INICIACIÓN A LA INVESTIGACIÓN
REVISTA ELECTRÓNICA
UNIVERSIDAD DE JAÉN

Ini Inv, 2:a9 (2007)

Q-Q Plot Normal. Los puntos de posición gráfica.

Sonia Castillo Gutiérrez* y Emilio Damián Lozano Aguilera**

Departamento de Estadística e Investigación Operativa.
Universidad de Jaén. Campus Las Lagunillas s/n, CP: 23071, Jaén.
Correo electrónico: socasti@ujaen.es

Resumen

En la construcción de los gráficos de probabilidad normal o Q-Q Plot Normal, intervienen como elementos fundamentales los denominados 'plotting positions' o puntos de posición gráfica. En este trabajo, describimos la construcción de un gráfico de probabilidad (y en particular el Q-Q Plot Normal) y analizamos las definiciones de los puntos de posición gráfica más relevantes que han sido introducidas a lo largo de la historia. Comprobamos que la elección de esos puntos influye en el gráfico, produciendo resultados sensiblemente distintos.

1. Introducción

El gráfico probabilístico normal nos permite comparar la distribución empírica de un conjunto de datos con la distribución Normal. Por tanto, dicho gráfico se puede considerar como una técnica gráfica para la prueba de normalidad de un conjunto de datos.

La construcción del gráfico de probabilidad normal se realizará a través de los cuantiles de la normal estándar, de forma que aceptaremos la hipótesis de normalidad de los datos, siempre que los puntos en el gráfico tengan un comportamiento "suficientemente rectilíneo".

En el gráfico de probabilidad son elementos fundamentales los conocidos como "puntos de posición gráfica"(plotting positions). Dedicamos parte del trabajo al análisis de dichos elementos, poniendo de manifiesto la diferencia existente entre las distintas propuestas de puntos de posición gráfica que han sido introducidas a lo largo de la historia.

El objetivo de este trabajo consiste en comprobar la diferencia que se produce al generar los gráficos de probabilidad normal, en función de la elección que se haga de los puntos de posición

*Doctoranda

**Director del Trabajo de Investigación "Consideraciones sobre los Puntos de Posición Gráfica en los Gráficos de Probabilidad" del Programa de Doctorado "Aproximación y Técnicas Estadísticas"

gráfica, quedando claro que dichos puntos son un elemento clave para los gráficos probabilísticos normales.

2. Gráfico de probabilidad (Q-Q Plot)

El gráfico de probabilidad se constituye en un método gráfico que nos permite comparar la distribución de un conjunto de datos con una distribución especificada.

Supongamos que disponemos de un conjunto de observaciones x_i, $(i = 1, 2, ..., n)$.
Sea $F(x)$ la función de distribución de una distribución especificada. El gráfico de probabilidad se construye siguiendo los siguientes pasos:

1) Ordenar las observaciones de menor a mayor en la forma

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

2) Determinar los valores

$$p_i = \frac{i - 0{,}5}{n} \quad i = 1, 2, ..., n.$$

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

$$x_{(i)} = Q_x(p_i) \quad i = 1, 2, ..., n.$$

3) Determinar los cuantiles de orden p_i, $i = 1, 2, ..., n$ de la distribución teórica representada por la función de distribución F, es decir:

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

4) Representar el conjunto de puntos $(Q_t(p_i), Q_x(p_i))$, $i = 1, 2, ..., n$, o lo que es lo mismo, los puntos $(F^{-1}(p_i), x_{(i)})$, $i = 1, 2, ..., n$.

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 Q-Q Plot Normal.

Si la distribución teórica constituye una buena aproximación de la distribución empírica, cabría esperar que los cuantiles de los datos estén muy próximos a los de la distribución teórica y, por tanto, los puntos del gráfico se dispondrían muy próximos a la bisectriz del primer cuadrante. En otros casos, puede ocurrir que los datos no estén próximos a la recta $y = x$, sino que simplemente los puntos se posicionen de forma rectilínea. En este caso, podemos encontrar las correspondientes constantes que produzcan un cambio de origen y/o escala en los datos, de forma que los valores resultantes produzcan una disposición de los puntos en el gráfico probabilístico, suficientemente próxima a la recta y=x. Suponiendo que una vez realizadas las transformaciones, obtenemos que los datos aparecen muy próximos a la recta y=x, concluiremos que la distribución empírica es compatible con la distribución teórica, salvo en los parámetros de localización y escala.

3. Puntos de posición gráfica

Los puntos de posición gráfica son elementos claves en la obtención de los gráficos de probabilidad, ya que permiten identificar las observaciones como los cuantiles de la distribución empírica de los datos, $x_{(i)} = Q_x(p_i)$, y por otra parte, constituyen los elementos fundamentales para determinar los cuantiles de la distribución teórica, representada por la función de distribución $F(x)$, de forma que $Q_t(p_i) = F^{-1}(p_i)$.

En esta sección, vamos a presentar distintas propuestas de definición de los puntos de posición gráfica encontrados a lo largo de la historia. Sin ánimo de ser exhaustivos, se han seleccionado aquellas definiciones que, desde el punto de vista histórico, o por su uso, se han considerado como más importantes. Proseguiremos en la siguiente sección con una comparación de las definiciones recogidas en ésta.

La primera elección de los p_i fue introducida y usada en el campo de la hidrología. Se conoce con el nombre de "Método California". Dicho método, consiste en definir los p_i como:

$$p_i = \frac{i}{n} \quad i = 1, 2, ..., n \tag{1}$$

Esta opción, que en un principio parece la más natural, fue pronto descartada debido a la imposibilidad de dibujar la observación n-ésima, que en muchas ocasiones resultaba de gran interés.

En 1930, Hazen (Hazen A., Flood Flows: A Study of Frequencies and Magnitudes, 1930, John Wiley & Sons, New York) propuso la siguiente definición:

$$p_i = \frac{i - 0{,}5}{n} \quad i = 1, 2, ..., n \tag{2}$$

donde el término 0.5 del numerador, aparece como un corrector de continuidad. De esta forma, se subsana el inconveniente comentado anteriormente, pudiendo dibujar todas las observaciones en el gráfico (esta definición viene recogida en el paquete estadístico SPSS como "Rankit").

Weibull (Weibull W., Ingeniors Ventenskaps Akademien Handlingar, 1939, 153, 17) propuso definir los p_i mediante la siguiente expresión:

$$p_i = \frac{i}{n+1} \quad i = 1, 2, ..., n \tag{3}$$

(Esta definición aparece en SPSS como propuesta por Van der Waerden).

Las dos propuestas anteriores de los p_i están recogidas en la expresión general:

$$p_i = \frac{i - c}{n - 2c + 1} \quad para \quad 0 \leq c \leq 1 \tag{4}$$

donde observamos que si c=0.5, aparece la propuesta de Hazen, y si c=0, obtenemos la propuesta de Weibull. Puede verse una extensa discusión sobre la elección óptima de c en (Barnett V.,

Applied Statistics, 1975, 24, 1:95-108).

En la literatura al uso, aparecen otras muchas definiciones distintas de los p_i. Una idea bastante aceptada y difundida es que los p_i deben ser determinados a partir de medidas de localización de los estadísticos de orden. Así, si consideramos los estadísticos de orden de la variable transformada en la forma $(x-\mu)/\sigma$, (donde μ y σ no son necesariamente la media y la desviación típica, sino parámetros de localización y escala respectivamente), y notamos por:

$$W_{(i)} = \frac{X_{(i)} - \mu}{\sigma} \quad i = 1, 2, ..., n$$

se propone tomar:

$$p_i = F\left[loc(W_{(i)})\right] \quad i = 1, 2, ..., n$$

donde loc($W_{(i)}$), simboliza a una medida de localización de $W_{(i)}$. En este sentido, varios autores (Shapiro et al., Biometrika, 1965, 52, 3-4:591-611), (Shapiro et al., Journal of the American Statistical Association, 1972, 67, 337:215-216) y (Breque J.L., Technometrics, 1977,19, 3:293-306) propusieron, para los gráficos de probabilidad, y por consiguiente, para los estadísticos que definen sus respectivos contrastes de normalidad, el uso de:

$$p_i = F\left[E(W_{(i)})\right] \quad i = 1, 2, ..., n$$

donde por E($W_{(i)}$) notamos al valor esperado del i-ésimo estadístico de orden transformado.

Por su parte, Benard y Bos-Levenbach (Benard et al., Statistica, 1953, 7:163-173) y Filliben (Filliben J., Technometrics, 1975, 17:111-117) definieron los p_i en los términos

$$p_i = F\left[Med(W_{(i)})\right] \quad i = 1, 2, ..., n$$

donde por Med($W_{(i)}$) se simboliza a la mediana del i-ésimo estadístico de orden transformado.

Vista la dificultad que en algunos casos aparece en la determinación de los p_i usando las definiciones anteriores (medidas de localización de los estadísticos de orden), algunos autores han propuesto otras definiciones basadas en meras aproximaciones de éstas.

Benard y Bos-Levenbach (Benard et al., Statistica, 1953, 7:163-173), que consideraron la elección de la mediana como medida de localización más apropiada, demostraron que para el gráfico probabilístico normal, una buena aproximación de los p_i, para tamaños muestrales intermedios, podía conseguirse con la expresión:

$$p_i = \frac{i - 0{,}3}{n + 0{,}4} \quad i = 1, 2, ..., n \tag{5}$$

Cabe observar que esta fórmula aparece al sustituir c=0.3 en la expresión (4).

Otros autores como Kimbal (Kimbal B., Journal of the American Statistical Association, 1960, 55:546-560) y Cunnane (Cunnane C., Journal of Hidrology, 1978, 37:205-222) recomendaron el uso de la esperanza como medida de localización, cuya aproximación ya fue propuesta

por Blom (Blom G., Statistical Estimates and Transformed Beta-Variables, 1958, Wiley, New York) y es usada particularmente cuando la distribución teórica es la distribución normal. La propuesta consiste en:

$$p_i = \frac{i - 3/8}{n + 1/4} \quad i = 1, 2, ..., n \tag{6}$$

Esta expresión se obtiene al sustituir c en la expresión (4) por 3/8. Esta elección de los p_i ha incrementado su aceptación entre muchos usuarios, y es la que aparece por defecto en el paquete estadístico SPSS.

La siguiente propuesta puede verse en (Tukey J. W., Annals of Mathematical Statistics, 1962, 33, 1:1-67), realizada al considerar, de nuevo en la expresión (4), c=1/3, obteniendo:

$$p_i = \frac{i - 1/3}{n + 1/3} \quad i = 1, 2, ..., n \tag{7}$$

Filliben (1975), como aproximación a la expresión

$$p_i = F\left[Med(W_{(i)})\right] \quad i = 1, 2, ..., n$$

sugirió el uso de:

$$p_i = \begin{cases} 1 - p_n & i = 1 \\ \dfrac{i - 0{,}3175}{n + 0{,}365} & i = 2, 3, ..., n - 1 \\ (0{,}5)^{(1/n)} & i = n \end{cases} \tag{8}$$

Si observamos, la aproximación dada por Filliben puede obtenerse de la expresión general (4), sin más que elegir c=0.3175 (para i=2,3,...,n-1).

En 1995, Lozano (Lozano E.D., Tesis Doctoral, 1995, Aportaciones a las técnicas gráficas para el estudio de normalidad y las causas de su pérdida. Universidad de Granada) obtuvo unos p_i usando como medida de localización la mediana de los estadísticos de orden (supuesta normalidad), tal y como propusieron Benard y Bos-Levenbach (1953) y Filliben (1975). La diferencia entre las tres propuestas es que estos dos últimos, abandonaron su cálculo directo en beneficio de las aproximaciones recogidas anteriormente; sin embargo, Lozano propone un procedimiento casi exacto, alcanzando, sin un coste elevado de tiempo una precisión tan grande como se quiera. El procedimiento consiste en elegir los puntos de posición gráfica del gráfico de probabilidad normal, como la mediana del i-ésimo estadístico de orden de una distribución uniforme estándar, sin necesidad de acudir a aproximaciones ($p_i = m_i$).

4. Comparación de los distintos gráficos de probabilidad según el punto de posición gráfica seleccionado.

En esta sección vamos a realizar una serie de gráficos, que nos servirán para poder apreciar la diferencia provocada por la elección de un p_i u otro.

La notación que utilizaremos para identificar las distintas definiciones de los puntos de posición gráfica es la siguiente (en todos los casos n es el número de observaciones e i=1,...n):

$$p_i^1 = \frac{i}{n} \qquad p_i^5 = \frac{i-0{,}375}{n+0{,}25}$$

$$p_i^2 = \frac{i-0{,}5}{n} \qquad p_i^6 = \frac{i-1/3}{n+1/3}$$

$$p_i^3 = \frac{i}{n+1} \qquad p_i^7 = \begin{cases} 1-p_n & i=1 \\ \dfrac{i-0{,}3175}{n+0{,}365} & i=2,3,...,n-1 \\ (0{,}5)^{(1/n)} & i=n \end{cases}$$

$$p_i^4 = \frac{i-0{,}3}{n+0{,}4} \qquad p_i^8 = m_i$$

Con estas ocho definiciones de los puntos de posición gráfica, obtenemos los correspondientes conjuntos de cuantiles de una N(0,1), que notaremos, respectivamente, como $q_i^1, q_i^2, ..., q_i^8$ y que serán los que aparecerán en los gráficos.

Se ha optado por fijar una de las definiciones y comparar todas las demás con ésta. En este sentido, hemos considerado que el punto de posición gráfica que tomaremos como referencia para realizar la comparación con los restantes, será p_i^5, es decir, la propuesta de Blom, debido a que es la que utiliza por defecto el paquete estadístico SPSS.

Para realizar la comparación de las distintas definiciones de puntos de posición gráfica, vamos a construir cada uno de los gráficos, con respecto al cuantil correspondiente al punto de posición gráfica p_i^5, es decir, q_i^5. En cada gráfico representamos los dos conjuntos de puntos con sus respectivas rectas de regresión de mínimos cuadrados ordinarios. En algunos de ellos, se apreciará una notable diferencia entre esas rectas. Para medir esa diferencia, que a simple vista observaremos, calculamos la mayor diferencia que se produce entre las rectas dentro del gráfico (en el recorrido de los puntos representados). Dicha diferencia máxima siempre se produce en los puntos de menor o mayor abscisa. Tomando ambos valores de las abscisas, evaluaremos las dos rectas en estos dos puntos, y haciendo la diferencia de las ordenadas correspondientes (menor y mayor) obtendremos la diferencia vertical de las dos rectas. De estas dos medidas, tomaremos la

mayor, y esta será la diferencia máxima entre las dos rectas. De esta forma, podremos comparar unos puntos de posición gráfica con otros.

Además de poder comparar unos puntos de posición gráfica con otros, vamos a poder compararlos según la distribución de la que provengan los datos, para intentar encontrar si existen más diferencias entre los puntos de posición gráfica, dependiendo de la distribución de los datos, o si la distribución de éstos no influye en las diferencias entre los distintos p_i.

De esta forma, consideramos tres grupos de gráficos. El primero para datos procedentes de una distribución normal, el segundo para datos procedentes de una distribución con asímetría y el tercero para datos que proceden de una distribución con una curtosis distinta a la de la normal.

El primer gráfico de cada grupo debemos hacerlo independientemente. Esto es debido a la propia definición del punto de posición gráfica p_i^1=i/n, porque cuando i=n, $p_i^1 = 1$ lo que imposibilita el cálculo del cuantil correspondiente, por lo que es necesario eliminar este último dato de p_i^1 y también de las observaciones para poder hacer el gráfico comparativo.

Los distintos gráficos se han realizado para un tamaño muestral 50 y los resultados obtenidos son los siguientes:

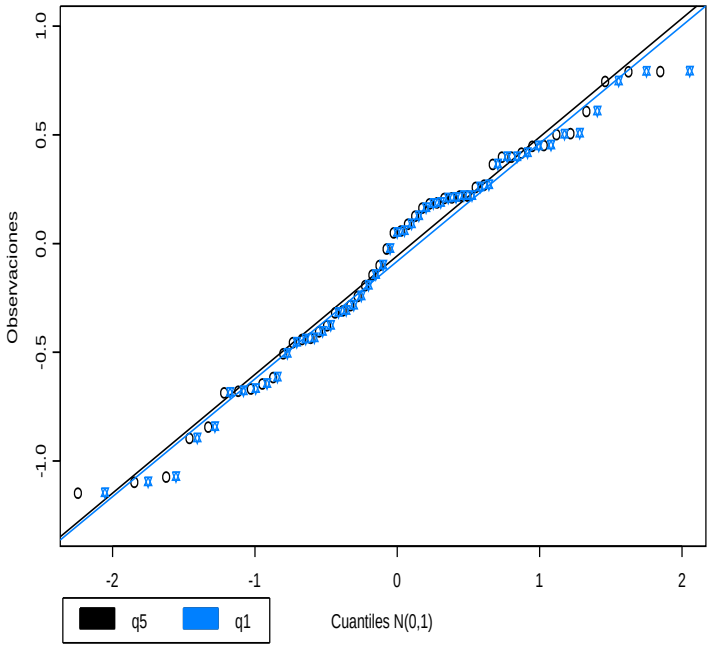

Figura 1: Distribución normal considerando p_i^1 y p_i^5

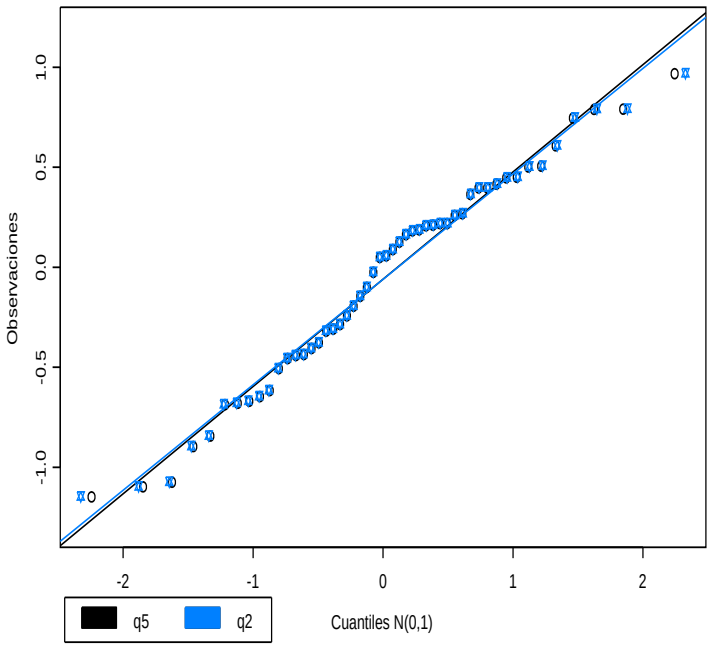

Figura 2: Distribución normal considerando p_i^2 y p_i^5

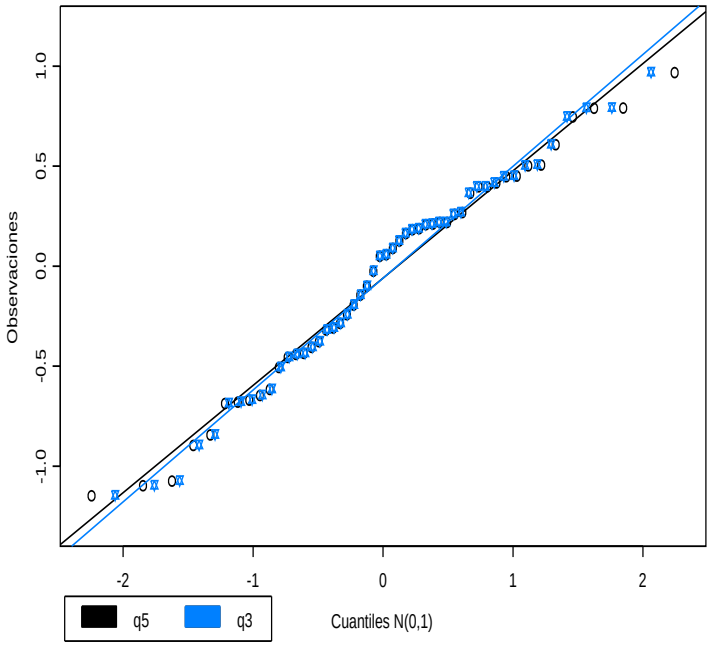

Figura 3: Distribución normal considerando p_i^3 y p_i^5

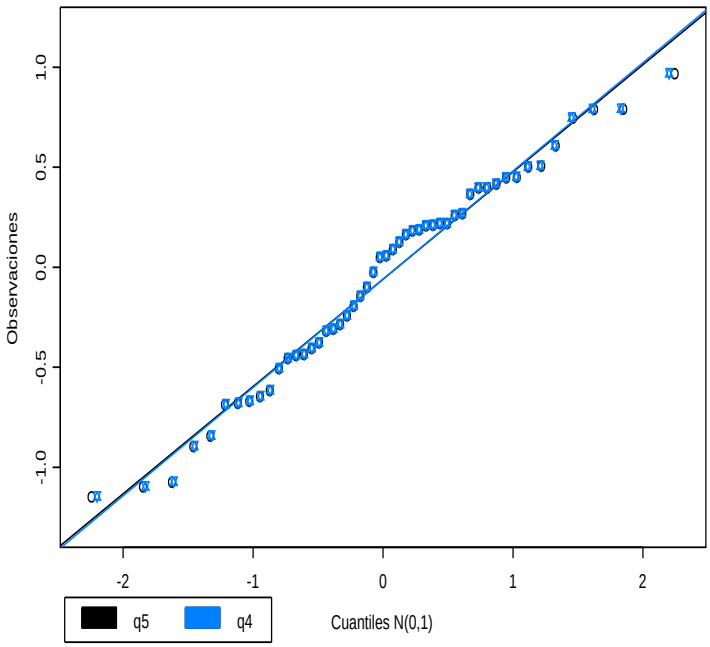

Figura 4: Distribución normal considerando p_i^4 y p_i^5

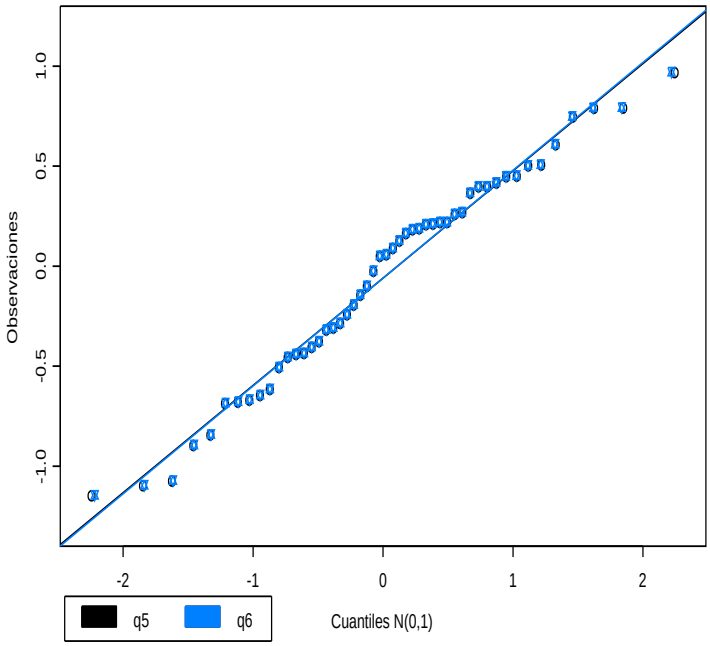

Figura 5: Distribución normal considerando p_i^6 y p_i^5

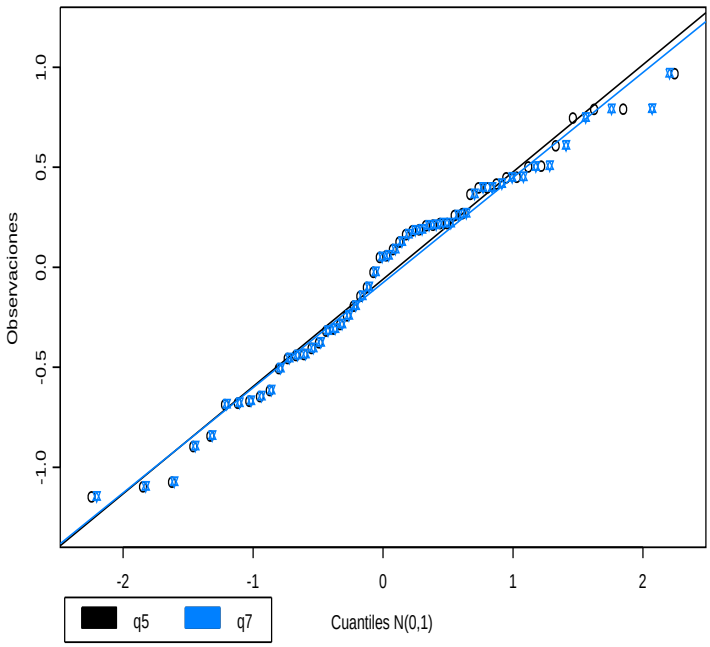

Figura 6: Distribución normal considerando p_i^7 y p_i^5

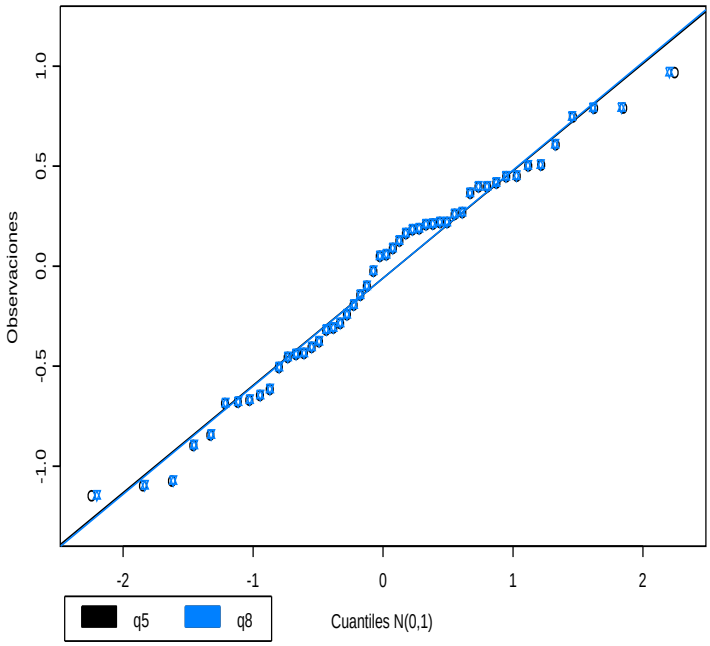

Figura 7: Distribución normal considerando p_i^8 y p_i^5

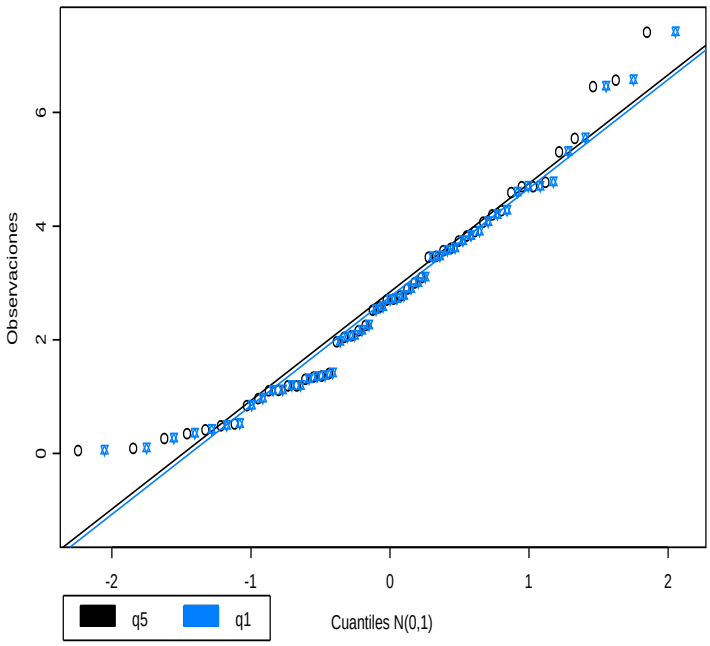

Figura 8: Distribución asimétrica a la derecha considerando p_i^1 y p_i^5

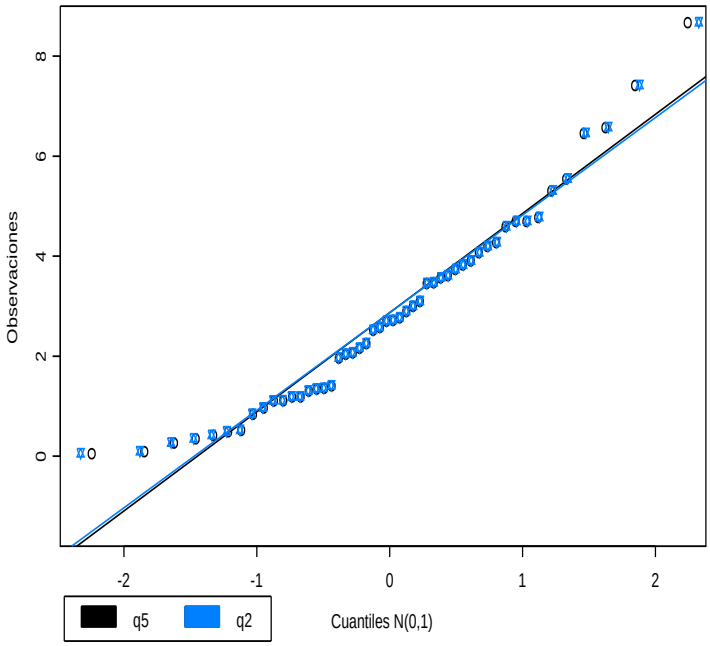

Figura 9: Distribución asimétrica a la derecha considerando p_i^2 y p_i^5

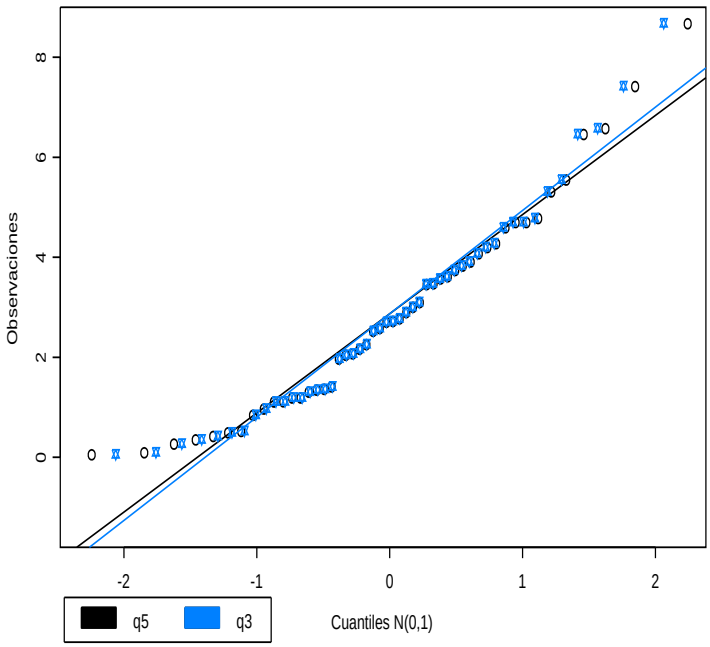

Figura 10: Distribución asimétrica a la derecha considerando p_i^3 y p_i^5

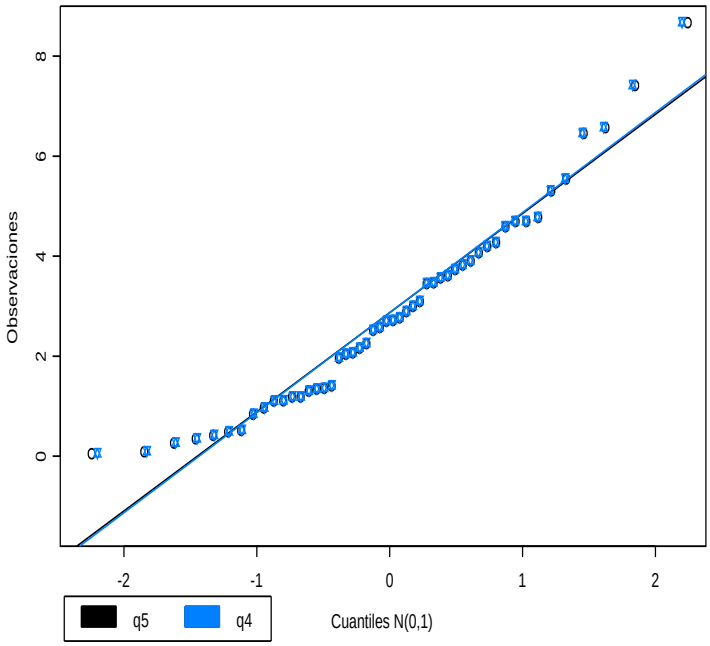

Figura 11: Distribución asimétrica a la derecha considerando p_i^4 y p_i^5

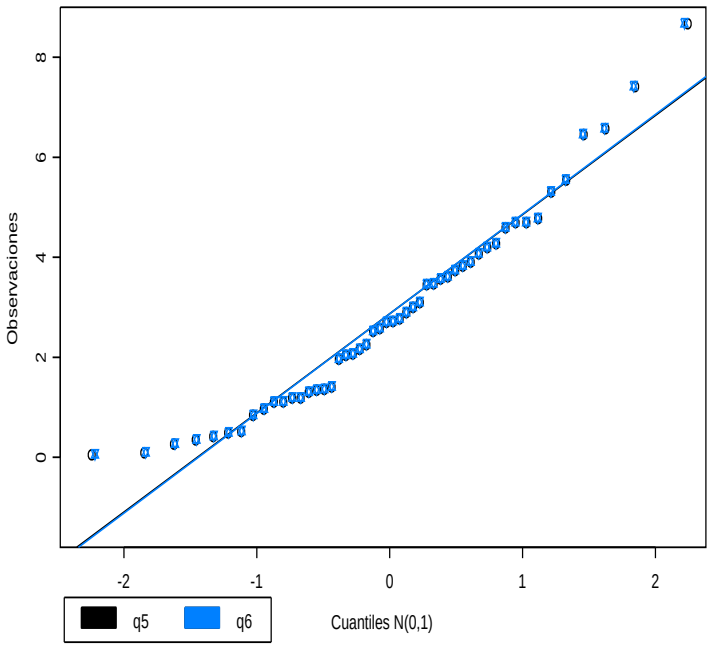

Figura 12: Distribución asimétrica a la derecha considerando p_i^6 y p_i^5

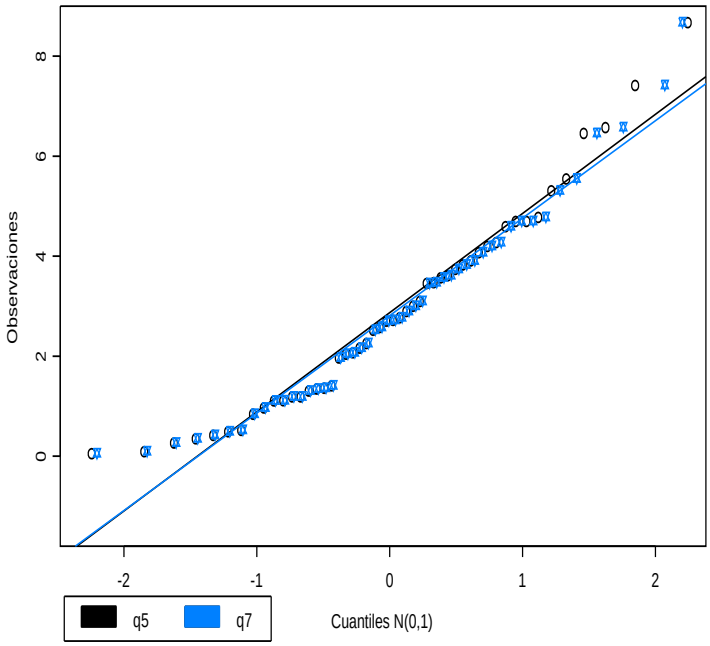

Figura 13: Distribución asimétrica a la derecha considerando p_i^7 y p_i^5

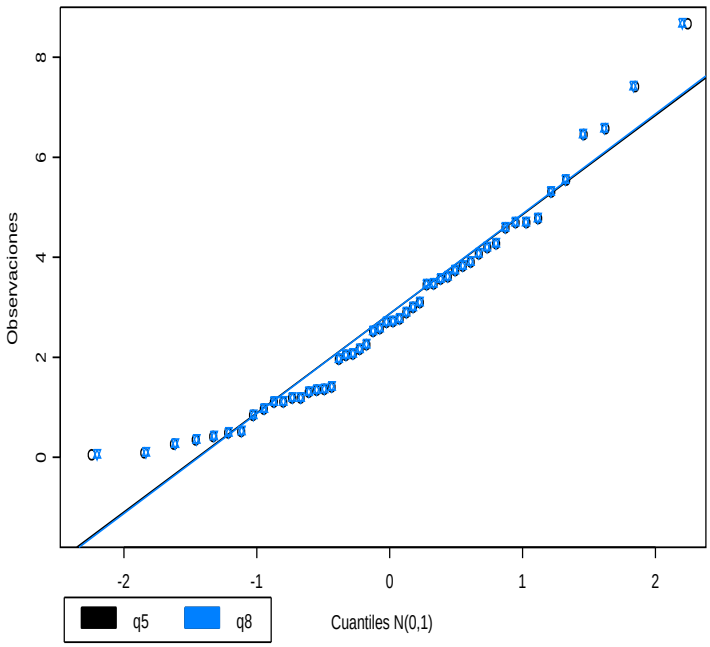

Figura 14: Distribución asimétrica a la derecha considerando p_i^8 y p_i^5

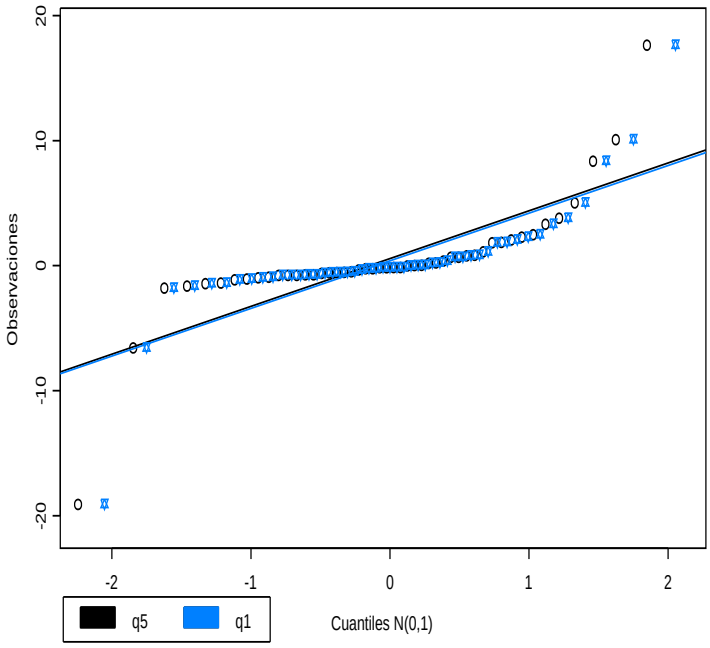

Figura 15: Distribución platicúrtica considerando p_i^1 y p_i^5

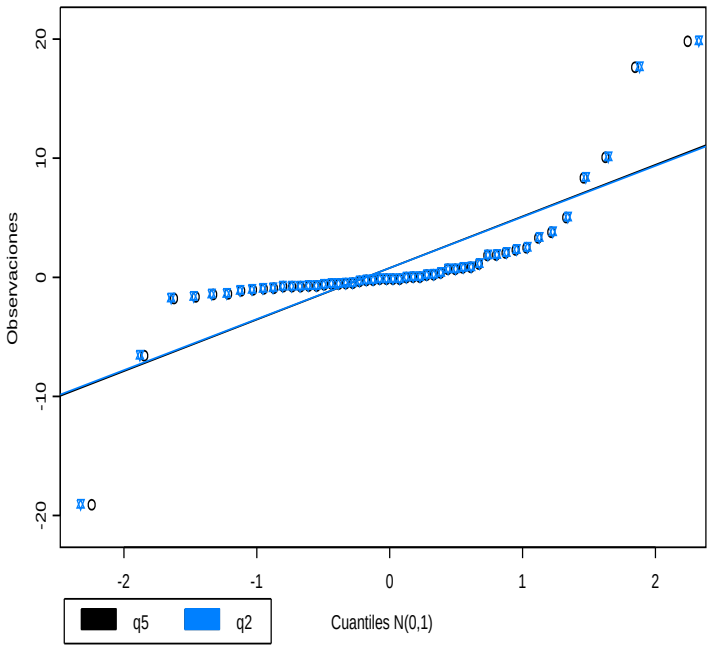

Figura 16: Distribución platicúrtica considerando p_i^2 y p_i^5

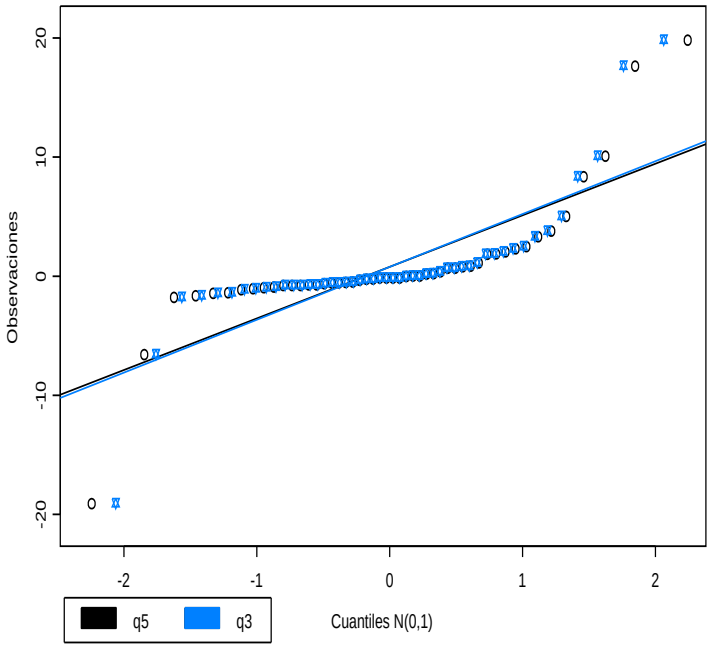

Figura 17: Distribución platicúrtica considerando p_i^3 y p_i^5

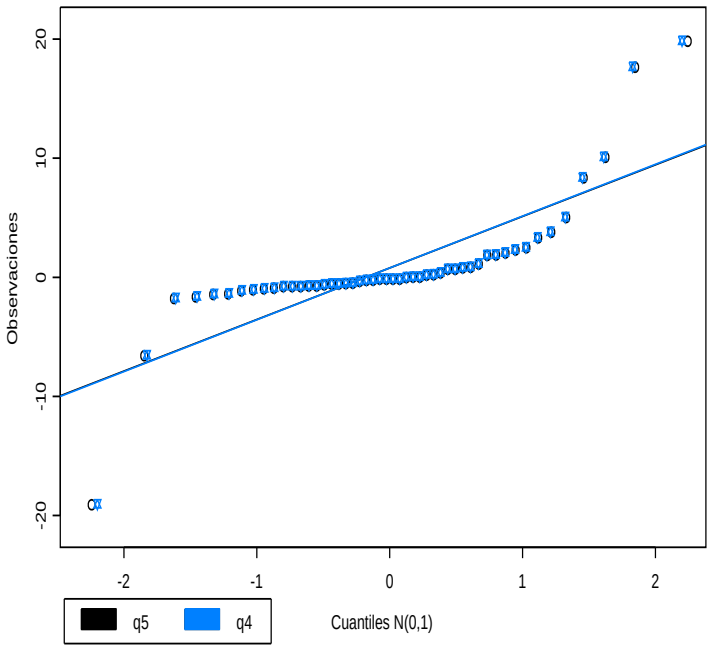

Figura 18: Distribución platicúrtica considerando p_i^4 y p_i^5

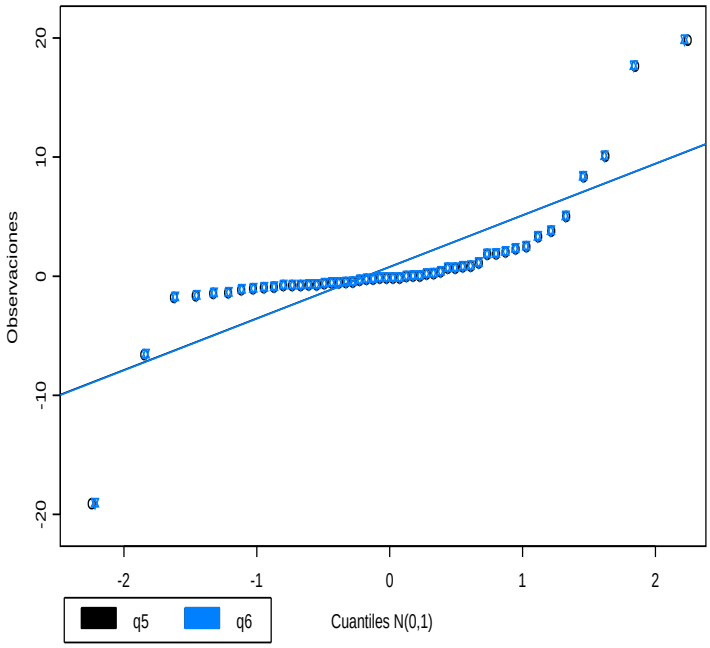

Figura 19: Distribución platicúrtica considerando p_i^6 y p_i^5

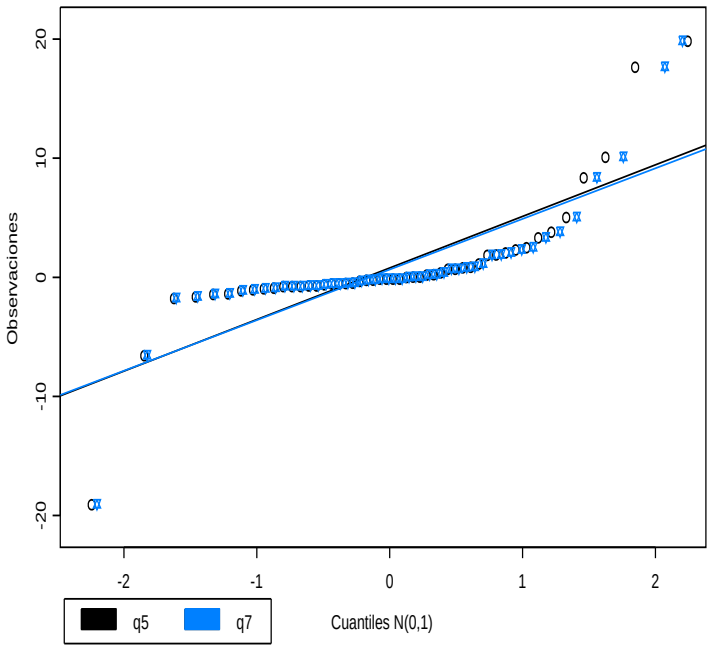

Figura 20: Distribución platicúrtica considerando p_i^7 y p_i^5

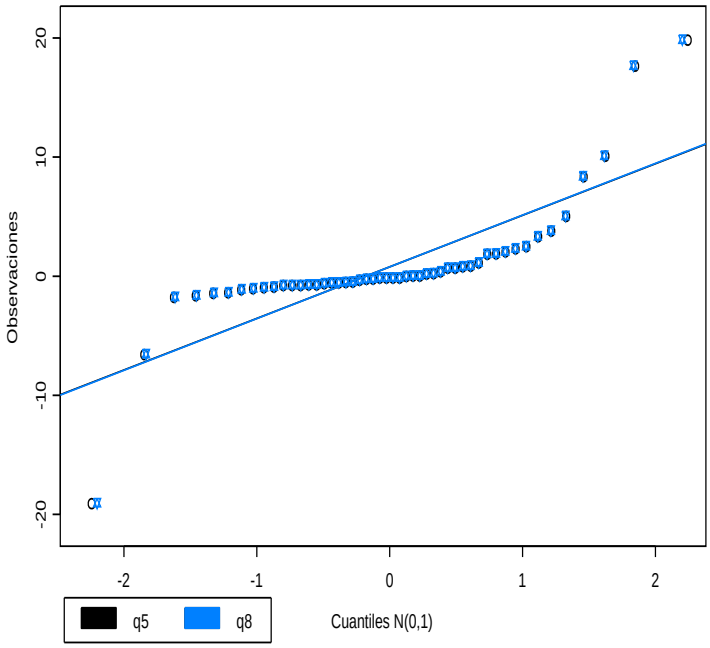

Figura 21: Distribución platicúrtica considerando p_i^8 y p_i^5

En los cuadros 1, 2 y 3, aparecen las diferencias entre las distintas rectas representadas.

En el cuadro 1, se muestran las diferencias correspondientes cuando las observaciones proceden de una distribución normal. El segundo cuadro corresponde a las observaciones generadas de una distribución asimétrica a la derecha. Y por último, aparecen las diferencias de las rectas correspondientes a las observaciones que presentan una curtosis distinta a la normal, más concretamente se trata de una distribución platicúrtica.

5. Conclusiones

En el primer grupo de gráficos, podemos observar que las diferencias entre las distintas rectas son como mucho del orden de centésimas, lo que nos puede llevar a pensar que en caso de normalidad, la elección de un punto de posición gráfica, aunque produzca alguna diferencia, no resulta altamente significativa.

En el segundo y tercer grupo de gráficos, hemos obtenido que la mayor diferencia se produce entre q_i^1 y q_i^5. Dado que la definición del punto de posición gráfica $p_i^1 = \frac{i}{n}$ fue descartada en

Cuantiles	Diferencia entre las rectas
$q_i^1 - q_i^5$	0.032793150
$q_i^2 - q_i^5$	0.020969461
$q_i^3 - q_i^5$	0.051784563
$q_i^4 - q_i^5$	0.011232398
$q_i^6 - q_i^5$	0.006310240
$q_i^7 - q_i^5$	0.040895369
$q_i^8 - q_i^5$	0.008818434

Cuadro 1: Diferencia entre las rectas para observaciones procedentes de una distribución normal

Cuantiles	Diferencia entre las rectas
$q_i^1 - q_i^5$	0.26329425
$q_i^2 - q_i^5$	0.07489709
$q_i^3 - q_i^5$	0.18569473
$q_i^4 - q_i^5$	0.04019290
$q_i^6 - q_i^5$	0.02257366
$q_i^7 - q_i^5$	0.13798318
$q_i^8 - q_i^5$	0.03125474

Cuadro 2: Diferencia entre las rectas para observaciones procedentes de una distribución asimétrica a la derecha

Cuantiles	Diferencia entre las rectas
$q_i^1 - q_i^5$	1.45575024
$q_i^2 - q_i^5$	0.09254288
$q_i^3 - q_i^5$	0.24789369
$q_i^4 - q_i^5$	0.05150238
$q_i^6 - q_i^5$	0.02876858
$q_i^7 - q_i^5$	0.30089663
$q_i^8 - q_i^5$	0.03166735

Cuadro 3: Diferencia entre las rectas para observaciones procedentes de una distribución platicúrtica

su momento rápidamente, como ya hemos indicado, vamos a obviar esta diferencia en los tres grupos de gráficos.

Por tanto, obviando el primer dato de cada grupo, observamos que las mayores diferencias se producen al comparar los cuantiles q_i^3 y q_i^7, respectivamente, con q_i^5. De forma, que las respectivas definiciones de los puntos de posición gráfica, son las que sugieren que difieren con respecto a la definición de Blom en un mayor grado.

En definitiva, y a modo de resumen, podemos indicar, como hemos observado, que la elección de la definición de puntos de posición gráfica influye en la forma final del gráfico de probabilidad, lo que nos podría llevar a la conclusión que, en aquellos casos en los que éste sea usado para la prueba de normalidad, podríamos obtener resultados diferentes, en función de la elección previa de los p_i. Esta conclusión sería igual de válida para aquellos contrastes cuya definición estuviese basada en el gráfico de probabilidad o directamente en los p_i. En definitiva, que el test usado sería sensible a la elección de los puntos de posición gráfica.