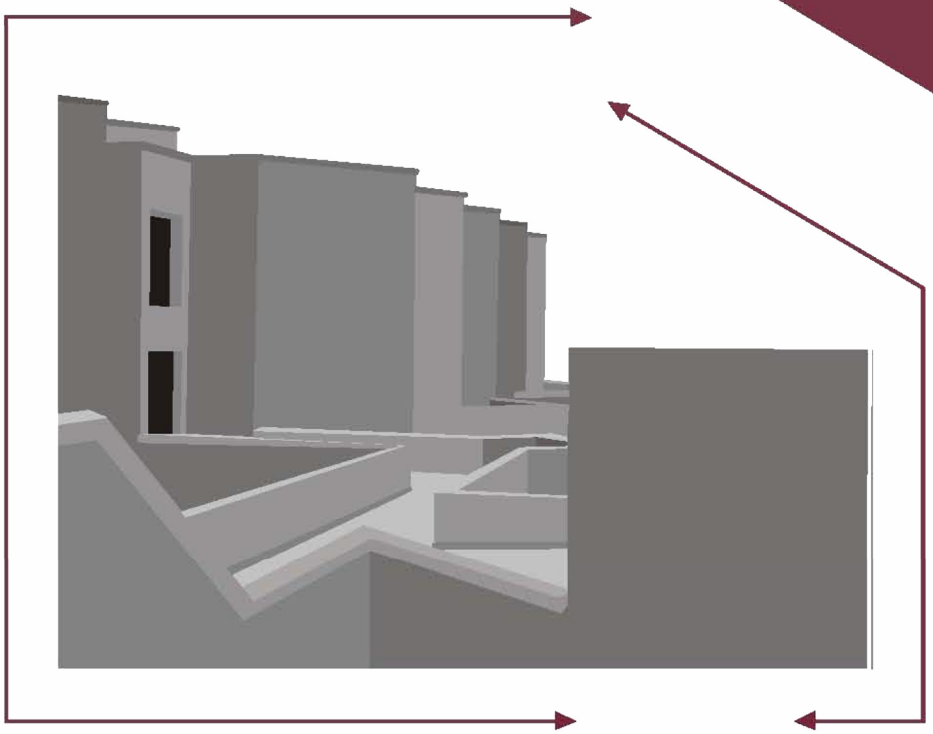

Comunicaciones del CIMAT

Intervalos de predicción simultáneos de cuantiles para validación de modelos estadísticos

Eloisa Díaz-Francés Murguía

Comunicación del CIMAT: I-18-01/07.02.2018
(PE/CIMAT)

CIMAT

Intervalos de predicción simultáneos de cuantiles para la validación de modelos estadísticos

Eloísa Díaz-Francés Murguía (CIMAT) PE/CIMAT I-18-01/07-02-2018

Febrero de 2018

Abstract

Se propone usar la unión de n intervalos de predicción para cuantiles de manera simultánea, basados en una cota inferior para la probabilidad global de todos ellos, para validar si un modelo estimado considerado para los datos es razonable. Para la construcción de los intervalos de predicción se aprovecha la distribución teórica de las estadísticas de orden transformadas bajo la distribución estimada. La banda de predicción propuesta es fácil de calcular y resulta un complemento visual muy útil para las gráficas de cuantiles y de probabilidad usuales.

1 Validación de un modelo estadístico para una muestra observada

Supóngase que se desea modelar un fenómeno natural aleatorio de interés y que para ello se ha observado una muestra de n variables continuas independientes $X_1, ..., X_n$, las cuales se supone se distribuyen idénticamente como $F_X(x;\theta)$. Se desea encontrar un buen candidato para F que esté respaldado por datos observados. Una parte importante de la modelación estadística, como describen Sprott (2000, Cap. 1) y Box (1980), consiste precisamente en el planteo y propuesta de una distribución particular F que se piense tenga todas las propiedades para describir bien el comportamiento aleatorio que rige a los datos observados. Después, con los datos observados se estiman los parámetros θ desconocidos de la distribución propuesta F a través de un estimador puntual óptimo $\hat{\theta}$ o de una región de estimación para θ. El siguiente paso en la modelación estadística recae en validar el modelo estimado $F\left(x;\hat{\theta}\right)$ con los datos observados y/o con muestras similares observadas. Las gráficas cuantil-cuantil, (QQ por sus siglas en inglés) y las de probabilidad descritas aquí permiten validar visualmente de manera práctica y fácil la bondad del ajuste del modelo estadístico esitmado propuesto para los datos observados. En ellas se comparan la distribución empírica de los datos contra la distribución estimada propuesta.

Recuérdese que dos distribuciones serán cercanas en tanto sus cuantiles también lo sean. Los datos observados, $x_1, ..., x_n$, son una realización de la muestra contemplada. A su vez, los datos resultan ser cuantiles de la distribución empírica y por ello se les llama también cuantiles empíricos. La distribución empírica se define en la Sección 2 y se denotará como $F_n(x)$. Debido a dos resultados límite que allí se describen, se tiene que la distribución empírica $F_n(x)$ converge a la verdadera distribución $F_X(x)$ bastante rápido, siendo ya muy cercanas incluso para muestras moderadas. Por ello, debe ocurrir que los cuantiles de ambas distribuciones sean cercanos. Como en una gráfica QQ se comparan los cuantiles empíricos contra los de la distribución estimada propuesta, cuando esta última sea razonable, los puntos graficados deberán caer cerca de la recta identidad y además seguir un patrón aproximadamente lineal.

Dichas gráficas QQ se han usado desde hace más de 50 años, véanse Wilks y Gnanadesikan (1968) y Lawless (2003, Sección 3.3). Sin embargo, aquí se propone agregarles una banda de probabilidad, conformada por la unión de n intervalos de predicción individuales de los cuantiles de la distribución estimada asociados a las mismas probabilidades que corresponden a los cuantiles empíricos. Cuando uno o más de los datos observados caiga fuera de la banda propuesta, entonces se tendrá evidencia, con cierta probabilidad, en contra del modelo propuesto. Un buen ajuste de la distribución estimada suele ir asociada a que todos los datos observados caigan dentro de la banda propuesta y además que se acomoden de manera aproximadamente lineal.

Para describir la propuesta con claridad, primero se presentan varios resultados teóricos que la sustentan y describen en las Secciones 2 a 5. En la Sección 3 se presenta con detalle cómo construir las gráficas de cuantiles y de probabilidades. Las bandas de probabilidad asociadas para ellas se presentan en la Sección 6. En la Sección 7 se presenta un ejemplo de tiempos entre llegadas de grupos de visitantes al Museo de las Momias de Guanajuato en un día festivo de 2017. Finalmente, las conclusiones se presentan en la Sección 8. Para más detalles sobre representaciones gráficas que ayudan a valorar si algún modelo estadístico considerado describe razonablemente bien a un conjunto de datos observados, se recomienda ver la Sección 3.3 del libro de Lawless (2003).

2 Cercanía entre la función de distribución empírica y la teórica

Sea $X_1, ...X_n$ una muestra de n variables aleatorias continuas independientes e idénticamente distribuídas con distribución $F(x;\theta_0)$, la cual se desconoce. Para la muestra de variables considerada, se define la función de distribución empírica como el promedio de las siguientes funciones indicadoras, las cuales resultan ser a su vez variables aleatorias Bernoulli,

$$F_n(x) = \frac{1}{n}\sum I_{(-\infty,x]}(X_i).$$

Nótese que la variable aleatoria $Z_i = I_{(-\infty,x]}(X_i)$ es una variable Bernoulli, puesto que toma solamente dos valores, cero y uno. El valor de uno lo toma con probabilidad igual a $P[X_i \leq x] = F_X(x)$. Vista así, para cada valor fijo x, la función de distribución empírica resulta ser un promedio de variables aleatorias Bernoulli, todas ellas con el mismo valor esperado. Es decir, nótese que

$$E\left[I_{(-\infty,x]}(X_i)\right] = P[X_i \leq x] = F_X(x).$$

Por ello, por la Ley Fuerte de los Grandes Núméros, resulta que este promedio de variables aleatorias, $F_n(x)$, debe converger a su valor esperado con probabilidad uno,

$$F_n(x) \overset{cp1}{\to} F_X(x),$$

para todo valor x. Glivenko y Cantelli en 1933 demostraron un resultado aún más fuerte que es la convergencia uniforme a la distribución teórica F_X, el cual se enuncia a continuación.

2.0.1 Teorema de Glivenko-Cantelli

Para una muestra $X_1, ...X_n$ de n variables aleatorias continuas independientes e idénticamente distribuídas con distribución $F_X(x;\theta)$, la función de distribución empírica correspondiente converge uniformemente a la distribución teórica,

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_X(x)| \overset{cp1}{\to} 0.$$

Ambos resultados mencionados implican también el resultado más débil que la función de distribución empírica converge en distribución a la teórica. La cercanía entre la distribución empírica y la teórica se suele dar incluso para muestras de tamaño moderado, por ejemplo donde $n \geq 20$. Todo ello se puede verificar a través de simulaciones.

Una vez que se cuente con una realización de la muestra, a la cual se denotará con minúsculas, $x_1, ..., x_n$, se podrá graficar una realización de la función de distribución empírica,

$$F_n(x) = \frac{1}{n}\sum I_{(-\infty,x]}(x_i).$$

La gráfica de esta función presenta brincos de tamaño $1/n$ en cada valor observado. Se trata de una función escalonada no decreciente que es similar a la función de distribución discreta uniforme que toma valores en $\{x_1, ..., x_n\}$ con probabilidades $1/n$ en cada uno de los n valores observados.

Por otra parte, defínase al cuantil Q_α de probabilidad α asociado a una distribución $F_X(x) = P[X \leq x]$, como el valor más pequeño x tal que la variable aleatoria X ha acumulado una probabilidad α hasta él. Generalizando este concepto, se suele definir la función de cuantiles como una función con dominio en el intervalo $[0,1]$ y contradominio en los reales, tal que es la función inversa generalizada de la función de distribución F_X,

$$Q_\alpha = F_X^{-1}(\alpha) = \acute{i}nf\{x : F_X(x) \geq \alpha\}. \tag{1}$$

Conviene definirla como una función inversa generalizada de la distribución de X, para así poderla aplicar a la distribución asociada a cualquier tipo de variable aleatoria, sea discreta o continua. En el caso de variables discretas las distribuciones son funciones escalonadas y calcular así los cuantiles será inmediato y simple. Si F_X es una distribución de una variable aleatoria continua, entonces la función de cuantiles es simplemente la función inversa usual de F_X.

Nótese que esta definición corrobora que los valores observados $x_1, ..., x_n$ son precisamente cuantiles de la función de distribución empírica $F_n(x)$. A una probabilidad α, le corresponde sólo un único valor observado como cuantil. Sin embargo, nótese que el i-ésimo valor observado ordenado, $x_{(i)}$, resulta ser cuantil de más de una probabilidad. De hecho, $x_{(i)}$ es cuantil de cualquier probabilidad que esté en el intervalo

$$\left(\frac{i-1}{n}, \frac{i}{n}\right].$$

En particular la probabilidad $\alpha_i = i/(n+1)$ está contenida en este intervalo. El lector puede verificar esto fácilmente. Convendrá decir que la i-ésima estadística de orden $x_{(i)}$ es el cuantil empírico de probabilidad α_i como se verá en las siguientes secciones. Así, se dirá que los cuantiles empíricos observados ordenados $x_{(1)}, x_{(2)}, ..., x_{(n)}$ corresponden a las probabilidades

$$\frac{1}{n+1}, \frac{2}{n+1}, ..., \frac{n}{n+1},$$

para identificar a cada cuantil empírico con tan sólo una probabilidad por sencillez.

3 Comparación gráfica de dos distribuciones

Hay varias maneras gráficas de comparar dos distribuciones continuas F y G. La más inmediata sería graficando directamente dichas funciones $F(x)$, $G(x)$ para una rejilla fina de valores de x en el eje horizontal. También se podrían graficar las dos funciones de densidad correspondientes, $f(x), g(x)$, en una misma gráfica. La cercanía de tales curvas indicaría visualmente la cercanía entre las distribuciones, si bien las escalas van a ser diferentes según se grafiquen las distribuciones o las densidades. A veces resulta más fácil notar discrepancias en la escala de las densidades que de las distribuciones para un mismo par de funciones F, G.

El recordar que dos distribuciones son iguales si acumulan la misma probabilidad para todo valor x en los reales o si sus cuantiles son iguales, lleva a considerar aún más opciones para compararlas gráficamente. Una de ellas es comparando los cuantiles de una distribución en un eje coordenado contra los cuantiles de la otra distribución en el otro eje. A manera de referencia, se agrega también la recta identidad en la misma gráfica, la cual denota los puntos donde ambas distribuciones son iguales. Al comparar los cuantiles de ellas se está comparando las funciones inversas de ambas distribuciones. Los puntos que se grafican estarán en la escala de los valores que toman las variables aleatorias asociadas.

Otra manera de comparar dos distribuciones es graficando puntos de coordenadas

$$[F(x), G(x)]$$

para una colección de puntos x en los reales. Los puntos graficados estarán contenidos en el cuadrado unitario, puesto que F y G son probabilidades acumuladas. Se suele agregar también la recta identidad. A este tipo de gráficas se les suele llamar gráficas de probabilidades. Las gráficas de cuantiles y las de probabilidades más usadas se describen enseguida con mayor detalle.

3.1 Gráficas Cuantil-Cuantil (QQ)

Al tipo de gráficas donde se comparan los cuantiles de dos distribuciones se les llama gráficas de cuantiles o gráfica QQ, debido a la abreviación del nombre en inglés (*QQ plot*). Como describen Martin Wilk y Ramanathan Gnanadesikan (1968), las graficas de cuantiles son una de las técnicas gráficas más usadas para la comparación de distribuciones. Si dos distribuciones son iguales, entonces sus cuantiles coinciden.

Se contrastan los cuantiles empíricos, en el eje vertical usualmente, contra los cuantiles estimados o teóricos asociados a las mismas probabilidades en el eje horizontal. Los cuantiles empíricos consisten de las estadísticas de orden observadas, $x_{(1)}, x_{(2)}, ..., x_{(n)}$ y estos cuantiles empíricos corresponden a las probabilidades $1/(n+1), 2/(n+1), ..., n/(n+1)$, respectivamente. En algunos libros se considera que el i-ésimo cuantil empírico, $x_{(i)}$ es cuantil de probabilidad $(i-0.5)/n$, pero como se verá más adelante, resulta más conveniente considerar que corresponde a la probabilidad $i/(n+1)$ como se hace aquí.

Por otra parte los cuantiles estimados se obtienen aplicando la función de distribución estimada inversa a las probabilidades mencionadas,

$$\alpha_1 = \frac{1}{n+1}, \alpha_2 = \frac{2}{n+1}, ..., \alpha_n = \frac{n}{n+1},$$

para así obtener

$$Q(\alpha_i) = F^{-1}\left(\alpha_i; \vec{\theta}\right),$$

para $i = 1, ..., n$, donde $\hat{\theta}$ es un estimador de los parámetros desconocidos que se obtiene con los datos observados.

Las coordenadas de los puntos que se grafican para obtener la gráfica QQ son

$$\left[F^{-1}\left(\alpha_i; \vec{\theta}\right), x_{(i)}\right] = \left[F^{-1}\left(\frac{i}{n+1}; \vec{\theta}\right), x_{(i)}\right], \text{ para } i = 1, ..., n.$$

Usualmente se grafica también la recta identidad para contrastar visualmente la cercanía de cada punto graficado con ella, la cual representa la situación en la que las dos distribuciones graficadas coinciden. Se graficarán así en el eje vertical los cuantiles empíricos de probabilidad $k/(n+1)$ contra los cuantiles estimados de las mismas probabilidades para $k = 1, ..., n$.

3.2 Gráficas de probabilidad (PP)

Para estas gráficas se aprovechan dos resultados teóricos muy útiles. Por el Teorema de la Transformada Integral de Probabilidad se sabe que si una variable aleatoria continua X tiene distribución $F_X(x; \theta)$ entonces la variable aleatoria definida como $U = F_X(X; \theta)$ tiene una distribución uniforme continua en $(0, 1)$. Por ello, los datos $x_1, ..., x_n$ transformados con la distribución estimada $F\left(x; \hat{\theta}\right)$,

$$u_1 = F\left(x_1; \hat{\theta}\right), ..., u_n = F\left(x_n; \hat{\theta}\right),$$

deberían seguir una distribución uniforme si el modelo propuesto F y estimado es razonable.

Es decir, si el modelo estimado F es razonable, las estadísticas de orden asociadas, $u_{(1)}, ..., u_{(n)}$, resultan ser estadísticas de orden de una muestra de n variables aleatorias uniformes en $(0, 1)$.

Por otra parte, por el Teorema 2.4.3. de Gibbons y Chakraborti (2011) se sabe que la estadística de orden transformada $u_{(k)}$, por ser una estadística de orden de una muestra de uniformes sigue una distribución Beta con parámetros k y $(n+1-k)$,

$$u_{(k)} \sim Beta(k, n+1-k), \text{ para } k = 1, ..., n.$$

El valor esperado de $u_{(k)}$ es

$$E\left(u_{(k)}\right) = \frac{k}{n+1}.$$

Véase la semejanza con las probabilidades asociadas a los cuantiles empíricos.

Para la gráfica PP se graficarán los puntos de coordenadas

$$\left[\left(\frac{k}{n+1}\right), u_{(k)}\right], \text{ para } k = 1, ..., n. \tag{2}$$

Con ello se estarán comparando los valores transformados de estadísticas de orden $u_{(k)}$, en el eje vertical, contra sus valores esperados en el eje horizontal. Usualmente se grafica también la recta identidad dentro del cuadrado unitario, que resulta la diagonal del cuadrado con pendiente positiva. La distancia vertical entre un punto graficado con respecto a la recta identidad describe la distancia entre $u_{(k)}$ y su valor esperado, lo que es igual a $\left|u_{(k)} - E\left[u_{(k)}\right]\right|$.

Nótese que se pueden interpretar a los puntos que se grafican en el eje horizontal de una gráfica QQ como la transformación de los valores esperados de $u_{(k)}$ al aplicarles la distribución inversa estimada $\hat{F}_X^{-1}$.

4 Intervalos de predicción para un cuantil

Como se mencionó en la sección anterior, la k-ésima estadística de orden de una muestra de n variables aleatorias uniformes en $(0, 1)$, denotada como $u_{(k)}$ sigue una distribución Beta cuya varianza es

$$V\left(u_{(k)}\right) = \frac{k\left(n+1-k\right)}{\left(n+1\right)^2\left(n+2\right)}.$$

Nótese que entre mayor sea el tamaño de muestra n, la varianza de cada estadística de orden será menor.

Por tanto, un intervalo de predicción de probabilidad $(1-\beta)$ para esta variable aleatoria Beta, $u_{(k)}$, se obtiene al considerar los cuantiles de probabilidades $\beta/2$ y $(1-\beta/2)$ para los extremos del intervalo

$$\left[Q^B_{\beta/2}, Q^B_{1-\beta/2}\right] \tag{3}$$

Este intervalo contiene al valor esperado $E\left[u_{(k)}\right] = k/\left(n+1\right)$. Entre mayor sea el tamaño de la muestra, más angosto será este intervalo de predicción.

Por ejemplo si $(1-\beta) = 0.95$, un intervalo de predicción a partir de (3) puede ser con los cuantiles de probabilidad $\beta/2 = 0.025$ y $(1-\beta/2) = 0.975$ puesto que se cumple

$$0.95 = P\left[Q^B_{0.025} \leq u_{(k)} \leq Q^B_{0.975}\right], \text{ para } k = 1, ..., n.$$

Nótese que también se pudieron usar los cuantiles de probabilidad 0.01 y 0.96 para calcular el intervalo de predicción. Usualmente se eligen los cuantiles tales que dan lugar al intervalo de predicción más corto, el cual siempre incluye al valor esperado $E\left(u_{(k)}\right)$.

Ahora, al notar que la estadística de orden en la escala original se obtiene aplicando la función de distribución estimada inversa,

$$x_{(k)} = \hat{F}^{-1}\left(u_{(k)}; \hat{\theta}\right),$$

entonces también se puede transformar un intervalo de predicción Beta (3) a la escala original de la variable X, aplicando a los extremos del dicho intervalo la función de distribución estimada inversa. Así, un intervalo de predicción de probabilidad $(1-\beta)$ para el cuantil empírico $x_{(k)}$ será

$$\left[\hat{F}_X^{-1}\left(Q^B_{\beta/2}\right), \hat{F}_X^{-1}\left(Q^B_{1-\beta/2}\right)\right]. \tag{4}$$

Nótese que los intervalos de predicción se pueden transformar de una escala a otra fácilmente según convenga puesto que se cumplen las siguientes igualdades,

$$\begin{aligned} 1-\beta &= P\left[Q^B_{\beta/2} \leq u_{(k)} \leq Q^B_{1-\beta/2}\right] = P\left[F_X^{-1}\left(Q^B_{\beta/2}\right) \leq F_X^{-1}\left(u_{(k)}\right) \leq F_X^{-1}\left(Q^B_{1-\beta/2}\right)\right] \\ &= P\left[F_X^{-1}\left(Q^B_{\beta/2}\right) \leq x_{(k)} \leq F_X^{-1}\left(Q^B_{1-\beta/2}\right)\right]. \end{aligned} \tag{5}$$

5 Cálculo de una cota para la probabilidad global de n intervalos de predicción

Si se desea usar de manera simultánea a n intervalos de predicción, de manera que cada uno de ellos tenga la misma probabilidad individual, pero pudiendo dar una cota inferior para la probabilidad global conjunta de todos ellos, conviene tomar en cuenta a la desigualdad de Bonferroni. Como presentan Casella y Berger (2002, $p.13$), esta desigualdad da una cota inferior para la probabilidad de la intersección de n eventos, A_i, con $i = 1, ..., n$,

$$P\left(\bigcap_{i=1}^{n} A_i\right) \geq \sum_{i=1}^{n} P\left(A_i\right) - n + 1.$$

En este caso, el evento i-ésimo correspondería a un intervalo de predicción para el cuantil i-ésimo, donde $i = 1, ..., n$. Denótese como γ a la probabilidad individual de un intervalo de predicción de un cuantil y supóngase que las probabilidades de los n eventos A_i son iguales,

$$\gamma = P\left(A_1\right) = ... = P\left(A_n\right).$$

Entonces si se desea garantizar que la probabilidad de la intersección de eventos sea igual o mayor que $1 - \beta$, por la desigualdad de Bonferroni, se puede afirmar que

$$(1 - \beta) \geq \sum_{i=1}^{n} \gamma - n + 1.$$

Esta desigualdad se simplifica aún más a

$$(1 - \beta) \geq n\gamma - n + 1,$$

con lo que finalmente se tiene que

$$\gamma \leq 1 - \frac{\beta}{n}.$$

Aplicando estas ideas al caso de n intervalos de predicción de cuantiles, cada uno con la misma probabilidad γ, entonces si se busca que la probabilidad de todos los n intervalos juntos sea al menos de $(1 - \beta)$, entonces convendrá tomar la probabilidad de cada intervalo individual como $\gamma = 1 - \beta/n$. Por ejemplo, si se tienen $n = 5$ intervalos de predicción y se considera que cada uno tenga probabilidad $\gamma = .99$, entonces se garantizará que la probabilidad global de predicción de los cinco intervalos de manera simultánea de al menos $0.95 = 1 - \beta$. Usualmente se toma n igual al número de datos en la muestra observada.

6 Bandas de predicción para gráficas PP y QQ

Para una gráfica PP, a cada valor esperado $E\left(u_{(k)}\right)$ que se grafica en el eje horizontal le corresponderá verticalmente la realización de la estadística de orden transformada $u_{(k)}$ junto con un intervalo de predicción para $u_{(k)}$ de probabilidad individual γ, como se describió en la Sección 5. Esto se hace para $k = 1, ..., n$, de manera que la probabilidad global conjunta de los n intervalos va a ser al menos de probabilidad $(1 - \beta)$. Si se unen los extremos de estos intervalos de predicción se formará así una banda como se muestra en las figuras del ejemplo de la siguiente sección. Si acaso alguno o más de los puntos para alguna muestra observada cayesen fuera de esta banda, entonces eso significará que hay evidencia en contra de la hipótesis de que el modelo estimado $F\left(x; \hat{\theta}\right)$ sea razonable, con un p-valor menor o igual a β.

Ahora, para el caso de las gráficas QQ, basta con transformar a la escala del cuantil las bandas de predicción calculadas para la gráfica PP como se acaba de indicar y entonces se tendrá una banda de predicción global de probabilidad conjunta para todos los cuantiles que al menos es de probabilidad $(1 - \beta)$. Nötese que la transformación inversa aplicada a $u_{(k)}$, arroja precisamente los cuantiles empíricos puesto que

$$F_X^{-1}\left[u_{(k)}\right] = F_X^{-1}\left[F_X\left(x_{(k)}\right)\right] = x_{(k)}.$$

Por otra parte, puesto que $\alpha_i = i/(n+1) = E\left(u_{(k)}\right)$, entonces

$$F_X^{-1}\left\{E\left[\left(u_{(k)}\right)\right]\right\} = Q\left(\alpha_i\right).$$

Así, los cuantiles estimados de probabilidad α_i se obtienen también transformando con F_X^{-1} el eje horizontal de la gráfica PP. Se sugiere primero calcular la banda para la gráfica PP y luego transformarla de regreso a la escala original de los cuantiles.

Quesenberry y Hales (1980) consideran un tipo de gráficas PP similares a la descrita aquí en la Sección 3.2, pero ellos colocaron los ejes coordenados al revés de como se sugiere aquí presentarlas. Ellos denominaron como *bandas de concentración* a una banda de confianza para las estadísticas de orden uniformes $u_{(k)}$, tomando en cuenta la distribución Beta que siguen y calculando intervalos de predicción individuales para las n estadísticas de orden $u_{(k)}$ (véase también el Capítulo 6 escrito por Charles Quesenberry en el libro de D 'Agostino y Stephens, 1986). Sin embargo, no consideraron la desigualdad de Bonferroni para relacionar la probabilidad de predicción individual de cada estadística de orden con su llamada banda de concentración global. En contraste, al hacerlo como se sugiere aquí, se tiene la ventaja de considerar correctamente que en realidad se están haciendo varias pruebas de hipótesis simultáneas sobre si cada cuantil observado es razonable bajo el modelo estimado o no. Tampoco propusieron regresar la gráfica PP a la escala original de los cuantiles como se sugiere hacer aquí.

7 Ejemplo

Considérense una muestra de $n = 47$ observaciones independientes e idénticamente distribuidas que corresponden a los tiempos entre llegadas de grupos de visitantes al Museo de las Momias de Guanajuato el Sábado Santo 15 de abril de 2017 de las 12:00 a las 12:30pm. Los 47 tiempos $x_1, ..., x_{47}$ fueron medidos como parte de su tesis por Ana Paulina Pérez Romero, exalumna graduada de la Maestría de Probabilidad y Estadística del CIMAT. Los datos registrados en segundos fueron:

$$54, 24, 54, 6, 14, 87, 42, 30, 21, 4, 43, 34, 25, 26, 79, 35, 15, 25, 33, 20, 12, 34, 24, 29, 28,$$
$$32, 83, 40, 32, 94, 27, 47, 52, 68, 14, 44, 34, 32, 32, 6, 18, 80, 90, 40, 100, 25, 30.$$

La suma de las observaciones es $T = \Sigma_{i=1}^{47} x_i = 1818$. Algunos de los datos aparecen como repetidos por haber sido redondeados al segundo más cercano. Sin embargo, debido a que el valor esperado bajo el modelo estadístico considerado es bastante mayor que la unidad de redondeo de un segundo, es razonable considerar la aproximación que estos tiempos fueron medidos con precisión muy fina. Es decir no es necesario estimar los parámetros considerando una verosimilitud con censura por intervalos puesto que se llegarían prácticamente a las mismas conclusiones para este ejemplo que al considerar a la verosimilitud como proporcional al producto de las densidades marginales de la muestra (véase Kalbfleisch, 1985, Sección 9.5). Las conclusiones sobre el ajuste del mejor modelo no se verán afectadas por ello..

Se consideraron dos modelos estadísticos para estos datos que se desean comparar a través de gráficas QQ y PP. El primer modelo es la distribución exponencial con tiempo medio de vida θ, cuya densidad es

$$f\left(x;\theta\right) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right) I_{(0,\infty)}\left(x\right).$$

El segundo modelo es la distribución Gama de parámetros de forma α y de escala β, cuya densidad es

$$g\left(x;\alpha,\beta\right) = \frac{1}{\Gamma\left(\alpha\right)\beta^{\alpha}} x^{\alpha-1}\exp\left(-\frac{x}{\beta}\right) I_{(0,\infty)}\left(x\right).$$

Para el modelo exponencial, el estimador de momentos de θ coincide con el de máxima verosimilitud y es $\hat{\theta} = T/n = 38.68$. Para el modelo Gama, unos estimadores de momentos son $\tilde{\alpha} = 2.55$ y $\tilde{\beta} = 15.17$. Los estimadores de máxima verosimilitud son $\hat{\alpha} = 2.52, \hat{\beta} = 15.33$.

Bajo el modelo exponencial y el gama estimados, el tiempo esperado de llegada coincide y es de 38.68 segundos. Bajo el modelo exponencial, las Figuras 1 y 2 muestran las gráficas PP y QQ descritas con base en los estimadores

de máxima verosimilitud. Las bandas de predicción son del 95% de probabilidad y para ello como $n = 47$, se tomó la probabilidad individual de predicción para cada estadística de orden como $\gamma = 0.9989$.

En la gráfica PP exponencial se nota que tres puntos muestrales caen fuera de la banda de predicción. Por tanto esto da evidencia de que el modelo exponencial no es razonable para los datos. La gráfica QQ exponencial vuelve a ratificar esta conclusión pues quedan excluidos los mismos 3 puntos en la escala original de las observaciones, pero además se resalta que varios tiempos de llegada yacen sobre la banda superior de confianza y además los puntos graficados están curveados y son cóncavos, en vez de más lineales como se esperaría estuvieran si el modelo exponencial fuese razonable. Varios tiempos chiquitos que se observaron resultaron ser más grandes que los tiempos estimados bajo el modelo exponencial y por el contraste los tiempos grandes observados son bastante más chicos que los que se esperaría ver con la distribución exponencial.

En gran contraste, las Figuras 3 y 4 muestran las gráficas PP y QQ bajo el modelo Gama estimado. Ambas gráficas muestran que los puntos son más lineales y cercanos a la recta identidad, además de que todos los puntos graficados están contenidos y muy al centro de las bandas de predicción. Por tanto a la luz de estas gráficas se concluye que la distribución exponencial no es un buen modelo para los tiempos interarribo observados. En contraste la distribución Gama estimada si parece describir razonablemente bien a estos tiempos.

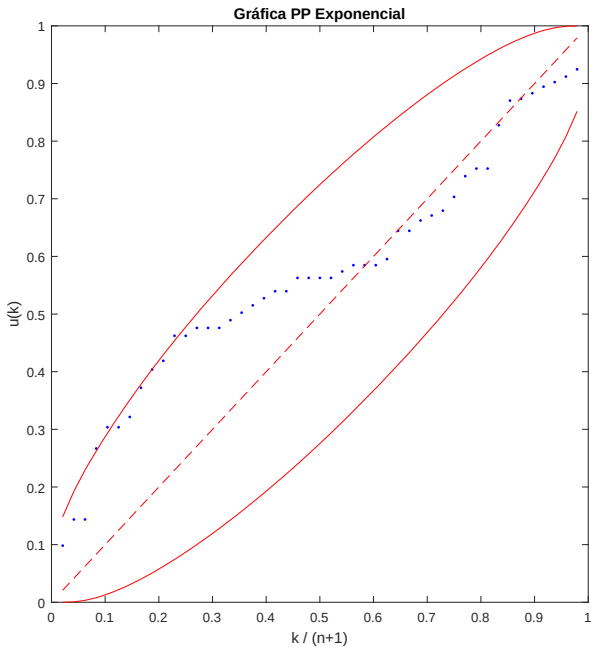

Figura 1. Gráfica PP Exponencial.

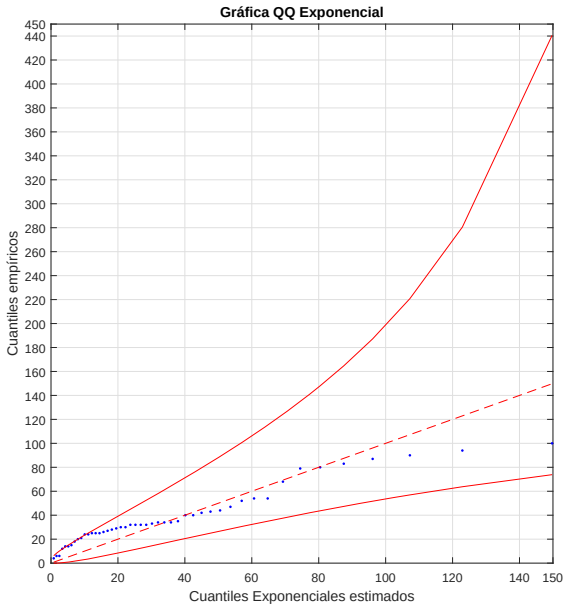

Figura 2. Gráfica QQ exponencial.

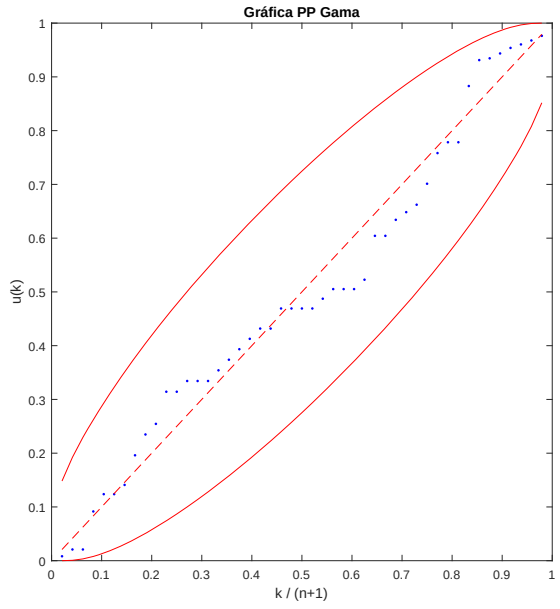

Figura 3. Gráfica PP Gama.

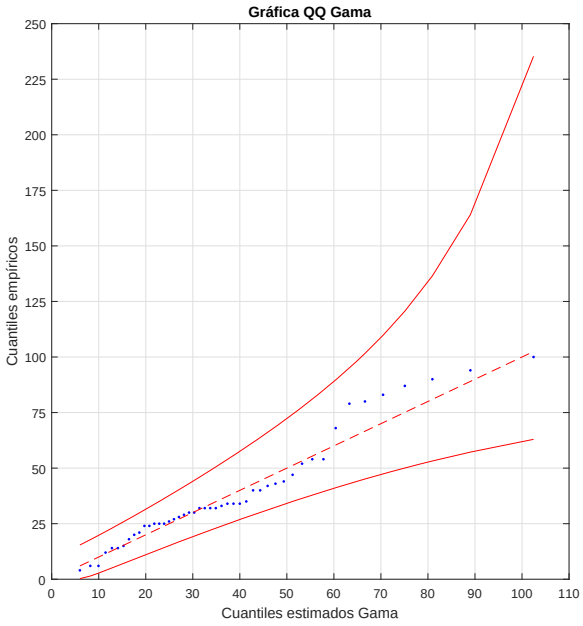

Figura 4. Gráfica QQ Gama.

8 Conclusiones

Las gráficas QQ y PP aquí descritas son complementarias entre si y permiten valorar la cercanía entre la distribución empírica y la distribución estimada planteada, tanto en las colas como en la parte central de las distribuciones. Se sugiere graficar ambas, ya que son muy sencillas de programar y sirven muy bien para validar el ajuste del modelo estimado propuesto a los datos.

Los intervalos de predicción para la estadística de orden uniforme $u_{(k)}$ similares a los descritos en la Sección 4 fueron propuestos originalmente por Quesenberry y Hales (1980). Sin embargo, dichos autores consideraron los ejes coordenados al revés y no relacionaron la probabilidad individual de cada intervalo de predicción con la probabilidad conjunta de la unión de todos los intervalos de predicción como se sugiere hacer aquí.

Intervalos similares de predicción para los cuantiles a los aquí propuestos se pueden obtener también a través de simulaciones, con la distribución empírica de la estadística de orden $x_{(k)}$ obtenida tras simular un número grande de veces muestras de tamaño n bajo el modelo F estimado. Cada muestra simulada de tamaño n se ordena y se obtiene así un valor simulado de $x_{(k)}$. Al contar con un número grande de M simulaciones, se podrán obtener los cuantiles empíricos de probabilidades $(1-\gamma/2)$ y $\gamma/2$ de la distribución empírica de la estadística de orden k-ésima $x_{(k)}$, a la manera en que sugirieron Hernández et al.(2002). A través de simulaciones, se puede corroborar que ese enfoque da resultados muy parecidos al aquí propuesto, dando lugar a bandas de predicción conjuntas similares, siempre que se tome en cuenta la cota sugerida aquí que se obtiene al usar la desigualdad de Bonferroni.

Al aumentar el tamaño de muestra n, los intervalos de predicción en ambas escalas, ya sea la de $u_{(k)}$ o la de $x_{(k)}$, serán más angostos. Para muestras más pequeñas, incluso si el modelo estimado es el correcto, se notará una mayor discrepancia entre los puntos graficados y la recta identidad tanto en la gráfica PP como en la QQ. En tanto aumenta n, los puntos se acercarán más a la recta identidad cuando el modelo estimado sea razonable.

Cuando para una muestra observada se tenga que alguno de los puntos observados no caiga en la banda de predicción de cuantiles, se tendrá evidencia en contra del modelo considerado. Sin embargo, las bandas de cuantiles son conservadoras y puede ser que todos los puntos muestrales estén contenidos en ellas a pesar que la distribución

considerada F no sea el mejor modelo que describa bien a los datos. En tal caso, lo que se puede decir es que F es cercana o parecida a la verdadera distribución atrás de los datos observados.

En resumen, cuando una gráfica QQ o PP rechace algún modelo, se trata de un caso donde el modelo considerado claramente NO es razonable para los datos observados. En contraste, cuando los puntos observados estén contenidos en las bandas de confianza, se dirá que el modelo considerado es razonable pero bien podría haber otros que sean aún mejores para los datos observados y habrá que estar alerta para dar con ellos.

9 Referencias

1. Box, G. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *JRSS*, Serie A, **143**, (4),383 − 430.

2. Casella, G. y R. L. Berger (2002, Segunda edición). Statistical Inference. E.U.A: Brooks Cole.

3. D 'Agostino y M. Stephens (1986). *Goodness of Fit Techniques*. Nueva York: Marcel Dekker, Inc.

4. Gibbons, J.D. y Chakraborti, S. (2011). *Nonparametric Statistical Inference*. Boca Raton: CRC Press.

5. Hernández-Campos, F, Marron, J. S., Samorodnitsky, G. y Smith, F. D. (2002). Variable Heavy Tails in Internet Traffic. *Performance Evaluation,* **58**, 261 − 284.

6. Kalbfleisch, J. G. (1985, 2a edición). *Probability and Statistical Inference*, V.2. Nueva York: Springer-Verlag.

7. Lawless, J.F. (2003, Segunda edición). *Statistical Models and Methods for Lifetime Data*. Nueva Jersey: John Wiley & Sons.

8. Quesenberry, C. P. y Hales, C. (1980). Concentration Bands for Uniformity Plots. J. Statist. Comput. Simul., **11**, 41 − 53.

9. Sprott, D. A. (2000). Statistical Inference in Science. Nueva York: Springer-Verlag.

10. Wilk, M. B. y Gnanadesikan, R. (1968). Probability Plotting Methods for the Analysis of Data. *Biometrika*, **55**, No. 1, 1 − 17.