ANÁLISIS DESCRIPTIVO Y EXPLORATORIO DE DATOS

José Manuel Rojo
Laboratorio de Estadística
Instituto de Economía y Geografía
Consejo Superior de Investigaciones Científicas
Madrid, 22 a 26 de Junio de 2006

ÍNDICE

Página

1- INTRODUCCIÓN

Uno de los objetivos de la Estadística es el de describir en unas pocas medidas resumen las principales características de un amplio conjunto de datos, de forma que estas medidas reflejen lo más fielmente las principales peculiaridades de dicho conjunto. A esta rama de la Estadística se la denomina Estadística Descriptiva.

Otro de los objetivos de la Estadística es realizar conjeturas acerca de las medidas resumen de un conjunto de datos conociendo tan sólo una parte del mismo; esta rama se denomina Estadística Inferencial.

2- POBLACIÓN Y MUESTRA

Población

Una población, desde un punto de vista estadístico, es un conjunto perfectamente definido de objetos. Por ejemplo: la población de ciudadanos españoles en un determinado intervalo de tiempo, los empleados de una empresa,…, etc. En cualquier caso, dado un objeto deberemos de tener una regla que determine sin ambigüedades de ningún tipo si un determinado objeto pertenece o no a la población.

Elementos de la población

Son todos y cada uno de los objetos de que esta constituida una población.

Muestra

Una muestra es un subconjunto de una determinada población; por ejemplo 10 trabajadores de una empresa constituyen una muestra de los trabajadores de la misma. Si la muestra abarca a toda la población se la denomina **censo**. En caso de que tratemos de conocer las principales características de una población a través de una muestra de la misma, esta muestra deberá de cumplir determinados requisitos respecto de su tamaño y el método de selección de sus elementos.

Variable de estudio

Es la característica a observar en cada uno de los elementos de la población o muestra. Se suele representar por una letra, p. e., X, o bien por un nombre corto y descriptivo, p. e. Edad.

Distribución de frecuencias de la variable

Son los valores observados y sus frecuencias relativas o absolutas.

Método de selección de la muestra

Es la técnica utilizada para seleccionar los elementos de la muestra. Los métodos de selección a utilizar deberán garantizar que la muestra sea representativa de la población.

Ejemplo

Consideremos la cuestión de determinar el número medio de personas por hogar en la Comunidad de Madrid.

Esta claro que la población de interés es la formada por los hogares de la Comunidad de Madrid, pero primero deberemos definir qué es un hogar; a continuación deberemos crear una regla para determinar si una persona vive o no en dicho hogar. En general estas cuestiones no tienen una solución sencilla y distintas reglas darán lugar a distintos resultados.

Enumeramos los distintos elementos:

- **Población**: la formada por todos los hogares de la Comunidad de Madrid.
- **Elementos de la población**: todos y cada uno de los hogares de la Comunidad de Madrid.
- **Muestra**: un determinado subconjunto de hogares.
- **Variable**: es la característica observada en cada elemento de la muestra en este caso será el número de personas que viven en cada hogar examinado.
- **Distribución de la variable**: son los valores observados y sus frecuencias.
- **Método de selección**: es el método empleado para seleccionar los hogares que van a ser observados, por ejemplo muestreo aleatorio simple sin reposición.

3- NIVELES DE MEDIDA DE LAS VARIABLES

En la práctica, la opción de un método estadístico depende en gran parte de la naturaleza de las observaciones que vayamos a realizar.
A continuación se muestran ordenados de menor a mayor los distintos niveles de medida, comenzando por el más débil y terminando por el más fuerte.

Nominal

Cada valor de una variable nominal se corresponde con una categoría de la variable; este emparejamiento es por lo general arbitrario. Como ejemplos de variables nominales podemos considerar el sexo de una persona, lugar de nacimiento etc. En este nivel de medida las categorías no pueden ser ordenadas en ningún sentido, y, por supuesto, no tiene sentido calcular medias, medianas,..., etc. Los estadísticos habituales serán frecuencias y porcentajes.

Ordinal

Cada valor representa la ordenación o el ranking; por ejemplo, el lugar de llegada a meta de los corredores, 1 significaría el primero, 2 significaría el segundo,... etc. Es muy común encontrarse este tipo de variables, por ejemplo, en la evaluación del gusto de los consumidores, a quienes se les suministra una serie de productos y ellos van indicando el más preferido,... etc. Sabremos cuál es el primero en preferencia, el segundo,..., etc., pero no sabremos cuánto es de preferido. En el ejemplo de la carrera sabremos cuál ha sido el primero, el segundo, pero no vamos a saber cual es la distancia entre el primero y el segundo. Los estadísticos a solicitar serán: frecuencias, porcentajes, moda y la mediana.

Intervalo

En variables de intervalo un incremento de una unidad en el valor numérico representa el mismo cambio en la magnitud medida, con independencia de donde ocurra en la escala. En este nivel de medida los estadísticos habituales son la media, la desviación típica y la mediana. La mayoría de los análisis asumen que las variables tienen, por lo menos, este nivel de medida. Un ejemplo de variable con nivel de intervalo podría ser el salario, la temperatura,…etc. Los estadísticos a emplear serán: la media, media recortada y la mediana.

Razón

Las variables de Razón tienen las mismas propiedades que las de intervalo, pero además tienen un punto cero significativo, que representa una ausencia completa de la característica medida. Por ejemplo, la edad o las ganancias anuales de una persona. Por ello, las variables de Razón tienen propiedades más fuertes que las de intervalo.

En función del nivel de medida de la variable de interés, se deberá de utilizar unas medidas resumen en vez de otras. Seguidamente vamos a enumerar las más usuales, indicando para cada una de ellas el nivel de medida adecuado y sus principales características.

4- TABLA DE FRECUENCIAS

La tabla de frecuencia es adecuada cuando estamos analizando una variable con nivel de medida nominal u ordinal; también se puede emplear sobre variables con nivel de medida de escala o razón, pero no es el método más adecuado y sus resultados deben ser examinados con mucho cuidado.

Frecuencia absoluta

Supongamos que tenemos una variable X con nivel de medida ordinal o nominal; esta variable tendrá k categorías: i=1 … k.

Llamamos frecuencia absoluta de la categoría i, y se representa por n_i, al número de veces que aparece la modalidad i en la muestra.

El listado de las categorías de la variable X junto con sus frecuencias absolutas se denomina tabla de frecuencias.

Frecuencia relativa

Si n es el número de casos observados, se denomina frecuencia relativa de la categoría i, y se representa como f_i, a:

$$f_i = \frac{n_i}{n}$$

Ejemplo

Sea una muestra de 294 personas, en cada persona se ha observado su empleo actual y el máximo nivel de estudios alcanzado. El Empleo es una variable con nivel de medida nominal y Estudios es ordinal, esto es, las categorías de la variable estudios pueden ordenarse de menor a mayor. Las tablas de frecuencias correspondientes a estas variables son:

empleo Tipo de empleo

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1 a tiempo completo	167	56,8	56,8	56,8
	2 a tiempo parcial	42	14,3	14,3	71,1
	3 desempleado	14	4,8	4,8	75,9
	4 jubilado	38	12,9	12,9	88,8
	5 trabajo en el hogar	27	9,2	9,2	98,0
	6 estudiante	2	,7	,7	98,6
	7 otros	4	1,4	1,4	100,0
	Total	294	100,0	100,0	

educa Estudios realizados

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1 sin estudios	5	1,7	1,7	1,7
	2 algo secundaria	61	20,7	20,7	22,4
	3 secundaria	114	38,8	38,8	61,2
	4 algo universitaria	48	16,3	16,3	77,6
	5 universitaria	43	14,6	14,6	92,2
	6 Master	14	4,8	4,8	96,9
	7 doctor	9	3,1	3,1	100,0
	Total	294	100,0	100,0	

La segunda columna de la tabla se corresponde con las frecuencias absolutas y la tercera con las relativas.

Es interesante observar que la variable Estudios ha sido codificada de forma que las distintas categorías se muestren de forma ascendente, en cambio en la variable Empleo es indistinto el orden en que pongamos las distintas categorías, pues no podemos establecer una relación de orden.

Si la muestra ha sido seleccionada mediante un método aleatorio adecuado, las frecuencias relativas se van a corresponder con la probabilidad de que al seleccionar una persona al azar, pertenezca a dicha categoría.

5- TABLAS DE FRECUENCIA DE DOBLE ENTRADA

Cuando trabajamos en un estudio estadístico y observamos simultáneamente dos caracteres en un mismo individuo obtenemos pares de resultados, por ejemplo, al observar en una persona el color de ojos y el color del pelo.
Los distintos valores de las modalidades que pueden adoptar estos caracteres forman un conjunto de pares, que representamos por (X, Y) y llamamos variable estadística bidimensional.

La forma más usual de estudiar variables estadísticas bidimensionales es representarlas en una tabla de doble entrada de la siguiente forma:

Y / X	y_1	y_2	.....	y_j	.....	y_k	n_i .
x_1	n11	n12	.....	n1j	.....	n1k	n_1 .
x_2	n21	n22	.....	n2j	.....	n2k	n_2 .
. . .	. .	. .	. .... .	. .	. .... .	. .	. .
x_i	ni1	ni2	.....	nij	.....	nik	n_i .
. . .	. . .	. . .	. .... .	. . .	. .... .	. . .	. . .
x_h	nh1	nh2	.....	nhj	.....	nhk	n_h .
$n._j$	$n._1$	$n._2$	.....	$n._j$	.....	$n._k$	N

En las columnas representamos las distintas modalidades de la variable X y en las filas las modalidades de la variable Y, como se muestra en la figura anterior.

En el cruce de la fila i con la columna j se representa en número de observaciones que han presentado simultáneamente la característica i de la variable X y la característica j de la variable Y, a dicha cantidad se le denomina frecuencia conjunta y se representa por n_{ij}

Frecuencias relativas

Las frecuencias relativas de la distribución bivariante son las frecuencias absolutas divididas por el número total de casos y se suelen representar por la letra f, por tanto $f_{ij} = \frac{n_{ij}}{n}$

Frecuencias marginales

Dada una variable bidimensional (X, Y) denominamos distribución de frecuencia marginal a cada una de las distribuciones de frecuencias de las dos variables estadísticas consideradas unilateralmente. Tradicionalmente las tablas de frecuencias marginales se muestran en los márgenes de la tabla, de donde deriva el nombre de frecuencia marginal.
La frecuencia marginal representa la distribución de cada una de las variables sin tener en consideración a la otra variable.

Frecuencia marginal de la variable X (fila):

$$n_{i.} = \sum_k n_{ik}$$

Frecuencia marginal de la variable Y (columna):

$$n_{.j} = \sum_k n_{kj}$$

Frecuencias condicionadas

De todos los elementos de la población, N, podemos estar interesados, en un momento dado, en un conjunto más pequeño y que está formado por aquellos elementos que han presentado la modalidad $Y = j$. El número de elementos de este conjunto es $n_{.j}$. La variable X definida sobre este conjunto se denomina variable condicionada y se suele denotar mediante $X/Y = j$. La distribución de frecuencias absolutas de esta nueva variable es exactamente la columna j de la tabla, pero sus frecuencias relativas, que denominaremos frecuencias relativas condicionadas son:

$$f(x = i / y = j) = \frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}}$$

En términos probabilísticos se puede decir que la frecuencia condicionada es la probabilidad de que X tome la modalidad i sabiendo que ha tomado la

modalidad j de la variable Y; por ejemplo: cuál es la probabilidad de que una persona rubia tenga los ojos azules.

Resumen:

Número total de observaciones: n

Frecuencia conjunta de las categorías (i,j): n_{ij}

Frecuencia relativa conjunta de las categorías (i,j): $f_{ij} = \frac{n_{ij}}{n}$

Frecuencia marginal de la variable fila: $n_{i\cdot} = \sum_j n_{ij}$

Frecuencia marginal de la variable columna: $n_{\cdot j} = \sum_i n_{ij}$

Frecuencia relativa marginal de la variable fila: $f_{i\cdot} = \sum_j \frac{n_{ij}}{n}$

Frecuencia relativa marginal de la variable columna: $f_{\cdot j} = \sum_i \frac{n_{ij}}{n}$

Frecuencias condicionadas: $f_{i/j} = \frac{f_{ij}}{f_{\cdot j}}$

Ejemplo:

Sean las variables nominales o categóricas, Color de pelo y Color de ojos. Tomamos una muestra de 68.000 individuos y observamos estas características. La tabla de frecuencias conjuntas es la siguiente:

Tabla de contingencia pelo Color pelo * ojos Color ojos

Recuento

			ojos Color ojos			
			1,00 azul	2,00 gris/verde	3,00 negro/pardo	Total
pelo Color pelo	1,00	rubio	1768	946	115	2829
	2,00	castaño	807	1387	438	2632
	3,00	negro	189	746	288	1223
	4,00	pelirojo	47	53	16	116
Total			2811	3132	857	6800

El número en cada casilla se corresponde con el número de personas que simultáneamente presenta las características fila y columna correspondientes; la primera celda indica que hay 1.768 personas que poseen simultáneamente el pelo rubio y los ojos azules.

La última columna indica las frecuencias absolutas de la variable Color de pelo, y la última fila contiene las frecuencias absolutas de la variable Color de ojos.

Para obtener las frecuencias relativas conjuntas, dividimos toda la tabla por el número total de casos, como se muestra a continuación:

Color pelo * Color ojos Crosstabulation

% of Total

		Color ojos			
		azul	gris/verde	negro/pardo	Total
Color pelo	rubio	26,0%	13,9%	1,7%	41,6%
	castaño	11,9%	20,4%	6,4%	38,7%
	negro	2,8%	11,0%	4,2%	18,0%
	pelirojo	,7%	,8%	,2%	1,7%
Total		41,3%	46,1%	12,6%	100,0%

Las frecuencias relativas conjuntas se pueden interpretar como la probabilidad de seleccionar a una persona con las características indicadas.

Nuevamente las frecuencias relativas marginales son mostradas en los márgenes de las tablas.

Tabla de frecuencias condicionadas

Al condicionar una variable por los valores de la otra, vamos a tener dos tablas $f_{x/y}$ y $f_{y/x}$.

Tabla de frecuencias del color del pelo condicionado por el color de los ojos $f_{Y/X}$.

pelo Color pelo * ojos Color ojos Crosstabulation

% within ojos Color ojos

			ojos Color ojos			
			1,00 azul	2,00 gris/verde	3,00 negro/pardo	Total
pelo Color pelo	1,00	rubio	62,9%	30,2%	13,4%	41,6%
	2,00	castaño	28,7%	44,3%	51,1%	38,7%
	3,00	negro	6,7%	23,8%	33,6%	18,0%
	4,00	pelirojo	1,7%	1,7%	1,9%	1,7%
Total			100,0%	100,0%	100,0%	100,0%

El sentido de esta tabla es responder a la pregunta **¿cual es la probabilidad de que una persona con los ojos azules tenga el pelo rubio?** o bien **¿cual es la probabilidad de que una persona con los ojos gris/verde tenga el pelo castaño?** La respuesta es 62.9% y 44.3% respectivamente.

Tabla de frecuencias del color de los ojos condicionado por el color del pelo $f_{x/y}$.

Análogamente, la primera celda indica la probabilidad de que una persona rubia tenga los ojos azules (62.5%), la segunda es la probabilidad de que una persona rubia tenga los ojos grises o verdes (33.4%), la tercera es la probabilidad de que una persona rubia tenga los ojos negros (4.1%) etc.

Tabla de contingencia pelo Color pelo * ojos Color ojos

% de pelo Color pelo

		ojos Color ojos			
		1,00 azul	2,00 gris/verde	3,00 negro/pardo	Total
pelo Color pelo	1,00 rubio	62,5%	33,4%	4,1%	100,0%
	2,00 castaño	30,7%	52,7%	16,6%	100,0%
	3,00 negro	15,5%	61,0%	23,5%	100,0%
	4,00 pelirojo	40,5%	45,7%	13,8%	100,0%
Total		41,3%	46,1%	12,6%	100,0%

6- COEFICIENTE χ^2

El objetivo de estudiar una tabla de contingencia o de doble entrada, además del puramente descriptivo, es determinar hasta qué punto existe asociación entre las dos variables consideradas. Si examinamos la tabla de frecuencias relativas de Color de ojos y Color de pelo que mostramos a continuación:

Tabla de contingencia pelo Color pelo * ojos Color ojos

% del total

		ojos Color ojos			
		1,00 azul	2,00 gris/verde	3,00 negro/pardo	Total
pelo Color pelo	1,00 rubio	26,0%	13,9%	1,7%	41,6%
	2,00 castaño	11,9%	20,4%	6,4%	38,7%
	3,00 negro	2,8%	11,0%	4,2%	18,0%
	4,00 pelirojo	,7%	,8%	,2%	1,7%
Total		41,3%	46,1%	12,6%	100,0%

Podemos observar que el color de ojos más frecuente (moda) es el gris/verde (frecuencia relativa marginal de 46.1) y el color de pelo más abundante es el rubio (frecuencia relativa marginal 41.6); sin embargo, lo más probable es que seleccionemos a una persona rubia con los ojos azules (frecuencia relativa conjunta 26.0). Esto nos lleva a la cuestión de dependencia e independencia estadística.

Independencia estadística

Si el suceso A tiene una probabilidad P, entonces en n repeticiones del experimento aleatorio el número esperado de ocurrencias del suceso A es:

Frecuencia esperada:

$$E(A) = n * P(A)$$

En general se dice que dos sucesos son independientes si la probabilidad de que ocurran simultáneamente es igual al producto de sus probabilidades, es decir:

$$P(A \cap B) = P(A) * P(B)$$

Por lo tanto la frecuencia esperada del suceso $\{A \cap B\}$ si ambos son independientes será:

$$E(\{A \cap B\}) = n * P(A) * P(B)$$

Sin embargo, la frecuencia observada de dicho suceso es n_{ab}.

Comparando las frecuencias esperadas con las observadas podemos hacernos una idea del grado de asociación existente entre las dos variables, a medida que las frecuencias observadas se van distanciando de las frecuencias esperadas va creciendo el grado de asociación entre las variables.

La frecuencia esperada bajo hipótesis de independencia estadística del suceso{pelo rubio y ojos azules} es:

$$6800 * P(pelo\_rubio) * P(ojos\_azules) = 6800 * \frac{2829}{6800} * \frac{2811}{6800} = 1169.45$$

Sin embargo la frecuencia observada es 1.768. Calculando de esta manera la frecuencia esperada para todas las celdas de la tabla podemos observar el grado de ajuste:

Tabla de contingencia pelo Color pelo * ojos Color ojos

			ojos Color ojos			
			1,00 azul	2,00 gris/verde	3,00 negro/pardo	Total
pelo Color pelo	1,00 rubio	Recuento	1768	946	115	2829
		Frecuencia esperada	1169,5	1303,0	356,5	2829,0
	2,00 castaño	Recuento	807	1387	438	2632
		Frecuencia esperada	1088,0	1212,3	331,7	2632,0
	3,00 negro	Recuento	189	746	288	1223
		Frecuencia esperada	505,6	563,3	154,1	1223,0
	4,00 pelirojo	Recuento	47	53	16	116
		Frecuencia esperada	48,0	53,4	14,6	116,0
Total		Recuento	2811	3132	857	6800
		Frecuencia esperada	2811,0	3132,0	857,0	6800,0

Las celdas subrayadas indican las mayores discrepancias observadas.

Para obtener una medida de asociación entre dos variables categóricas, podemos utilizar el estadístico χ^2, que calcula la probabilidad de obtener estas diferencias teniendo en cuenta que las variables son independientes; si esta probabilidad es muy pequeña se deberá de concluir que existe asociación entre las variables.

El coeficiente χ^2 cuadrado se define como:

$$\chi^2 = \sum \sum \frac{\left(n_{i,j} - n'_{i,j}\right)^2}{n'_{i,j}}$$

Notación:

- $n_{i,j}$ es la frecuencia conjunta observada de la fila i y columna j.
- $n'_{i,j}$ es la frecuencia conjunta esperada de la fila i y columna j.

Entonces, para n grande, el estadístico

$$\chi^2 = \sum \sum \frac{\left(n_{i,j} - n'_{i,j}\right)^2}{n'_{i,j}}$$

tiene una distribución aproximada ji-cuadrado con (Columnas-1)*(Filas-1) grados de libertad si la hipótesis nula es verdadera, es decir, no existe asociación entre las variables. Por consiguiente, la hipótesis de independencia debe rechazarse si el valor del estadístico de prueba X2 calculado es mayor que X2 crítico de tabla. En este ejemplo el valor del estadístico es: $\chi^2 = 1073.5$. Si miramos en la tabla de distribución de χ^2_6, vemos que el valor critico para una distribución χ^2 con 6 grados de libertad y un $\alpha = 0.05$ es de 12.592 por lo tanto debemos rechazar la hipótesis de que no existe dependencia entre estas variables.

Resumiendo:

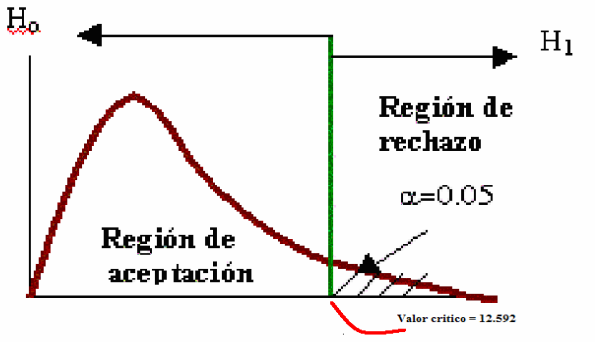
- Ho: El color de los ojos no influye en el color del pelo.
- H1: El color de los ojos si influye en el color del pelo.

Grados de libertad: (r-1)(c-1) = (3-1)(4-1)=(2)(3) = 6.

Valor crítico: 12.592

Regla de decisión:

- Si $\chi^2 \leq$ 12.592 no se rechaza Ho.
- Si $\chi^2 >$ 12.592 se rechaza Ho.



7- MEDIDAS DE POSICIÓN

Si bien la tabla de frecuencias ofrece toda la información disponible, el analista se encuentra difícil, en numerosos casos, la interpretación de toda esta extensa información. Por ello, es conveniente sinterizarla en unas pocas medidas de resumen.

En este proceso de resumen buscamos unos valores que nos fijen el comportamiento global del fenómeno que estamos estudiando. Estos valores sintéticos globales son llamados medidas de posición.

Media aritmética

Se define la media aritmética como la suma de todos los valores de la distribución dividida por el número total de datos.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Propiedades de la media aritmética:

- La media aritmética va en las mismas unidades que la variable observada.
- Si a todos los valores de la variable les sumamos una constante, la media aritmética queda aumentada en dicha constante.
- Si a todos los valores de la variable les multiplicamos por una constante, la media aritmética quedara multiplicada por dicha constante.
- La suma de las desviaciones de los valores de la variable a su media es 0

$$\sum \frac{(x_i - \bar{x})}{n} = 0$$

Ventajas de la media aritmética:

- En su cálculo intervienen todos los valores de la variable.
- Es única.
- Siempre es calculable.
- Es de fácil interpretación.

Inconvenientes de la media aritmética:

- Es muy sensible a valores anormalmente altos o bajos pudiendo inducir a conclusiones poco atinadas.

La media aritmética es la medida más adecuada para el resumen de variables de escala y proporciones.

Media geométrica

Se define la media geométrica como:

$$G(x) = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Es decir, la raíz n-ésima del producto de todos los valores de la variable.

Vamos a ver su significado a través de un **ejemplo**.

Ejemplo

Supongamos que hemos invertido un capital C a tres años, con un interés i_1, i_2 y i_3.
El interés medio, entendiendo el interés que permaneciendo constante y partiendo de un capital C, en tres años nos hubiera generado un capital de C_3 no es:

$$\frac{i_1 + i_2 + i_3}{3}$$

Si hacemos cálculos:

- Primer año: $C_1 = C * (1 + i_1)$
- Segundo año: $C_2 = C_1 * (1 + i_2) = C * (1 + i_1) * (1 + i_2)$
- Tercer año: $C_3 = C_2 * (1 + i_3) = C * (1 + i_1) * (1 + i_2) * (1 + i_3)$

Si el interés hubiera permanecido constante, es decir, el interés medio tendríamos:

$$C_3 = C * (1 + i)^3$$

Igualando estas expresiones y simplificando obtenemos:

$$C*(1+i)^3 = C*(1+i_1)*(1+i_2)*(1+i_3)$$

$$(1+i) = \sqrt[n]{(1+i_1)*(1+i_2)*(1+i_3)}$$

De donde se deduce que el interés medio es la media geométrica de los intereses.

La media geométrica se debe de emplear cuando estemos estudiando incrementos porcentuales acumulativos.

Ventajas de la media geométrica:

- En su cálculo intervienen todos los valores de la distribución.
- Es menos sensible que la media aritmética a los valores extremos.

Inconvenientes de la media geométrica:

- En general es de significado menos intuitivo que la media aritmética.
- Cuando existe algún valor igual a cero o negativo queda indeterminada.

Media armónica

La media armónica H(x) de una distribución se define como:

$$H(x) = \frac{n}{\sum \frac{1}{x_i}}$$

Mediana

Como es posible que la definición entrañe alguna dificultad vamos a dar varias definiciones y un ejemplo aclaratorio.

a) Es aquel valor de la distribución (ordenada de menor a mayor) que deja a ambos lados el mismo número de casos, es decir el valor central.

b) Es aquel valor cuya frecuencia acumulada es: $n/2$

Ejemplo

Supongamos que tenemos la altura de 11 personas:

1	2	3	4	5	6	7	8	9	10	11
1.61	1.88	1.65	1.88	1.61	1.87	1.71	1.68	1.5	1.6	1.54

En primer lugar, ordenamos los valores de menor a mayor:

1	2	3	4	5	6	7	8	9	10	11
1.50	1.54	1.60	1.61	1.61	**1.65**	1.68	1.71	1.87	1.88	1.88

La mediana es el valor que ocupa la casilla central, por lo tanto la mediana de esta distribución es 1.65.

Propiedades de la mediana:

- La mediana hace mínima las sumas de las diferencias en valor absoluto de los valores de la distribución: $Min\sum|x_i - Med|$

Ventajas de la mediana

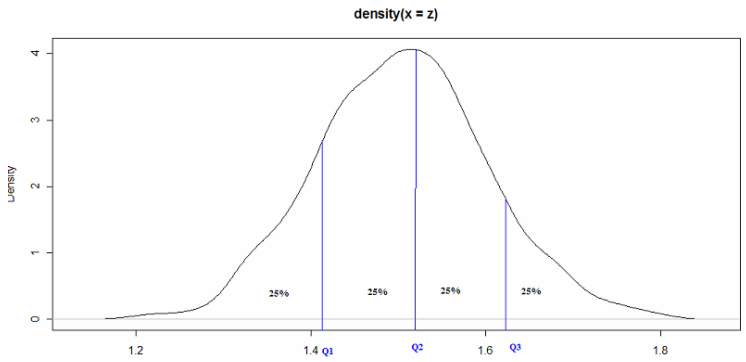
- La mediana tiene las mismas unidades que la variable.
- La mediana es una medida robusta frente a valores atípicos en la distribución.
- Tiene fácil interpretación.

Inconvenientes de la mediana

- No todos los valores de la distribución entran en su cálculo.
- Su estimación es menos precisa que la estimación de la media.

Cuartiles

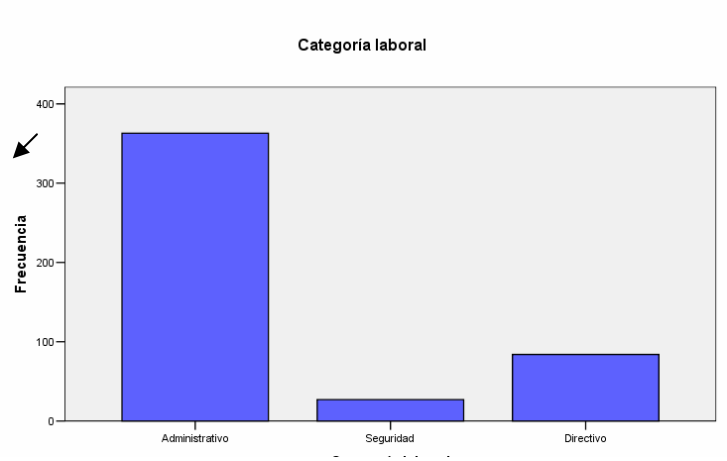
Son tres valores de la distribución que la dividen en cuatro partes iguales, es decir, cuatro intervalos, dentro de cada cual se encuentra un 25% de los casos. Se suelen representar mediante la letra Q acompañada del subíndice correspondiente, esto es Q_1, Q_2 y Q_3.



Moda

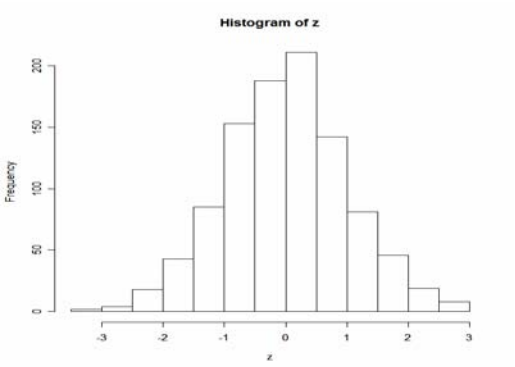
Es aquel valor de la distribución que más se repite, es decir, aquél con mayor frecuencia.
La moda es la medida más representativa en caso de distribuciones en escala nominal u ordinal.
En las variables con nivel de medida de escala o proporción se deben de emplear técnicas especiales para su cálculo, las cuales no suelen estar implementadas en los paquetes estadísticos ordinarios.

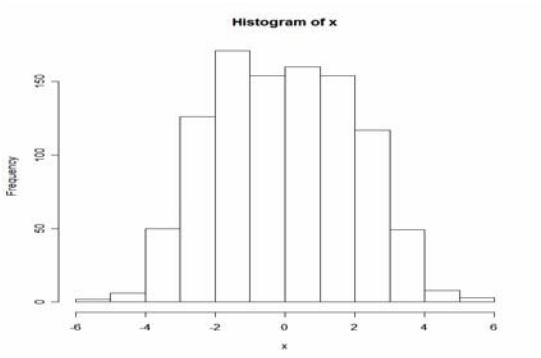
En las variables con nivel de medida nominal u ordinal, la moda o valor modal puede ser deducido sin más que mirar el diagrama de frecuencias correspondiente.



8- MEDIDAS DE DISPERSIÓN

Si bien las medidas de tendencia central (media, mediana y moda) informan sobre los valores medios de la distribución; la representatividad de estos valores está relacionada por cómo están de próximos a estas medidas. Lógicamente, cuanto más próximos, más representativas serán estas medidas.
A la proximidad de unos valores respecto de otros se le denomina dispersión o variabilidad.





8.1- Medidas de dispersión absolutas

Recorrido: Max(x)-Min(x)

Rango intercuartilico: Q3(x)-Q1(x)

Varianza: $$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \overline{X})^2}{N}$$

Desviación típica: $$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \overline{X})^2}{N}}$$

Propiedades de la Desviación típica:

- La desviación típica va expresada en las mismas unidades de medida que la distribución, lo cual la hace más apta como medida de dispersión.
- La desviación típica siempre es positiva.
- Si la desviación típica es 0, entonces todos los valores de la distribución son iguales.
- Si multiplicamos todos los valores de la variable por una constante, la desviación típica queda multiplicada por dicha constante.
- Si sumamos o restamos una constante a todos los valores de la distribución la desviación típica permanece inalterable.

Tipificación

Una variable estadística se denomina tipificada, reducida o estandarizada si su media es cero y su varianza 1.
Dada una variable estadística X con varianza distinta de cero, para estandarizarla basta con restarle su media y dividirla por su desviación típica:

$$Z_i = \frac{X_i - \overline{X}}{\sigma^2}$$

La variable tipificada carece de unidades de medida, lo cual permite comparar los valores individuales sin tener en cuenta en qué escala fueron medidos; por tanto los valores estandarizados representan la distancia a la media de la población pero medida en desviaciones típicas.

8.2- Medidas de dispersión relativas

Coeficiente de apertura

Es el cociente entre el mayor y menor valor de la distribución:

$$Ca(X) = \frac{Max(X)}{Min(X)}$$

Este coeficiente sólo tiene sentido cuando la variable tiene el mismo signo y además no puede tomar el valor 0. Su ámbito de aplicación está en variables del tipo salarios, ingresos, …, etc.

Coeficiente de variación de Pearson

Es el cociente de la desviación típica entre su media:

$$Cv(X) = \frac{\sigma}{\overline{X}}$$

Al igual que el coeficiente de apertura, sólo se debe emplear cuando los valores de la distribución no cambian de signo. Su significado, en cambio, es muy diferente, pues es una medida de la variación de los datos (desviación típica) respecto a su media.
Este coeficiente no tiene unidades de medida.

Ejemplo de aplicación

Supongamos que hemos observado el salario anual de una serie de personas en dos momentos distintos de tiempo: en 1996 y en 2006; como los salarios han aumentado, en la misma medida habrá aumentado, en líneas generales,

su dispersión, y su varianza; por lo tanto, para comparar la variabilidad de los datos deberemos de utilizar el coeficiente de variación de Pearson.

	1996	2006
Media	2,8305	5,5578
Desv. típ.	1,25766	2,66711

$$Cv(X_{1996}) = \frac{1.257}{2.8305} = 0.44$$

$$Cv(X_{2006}) = \frac{2.667}{5.558} = 0.48$$

9- MEDIDAS DE FORMA

Las medidas vistas anteriormente sintetizan la información (medidas de tendencia central) y, además, tratan de indicar cómo están de concentrados los valores en torno a dichas medidas (medidas de dispersión).
A continuación vamos a ver qué medidas debemos emplear para caracterizar el comportamiento de la distribución, esto es, su forma.

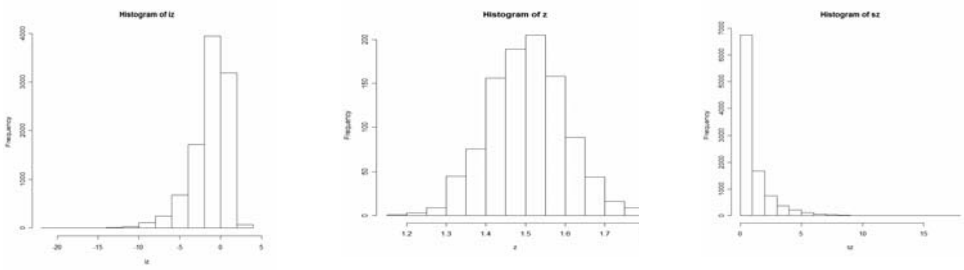
Las medidas de la forma de la distribución se pueden clasificar en dos grandes grupos:

- Asimetría
- Curtosis

9.1- MEDIDAS DE ASIMETRÍA

Las medidas de asimetría están encaminadas a determinar hasta qué punto una distribución es simétrica o no.

Generalizando podemos distinguir entre tres casos típicos:

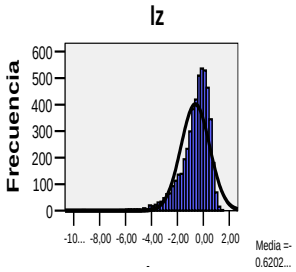


La medida utilizada para determinar el grado de asimetría es el **coeficiente de asimetría o SKEWNESS**. La definición matemática de este coeficiente es:

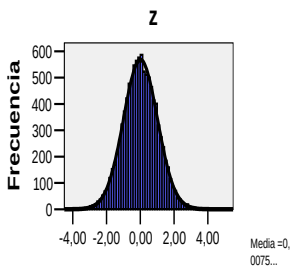
$$Casimetria = \sum \frac{(x_i - \bar{x})^3}{\sigma^3 * n}$$

Propiedades del coeficiente de asimetría:

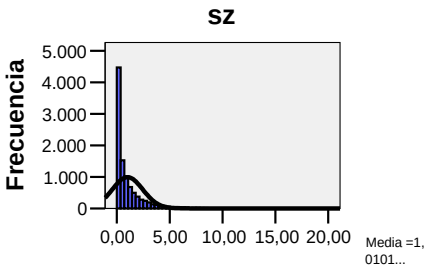
- Carece de unidades de medida.
- Si la distribución es simétrica, toma valores cercanos a 0.
- Si la cola izquierda es más corta que la derecha, tomará valores positivos.
- Si la cola derecha es más corta que la izquierda, tomará valores negativos.



Ca= -1.52



Ca=0.009



Ca=2.82

9.2- MEDIDAS DE APUNTAMIENTO

Las medidas de apuntamiento, o concentración central, se deberían de aplicar únicamente a distribuciones simétricas, o con ligera asimetría.Tratan de estudiar la distribución de frecuencias de la zona central; la mayor o menor concentración dará lugar a un mayor o menor apuntamiento.

En el caso de asimetría estaba claro cuándo una distribución era simétrica o no, pero en el caso de apuntamiento se suele tomar una distribución de referencia para comparar si el apuntamiento es mayor o menor.
En general la distribución de referencia suele ser la distribución normal.

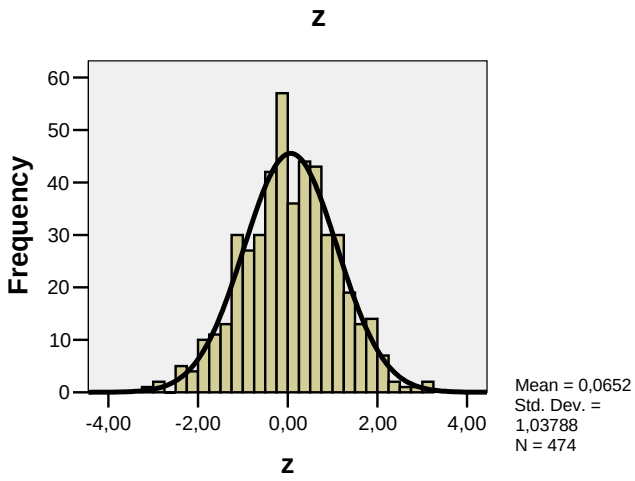
La formulación matemática del coeficiente de apuntamiento es:

$$Ca = \frac{\sum (x_i - \bar{x})^4}{n * \sigma^4} - 3$$

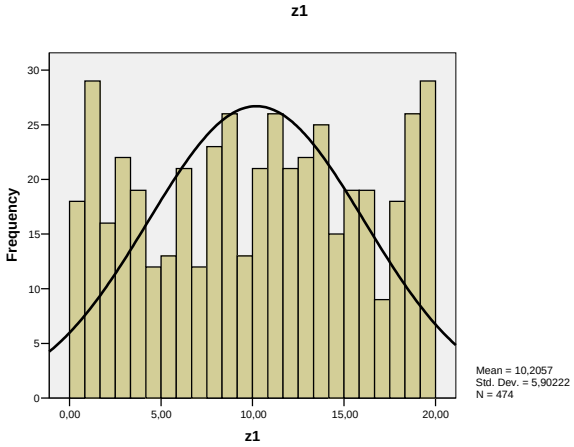
Propiedades del coeficiente de apuntamiento:

- Es independiente de la unidad de medida.
- Valores cercanos a cero indican que la distribución de frecuencias de la variable en cuestión tiene una curtosis equiparable con una distribución normal de igual media y varianza.
- Valores positivos indican una curtosis superior a una distribución normal.
- Valores negativos indican una curtosis inferior a la normal.

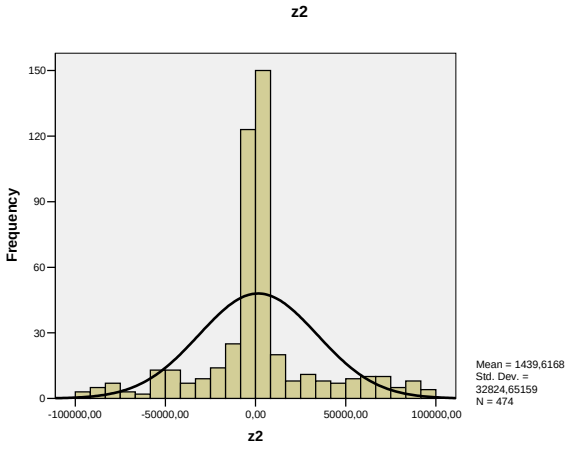
Ejemplo



Simetría-0.047
Curtosis=0.018



Simetría-0.037
Curtosis=-1.17



Simetría =2.32
Curtosis=1.81

10- DISTRIBUCIONES BIDIMENSIONALES

En los capítulos precedentes habíamos considerado el estudio de una única variable que reflejaba una determinada característica, como por ejemplo la edad. Pero para una población dada es posible estudiar simultáneamente dos o más características o variables. Por ejemplo, podemos estar interesados en estudiar la altura y el peso de las personas de forma simultánea.

De forma general, si se estudian dos características X, Y sobre una misma población, y ambas características son cuantitativas, podemos considerar para cada unidad muestral i, el par (x_i, y_i), podemos estudiar por separado cada variable, pero también es posible su estudio conjunto. El estudio conjunto de estas dos variables tiene como objetivo el determinar si existe o no algún grado de asociación entre ellas.

El coeficiente más ampliamente utilizado para estudiar la asociación entre dos variables cuantitativas es el coeficiente de **correlación lineal de Pearson**.

11- COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

El coeficiente de correlación de Pearson, o familiarmente denominado Correlación, mide el grado de asociación lineal entre dos variables, es decir, hasta qué punto dos variables son proporcionales.
El valor de este coeficiente no depende de las unidades de medida utilizadas; el valor de la correlación (es decir el valor del coeficiente) entre la altura y el peso de las personas será idéntico tanto si hemos medido la altura de las personas en metros, centímetros o pies y el peso en gramos, kilos o libras.

La formulación matemática de este coeficiente es:

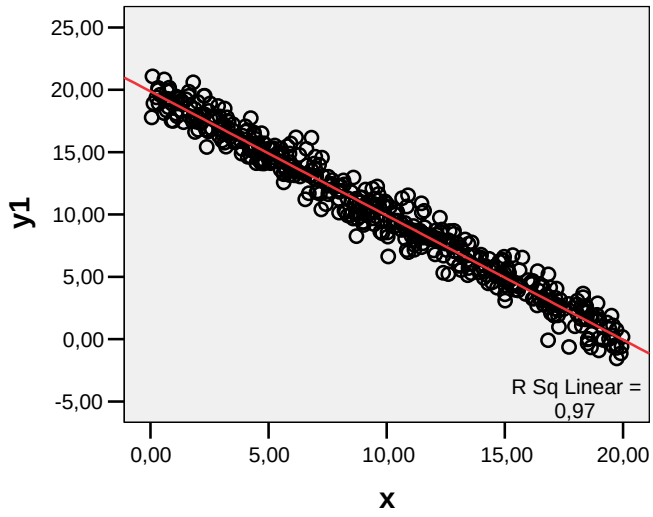
$$R = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{n * \sigma_x * \sigma_Y}$$

El recorrido de este coeficiente esta acotado en el intervalo $[-1,1]$

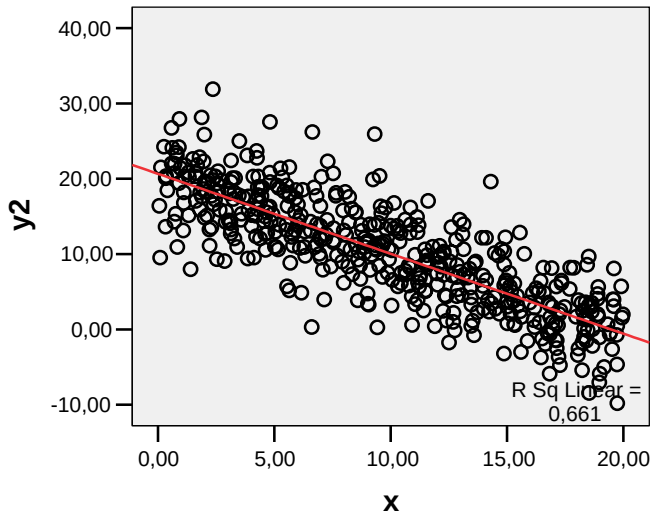
La interpretación de este coeficiente es bien sencilla:

- Valores cercanos a 1 ó -1 indican una fuerte asociación lineal.
- Valores cercanos a 0 indican falta de asociación lineal.
- Si el coeficiente es positivo, indica asociación positiva entre las variables, es decir, valores altos de la variable X se corresponderán con valores altos de la variable Y; igualmente, valores bajos de la variable X se corresponderán con valores bajos de la variable Y.
- Si el coeficiente es negativo, indica asociación negativa entre las variables, es decir valores altos de la variable X se corresponderán con valores bajos de la variable Y; igualmente valores bajos de la variable X se corresponderán con valores altos de la variable Y.

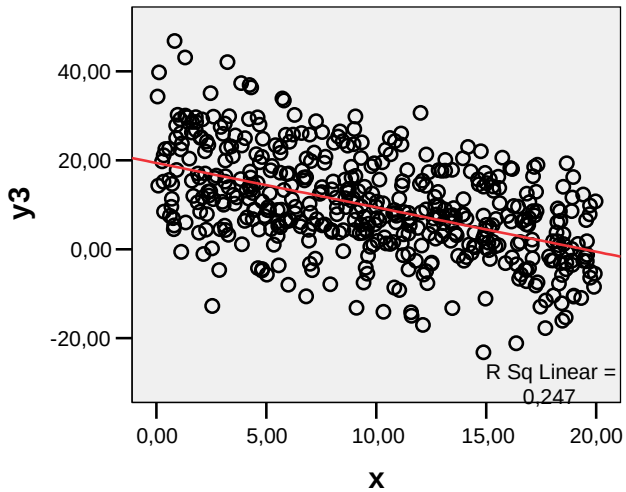
Ejemplos de correlación negativa:



R=-0.98

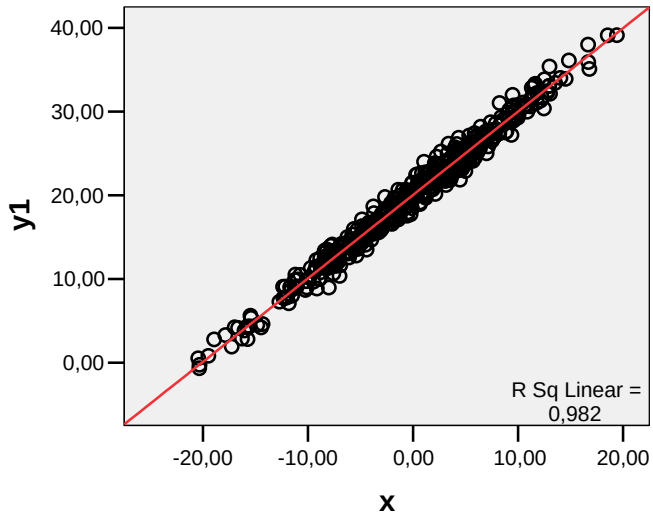


R=-0.81

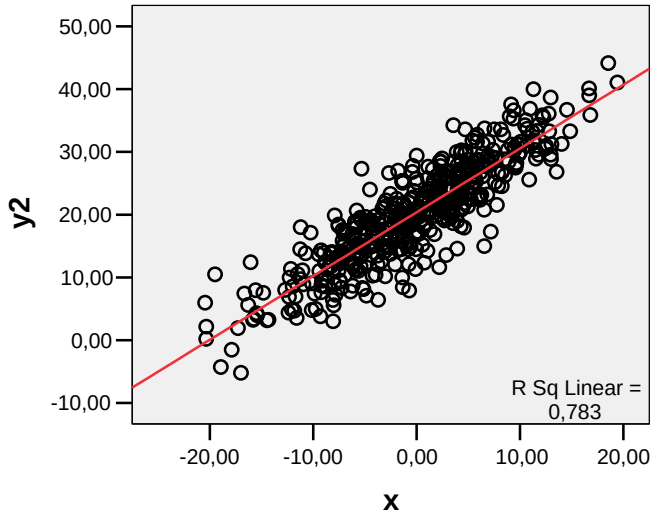


R=-0.49

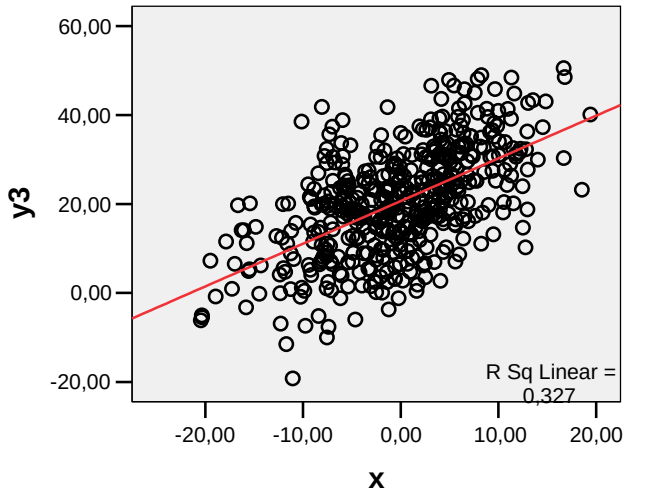
Ejemplos de correlación positiva:



R= 0.991

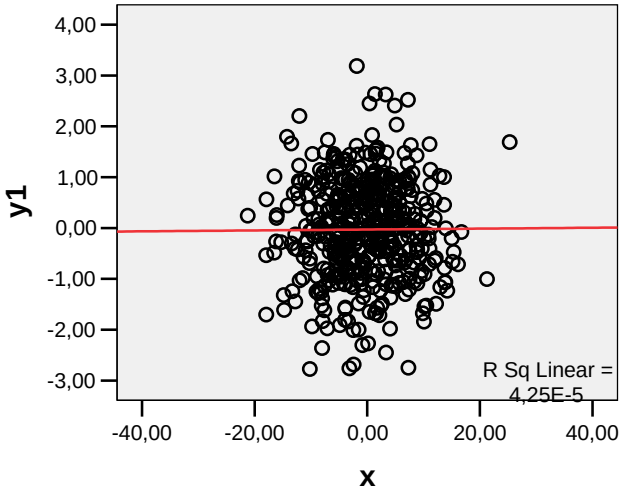


R=0.885

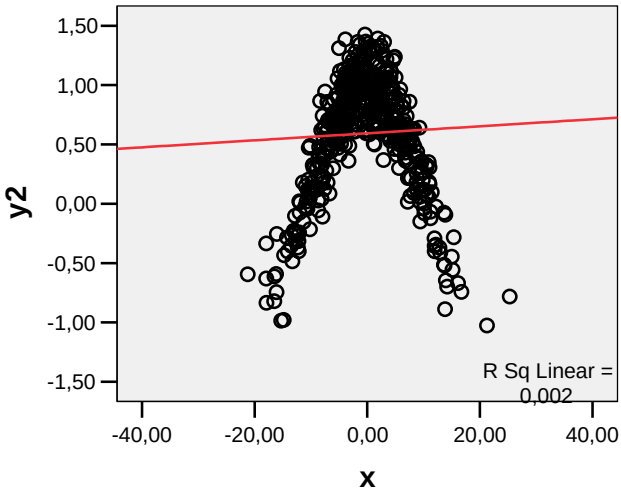


R=0.557

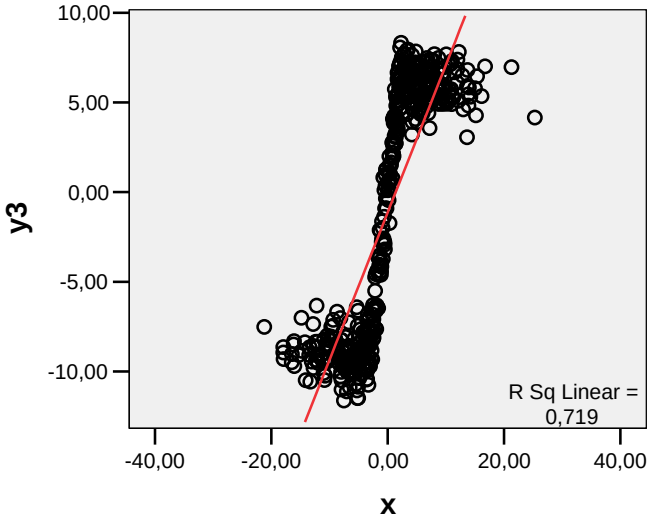
Ejemplos de falta de asociación lineal:



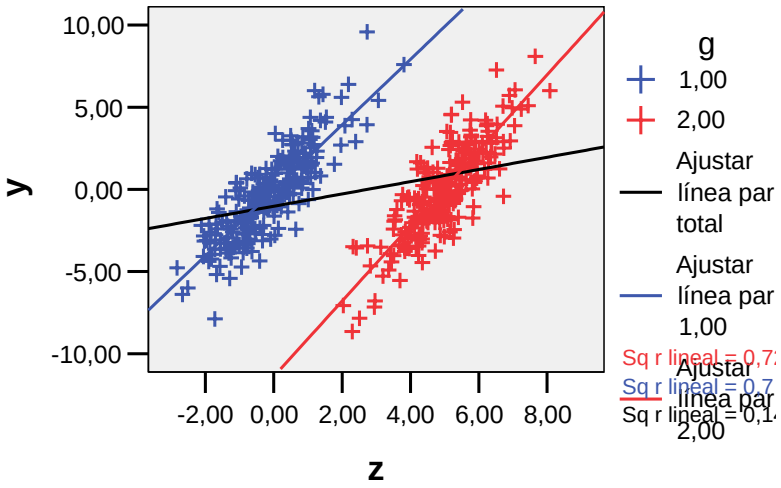
R=0.007



R=0.043



R=0.848


10,00
5,00
0,00
-5,00
-10,00
y
-2,00
0,00
2,00
4,00
6,00
8,00
z
g
1,00
2,00
Ajustar línea par total
Ajustar línea par 1,00
Ajustar línea par 2,00
Sq r lineal = 0,7
Sq r lineal = 0,7
Sq r lineal = 0,1

R=0.379

12- ANÁLISIS EXPLORATORIO DE DATOS

Antes de iniciar el análisis estadístico, es conveniente realizar una fase previa, encaminada a que el analista vaya tomando contacto con los datos que va a analizar y se familiarice con la naturaleza de los mismos.

Este estudio previo se denomina **análisis exploratorio de datos**, y se realiza sin ninguna hipótesis a priori, utilizando técnicas muy sencillas, donde abundan las representaciones graficas.

Es en esta fase donde se empezarán a revelar las relaciones más evidentes existentes entre las variables que posteriormente se estudiarán con el rigor correspondiente.

13- OBJETIVOS DEL ANÁLISIS EXPLORATORIO DE DATOS

En resumen, los objetivos del análisis exploratorio de datos son los siguientes:

- Familiarizarse con la naturaleza de los datos a analizar.
- Estudiar las principales características de la distribución de las variables.
- Tratar de poner de manifiesto las relaciones más evidentes que pudieran existir entre las variables.
- Detectar los valores atípicos.

14- FAMILIARIZÁNDOSE CON LA NATURALEZA DE LOS DATOS

Origen de los datos

El primer paso para empezar a familiarizarse con la naturaleza de los datos, es conocer la fuente de la que han sido obtenidos, a que año corresponden, cual ha sido el método de muestreo utilizado para la obtención de los mismos, así mismo debemos de identificar cual es la unidad muestral; la unidad muestral es lo que representa cada registro: si es una parcela de terreno, una persona, un departamento, una empresa, …, etc.
A continuación deberemos examinar cuantos casos y variables tiene el fichero (largo y ancho) y comprobar que es lo que representa cada variable. En este paso no debemos dejar lugar a la ambigüedad, además pondremos especial atención a las unidades de medida cuando las haya; por ejemplo, si los ingresos son mensuales o anuales, si van medidos en euros o en dólares, …, etc.

Nivel de medida de las variables

El siguiente paso es identificar el nivel de medida de cada variable:

En la práctica, la opción de un método estadístico depende en gran parte de la naturaleza de las observaciones que vayamos a estudiar. A continuación se muestran ordenados de menor a mayor los distintos niveles de medida, comenzando por el mas débil y terminando por el mas fuerte.

Nominal: Cada valor de una variable *nomina*l se corresponde con una categoría de la variable; este emparejamiento es por lo general arbitrario. Como ejemplos de variables nominales podemos considerar el sexo de una persona, lugar de nacimiento etc. En este nivel de medida las categorías no pueden ser ordenadas en ningún sentido, y, por supuesto, no tiene sentido calcular medias, medianas, ..., etc. Los estadísticos habituales serán frecuencias y porcentajes.

Ordinal: Cada valor representa la ordenación o el ranking; por ejemplo, el lugar de llegada a meta de los corredores, 1 significaría el primero, 2 significaría el segundo, ..., etc. Es muy común encontrarse este tipo de variables en la evaluación del gusto de los consumidores: se les suministra una serie de productos y ellos van indicando el más preferido, ..., etc. Sabremos cuál es el más preferido, el segundo más preferido, ..., etc., pero no sabremos cuanto es de preferido. En el ejemplo de la carrera sabremos cuál ha sido el primero, el segundo, pero no vamos a saber cuál es la distancia entre el primero y el segundo. Los estadísticos a solicitar serán: frecuencias, porcentajes, moda y la mediana.

Intervalo: En variables de *intervalo* un incremento de una unidad en el valor numérico representa el mismo cambio en la magnitud medida, con independencia de donde ocurra en la escala. En este nivel de medida los estadísticos habituales son la media, la desviación típica y la mediana. La mayoría de los análisis asumen que las variables tienen por lo menos este nivel de medida. Un ejemplo de variable con nivel de intervalo podría ser el salario, la temperatura, …, etc. Los estadísticos a emplear serán: la media, media recortada y la mediana.

Razón: Las variables de *razón* tienen las mismas propiedades que las de intervalo, pero además tienen un punto cero significativo; dicho punto representa una ausencia completa de la característica medida; por ejemplo, la edad o las ganancias anuales de una persona. Por ello, las variables de *razón* tienen propiedades más fuertes que las de intervalo.

Tipos de variables

Aparte del nivel de medida de las variables, también tenemos que ver el papel que van a jugar las variables en los diseños a realizar. Podemos distinguir las siguientes categorías:

a) Variables de identificación
Sirven para identificar un caso concreto; por ejemplo, el nombre de la persona. No siempre será posible o práctico tener la variable de identificación; en ese caso nos limitaremos a identificarla por el número que ocupa en el fichero de datos.

b) Variables de agrupamiento o factores

Permiten agrupar los casos en función a determinadas características; por ejemplo, el nivel de estudios alcanzado, el género de una persona, su estado civil, …, etc. En la práctica es usual solicitar estadísticos separados por **factores**, para comparar la posible influencia de los mismos en las variables de interés.

Pongamos, por caso, el factor género y la variable salario; podríamos solicitar los estadísticos por separado y comparar el salario de los hombres con el de las mujeres, como se muestra en la siguiente tabla:

		sexo Sexo del entrevistado	
		1 Hombre	2 Mujer
salario	Media	2347419,1	1916781
	Mediana	2490000,0	1992000
	Moda	2656000,0	2490000
	N total	641	859

c) Variables de ponderación

Se utilizan para elevar los resultados de la muestra (en general medias, frecuencias y porcentajes) y hacerlos representativos de la población que han sido seleccionados. En general, debemos examinar el diseño muestral para decidir cómo vamos a utilizar la variable de ponderación.

15- ESTUDIO DE LAS PRINCIPALES CARACTERÍSTICAS DE LA DISTRIBUCIÓN DE LAS VARIABLES

Valores en rango

El primer paso consiste en tratar de detectar si existen valores mal codificados, para lo cual vamos a distinguir dos grupos de variables en función a su nivel de medida; el primero va a estar formado por variables con nivel de medida de razón y de intervalo, el segundo va a estar formado por las variables con nivel de medida ordinal y nominal.

Para las variables de tipo ordinal y nominal, solicitamos la tabla de frecuencias, incluyendo las etiquetas y sus valores, con el fin de aseguramos de que están correctamente asignadas las etiquetas a los valores y que los valores están en rango.

ecivil Estado civil

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	1 Casado	795	53,0	53,0	53,0
	2 Viudo	165	11,0	11,0	64,0
	3 Divorciado	213	14,2	14,2	78,2
	4 Separado	40	2,7	2,7	80,9
	5 Soltero	286	19,1	19,1	99,9
	90	1	,1	,1	100,0
	Total	1500	99,9	100,0	
Perdidos	9 No contesta	1	,1		
Total		1501	100,0		

En esta tabla podemos observar que aparece un caso con el valor 90 y sin etiqueta, y que el valor *missing* definido por el usuario es 9, podría tratarse de un simple error de codificación.

Para las variables *continuas* solicitamos una tabla de estadísticos descriptivos, que incluya los valores mínimos y máximos, con el fin de detectar si los valores máximos y mínimos son razonables.

Estadísticos descriptivos

	N	Mínimo	Máximo	Media	Desv. típ.
salario	994	166000,00	3652000	2125602	933110,8
edad Edad del encuestado	1496	18	480	46,52	20,711
hijos Número de hijos	1495	0	8	1,85	1,682
N válido (según lista)	989				

En esta tabla podemos observar que en la variable edad hay un individuo con una edad de **480** años; este valor indica, con toda seguridad, un error de codificación. Por otra parte la existencia de una persona con ocho hijos, siendo un valor alto, no deja de ser un valor valido en esta fase del estudio.

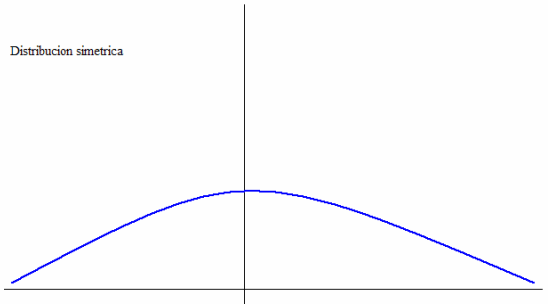
Características de forma

Una vez corregidos los valores mal codificados, bien sea poniendo su valor correcto o declarándolos como perdidos, ya podemos empezar a estudiar las principales características de las variables.
En primer lugar estudiaremos la forma de su distribución mediante técnicas gráficas y a continuación realizaremos un análisis numérico univariante para crear una tabla resumen con sus principales características. Las técnicas graficas a emplear son el histograma y el grafico de cajas.

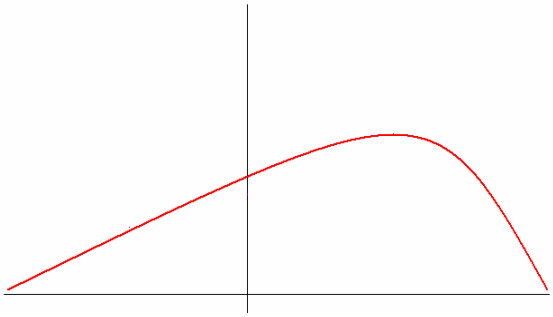
En el histograma debemos observar si la distribución es simétrica o no; en caso de ser simétrica, la media y la mediana deberán de coincidir, siendo la media el estadístico resumen de tendencia central mas adecuado. En caso de no ser simétrica, la media y la mediana no coincidirán, y deberemos de considerar, según sea el caso, la conveniencia de utilizar un estadístico u otro, pues cada uno esta midiendo conceptos distintos.

Si tenemos el suficiente número de casos, también deberemos observar el número de modas, pudiendo indicar la existencia de distintos grupos.

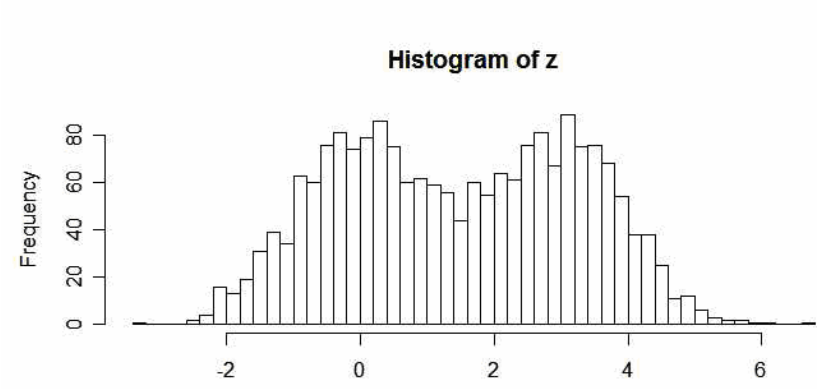
Ejemplo de distribución simétrica:



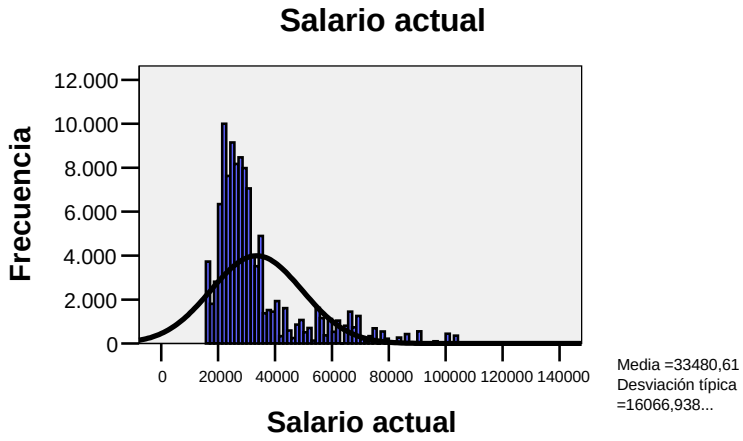
Ejemplo de distribución asimétrica:

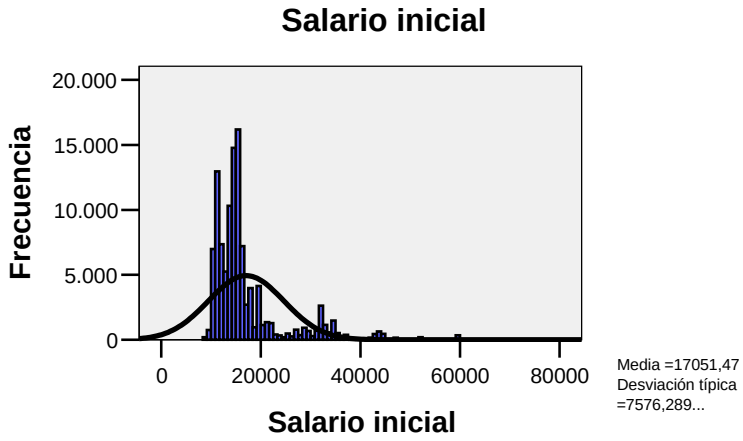


Ejemplo de una distribución bimodal:

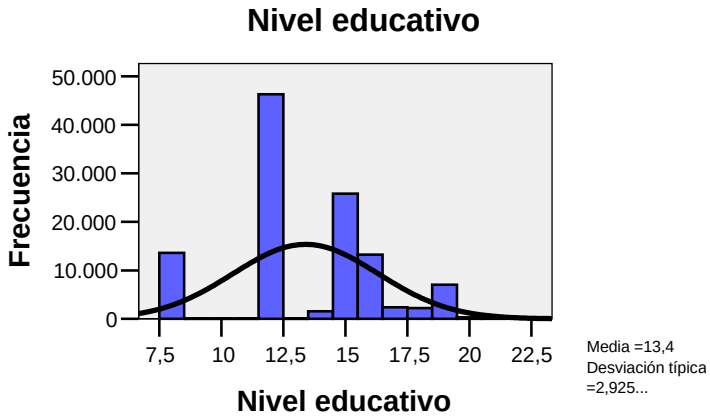


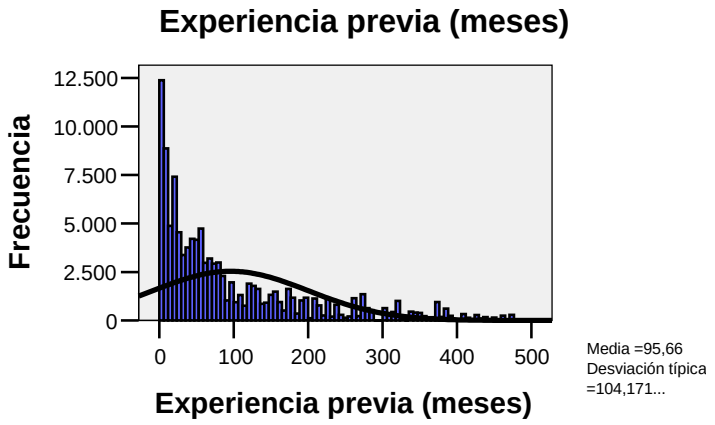
Examinando los histogramas nos hacemos idea de las formas de las distribuciones de las variables.





La forma de la cola derecha parece indicar la existencia de una o varias subpoblaciones.

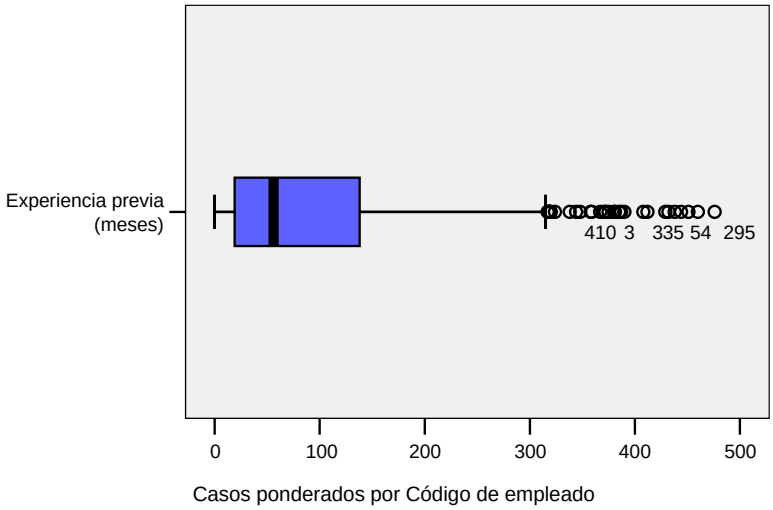


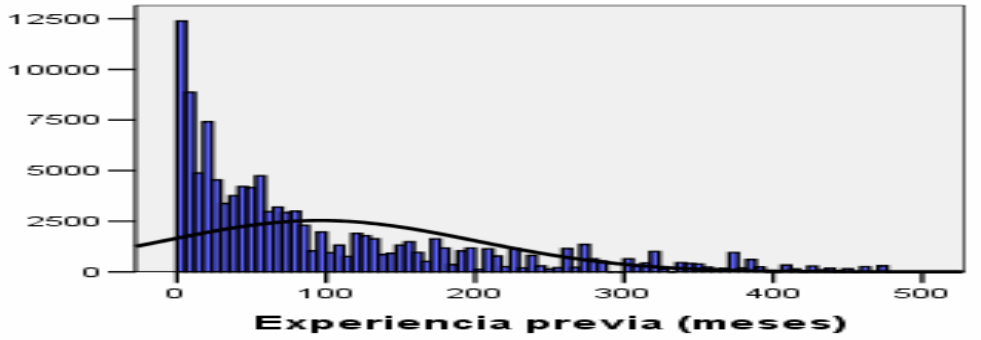


En este histograma vemos que el valor 476 no parece ser un error de codificación.

Gráfico de cajas

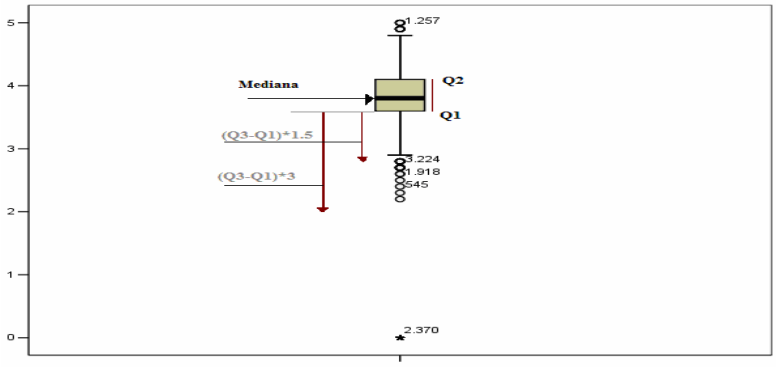
El gráfico de cajas, o box-plot, es un gráfico que representa de forma no parametrica la distribución de una variable, permite, además de observar existencia o no de simetría en los datos y detectar valores atípicos, estudiar la influencia que los factores pudieran tener en la distribución de la variable.



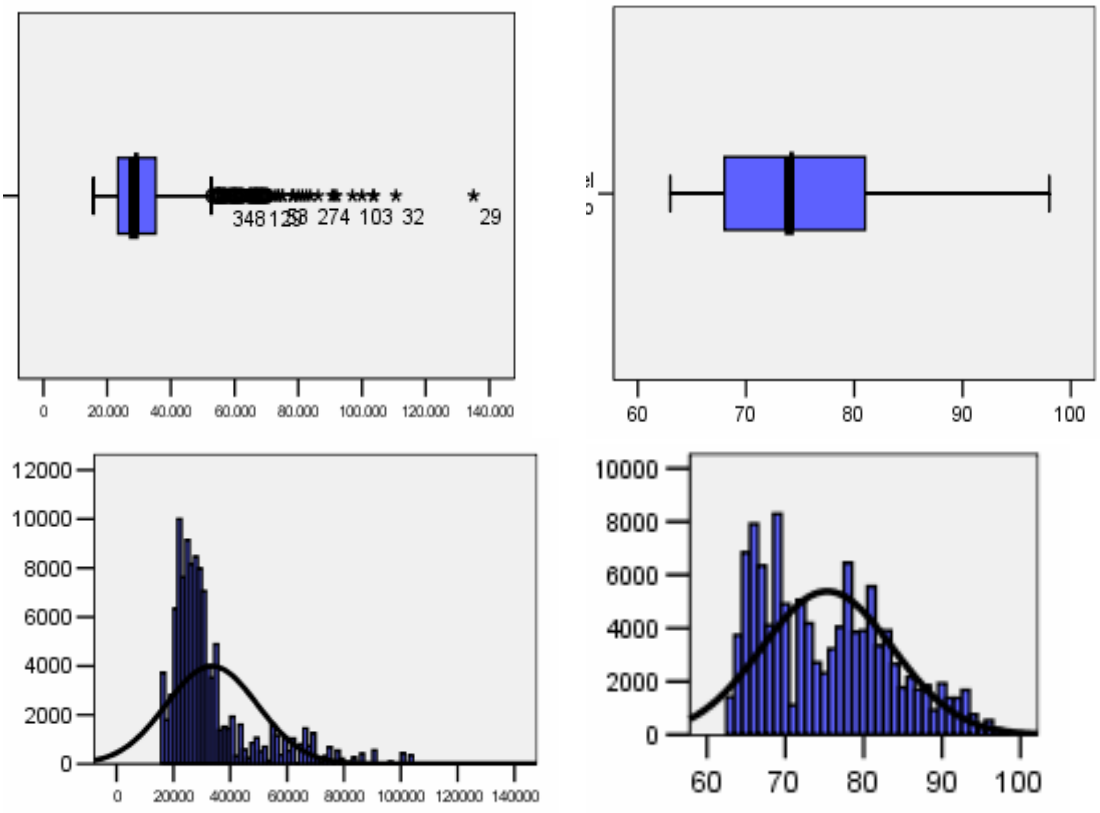


El gráfico de cajas es una representación gráfica no parametrica de los datos (ver grafico); los limites inferior y superior de la caja son los cuartiles primero

(Q_1) y tercero (Q_3); por lo tanto la caja contiene el 50% de los datos, la línea dentro de la caja indica cuál es la posición de la mediana: si esta línea no esta en el centro de la caja, indicaría la falta de simetría. Cuanto mayor es la longitud de la caja, mayor es la variabilidad de las observaciones.
Las líneas que se extienden desde cada lado de la caja se llaman bigotes; los bigotes van desde cada lado de la caja hasta la ultima observación cuyo valor es inferior a 1.5 veces el rango intercuartilico.
Los valores comprendidos entre 1.5 y 3 veces el rango intercuartilico se consideran valores atípicos moderados y se representan mediante el símbolo "**O**"; los valores a más de 3 veces el rango intercuantilico se consideran valores atípicos fuertes y se representan mediante el símbolo "*****".



Podemos comparar los histogramas con los gráficos de cajas para hacernos una idea de cómo van a resumir la información:



En el análisis numérico univariante de datos, podemos solicitar los siguientes estadísticos:

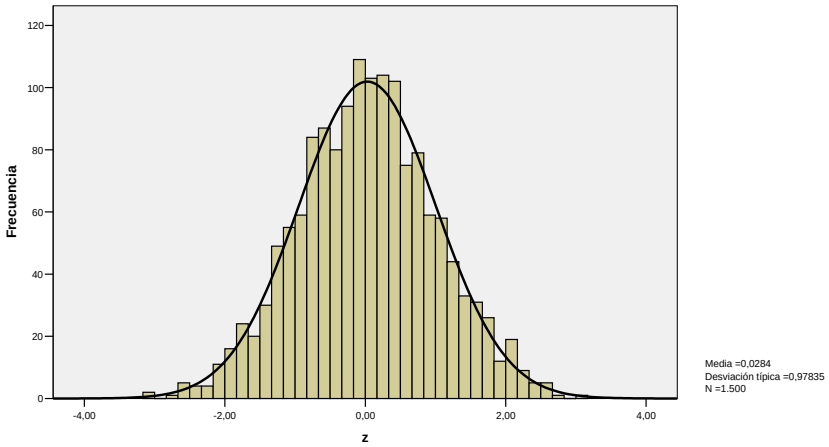
- Medidas de tendencia central: moda, media, media recortada y mediana.
- Medidas de dispersión: desviación típica, amplitud.
- Medidas de forma: coeficiente de asimetría y de curtosis.

El coeficiente de asimetría indica hasta qué punto una distribución es simétrica. Cuando la distribución es perfectamente simétrica el coeficiente de asimetría toma el valor 0, a medida que en valor absoluto el coeficiente de asimetría se va alejando de 0, la distribución va dejando de ser simétrica; en general si el valor absoluto del coeficiente de asimetría es mayor que 1, se puede considerar que la distribución no es simétrica.

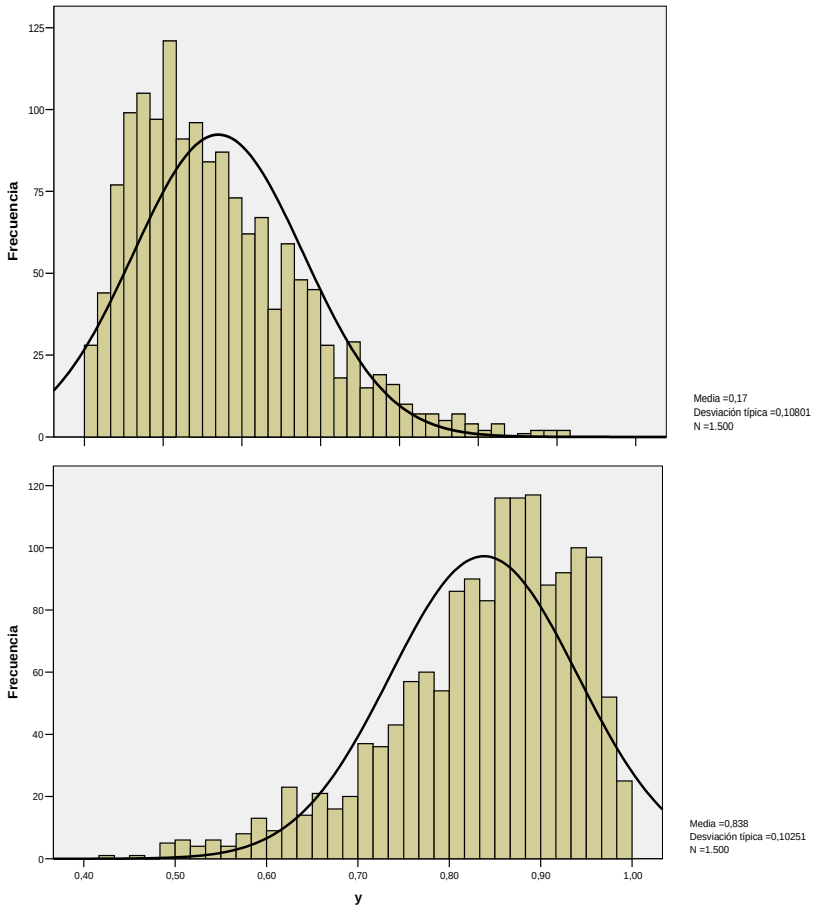
Estadísticos

		salario Salario actual	salini Salario inicial	educ Nivel educativo	expprev Experiencia previa (meses)	tiempemp Meses desde el contrato
N	Válidos	112575	112575	112575	112575	112575
	Perdidos	0	0	0	0	0
Media		33480,61	17051,47	13,40	95,66	75,33
Mediana		28350,00	15000,00	12,00	56,00	74,00
Desv. típ.		16066,938	7576,289	2,925	104,171	8,342
Asimetría		1,942	2,368	-,074	1,468	,474
Error típ. de asimetría		,007	,007	,007	,007	,007
Curtosis		3,874	6,932	-,347	1,518	-,723
Error típ. de curtosis		,015	,015	,015	,015	,015
Mínimo		15750	9000	8	0	63
Máximo		135000	79980	21	476	98

En los siguientes histogramas podemos comparar el coeficiente de asimetría con la forma del histograma.



CA = 0.012



CA = 0.97

CA= -0.97

El coeficiente de curtosis mide como se concentran los valores de la variable en torno al valor central, comparándolo con la distribución normal.

Un valor positivo del coeficiente de curtosis indica que los valores están más concentrados en torno a la media que en una distribución normal de los mismos parámetros (media y varianza); valores negativos indican que los valores están menos concentrados en torno a la media que en una distribución normal.

16- CONTRASTE DE HIPÓTESIS

Conocida la forma de la distribución, ya podemos contrastar una serie de hipótesis requeridas habitualmente para realizar análisis más complejos, como son normalidad y homocedasticidad.

Normalidad

En general, es muy buena condición para la aplicación de técnicas más sofisticadas el que la forma de la variable sea normal. La normalidad significa que la forma de la distribución de la variable se corresponde con la forma de una distribución normal; en la práctica se puede rebajar este requerimiento a dos condiciones:

- Que la distribución sea simétrica.
- Que la distribución sea unimodal.

Con un poco de práctica, es posible determinar el grado de normalidad de una variable simplemente examinando el histograma de la variable.

Existen contrastes de hipótesis para determinar el grado de confianza de que los valores obtenidos de una variable provengan de una distribución normal; el más empleado es el test de Kolgomorov-Smirnov.

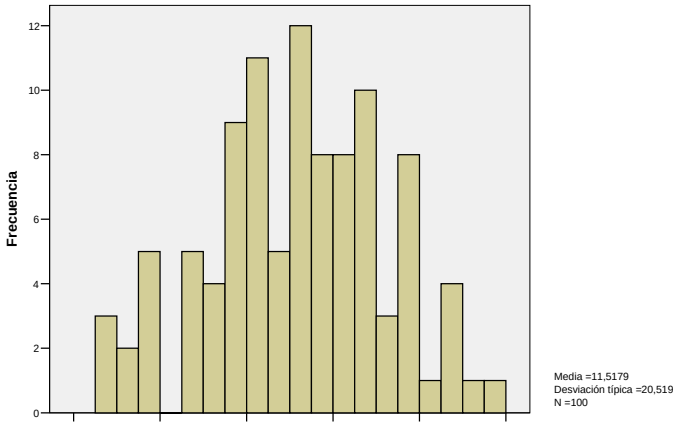
Para muestras pequeñas, el contraste de Kolgomorov pierde eficacia y es conveniente utilizar el test de Shapiro Wilks.

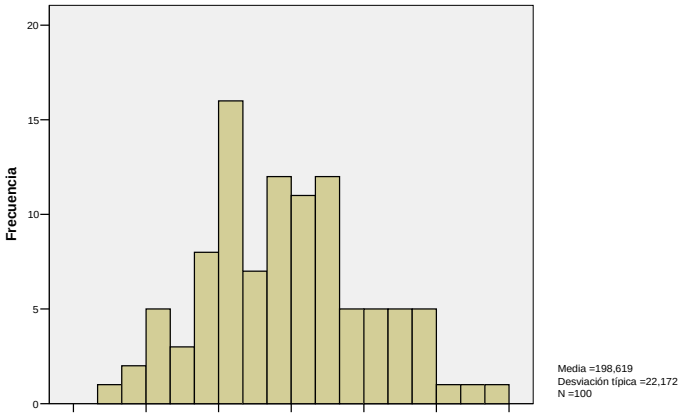
En ambos casos, la hipótesis nula de los test es que los valores provienen de una distribución normal, la hipótesis alternativa es que no siguen una distribución normal, es decir:

$$H_0 . X \approx N(\mu, \sigma)$$
$$H_1 . X \neq N(\mu, \sigma)$$

Por lo tanto, si rechazamos la hipótesis nula, estamos aceptando que la variable no sigue una distribución normal.

Sin embargo, en caso de no rechazar la hipótesis nula, no estamos aceptando que la variable sigue una distribución normal, sino que no hemos encontrado diferencias estadísticamente significativas con una distribución normal como para rechazar la hipótesis de normalidad.





Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
z	,043	100	,200*	,987	100	,452
v	,068	100	,200*	,989	100	,554

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

Homocedasticidad

Este concepto hace referencia a la variabilidad de la variable a través de los grupos definidos por los factores.
Se dice que una variable es homocedástica cuando dicha variabilidad (varianza) permanece constante a través de los distintos grupos definidos por los factores. En general, la variabilidad va a depender de la media, y ésta es posible que dependa de los grupos definidos por los factores.

Para verificar si se cumple esta propiedad, podemos utilizar el test de Levene y también observar el grafico de cajas; la amplitud de las cajas está relacionada directamente con la varianza de la variable.

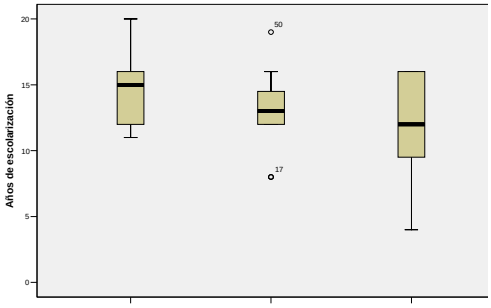
Existe una familia de transformaciones que aplicadas a una variable heterocedástica, la variable transformada resulta ser homocedástica, aunque debido a la gran dificultad para interpretar los resultados, en general, no se aplicará esta técnica salvo en casos muy concretos.

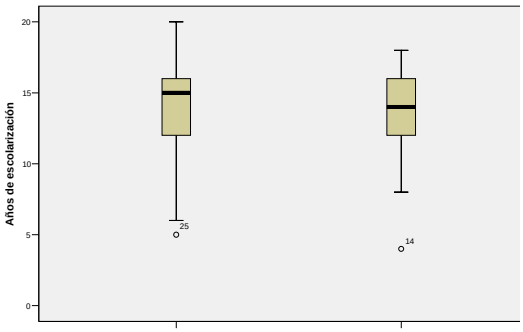
Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
educ Años escolarizac	Basándose en la	4,604	2	97	,012
	Basándose en la mediana.	4,283	2	97	,017
	Basándose en la mediana y con g corregido	4,283	2	72,045	,017
	Basándose en la recortada	4,404	2	97	,015

Prueba de homogeneidad de la varianza

		Estadístico de Levene	gl1	gl2	Sig.
educ Años escolarizaci	Basándose en la	,572	1	98	,451
	Basándose en la mediana.	,367	1	98	,546
	Basándose en la mediana y con gl corregido	,367	1	91,748	,546
	Basándose en la recortada	,507	1	98	,478





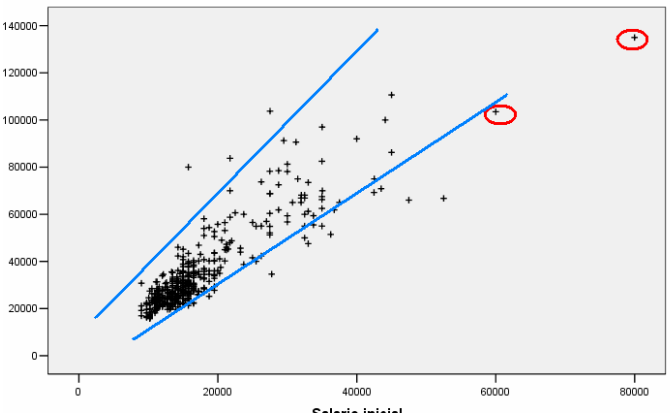
17- RELACIONES ENTRE LAS VARIABLES

Una vez realizados los análisis gráficos y numéricos unívariantes, vamos a estudiar si existen relaciones entre las variables mediante procedimientos gráficos muy sencillos. Existen técnicas multivariantes para detectar distintos tipos de relaciones entre conjuntos de variables, pero por su complejidad quedan fuera de esta etapa.

Las técnicas a utilizar van a depender del nivel de medida de las variables; enumeramos los casos más comunes.

Continua por continua

La forma más sencilla de determinar si existe relación entre dos variables con nivel de medida de escala o proporción es examinar, sin más, el diagrama de dispersión, como se muestra en el **siguiente ejemplo**.



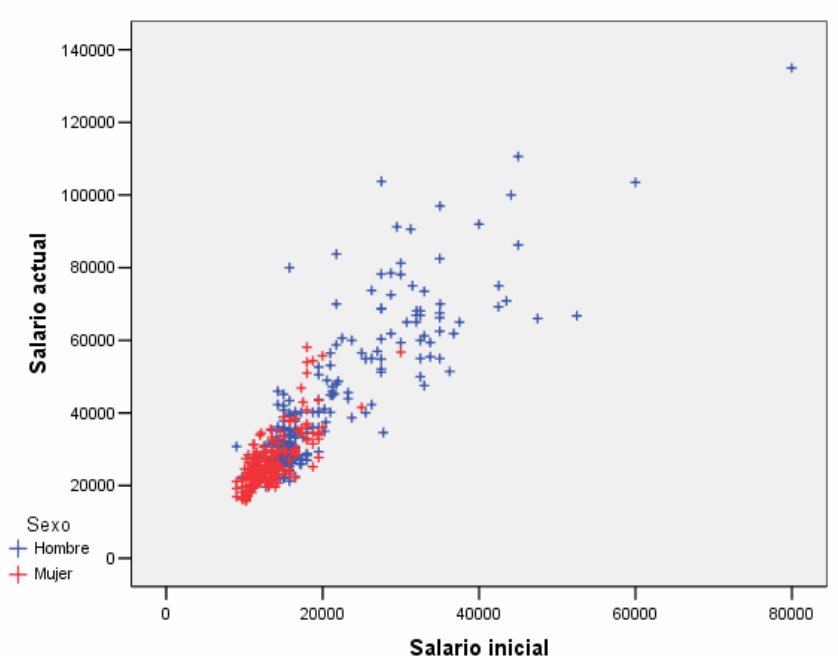
Este diagrama corresponde a la variable salario inicial (eje x) y al salario actual (eje y); podemos observar una clara tendencia lineal positiva, de donde se deduce que las personas que empezaron cobrando un salario bajo siguen cobrando un salario bajo y las personas que empezaron con un salario alto actualmente siguen cobrando un salario alto.
Además se pueden observar dos casos atípicos, que sin embargo siguen la tendencia general.

Continua por continua más una categórica

Para estudiar el efecto de una variable categórica sobre dos continuas solicitamos el diagrama de dispersión pero con 'marcas'; esto significa que los puntos del gráfico se colorean para poder identificar a qué categoría pertenecen.

Ejemplo

Supongamos que deseamos estudiar las variables salario inicial y salario actual, y, además, el género de las personas (sexo). Por lo dicho anteriormente realizamos un grafico de dispersión como el anterior pero esta vez añadiendo marcas.



Podemos observar que los puntos correspondientes a la categoría Mujer se sitúan en la esquina inferior izquierda, es decir, bajos salarios iniciales y bajos salarios actuales; en cambio los puntos correspondientes a los hombres están

repartidos más uniformemente, de donde se puede deducir que, en general, las mujeres cobran salarios bajos y medios pero no altos.

Categórica por categórica

Una forma de estudiar la asociación entre dos variables categóricas es estudiar la tabla de frecuencias de doble entrada y el coeficiente χ^2.

Ejemplo

Deseamos estudiar de forma conjunta la variable género y la variable catalogación laboral.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	79,277[a]	2	,000
Razón de verosimilitud	95,463	2	,000
N de casos válidos	474		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 12,30.

El p-value es inferior a 0.05, por lo que rechazamos la independencia estadística.
Si examinamos la tabla de frecuencias de doble entrada:

Tabla de contingencia sexo Sexo * catlab Categoría laboral

Recuento

		catlab Categoría laboral			
		1 Administrativo	2 Seguridad	3 Directivo	Total
sexo Sexo	h Hombre	157	27	74	258
	m Mujer	206	0	10	216
Total		363	27	84	474

En general es bastante complicado sacar conclusiones examinado la tabla de frecuencias directamente, es mucho mas practico el examinar la tabla de frecuencias condicionadas.

Tabla de contingencia sexo Sexo * catlab Categoría laboral

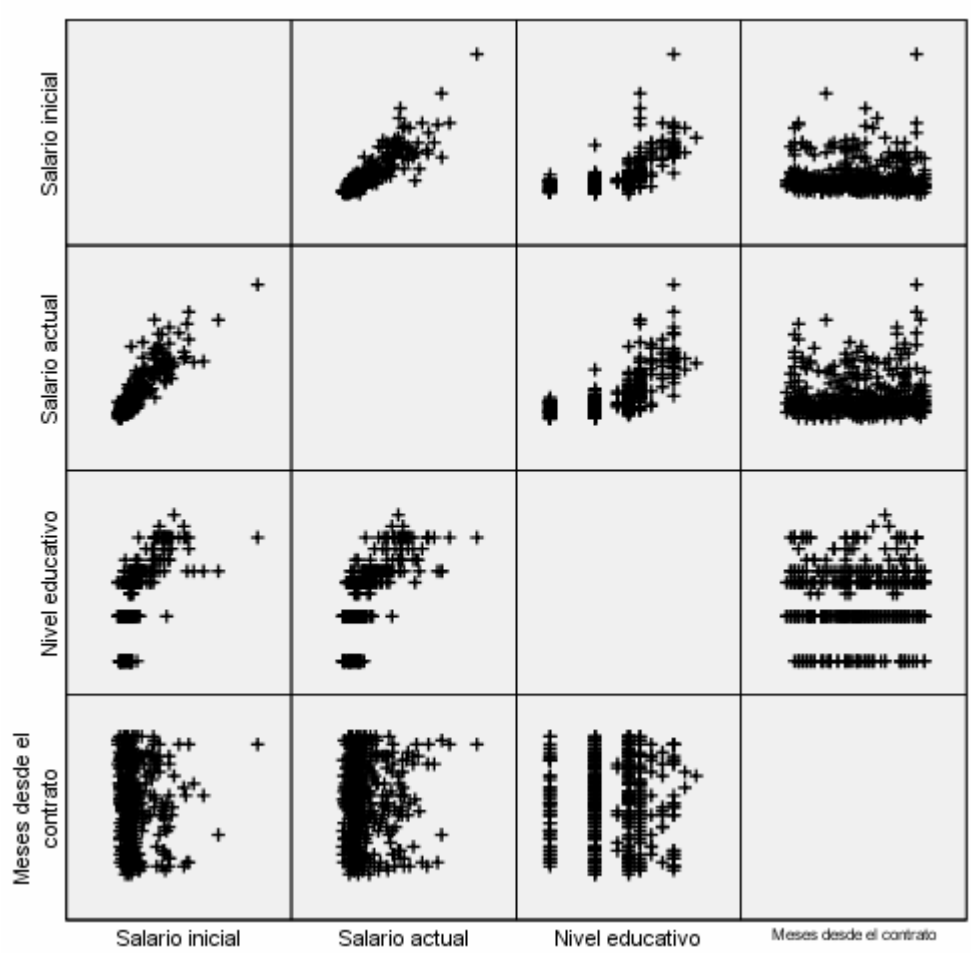
			catlab Categoría laboral			
			1 Administrativo	2 Seguridad	3 Directivo	Total
sexo Sexo	h Hombre	Recuento	157	27	74	258
		% de sexo Sexo	60,9%	10,5%	28,7%	100,0%
	m Mujer	Recuento	206	0	10	216
		% de sexo Sexo	95,4%	,0%	4,6%	100,0%
Total		Recuento	363	27	84	474
		% de sexo Sexo	76,6%	5,7%	17,7%	100,0%

Ahora es mucho más sencillo; podemos observar que:
- el 60.9% de hombres trabaja de administrativo frente al 95.4% de las mujeres.
- el 28.7% de los hombres trabajan de directivos frente al 4.6% de las mujeres.
- en seguridad, todos son hombres.

De ello se puede deducir que si bien hay más cargos de administrativos que de directivos, éstos, en general, suelen ser ocupados por hombres en mayor proporción que por mujeres.

Más de dos variables continuas

El gráfico matricial sirve para representar varios gráficos de dispersión en un mismo marco; se suele utilizar para estudiar la relación de más de dos variables continuas, aunque a medida que vamos añadiendo variables se va volviendo más difícil de interpretar.

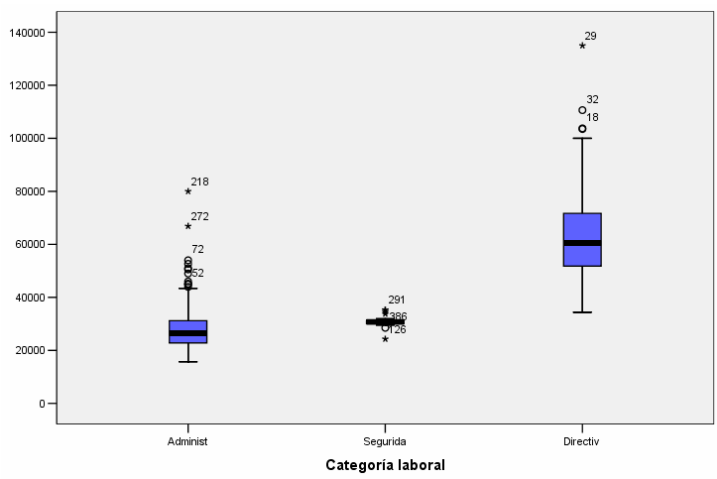


Continua por categórica

Para determinar como una variable categórica (puede ser ordinal) tiene sobre una variable continua podemos utilizar el gráfico de cajas con un factor. Este gráfico consiste en representar en un mismo cuadro las distribuciones definidas por las distintas categorías de la variable categórica.

Ejemplo

Deseamos estudiar la relación de la categoría laboral en el salario, para lo que realizamos un gráfico de cajas con un factor.



Estas tres cajas se corresponden con las tres categorías que tiene la variable ‘clasificación laboral’.

En primer lugar, podemos observar que, en general, los ‘directivos’ tienen un salario bastante superior a los ‘administrativos’ y al ‘personal de seguridad’, pues las cajas aparecen a distintas alturas.

También llama la atención la estrechez de la caja correspondiente al ‘personal de seguridad’ indicando que los sueldos de esa categoría son muy uniformes.

Podemos observar también la existencia de valores extremos tanto en la categoría de ‘administrativos’ como en la categoría de ‘directivos’.

18- VALORES ATÍPICOS

Los valores atípicos son valores (o grupos de valores) concretos que se destacan claramente de los demás por ser sucesos con poca probabilidad de ocurrencia y por tanto con poca probabilidad de aparecer en la muestra. Pongamos, por ejemplo, que en una muestra de 10 alumnos de secundaria hay 3 alumnos con una altura superior a 1.85m; éstos podrían ser considerados como valores atípicos.

Las causas por las que pueden aparecer valores atípicos suelen ser las siguientes:

- Errores de codificación: son errores producidos en el proceso de recogida de datos; por ejemplo, poner un cero de más accidentalmente. No siempre este tipo de errores se podrá detectar.
- Ocurrencia de un suceso extraordinario: en este caso hay que distinguir si no tiene explicación, o bien han cambiado las condiciones del experimento.

En los errores de codificación, si no es posible identificar el valor original de la variable, lo mejor es declararlos como valores perdidos por el usuario, pues pueden afectar muy negativamente a los análisis posteriores.

En cuanto a los valores atípicos debidos a la ocurrencia de un suceso con poca probabilidad de ocurrencia, hay que determinar hasta qué punto son o no son representativos de la población para su permanencia en la muestra; en caso que se decida que son representativos, es conveniente realizar un estudio de cómo esos casos con valores atípicos están afectando a los análisis, calculando su medida de influencia, incluso realizar dos análisis, uno con los valores atípicos incluidos y otro sin incluir.

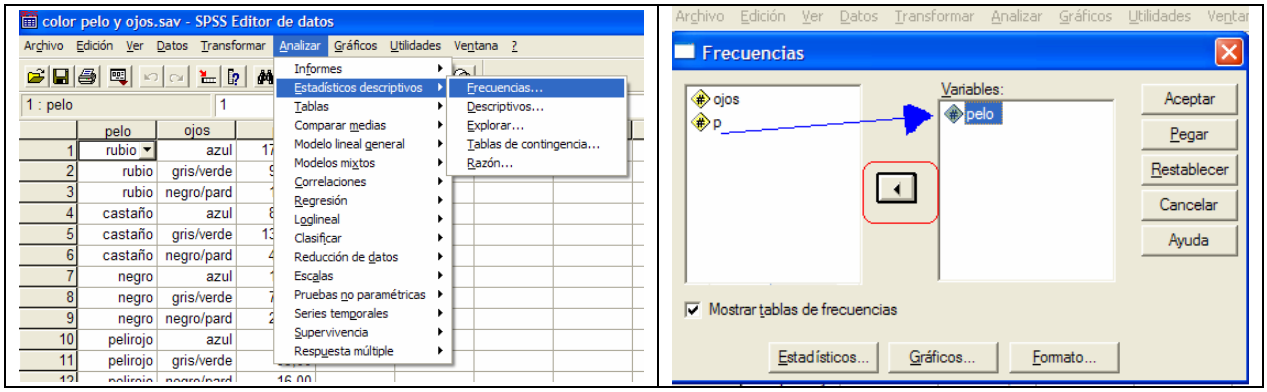
Cuando los valores atípicos son debidos a que han cambiado inadvertidamente las condiciones del experimento o de toma de datos, en realidad lo que tenemos son dos o más grupos de casos distintos, pues cada grupo esta asociado a distintas condiciones experimentales. En este caso no debemos de mezclar ambas situaciones, pudiendo ser recomendable estudiarlas por

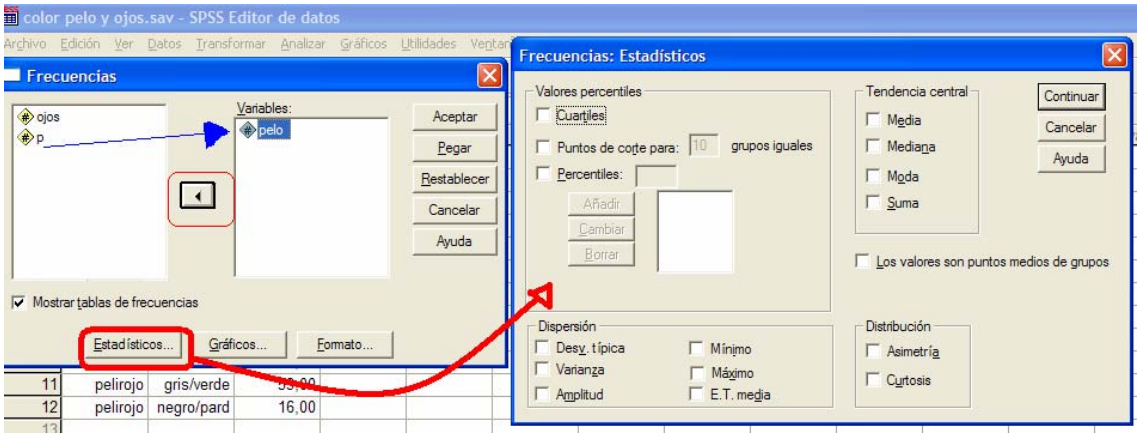
separado o bien eliminar los casos procedentes de las condiciones experimentales no controladas.

La detección de valores atípicos se puede realizar mediante el grafico de cajas y también calculando las puntuaciones tipificadas o puntuaciones Z.
Las puntuaciones tipificadas son transformaciones de las variables, eliminado las escalas de las mismas, lo cual permite comparar estos valores con los valores de una distribución normal y determinar cuál seria su probabilidad de ocurrencia.

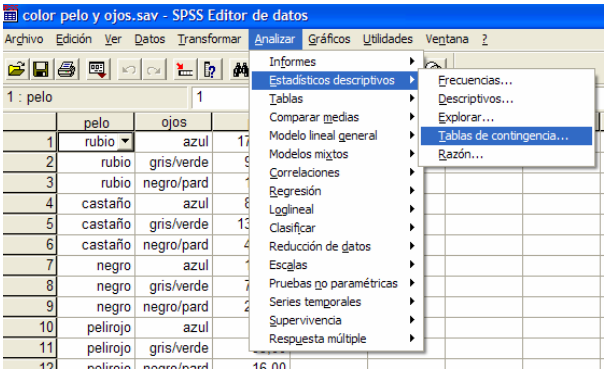
19- GUÍA VISUAL DE PROCEDIMIENTOS ESTADÍSTICOS CON SPSS- V13

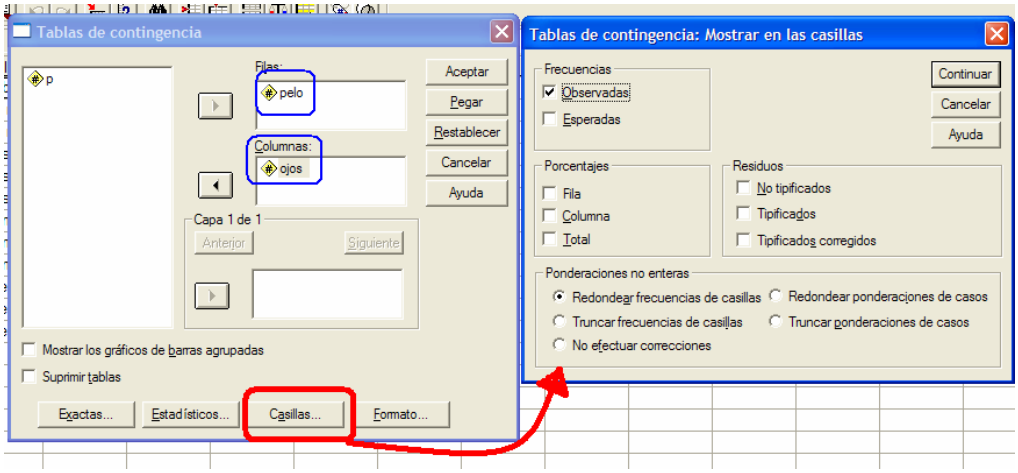
Tablas de frecuencias

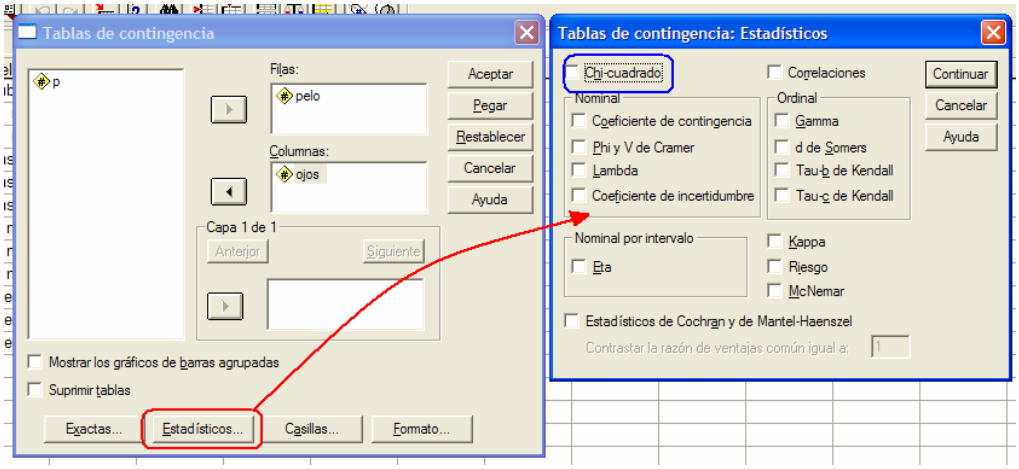




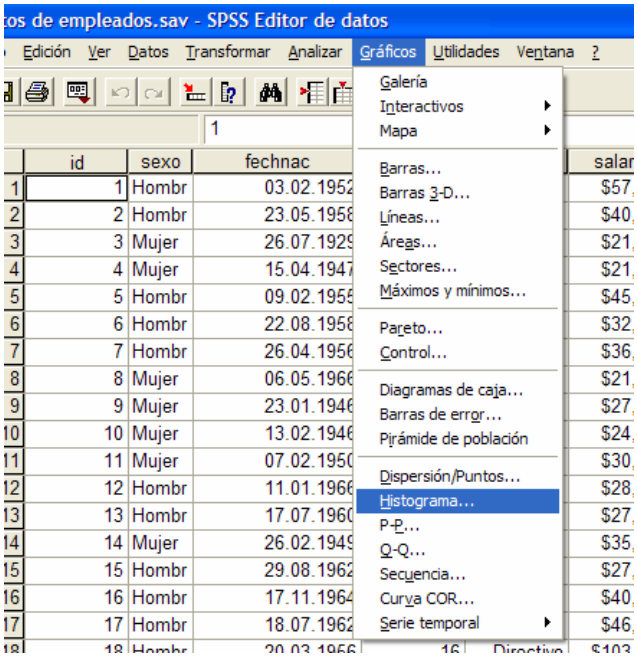
Tablas de frecuencias de doble entrada







Histograma


os de empleados.sav - SPSS Editor de datos
Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?
Galería
Interactivos
Mapa
Barras...
Barras 3-D...
Líneas...
Áreas...
Sectores...
Máximos y mínimos...
Pareto...
Control...
Diagramas de caja...
Barras de error...
Pirámide de población
Dispersión/Puntos...
Histograma...
P-P...
Q-Q...
Secuencia...
Curva COR...
Serie temporal
id sexo fechnac salar

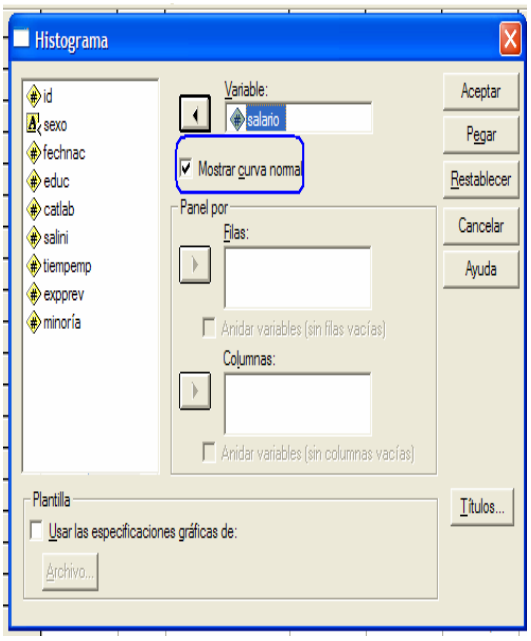
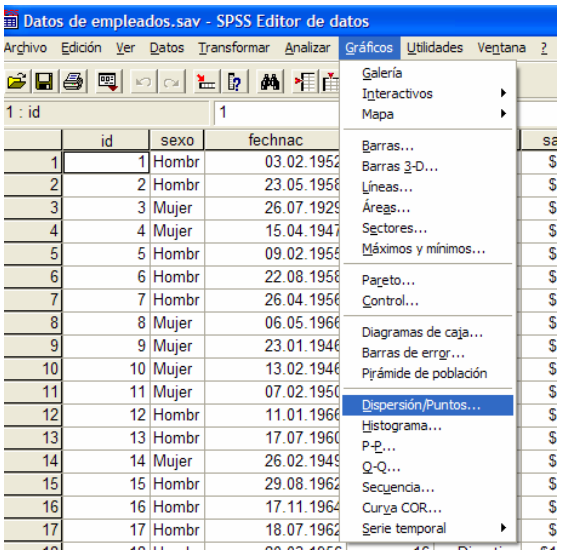
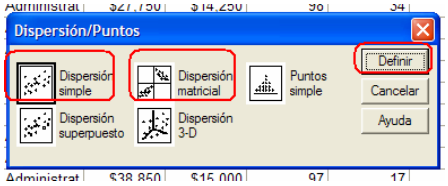

Histograma
id
sexo
fechnac
educ
catlab
salini
tiempemp
expprev
minoría
Variable:
salario
Mostrar curva norma
Panel por
Filas:
Anidar variables (sin filas vacías)
Columnas:
Anidar variables (sin columnas vacías)
Plantilla
Usar las especificaciones gráficas de:
Archivo...
Aceptar
Pegar
Restablecer
Cancelar
Ayuda
Títulos...

Diagrama de dispersión con o sin marcas





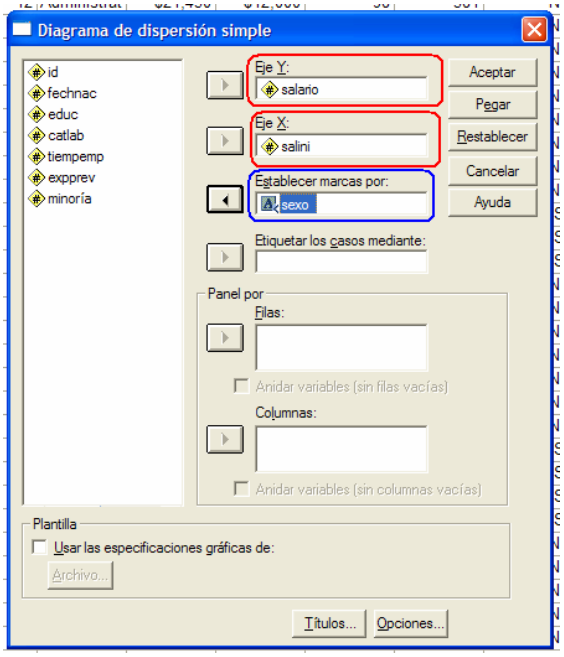
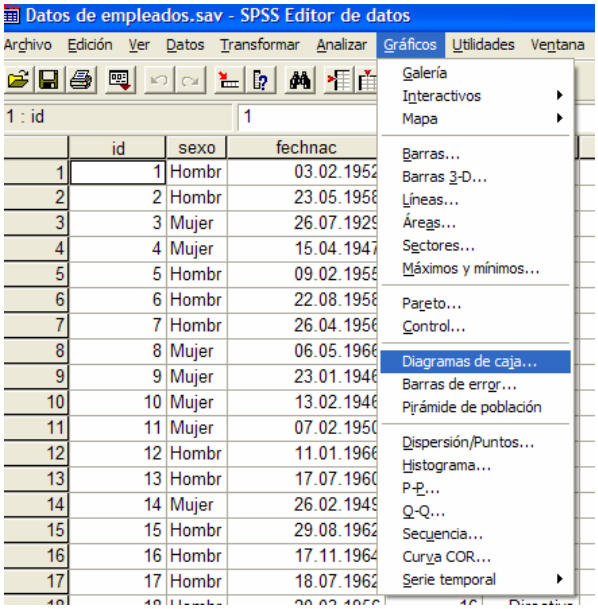
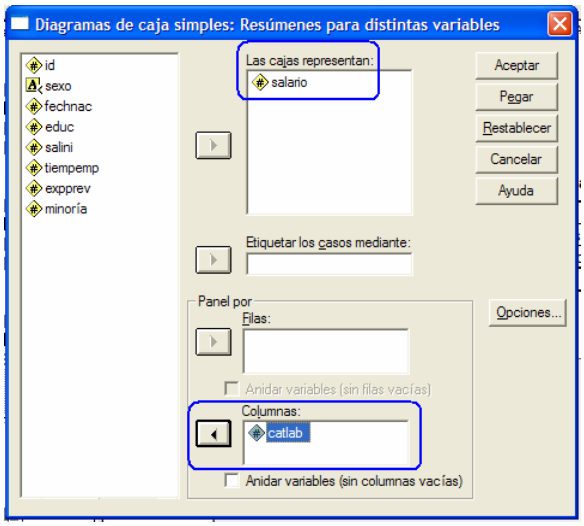


Gráfico de cajas





Fin