FACULTAD DE CIENCIAS ECONÓMICAS

UNC Universidad Nacional de Córdoba

REPOSITORIO DIGITAL UNIVERSITARIO (RDU-UNC)

Modelos Zero Inflated Binomial y Poisson en el control estadístico de procesos de alta calidad

Andrea Fabiana Righetti, Silvia Joekes, Cristián Abrego, María Rosa Yacci

Ponencia presentada en XLI Coloquio Argentino de Estadística realizado en 2013 en la Facultad de Ciencias Económicas - Universidad Nacional de Cuyo. Mendoza, Argentina

Esta obra está bajo una Licencia Creative Commons Atribución-NoComercial 4.0 Internacional

XLI COLOQUIO ARGENTINO DE ESTADÍSTICA
15 A 18 DE OCTUBRE 2013
MENDOZA - ARGENTINA

MODELOS ZERO INFLATED BINOMIAL Y POISSON EN EL CONTROL ESTADISTICO DE PROCESOS DE ALTA CALIDAD

RIGHETTI ANDREA F [1], JOEKES SILVIA [2], ABREGO CRISTIAN[3]; YACCI, M. ROSA[4]

Instituto de Estadística y Demografía .FCE. UNC 1 analizamos@yahoo.com.ar

2 joekess@eco.unc.edu.ar; 3 abrego_cristian@hotmail.com ; 4 mryacci@yahoo.com.ar

RESUMEN

Las necesidades del mercado y la evolución tecnológica ha promovido el desarrollado de nuevas propuestas para el monitoreo y control de procesos con un alto estándar de calidad. Estos procesos se caracterizan por tener gran cantidad de muestras con cero no conformidades. Para el caso particular del control de procesos por atributos, cuando n, p o lambda son muy pequeños, la aproximación normal ya no es adecuada para el control de los procesos. En consecuencia se ha desarrollado nueva metodología para el caso de modelos Binomial y Poisson, cuando los procesos están contaminados por un exceso de ceros. Estos modelos se conocen como modelos ZIB (Zero-Inflated Binomial) y ZIP (Zero-Inflated Poisson). En este trabajo se sintetizan algunos de los procedimientos factibles de ser utilizados en los gráficos de control de atributos basados en estas distribuciones. Para el primer caso, modelos ZIB, se incluyen gráficos de control desarrollados sobre la base de métodos de intervalos de confianza tales como: gráficos np_{ZIB}, gráficos np_J de Jeffreys, gráficos np_w de Wilson, gráficos np_{AC} de Agresti-Coull, gráficos np_{BS} de Blyth-Still y gráficos de control basados en la distribución ZIP truncada (TZIB). Para el segundo caso, modelos ZIP, se presentan los gráficos basados en la distribución Chi-cuadrado: C_{Chi}, C_{CChi} y C_{MChi} y basados en la distribución geométrica: C_g, C_{mg} y C_{me}. En ambas casos se indica la performance de cada uno de ellos cuando los procesos se encuentran dentro o fuera de control.

***PALABRAS CLAVE*:** Distribución Zero Inflated Binomial (ZIB), Distribución Zero Inflated Poisson (ZIP), Control Estadístico de Procesos, Longitud Promedio de Corrida (ARL), Procesos de Alta Calidad.

1. INTRODUCCIÓN

Los avances tecnológicos y la automatización de los procesos de fabricación han generado procesos de alta calidad que se caracterizan por tener muy bajas tasas de ocurrencia de defectos y los tradicionales gráficos de atributos de Shewhart, pueden no ser adecuados para el control de procesos y la toma de decisiones.

Los gráficos *C* y *np* a menudo subestiman la dispersión observada, con lo cual los límites de control pueden resultar inadecuadamente estrechos y posteriormente conducen a una mayor tasa de falsas alarmas en la detección de señales fuera de control. Por lo tanto, la modificación de una distribución Poisson o binomial podría ser utilizada para tratar la sobredispersión.

En la distribución Poisson, por ejemplo la sobredispersión ocurre cuando la varianza es superior a la media en una muestra, y el "Zero Inflated" o "Excesos de Ceros" se observa cuando los datos muestrales presentan una frecuencia más elevada para la ocurrencia de cero eventos que la que se debería esperar si la muestra hubiera sido generada mediante una distribución Poisson o binomial.

La sobredispersión por lo tanto puede ser originada por una gran cantidad de ceros de los cuales alguno de ellos son producidos por eventos extraños al proceso, denominados como "perturbaciones aleatorias". La variabilidad que excede la explicada por los modelos de Poisson o binomial se pueden modelar a través de un efecto aleatorio en un modelo estadístico.

Estos modelos pueden basarse en una distribución denominada Zero Inflated Poisson (ZIP), que es una combinación de la distribución Zero Inflated y la Poisson, y la Zero Inflated binomial (ZIB), que es la combinación de la distribución Zero Inflated y la binomial. Ambos modelos suponen que las perturbaciones aleatorias se producen con cierta probabilidad, y la ocurrencia de tales efectos aleatorios (número de no conformidad en cada subgrupo de muestreo) sigue una distribución Poisson o binomial.

Muchos de los problemas que se presentan tales como la alta probabilidad de falsas alarmas en la detección de puntos fuera de control, la incapacidad de detectar la mejora de procesos, la baja probabilidad de cobertura, entre otros, han sido identificados por varios investigadores y para resolverlos se han desarrollado recientemente, nuevos modelos y técnicas de monitoreo. Algunos de estos nuevos métodos se resumen en este artículo.

Para la distribución binomial, Sim y Lim (2008) compararon los gráficos de Shewhart (gráficos *np*), con gráfico *np* basados en el modelo ZIB (gráficos np_{ZIB}), un gráfico *np* con un intervalo preferente de Jeffreys (gráfico np_J) y un gráfico *np* con un intervalo de Blyth-Still (gráficos np_{BS}). Se comparó la performance de los distintos gráficos utilizando la longitud media de corrida (ARL).

Fatai y otros (2010) investigaron sobre gráfico de control basada en una distribución ZIB truncada (TZIB) aplicando límites de probabilidad en lugar de los límites de control basados en Shewhart para el seguimiento de las observaciones distribuidas como ZIB. Para evaluar la performance del grafico propuesto también utilizaron el enfoque del ARL.

Yawsaeng y Mayureesawan (2012), estudiaron los gráficos de control anteriormente mencionados tales como: gráficos np_{ZIB}, gráficos np_J y además los gráficos np_w de Wilson, gráficos np_{AC} de Agresti-Coull y gráficos np_{BS} de Blyth-Still. La performance de los gráficos de control se comparó en términos de ARL y también en términos de probabilidad promedio de cobertura (ACP), asignando diferentes valores para la proporción de ceros observados (φ), distintos niveles de cambio en el porcentaje de no conformidad (θ) y en los valores para la varianzas (σ^2).

Para los gráficos *C*, Cohen (1991) desarrolló un modelo ZIP en el cual estimó el valor de la media lambda mediante el estimador máximo de verosimilitud (MLE).

Xie y otros, (2001) desarrollaron un Gráfico *C* para el modelo ZIP que se denominó gráficos C_{ZIB}. Estudiaron la eficiencia de éste gráfico de control para detectar desplazamientos del aumento del valor medio en el número de no conformidades en un proceso.

Sim y Lim (2008) propusieron gráficos de control para datos con excesos de ceros, tanto en una distribución de Poisson como en una distribución binomial. Para el caso de la distribución Poisson propusieron un gráfico *C* a partir del intervalo de Jeffreys para detectar aumentos de la media, denominado gráfico C_J. Compararon los gráficos C_J y C_{ZIP} con los tradicionales gráficos *C*.

Peerajit and Mayureesawan (2010) continuaron las investigaciones de Sim y Lim y compararon la performance de los gráfico C_J; C_{ZIB} con el gráfico *C* de Shewhart, en procesos con diferentes proporciones de exceso de ceros, utilizando como indicadores el ARL y ACP.

Katemee y Mayureesawan (2012) propusieron en una de sus investigaciones, gráficos basados en la distribución Chi-cuadrado denominados: C_{Chi}, C_{CChi} y C_{MChi} y en otra, gráficos basados en la distribución geométrica, denominados: C_g, C_{mg} y C_{me}. En ambos casos se compararon con los gráfico *C*; C_J y C_{ZIB} y se indicó la performance de cada uno de ellos cuando los procesos se encontraron dentro o fuera de control.

La Distribución ZIP se aproximó mediante una distribución chi cuadrado no central con parámetro λ_{Chi}. El mejor valor de ajuste λ_{Chi} se utilizó para sustituir la media y la varianza estimada en los límites de control de los gráficos C de Shewhart por tres métodos diferentes. En el grafico C_{Chi}, los valores estimados de la media y la varianza de los gráficos *C* se sustituyeron por $\lambda_{Chi,}$ mientras que en el grafico C_{CChi} fueron con los estimadores de la media y la varianza, respectivamente, de la distribución chi cuadrado no central. En el gráfico C_{MChi}, la media fue reemplazada por la media estimada de la chi cuadrado no central y la varianza por el rango intercuartílico.

En el grafico C_g, los límites de control se construyeron con la media y la varianza de la distribución geométrica. En los gráficos C_{mg}, la media de la distribución geométrica se utilizó para sustituir tanto la media como la varianza en los límites de control. En el grafico C_{me}, se utilizó la mediana y la varianza geométrica en los límites de control.

2. METODOLOGÍA

1. Modelos ZIB

A continuación se resume en el cuadro Nº1 los limites superiores de control para los gráficos *np*, según los diferentes modelos propuestos cuando existe exceso de ceros o "Zero Inflated".

Cuadro Nº1: límite superior de control para gráficos *np* con Zero Inflated.

Gráficos	Límite superior de control	Comentarios
np	$UCL = np + 3\sqrt{np(1-p)}$ (1)	*np* número de unidades no conformes en una muestra de tamaño n
np_{ZIB}	$UCL = n\hat{p} + 3\sqrt{n\hat{p}(1-\hat{p})}$ (2)	$\hat{p} = \frac{[1-(1-\hat{p})^n]\bar{Z}^+}{n}$; $\bar{Z}^+$ es la media de m muestras
np_j	$UCL = \max[\lvert x\rvert p_0 \geq B(\alpha; x+0.5.n-x+0.5]$ (3)	*X* es una VA con distribución binomial con parámetro *n, p* $B(\alpha; a, b)$ es el percentil 100 de una distribución beta con parámetro a y b p_0 es el valor estimado de $\hat{p}$
np_w	$UCL = \max[\lvert x\rvert p_0 \geq w(x)]$ (4)	$w(x) = \frac{x+\frac{z_\alpha^2}{2}}{n+z_\alpha^2} - \frac{z_\alpha\sqrt{n}}{n+z_\alpha^2}\sqrt{\tilde{p}\tilde{q}+z_\alpha^2/4n}$

np_{AC}	$UCL = max[\|X\|p_0 \geq ac(x)]$ (5)	$ac(x) = \tilde{p} - z_\alpha(\tilde{p}\tilde{q})^{1/2}\tilde{n}.^{-1/2}$ $\tilde{p} = \tilde{x}/\tilde{n}$
np_{BS}	$UCL = max[\|X\|p_0 \geq a(x)]$ (6)	$a(x)$ $= \frac{(x-0.5) + 0.5z_\alpha^2 - z_\alpha\sqrt{(x-0.5) - \frac{(x-0.5)^2}{n} + 0.25z_\alpha^2}}{n + z_\alpha^2}$
T_{ZIB}	$\sum_{x=[UPL_{rZIB}]^R}^{n} \theta\binom{n}{x}p^x(1-p)^{n-x} = \frac{\alpha}{2}$ (7)	$\hat{\theta} = \frac{(n-1)(\sum_{i=1}^{m} x_i)^2}{n.m(\sum_{i=1}^{m} x_i{}^2 - \sum_{i=1}^{m} x_i\ )}$ *X* es la variable con distribución binomial $\hat{\theta}$ probabilidad de exceso de ceros

2. Modelos ZIP

En el cuadro Nº2 se resume los limites superiores de control para los gráficos *C* según los diferentes modelos propuestos cuando existe exceso de ceros o "Zero Inflated".

Cuadro Nº2 límite superior de control para gráficos *C* con Zero Inflated.

Gráficos	Límite superior de control	Comentarios
C	$UCL = c + 3\sqrt{c}$ (8)	C cantidad promedio de no conformidades por muestra puede estimarse como la cantidad promedio de no conformidades en una muestra de ítems observados
C_{ZIP}	$UCL = \hat{\lambda} + 3\sqrt{\hat{\lambda}}$ (9) $\hat{\lambda} = \bar{y}^+[1 - e^{\hat{\lambda}}]$	Y variable aleatoria cantidad de no conformidades por ítem λ promedio de no conformidades por ítem $\bar{y}^+$ cantidad promedio de no conformidades por ítems que tienen una cantidad no nula de no conformidades
C_j	$UCL = \max[\|y\|\lambda > G(\alpha; y + 0.5,1)]$ (10)	λ promedio de no conformidades por ítem en la muestra $G(\alpha; a, b)$ es el percentil 100 de una distribución gama con parámetros $a = y + 05$ y $b = 1$
C_{Chi}	$UCL = \hat{\lambda}_{Chi} + 3\sqrt{\hat{\lambda}_{Chi}}$ (11)	$\hat{\lambda}_{Chi} = \sum_{i=1}^{k} (\frac{\mu_i}{\sigma_i})^2$
C_{CChi}	$UCL = E(y) + 3\sqrt{V(y)}$ (12)	$E(y) = \lambda_{Chi}\ \ V(y) = 4\lambda_{Chi}$
C_{MChi}	$UCL = E(y) + 3\sqrt{IQR(c)}$ (13)	$IQR(c) = Q_3 - Q_1$
C_g	$UCL = E(y) + 3\sqrt{V(y)}$ (14)	y es la variable aleatoria del numero de fallas antes de que el primer éxito ocurra.

		p_g es la probabilidad de éxito en cada ensayo $\hat{p}_g = (1+\frac{1}{n}\sum_{i=1}^{n} k_i)^{-1}$ $E(y) = \frac{1-\hat{p}_g}{\hat{p}_g}$ $V(y) = \frac{1-\hat{p}_g}{\hat{p}_g^2}$
C_{mg}	$UCL = E(y) + 3\sqrt{E(y)}$ (15)	$\hat{p}_g = (1+\frac{1}{n}\sum_{i=1}^{n} k_i)^{-1}$ $E(y) = \frac{1-\hat{p}_g}{\hat{p}_g}$
C_{me}	$UCL = M + 3\sqrt{V(y)}$ (16)	$Mediana = \left[\frac{-1}{\log_2(1-\hat{p}_g)}\right] - 1$

3. Comparación de la performance de los gráficos de control

En las distintas investigaciones la performance de los diferentes gráficos de control se compararon utilizando el ARL y /o ACP.

La longitud media de corrida o ARL, es el número promedio de puntos que deben ser trazados hasta encontrar un punto que indique una situación fuera de control. Para procesos bajo control, el método más efectivo es aquel cuya longitud promedio de corrida es lo más grande posible. Si no están correlacionadas las observaciones del proceso, entonces para cualquier gráfico de control, el ARL puede calcularse fácilmente a partir de: ARL = 1/(probabilidad de encontrar un punto fuera de control).

El ACP es la probabilidad promedio de cobertura. Para un proceso dado, el valor de ACP más cercano al coeficiente de confianza (1-α) es el gráfico recomendado.

4. Reglas de control K de K

Algunas investigaciones como las de Sim y Lim (2008) y Yawsaeng y Mayureesawan (2012), incorporan las reglas K de K para indicar un proceso fuera de control, asignando valores a K que van de 1 a 5. Indica que cuando K observaciones consecutivas superan el límite de control superior, es señal que el proceso se encuentra fuera de control. Así una regla de control “dos de dos”, indica que, un proceso es declarado fuera de control si dos puntos sucesivos caen fuera de los límites de control.

3. RESULTADOS Y DISCUSIÓN

En el caso de la variable binomial el estudio de Sim y Lim (2008) reveló que los gráfico np_J (Eq.3) y np_{BS} (Eq.6) tienen un valor de ARL mucho más pequeño que el valor deseado debido a la baja probabilidad de cobertura del límite de control. Al incorporar una simple regla de control "dos de dos" funcionan mejor que los gráficos np_J y np_{ZIB} en todas las situaciones simuladas.

Fatai y otros (2010) mostraron que los resultados de la aplicación de limites de probabilidad para los gráficos de control a partir de una distribución ZIB (Eq.7) truncada no pueden generalizarse ya que funcionan bastante bien para los gráficos *p* pero no para los gráficos *np*, al menos no resultó adecuado para el caso particular presentado en la investigación.

Yawsaeng y Mayureesawan (2012) compararon el desempeño de los gráficos de control mediante simulación con diferentes proporciones observadas de cero (φ) y niveles de cambio en las proporciones de no conformidad (θ), para diferentes valores de σ^2. Además utilizaron las reglas K de K asignando diferentes valores a K.

Para procesos bajo control los gráficos np_w (Eq.4) fueron preferibles para valores bajos de K y alta proporción de ceros (φ= 0,9), mientras que los gráficos np_J (Eq.3) se comportaron mejor que los otros gráficos para K = 1, 3 y 5 y proporción de ceros relativamente baja (φ = 0,3) para los distintos valores de la varianza. El gráfico np_{ZIB} (Eq.2) pareció funcionar mejor que los otros, para K = 2 y 4 y proporción de ceros bajos (φ = 0.3 y 0.4). Cabe señalar que el gráfico *np* (Eq.1) resultó más adecuado para altos valores de K (K = 3, 4, 5) y alta proporción de ceros (φ = 0,8-0,9) en todos los niveles de varianza.

En el caso de procesos fuera de control, en situaciones en las que se observaron una baja proporción de ceros (φ = 0,3) y una baja varianza (σ^2=2), lo recomendable seria usar el gráfico np_J con la aplicación de la norma de control K de K cuando K es 2 o mayor. Sin embargo, si la proporción de ceros es extremadamente alta (0,9), se sugiere el gráfico *np* con la aplicación de la regla de control K de K si K = 3 o superior. Cabe señalar que para proporciones de ceros moderadamente altas (0,4-0,8), los gráficos de control estudiados no funcionan bien. Por lo tanto, este problema podría ser un tema interesante para su estudio.

Para el caso de la variable Poisson, Xie y otros (2001), llegaron a la conclusión que el ARL es relativamente sensible a incrementos de la proporción de ítems no conformes. A medida

que ésta proporción aumenta el ARL disminuye rápidamente, pero la misma situación no se da a partir de ciertos valores de la media y esto es un problema del método, debido a que no permite detectar rápidamente el incremento en los valores promedios. Por lo tanto se requieren de otros métodos para detectar esta situación.

Sim y Lim (2008), concluyeron que los gráficos C_J (Eq.10) son apropiados cuando la media está bajo control, pero en el caso que la media esté fuera de control los gráficos C (Eq.8) presentan mejor performance que los otros gráficos pero con una menor probabilidad de cobertura.

En los estudios de Peerajit and Mayureesawan (2010), los resultados revelaron que el gráfico C_J tiene una eficacia más alta que los otros gráficos de control cuando el proceso está bajo control y cuando está fuera, el gráfico C es el más eficaz para detectar los cambios. Pero al igual que Sim y Lim, concluyeron que el ACP del gráfico C fue demasiado bajo para ser un buen estimador de los parámetros del proceso.

En la investigación desarrollada por Katemee y Mayureesawan, (2012a) en la cual propusieron gráficos de control C_g (Eq.14), C_{mg} (Eq.15) y C_{me} (Eq.16), modelando el número de no conformidades con una distribución geométrica, utilizaron la media, la mediana y la varianza de esta distribución para construir los limites de control. A partir de las simulaciones llevadas a cabo efectuaron comparaciones con los gráficos C, C_{ZIP} (Eq.9) y C_J e indicaron que: para un proceso bajo control, el grafico C_g fue superior para valores bajo de la media y diferentes proporciones de ceros, mientras que para un proceso fuera de control, los gráficos c_{mg} y C_{me} fueron los mejores para valores de la media de 2, 3 y 4 en todos los niveles de los otros parámetros simulados.

Otra de las investigaciones de Katemee y Mayureesawan, (2012b) en la que analizaron los gráficos de control basados en la distribución chi cuadrado, concluyeron que para un proceso bajo control, el gráfico C_{CChi} (Eq.12) es superior a todos los otros gráficos considerados con una baja proporción de cero no conformidades. Para un proceso fuera de control, el gráfico c_{MChi} (Eq.13) se comportó mejor que los otros gráficos, para valores bajos de la proporción de cero no conformidades para diferentes medias y valores de p. Sin embargo, para valores altos de la proporción de cero no conformidades, el grafico C_{Chi} (Eq.11) resultó mejor que los otros gráficos considerados.

4. CONCLUSION

Al monitorear procesos de alta calidad frecuentemente se observan excesos de ceros que producen sobredispersión. Estos procesos suelen ajustarse con modelos basados en las distribuciones binomial o Poisson conocidas como distribuciones ZIB y ZIP. En este trabajo se ha realizado una breve reseña de los últimos avances en materia de investigación de los procedimientos estadísticos adecuados para la implementación de los correspondientes gráficos de control, resaltando sus alcances y limitaciones.

En todos los casos la performance de los diferentes gráficos ha sido evaluada mediante ARL o ACP o ambos para seleccionar el grafico más adecuado para cada situación particular. No obstante las respuestas son muy variables debido a que dependen de la magnitud de la proporción de ceros (sobredispersión), los cambios en la media o en la proporción de no conformidades y del estado en el que se encuentra el proceso (bajo o fuera de control), lo que conlleva a ser precavido a la hora de seleccionar un procedimiento de control.

5. REFERENCIAS

- COHEN, A.C. (1991). *Truncated and Censored Samples: Theory and Applications*. New York, Marcel Dekker, 1.
- FATAHI, A.A; NOOROSANA, R; DOKOUHAKI, P. y BABAKHANI, M. (2010). Truncated Zero Inflated Binomial Control Charts for monitoring rare health events. *International Journal of Research and Reviews in Applied Sciences* (IJRRAS). 4. 4. 380-387.
- KATEMEE, N y MAYUREESAWAN, T (2012a). Nonconforming Control Charts for Zero Inflated Poisson distribution. *World Academy of Science, Enginiering and Technology*. 69. 149-155.
- KATEMEE, N y MAYUREESAWAN, T (2012b). Control Charts for Zero-Inflated Poisson Models. *Applied Mathematical Sciences*. 6. 56. 2791 – 2803.
- PEERAJIT, V. and MAYUREESAWAN, T. (2010). Nonconforming Control Charts for Zero - Inflated Processes, In *Proceeding of 11th Conf. on Statistic and Applied Statistic*, Holiday In, Chiangmai, Thailand. 76. 61-73

- SIM, C.H. and LIM, M.H. (2008). Attribute Charts for Zero – Inflated Processes. *Journal of Communications in Statistics-Simulation and Computation.* 37. 1440 - 1452.
- XIE, M. HE,B. and GOH, T. N. (2001). Zero-inflated Poisson model in Statistical Process Control. *Computational Statistics & Data Analysis*. 38. 191 - 201.
- YAWSAENG, B. and MAYUREESAWAN, T. (2012) Control Charts for Zero-Inflated Binomial Models. *Thailand Statistician*. 10(1). 107-120.