Trabajo final - Métodos de Estimación 2023

Jesica Charaf e Ignacio Spiousas

2023-07-08

Presentación

La distribución elegida para trabajar es la binomial negativa. Para el desarrollo utilizaremos la parametrización con r y μ que se muestra a continuación:

Sea $X \sim BN(r, \mu)$, entonces, su función de probabilidad puntual puede expresarse como:

$$p_X(x) = \frac{\Gamma(x+r)}{x!\Gamma(r)} \left(\frac{r}{r+\mu}\right)^r \left(\frac{\mu}{r+\mu}\right)^x,$$

con $x = 0, 1, 2, 3, ...$

Además, se tiene que: $\mathbb{E}(X) = \mu$ y $V(X) = \mu + \frac{\mu^2}{r}$.

Estimadores de momentos para r y μ

Obtención de los estimadores

Consideremos al conjunto de parámetros $\theta = (r, \mu)$ y que $X_1, ..., X_n \overset{iid}{\sim} BN(r, \mu)$. Como tenemos dos parámetros, para obtener los estimadores de momentos tendremos que utilizar el primer y segundo momento.

Primer momento

De plantear el primer momento obtenemos:

$$\overline{X} = \mathbb{E}_{\hat{\theta}}(X)$$
$$\overline{X} = \hat{\mu}.$$

Segundo momento

Para el segundo momento tenemos que plantear:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathbb{E}_{\hat{\theta}}(X^2).$$

Y para esto vamos a recordar que $V_{\hat{\theta}}(X) = \mathbb{E}_{\hat{\theta}}(X^2) - (\mathbb{E}_{\hat{\theta}}(X))^2$, por lo tanto:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathbb{E}_{\hat{\theta}}(X^2)$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = V_{\hat{\theta}}(X) + (\mathbb{E}_{\hat{\theta}}(X))^2$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\mu} + \frac{\hat{\mu}^2}{\hat{r}} + \hat{\mu}^2.$$

Resumiendo

Entonces, el sistema de ecuaciones que tenemos que resolver para hallar $\hat{r}$ y $\hat{\mu}$ es:

$$\overline{X} = \hat{\mu}$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\mu} + \frac{\hat{\mu}^2}{\hat{r}} + \hat{\mu}^2.$$

El despeje

En el caso del estimador de $\hat{\mu}$, no hace falta despejar nada ya que queda directamente definido por el primer momento. Por otro lado, a $\hat{r}$ lo podemos despejar fácilmente del segundo momento reemplazando $\hat{\mu}$ por $\overline{X}$:

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu} - \hat{\mu}^2 = \frac{\hat{\mu}^2}{\hat{r}}$$
$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X} - \overline{X}^2 = \frac{\overline{X}^2}{\hat{r}}$$
$$\hat{r} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X} - \overline{X}^2}$$

Entonces los estimadores de momentos son:

$$\hat{\mu}_{mo} = \overline{X}$$
$$\hat{r}_{mo} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X} - \overline{X}^2}$$

Consistencia

De nuevo, probar la consistencia del estimador $\hat{\mu}_{mo}$ es directo, ya que por LGN $\overline{X} \overset{c.s.}{\to} \mu$. Entonces, $\hat{\mu}_{mo}$ es un estimador consistente para μ.

Para estudiar la consistencia de $\hat{r}_{mo}$ también vamos a utilizar LGN, pero vamos a tener que utilizar algunas propiedades. Por LGN sabemos que:

$$\sum_{i=1}^{n} X_i^2 \overset{c.s.}{\to} \mathbb{E}(X^2) = \mu + \frac{\mu^2}{r} + \mu^2.$$

Además, como $\overline{X} \overset{c.s.}{\to} \mu$ y $g(t) = t^2$ es una función continua, podemos afirmar que $\overline{X}^2 \overset{c.s.}{\to} \mu^2$. Luego, utilizando propiedades de convergencia casi segura:

$$\frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X} - \overline{X}^2} \xrightarrow{c.s.} \frac{\mu^2}{\mu + \frac{\mu^2}{r} + \mu^2 - \mu - \mu^2}$$
$$\frac{\overline{X}^2}{\frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X} - \overline{X}^2} \xrightarrow{c.s.} \frac{\mu^2}{\frac{\mu^2}{r}} = r.$$

Por lo tanto, $\hat{r}_{mo}$ es un estimador consistente para r.

Distribución asintótica

Para empezar, podemos ver que el estimador de momentos de μ es asintóticamente normal a partir del Teorema Central del Límite. Dado que $\hat{\mu}_{mo} = \overline{X}$, $E(\overline{X}) = \mu$ y $V(\overline{X}) = \frac{V(X_i)}{n}$, por TCL tenemos que:

$$\frac{\overline{X} - \mu}{\sqrt{\frac{1}{n}\left(\mu + \frac{\mu^2}{r}\right)}} \xrightarrow{D} \mathcal{N}(0,1).$$

En el caso del estimador de momentos de r el cálculo de la distribución asintótica no es tan inmediato, por lo que haremos un estudio de simulación para analizar la distribución asintótica de $\hat{r}_{mo}$ y, a su vez, verificar el comportamiento del estimador $\hat{\mu}_{mo}$.

Para estudiar la distribución asintótica vamos a generar simulaciones de datos con distribución binomial negativa y parámetros poblacionales $r = 1$ y $\mu = 5$ y para cuatro tamaños de muestra, con $n = 10, 50, 100, 500$. Con cada simulación vamos a estimar los parámetros utilizando los estimadores de momentos ($\hat{\theta}$) y, a partir de ellos, vamos a calcular las estimaciones estandarizadas de acuerdo a:

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})}.$$

Una vez obtenidos los valores estandarizados para $\hat{r}$ y $\hat{\mu}$ graficaremos los histogramas de ambas distribuciones comparados con la densidad de una distribución $\mathcal{N}(0,1)$.

Primero definimos las funciones que generarán las estimaciones de momentos

```
mu_mo <- function(x) {
  mean(x)
}

r_mo <- function(x) {
  (mean(x))^2/(mean(x^2) - mean(x) - mean(x)^2)
}
```

Luego, vamos a realizar `Nrep=10000` simulaciones del vector aleatorio $(X_1, ..., X_n)$ con $X_i \sim BN(\mu = 5, r = 1)$:

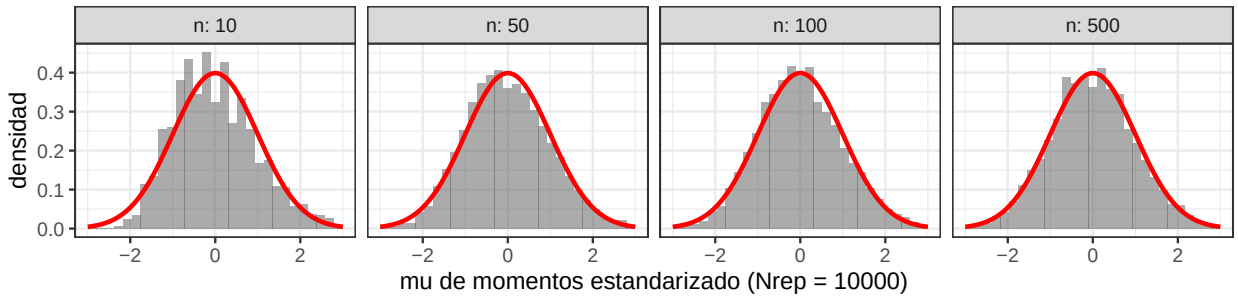
```
ns <- c(1e1, 5e1, 1e2, 5e2)
Nrep <- 1e4
mu_pob <- 5
r_pob <- 1

est_rs <- vector(length = length(ns)*Nrep)
est_mus <- vector(length = length(ns)*Nrep)
est_ns <- vector(length = length(ns)*Nrep)
```

```
set.seed(12)
for (i in 1:length(ns)) {
  for (j in 1:Nrep) {
    data <- rnbinom(n = ns[i], size = r_pob, mu = mu_pob)
    est_ns[(i-1)*Nrep+ j] <- ns[i]
    est_mus[(i-1)*Nrep + j] <- mu_mo(data)
    est_rs[(i-1)*Nrep + j] <- r_mo(data)
  }
}
```

Luego estandarizamos las estimaciones obtenidas con las muestras simuladas:

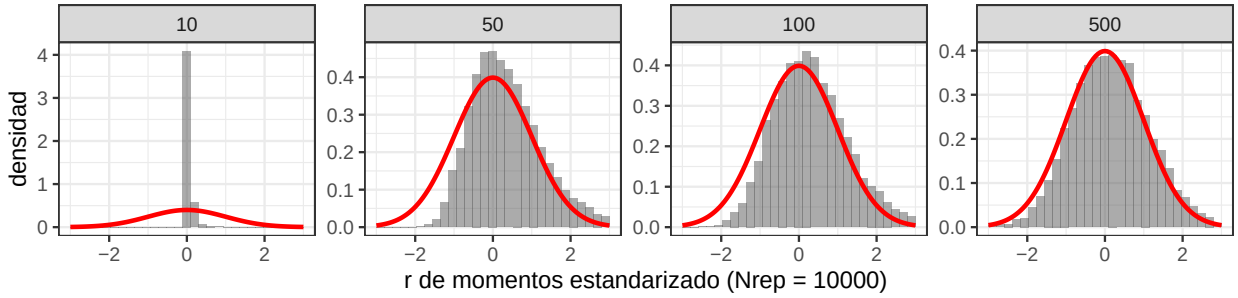
```
sim_mo <- tibble(n = est_ns,
                 mu = est_mus,
                 r = est_rs) %>%
  filter(r!=Inf) %>%
  group_by(n) %>%
  mutate(mu_est = (mu - mu_pob)/sd(mu),
         r_est = (r - r_pob)/sd(r))
```

Y, finalmente, vemos los histogramas de estas estimaciones estandarizadas comparados con una distribución $\mathcal{N}(0,1)$. Primero $\hat{\mu}_{mo,est}$:



En los histogramas puede verse que, aún para valores chicos de n, la distribución del estimador estandarizado $\hat{\mu}_{mo,est}$ es muy similar a la línea roja, es decir, su distribución asintótica es una $\mathcal{N}(0,1)$. Por otro lado, algo que podemos observar es que el estimador no pareciera tener un sesgo para ningún valor de n.

Y para $\hat{r}_{mo,est}$:



En este caso, a diferencia de lo visto para $\hat{\mu}_{mo,est}$, vemos que sólo para valores de n grandes la distribución de $\hat{r}_{mo,est}$ se aproxima a la $\mathcal{N}(0,1)$, mientras que para valores chicos de n muestra una distribución más concentrada y sesgada hacia la derecha.

Una cosa extra que podemos hacer, ya que tenemos simulaciones, es verificar la consistencia calculando el error cuadrático medio empírico (ECME):

```
sim_mo %>%
  group_by(n) %>%
  summarise(ECME_r = mean(r - r_pob)^2,
            ECME_mu = mean(mu - mu_pob)^2) %>%
  knitr::kable()
```

n	ECME_r	ECME_mu
10	0.2095478	7.51e-05
50	0.0244619	1.90e-06
100	0.0069814	1.77e-05
500	0.0002816	1.90e-06

Puede verse que ambos estimadores son consistentes dado que el ECME tiende a cero cuando n crece. A a su vez, a partir de los histogramas, pareciera que $\hat{\mu}_{mo}$ es un estimador insesgado mientras que $\hat{r}_{mo}$ uno asintóticamente insesgado, pero con un sesgo a derecha para valores chicos de n.

Estimadores de máxima verosimilitud para r y μ

Para hallar los estimadores de máxima versosimilitud utilizaremos el comando `fitdistr` del paquete {MASS}.

Obtención de los estimadores

Las funciones que nos devuelven los estimadores de máxima verosimilitud ($\hat{r}_{mv}$ y $\hat{\mu}_{mv}$) son:

```
mu_mv <- function(x) {
  fitdistr(x, densfun = "negative binomial")$estimate[2]
}

r_mv <- function(x) {
  fitdistr(x, densfun = "negative binomial")$estimate[1]
}
```

Consistencia

Para la consistencia vamos a realizar un procedimiento similar al utilizado para estudiar la distribución asintótica y la consistencia (a partir del ECME) de los estimadores de momentos. Primero hacemos la simulación de las muestras y obtenemos las estimaciones:

```
ns <- c(1e1, 5e1, 1e2, 5e2)
Nrep <- 1e4
mu_pob <- 5
r_pob <- 1

est_rs <- vector(length = length(ns)*Nrep)
est_mus <- vector(length = length(ns)*Nrep)
est_ns <- vector(length = length(ns)*Nrep)

set.seed(12)
for (i in 1:length(ns)) {
  cat(paste("n =", ns[i], "\n"))
  for (j in 1:Nrep) {
```

```
    data <- rnbinom(n = ns[i], size = r_pob, mu = mu_pob)
    est_mv <- fitdistr(data, densfun = "negative binomial")
    est_ns[(i-1)*Nrep+ j] <- ns[i]
    est_mus[(i-1)*Nrep + j] <- est_mv$estimate[2]
    est_rs[(i-1)*Nrep + j] <- est_mv$estimate[1]
  }
}
```

```
## n = 10
## n = 50
## n = 100
## n = 500
```

Luego, calculamos el ECME:

```
sim_mv <- tibble(n = est_ns,
                 mu = est_mus,
                 r = est_rs) %>%
  filter(r!=Inf)

ECME_mv <- sim_mv %>%
  group_by(n) %>%
  summarise(ECME_r = mean(r - r_pob)^2,
            ECME_mu = mean(mu - mu_pob)^2)

ECME_mv %>%
  knitr::kable()
```

n	ECME_r	ECME_mu
10	4.4644000	5.56e-05
50	0.0067779	1.90e-06
100	0.0018059	1.77e-05
500	0.0000514	1.80e-06

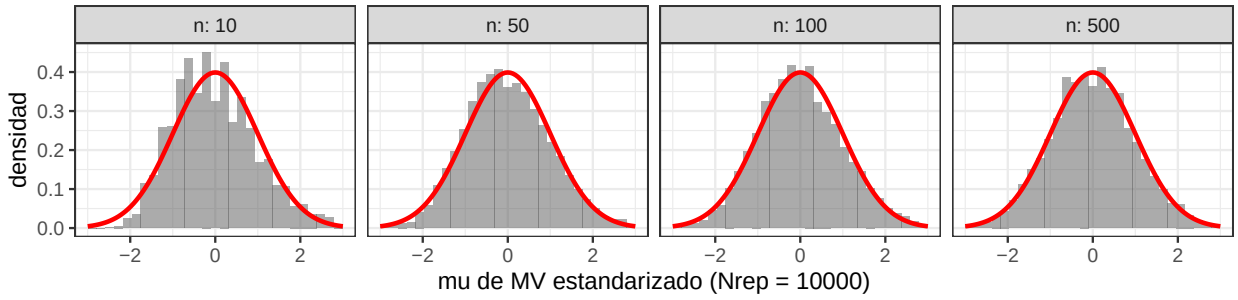
De la misma forma que en los estimadores de momentos, $\hat{\mu}_{mv}$ y $\hat{r}_{mv}$ parecen ser consistentes, aunque $\hat{\mu}_{mv}$ parece ser insesgado mientras que $\hat{r}_{mv}$ asintóticamente insesgado.

Distribución asintótica

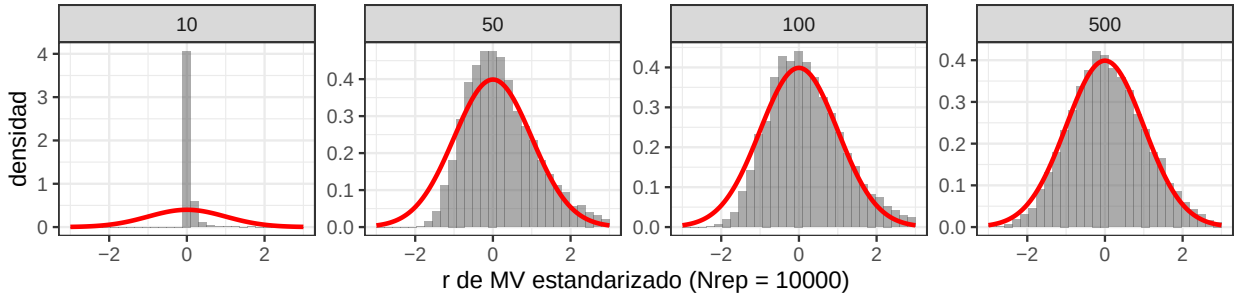
Lo primero que hacemos es estandarizar las estimaciones de los parámetros obtenidas a partir de las muestras simuladas:

```
sim_mv <- sim_mv %>%
  group_by(n) %>%
  mutate(mu_est = (mu - mu_pob)/sd(mu),
         r_est = (r - r_pob)/sd(r))
```

Después de estandarizar podemos ver los histogramas para $\hat{\mu}_{mv,est}$:



Y para $\hat{r}_{mv,est}$:



En estos, al igual que en los estimadores de momentos se ve que la distribución asintótica para ambos parámetros se parece a una $\mathcal{N}(0,1)$, pero con diferencias para valores de n pequeños.

Aplicaciones y datos

En la parametrización que se presenta en este trabajo no tenemos a la distribución negativa como una distribución que se utiliza para modelar la cantidad de fracasos hasta tener k éxitos, sino como un caso general de la distribución de Poisson en el que la media no es igual a la varianza. Al igual que la distribución de Poisson, la binomial negativa con esta parametrización se utiliza para modelar el conteo de un proceso aleatorio, en particular en casos donde hay sobredispersión, es decir, donde la varianza es más grande que la esperanza (Stoklosa, Blakey, and Hui 2022).

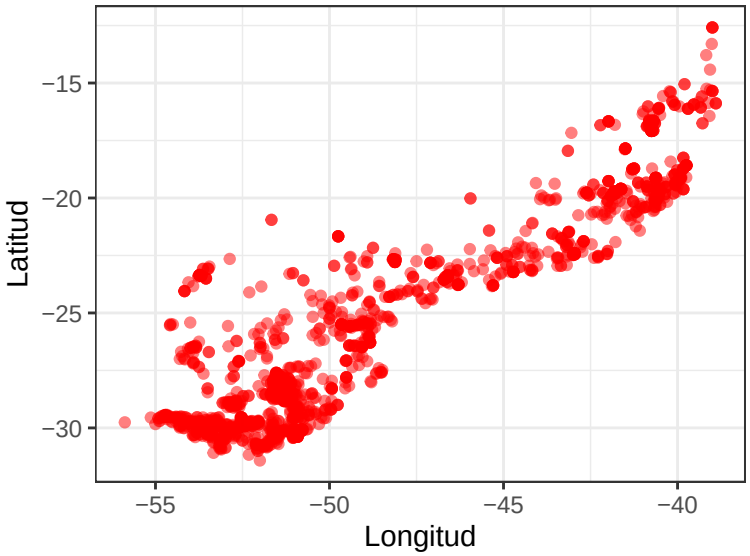
Recordando los valores de la esperanza y la varianza, $\mathbb{E}(X) = \mu$ y $V(X) = \mu + \frac{\mu^2}{r}$, podemos ver que es el parámetro r el que gobierna la relación entre esperanza y varianza, donde con valores de r más pequeños se tienen poblaciones más dispersas (en la literatura se conoce a r como factor de dispersión) y con valores de r grandes se tienen poblaciones más similares a una Poisson (se puede demostrar que cuando $r \to \infty$ la binomial negativa se transforma en una Poisson). Como ejemplos de sobredispersión tenemos el caso de los días de hospitalización de un individuo (Weaver et al. 2015) o el censo de alguna especie en una grilla geográfica (Stoklosa, Blakey, and Hui 2022).

En particular cuando lidiamos con el conteo de una especie podemos esperar que no esté distribuido uniformemente, sino que los individuos aparezcan en grupos o comunidades. Como se menciona en (Stoklosa, Blakey, and Hui 2022), esto hace que la varianza de la muestra sea muy superior a la media, lo que indicaría que debemos abandonar el modelo de Poisson y considerar modelar los datos asumiendo que las variables aleatorias en juego provienen de una distribución binomial negativa (argumento que defienden fuertemente en (Stoklosa, Blakey, and Hui 2022)).

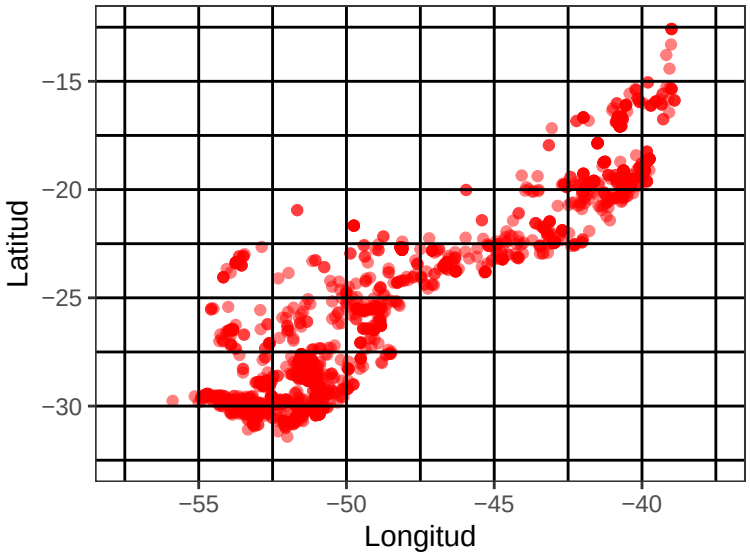
Ejemplo 1: Un caso de conteo de individuos de una especie

Los datos que vamos a modelar en este primner ejemplo fueron obtenidos de (Culot et al. 2019) y representan la observación de individuos de mono Carayá Marrón (*Alouatta guariba*) en el sudeste de Brasil. Si vemos los

datos en función de su longitud y latitud (podemos imaginarnos la costa de Brasil), su distribución geográfica es la siguiente.

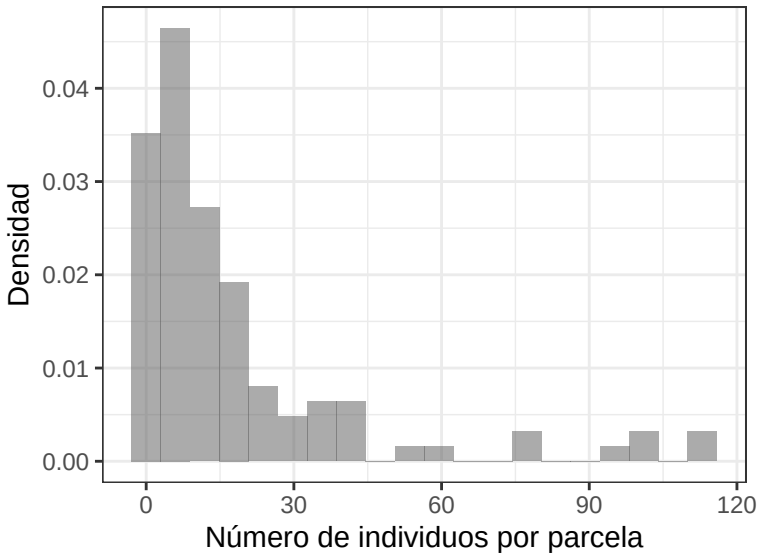


Como podemos ver, hay agrupamientos de individuos y los mismos no se distribuyen uniformemente. Ahora imaginemos qué pasa si insertamos una grilla en la que, por algún motivo, queremos subdividir nuestras observaciones (esto está muchas veces relacionado con la forma en la que se censan las especies en estudios de biodiversidad).



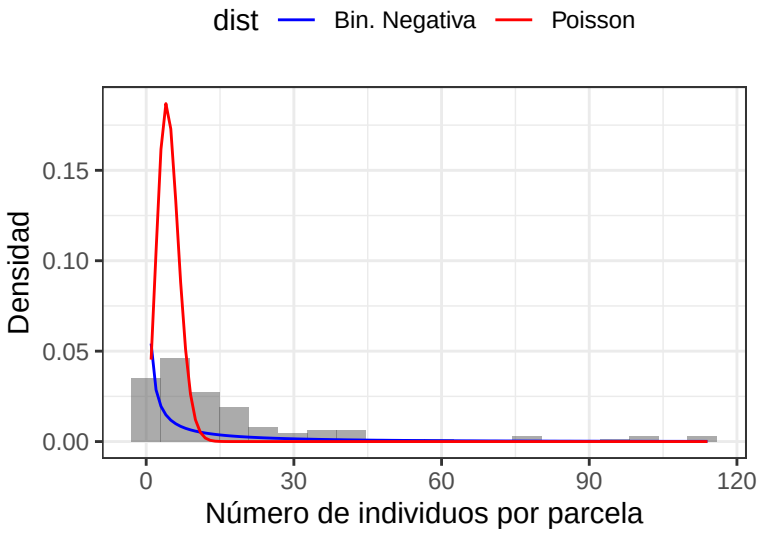
En este caso, vamos a realizar el experimento de contabilizar los individuos observados en cada parcela (no necesariamente las de la figura, sino unas más pequeñas). Para este ejercicio en particular vamos a quitar los ceros de esa lista[1] y a graficar un histograma para ver cómo se distribuyen esos conteos.

[1] Habría que tener otras consideraciones que se van del alcance del presente trabajo



En esta muestra, se observa que la media muestral es 4.62 y la varianza muestral 220.22, siendo la segunda notablemente más grande que la primera.

Ahora, veamos qué pasa si superponemos la densidad obtenida a partir de las estimaciones de máxima verosimilitud de los parámetros de una distribución de Poisson (en rojo) y de una binomial negativa (en azul).



Podemos ver que, si bien ninguno de los dos modelos se ajusta perfectamente a los datos, pareciera que el modelo que asume una distribución binomial negativa es más acorde a la muestra observada.

Ejemplo 2: Un caso de días de hospitalización

A continuación vamos a analizar un conjunto de datos (proporcionado por la cátedra) sobre los días de internación de pacientes que llegan a la guardia de un hospital. Los datos contienen 1000 historias clínicas elegidas al azar, con el registro de la cantidad de días de internación de cada paciente, es decir una variable de conteo.

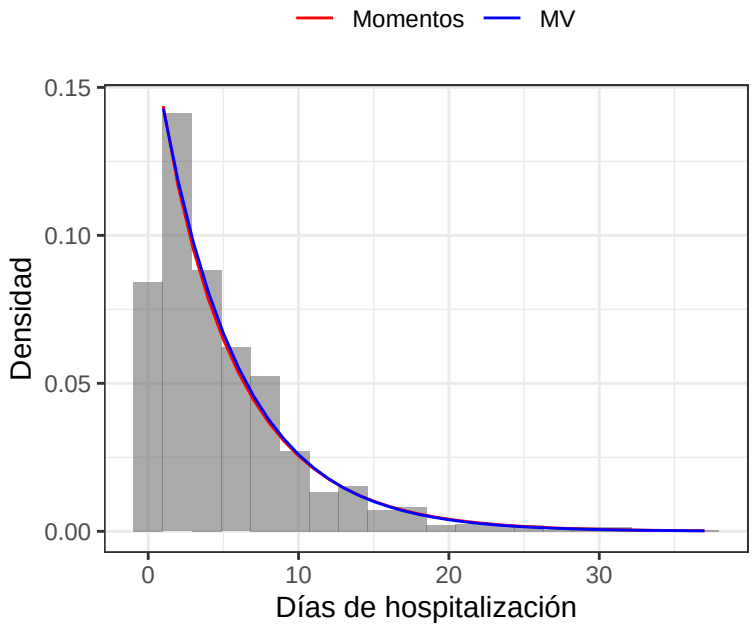
A partir de la muestra, vemos que la media muestral es 4.79 y la varianza muestral 29.36. Es decir, estamos

frente a una muestra de una variable de conteo con varianza considerablemente mayor a la media, por lo que resulta razonable modelar la cantidad de días de internación de cada paciente como una variable aleatoria con distribución binomial negativa.

Para empezar, ajustamos un modelo correspondiente a una distribución binomial negativa y calculamos los estimadores de momentos y de máxima verosimilitud, obteniendo los siguientes resultados.

Método	mu	r
Momentos	4.792000	0.9357965
MV	4.792088	1.0006748

Además, realizamos un histograma de densidad de los datos y, a su vez, graficamos de forma superpuesta las densidades estimadas a partir de las estimaciones previas de los parámetros.



Podemos ver que ambos estimadores tienen un desempeño similar y parecen ajustarse adecuadamente al conjunto de datos.

Referencias

Culot, Laurence, Lucas Augusto Pereira, Ilaria Agostini, Marco Antônio Barreto de Almeida, Rafael Souza Cruz Alves, Izar Aximoff, Alex Bager, et al. 2019. "ATLANTIC-PRIMATES: A Dataset of Communities and Occurrences of Primates in the Atlantic Forests of South America." Wiley Online Library.

Stoklosa, Jakub, Rachel V Blakey, and Francis KC Hui. 2022. "An Overview of Modern Applications of Negative Binomial Modelling in Ecology and Biodiversity." *Diversity* 14 (5): 320.

Weaver, Colin G, Pietro Ravani, Matthew J Oliver, Peter C Austin, and Robert R Quinn. 2015. "Analyzing Hospitalization Data: Potential Limitations of Poisson Regression." *Nephrology Dialysis Transplantation* 30 (8): 1244–49.