Algunas observaciones sobre los Intervalos de Confianza

Rodrigo Zepeda (rodrigo.zepeda@itam.mx)

16 de julio 2020

Intervalos asintóticos

Los intervalos de confianza que hemos construido hasta ahora son con la idea de normalidad asintótica; se basan en la idea de que:

$$Z = \lim_{N-n,n\to\infty} \sqrt{\frac{1}{\text{Var}(\bar{x}_{\mathcal{S}})}} \cdot \left(\frac{1}{n}\sum_{i=1}^{N} x_i \mathbb{I}_{\mathcal{S}}(x_i) - \bar{x}_{\mathcal{U}}\right) \sim \text{Normal}(0,1)$$

Es decir, si tuviéramos una población infinita N y una muestra infinita n entonces la diferencia entre la media muestral y la media verdadera (normalizadas por la varianza) tienen una distribución normal. Empero, esto no siempre funciona como el siguiente con $n = 15$ muestras nos enseña:

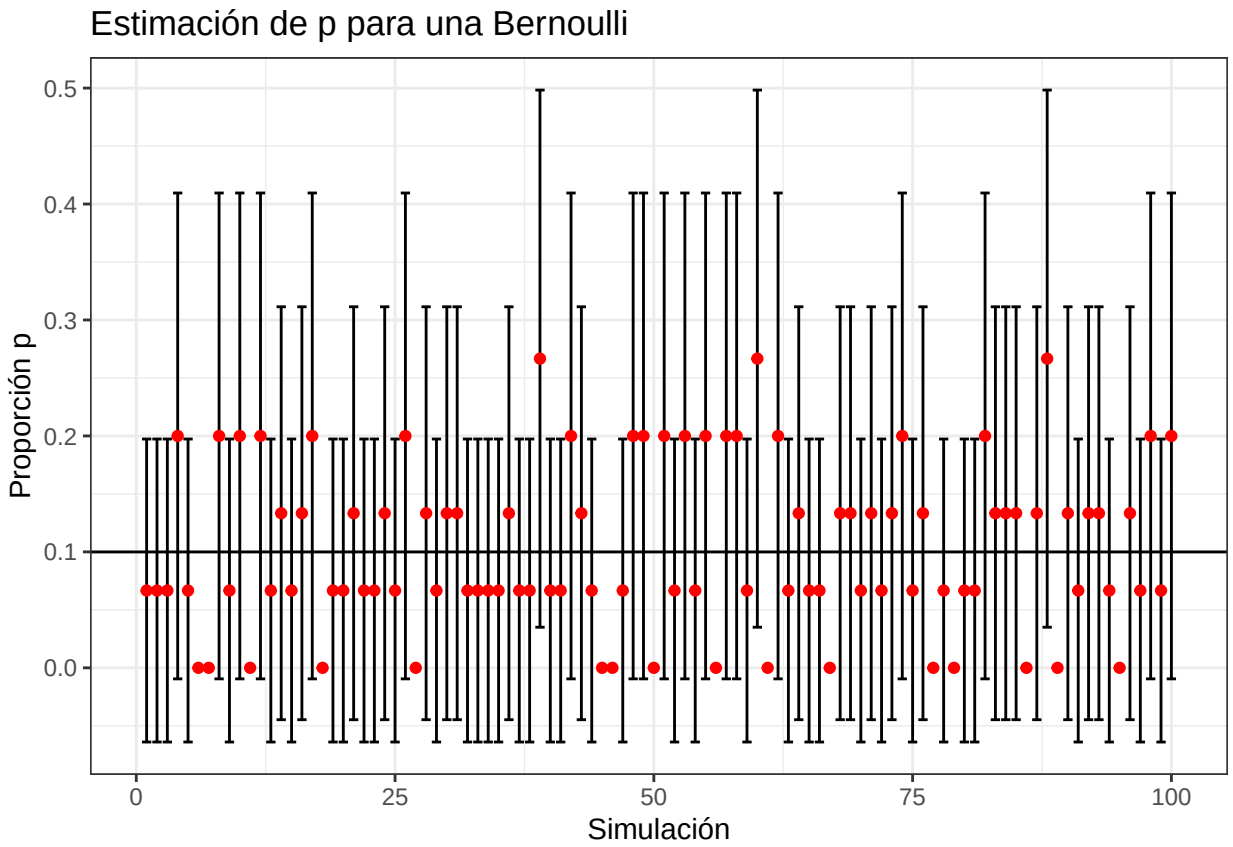
```
#Tomar muestras de tamaño pequeño
n       <- 15
lambda  <- 0.1
Z       <- qnorm(0.975)
nsim    <- 100    #Cantidad sims
N       <- 1000000 #Tamaño poblaciím
pop     <- sample(c(0,1), N, replace = TRUE, prob = c(1 - lambda, lambda))


#Creamos las bases donde guardar los datos
datos <- data.frame(matrix(NA, ncol = 3, nrow = nsim))
colnames(datos) <- c("IC_Bajo","Media","IC_Alto")
datos$Sim       <- 1:nsim

#Obtenemos 100 muestras
for (i in 1:nsim){
  muestra        <- sample(pop, n)
  datos$Media[i] <- mean(muestra)
  varianza       <- (1 - n/N)/n*var(muestra)
  datos$IC_Bajo[i]  <- datos$Media[i] - Z*sqrt(varianza)
  datos$IC_Alto[i]  <- datos$Media[i] + Z*sqrt(varianza)
}

ggplot(datos, aes(x = Sim)) +
  geom_errorbar(aes(ymin = IC_Bajo, ymax = IC_Alto)) +
  geom_point(aes(y = Media), color = "red") +
  geom_hline(aes(yintercept = mean(pop) )) +
  theme_bw() +
  labs(
    x = "Simulación",
```

```
    y = "Proporción p",
    title = "Estimación de p para una Bernoulli"
  )
```



En el caso de la Bernoulli, por ejemplo, se puede un mejor intervalo de confianza para cuando la proporción $\hat{p} = 0$ **bajo una hipótesis distribucional**. Esta se encuentra en el paquete `binom` de `R`:

```
library(binom)
binom.confint(0, n, conf.level = 0.95, method = 'exact')
```

```
##   method x  n mean lower     upper
## 1  exact 0 15    0     0 0.2180194
```

La deducción de dichos intervalos se hace a través de un esquema distinto de muestreo: **muestreo basado en modelos**.

Muestreo basado en modelos

Quizá en otras clases de estadística viste muestreo de otra manera. Lo usual en Estadística Matemática es suponer se tienen un conjunto de n variables aleatorias independientes idénticamente distribuidas que representan la cantidad de interés $\{Y_1, Y_2, \ldots, Y_n\}$. De dichas variables se observa que toman los valores $\{y_1, y_2, \ldots, y_n\}$ que son los medidos en la muestra. Aquí no se considera que haya un conjunto finito (universal) dado por la **población** sino que todas las variables provienen de una **metapoblación** (universo infinito de posibilidades). Así, por ejemplo, las alturas de individuos podrían tener una distribución gamma truncada y las alturas de personas en particular corresponden a realizaciones *independientes* de dichas Y_i. Este enfoque lo que hace es convertir el problema de estimación en un problema de predicción. No es que *la altura* de un individuo sea aleatoria en el sentido de que siempre cambie sino que *por nuestra ignorancia* nosotros modelamos dicha altura con una variable aleatoria que representa, en esta **metapoblación** la altura de una

cantidad infinita de individuos posibles. Bajo este esquema, una muestra aleatoria está dada por:

$$\mathcal{Y}_{(n)} = \{Y_1, Y_2, \ldots, Y_n\}$$

y la muestra observada es:

$$\dagger_{(n)} = \{y_1, y_2, \ldots, y_m\}$$

Donde suponemos que lo que se observó fue que $Y_i = y_i$. Un punto relevante aquí es que como las $\{Y_i\}$ son variables aleatorias éstas siguen algún modelo. En particular, se eligen modelos paramétricos para estos casos en los que la distribución asintótica no funciona. El problema de estimación se convierte ya sea en estimar el parámetro de la función de densidad (por ejemplo el $\Theta = (\mu, \sigma)^T$ de la normal o el λ de una Poisson) o una función del mismo (como puede ser la mismísima función de densidad o distribución acumulada).

> **Nota** El enfoque basado en modelos es muy bueno cuando se tienen pocos dados (o mal medidos) y las observaciones por sí mismas no son suficientes para poder generar información. En ese caso se hacen hipótesis adicionales (como un modelo) para los procesos de estimación.

Veamos un ejemplo de generación de intervalos para la media (λ) de una variable Poisson.

Ejemplo Poisson

Consideremos una muestra aleatoria $\mathcal{Y}_{(n)} = \{Y_1, Y_2, \ldots, Y_n\}$ de variables que se distribuyen Poisson(λ). Sabemos de proba (y si no lo sabemos no te apures, este ejemplo es sólo para ilustrar, no vamos a hacer más de esto) que la suma de variables aleatorias Poisson es Poisson, de donde:

$$\sum_{i=1}^{n} Y_i \sim \text{Poisson}(n \cdot \lambda)$$

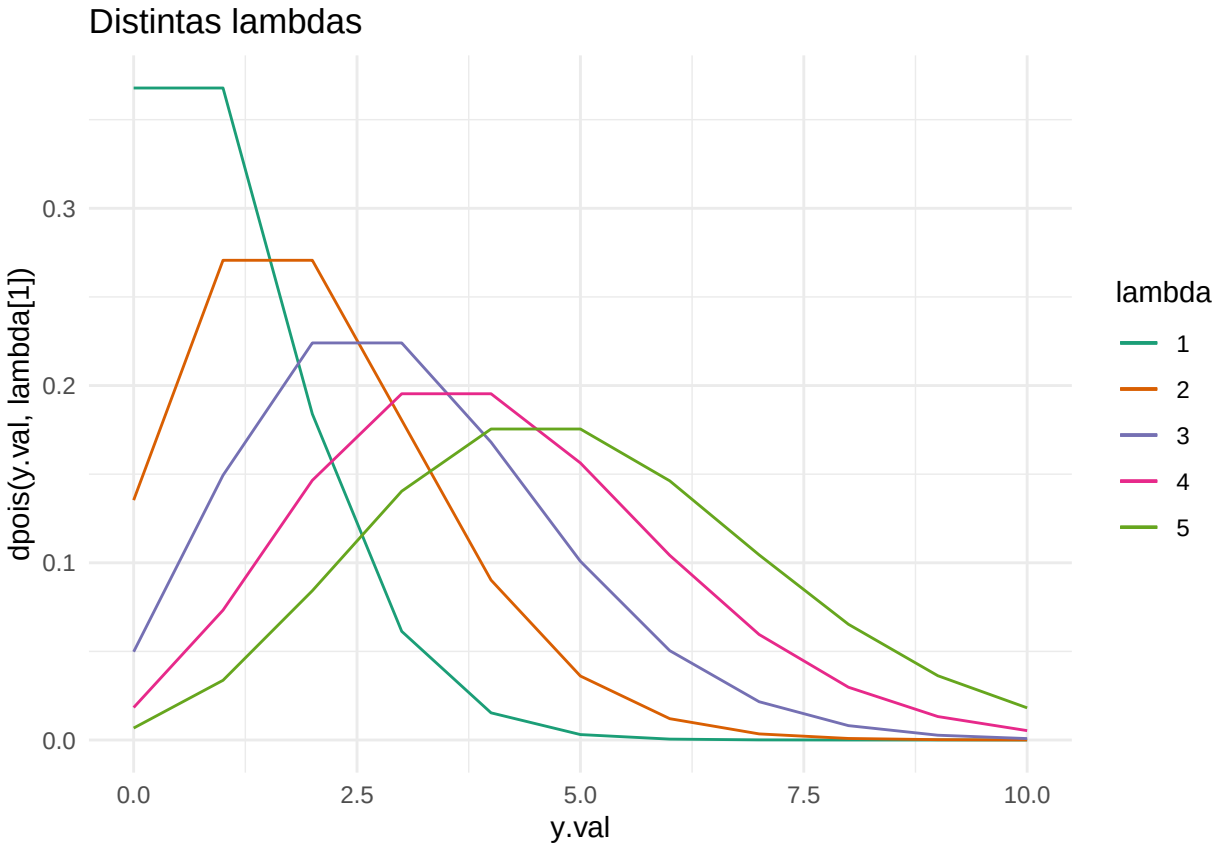
por lo cual:

$$n\bar{Y} = \sum_{i=1}^{n} Y_i \sim \text{Poisson}(n\lambda)$$

tenemos entonces que $\mathbb{E}[\bar{Y}] = \lambda$ (es insesgado). En este caso tomaremos que en particular observamos

$$\hat{y} = n\bar{y} = \sum_{i=1}^{n} y_i$$

Observamos que para y fijo, la distribución Poisson es decreciente antes de la media en λ y creciente después de



la media:

Podemos entonces obtener un estimador λ consideramos los valores λ_1 y λ_2 tales que:

$$\sum_{k\leq\hat{y}} \frac{(n\lambda_1)^{\hat{y}}e^{-n\lambda_1}}{\hat{y}!} = \alpha/2 \quad \text{y} \quad \sum_{k\geq\hat{y}} \frac{(n\lambda_2)^{\hat{y}}e^{-n\lambda_2}}{\hat{y}!} = \alpha/2$$

El cual podemos encontrar facilmente con ayuda de `R`:

```
#Obtenemos la muestra
n         <- 20
ybarra    <- sum(rpois(n, 1/5))
alpha.val <- 0.05

#Optimizamos
func.opt.1 <- function(lambda){ppois(ybarra, n*lambda) - alpha.val/2}
lambda.1   <- uniroot(func.opt.1, lower = 0, upper = 10, tol = 1.e-10)$root

func.opt.2 <- function(lambda){1 - ppois(ybarra, n*lambda) + dpois(ybarra, n*lambda) - alpha.val/2}
lambda.2   <- uniroot(func.opt.2, lower = 0, upper = 10, tol = 1.e-10)$root

c("Lower" = lambda.2, "Upper" = lambda.1)
```

```
##      Lower      Upper
## 0.03093361 0.43836365
```

> **Nota** Estos intervalos se conocen como *fiduciarios/fiduciales*. La idea es obtener los valores de λ tales que $\bar{Y} = Y$ con un intervalo de probabilidad $(1-\alpha)\times 100\%$.

Ejercico en clase

Obtén los intervalos fiduciarios para una variable aleatoria Bernoulli. Considera que la muestra está dada por:

```
set.seed(646)
muestra <- sample(c(0,1), 13, replace = TRUE, prob = c(0.2, 0.8))
```

Recuerda Si $X_i \sim \text{Bernoulli}(p)$ entonces $\sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$.