

Metodología de la investigación

Pruebas de bondad de ajuste a una distribución normal

Autores

Romero-Saldaña, Manuel
 Doctor en Biomedicina
 Enfermero del Trabajo

Las pruebas de bondad de ajuste se utilizan para contrastar si los datos de la muestra pueden considerarse que proceden de una determinada distribución o modelo de probabilidad. Por ejemplo, cuando deseamos saber si los datos que manejamos proceden de una distribución nor-

mal, binomial, de Poisson, exponencial, etc.

En definitiva, las pruebas de bondad de ajuste permiten verificar qué tipo de distribución siguen nuestros datos y, por tanto, qué pruebas (paramétricas o no) podemos llevar a cabo en el contraste estadístico.

Tabla 1. Tipos de distribuciones de probabilidad

Tipo de variable	Tipo de distribución	Ejemplo
Cualitativa dicotómica	Bernoulli	Lanzar una moneda al aire: resultados posibles, cara o cruz.
	Binomial	Número de veces que sale cara cuando lanzamos la moneda 100 veces
	De Poisson	Número de mensajes que llegan a un servidor en una hora
Cuantitativa	Normal	Distribución del nivel de colesterol en una muestra de 100 personas
	Exponencial	Tiempo transcurrido entre la llegada de dos clientes consecutivos a una tienda

En definitiva, las pruebas de bondad de ajuste permiten verificar qué tipo de distribución siguen nuestros datos y, por tanto, qué pruebas (paramétricas o no) podemos llevar a cabo en el contraste estadístico.

A continuación, se exponen las principales pruebas de bondad de ajuste para contrastar la normalidad de nuestros datos (o sea, si nuestra muestra está extraída aleatoriamente de una población que sigue un modelo de probabilidad ajustado a la distribución normal).

Pruebas de bondad de ajuste para el contraste de distribución normal:

1. Prueba de Kolmogorov-Smirnov (K-S).
2. Test de Lilliefors (prueba de corrección para Kolmo

gorov).

3. Prueba de gráficos: Histograma, Q-Q Plots,
4. Prueba de Shapiro-Wilks.

1. PRUEBA DE KOLMOGOROV-SMIRNOV

Conocida como prueba K-S, es una prueba de significación estadística para verificar si los datos de la muestra proceden de una distribución normal. Se emplea para variables cuantitativas continuas y cuando el tamaño muestral es mayor de 50.

Ejemplo número 1. Se tiene una muestra de 636 trabajadores en la que deseamos saber si la variable edad sigue una distribución normal. La hipótesis nula (H₀) es que la

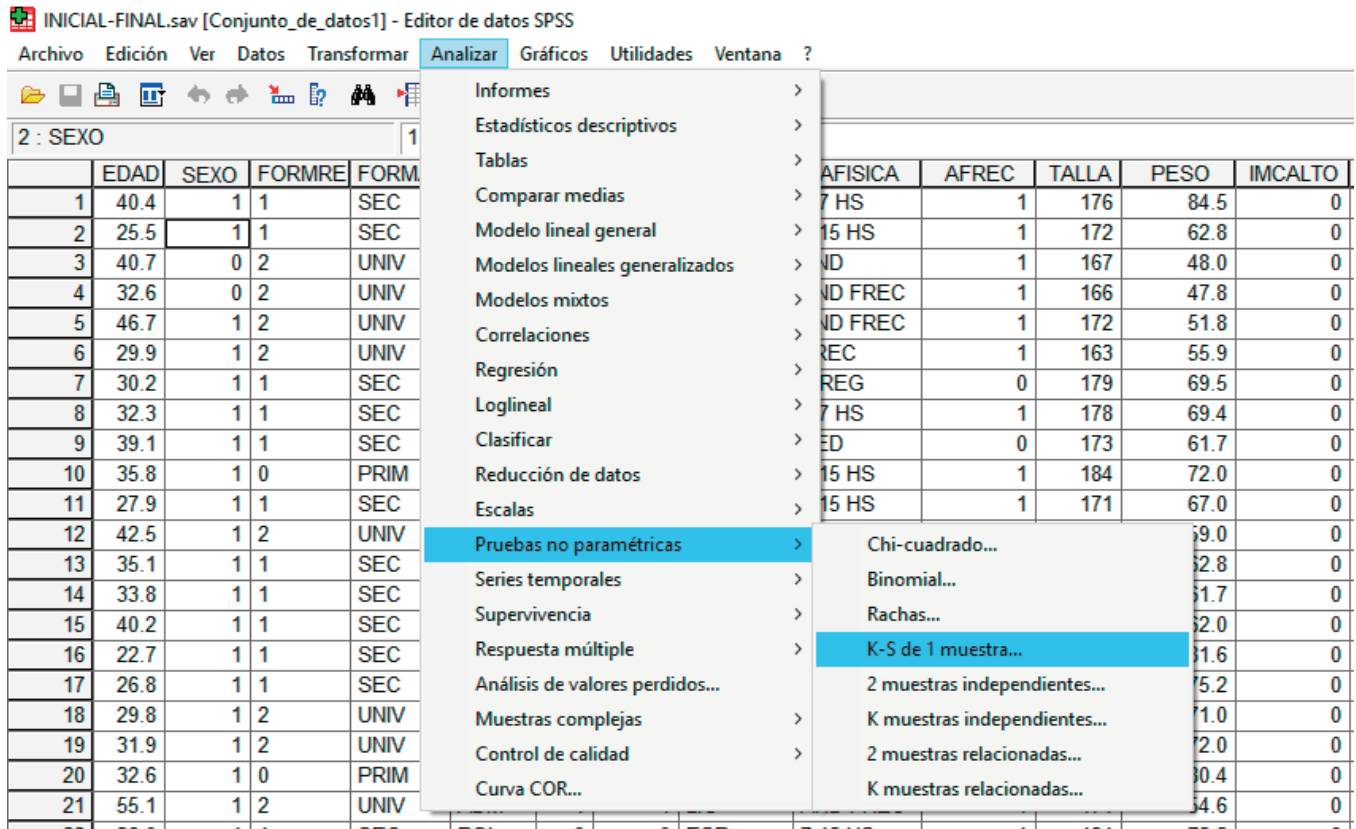


muestra procede de una distribución normal, mientras que la hipótesis alternativa (Ha) es que los datos no se distribuyen según un modelo de probabilidad normal. Por tanto, para aceptar Ho el

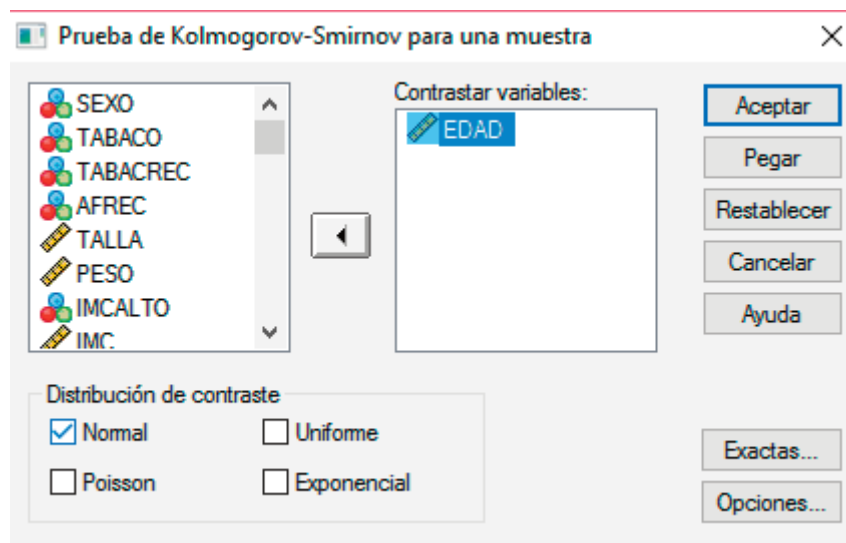
valor de la significación estadística (valor de p) deberá ser mayor de 0,05.

Para llevar a cabo el contraste, emplearemos el programa SPSS. El procedimiento sería el siguiente:

Paso 1. Hacemos click en ANALIZAR, Pruebas no paramétricas, K-S de 1 muestra



Paso 2. Introducimos la variable EDAD en “Contrastar variables” y seleccionamos la distribución de contraste “Normal”



Paso 3. Interpretamos la salida de resultados. Observamos los estadísticos descriptivos para la variable EDAD: N (número de registros analizados), media, desviación típica, mínimo y máximo y los percentiles 25, 50 y 75 que serían los cuartiles Q1, Q2 y Q3.

Estadísticos descriptivos

	N	Media	Desviación típica	Mínimo	Máximo	Percentiles		
						25	50 (Mediana)	75
EDAD	636	45.098	8.8007	22.7	66.8	39.061	44.562	51.531

Prueba de Kolmogorov-Smirnov para una muestra

		EDAD
N		636
Parámetros normales ^{a,b}	Media	45.098
	Desviación típica	8.8007
Diferencias más extremas	Absoluta	.037
	Positiva	.037
	Negativa	-.025
Z de Kolmogorov-Smirnov		.922
Sig. asintót. (bilateral)		.363

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

A continuación, valoramos los resultados de la prueba de contraste Kolmogorov-Smirnov para una muestra. Observamos que el estadístico Z ha obtenido un valor de 0.922 y que la significación estadística (sig. asintót. Bilateral) ha sido 0.363. Este sería el valor de p, y como es mayor de 0,05 se acepta Ho, o sea, la variable edad sigue una distribución normal y, en consecuencia, podremos aplicar pruebas paramétricas para el contraste estadístico.

2. TEST DE LILLIEFORS (PRUEBA DE CORRECCIÓN PARA KOLMOGOROV-SMIRNOV)

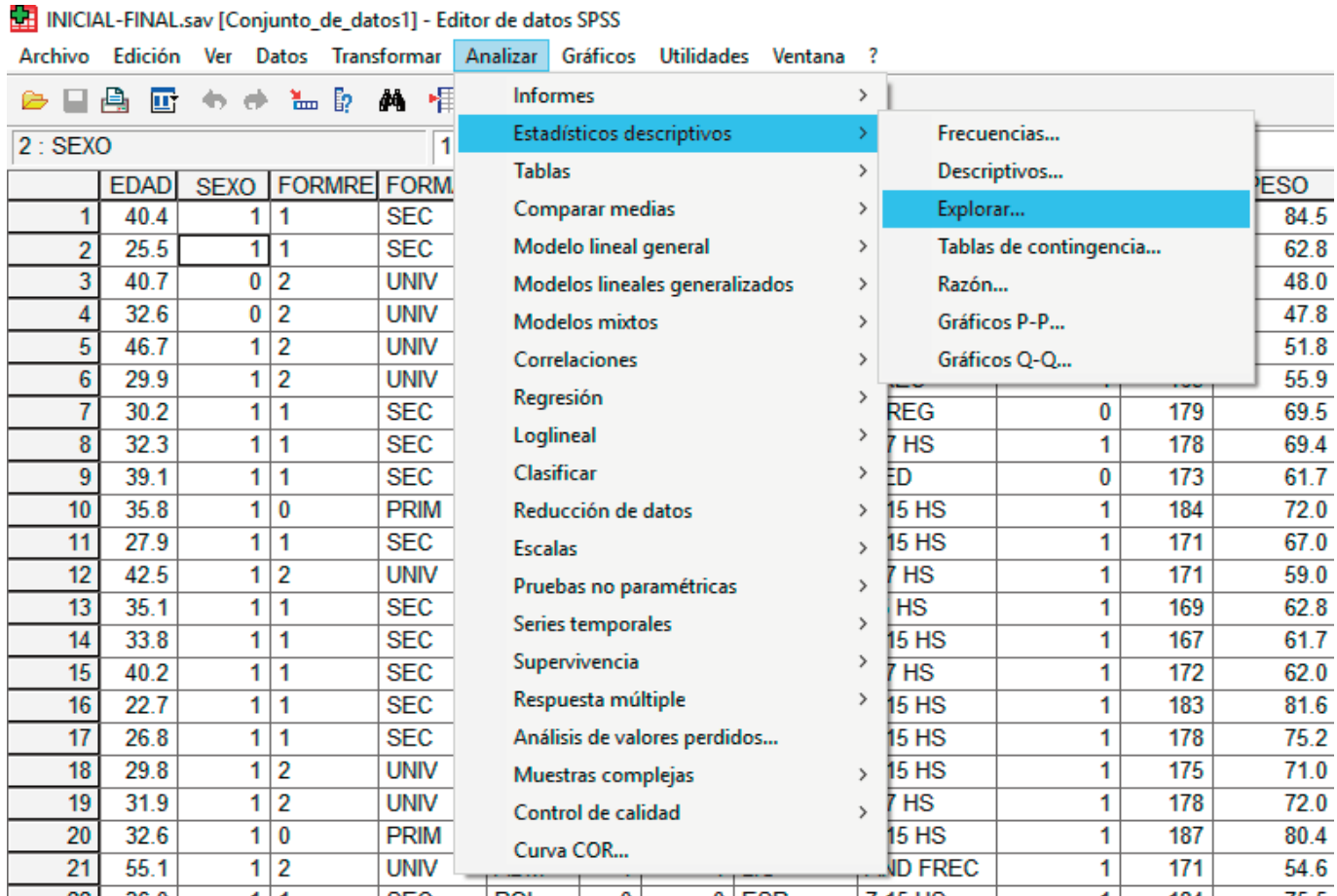
La prueba K-S para una muestra no es muy útil en la práctica, ya que en la gran mayoría de las veces descono-

ceamos cuál es la media y desviación estándar de la población, y por tanto, se deben estimar para la distribución teórica de comparación. Esto genera que la prueba K-S sea muy conservadora, aceptando la hipótesis nula en la mayoría de las ocasiones.

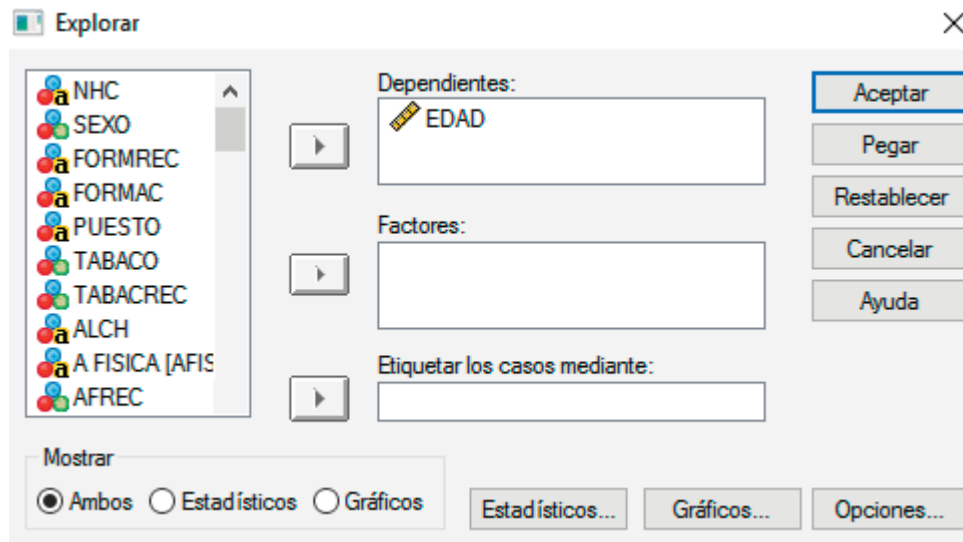
Para solventar este problema Lilliefors tabuló el estadístico de Kolmogorov-Smirnov para el caso más habitual en el que desconocemos la media y la varianza poblacional y se estiman a través de los datos muestrales.

Siguiendo con el ejemplo anterior, vamos a realizar la corrección de Lilliefors en la prueba K-S según el programa SPSS:

Paso 1. Abrimos el menú ANALIZAR, hacemos click en “Estadísticos descriptivos” y a continuación en “Explorar”



Paso 2. Introducimos la variable EDAD en “Dependientes” y en “Mostrar” activamos la pestaña de “Ambos”, o sea, que muestre estadísticos y gráficos.



Paso 3. A continuación valoramos la salida de resultados. Observamos los resultados de la estadística descriptiva (media, intervalo de confianza, mediana, desviación típica, asimetría, curtosis, etc.) y los principales percentiles

Descriptivos

			Estadístico	Error típ.
EDAD	Media		45.098	.3490
	Intervalo de confianza para la media al 95%	Límite inferior	44.412	
		Límite superior	45.783	
	Media recortada al 5%		45.101	
	Mediana		44.562	
	Varianza		77.452	
	Desv. típ.		8.8007	
	Mínimo		22.7	
	Máximo		66.8	
	Rango		44.1	
	Amplitud intercuartil		12.5	
	Asimetría		.019	.097
	Curtosis		-.523	.194

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Promedio ponderado(definición 1)	EDAD	30.130	33.216	39.061	44.562	51.531	57.224	60.268
Bisagras de Tukey	EDAD			39.064	44.562	51.514		

Por último, nos fijamos en las pruebas de normalidad “Kolmogorov-Smirnov” donde ya se ha realizado la corrección de Lilliefors.

Pruebas de normalidad

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
EDAD	.037	636	.042	.993	636	.005

a. Corrección de la significación de Lilliefors

Se observa en esta ocasión, que el estadístico ha obtenido un valor de 0.037 y que el valor de la significación estadística (p) ha sido 0.042, o sea, menor de 0,05 y, por tanto, se rechaza Ho. Aceptamos Ha, que significa que la variable edad no sigue una distribución normal.

En esta situación, donde la prueba simple K-S y la corrección de Lilliefors son contradictorias, es aconsejable realizar la prueba gráfica, donde podremos observar “gráficamente” si los datos de nuestra variable se distribuyen con normalidad o no.

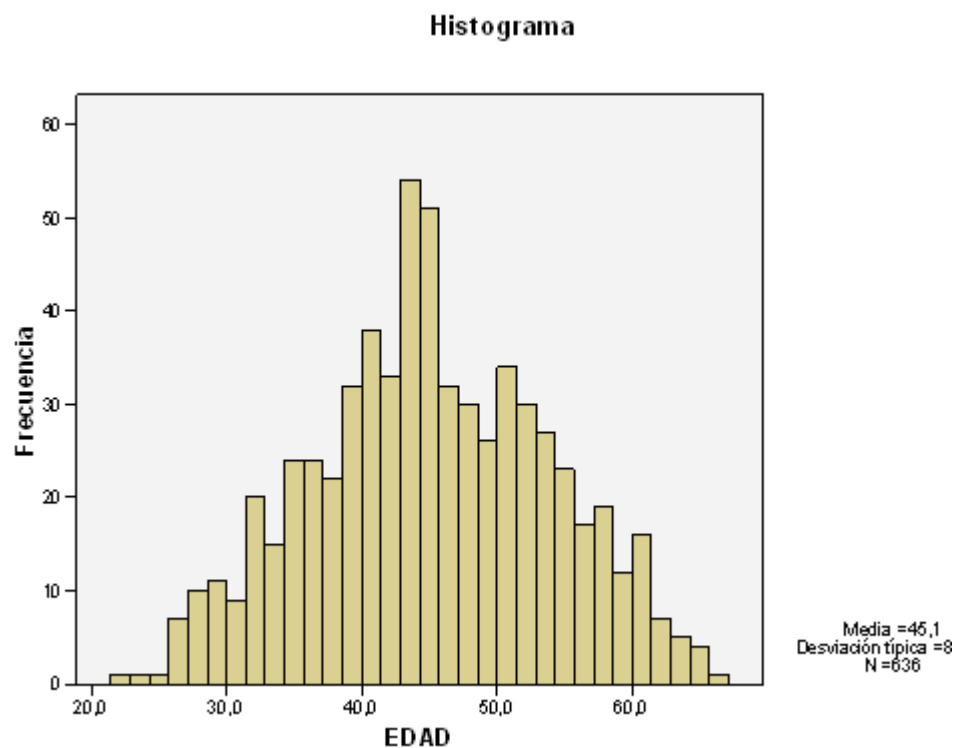
3. PRUEBA DE GRÁFICOS: HISTOGRAMA, Q-Q PLOTS

Hasta ahora hemos visto como, a través de pruebas de significación estadística, podemos determinar la bondad de ajuste de una muestra a una distribución normal. Por otra parte, podemos hacer algo parecido de una manera visual, mediante la observación de gráficos, los cuales nos orientan sobre la normalidad o no de la muestra. El uso de gráficos presenta varias ventajas, como por ejemplo, la sencillez de interpretación o la facilidad para obtener el diagrama a través de los propios paquetes es

tadísticos. Sin embargo, el principal inconveniente es la subjetividad de la interpretación visual, ya que al contrario de las pruebas de significación estadística, las pruebas gráficas no incluyen ningún valor de “p”.

3.1. HISTOGRAMA

El histograma de la variable EDAD está incluido en los resultados que SPSS genera cuando realizamos la prueba Kolmogorov-Smirnov con la corrección de Lilliefors. Por tanto, no tenemos que hacer nada nuevo.



Se puede observar como el histograma para la variable edad presenta una morfología muy parecida a una curva de Gauss, sin simetrías y, por tanto, podemos concluir que sigue una distribución normal.

3.2. GRÁFICOS Q-Q

Muestra los diagramas de probabilidad normal y de probabilidad sin tendencia. Los gráficos Q-Q representan los cuantiles de la distribución de una variable respecto a los cuantiles de cualquiera de las integrantes en una serie de distribuciones de contraste. Los gráficos de probabilidad se suelen utilizar para determinar si la distribución de una variable coincide con otra distribución especificada. Si la variable seleccionada coincide, los puntos se agruparán en torno a una línea recta.

La aplicación de la prueba de normalidad muestra dos gráficos:

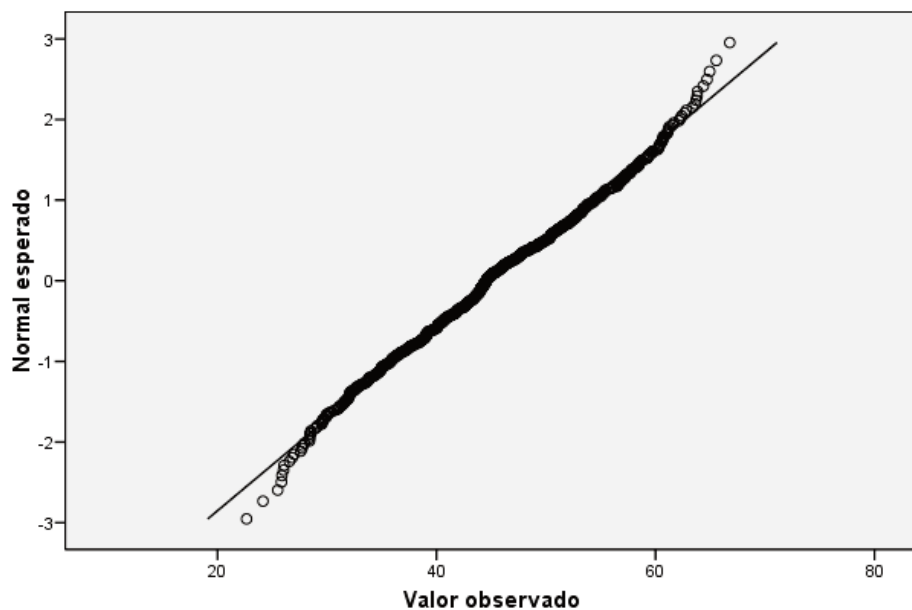
A. Gráfico Q-Q normal o Normal Probability Plot

Este gráfico se construye con parejas de valores, donde a cada valor observado se le empareja con su valor esperado, procedente éste último de una distribución normal. Si la muestra es extraída de una población normal ambos valores se encontrarán en la misma línea recta.

Este gráfico también se obtiene en los resultados de SPSS en la corrección de Lilliefors para K-S. Como podemos observar, prácticamente todos los puntos representados coinciden sobre la línea recta y, en consecuencia, podemos afirmar que la variable EDAD sigue una distribución normal.



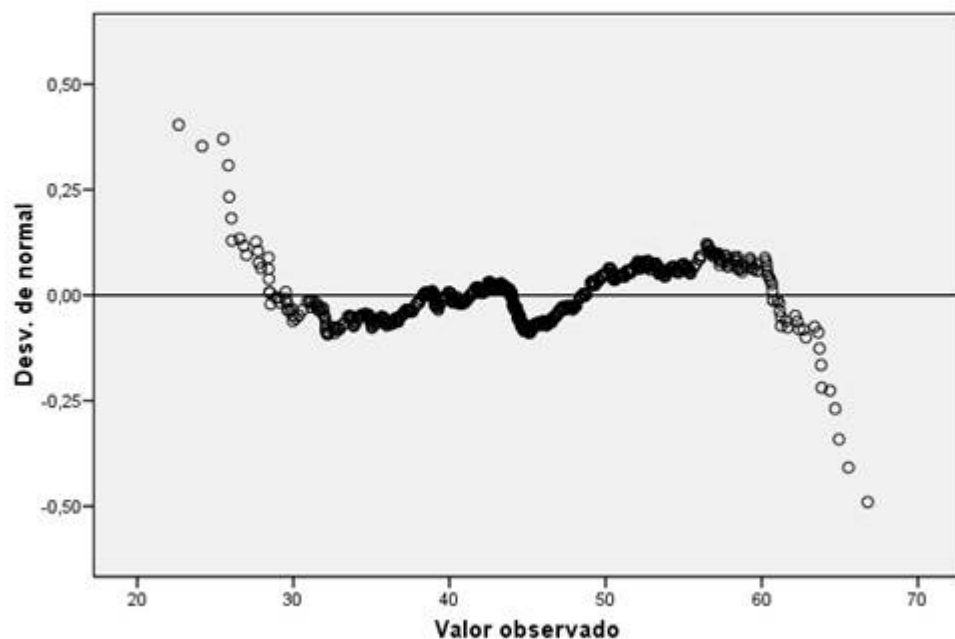
Gráfico Q-Q normal de EDAD



B. Gráfico Q-Q normal sin tendencias o Detrended normal Plot

En este caso, el gráfico muestra las desviaciones de los puntos con relación a una línea recta. Si la muestra ha sido extraída de una población normal los puntos deben situarse alrededor de una línea horizontal con el origen en el punto 0,00.

Gráfico Q-Q normal sin tendencias de EDAD



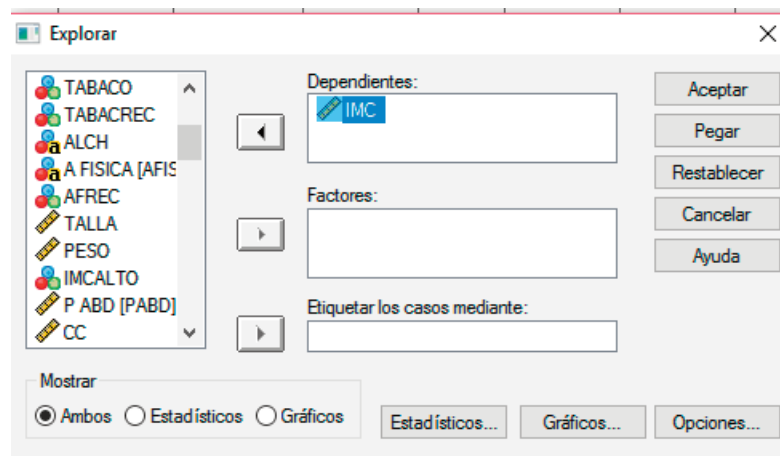
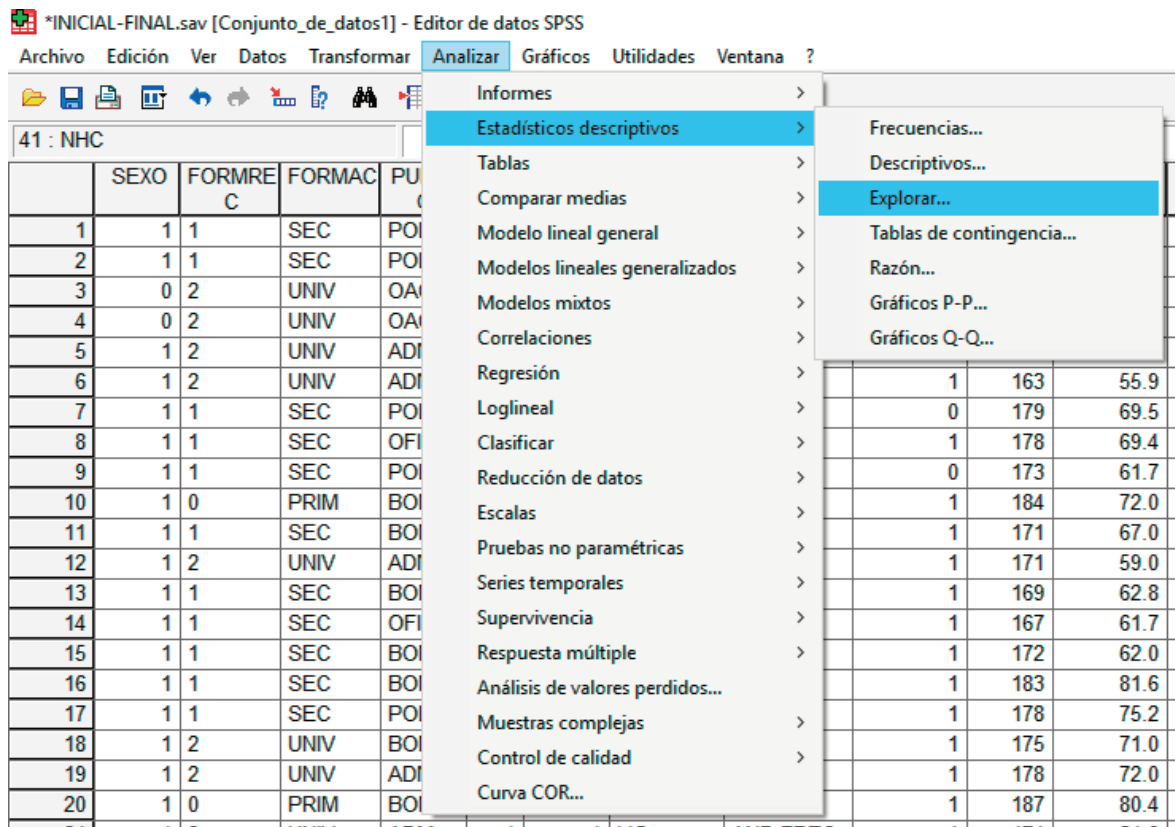


Como podemos observar, la gran mayoría de los datos de la variable edad se posicionan muy próximos a la línea central, y sólo los datos más bajo y los más altos (extremos de la distribución) se alejan de la esta línea. Podemos considerar que la variable EDAD sigue una distribución normal.

4. PRUEBA DE SHAPIRO-WILKS

Cuando el tamaño muestral es igual o inferior a 50 la prueba de contraste de bondad de ajuste a una distribución normal es la prueba de Shapiro-Wilks.

Ejemplo número 2. Tenemos una muestra de 40 trabajadores a los que hemos determinado su IMC. Queremos saber si esta variable sigue una distribución normal. Para ello, emplearemos el programa SPSS. Se seguirán los pasos 1, 2 y 3 recogidos en el apartado nº 2 (Test de Lilliefors para la corrección de la prueba K-S).





A continuación, valoramos los resultados obtenidos.

Resumen del procesamiento de los casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
IMC	40	100.0%	0	.0%	40	100.0%

Descriptivos

			Estadístico	Error típ.
IMC	Media		21.7988	.32121
	Intervalo de confianza para la media al 95%	Límite inferior	21.1491	
		Límite superior	22.4486	
	Media recortada al 5%		21.8190	
	Mediana		21.8702	
	Varianza		4.127	
	Desv. típ.		2.03150	
	Mínimo		17.23	
	Máximo		27.25	
	Rango		10.02	
	Amplitud intercuartil		2.62	
	Asimetría		-.216	.374
	Curtosis		.939	.733

El análisis descriptivo muestra los principales indicadores (media, intervalo de confianza, mediana, desviación típica, asimetría, curtosis, etc.), así como los principales percentiles.

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Promedio ponderado(definición 1)	IMC	17.3536	18.7420	20.6538	21.8702	23.2711	23.7581	24.6336
Bisagras de Tukey	IMC			20.6684	21.8702	23.2597		

Pruebas de normalidad

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
IMC	.109	40	.200*	.956	40	.124

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

a. Corrección de la significación de Lilliefors





Revista Enfermería del Trabajo 2016; 6:3 (105-114)
Romero Saldaña, Manuel
Pruebas de bondad de ajuste a una distribución normal

114

En este caso, nos fijamos en la prueba Shapiro-Wilks, donde el estadístico ha obtenido un valor de 0.956 y la significación estadística ($p=0.124$). Como p es mayor de 0.05, entonces, se acepta H_0 , y se afirma que la variable IMC sigue una distribución normal.

Referencias

1. Distribuciones de probabilidad más usuales. Disponible en: http://personales.unican.es/gonzaleof/Itop/teoria/Teoria_distribuciones.pdf. Consultado el 07-07-2016.
2. Universidad de Valencia. Las pruebas paramétricas. Disponible en: <http://www.uv.es/~friasnav/SupuestosParametrica.pdf>. Consultado el 07-07-2016.
3. Manual de ayuda de SPSS.

