Cálculo de probabilidades

Capítulo 3: Familias paramétricas especiales de distribuciones univariadas

Mg. Sc. J. Eduardo Gamboa U.

Familia paramétrica

Es una colección de densidades que están indexadas por una o más cantidades llamadas parámetros.

Por ejemplo, $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) I_{(-\infty,\infty)}(x)$ es una familia paramétrica indexada con los parámetros σ^2 y μ. Si $\sigma^2 = 4$ y $\mu = 10$ se tiene el siguiente miembro de la familia paramétrica:

$$f(x) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2 \times 4}(x-10)^2\right) I_{(-\infty,\infty)}(x)$$

Distribuciones discretas

1. Distribución uniforme discreta
2. Distribución Bernoulli
3. Distribución Binomial
4. Distribución hipergeométrica
5. Distribución Poisson
6. Distribución geométrica
7. Distribución Pascal
8. Distribución zeta

Distribuciones continuas

1. Distribución uniforme continua
2. Distribución triangular
3. Distribución normal
4. Distribución lognormal
5. Distribución gamma
6. Distribución beta
7. Distribución Weibull
8. Distribución Cauchy
9. Distribución Laplace
10. Distribución logística
11. Distribución Pareto

Distribución uniforme discreta

Función de probabilidad

Una variable aleatoria discreta X que toma los valores x_1, ..., x_N tiene distribución uniforme discreta si su función de probabilidad es:

$$f(x) = \frac{1}{N} I_{\{x_1, \ldots, x_N\}}(x)$$

Notación: $X \sim Unif\ Discreta(x_1, ..., x_N)$

Ejemplo:

- Resultado al lanzar un dado justo
- Resultado al sacar al azar un ticket, numerado del 1 al n
- Resultado de seleccionar una carta de corazones

Valor esperado

$$E(X) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Varianza

$$V(X) = \frac{1}{N} \sum_{i=1}^{N} (x_i - E(X))^2$$

Función generatriz de momentos

$$M_X(t) = \frac{1}{N}\sum_{i=1}^{N} e^{tx_i}$$

Función característica

$$\phi_X(t) = \frac{1}{N}\sum_{j=1}^{N} e^{itx_j}$$

Implementación en R

- Paquete: distr
- Funciones:
  - DiscreteDistribution(supp, prob)
  - d(X): probabilidad en un punto
  - p(X): probabilidad acumulada
  - r(X): muestra aleatoria

Ejemplo 1

La variable aleatoria X es la cara de un dado justo obtenida luego de ser lanzado.

$$f(x) = \frac{1}{6} I_{\{1,2,3,4,5,6\}}(x)$$

$$E(X) = \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

$$V(X) = \frac{1}{6}\left((1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2\right)$$

$$V(X) = 2.9167$$

$$M_X(t) = \frac{1}{6}\left(e^t + e^{2t} + e^{3t} + e^{4t} + e^{5t} + e^{6t}\right)$$

$$M_X^{(1)}(t) = \frac{1}{6}\left(e^t + 2e^{2t} + 3e^{3t} + 4e^{4t} + 5e^{5t} + 6e^{6t}\right)$$

$$M_X^{(1)}(0) = \frac{1}{6}\left(1 + 2 + 3 + 4 + 5 + 6\right) = 3.5 = E(X)$$

$$M_X^{(2)}(t) = \frac{1}{6}\left(e^t + 4e^{2t} + 9e^{3t} + 16e^{4t} + 25e^{5t} + 36e^{6t}\right)$$

$$M_X^{(2)}(0) = \frac{1}{6}\left(1 + 4 + 9 + 16 + 25 + 36\right) = 15.1667 = E(X^2)$$

$$V(X) = 15.1667 - 3.5^2 = 2.9167$$

$$\phi_X(t) = \frac{1}{6}\left(e^{it} + e^{2it} + e^{3it} + e^{4it} + e^{5it} + e^{6it}\right)$$

$$\phi_X^{(1)}(t) = \frac{1}{6}\left(ie^{it} + 2ie^{2it} + 3ie^{3it} + 4ie^{4it} + 5ie^{5it} + 6ie^{6it}\right)$$

$$\phi_X^{(1)}(0) = \frac{1}{6}\left(i + 2i + 3i + 4i + 5i + 6i\right) = 3.5i \Rightarrow E(X) = \frac{3.5i}{i} = 3.5$$

$$\phi_X^{(2)}(t) = \frac{1}{6}\left(i^2e^{it} + 4i^2e^{2it} + 9i^2e^{3it} + 16i^2e^{4it} + 25i^2e^{5it} + 36i^2e^{6it}\right)$$

$$\phi_X^{(2)}(0) = \frac{1}{6}\left(i^2 + 4i^2 + 9i^2 + 16i^2 + 25i^2 + 36i^2\right) = 15.1667i^2$$

$$E(X^2) = \frac{15.1667i^2}{i^2} = 15.1667$$

$$V(X) = 15.1667 - 3.5^2 = 2.9167$$

```r
library(distr)
library(distrEx)
X <- DiscreteDistribution(supp = 1:6,
                          prob = rep(1/6,6))
X |> E()
```

```
[1] 3.5
```

```r
X |> var()
```

```
[1] 2.916667
```

```r
3 |> d(X)()
```

```
[1] 0.1666667
```

```r
3 |> p(X)()
```

```
[1] 0.5
```

```r
3 |> r(X)()
```

```
[1] 5 1 5
```

Ejemplo 2

Un laboratorio de cómputo puede ser usado en alguno de los siguientes horarios, con la misma probabilidad: 8, 10, 14 o 17 horas.

$$f(x) = \frac{1}{4} I_{\{8,10,14,17\}}(x)$$

$$E(X) = \frac{1}{4}(8 + 10 + 14 + 17) = 12.25$$

$$V(X) = \frac{1}{4}\left((8 - 12.25)^2 + (10 - 12.25)^2 + (14 - 12.25)^2 + (17 - 12.25)^2\right) = 12.1875$$

$$M_X(t) = \frac{1}{4}\left(e^{8t} + e^{10t} + e^{14t} + e^{17t}\right)$$

$$M_X^{(1)}(t) = \frac{1}{4}\left(8e^{8t} + 10e^{10t} + 14e^{14t} + 17e^{17t}\right)$$

$$M_X^{(1)}(0) = \frac{1}{4}(8 + 10 + 14 + 17) = 12.25 = E(X)$$

$$M_X^{(2)}(t) = \frac{1}{4}\left(64e^{8t} + 100e^{10t} + 196e^{14t} + 289e^{17t}\right)$$

$$M_X^{(2)}(0) = \frac{1}{4}(64 + 100 + 196 + 289) = \frac{649}{4} = 162.25 = E(X^2)$$

$$V(X) = 162.25 - 12.25^2 = 12.1875$$

$$\phi_X(t) = \frac{1}{4}\left(e^{8it} + e^{10it} + e^{14it} + e^{17it}\right)$$

$$\phi_X^{(1)}(t) = \frac{1}{4}\left(8ie^{8it} + 10ie^{10it} + 14ie^{14it} + 17ie^{17it}\right)$$

$$\phi_X^{(1)}(0) = \frac{1}{4}\left(8i + 10i + 14i + 17i\right) = 12.25i \Rightarrow E(X) = \frac{12.25i}{i} = 12.25$$

$$\phi_X^{(2)}(t) = \frac{1}{4}\left(64i^2e^{8it} + 100i^2e^{10it} + 196i^2e^{14it} + 289i^2e^{17it}\right)$$

$$\phi_X^{(2)}(0) = \frac{1}{4}\left(64i^2 + 100i^2 + 196i^2 + 289i^2\right) = 162.25i^2 \Rightarrow E(X^2) = \frac{162.25i^2}{i^2} = 162.25$$

$$V(X) = 162.25 - 12.25^2 = 12.1875$$

```r
X <- DiscreteDistribution(supp = c(8,10,14,17),
                          prob = rep(1/4,4))
X |> E()
```

```
[1] 12.25
```

```r
X |> var()
```

```
[1] 12.1875
```

```r
14 |> d(X)()
```

```
[1] 0.25
```

```r
14 |> p(X)()
```

```
[1] 0.75
```

```r
5 |> r(X)()
```

```
[1] 14  8 10 14 14
```

Distribución Binomial

Experimento Bernoulli

- Consiste en una secuencia de n ensayos (muestreo con reemplazo, o sin reemplazo de una población infinita o muy grande), donde n se fija antes del experimento.
- Los ensayos son idénticos e independientes, y cada uno de ellos solo tiene dos posibles resultados: éxito (E) o fracaso (F), por ello se dice que es dicotómico.
- La probabilidad de éxito es conocida y constante de un ensayo a otro; se denota esta probabilidad por $P(E) = \pi$ y la probabilidad de fracaso es $P(F) = 1 - \pi$.
- La distribución Binomial es la suma de n variables de Bernoulli independientes:

$$X = \sum_{i=1}^{n} B_i \text{ donde } B_i \sim Bern(\pi)$$

Función de probabilidad

Dado un experimento binomial, entonces se define la V.A.D. X $=$ El número de éxitos en una secuencia de n ensayos Bernoulli independientes. Su función de probabilidad está dada por:

$$f(x) = P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x} I_{\{0,1,2,\ldots,n\}}(x)$$

Notación: $X \sim Bin(n, \pi)$

Ejemplos

- X $=$ Número de correos promocionales abiertos $\rightarrow X \sim Bin(n = 1000, \pi = 0.20)$
- Y $=$ Número de piezas defectuosas producidas $\rightarrow Y \sim Bin(n = 30, \pi = 0.05)$
- L $=$ Número de respuestas correctas en un examen de opción múltiple (cuando se elige al azar) $\rightarrow L \sim Bin(n = 10, \pi = 0.25)$
- S $=$ Número de semillas que germinan $\rightarrow S \sim Bin(n = 200, \pi = 0.88)$

Media

$$\mu_X = E(X) = n\pi$$

Varianza

$$\sigma_X^2 = Var(X) = n\pi(1-\pi)$$

Asimetría

- Cuando $\pi = 0.5$, la distribución es simétrica
- Cuando $\pi < 0.5$, la distribución es sesgada a la derecha
- Cuando $\pi > 0.5$, la distribución es sesgada a la izquierda

Función generatriz de probabilidades

$$G_X(s) = E(s^x) = (1 - p + ps)^n$$

Función generatriz de momentos

$$M_X(t) = E(e^{tX}) = (1 - p + pe^t)^n$$

Función característica

$$\phi_X(t) = E(e^{itX}) = (1 - p + pe^{it})^n$$

Implementación en R

- Paquete: stats
- Funciones:
  - dbinom: probabilidad en un punto
  - pbinom: probabilidad acumulada
  - qbinom: cuantil
  - rbinom: muestra aleatoria

Ejemplo 3

Una tienda de ropa implementó una nueva estrategia de marketing basada en recomendaciones personalizadas. En cada una de las semanas previas, se observó que el 75% de los clientes que recibieron la recomendación terminaron comprando al menos una prenda.

Esta semana, llegaron 20 nuevos clientes, cada uno expuesto a la misma estrategia.

Experimento Bernoulli:

- Ensayo: llegada de un cliente a la tienda. Se tienen $n = 20$ ensayos
- Cada cliente que llega es independiente y tiene dos posibilidades (E=éxito o F=fracaso): E $=$ $\{$un cliente compra$\}$ y F $=$ $\{$un cliente no compra$\}$
- La probabilidad de éxito es $\pi = 0.75$, y la de fracaso es $1 - \pi = 0.25$

Variable aleatoria:

X : número de clientes que hacen una compra, $X \sim Bin(n = 20, \pi = 0.75)$

Distribución de probabilidades:

$$f(x) = P(X = x) = \binom{20}{x} 0.75^x \times 0.25^{20-x} I_{\{0,1,2,\ldots,20\}}(x)$$

La cantidad media de clientes que hacen una compra:

$$\mu_X = E(X) = 20 \times 0.75 = 15$$

Su varianza:

$$\sigma_X^2 = V(X) = 20 \times 0.75 \times 0.25 = 3.75$$

Función generatriz de probabilidades:

$$G_X(s) = E(s^x) = (0.25 + 0.75s)^{20}$$

Función generatriz de momentos:

$$M_X(t) = E(e^{tX}) = (0.25 + 0.75e^t)^{20}$$

Función característica

$$\phi_X(t) = E(e^{itX}) = (0.25 + 0.75e^{it})^{20}$$

La probabilidad de que el número de clientes que hacen una compra sea mayor que 18:

$$P(X > 18) = P(X = 19) + P(X = 20) = \binom{20}{19} 0.75^{19} 0.25 + \binom{20}{20} 0.75^{20} 0.25^{0} = 0.0243$$

```r
(px19 <- dbinom(x = 19, size = 20, prob = 0.75))
```

```
[1] 0.02114141
```

```r
(px20 <- dbinom(x = 20, size = 20, prob = 0.75))
```

```
[1] 0.003171212
```

```r
(px   <- px19 + px20)
```

```
[1] 0.02431262
```

La probabilidad de que el número de clientes que hacen una compra sea como máximo 14 es:

$$P(X \leq 14) = \sum_{i=1}^{14} P(X = i) = 0.383$$

```r
pbinom(q = 14, size = 20, prob = 0.75)
```

```
[1] 0.3828273
```

La cantidad mediana de clientes que hacen una compra es:

```r
qbinom(p = 0.5, size = 20, prob = 0.75)
```

```
[1] 15
```

Esto significa que en al menos la mitad de las semanas, 15 o menos clientes realizan una compra.

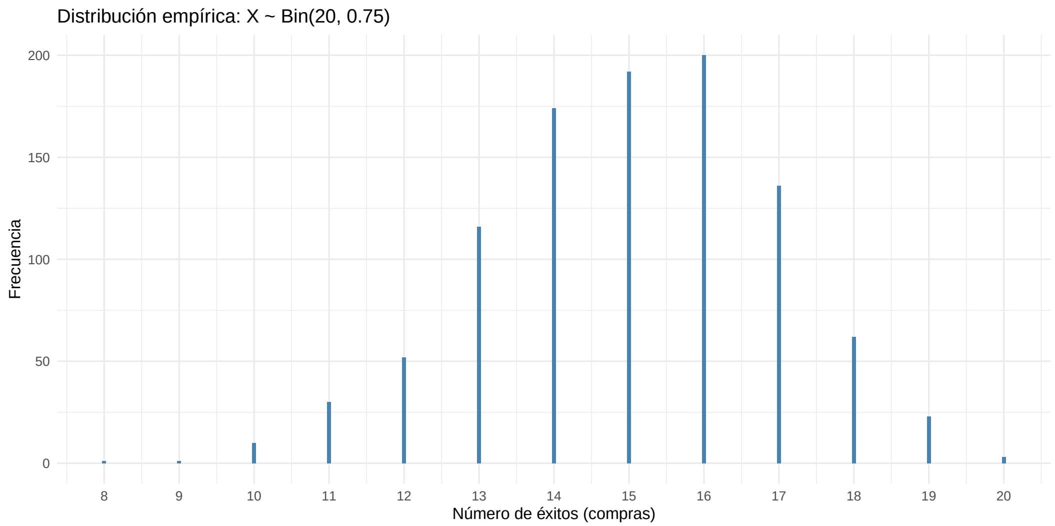
Simulación de una muestra aleatoria de la cantidad de clientes que hace una compra:

```r
set.seed(159)
rbinom(n = 10, size = 20, prob = 0.75)
```

```
 [1] 17 15 14 17 19 18 17 13 16 17
```

Se ha simulado 10 semanas, y en cada una de ellas la cantidad de clientes que realizaron una compra fue 17, 15, 14, etc.

```r
library(ggplot2)
set.seed(123)
x <- rbinom(n = 1000, size = 20, prob = 0.75)
df <- as.data.frame(table(x))
colnames(df) <- c("Éxitos", "Frecuencia")
df$Éxitos <- as.numeric(as.character(df$Éxitos))
ggplot(df, aes(x = Éxitos, y = Frecuencia)) +
  geom_bar(stat = "identity", fill = "steelblue", width = .05) +
  scale_x_continuous(breaks=8:20)+
  labs(title = "Distribución empírica: X ~ Bin(20, 0.75)",
       x = "Número de éxitos (compras)",
       y = "Frecuencia") +
  theme_minimal()
```


Distribución empírica: X ~ Bin(20, 0.75)
Frecuencia
200
150
100
50
0
8
9
10
11
12
13
14
15
16
17
18
19
20
Número de éxitos (compras)

Distribución Hipergeométrica

Los supuestos que se consideran para una distribución hipergeométrica son:

- Se tiene una población de N elementos, individuos u objetos (una población finita)
- Cada elemento tiene dos posibles resultados: éxito (E) o fracaso (F). Existen A éxitos y (N-A) fracasos en la población.
- Se selecciona una muestra aleatoria sin reemplazo de tamaño n.

Al no reponer elementos, la probabilidad de selección varía en cada ensayo. Se aplica en poblaciones pequeñas, como en la prueba exacta de Fisher y en muestreos de aceptación por atributos.

Función de probabilidad

Se define la V.A. Discreta X=Número de éxitos en la muestra de tamaño n. Su distribución de probabilidad es:

$$f(x) = P(X = x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}} I_{\{\mathbf{máx}(0,n+A-N),\ldots,\mathbf{mín}(n,A)\}}(x)$$

Notación: $X \sim Hiper(N, n, A)$

Ejemplos

- X = Número de focos defectuosos en una muestra
  $X \sim Hiper(N = 100, n = 15, A = 10)$
- Z = Número de expedientes con errores detectados en una auditoría
  $Z \sim Hiper(N = 200, n = 20, A = 30)$
- N = Número de estudiantes que recibieron una beca
  $N \sim Hiper(N = 80, n = 10, A = 25)$

Media

$$\mu_X = E(X) = n\frac{A}{N}$$

Varianza

$$\sigma_X^2 = Var(X) = n\frac{A}{N}\left(1 - \frac{A}{N}\right)\left(\frac{N-n}{N-1}\right)$$

Asimetría

- Cuando $\frac{A}{N} \approx 0.5$ y $n \approx \frac{N}{2}$, la distribución es simétrica
- Cuando $\frac{A}{N}$ es pequeña, la distribución es sesgada a la izquierda.
- Cuando $\frac{A}{N}$ es grande, la distribución es sesgada a la derecha.

Implementación en R

- Paquete: stats
- Funciones:
  - dhyper: probabilidad en un punto
  - phyper: probabilidad acumulada
  - qhyper: cuantil
  - rhyper: muestra aleatoria

Diferencia entre Distribuciones Hipergeométrica y Binomial

- D. Binomial: La probabilidad del evento es constante en cada ensayo.
- D. Hipergeométrica: La probabilidad cambia en cada ensayo porque no hay reemplazo.
- Ejemplo (población de 5 personas, 3 con sangre O+):
  - $P(1^a \text{ persona O+}) = \frac{3}{5} = 0.6$
  - Si la primera persona es O+, $P(2^a \text{ persona O+}) = \frac{2}{4} = 0.5$
  - En la distribución hipergeométrica, cada selección altera las probabilidades, especialmente en poblaciones pequeñas.
- Ejemplo (población de 500 personas, 300 con sangre O+):
  - $P(1^a \text{ persona O+}) = \frac{300}{500} = 0.6$
  - Si la primera persona es O+, $P(2^a \text{ persona O+}) = \frac{299}{499} = 0.5992 \approx 0.6$
  - La diferencia es imperceptible, por el tamaño grande de la población. Podría considerarse como un caso de distribución Binomial con prob. de éxito $\pi = 0.6$.

Ejemplo 4

Un embarque internacional de sustancias químicas ha llegado al puerto en 15 contenedores sellados. Por protocolo de bioseguridad, se debe realizar un muestreo aleatorio sin reemplazo para evaluar la pureza del producto antes de autorizar su ingreso al país.

Tras un informe preliminar, se sospecha que 2 de los 15 contenedores no cumplen con los requisitos de pureza. Para verificar la situación, se seleccionan aleatoriamente 3 contenedores para análisis de laboratorio.

Variable aleatoria:

Éxito=Contenedor que no cumplen con los estándares de pureza.

La V.A. Discreta: X=Número de contenedores que no cumplen con los estándares de pureza

$$X \sim Hiper(N = 15, n = 3, A = 2)$$

Distribución de probabilidades:

$$f(x) = P(X = x) = \frac{\binom{2}{x}\binom{13}{3-x}}{\binom{15}{3}} I_{\{\textbf{máx}(0,-10),\ldots,\textbf{mín}(3,2)\}}(x)$$

La cantidad media de contenedores que no cumplen con los estándares de pureza:

$$\mu_X = E(X) = 3 \times \frac{2}{15} = 0.4$$

Su varianza:

$$\sigma_X^2 = V(X) = 3 \times \frac{2}{15} \times \left(1 - \frac{2}{15}\right) \times \left(\frac{15-3}{15-1}\right) = 0.0248$$

La probabilidad de encontrar 1 contenedor que no cumpla con los estándares de pureza:

$$P(X=1) = \frac{\binom{2}{1}\binom{13}{2}}{\binom{15}{3}} = 0.3428571$$

```r
dhyper(x = 1, m = 2, n = 13, k = 3) # m = A, n = N-A, k = n
```

```
[1] 0.3428571
```

La probabilidad de encontrar a lo más 1 contenedor que no cumpla con los estándares de pureza:

$$P(X \leq 1) = \frac{\binom{2}{0}\binom{13}{3}}{\binom{15}{3}} + \frac{\binom{2}{1}\binom{13}{2}}{\binom{15}{3}} = 0.9714285$$

```r
dhyper(x = 0, m = 2, n = 13, k = 3) + dhyper(x = 1, m = 2, n = 13, k = 3)
```

```
[1] 0.9714286
```

```r
phyper(q = 1, m = 2, n = 13, k = 3)
```

```
[1] 0.9714286
```

El percentil 65 de la cantidad de contenedores que no cumple con los estándares de pureza es:

```r
qhyper(p = 0.65, m = 2, n = 13, k = 3)
```

```
[1] 1
```

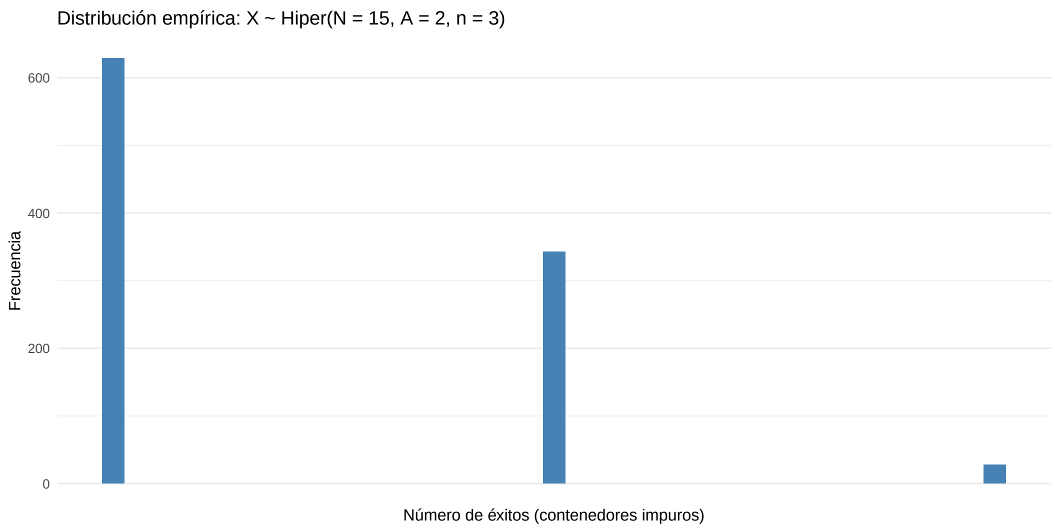
Es decir en al menos el 65% de las inspecciones se encontrará como máximo un contenedor que no cumple con los estándares de pureza.

Simulación de una muestra aleatoria de la cantidad de contenedores que no cumple con los estándares de pureza:

```
set.seed(159)
rhyper(nn = 15, m = 2, n = 13, k = 3)
```

```
 [1] 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0
```

```r
library(ggplot2)
set.seed(123)
x <- rhyper(nn = 1000, m = 2,n = 13, k = 3)
df <- as.data.frame(table(x))
colnames(df) <- c("Éxitos", "Frecuencia")
df$Éxitos <- as.numeric(as.character(df$Éxitos))
ggplot(df, aes(x = Éxitos, y = Frecuencia)) +
  geom_bar(stat = "identity", fill = "steelblue", width = .05) +
  scale_x_continuous(breaks=8:20)+
  labs(title = "Distribución empírica: X ~ Hiper(N = 15, A = 2, n = 3)",
       x = "Número de éxitos (contenedores impuros)",
       y = "Frecuencia") +
  theme_minimal()
```


Distribución empírica: X ~ Hiper(N = 15, A = 2, n = 3)
Frecuencia
600
400
200
0
Número de éxitos (contenedores impuros)

Distribución Poisson

- Distribución discreta propuesta por Siméon-Denis Poisson (1838)
- Modela la cantidad de eventos en un intervalo de tiempo, espacio o volumen, dada una frecuencia media.
- Se usa cuando los sucesos son independientes y tienen baja probabilidad individual.

Proceso de Poisson

Es un experimento aleatorio en el que ocurren sucesos en un intervalo de longitud t:

- Los sucesos son de la misma clase u homogéneos.
- Los sucesos en un intervalo son independientes de los sucesos en otros intervalos no superpuestos.
- La probabilidad de más de un suceso en un intervalo muy pequeño es despreciable.
- El promedio de sucesos (v) por unidad de intervalo (t), es denotado por $\lambda = v \times t$

Función de probabilidad

Se define la V.A.D. X=El número de sucesos que ocurren en intervalos de tamaño t. Su función de probabilidades es:

$$f(x) = P(X = x) = \begin{cases} \frac{e^{-vt}(vt)^x}{x!}, & x \in \{0, 1, 2, ...\} \\ 0, & \text{de otro modo} \end{cases}$$

v = promedio de sucesos por unidad de intervalo.
t = tamaño del intervalo (ejemplo: t = 2.3, t = 5.8, etc.).
vt = promedio de sucesos por intervalo de tamaño t (tasa de ocurrencia)
Notación: $X \sim Pois(vt)$
También se puede expresar:

$$f(x) = P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x \in \{0, 1, 2, ...\}$$

donde $\lambda = vt$, entonces $X \sim Pois(\lambda)$

Media

$$\mu_X = E(X) = \lambda$$

Varianza

$$\sigma_X^2 = Var(X) = \lambda$$

La igualdad de media y varianza es un rasgo característico que permite identificar si una variable puede ajustarse a esta distribución.

Función generatriz de probabilidades

$$G_X(s) = e^{\lambda(s-1)}$$

Función generatriz de momentos

$$M_X(t) = e^{\lambda(e^t-1)}$$

Función característica

$$\phi_X(t) = e^{\lambda(e^{it}-1)}$$

Implementación en R

- Paquete: stats
- Funciones:
  - dpois: probabilidad en un punto
  - ppois: probabilidad acumulada
  - qpois: cuantil
  - rpois: muestra aleatoria

Ejemplos

- X = Número de llamadas que recibe una central de emergencias en 1 minuto
  $X \sim Pois(4)$
- R = Número de picaduras en una persona durante una noche en una zona tropical
  $R \sim Pois(2)$
- E = Número de errores en una página de un libro impreso $E \sim Pois(0.3)$
- A = Cantidad de pacientes que llegan a emergencias en una hora $A \sim Pois(6)$

Ejemplo 5

Se cree que el número promedio de individuos por cada 2 km^2 de cierta especie de mamífero que habita en las alturas de cierta región es de 1.2.

Variable aleatoria:

V.A. X=Número de individuos en 2 km^2, $X \sim Pois(1.2)$

Distribución de probabilidades:

$$f(x) = P(X = x) = \frac{e^{-1.2} 1.2^x}{x!} \quad x \in \{0, 1, 2, ...\}$$

La cantidad media de individuos en un área de $3km^2$ es:

$$\mu_X = E(X) = 1.8$$

Su varianza:

$$\sigma_X^2 = V(X) = 1.8$$

Función generatriz de probabilidades:

$$G_X(s) = e^{1.8(s-1)}$$

Función generatriz de momentos:

$$M_X(t) = e^{1.8(e^t-1)}$$

Función característica:

$$\phi_X(t) = e^{1.8(e^{it}-1)}$$

Si se observa un área de 3 km^2 en dicha región, ¿cuál es la probabilidad de que se encuentren 3 individuos de esta especie?

Para 3 km^2: $\lambda = 3 \times 1.2/2 = 1.8$

X = Número de individuos en 3 km^2, $X \sim Pois(1.8)$

$$P(X = 3) = \frac{e^{-1.8}(1.8)^3}{3!} = 0.1606705$$

```r
dpois(x = 3, lambda = 1.8)
```

```
[1] 0.1606705
```

Si se observa un área de 3 km^2 en dicha región, ¿cuál es la probabilidad de que se encuentren como máximo 3 individuos de esta especie?

Para 3 km^2: $\lambda = 3 \times 1.2/2 = 1.8$

X = Número de individuos en 3 km^2, $X \sim Pois(1.8)$

$$P(X \leq 3) = \sum_{x=0}^{3} \frac{e^{-1.8}(1.8)^x}{x!} = \frac{e^{-1.8}(1.8)^0}{0!} + \frac{e^{-1.8}(1.8)^1}{1!} + \frac{e^{-1.8}(1.8)^2}{2!} + \frac{e^{-1.8}(1.8)^3}{3!} = 0.8913$$

```r
ppois(q = 3, lambda = 1.8)
```

```
[1] 0.8912916
```

Si se observa un área de 3 km^2 en dicha región, ¿cuál es la probabilidad de que se encuentren más de 3 individuos de esta especie?

Para 3 km^2: $\lambda = 3 \times 1.2/2 = 1.8$

X = Número de individuos en 3 km^2, $X \sim Pois(1.8)$

$$P(X > 3) = 1 - P(X \leq 3) = 1 - \sum_{x=0}^{3} \frac{e^{-1.8}(1.8)^x}{x!} = 1 - 0.8912 = 0.1087$$

```
1 - ppois(q = 3, lambda = 1.8)
```

```
[1] 0.1087084
```

```
ppois(q = 3, lambda = 1.8, lower.tail = FALSE)
```

```
[1] 0.1087084
```

El percentil 18 de la cantidad de individuos en un área de 3 km^2 es:

```
qpois(p = 0.18, lambda = 3)
```

```
[1] 1
```

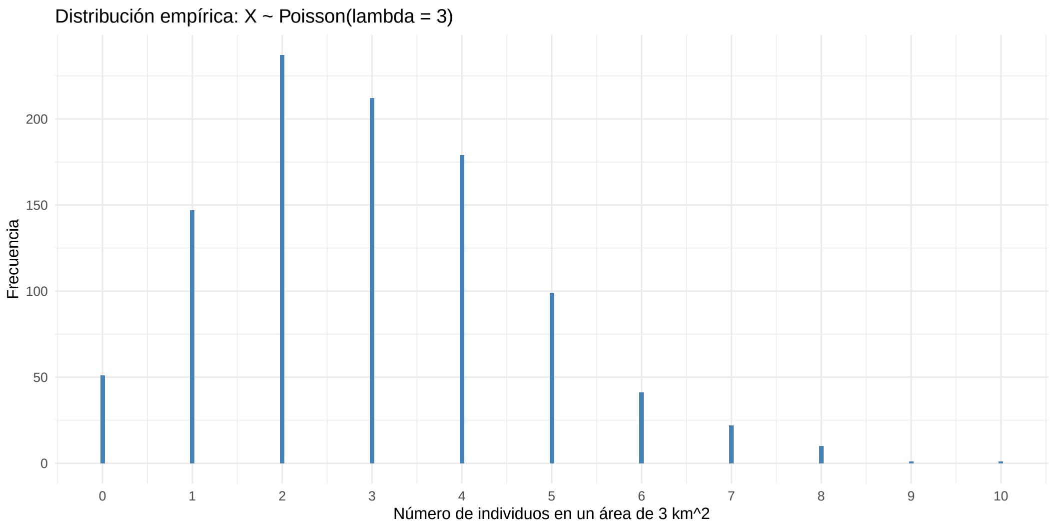
Es decir en al menos el 18% de las áreas de 3 km^2 se ecuentra como máximo un individuo.

Simulación de una muestra aleatoria de la cantidad de individuos en áreas de 3 km^2 es:

```r
set.seed(555)
rpois(n = 12, lambda = 3)
```

```
 [1] 2 5 4 3 3 0 7 2 0 5 4 2
```

```r
library(ggplot2)
set.seed(123)
x <- rpois(n = 1000, lambda = 3)
df <- as.data.frame(table(x))
colnames(df) <- c("Sucesos", "Frecuencia")
df$Sucesos <- as.numeric(as.character(df$Sucesos))
ggplot(df, aes(x = Sucesos, y = Frecuencia)) +
  geom_bar(stat = "identity", fill = "steelblue", width = .05) +
  scale_x_continuous(breaks=0:10)+
  labs(title = "Distribución empírica: X ~ Poisson(lambda = 3)",
       x = "Número de individuos en un área de 3 km^2",
       y = "Frecuencia") +
  theme_minimal()
```


Distribución empírica: X ~ Poisson(lambda = 3)
Frecuencia
200
150
100
50
0
0
1
2
3
4
5
6
7
8
9
10
Número de individuos en un área de 3 km^2

La distibución Poisson surge como límite de una binomial cuando $n \to \infty$, $\pi \to 0$ y $\lambda = n\pi$ es constante.

Por ejemplo, si $X \sim Bin(n = 2000, \pi = 0.001)$, entonces $X \to Y$, donde $Y \sim Pois(2)$

$$P(X = 3) = P(Y = 3) = 0.18$$

```
dbinom(x = 3, size = 2000, prob = 0.001)
```

```
[1] 0.1805373
```

```
dpois(x = 3, lambda = 2)
```

```
[1] 0.180447
```

Distribución Geométrica

- Distribución discreta que modela el número de intentos hasta obtener el primer éxito en una secuencia de ensayos de Bernoulli independientes.
- Fue propuesta en el contexto de los primeros estudios de probabilidad con dados y juegos de azar, y es fundamental para modelar tiempos de espera discretos.
- Se aplica cuando los ensayos son independientes, con dos posibles resultados (éxito o fracaso) y una probabilidad constante de éxito.

Función de probabilidad

Se define la V.A.D. X = el número de ensayos necesarios antes del primer éxito (es decir el número de fracasos). Su función de probabilidad es:

$$f(x) = P(X = x) = \pi(1-\pi)^x I_{\{0,1,2,3,\ldots\}}(x)$$

Notación: $X \sim Geom(\pi)$

Propiedad de falta de memoria

$$P(X > s + t | X > s) = P(X > t)$$

Media

$$\mu_X = E(X) = \frac{1-\pi}{\pi}$$

Varianza

$$\sigma_X^2 = V(X) = \frac{1-\pi}{\pi^2}$$

Asimetría

La distribución geométrica no es simétrica, y se sesga a la derecha, especialmente cuando π es pequeña.

Función generatriz de probabilidades

$$G_X(s) = \frac{\pi s}{1-(1-\pi)s}, \quad |s| < \frac{1}{1-\pi}$$

Función generatriz de momentos

$$M_X(t) = \frac{\pi e^t}{1-(1-\pi)e^t}$$

Función característica

$$\phi_X(t) = \frac{\pi e^{it}}{1-(1-\pi)e^{it}}$$

Implementación en R

- Paquete: stats
- Funciones:
  - dgeom: probabilidad en un punto
  - pgeom: probabilidad acumulada
  - qgeom: cuantil
  - rgeom: muestra aleatoria

Ejemplos

- X $=$ Número de intentos antes de que un cliente realice su primera compra en una tienda online $\rightarrow X \sim Geom(0.2)$
- D $=$ Número de inspecciones antes de encontrar el primer producto con algún defecto $\rightarrow D \sim Geom(0.05)$
- E $=$ Cantidad de intentos antes de que un estudiante resuelva correctamente un ejercicio sin ayuda $\rightarrow E \sim Geom(0.1)$

Ejemplo 6

Se estima que el 15% de los correos enviados por una empresa reciben una respuesta.

Variable aleatoria

Éxito = Recibir una respuesta

La V.A. Discreta: X= Número de correos enviados antes de recibir la primera respuesta

$$X \sim Geom(\pi = 0.15)$$

Distribución de probabilidades

$$f(x) = P(X = x) = 0.85^{x-1}0.15 \quad x \in \{1, 2, 3, ...\}$$

La cantidad media de correos enviados antes de recibir la primera respuesta es:

$$\mu_X = E(X) = \frac{1-0.15}{0.15} = 5.67$$

Su varianza:

$$\sigma_X^2 = V(X) = \frac{1-0.15}{0.15^2} = 37.778$$

Función generatriz de probabilidades:

$$G_X(s) = \frac{0.15s}{1-0.85s}, \quad |s| < \frac{1}{0.15}$$

Función generatriz de momentos:

$$M_X(t) = \frac{0.15e^t}{1-0.85e^t}$$

Función característica:

$$\phi_X(t) = \frac{0.15e^{it}}{1-0.85e^{it}}$$

Probabilidad de que se reciba la primera respuesta en el tercer correo enviado:

$$P(X = 2) = 0.85^2 \times 0.15 = 0.108375$$

```r
dgeom(x = 2, prob = 0.15)
```

```
[1] 0.108375
```

Probabilidad de que se reciba la primera respuesta a más tardar el tercer correo enviado:

$$P(X \leq 2) = 0.85^0 \times 0.15 + 0.85^1 \times 0.15 + 0.85^2 \times 0.15 = 0.108375$$

```r
dgeom(x = 0, prob = 0.15) + dgeom(x = 1, prob = 0.15) +
  dgeom(x = 2, prob = 0.15)
```

```
[1] 0.385875
```

```r
pgeom(q = 2, prob = 0.15)
```

```
[1] 0.385875
```

Probabilidad de que se requieran más de 5 intentos para lograr obtener una respuesta:

$$P(X \geq 5) = \sum_{i=5}^{\infty} P(X = i) = P(X = 5) + P(X = 6) + ... = 0.4437$$

```r
pgeom(q = 4, prob = 0.15, lower.tail=FALSE)
```

```
[1] 0.4437053
```

```r
1-pgeom(q = 4, prob = 0.15)
```

```
[1] 0.4437053
```

Primer cuartil de la cantidad de intentos antes de recibir la primera respuesta

```r
qgeom(p = 0.25, prob = 0.15)
```

```
[1] 1
```

Esto significa que en al menos el 25% de los casos se tiene como máximo un intento fallido antes de recibir la respuesta por correo.

Simulación de una muestra aleatoria de la cantidad de correos enviados antes de recibir la respuesta.

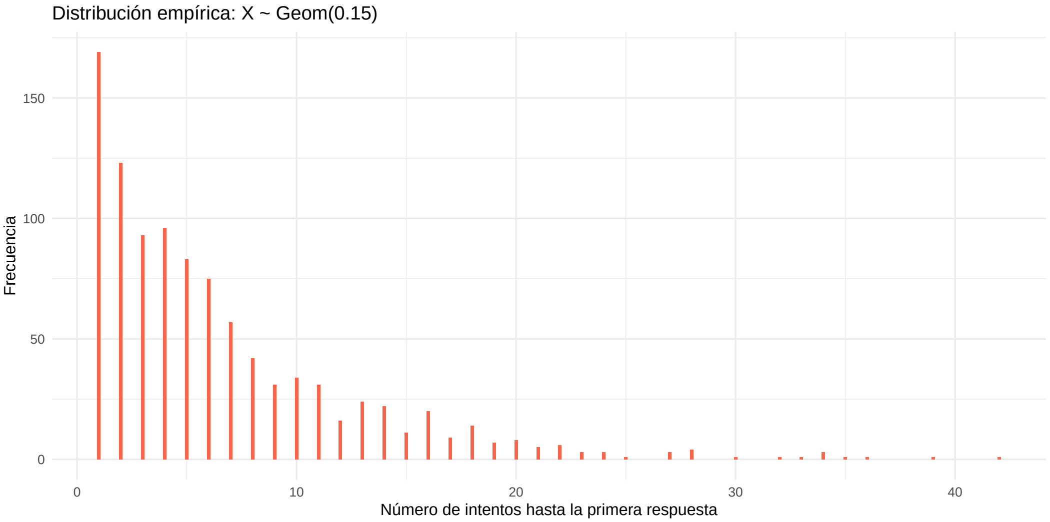
```r
set.seed(2025)
rgeom(n = 10, prob = 0.15)
```

```
 [1]  2  0  3  0  2  4  1  8 19  2
```

El primer correo se tuvo que enviar dos veces y se obtuvo respuesta en el tercer envío, el segundo correo obtuvo respuesta en el envío original, el tercer correo obtuvo respuesta en el cuarto correo enviado (3 correos fallidos previos), etc.

```r
library(ggplot2)
set.seed(2025)
x <- rgeom(1000, prob = 0.15)
df <- as.data.frame(table(x))
df$x <- as.numeric(as.character(df$x))

ggplot(df, aes(x = x, y = Freq)) +
  geom_bar(stat = "identity", fill = "tomato", width = 0.15) +
  labs(title = "Distribución empírica: X ~ Geom(0.15)",
       x = "Número de intentos antes de la primera respuesta",
       y = "Frecuencia") +
  theme_minimal()
```


Distribución empírica: X ~ Geom(0.15)
Frecuencia
150
100
50
0
0
10
20
30
40
Número de intentos hasta la primera respuesta

Distribución Binomial negativa

- Distribución discreta que modela el número de fracasos antes de alcanzar un número fijo de éxitos en una secuencia de ensayos de Bernoulli independientes.
- Se aplica cuando se desea conocer cuántos intentos fallidos ocurren antes de lograr el r-ésimo éxito, con una probabilidad constante de éxito π
- Es una generalización de la distribución geométrica (que corresponde al caso particular donde $r = 1$).
- También es conocida como distribución de Pascal.

Función de probabilidad

Se define la V.A.D. X = número de fracasos antes del r-ésimo éxito. Su función de probabilidad es:

$$f(x) = P(X = x) = \binom{x + r - 1}{x} (1 - \pi)^x \pi^r I_{\{0,1,2,3,\ldots\}}(x)$$

Notación: $X \sim BinNeg(r, \pi)$
Si $r = 1$, se trata de una $Geom(\pi)$

Media

$$\mu_X = E(X) = r \times \frac{1-\pi}{\pi}$$

Varianza

$$\sigma_X^2 = V(X) = r \times \frac{1-\pi}{\pi^2}$$

Función generatriz de probabilidades

$$G_X(s) = \left(\frac{\pi s}{1-(1-\pi)s} \right)^r, \quad |s| < \frac{1}{1-\pi}$$

Función generatriz de momentos

$$M_X(t) = \left(\frac{\pi e^t}{1-(1-\pi)e^t} \right)^r$$

Función característica

$$\phi_X(t) = \left(\frac{\pi e^{it}}{1-(1-\pi)e^{it}} \right)^r$$

Implementación en R

- Paquete: stats
- Funciones:
  - dnbinom: probabilidad en un punto
  - pnbinom: probabilidad acumulada
  - qnbinom: cuantil
  - rnbinom: muestra aleatoria

Ejemplos

- X $=$ Número de intentos fallidos antes de que un estudiante acierte 3 respuestas correctas en una trivia $\rightarrow X \sim BinNeg(r = 3, \pi = 0.25)$
- V $=$ Número de ventas fallidas antes de conseguir 5 ventas exitosas $\rightarrow V \sim BinNeg(r = 5, \pi = 0.18)$
- R $=$ Número de fallas que ocurren antes de que un robot logre completar correctamente 6 ensamblajes exitosos $R \sim BinNeg(r = 6, \pi = 0.3)$

Ejemplo 7

Un agente de ventas logra cerrar un trato con probabilidad 0.2. Se desea saber cuántos intentos fallidos ocurren, en promedio, antes de cerrar 3 ventas exitosas.

Variable aleatoria

Éxito = Cerrar una venta

V.A. X = Número de intentos fallidos antes de lograr 3 ventas

$$X \sim BinNeg(r = 3, \pi = 0.2)$$

Distribución de probabilidad

$$f(x) = P(X = x) = \binom{x+2}{x} 0.8^x 0.2^3 \qquad x \in \{0, 1, 2, 3, ...\}$$

La cantidad media de intentos fallidos antes de lograr las 3 ventas es

$$\mu_X = E(X) = 3 \times \frac{1 - 0.2}{0.2} = 12$$

Su varianza:

$$\sigma_X^2 = V(X) = 3 \times \frac{1 - 0.2}{0.2^2} = 60$$

Función generatriz de probabilidades:

$$G_X(s) = \left(\frac{0.2s}{1-0.8s}\right)^3, \quad |s| < \frac{1}{0.8}$$

Función generatriz de momentos:

$$M_X(t) = \left(\frac{0.2e^t}{1-0.8e^t}\right)^3$$

Función característica:

$$\phi_X(t) = \left(\frac{0.2e^{it}}{1-0.8e^{it}}\right)^3$$

Probabilidad de que ocurran exactamente 5 intentos fallidos antes de cerrar exitosamente 3 ventas:

$$P(X=5) = \binom{5+2}{5} \times 0.8^5 \times 0.2^3 = 0.05505024$$

```r
dnbinom(x = 5, size = 3, prob = 0.2)
```

```
[1] 0.05505024
```

Probabilidad de que ocurran como máximo 5 intentos fallidos antes de cerrar exitosamente 3 ventas:

$$P(X \leq 5) = \sum_{x=0}^{5} \binom{x+2}{x} \times 0.8^x \times 0.2^3 = 0.20308$$

```r
dnbinom(x=0, size=3, prob=0.2) + dnbinom(x=1, size=3, prob=0.2) +
  dnbinom(x=2, size=3, prob=0.2) + dnbinom(x=3, size=3, prob=0.2) +
  dnbinom(x=4, size=3, prob=0.2) + dnbinom(x=5, size=3, prob=0.2)
```

```
[1] 0.2030822
```

```r
pnbinom(q = 5, size = 3, prob = 0.2)
```

```
[1] 0.2030822
```

Tercer cuartil de la cantidad de fracasos antes de lograr 3 ventas

```
qnbinom(p = 0.75, size = 3, prob = 0.2)
```

```
[1] 16
```

En al menos el 75% de los casos, el vendedor fallará como máximo 16 veces antes de cerrar existosamente 3 ventas.

Simulación de una muestra aleatoria de 10 casos de vendedores:

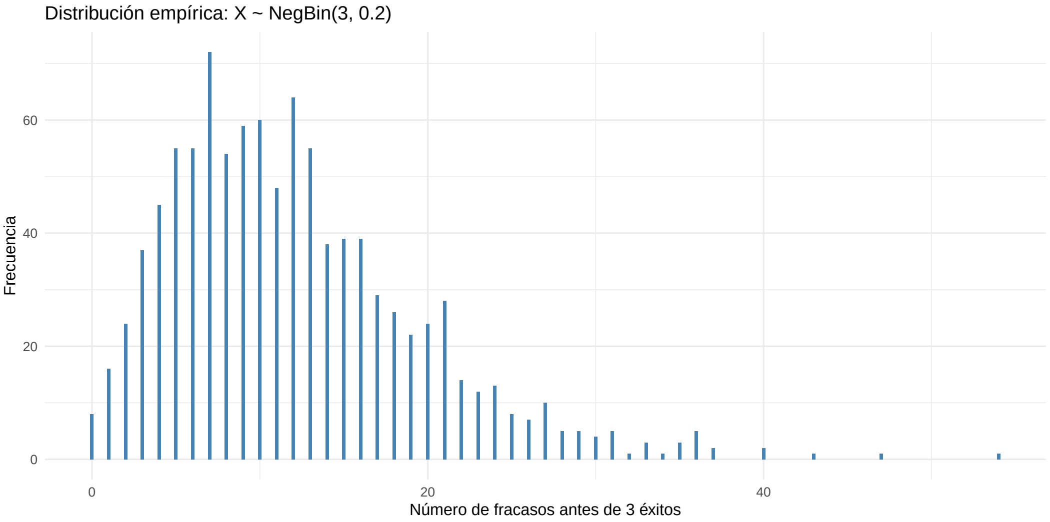
```r
set.seed(75321)
rnbinom(n = 10, size = 3, prob = 0.2)
```

```
 [1] 16 13 12 17 31  2 17 24  9 14
```

El primer vendedor falla 16 veces antes de conseguir su tercera venta exitosa, el segundo falla 13 veces antes de las 3 ventas exitosas, etc.

```r
library(ggplot2)
set.seed(75321)
x <- rnbinom(1000, size = 3, prob = 0.2)
df <- as.data.frame(table(x))
df$x <- as.numeric(as.character(df$x))

ggplot(df, aes(x = x, y = Freq)) +
  geom_bar(stat = "identity", fill = "steelblue", width = 0.2) +
  labs(title = "Distribución empírica: X ~ NegBin(3, 0.2)",
       x = "Número de fracasos antes de 3 éxitos",
       y = "Frecuencia") +
  theme_minimal()
```


Distribución empírica: X ~ NegBin(3, 0.2)
Frecuencia
0
20
40
60
0
20
40
Número de fracasos antes de 3 éxitos

Distribución zeta

Una variable aleatoria discreta X tiene distribución Zeta si toma valores enteros positivos y su función de probabilidad decrece según una ley de potencia. Es una distribución de cola pesada.

Función de probabilidad

$$f(x) = P(X = x) = \frac{1}{\zeta(s+1)x^{s+1}} I_{\{1,2,3,\ldots\}}(x), \quad s > 0$$

donde:

- s es el parámetro de forma
- $\zeta(s) = \sum_{k=1}^{\infty} \frac{1}{k^s}$ es la función zeta de Riemann.

Notación: $X \sim Zeta(s)$

Media

$$E(X) = \frac{\zeta(s)}{\zeta(s+1)} \qquad s > 3$$

Si $0 < s \leq 1$, $E(X) = \infty$

Varianza

$$V(X) = \frac{\zeta(s-1)}{\zeta(s+1)} - \left(\frac{\zeta(s)}{\zeta(s+1)}\right)^2 \qquad s > 2$$

Si $0 < s \leq 2$, $V(X) = \infty$

Implementación en R

- Paquete: VGAM
- Funciones:
  - zeta: Función zeta
  - dzeta: probabilidad en un punto
  - pzeta: probabilidad acumulada
  - qzeta: cuantil

Ejemplos

- En el estudio de frecuencia de palabras en un texto, X es el rango de frecuencia para una palabra elegida al azar, entonces $X \sim Zeta(s = 1.1)$
- En una plataforma de videos como YouTube, muchos videos tienen pocas vistas, y pocos acumula muchísimas visualizaciones. X es el rango de popularidad de un video elegido al azar, entonces $X \sim Zeta(s = 2)$

Ejemplo 8

En un servidor web, hay muchos archivos pequeños y pocos archivos muy grandes.

Variable aleatoria

V.A. X = Tamaño ordenado por rango

$$X \sim Zeta(s = 4.1)$$

Distribución de probabilidad

$$f(x) = P(X = x) = \frac{1}{\zeta(5.1)x^{5.1}} I_{\{1,2,3,\ldots\}}(x)$$

El rango medio:

$$\mu_X = \frac{\zeta(4.1)}{\zeta(5.1)}$$

```
library(VGAM)
zeta(4.1)/zeta(5.1)
```

```
[1] 1.040186
```

```
x  <- 1:100
fx <- dzeta(1:100, shape = 4.1)
sum(x*fx)
```

```
[1] 1.040186
```

Su varianza:

$$V(X) = \frac{\zeta(3.1)}{\zeta(5.1)} - \left(\frac{\zeta(4.1)}{\zeta(5.1)}\right)^2$$

```
zeta(3.1)/zeta(5.1) - (zeta(4.1)/zeta(5.1))**2
```

```
[1] 0.06227857
```

```
x  <- 1:100
fx <- dzeta(1:100, shape = 4.1)
sum(x**2*fx) - (sum(x*fx))**2
```

```
[1] 0.06225022
```

Probabilidad de seleccionar el segundo archivo más grande

$$P(X=2) = \frac{1}{2^{5.1}} * \frac{1}{\zeta(5.1)} = 0.054$$

```
dzeta(x = 2, shape = 4.1)
```

```
[1] 0.02819347
```

```
1/2**5.1 * 1/zeta(5.1)
```

```
[1] 0.02819347
```

Probabilidad de seleccionar el primero, segundo, tercero, cuarto y quinto archivos más grandes

```
dzeta(x = 1:5, shape = 4.1) |> round(3)
```

```
[1] 0.967 0.028 0.004 0.001 0.000
```

Probabilidad de seleccionar hasta el tercer archivo más grande:

$$P(X \leq 3) = 0.994$$

```r
pzeta(q = 3, shape = 4.1)
```

```
[1] 0.9987032
```

```r
dzeta(x = 1:3, shape = 4.1) |> sum()
```

```
[1] 0.9987032
```

```r
1/(1**5.1*zeta(5.1)) + 1/(2**5.1*zeta(5.1)) + 1/(3**5.1*zeta(5.1))
```

```
[1] 0.9987032
```

Percentil 98 del rango de elección de archivos

```r
qzeta(p = 0.98, shape = 4.1)
```

```
[1] 2
```

Ejemplo de simulación aleatoria.

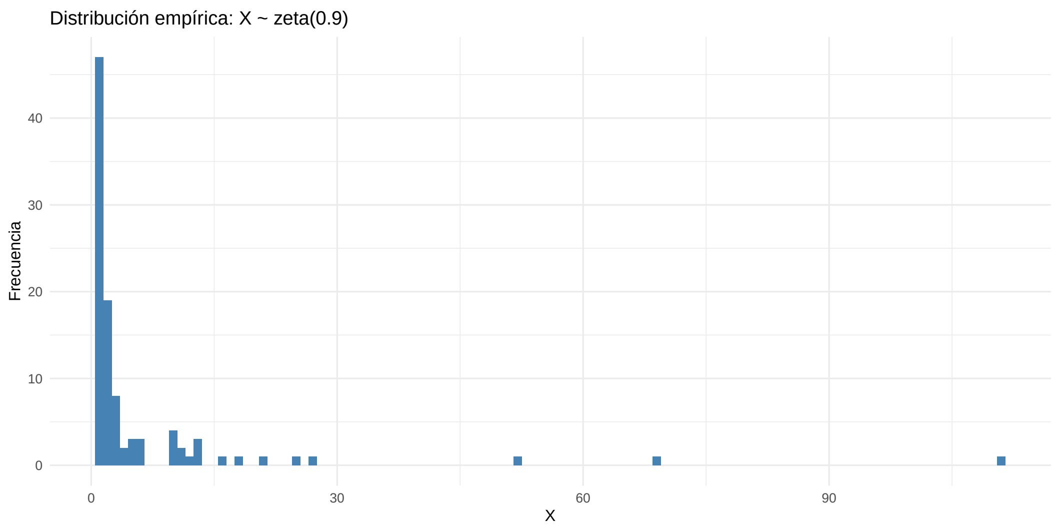
Supongamos que deseamos simular 12 datos de $Zeta(s = 0.9)$

```r
rzeta(n = 12, shape = 0.9)
```

```
 [1] 1 1 2 1 1 1 1 7 2 1 1 1
```

```r
library(ggplot2)
set.seed(21)
x <- rzeta(n = 100, shape = 0.9)
df <- as.data.frame(table(x))
df$x <- as.numeric(as.character(df$x))

ggplot(df, aes(x = x, y = Freq)) +
  geom_bar(stat = "identity", fill = "steelblue", width = 1) +
  labs(title = "Distribución empírica: X ~ zeta(0.9)",
       x = "X",
       y = "Frecuencia") +
  theme_minimal()
```


Distribución empírica: X ~ zeta(0.9)
Frecuencia
40
30
20
10
0
0
30
60
90
X