Práctica 5

MÉTODOS DESCRIPTIVOS PARA DETERMINAR LA NORMALIDAD

Objetivos:

En esta práctica utilizaremos el paquete SPSS para determinar si los datos de una muestra dada provienen de una población normal, para así poder aplicar técnicas que se basan en el supuesto de que la población presenta una distribución normal aproximada.

Índice:

1. Histograma

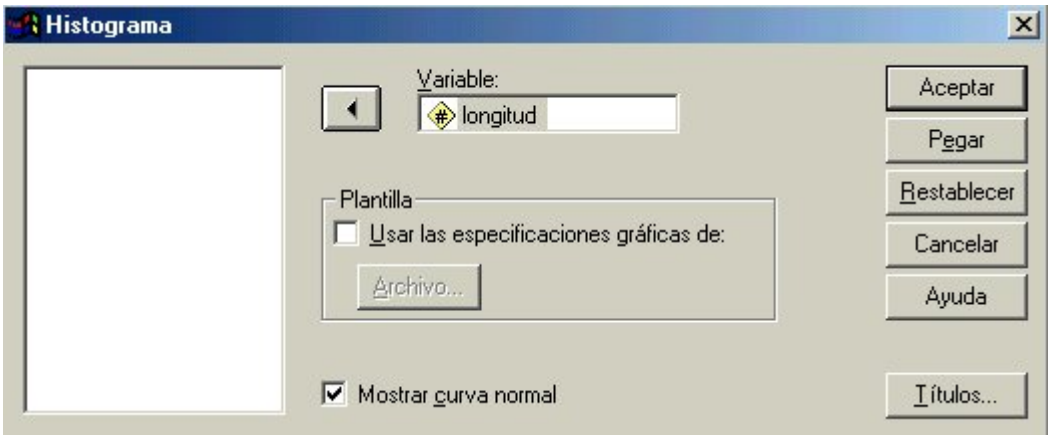
En este apartado dibujaremos el histograma de una muestra de una población con distribución desconocida. Para ilustrar el procedimiento que determina si esta muestra proviene de una distribución normal trabajaremos con el siguiente supuesto:

"Los valores sobre las longitudes en micras de 50 filamentos de la producción de una máquina son los siguientes:"

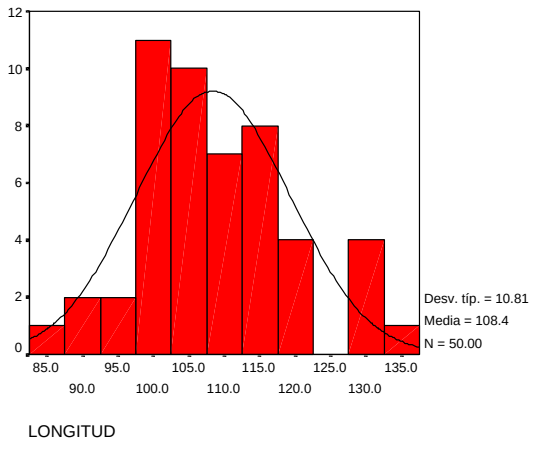
102	98	93	100	98	105	115	110	99	120
115	130	100	86	95	103	105	92	99	134
116	118	89	102	128	99	119	128	110	130
112	114	106	114	100	116	108	113	106	105
120	106	110	100	106	117	109	108	105	106

Determine si los datos de la muestra provienen de una distribución aproximadamente normal.

En primer lugar, necesitamos crear un nuevo banco de datos (`Archivo/Nuevo/Datos`), al que podemos llamar *Mediciones*. En el editor de datos introducimos los datos en una variable que llamaremos *longitud*. Para generar el histograma asociado a la variable longitud, seleccionamos `(Gráficos/Histograma)`, tomamos la variable *longitud* y seleccionamos `(Mostrar curva normal)`



El resultado que genera el SPSS es



Esta es la primera prueba que realizamos para comprobar si los datos proceden de una distribución normal. Si los datos son aproximadamente normales, la forma de la gráfica será similar a la de la curva normal superpuesta (esto es, con forma de joroba y simétrica alrededor de la media). En nuestro caso hay un cierto parecido, por lo que en principio, aunque no podemos afirmar con rotundidad que la muestra proviene de una población normal a la espera de otros resultados, podemos concluir que se asemeja a la curva normal teórica.

2. Cálculo IQR/S

El segundo paso es el de calcular el intervalo intercuartiles, IQR, la desviación estándar, s para la muestra y luego calcular el cociente IQR/S. Si los datos son aproximadamente normales, IQR/S $\approx$1.3. Puede verse que esta propiedad se cumple para las distribuciones normales si se observa que los valores z que corresponden a los percentiles 75o. y 25o. son 0.67 y -0.67, respectivamente. Puesto que σ = 1 para una distribución normal estándar (z),

$$IQR/\sigma = [.67- (-.67)]/1 = 1.34.$$

En esta segunda verificación debemos obtener el intervalo intercuartiles (es decir la diferencia entre los percentiles 75º. Y 25º.) y la desviación estándar del conjunto de datos. Con éstos, calcularemos el cociente anterior. Para determinar los cuartiles y la desviación típica vamos a `(Analizar / Estadísticos descriptivos/ Frecuencias / Estadísticos...)` allí seleccionamos los Cuartiles y la Desv típica. Los resultados para nuestro ejemplo son:

Estadísticos
LONGITUD

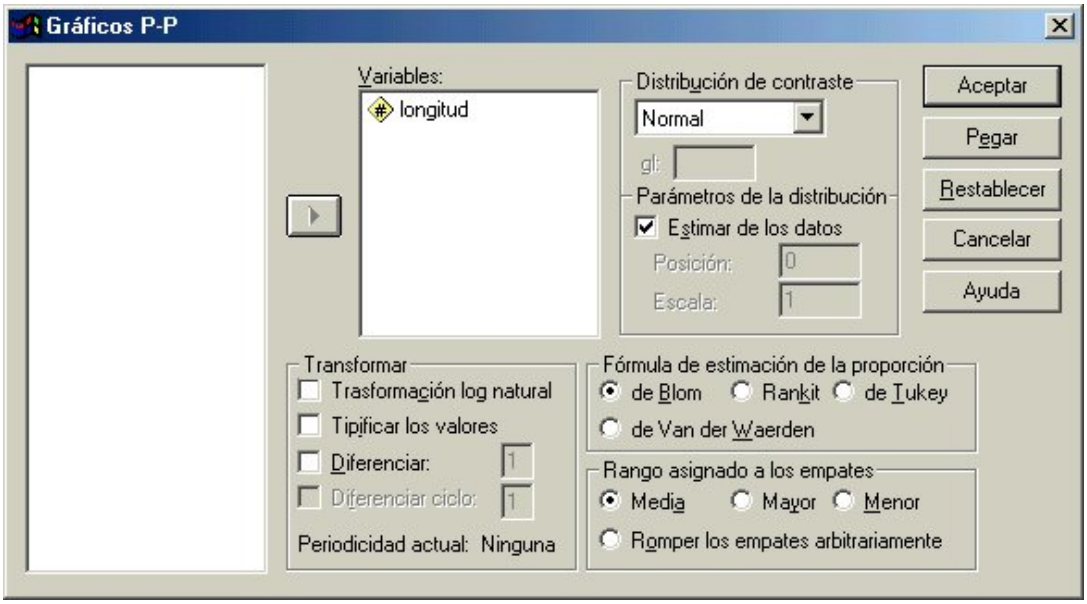
N	Válidos	50
	Perdidos	0
Desv. típ.		10.81
Percentiles	25	100.00
	50	106.00
	75	115.25

Haciendo el cálculo obtenemos que IQR/S = (115.25-100.00) / 10.81 = 1.41. Puesto que este valor es aproximadamente igual a 1.3, tenemos una confirmación adicional de que los datos son aproximadamente normales.

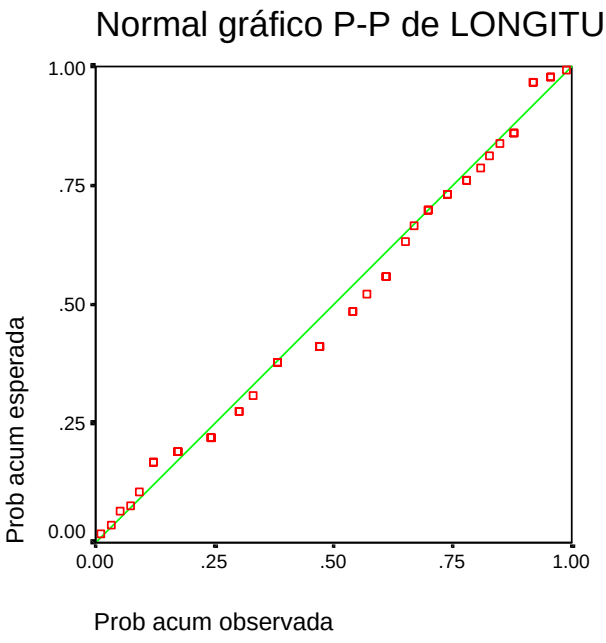
3. Gráfico de probabilidad normal

Una tercera técnica descriptiva para comprobar la normalidad es la gráfica de probabilidad normal. En una gráfica de probabilidad normal, las observaciones de un conjunto de datos se ordenan y luego se grafican contra los valores esperados estandarizados de las observaciones bajo el supuesto de que los datos están distribuidos normalmente. Si los datos en verdad tienen una distribución normal, una observación será aproximadamente igual a su valor esperado. Por tanto, una tendencia lineal (de línea recta) en la gráfica de probabilidad normal sugiere que los datos provienen de una distribución aproximadamente normal, en tanto que una tendencia no lineal indica que los datos no son normales.

En el SPSS podemos generar dos tipos de gráficos de este tipo, el Gráfico P-P y el Gráfico Q-Q. El P-P crea un gráfico de las proporciones acumuladas de una variable respecto a las de una distribución cualquiera de prueba (en nuestro caso de una normal). Si la variable seleccionada coincide con la distribución de prueba, los puntos se concentran en torno a una línea recta. Para generar este gráfico iremos a `(Gráficos/P-P)` y marcamos las casillas siguientes:



Damos a aceptar y la gráfica que obtenemos sería:



Observe que los puntos se ajustan relativamente bien a una línea recta. Por tanto, la verificación 3 también sugiere que los datos probablemente tienen una distribución normal.

El otro gráfico que genera el SPSS es el Gráfico Q-Q. La forma de obtener dicho gráfico es similar al anterior únicamente que debemos seleccionar (Gráficos/Q-Q). Este procedimiento crea un gráfico con los cuantiles de distribución de una variable respecto a los cuantiles de cualquiera de varias distribuciones de prueba (en nuestro caso de una normal). Si la variable

seleccionada coincide con la distribución de prueba, los puntos se concentran en torno a una línea recta.

Ejercicio:

- Generar el gráfico Q-Q para la variable longitud del problema inicial.

Estas verificaciones de normalidad dadas son técnicas fáciles de aplicar y de gran utilidad, pero sólo son descriptivas. Es posible (aunque poco probable) que los datos no sean normales a pesar de que las verificaciones se satisfacen razonablemente. Por tanto, debemos tener cuidado de no asegurar que las 50 longitudes de los filamentos están, de hecho, distribuidas normalmente. Sólo podemos decir que es razonable pensar que los datos provienen de una distribución normal. Pruebas que aportan una mayor confiablilidad a la inferencia son los test de hipótesis de Kolmogorv-Smirnov y el de Shapiro-Wilk en los cuales se contrasta la normalidad de la población de una muestra..

4. Ejercicios

1. Los silvicultores recorren periódicamente los bosques para determinar el tamaño (que por lo regular se mide como el diámetro a la altura del pecho) de una especie de árbol determinada. Aquí se presentan los diámetros ala altura del pecho (en metros) de una muestra de 28 álamos temblones del bosque boreal de la Columbia Británica.

12.4	17.3	27.3	19.1	16.9	16.2	20.0
16.6	16.3	16.3	21.4	25.7	15.0	19.3
12.9	18.6	12.4	15.9	18.8	14.9	12.8
24.8	26.9	13.5	17.9	13.2	23.2	12.7

   a. Determine si los datos de la muestra provienen de una distribución aproximadamente normal. En ese caso dar los parámetros aproximados de dicha distribución.
   b. ¿Qué porcentaje de álamos temblones podríamos encontrar con un diámetro entre 20 y 30 metros?
   c. ¿Qué probabilidad hay de encontrar álamos con diámetros mayores de 25 metros?

2. Investigadores del Massachusetts Institute of Technology (MIT) estudiaron las propiedades espectroscópicas de asteroides de la franja principal con un diámetro menor a los 10 kilómetros. Los asteroides se observaron con el telescopio Hiltner del observatorio del MIT; se registró el número N de exposiciones de imagen espectral independientes para cada observación. Aquí se presentan los datos de 40 observaciones de asteroides obtenidas de Science (9 de abril de 1993).

Número de observaciones de exposiciones de imagen espectral de 40 asteroides

3	4	3	3	1	4	1	3	2	3
1	1	4	2	3	3	2	6	1	1
3	3	2	2	2	2	1	3	2	1
6	1	3	2	2	1	2	2	4	2

a. Establece quién es la variable aleatoria correspondiente y determina si los datos son aproximadamente normales así como los parámetros para dicha v.a. normal.
b. Nos interesan aquellos asteroides con un número de observaciones de exposiciones de imagen espectral superior a 5. ¿Es común este tipo de asteroides?¿Por qué?
c. ¿Qué asteroides, en función del número de observaciones de exposiciones de imagen, aparecen en el 94 % de los casos?

3. La Harris Corporation y la University of Florida emprendieron un estudio para determinar si un proceso de fabricación efectuado en un lugar lejano se podría establecer localmente. Se instalaron dispositivos de prueba (pilotos) tanto en la ubicación antigua corno en la nueva y se tomaron lecturas de voltaje del proceso. Se considera que un proceso "bueno" produce lecturas de por lo menos 9.2 volts (y las lecturas mayores son mejores que las menores). La tabla contiene lecturas de voltaje para 30 series de producción en cada lugar.

Primer sitio			Segundo sitio		
9.98	10.12	9.84	9.19	10.01	8.82
10.26	10.05	10.15	9.63	8.82	8.65
10.05	9.80	10.02	10.10	9.43	8.51
10.29	10.15	9.80	9.70	10.03	9.14
10.03	10.00	9.73	10.09	9.85	9.75
8.05	9.87	10.01	9.60	9.27	8.78
10.55	9.55	9.98	10.05	8.83	9.35
10.26	9.95	8.72	10.12	9.39	9.54
9.97	9.70	8.80	9.49	9.48	9.36
9.87	8.72	9.84	9.37	9.64	8.68

a. Determine si las lecturas de voltaje en cada uno de esos sitios son aproximadamente normales. En caso afirmativo establezca las distribuciones correspondientes.
b. ¿En qué porcentaje las lecturas de voltaje en el primer sitio estaban por debajo de 9.2? ¿Y en el segundo? Según esto, ¿qué proceso de fabricación es mejor?
c. Nos interesa saber si es probable que las lecturas de voltaje estén entre 9.4 y 11 voltios en el primer sitio. ¿Qué conclusión puedes sacar?
d. Una nueva ley nos obliga a que los procesos de fabricación deberán de superar los 9.4 voltios en una serie de producción. ¿Podremos abrir nuestro proceso en el segundo sitio? ¿Podemos dejar abierto el primero?