SIMULACIÓN EN EL AULA: PRUEBA DE NORMALIDAD PARA MUESTRAS PEQUEÑAS USANDO TEST GRÁFICOS.

CALANDRA, MARÍA VALERIA [1,2]; VERICAT, FERNANDO [1,2]

[1] Departamento de Ciencias Básicas.
[2] Grupo de Aplicaciones Matemáticas y Estadísticas de la Facultad de Ingeniería (Gamefi).
Facultad de Ingeniería - Universidad Nacional de La Plata.
E-mail: mava@mate.unlp.edu.ar

RESUMEN

Los métodos gráficos son populares para chequear modelos, un gráfico cuantil- cuantil (qq-plot) permite observar cuan cerca está la distribución de un conjunto de datos a alguna distribución ideal o comparar la distribución de dos conjuntos de datos. La forma del gráfico debería ser idealmente una línea recta específica. Si interesa comparar con la distribución Gaussiana se llama gráfico de probabilidad Normal. El objetivo del presente trabajo es testear normalidad de una muestra en especial para el caso de tamaños muestrales pequeños para los cuales el comportamiento de estos gráficos suele ser errático y conducir a falsas interpretaciones, mostraremos además que no suele ser así con tamaños muestrales más grandes. Proponemos, también establecer una banda de probabilidad o banda envolvente basada en un método empírico, específicamente mediante el método de Monte Carlo, dicha banda nos establecerá un marco de referencia probabilístico para evitar falsas interpretaciones .Se presenta un código computacional, de fácil implementación, empleado por los alumnos para la aplicación de esta metodología de análisis de normalidad, el cual es utilizado para la enseñanza de la temática en cuestión.

Palabras clave: test gráficos, normalidad, simulación, monte carlo.

GRÁFICO DE PROBABILIDAD (QQ-PLOT)

El gráfico de probabilidad, qq-plot o gráfico cuantil-cuantil, constituye un método gráfico que nos permite comparar la distribución de un conjunto de datos con una distribución específica. Existen también test teóricos ¨test de bondad de ajuste¨ para chequear modelos pero estos en general requieren más formación por parte del alumno. Pearson (1900) fue uno de los primeros que introdujo uno de los más populares test de bondad de ajuste, más adelante Kolmogorov (1933).(Castro Kuriss, 2007). Para realizar un qq-plot se ordenan los datos de menor a mayor y se compara el i-ésimo dato con el correspondiente cuantil teórico ó poblacional. Si la distribución teórica propuesta constituye una buena aproximación a la distribución empírica cabría esperar que los cuantiles muestrales estén muy cerca a los de la distribución teórica propuesta y, por lo tanto, los puntos de la gráfica deberían estar muy próximos a la bisectriz del primer cuadrante. Sea $F(x)$ la función de distribución de una distribución específica. El gráfico de probabilidad se construye siguiendo los siguientes pasos (Castillo Gutierrez y Lozano Aguilera, 2007):

a) Se ordenan las observaciones de menor a mayor de la siguiente forma

$$x(1) \leq x(2) \leq \cdots \leq x(n)$$

b) Se determinan los valores

$$p_i = \frac{i - 0.5}{n} \quad i = 1, 2, \cdots, n.$$

Si por $q_x(p)$ notamos al cuantil de orden p ($0 < p < 1$) de las observaciones, tenemos que:

$$x(i) = q_x(p_i) \quad i = 1, 2, \cdots, n$$

c) Se determinan los cuantiles de orden p_i, $i = 1, 2, \ldots, n$ de la distribución teórica representada por la función de distribución F *(cuantiles teóricos)*, es decir:

$$q_t(p_i) = F^{-1}(p_i) \quad i = 1, 2, \cdots, n$$

d) Se representa el conjunto de puntos $(q_t(p_i), q_x(p_i))$, $i = 1, 2, \ldots, n$, o lo que es lo mismo, los puntos $(F^{-1}(p_i), x(i))$, $i = 1, 2, \ldots, n$.

En el caso en que F represente la función de distribución de una Normal, al gráfico de probabilidad resultante lo denominaremos gráfico probabilístico Normal o qq-plot Normal.

El gráfico cuantil-cuantil puede realizarse cualquiera sea la distribución hipotética. Mediante este gráfico es posible estudiar visualmente: asimetría hacia la derecha o izquierda, colas pesadas respecto de la distribución elegida, colas livianas respecto a la distribución elegida.

CONFUSIÓN A LA HORA DE ESTABLECER NORMALIDAD

La distribución Normal es una de las distribuciones más usadas e importantes, se ha convertido en una herramienta indispensable en cualquier rama de la ciencia, de la industria y del comercio. Muchos eventos naturales y reales tienen una distribución de frecuencias muy parecida a la Normal.

La Figura 1 muestra como una muestra aleatoria con distribución Normal Estándar de tamaño pequeño puede apartarse suficientemente de la bisectriz del primer cuadrante, sugiriendo falsamente falta de adecuación del modelo pese a que fue generada mediante la distribución Normal correspondiente. Y a su vez muestra como a medida que se consideran tamaños muetrales más grandes esa falsa falta de adecuación disminuye.

Para estudiar la situación de falsa interpretación gráfica a la hora de establecer normalidad en una muestra aleatoria Normal, se propuso analizar cuatro casos de distinto tamaño muestral. Dichas muestras fueron generadas mediante el lenguaje Matlab a partir de una distribución Normal Standard. Para cada caso se calcularon los cuantiles muestrales y los exactos propuestos por la fórmula de Hazen $q_i = \Phi^{-1}\left(\frac{i-0.5}{n}\right)$ (Hanzen, 1930), quedando plasmado en el qq-plot de la Figura 1.

Particularmente para poder contabilizar, de alguna manera, dicha discrepancia nos focalizamos en los cuantiles o percentiles p_i= 0,025, 0,05, 0,95 y 0,975, tanto para las muestras generadas como para la distribución exacta, que corresponden a las colas de la distribución y a los valores extremos de la muestra respectivamente. El objetivo de ello consiste en analizar, en función del tamaño de la muestra (n), que diferencia se presenta entre los cuantiles exactos y los estimados. En general se observa una mayor fluctuación en las colas de la distribución, es decir en los cuantiles inferiores y en los superiores, que en los cuantiles intermedios. Para llevar adelante este estudio se realizaron simulaciones para distintos tamaños de muestra (n=39, 99, 399 y 999). En la Figura1 se muestran los gráficos qq-plot correspondientes. En la situación ideal todos los puntos deberían de caer sobre la recta azul .En particular para los 4 cuantiles específicos en que nos focalizamos la intersección de las rectas horizontal y vertical del color correspondiente debería caer sobre la recta azul, esto nos indicaría que el valor del cuantil de la muestra coincide con el valor exacto del cuantil correspondiente. Cuanto más cerca, mejor resulta la aproximación. Repitiendo estas simulaciones, pudo observarse que para n=39 y n=99 los resultados no son adecuados, un valor muestral de n=399 no parece ser suficiente y por el contrario un valor n=999 resulta razonable.

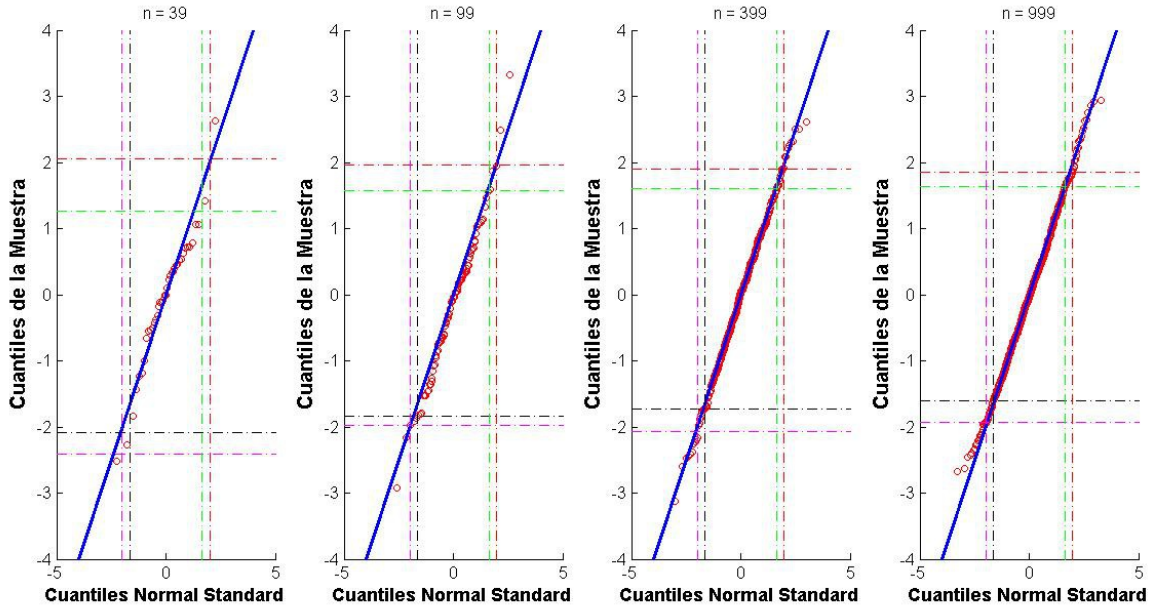

Figura 1. Cuantiles de las Muestras vs. Cuantiles de la Normal Standard en función del número de réplicas n.

En la Tabla 1 se muestran los porcentajes basados en 5.000 réplicas de la distribución Normal Estándar, para cada tamaño muestral, donde el cuantil estimado cae en un rango de error del 5% respecto al valor exacto. En la misma se puede observar como a medida que aumenta el número de muestras utilizadas los valores de los cuantiles se aproximan mejor al valor esperado.

	Cuantiles			
Muestras	**0,025**	**0,05**	**0,95**	**0,975**
39	17,32	19,52	19,26	17,52
99	28,54	29,06	29,98	28,38
399	52,80	55,64	57,86	53,08
999	74,50	76,72	78,14	75,68

Tabla 1. Porcentajes de valores, según la muestra, que se acercan al valor exacto con un 5% de error respecto al mismo.

Figura 2. Histogramas de los cuantiles simulados para cada tamaño de muestra.

En la Figura 2 se muestran los Histogramas correspondientes a las 5.000 réplicas de muestra Normales Estándar correspondientes a los distintos tamaños de muestras y para el cuantil o percentil p_i=0,025. Las franjas verticales corresponden al:

Cuantil Exacto ± 0,05 x Cuantil Exacto

Para *n*=39 y *n*=99 el sesgo es evidente y la variabilidad es grande. Ésta disminuye al aumentar n.

DESARROLLO DEL EJEMPLO Y BANDA DE PROBABILIDAD

En la Tabla 2 se muestra una serie de datos, que se corresponden a la medición de la desviación de la aceleración de la gravedad (g) respecto a un valor de $980.000*10^{-3}$ cm/seg^2, en unidades de cm/seg^2 $*10^{-3}$.

z1	z2	z3	z4	z5	z6	z7	z8	z9	z10	z11	z12	z13
84	86	85	82	77	76	77	80	83	81	78	78	78

Tabla 2. Muestra de mediciones de la desviación de la aceleración de la gravedad respecto del valor $g=980.000*10^{-3}$ [cm/seg^2].

En particular nosotros queremos averiguar si esta muestra de n = 13 observaciones pueden ser o no interpretadas correctamente con distribución Normal Estándar, se eligió un tamaño de

muestra bastante chico. En la Figura 3 se observa un gráfico de normalidad para estos datos, en la misma encontramos los cuantiles Normales o cuantiles teóricos $\Phi^{-1}\left(\frac{i-0.5}{13}\right)$ *i=1, 2, …, 13* contra los valores observados ordenados *z(i) i=1, 2, … , 13* que corresponden a lo que se denomina cuantiles muestrales . La línea a trazos azul es el patrón esperado, y la cuestión es identificar dónde ó no los puntos se apartan suficientemente del mismo para sugerir que la muestra no se distribuye normalmente.

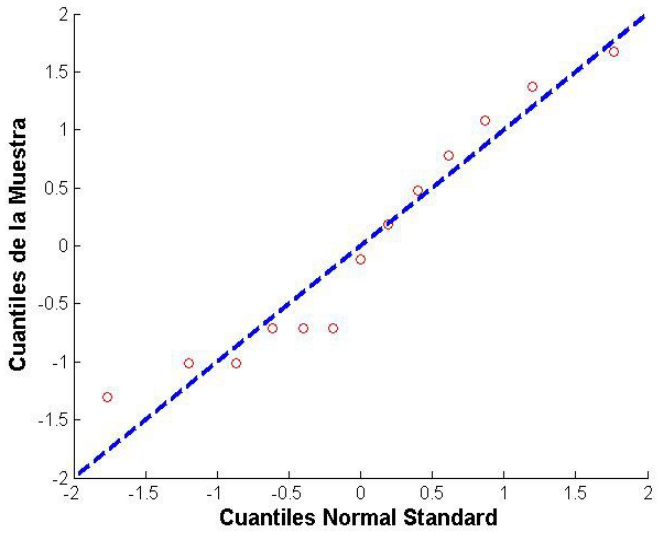

Figura 3. Cuantiles de la Normal Standard vs. Cuantiles de la Muestra.

Evidentemente, se observa una cierta cantidad de puntos desajustados, pero como nosotros sabemos que en realidad dicha muestra tiene distribución Normal Estándar, se nos hace necesario tener algún otro criterio que nos ayude a ver si dichos puntos están suficientemente alejados del modelo propuesto como para sugerir la adecuación o no del mismo. La idea es encontrar una banda de probabilidad, empírica, simulada por el método de Monte Carlo (también denominado método paramétrico de remuestreo)(Davidson y Hinkley, 1998) donde debería encuadrarse nuestra muestra original para ser considerada con distribución Normal Estándar.

Dada la muestra original $x_{(1)}, x_{(2)}, \ldots, x_{(13)}$ ya ordenada de menor a mayor, se toman *R* réplicas (R es el número de muestras Normales simuladas mediante un paquete estadístico, en general tomaremos R=999) de tamaño muestral *n*=13 de distribución Normal Estándar, es decir coincidente con el tamaño de la muestra original, también ordenadas de menor a mayor:

$$\begin{matrix} z(1)_1^*, & z(2)_1^*, & \cdots & z(13)_1^* \\ z(1)_2^*, & z(2)_2^*, & \cdots & z(13)_2^* \\ \vdots & \vdots & \vdots & \vdots \\ z(1)_R^*, & z(2)_R^*, & \cdots & z(13)_R^* \end{matrix}$$

La primera fila de este arreglo corresponde a la primera réplica, la segunda fila a la segunda réplica, y así sucesivamente hasta la R-ésima réplica, todas ordenadas de menor a mayor.

La idea es tomar el primer dato de cada una de las réplicas, esto es [illegible], (primera columna del arreglo) y ordenarlos de menor a mayor: [illegible], .Con esta muestra formaremos un primer intervalo empírico que contendrá un 90% de estos datos.

Para ello debemos elegir un valor *k* entero, entre 1 y *R,* tal que *k/(R+1)=p*, siendo *p* un número igual a la mitad de la proporción de datos que no contendrá el intervalo, por lo tanto en nuestro caso tenemos *p = 0,05* y como fue sugerido tomaremos *R = 999*, lo que nos dará *k=50* y una banda de probabilidad puntual correspondiente al *(1-2p)*100%*, es decir *90%*.

Luego formamos el intervalo de probabilidad con *90%* de confianza correspondiente al primer valor muestral [illegible], llamado también intervalo puntual empírico estimado por Monte Carlo para el primer dato muestral, de esta manera nos aseguramos que fuera de dicho intervalo caen sólo un 10% de valores. Procedemos de igual forma con el segundo valor muestral y con sus respectivas réplicas, quedándonos formado el intervalo [illegible], y así sucesivamente nos quedan formados los intervalos de referencia en donde deberían de caer cada uno de los datos de la muestra original si tuvieran distribución Normal al igual que las muestras.

En la Figura 4 se muestran, tanto las réplicas de Monte Carlo como así también las bandas de confianza marcadas en trazo grueso que corresponden a las bandas que unen los extremos inferiores y superiores de los intervalos puntuales.

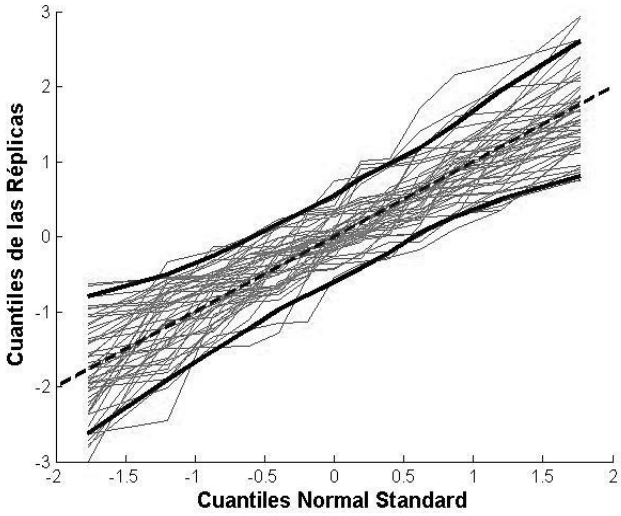

Figura 4. Líneas correspondientes a las réplicas y las bandas de confianza.

Como se ve en la Figura 4 la mayoría de las muestras simuladas caen dentro de las bandas envolventes.

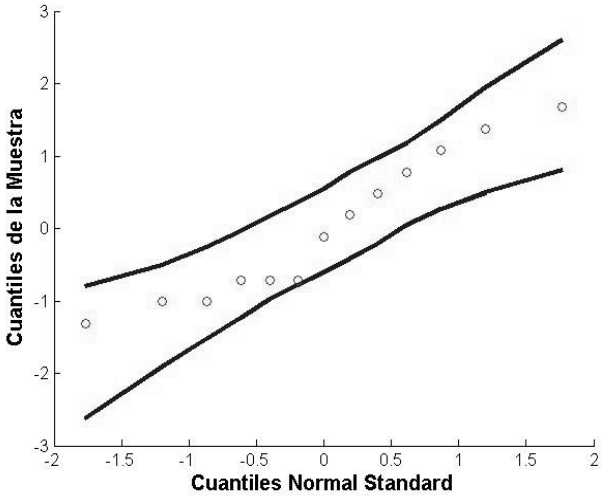

Figura 5. La muestra original con las bandas de confianza.

Como se observa, claramente, en la Figura 5 todos nuestros datos originales caen dentro de la banda propuesta, lo que implicaría que el modelo Normal Estándar resulta adecuado.

Si quisiéramos una banda con un nivel de confianza superior al *90%* deberíamos tomar un valor de *p* más pequeño, aunque la cantidad de réplicas *R* no deberían ser modificadas debido a que las mismas son suficientes para estimar el comportamiento de la muestra.
Presentamos el script realizado en Matlab correspondiente a los cálculos definidos para la determinación de las bandas de confianza, el gráfico correspondiente de las mismas y los datos de la muestra original.

Script en Matlab

==========

```
%Lectura de datos y cálculos iniciales
Datos = [84 86 85 82 77 76 77 80 83 81 78 78 78];
d2 = (Datos - mean(Datos))/std(Datos);
d2s=sort(d2,2);
p=0.05;
R=999;
k=p*(R+1);
[M,N]=size(Datos);
%Generación de las Réplicas
 for i=1:R;
    num(i,1)=i;
 end;
for j=1:R;
    g(j,:)=normrnd(0,1,1,N);
end;
B1=sort(g,2);
Rep1=cat(2,num,B1);
Margeninf=k;
Margensup=R+1-k;
%Determinación de las bandas inferior y superior
for f=2:N+1;
    [B,index]=sortrows(Rep1,f);
    inferior(f-1)=B(Margeninf,f);
    superior(f-1)=B(Margensup,f);
end
for t=1:N;
   yi(t)=(t-0.5)/N;
end;
for d=1:N;
    ys(d)=norminv(yi(d),0,1);
```

```
end;
%Gráfico de los datos y las bandas inferior y superior
figure(1)
hold on
plot(ys,inferior,'LineWidth',3,'MarkerEdgeColor','r','MarkerFaceColor','g','MarkerSize',1)
plot(ys,superior,'LineWidth',3,'MarkerEdgeColor','r','MarkerFaceColor','g','MarkerSize',1)
plot(ys,d2s,'or','MarkerSize',6),xlabel('Cuantiles Normal Standard', 'fontsize', 12, 'fontweight','b'),
ylabel ('Cuantiles de la Muestra','fontsize',12,'fontweight','b');
hold off
```

==========

CONCLUSIONES

- En la práctica se pueden graficar las bandas envolventes para predecir tendencias, manifestadas por secuencias de puntos que se salen fuera de las bandas. O para observar el comportamiento de las colas de la distribución.

- Estos gráficos son usados para observar posibles desviaciones de un modelo hipotético.

- Son muy útiles a la hora de chequear normalidad de residuos en regresión.

- Nosotros hemos tratado muestras de tamaño chico, dado su comportamiento errático lo cual puede llevar a falsas interpretaciones con métodos gráficos convencionales.

- Los test gráficos pueden ser abordados por alumnos sin mucha formación estadística, en contraste con los test teóricos.

- Estas bandas envolventes también se pueden hacer para chequear otros modelos distribucionales, no necesariamente Normal.

- Los métodos de simulación tanto paramétricos como no paramétricos van a la vanguardia a la hora de testear modelos teóricos sin necesidad de una teoría engorrosa y con un tiempo computacional bajo.

- Dicho método fue implementado en las carreras de grado de la Facultad de Ingeniería en la materia Estadística y facilitó enormemente la interpretación de los gráficos cuantil-cuantil.

- El nivel de simulación utilizado en este trabajo, estimula a los alumnos de grado a trabajar en proyectos de programación de herramientas estadísticas para su aplicación a la solución de problemas específicos de su carrera.

- Además esta aplicación nos brinda la posibilidad de incorporar el uso de simulaciones como herramienta de enseñanza y aprendizaje.

REFERENCIAS BIBLIOGRÁFICAS

Castillo Gutiérrez, S y Lozano Aguilera, E.D. (2007). Q-Q Plot Normal. Los puntos de posición gráfica. *Revista electrónica Iniciación a la Investigación, Universidad de Jaen, España*, Nº2, ISSN: 1988-415X.

Castro Kuriss, C.A. (2007). Tests de bondad de ajuste basados en la distribución empírica para datos con y sin censura. *Tesis de la Maestría en Estadística Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires*.

Davidson, A.C. y Hinkley, D.V. (1998). *Bootstrap methods and their application*. Cambridge University Press, Reino Unido.

Hanzen, A. (1930). *Flood Flows: A Study of Frecuencies and Magnitudes.* Ed. John Wiley and Sons, New York.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell´ Istituto Italiano degli Attuari*, 4: 83-91.

Pearson K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, 50 (5): 157-175.