Publicado en la Rev. Fac. Nac. Salud Pública 1999; 17(1): 30-36

# Estimación de los parámetros de las distribuciones Bernoulli y Poisson bajo cero eventos

*Estimating parameters of the Bernoulli and Poisson distributions for zero events*

***Juan Carlos Correa M.***[1]
***Esperanza Sierra L.***[1]

## Resumen

Cuando los parámetros $p$, en la distribución de Bernoulli y $\lambda$, en la de Poisson, son muy pequeños, su estimación es un problema díficil porque, a menudo, en muestras aleatorias no se presenta el caso de interés. En estudios epidemiológicos sobre enfermedades poco comunes es frecuente encontrar este tipo de resultados. Se presentan algunas soluciones a este problema.

## Palabras clave

Intervalo de confiaza, distribución de Poisson, distribución de Bernoulli, muestras con cero eventos.

## Abstract

When the parameters p, in the Bernoulli distribution, and $\lambda$, in the Poisson distribution, are very small their estimation presents difficulties when random samples shows no events or successes. In epidemiological studies of rare diseases it is common to obtain these kinds of results. Several solutions to this problem are presented.

## Keywords

Confidence interval, Poisson distribution, Bernoulli distribution, zero-events sample, zero-success sample.

[1] Profesores, Departamento de Matemáticas. Universidad Nacional, Medellín, Colombia. E-mail: jccorrea@perseus.unalmed.edu.co. esierra@perseus.unalmed.edu.co

# Introducción

Un problema frecuente en estudios epidemiológicos es la estimación de $p$, la proporción de sujetos que en una población tienen una característica dada. El problema se complica cuando en la muestra que se tiene no se observa un solo caso favorable, situación que se presenta cuando la característica de interés es muy rara, esto es, $p$ es muy pequeña.

Sea $x_1, x_2, \ldots x_n$ una muestra aleatoria de una distribución de Bernoulli con parámetro desconocido $p$. Se quiere estimar $p$ cuando las observaciones son $x_1=0, x_2=0,\ldots x_n=0$. Se sabe que el estimador máximo verosímil de $p$ es $\hat{p} = \sum_{i=1}^{n} x_i$ ; en este caso $\hat{p} = 0$.

Por ejemplo, se toma una muestra de *n* personas de una comunidad para determinar *p,* la proporción de personas que tienen una enfermedad poco común, y ninguna de ellas tienen tal enfermedad. Puede entonces concluirse, sin ninguna medida de confiabilidad, que en la comunidad no existe la enfermedad, ó puede determinarse con un nivel de confianza dado, el mayor valor que *p* podría tomar, sabiendo que en la muestra no se presentó ningun enfermo.

Otro caso similar se presenta cuando se toma una muestra aleatoria para estudiar un evento, que en un tiempo dado ocurre pocas veces, esto es, un evento con una tasa de presentación $\lambda$ muy baja. Es común que el número de eventos en la muestra sea cero. Por ejemplo, en una unidad de salud, en una muestra de *n* semanas no se encontró ninguna mujer embarazada portadora del VIH. Se puede concluir que semanalmente no consulta ninguna embarazada portadora de VIH; ó encontrar con un nivel de confianza dado, el mayor valor posible de $\lambda$ -el número promedio de embarazadas portadoras que semanalmente consultrían- sabiendo que en la muestra estudiada hay cero casos.

Se a $y_1, y_2, \ldots y_n$ una muestra aleatoria de una distribución de Poisson con parámetro desconocido $\lambda$, se va a estimar este parámetro cuando los datos observados son $y_1 = 0, y_2 = 0, \ldots, y_n = 0$. Como el estimador máximo verosímil de $\lambda$ es $\hat{\lambda} = \sum_{i=1}^{n} y_i$ entonces el estimador de $\lambda$ es 0.

En los dos casos la estimación máximo verosímil del parámetro es cero, y esto impide construir los intervalos de confianza con base en los métodos usuales, porque la varianza estimada de dichos estimadores es cero.

Como $p$ y $\lambda$ son parámetros positivos, que a priori se suponen pequeños, para estimar estos parámetros se construirán, intervalos unilaterales superiores $(0, Ls)$ con $(1-\alpha)100\%$ de confianza. Es decir, con una confianza de $(1-\alpha)100\%$ se determinará *Ls*, el valor más

grande que podría tomar el parámetro. Se presentan algunos métodos para calcular la cota $Ls$, para el parámetro de cada distribución.

# Métodos para la construcción de los intervalos de confianza

## 1. Estimación del parámetro $p$ de la distribución de Bernoulli

Jovanovic y Levy (1997) deducen 'la regla del tres', una regla fácil y rápida para determinar la cota superior *Ls* del intervalo unilateral de $(1-\alpha)100\%$ de confianza para $p$; y encuentran que $Ls = -\ln(\alpha)/n$. Si el nivel de confianza es del 95%, se obtiene $Ls = 3/n$.

Haciendo uso de la estadística Bayesiana se puede modificar esta regla. Suponiendo que $p$ tiene una distribución a priori $Beta\ (1, b)$, donde $b \geq 1$ se escoge de acuerdo al conocimiento previo que se tenga sobre *p*; se puede establecer la 'regla del tres bayesiana': $Ls = 3/(n+b)$.

Para $b = 1$ se tiene la mayor cota superior, pero si se han realizado *k* estudios anteriores, bajo las mismas circunstancias del estudio actual, en los cuales no se presentó ningún caso favorable, se puede tomar $b = \sum_{i=1}^{k} n_i + 1$, donde $n_1, n_2, ....., n_k$ son los tamaños de las muestras de los *k* estudios anteriores. [1]

La tabla 1 muestra las cotas superiores para *p,* que se obtienen al aplicar las reglas descritas anteriormemte con distintos tamaños de muestra.

**Tabla 1.** Estimación del límite superior del intervalo unilateral del 95% de confianza $(0, Ls)$ para $p$ con distintos valores de $n$.

| **Tamaño de muestra** | **Regla del tres** | **Regla del tres Bayesiana con b=1** | **Regla del tres Bayesiana con b=25** |
|---|---|---|---|
| n | 3/n | 3/n | 3/(n+25) |
| 10 | 0,3000 | 0,3000 | 0,0857 |
| 20 | 0,1500 | 0,1500 | 0,0667 |
| 50 | 0,0600 | 0,0600 | 0,0400 |
| 100 | 0,0300 | 0,0300 | 0,024 |
| 500 | 0,0060 | 0,0060 | 0,0057 |

La última columna se calculó suponiendo que en un estudio anterior, con una muestra de tamaño 24, tampoco se observó ningún caso favorable.

## 2. Estimación del parámetro $\lambda$ de la distribución de Poisson

### 2.1 Intervalo de máxima verosímilitud

La probabilidad de observar una muestra $0,0,\cdots,0$ de tamaño $n$ de una distribución de Poisson con parámetro $\lambda$ es $P(y_1 = 0, y_2 = 0, \ldots, y_n = 0 / \lambda) = e^{-n\lambda}$

Hallando $\lambda$ tal que $e^{-n\lambda} \geq \alpha$, se puede establecer un límite superior, *Ls* para $\lambda$ de tal manera que $(1-\alpha)100\%$ de los $\lambda$ que puedan generar esa muestra sean a lo sumo iguales a $Ls$, $Ls = -\ln(\alpha)/n$. Si $\alpha = 0{,}05$ entonces este límite superior es $3/n$.

Por ejemplo, si en una muestra de 20 semanas no se presentó ninguna paciente portadora del VIH, se puede estimar que la tasa semanal de portadoras es a lo sumo igual a 0,15 casos por semana.

### 2.2 Usando un factor de corrección

Se sabe que $\hat{\lambda} + 1/a_n$ se distribuye asintóticamente normal, con media $\lambda + 1/a_n$ y varianza $\lambda/n + 1/a_n$ donde $\hat{\lambda}$ es el estimador de máxima verosimilitud de $\lambda$ y $\{1/a_n\}$ es una sucesión de constantes que cumplen ciertas condiciones y que converge a cero[2]. La sucesión $\{1/n^m\}$ con $m>1$ cumple estas condiciones.

Usando la distribución asintótica y tomando $\hat{\lambda}=0$, se construye el intervalo unilateral del $(1-\alpha)100\%$ de confianza:

$\left(0, -1/n^m + z_\alpha \sqrt{1/n^m}\right)$ donde $z_\alpha$ es el percentil $(1-\alpha)100$ de la normal estándar.

En una simulación[3] con $m$ = 1,1; 1,5; 2,0 y 3,0 y $\alpha = 0{,}05$ para distintos valores de $\lambda$ y $n$ se observó que por este método los mejores intervalos se obtienen cuando $m$=1,5.

Por ejemplo, si en una muestra de 20 semanas no se presentó ninguna paciente portadora del VIH, se puede estimar que la tasa semanal de portadoras es a lo sumo igual a 0.16276 casos por semana.

### 2.3 Método bayesiano

Con miras a obtener las estimaciones, el método bayesiano permite considerar la información previa que se tenga sobre el parámetro; esto lo hace muy atractivo para los investigadores.

Este método requiere, para estimar un parámetro $\theta$, que antes de tomar la muestra, $y_1$, $y_2$, ...$y_n$ se asigne al parámetro una distribución "a priori" $\xi(\theta)$ y luego de recoger la muestra, tal distribución se actualice mediante el teorema de Bayes; esta distribución actualizada, $\xi(\theta/y)$, recibe el nombre de "a posteriori", y se sabe que, $\xi(\theta/y)$ es proporcional a $f(y/\theta)\,\xi(\theta)$ siendo $f(y/\theta)$ la distribución que genera la muestra que se tiene, si el parámetro fuera $\theta$.

Para determinar la distribución "a posteriori" se pueden usar las familias conjugadas de distribuciones. Una familia de distribuciones es conjugada si tanto la distribución a priori como la a posteriori pertenecen ella[4].

Para la distribución de Poisson la familia conjugada es la familia de distribuciones *Gamma* que por ser tan amplia permite representar el conocimiento previo que se tiene sobre la tasa $\lambda$, a la que ocurre el evento.

Tomando como distribución "a priori" de $\lambda$ una $Gamma(\alpha,\beta)$ entonces la distribución "a posteriori" será una $Gamma\left(\alpha+\sum_{i=1}^{n} y_i, \beta+n\right)$, con media $\mu=\left(\alpha+\sum_{i=1}^{n} y_i\right)\Big/\left(\beta+n\right)$. Se puede usar $\mu$ como un estimador de $\lambda$ escogiendo valores apropiados para $\alpha$ y $\beta$. En el caso de la muestra con ningún evento, $\xi(\lambda/y)$ es una $Gamma(\alpha,\beta+n)$ y la estimación sería $\hat{\lambda}=\alpha/(\beta+n)$.

Teniendo en cuenta que la tasa a la que ocurre el evento es muy baja y la distribución *Gamma* es sesgada a la derecha, se puede escoger los valores de $\alpha$ y $\beta$ iguales y pequeños; de esta forma los $\lambda$ tendrían media uno y una varianza grande.

El estudio de simulación[3] citado anteriormente muestra que si el parámetro verdadero es muy pequeño, tomando valores de $\alpha$ pequeños, como 0,5 ó 0,25, se obtienen intervalos de confianza más estrechos, aún en caso de muestras pequeñas.

Para hallar el intervalo de confianza se encuentra $Ls$ tal que: $P\left((0\le\lambda\le Ls)/y\right)=1-\alpha$

El intervalo de probabilidad obtenido $(0, Ls)$ sirve como una buena aproximación aunque estrictamente no es un intervalo de confianza.

## 2.4 Resultados númericos

Con los metódos descritos en las secciones 2.1 a 2.3 se calcularon los intervalos unilaterales superiores del 95% de confianza $(0, Ls)$ para algunos tamaños de muestra. En la tabla 2 se muestran esos valores de *Ls*.

En el anexo figura el programa en SAS que se utilizó para resolver la ecuación planteada en 2.3. que permite calcular la última columna de la tabla 2.

**Tabla 2.** Estimación del límite superior del intervalo unilateral del 95% de confianza $(0\,,\,Ls)$ para $\lambda$ con distintos valores de $n$.

| **Tamaño** $n$ | **Máxima Verosimilitud** $3/n$ | **Factor de Corrección** $\mathbf{1/n^{1,5}}$ | **Intervalo Bayesiano** $\alpha = \beta = 0{,}5$ |
|---|---|---|---|
| 10 | 0.3000 | 0.26090 | 0.18293 |
| 20 | 0.1500 | 0.16276 | 0.09369 |
| 50 | 0.0600 | 0.08466 | 0.03803 |
| 100 | 0.0300 | 0.05102 | 0.01911 |
| 500 | 0.0060 | 0.01547 | 0.00384 |

# Conclusiones

Cuando se quiere estimar $p$, en el caso de una muestra que no presentó ningún caso favorable todavía es posible determinar, con un nivel de confianza dado, cual sería el mayor valor de $p$ que podría esperarse.

Los resultados mostrados en la tabla 1 permiten concluir que los intervalos construídos con la regla del tres bayesiana son mejores y si se tienen $k$ estudios anteriores similares al actual, donde las muestras no presentaron casos favorables es recomendable tomar $b = \sum_{i=1}^{k} n_i + 1$, especialmente si la muestra es pequeña. Para muestras grandes es suficiente tomar $b = 1$.

En el caso de la distribución de Poisson con parámetro $\lambda$, cuando la muestra presenta cero eventos también es posible estimar, dado un nivel de confianza, el mayor valor que $\lambda$ podría alcanzar.

Los resultados que se muestran en la tabla 2 permiten concluir que los intervalos calculados con el método bayesiano resultan ser los mejores, cuando se elijen adecuadamente los parámetros de la distribución a priori. Una buena elección es tomar $\alpha = \beta = 0{,}5$, aunque si se tiene información pasada sobre el proceso, se puede refinar esta elección teniendo en cuenta que mientras más pequeña sea $\lambda$, más pequeños deben ser los parámetros escogidos.

# Anexo

Programa SAS que calcula, mediante el método bayesiano, el límite superior del intervalo unilateral del 95% de confianza para $\lambda$ —el parámetro de la Poisson— cuando en una muestra de tamaño 500 hay cero eventos y con $\alpha = \beta = 0{,}5$

```
data uno;

beta=0.5;*coloque aquí el valor de beta;

alfa=beta;

n=500;*coloque aquí el tamaño de la muestra;

b1=beta+n;

ns=0.95;*coloque aquí el nivel de confianza;

ls=gaminv(ns,alfa)/b1;*este es límite superior;
proc print data=uno;

var alfa n ns ls;

run;

quit;
```

# Referencias


1. Jovanovic BJ, Levy PS. A look at the rule of tree. Am Stat 1997; 51(2); 137-139.

2. Serfling RJ. Approximation theorems of mathematical statistics. New YorK: John Wiley & Sons; 1980.

3. Correa, JC, Sierra E. Sobre la estimación de la distribución poisson con tasa pequeña. Reporter Técnico. Medellín: Universidad Nacional; 1999.

4. DeGroot M.H. Optimal statistical decisions. New York: McGraw-Hill; 1970.