Actividad 2: SPSS

Materia: Entornos de computacion estadística

Autor: Francisco Márquez

contacto: franmarq@gmail.com
(mailto:franmarq@gmail.com)

Ejercicio 1: Análisis exploratorio de datos con SPSS

En el fichero *empleados.sav* se encuentra información relativa a 474 individuos.
Realizar un análisis exploratorio de las variables salario actual (*salario*) y meses desde el contrato (*tiempemp*), según categoría laboral (*catlab*) y etiquetando los casos según nivel educativo (*educ*)
Especificar en cada caso los análisis realizados e interpretar los resultados obtenidos
Definir y explicar el comportamiento y uso de las órdenes y reglas de sintaxis empleadas por SPSS

SOLUCION

Iniciamos el análisis observando algunas medidas de tendencia centrak y de dispersión para las variables de interés 'salario' y 'meses desde el contrato' agrupada por categoría ocupacional. Ara ello usamos la siguiente sintaxis:

```
EXAMINE
VARIABLES=salario tiempemp BY catlab
/PLOT BOXPLOT STEMLEAF HISTOGRAM NPPLOT
/COMPARE GROUP
/PERCENTILES (5,10,25,50,75,90,95) HAVERAGE
/STATISTICS DESCRIPTIVES EXTREME
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

El resultado Podemos resumirlo en el siguiente cuadro, algunas estadísticas fueron omitidas por ser poco relevantes:

Descriptives

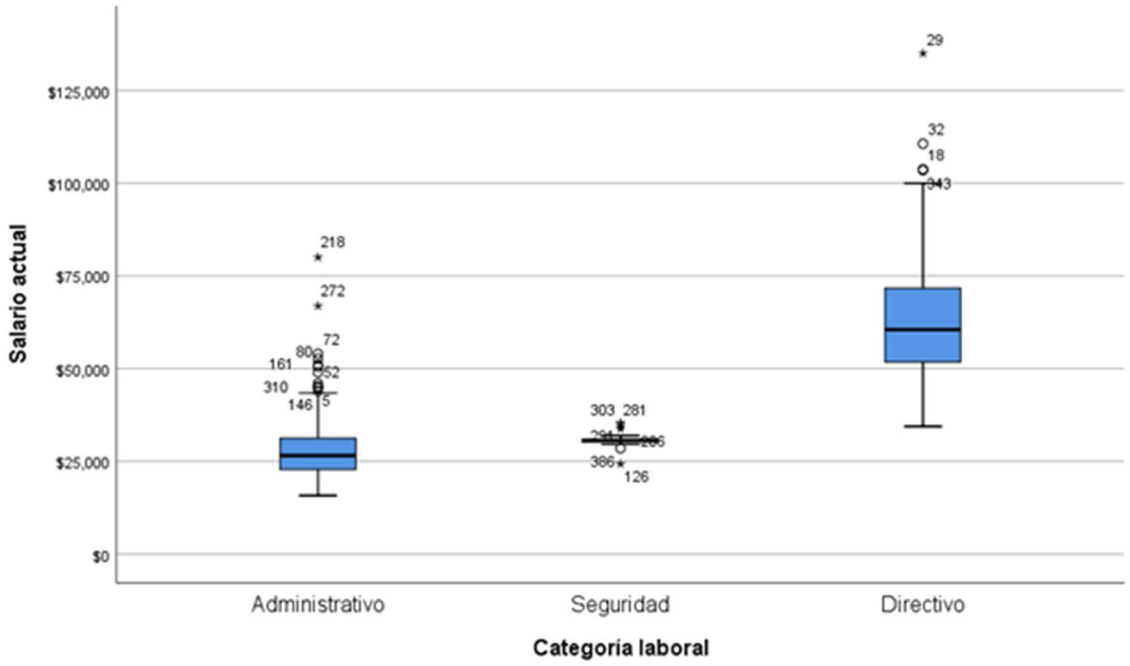
	Categoría laboral		Statistic	Std. Error
Salario actual	Administrativo	Mean	$27,838.54	$397.217
		Variance	57274547.724	
		Std. Deviation	$7,567.995	
		Minimum	$15,750	
		Maximum	$80,000	
	Seguridad	Mean	$30,938.89	$406.958
		Variance	4471602.564	
		Std. Deviation	$2,114.616	
		Minimum	$24,300	
		Maximum	$35,250	
	Directivo	Mean	$63,977.80	$1,990.668
		Variance	332871850.212	
		Std. Deviation	$18,244.776	
		Minimum	$34,410	
		Maximum	$135,000	
Meses desde el contrato	Administrativo	Mean	81.07	.531
		Variance	102.222	
		Std. Deviation	10.110	
		Minimum	63	
		Maximum	98	
	Seguridad	Mean	81.56	1.633
		Variance	72.026	
		Std. Deviation	8.487	
		Minimum	67	
		Maximum	95	
	Directivo	Mean	81.15	1.136
		Variance	108.373	
		Std. Deviation	10.410	
		Minimum	64	
		Maximum	98	

Entre los aspectos observados más destacados se encuentra que el *salario* promedio es notablemente diferente para los *Directivos* comparando con las otras categorías ocupacionales.

El tiempo medio de *meses desde el contrato* es muy similar en las tres categorías ocupacionales.

Variable Salario.

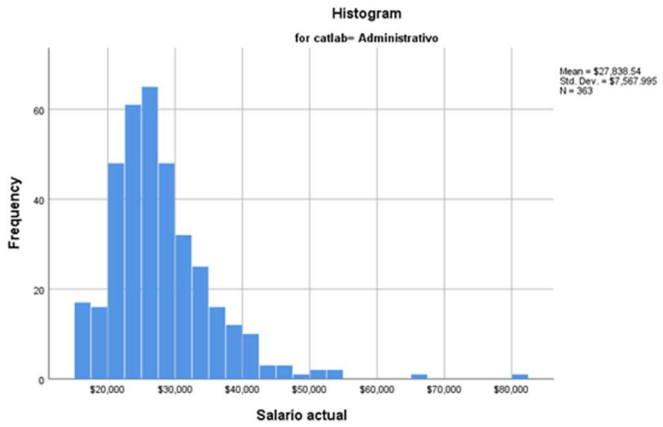
En primer lugar vamos comparar la distribución de la variable salario desde los grupos de la variable categoría ocupacional:



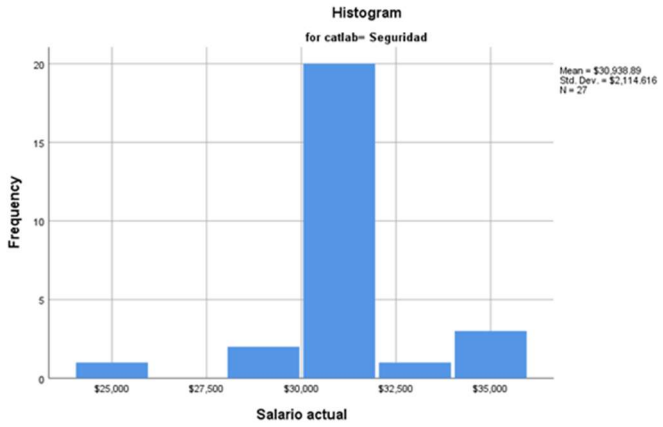
Se observan valores atípicos de *salario* para los tres grupos de categorías ocupacional. También se ve como la distribución más concentrada, y más pequeña, es la del grupo de Seguridad.

Se produjeron los siguientes Histogramas para entender la distribución del *salario* y *meses de contrato* por cada *categoría ocupacional.*

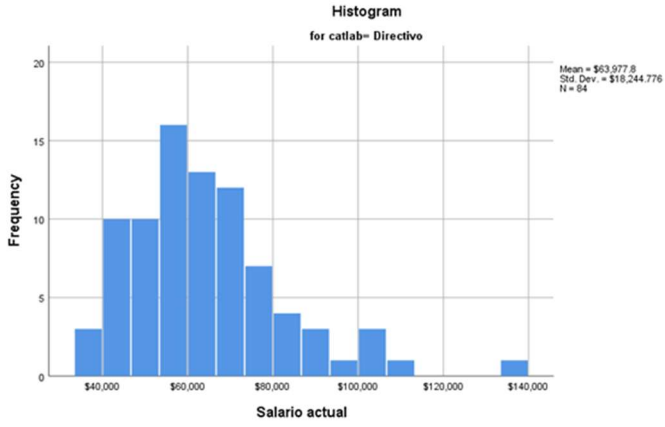
Salario vs categoría Ocupacional: Administrativo



Salario vs categoría Ocupacional: Seguridad



Salario vs categoría Ocupacional: Directivo



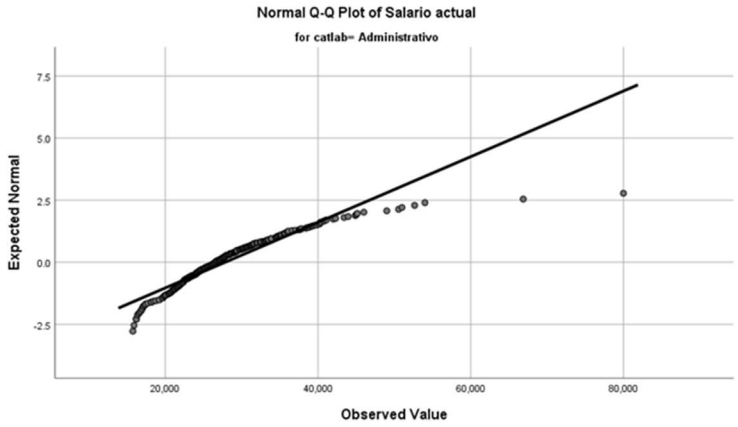
Podemos observar que las distribuciones del salario en el caso de Administrativo y Directivos son acampanadas, la distribución para seguridad es la que muestra menor dispersión.

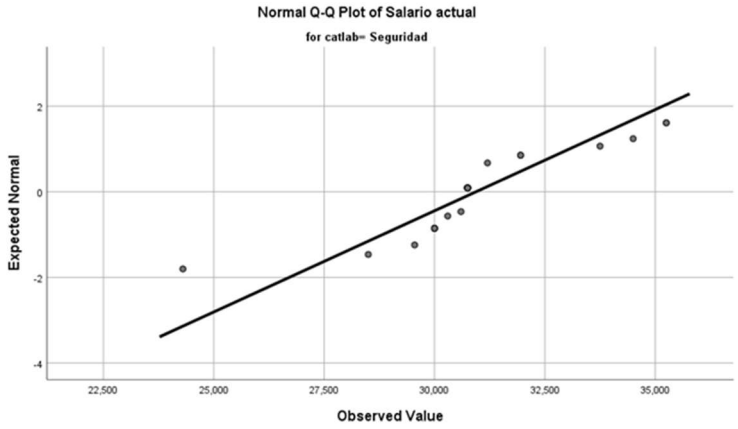
Para completar el análisis exploratorio, se generaron gráficos de tallo y hojas, resultando los siguientes:

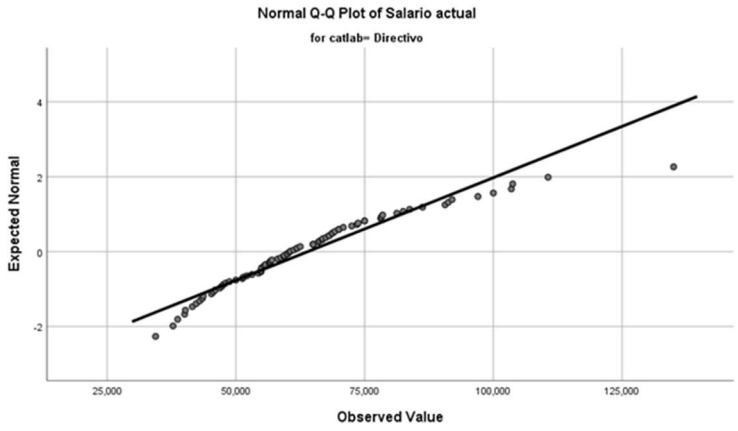
```
Salario actual Stem-and-Leaf Plot for
catlab= Administrativo

 Frequency    Stem &  Leaf

     2.00        1 .  55
    16.00        1 .  6666666666777777
    15.00        1 .  888899999999999
    35.00        2 .  00000000000001111111111111111111111
    44.00        2 .  22222222222222222222222222223333333333333333
    53.00        2 .  44444444444444444444444444445555555555555555555555
    55.00        2 .  66666666666666666666666666677777777777777777777777777777
    35.00        2 .  88888888888888889999999999999999999
    30.00        3 .  000000000000000111111111111111
    19.00        3 .  2222223333333333333
    17.00        3 .  44444444555555555
    11.00        3 .  66666677777
     8.00        3 .  88889999
     8.00        4 .  00000001
     3.00        4 .  223
    12.00 Extremes    (>=43950)

 Stem width:     10000
 Each leaf:        1 case(s)


Salario actual Stem-and-Leaf Plot for
catlab= Seguridad

 Frequency    Stem &  Leaf

     2.00 Extremes    (=<28500)
     1.00       29 .  5
     5.00       30 .  00003
    12.00       30 .  677777777777
     1.00       31 .  2
     2.00       31 .  99
     4.00 Extremes    (>=33750)

 Stem width:      1000
 Each leaf:        1 case(s)


Salario actual Stem-and-Leaf Plot for
catlab= Directivo

 Frequency    Stem &  Leaf

     3.00        3 .  478
    15.00        4 .  001233355667788
    21.00        5 .  0112344455555566678899
    21.00        6 .  000011125556666788889
    11.00        7 .  00023355888
     4.00        8 .  1236
     4.00        9 .  0127
     1.00       10 .  0
     4.00 Extremes    (>=103500)

 Stem width:     10000
 Each leaf:        1 case(s)
```

Puede verse muy claro que la cateroria Directivos acumula muchas más observaciones que las otras dos.

Ahora validemos supuesto de Normalidad, para ello generamos el grafico cuantil cuantil:



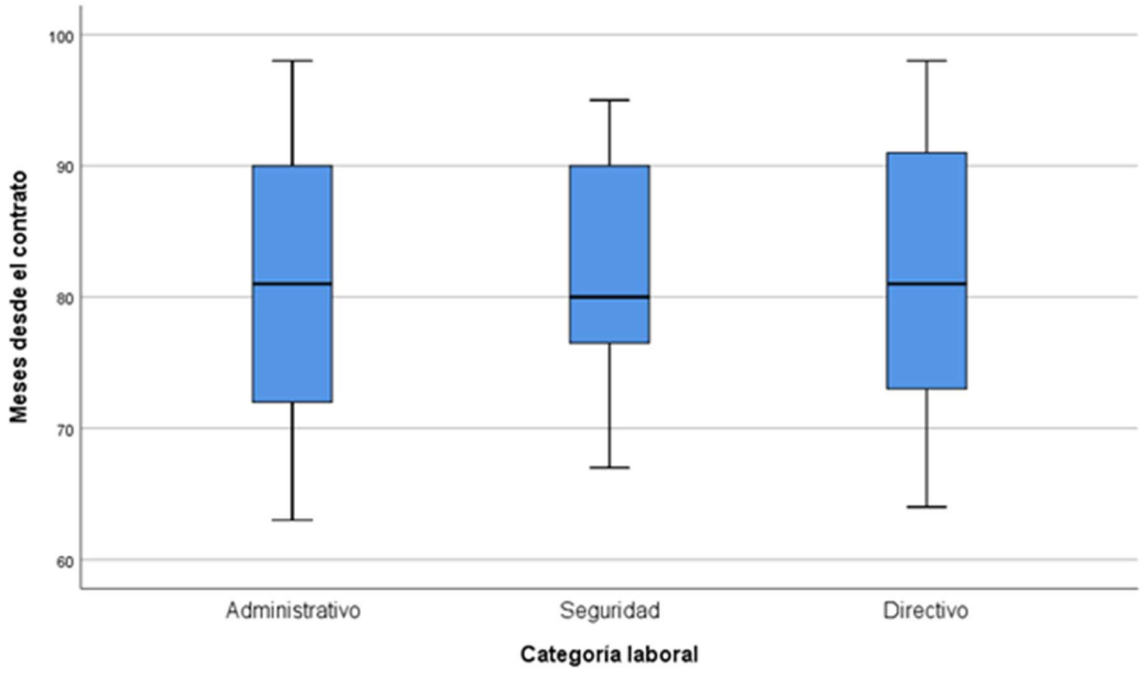




Podemos ver que en los tres grupos de categoría ocupacional, la variable salario presenta un comportamiento ajustado a la reca lo que sugiere similitud de la distribuciónes con la distribución normal.

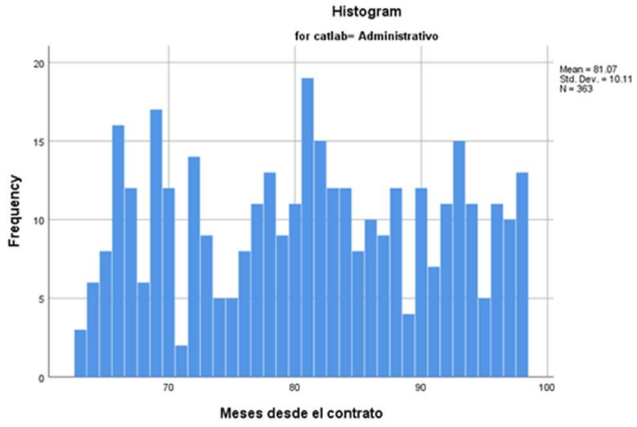
Variable Meses de contrato

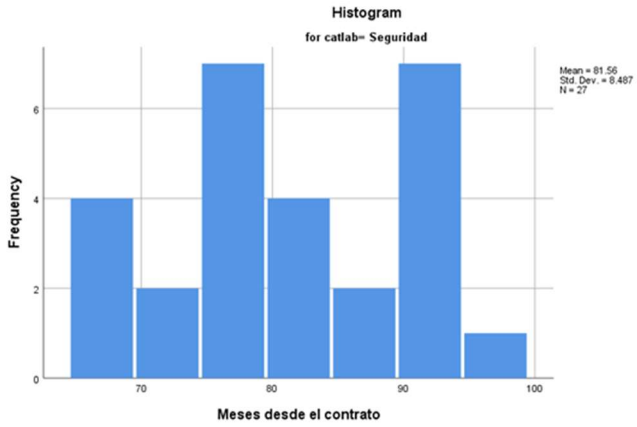
Comparamos la distribución de la variable de acuerdo a las categorías ocupacionales.

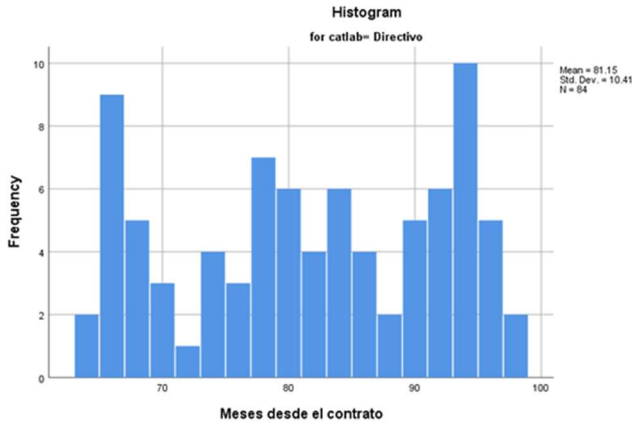


Se ve de forma muy clara que la distribución de los meses de contrato es similar en los tres grupos a diferencia de la variable salario.

A continuación se generaron los Histogramas de las variables para grupo de categoría ocupacional.







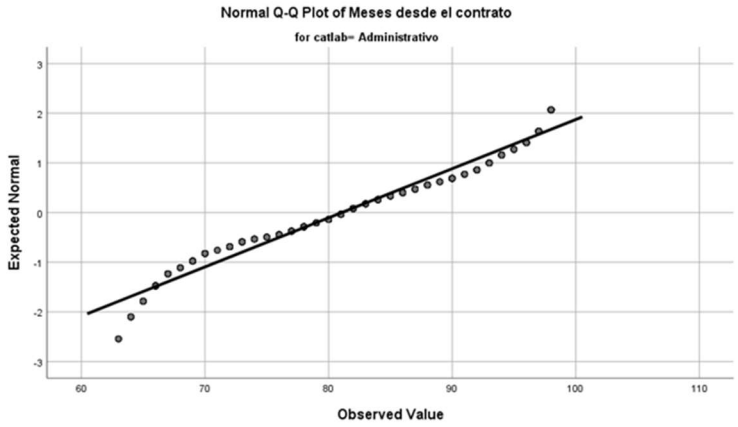
Podemos observar que en la distribución en los tres casos no presenta forma unimodal y por su forma no es acampanada. Esto pudiera sugerir ausencia de comportamiento ‘’normal’.

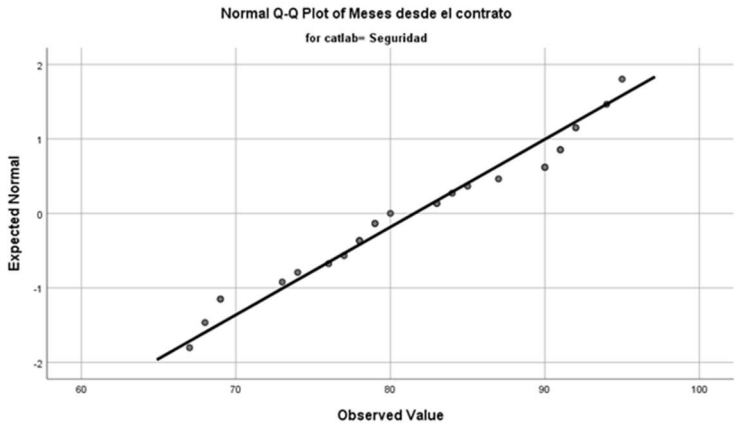
A continuación fueron generados gráficos de tallo y hojas para los tres grupos.

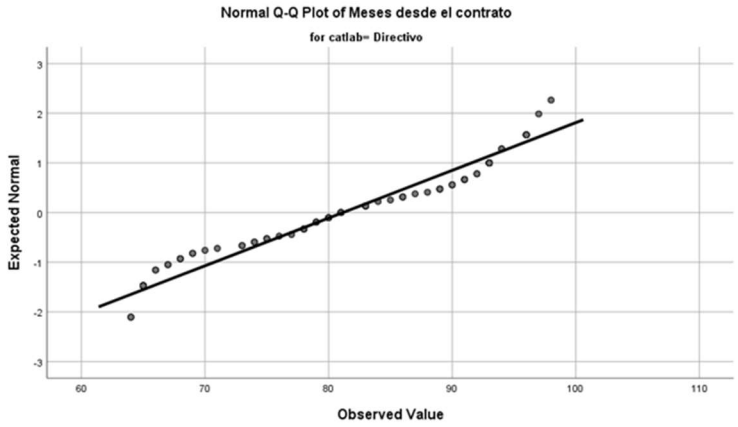
```
Meses desde el contrato Stem-and-Leaf Plot for
catlab= Administrativo

 Frequency    Stem &  Leaf

     3.00        6 .  333
    14.00        6 .  44444455555555
    28.00        6 .  6666666666666666777777777777
    23.00        6 .  88888899999999999999999
    14.00        7 .  00000000000011
    23.00        7 .  22222222222222333333333
    10.00        7 .  4444455555
    19.00        7 .  6666666677777777777
    22.00        7 .  88888888888999999999
    30.00        8 .  000000000001111111111111111111
    27.00        8 .  222222222222222333333333333
    20.00        8 .  44444444444455555555
    19.00        8 .  66666666677777777
    16.00        8 .  8888888888889999
    19.00        9 .  0000000000001111111
    26.00        9 .  22222222222333333333333333
    16.00        9 .  4444444444455555
    21.00        9 .  666666666667777777777
    13.00        9 .  8888888888888

 Stem width:  10
 Each leaf:        1 case(s)
```

```
Meses desde el contrato Stem-and-Leaf Plot for
catlab= Seguridad

 Frequency    Stem &  Leaf

     4.00        6 .  7899
     2.00        7 .  34
     7.00        7 .  6788899
     4.00        8 .  0334
     2.00        8 .  57
     7.00        9 .  0011224
     1.00        9 .  5

 Stem width:  10
 Each leaf:        1 case(s)


Meses desde el contrato Stem-and-Leaf Plot for
catlab= Directivo

 Frequency    Stem &  Leaf

     2.00        6 .  44
    16.00        6 .  5555555667788899
     6.00        7 .  013344
    13.00        7 .  5567888888999
    13.00        8 .  0001111333334
     9.00        8 .  566678999
    18.00        9 .  001111223333333344
     7.00        9 .  6666678

 Stem width:  10
 Each leaf:        1 case(s)
```

Para comprobar supuesto de normalidad fueron generados gráficos cuantil cuantil.







Aún cuando la distribución inicial no sugiere normalidad en estos grupos, los puntos cerca de la recta en el gráfico pueden indicar que es válido este supuesto.

Ejercicio 2: Cubos OLAP con SPSS

Analizar los datos del l fichero EncuestaUSA 1991.sav

Dar un resumen estadístico de la información almacenada en este fichero mediante la utilización de Cubos OLAP. Analizar las variables:

▪ edad (Edad del encuestado)

▪ educ (Número de años de escolarización)

▪ educpad (Número de años de escolarización del padre)

▪ educesp (Número de años de escolarización del cónyugue) prestg80

(Puntuación de prestigio profesional (1980) agrupadas según:

▪ sexo (Sexo del encuestado)

▪ catocu80 (Categoría ocupacional)

▪ obedecer (Obedecer es)

▪ trabajar (Trabajar duro es)

Definir y explicar el comportamiento y uso de las órdenes y reglas de sintaxis empleadas por SPSS

SOLUCION

Para generar los cubos OLAP se utilizó la siguiente sintaxis:

'Ejercicio 2: Cubos OLAP con SPSS'
'Data fuente'.
GET
FILE='C:\Users\santi\OneDrive\Desktop\Entornos de Computación Estadística\Actividad 2\Ficheros datos2 SPSS\EncuestaUSA 1991.sav'.
DATASET NAME DataSet2 WINDOW=FRONT.

'1. Construccion de los cubos'.
DATASET ACTIVATE DataSet2.
OLAP CUBES
edad educ educpad educesp by sexo catocu80 obedecer trabajar
/CELLS=SUM NPCT.

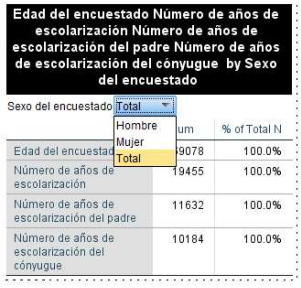
En primer lugar, este proceso genera una estadísticas de los casos considerados de acuerdo a las variables de cruce en donde ambas presentan valores:

(Parte de la tabla)

Case Processing Summary

	Cases					
	Included		Excluded		Total	
	N	Percent	N	Percent	N	Percent
Edad del encuestado * Sexo del encuestado	1514	99.8%	3	0.2%	1517	100.0%
Número de años de escolarización * Sexo del encuestado	1510	99.5%	7	0.5%	1517	100.0%
Número de años de escolarización del padre * Sexo del encuestado	1069	70.5%	448	29.5%	1517	100.0%
Número de años de escolarización del cónyugue * Sexo del encuestado	790	52.1%	727	47.9%	1517	100.0%
Edad del encuestado * Categoría ocupacional	1416	93.3%	101	6.7%	1517	100.0%

Luego se generan cada uno de los OLAP o reportes tabulares de varias dimensiones de acuerdo a los cruces indicados. A continuación se muestra el caso de los totales de la variables Edad, Educ, Edupad y Educesp la cual puede ser consultadas por sexo.



Edad del encuestado Número de años de escolarización Número de años de escolarización del padre Número de años de escolarización del cónyugue by Sexo del encuestado

Sexo del encuestado: Total (Hombre / Mujer / Total)

	um	% of Total N
Edad del encuestad	9078	100.0%
Número de años de escolarización	19455	100.0%
Número de años de escolarización del padre	11632	100.0%
Número de años de escolarización del cónyugue	10184	100.0%

Ahora en vez de Sexo obtenemos la tabla por:

Categoria ocupacional:

Edad del encuestado Número de años de escolarización Número de años de escolarización del padre Número de años de escolarización del cónyugue by Categoría ocupacional

Categoría ocupacional: Total

	Sum	% of Total N
Edad del encuestado	64840	100.0%
Número de años de escolarización	18423	100.0%
Número de años de escolarización del padre	11010	100.0%
Número de años de escolarización del cónyugue	9862	100.0%

La pregunta: Obedecer es

Edad del encuestado Número de años de escolarización Número de años de escolarización del padre Número de años de escolarización del cónyugue by Obedecer es

Obedecer es: Total

	Sum	% of Total N
Edad del encuestado	44439	100.0%
Número de años de escolarización	12637	100.0%
Número de años de escolarización del padre	7496	100.0%
Número de años de escolarización del cónyugue	6909	100.0%

La pregunta: Trabajar duro es

Edad del encuestado Número de años de escolarización Número de años de escolarización del padre Número de años de escolarización del cónyugue by Trabajar duro es

Trabajar duro es: Total

	Sum	% of Total N
Edad del encuestado	44439	100.0%
Número de años de escolarización	12637	100.0%
Número de años de escolarización del padre	7496	100.0%
Número de años de escolarización del cónyugue	6909	100.0%

Ejercicio 3: Regresión lineal con SPSS

En el fichero *Hatco.sav* se dispone de 100 observaciones referentes a 10 variables obtenidas a partir de encuestas realizadas a clientes de un distribuidor industrial.

A partir de las variables:

Percepciones de HATCO:	Resultados de compra:	Características del comprador:
X1 Velocidad de entrega **X2** Nivel de precios **X3** Flexibilidad de precios **X4** Imagen del fabricante **X5** Servicio conjunto **X6** Imagen de fuerza de ventas **X7** Calidad de producto	**X9 (Y)** Nivel de fidelidad	**X8** Tamaño de la empresa (variable codificada 0 – 1)

Predecir los niveles de fidelidad a los productos por parte de los clientes basándose en las percepciones que estos tienen de la actividad de HATCO, así como identificar los factores que llevan al aumento de la utilización del producto para su aplicación en campañas de marketing diferenciadas.

Definir y explicar el comportamiento y uso de las órdenes y las reglas de sintaxis empleadas por

SPSS

SOLUCION

Para generar el análisis de regresión, considerando las indicaciones del ejercicio, se ejecutó la siguiente sintaxis:

```
'Ejercicio 3: Regresion lineal con SPSS'.
'Data fuente'.
GET
FILE='C:\Users\santi\OneDrive\Desktop\Entornos de Computación Estadística\Actividad 2\Ficheros datos2
SPSS\Hatco.sav'.
DATASET NAME DataSet3 WINDOW=FRONT.
'1. Analisis de Regresion'.
DATASET ACTIVATE DataSet3.
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT y
/METHOD=ENTER x1 x2 x3 x4 x5 x6 x7 x8.
```

La salida nos muestra en primer lugar las variables consideradas en la construcción del modelo, siguiendo el método de selección de entrada.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	x8, x4, x5, x3, x7, x6, x2, x1[b]	.	Enter

a. Dependent Variable: y

b. All requested variables entered.

A continuación tenenos un resumen de los valores de análisis más relevantes asociados al modelo. El coeficiente de determinación R^2 nos sugiere que el modelo obtenido puede explicar de forma apropiada la variabilidad de los *niveles de fidelidad de los productos* a partir de las *percepciones de los clientes*.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.896[a]	.802	.785	4.1680

a. Predictors: (Constant), x8, x4, x5, x3, x7, x6, x2, x1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6418.138	8	802.267	46.181	.000[b]
	Residual	1580.862	91	17.372		
	Total	7999.000	99			

a. Dependent Variable: y

b. Predictors: (Constant), x8, x4, x5, x3, x7, x6, x2, x1

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		

1	(Constant)	-14.804	4.866		-3.043	.003
	x1	.689	1.908	.101	.361	.719
	x2	-.405	1.971	-.054	-.205	.838
	x3	3.990	.425	.615	9.385	.000
	x4	-.112	.629	-.014	-.178	.859
	x5	7.842	3.695	.655	2.123	.036
	x6	1.748	.902	.150	1.938	.056
	x7	-.159	.392	-.028	-.406	.686
	x8	5.281	1.486	.289	3.554	.001

a. Dependent Variable: y