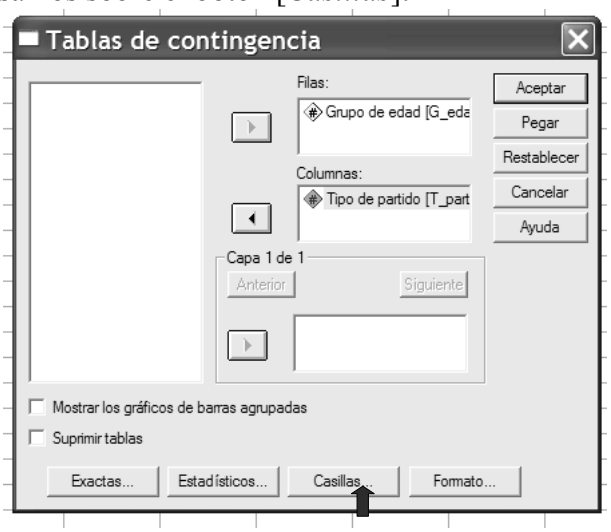
$$AR_{21} = \frac{-0{,}5722}{\sqrt{\left(1-\frac{60}{220}\right)\times\left(1-\frac{40}{220}\right)}} = 3{,}6 \ldots AR_{22} \frac{2{,}0731}{\sqrt{\left(1-\frac{90}{220}\right)\times\left(1-\frac{40}{220}\right)}} = -2{,}3$$

	Jóvenes	Adultos	Mayores
Izquierda	-0,6	2,1	-1,6
Centro	3,6	-2,3	-1,0
Derecha	-2,2	-0,3	2,4

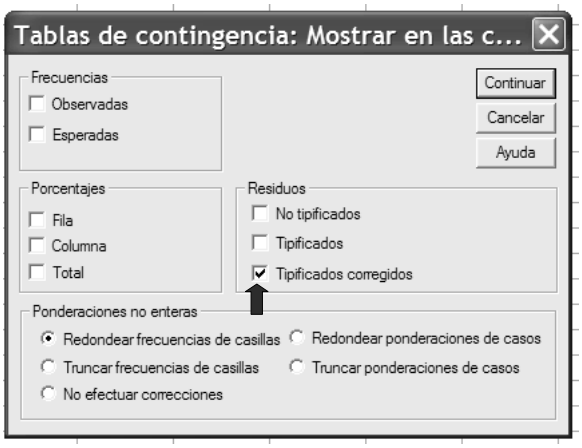
- Al distribuirse normalmente con μ=0 y σ=1 sabemos que los valores superiores en valor absoluto a ±1,96 dejan tras ellos el 95% de los casos, por tanto son significativos a un nivel α=0,05 las siguientes parejas de categorías:
  - Los jóvenes y adultos votan significativamente más opciones de centro e izquierda
  - Los mayores votan significativamente más las opciones de derechas

Ejemplo con el SPSS

- Con los datos introducidos entramos en *«Analizar»* → *«Estadíscos descriptivos»* → *«Tablas de contingencia»*
  - Allí pulsamos sobre el botón [*Casillas*]:



- *Activamos las casillas de residuos «Tipificados corregidos»*



- En la pantalla de resultados nos muestra los residuos corregidos *«SR_{ij}»* que coinciden con los calculados antes manualmente

➜ **Tablas de contingencia**

Resumen del procesamiento de los casos

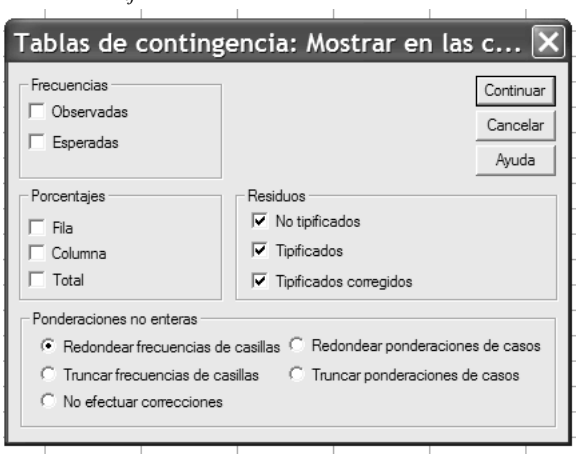
	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Grupo de edad * Tipo de partido	220	100,0%	0	,0%	220	100,0%

Tabla de contingencia Grupo de edad * Tipo de partido

Residuos corregidos

		Tipo de partido		
		Izquierda	Centro	Derecha
Grupo de edad	Jóvenes (18/25 años)	-,6	2,1	-1,6
	Adultos (25/45 años)	3,6	-2,3	-1,0
	Mayores > 65 años	-2,2	-,3	2,4

- Si activamos los tres tipos de residuos como muestra el gráfico el programa nos mostrara los Residuos *«R_{ij}»,* los Residuos tipificados *«AR_{ij}»* y los Residuos ajustados *«SR_{ij}»*



- La opciones d *«Tablas* de contingencia» en el SPSS nos permiten además obtener frecuencias observadas, esperadas o porcentajes de columnas o filas

➜ Tablas de contingencia

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
Grupo de edad * Tipo de partido	220	100,0%	0	,0%	220	100,0%

Tabla de contingencia Grupo de edad * Tipo de partido

			Tipo de partido		
			Izquierda	Centro	Derecha
Grupo de edad	Jóvenes (18/25 años)	Residuo	-1,8	7,3	-5,5
		Residuos tipificados	-,4	1,3	-1,1
		Residuos corregidos	-,6	2,1	-1,6
	Adultos (25/45 años)	Residuo	9,1	-6,4	-2,7
		Residuos tipificados	2,8	-1,6	-,8
		Residuos corregidos	3,6	-2,3	-1,0
	Mayores > 65 años	Residuo	-7,3	-,9	8,2
		Residuos tipificados	-1,4	-,1	1,5
		Residuos corregidos	-2,2	-,3	2,4

Medidas de asociación *(variables no métricas)*

- Podemos definir la asociación entre dos variables como la intensidad con la que unas categorías de una variable diferencian las frecuencias obtenidas en el cruce con la otra
  - Una primera mediada podría ser la diferencia de porcentajes Para Sánchez Carrión, J. (1995) es la mejor de todos ellas.

	M	V	Total
Opción A	15	35	50
Opción B	35	15	50
Total	50	50	100

En la tabla hay un diferencial de 20% entre Mujeres y Varones entre las opciones A y B

El diferencial porcentual varía entre: 0 <*d*> 100

El problema es que hay que calcularlo para cada casilla, de ahí que se busque un indicador único

  - El Ji-cuadrado además de determinar si son significativas estadísticamente las diferencias ya constituye por si mismo un indicador, su problema es que el valor no es estándar, depende de las frecuencias y del tamaño de la tabla

(a)

30	20	50
20	30	50
50	50	100

(b)

60<	40	50
40	60	50
100	100	200

En ambas tablas existe la misma relación un diferencia porcentual del 10% solo que la b tiene el doble de frecuencia y sus Ji-cuadrados:

$\chi_a^2 = 4{,}0$ $\chi_b^2 = 8{,}0$

- Para evitar estos problemas del Ji-Cuadrado se utilizan algunas modificaciones:
  - El ***«Phi»*** consiste en hacer la raíz cuadrada del Ji-Cuadrada dividida por el número total de casos de la tabla a fin de eliminar el problema de las frecuencias altas

$$Ø = \sqrt{\frac{x^2}{n}}$$

Su valor oscila entre 0 y 1 y es igual al coeficiente de correlación de Pearson para tablas de 2x2, pero si la tabla es mayor no tiene máximo

  - El ***«Coeficiente de contingencia»*** Intenta solucionar ese problema poniendo en el denominador de la fórmula de Phi la suma de χ^2+n

$$C = \sqrt{\frac{x^2}{x^2 + n}}$$

Plantea a su vez el problema de que nunca llega a valer 1 ni siquiera con asociación perfecta en tablas cuadradas (igual número de filas y columnas *«I»*) su valor máximo es:

$$C_{máximo} = \sqrt{(I-1)/I}$$

Por lo que se puede calcular un C ajustado de la siguiente forma: $C_{ajus}=C/C_{max}$

  - El ***«Coeficiente V de Cramer»*** Sustituye en el denominador de ***«Phi»*** el valor mínimo de (I-1) o (J-1)

$$V = \sqrt{x^2 / \text{mínimo de (I-1) o (J-1)}}$$

Asociación. *Indicadores basados en la reducción de error de predicción*

- A diferencia de los anteriores basados en Ji-cuadrado. Estos tratan de ver la relación entre variables intentando predecir como se clasifica un sujeto en la variable «*Y*» a partir de conocer su clasificación en la «*X*»

Coeficiente Lambda de Goodman y Kruskal

- Llamado también «*Coeficiente de predictibilidad de Guttman*» se basa en la reducción de error en la predicción conociendo la distribución de una variable bajo la fórmula:

$$\lambda_{yx} = \frac{(N - M_y) - (N - \sum m_y)}{N - M_y} = \frac{M_y - \sum m_y}{N - M_y}$$

Siendo:
M_y = la frecuencia modal global
M_y = la suma de frecuencias modales
N = Total de casos

El numerador sería pues el número de no errores cometidos bajo la predicción II (conociendo la distribución de segunda variable) que es igual a la diferencia de los errores de la predicción I (sin conocer la distribución) menos los de la predicción II. Al dividir por la predicción I me debe dar una cifra entre 0 ninguna reducción (independencia total ya que una variable no predice la otra o 1 si la puede predecir de forma total.

- Tras el hundimiento del Titanic de las 1285 personas que viajaban en él perecieron 800 y murieron 485 en función del sexo la distribución fue:

	V	M	Total	%
Mueren	637	163	800	62,3
Sobreviven	138	347	485	37,7
Total	775	510	1285	100
%	60,3	39,7		

Si pretendo acertar el destino de un pasajero cualquiera, sin saber nada más, me aventuraría por decir que murió, ya que fueron mayoría los que perecieron (intervalo modal) y tendría una posibilidad de errar de M_y=485

➡ Sabiendo que es hombre la posibilidad de que fallara mi pronóstico sería m_1= 138 Por el contrario si se que es mujer la posibilidad de errar es m_2=163. El error al conocer la distribución de la segunda variable es menor que si no la conozco. Aplicando la formula de Lamda:

$$\lambda_{yx} = \frac{(N - M_y) - (N - \sum m_y)}{N - M_y} = \frac{M_y - \sum m_y}{N - M_y}$$