Revista Colombiana de Estadística
Volumen 30 No. 1. pp. 69 a 75. Junio 2007

Estimación por intervalo del parámetro de la distribución de Poisson con una sola observación

Interval Estimation for the Poisson Distribution Parameter with a Single Observation

Juan Carlos Correa[a]

Escuela de Estadística, Universidad Nacional de Colombia, Medellín

Resumen

La estimación del parámetro de la distribución de Poisson, digamos λ, es un problema importante en el trabajo estadístico aplicado. En muchas ocasiones solo disponemos de un único dato para construir un intervalo de confianza. Se muestra cuándo se pueden construir intervalos de confianza basados en el teorema central del límite, el método exacto y la razón de verosimilitud cuando se tiene una sola observación. Se ilustra este caso construyendo un intervalo para la tasa de suicidios en Colombia.

***Palabras clave*:** estimación, intervalo de confianza, tamaño de muestra pequeño, teorema central del límite, razón de verosimilitud.

Abstract

The estimation of the parameter of the Poisson distribution, say λ, is an important task in applied statistics. Frequently we only have available a single observation and our goal is to construct a confidence interval. We illustrate under what conditions we can construct a confidence interval based on three methods: central limit theorem, exact method, and the likelihood ratio method. We also illustrate this problem constructing a confidence interval for the rate of suicides in Colombia.

***Key words*:** Estimation, Confidence interval, Small sample size, Central limit theorem, Likelihood ratio.

1. Introducción

La distribución de Poisson juega un papel de fundamental importancia en el trabajo aplicado para modelar problemas de conteo en muchas áreas.

[a]Profesor asociado. E-mail: jccorrea@unalmed.edu.co

Asumamos que X_1 es una observación de una distribución de Poisson con función de masa dada por:

$$P_X(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

para $\lambda > 0$ y $x = 0, 1, 2, \ldots$, la media de la distribución es λ y su varianza $\sigma^2 = \lambda$.

Sea $Y_1, Y_2, \ldots, Y_N$ una muestra aleatoria de tamaño n, el estimador de máxima verosimilitud para λ es $\widehat{\lambda} = \overline{Y} = 1/N \sum_{i=1}^{N} Y_i$, donde $\sum_{i=1}^{N} Y_i$ es suficiente minimal para λ.

Para poder usar el teorema del límite central (TLC) debemos tener una muestra aleatoria de tamaño N, grande, de una población con varianza $\sigma^2 < \infty$ y media μ_Y. Entonces

$$\frac{\overline{Y} - \mu_Y}{\sigma/\sqrt{N}} \longrightarrow N(0, 1)$$

cuando $N \to \infty$.

En el caso de X_1, aparentemente solo tenemos una única observación, pero si representamos X_1 como

$$X_1 = \sum_{i=1}^{N} Y_i$$

donde las Y_i's son i.i.d. (independientes e idénticamente distribuidas), y si N es grande, entonces podríamos aplicar el TLC. Esta descomposición es posible y justificada por el teorema de Skorohod (Billingsley 1986). En el caso Poisson $Y_1 \sim Poisson(\lambda^* = \lambda/N)$.

El intervalo de confianza para λ^* basado en el TLC es:

$$\left(\widehat{\lambda}^* - z_{\alpha/2}\sqrt{\frac{\widehat{\lambda}^*}{N}}, \widehat{\lambda}^* + z_{\alpha/2}\sqrt{\frac{\widehat{\lambda}^*}{N}} \right)$$

donde $\widehat{\lambda}^* = \overline{Y}$ y $z_{\alpha/2}$ es el percentil $\alpha/2$ superior de la normal estándar. Ya que el interés no es λ^* sino λ, lo único que debemos hacer es multiplicar por N ambos límites del intervalo anterior. Con esto llegamos a que el intervalo de confianza basado en el TLC para λ con una sola observación es

$$\left(\widehat{\lambda} - z_{\alpha/2}\sqrt{\widehat{\lambda}}, \widehat{\lambda} + z_{\alpha/2}\sqrt{\widehat{\lambda}} \right)$$

donde $\widehat{\lambda} = X_1$. Note que el resultado final no depende de N y depende exclusivamente del valor del estadístico suficiente. Esto es claro en el caso donde alguien decide observar un conteo por unidad de tiempo en minutos en lugar de horas: una observación de una hora equivale a 60 observaciones de un minuto. No se tiene más información al hacerlo por minutos.

Uno de los problemas que se enfrenta con aproximaciones de este tipo es saber cuándo N es lo suficientemente grande. Para ello recurrimos al teorema de Berry-Esséen (Serfling 1980, p. 33; Lehmann 1999, p. 78): Si $Y_1, \ldots, Y_N$ son variables

aleatorias i.i.d. con media μ y varianza $\sigma^2 > 0$ y tercer momento central finito, y si $G_N(t) = P\left(S_N \leq t\right)$, donde

$$S_N = \frac{\overline{Y} - \mu}{\sigma/\sqrt{N}}$$

entonces

$$\sup_t |G_N(t) - \Phi(t)| \leq \frac{C}{\sqrt{N}} \frac{|Y_1 - \mu|^3}{\sigma^3}, \qquad \forall\ N$$

$\Phi(z)$ es la función de distribución acumulada de una variable aleatoria normal estándar. La determinación de la constante óptima C ha sido motivo de una intensa investigación; se sabe que existe la constante pero no se conoce, y se ha logrado reducir hasta $C = 0.7975$ (Lehmann 1999). Serfling (1980) en la presentación de este resultado tiene $C = 33/4$.

En el caso de la distribución Poisson tenemos que $\mu = \sigma^2 = E(Y_1 - \lambda)^3 = \lambda$. Por lo tanto la cota se reduce a $0.7975/\sqrt{N\lambda}$.

Tabla 1: Tamaños muestrales mínimos para estimar el parámetro de la Poisson usando el teorema de Berry-Essén.

	Error máximo			
λ	0.1	0.05	0.01	0.005
0.010	6361	25441	636007	2544025
0.025	2545	10177	254403	1017610
0.050	1273	5089	127202	508805
0.750	85	340	8481	33921
1.000	64	255	6361	25441
2.000	32	128	3181	12721
3.000	22	85	2121	8481
4.000	16	64	1591	6361
5.000	13	51	1273	5089
10.000	7	26	637	2545
20.000	4	13	319	1273
30.000	3	9	213	849
40.000	2	7	160	637
50.000	2	6	128	509
60.000	2	5	107	425
70.000	1	4	91	364
80.000	1	4	80	319
90.000	1	3	71	283
100.000	1	3	64	255
200.000	1	2	32	128
500.000	1	1	13	51

La tabla 1 puede leerse así: si permitimos un error máximo en la aproximación a la normal de 0.1 (diferencia entre la distribución acumulada real y la normal estándar acumulada), dado un λ específico, por ejemplo 1.0, la muestra mínima en este caso es 64. Si la diferencia máxima permitida la rebajamos a 0.05, el tamaño muestral mínimo se incrementa a 255. De la tabla 1 se observa que para valores muy grandes de λ solo es necesaria una observación, lo cual nos garantiza

que la aproximación usando la distribución normal para la media muestral es lo suficientemente buena.

2. Otros intervalos

2.1. Método exacto

Un intervalo de confianza exacto para λ, en el caso de una sola observación, se obtiene resolviendo las siguientes ecuaciones para λ_L y λ_U:

$$\exp\left(-\lambda_L\right)\sum_{i=0}^{X_1}\frac{\left(\lambda_L\right)^i}{i!}=1-\frac{\alpha}{2}$$

y

$$\exp\left(-\lambda_U\right)\sum_{i=0}^{X_1}\frac{\left(\lambda_U\right)^i}{i!}=\frac{\alpha}{2}$$

Observe que la solución existe sin importar que X_1 sea discreta, ya que λ_L y λ_U toman valores en $(0,\infty)$.

2.2. Intervalos basados en la razón de verosimilitud relativa

Kalbfleish (1985) presenta la metodología para construir intervalos de verosimilitud. Si $L(\mu)$ es una función de verosimilitud, se define la *función de verosimilitud relativa* como

$$R(\lambda)=\frac{L(\lambda)}{L\big(\widehat{\lambda}\big)}$$

El conjunto de valores de λ para los cuales $R(\lambda)\geq p$ es llamado el *intervalo de p100% de verosimilitud* para λ. Los intervalos del 14.7% y del 3.6% de verosimilitud corresponden a intervalos de confianza aproximadamente de niveles del 95% y del 99%, respectivamente.

Lo que se debe hacer entonces es hallar las raíces que nos dan los límites del intervalo. Para el caso del parámetro de la Poisson λ, tenemos que un intervalo de confianza del 95% se halla encontrando el par de raíces tal que

$$R(\lambda)=\frac{L(\lambda)}{L\big(\widehat{\lambda}\big)}=\left(\frac{\lambda}{X_1}\right)^{X_1}\exp\Big(-\Big(\lambda-\widehat{\lambda}\Big)\Big)\geq 0.147$$

donde $\widehat{\lambda}=X_1$. Esto se resuelve numéricamente. Las raíces existen dada la log-concavidad de la función de verosimilitud, asumiendo que el estadístico suficiente sea mayor que cero.

2.3. Método basado en la máxima verosimilitud

Se sabe que si $\widehat{\theta}$ es el estimador máximo verosímil para θ (el cual puede ser un vector), bajo ciertas condiciones suaves (Serfling 1980), entonces $\widehat{\theta} \sim N\left(\theta, I^{-1}(\theta)\right)$, con $I(\theta)$ siendo la matriz de información de Fisher. Entonces, en el caso Poisson

$$\left(\overline{X} - z_{\alpha/2}\frac{\sqrt{\overline{X}}}{\sqrt{n}},\ \overline{X} + z_{\alpha/2}\frac{\sqrt{\overline{X}}}{\sqrt{n}}\right)$$

en el caso de una observación se tiene

$$\left(X_1 - z_{\alpha/2}\sqrt{X_1},\ X_1 + z_{\alpha/2}\sqrt{X_1}\right)$$

Este método produce el mismo resultado que el basado en el TLC, ya que $\widehat{\lambda} = X_1$.

3. Resultados de simulación

Se realizó una simulación para comparar tanto la longitud de los intervalos como el nivel de confianza real (el porcentaje de veces que el intervalo cubre el parámetro) alcanzado por los tres métodos considerados cuando la muestra es de tamaño uno. El nivel de confianza nominal o teórico fue del 95%. La tabla 2 presenta algunos estadísticos de la distribución de la amplitud de los intervalos como son el percentil 5%, la mediana, la amplitud media y el percentil 95%. Esto nos da una idea de la dispersión de las amplitudes. La última columna hace referencia al nivel real de confianza logrado. Para diferentes valores de λ se generaron 1000 muestras de tamaño uno. A cada muestra se le aplicó cada uno de los métodos para construir los intervalos.

Los tres métodos producen intervalos con niveles reales cercanos al nivel nominal; sin embargo, el método exacto tiende a producir intervalos con amplitudes mayores que los otros dos métodos, los cuales producen resultados bastante similares.

4. Ilustración

El número de suicidios en Colombia fue 1786 casos en el año 2005 (Sarmiento 2007). Asumiendo que el número de suicidios en un año puede distribuirse Poisson, y dado que solo tenemos un dato, aplicamos el método anterior el cual nos lleva a concluir que el número esperado de suicidios está en el intervalo (1703.16, 1868.83) a un nivel de confianza del 95%. Este intervalo se construyó utilizando el método basado en la máxima verosimilitud. Si la población a mitad de año era de 45795000 habitantes, entonces la tasa de suicidios por cada 100000 habitantes puede estimarse entre 3.7191 y 4.0808 con una confianza del 95%.

5. Conclusiones

Bajo ciertas condiciones es posible construir intervalos de confianza a partir de muestras de tamaño uno, lo cual es ilustrativo en los cursos básicos de estadística donde una inquietud general por parte de los estudiantes es determinar un n mínimo. El resultado es además útil para epidemiólogos y demógrafos, para quienes

Tabla 2: Longitud y nivel de confianza real de los tres tipos de intervalos: TLC, exacto y razón de verosimilitud al 95% de confianza nominal.

	Longitud del intervalo				
$\lambda = 10$	Perc. 0.05	Mediana	Media	Perc. 0.95	Nivel real
TLC	8.7654	12.396	12.266	15.680	0.9278
Exacto	10.0450	13.595	13.478	16.838	0.9752
R.V.	8.9516	12.526	12.407	15.781	0.9444
$\lambda = 20$					
TLC	14.134	17.531	17.411	20.743	0.9416
Exacto	15.308	18.672	18.555	21.862	0.9506
R.V.	14.247	17.621	17.503	20.817	0.9506
$\lambda = 50$					
TLC	24.165	27.719	27.599	30.866	0.9474
Exacto	25.267	28.808	28.689	31.946	0.9500
R.V.	24.227	27.772	27.653	30.913	0.9500
$\lambda = 75$					
TLC	30.616	33.948	33.883	36.981	0.9508
Exacto	31.697	35.021	34.956	38.048	0.9508
R.V.	30.663	33.989	33.925	37.018	0.9508
$\lambda = 100$					
TLC	35.927	39.200	39.150	42.220	0.9476
Exacto	36.996	40.263	40.213	43.278	0.9566
R.V.	35.966	39.234	39.184	42.250	0.9500
$\lambda = 150$					
TLC	44.695	48.010	47.980	51.111	0.9502
Exacto	45.750	49.061	49.031	52.158	0.9498
R.V.	44.722	48.034	48.004	51.132	0.9446
$\lambda = 200$					
TLC	52.152	55.437	55.402	58.538	0.9472
Exacto	53.199	56.481	56.446	59.580	0.9556
R.V.	52.173	55.455	55.420	58.554	0.9556
$\lambda = 500$					
TLC	84.440	87.654	87.631	90.840	0.9480
Exacto	85.467	88.680	88.658	91.865	0.9508
R.V.	84.441	87.654	87.632	90.838	0.9508
$\lambda = 1000$					
TLC	120.630	123.900	123.920	127.140	0.9480
Exacto	121.650	124.920	124.940	128.160	0.9488
R.V.	120.620	123.890	123.910	127.130	0.9476

Perc.: Percentil
TLC: Método basado en el TLC
Exacto: Método exacto
R.V.: Método basado en la razón de verosimilitudes

no es inusual obtener al final de un período una única cifra sobre las ocurrencias de eventos de interés.

Una diferencia que cabe anotar entre el resultado del teorema de Berry-Essen y los resultados obtenidos en la simulación es que el teorema de Berry-Essen presenta una cota uniforme para la diferencia entre la distribución de la media muestral y la distribución normal, mientras que en el caso tradicional la construcción del intervalo es más importante que la aproximación de las distribuciones en las colas.

Agradecimientos

Al profesor Francisco Díaz, quien leyó cuidadosamente este documento y sugirió correcciones que resultaron en una mejora sustancial.

Recibido: octubre de 2006
Aceptado: marzo de 2007

Referencias

Billingsley, P. (1986), *Probability and Measure*, 2nd edn, John Wiley & Sons, New York.

Kalbfleish, J. G. (1985), *Probability and Statistical Inference*, Vol. 2, 2nd edn, Springer-Verlag, New York.

Lehmann, E. L. (1999), *Elements of Large-Sample Theory*, Springer-Verlag, New York.

Sarmiento, L. (2007), Jóvenes: ¿Por qué se suicidan?, Web, Red de Prensa No Alineados.
*http://www.voltairenet.org/image/article139303.html#article139303

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley, New York.