Lectura 3

Variables aleatorias y distribución de probabilidad

Variables aleatorias

-> Variable aleatoria continua: Es una variable que puede tomar cualquiera de los infinitos valores posibles dentro de un intervalo

-> Variable aleatoria discreta: Sólo puede tomar un conjunto finito de valores (Lanzamiento de un dado)

El **valor esperado**, denotado como $E(X)$ o μ, corresponde al resultado promedio de una variable aleatoria. Para una variable aleatoria discreta, se calcula sumando los valores posibles ponderados por su probabilidad, como muestra la ecuación 3.1.

$$E(X) \equiv \mu = \sum_{i=1}^{n} x_i P(X = x_i) \tag{3.1}$$

También podemos calcular qué tan alejado podría estar el valor obtenido del valor esperado por medio de la varianza general, denotada por $Var(X)$ o σ^2, que se calcula como la media de los cuadrados de la diferencia con respecto a la media ponderada según la probabilidad de ocurrencia, como muestra la ecuación 3.2. Una vez más, la desviación estándar corresponde a la raíz cuadrada de la varianza.

$$Var(X) \equiv \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i) \tag{3.2}$$

Script 3.1: variables aleatorias discretas en R.

```r
library(discreteRV)

# Crear una variable discreta para representar el dado
# adulterado de la tabla 3.1.
resultados <- 1:6
probabilidades = c(0.25, 0.125, 0.125, 0.125, 0.125, 0.25)
X <- RV(outcomes = resultados, probs = probabilidades)

# Calcular el valor esperado.
esperado <- E(X)
cat("Valor esperado:", esperado, "\n")

# Calcular la varianza.
varianza <- V(X)
cat("Varianza:", varianza, "\n")

# Calcular la desviación estándar.
desviacion <- SD(X)
cat("Desviación estándar:", desviacion, "\n")
```

Conocer la distribución de probabilidad de una variable discreta nos ayuda a hacer estimaciones útiles

Cuando las variables de una combinación lineal son independientes[1], podemos calcular el valor esperado y la varianza de la combinación lineal usando las ecuaciones 3.5 y 3.6. Una vez más, la desviación estándar está dada por la raíz cuadrada de la varianza.

$$E\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i E(X_i) \tag{3.5}$$

$$Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 Var(X_i) \tag{3.6}$$

Por supuesto, en R también podemos trabajar con combinaciones lineales de variables aleatorias discretas, como muestra el script 3.3.

Script 3.3: combinación lineal de variables aleatorias discretas en R.

```r
library(discreteRV)

# Crear una variable discreta para representar el dado adulterado de la tabla
# 3.1, y calcular su valor esperado, varianza y desviación estándar.
resultados <- 1:6
probabilidades = c(0.25, 0.125, 0.125, 0.125, 0.125, 0.25)
X <- RV(outcomes = resultados, probs = probabilidades)
esperado_x <- E(X)
varianza_x <- V(X)
desviacion_x <- SD(X)
cat("E(X):", esperado_x, "\n")
cat("V(X):", varianza_x, "\n")
cat("SD(X):", desviacion_x, "\n\n")

# Crear una variable aleatoria para un dado balanceado, y calcular su valor
# esperado, varianza y desviación estándar.
Y <- RV(outcomes = resultados, probs = 1/6)
```

[1]Si las variables no son independientes, se requieren métodos más complejos fuera del alcance de este libro.

```r
esperado_y <- E(Y)
varianza_y <- V(Y)
desviacion_y <- SD(Y)
cat("E(Y):", esperado_y, "\n")
cat("V(Y):", varianza_y, "\n")
cat("SD(Y):", desviacion_y, "\n\n")

# Crear una combinación lineal de variables aleatorias, y calcular su valor
# esperado, varianza y desviación estándar.
Z <- 0.5 * X + 0.5 * Y
esperado_z <- E(Z)
varianza_z <- V(Z)
desviacion_z <- SD(Z)
cat("E(Z):", esperado_z, "\n")
cat("V(Z):", varianza_z, "\n")
cat("SD(Z):", desviacion_z)
```

Cuando el histograma se asemeja cada vez más a una curva continua, esta recibe el nombre de **función de densidad de probabilidad** o simplemente **distribución o densidad**

-> El área bajo la curva siempre es 1

Distribuciones continuas

Distribución Normal

Así, denotamos este tipo de distribución por $N(\mu, \sigma)$. La figura 3.4, creada mediante el script 3.4, muestra dos ejemplos superpuestos de distribución normal: $N(\mu = 0, \sigma = 1)$ en azul y $N(\mu = 10, \sigma = 6)$ en rojo.

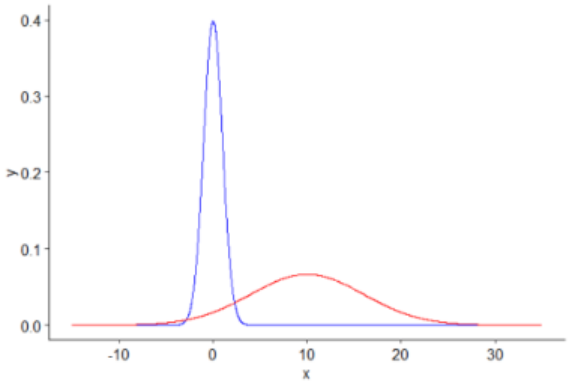

Figura 3.4: dos ejemplos superpuestos de distribución normal.

Script 3.4: graficando dos ejemplos de distribución normal.

```r
library(ggpubr)

# Generar valores para una distribución normal con media 0 y
# desviación estándar 1.
media <- 0
desv_est <- 1
x <- seq(-15, 35, 0.01)
y <- dnorm(x, mean = media, sd = desv_est)
normal_1 <- data.frame(x, y)

# Repetir el proceso para una distribución normal con media 10
# y desviación estándar 6.
media <- 10
desv_est <- 6
x <- seq(-15, 35, 0.01)
y <- dnorm(x, mean = media, sd = desv_est)
normal_2 <- data.frame(x, y)

# Graficar ambas distribuciones.
g <- ggplot(normal_1, aes(x, y)) + geom_line(color = "blue")
g <- g + geom_line(data = normal_2, color = "red")
g <- g + theme_pubr()

print(g)
```

-> Unimodal y simétrica

->Distribución Gaussiana (Forma de campana)

Una **regla empírica** muy útil al momento de trabajar con distribuciones normales es la llamada regla 68-95-99.7, ilustrada en la figura 3.5, la cual establece que:

- Cerca de 68 % de las observaciones se encuentran a una distancia de una desviación estándar de la media.
- Alrededor de 95 % de las observaciones se encuentran a una distancia de dos desviación estándar de la media.
- Aproximadamente 99.7 % de las observaciones se encuentran a una distancia de tres desviación estándar de la media.

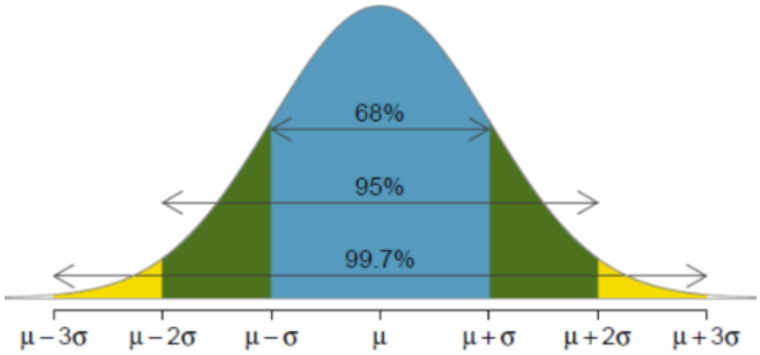

Figura 3.5: regla empírica de la distribución normal. Fuente: Diez y col. (2017, p. 136).

Gráfico Q-Q

Se debe verificar que el supuesto de una distribución normal sea aceptable , por lo tanto usamos el gráfico cuantil- cuantil (Q-Q).

-> Puntos observaciones

->Recta representa distribución normal

-> Mientras más se asemeje el patrón que forman los puntos a la recta, más parecida será la distribución normal

-> La banda coloreada establece el margen aceptable para suponer normalidad e el conjunto de datos

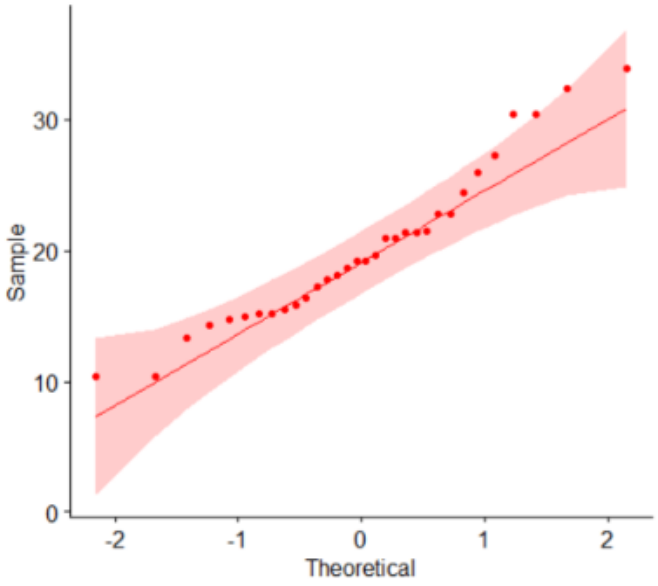

Figura 3.6: gráfico cuantil-cuantil.

Script 3.5: creación de un gráfico cuantil-cuantil.

```
library(ggpubr)

# Cargar datos.
datos <- read.csv2("C:/Inferencia/Mtcars.csv", stringsAsFactors = TRUE,
                   row.names = 1)

# Gráfico Q-Q para la variable Rendimiento.
g <- ggqqplot(datos,
              x = "Rendimiento",
              color = "red")

print(g)
```

Distribución Z

Al trabajar con distribuciones, especialmente las simétricas, a menudo usaremos **técnicas de estandarización** para determinar qué tan usual o inusual es un determinado valor en una escala única. Así, para la distribución normal usamos como estandarización la **distribución Z** o **distribución normal estándar**, que no es más que una distribución normal centrada en 0 y con desviación estándar 1, que podemos obtener de manera bastante sencilla como muestra la ecuación 3.7.

$$Z = \frac{x - \mu}{\sigma} \tag{3.7}$$

Al aplicar la ecuación 3.7 a una observación x en una distribución normal obtenemos, entonces, su **valor**

z, que determina cuán por encima o por debajo de la media (en términos de la desviación estándar) se encuentra dicha observación x. Así, observaciones cuyos valores z sean negativos estarán por debajo de la media. Análogamente, un valor Z positivo indica que la observación está por sobre la media. Mientras mayor sea el valor absoluto de su valor z ($|z|$), más inusual será la observación.

Distribución chi-cuadrado

También llamada **ji-cuadrado** o χ^2, la distribución **chi-cuadrado** (Devore, 2008) se usa para caracterizar valores siempre positivos y habitualmente desviados a la derecha. El único parámetro de esta distribución corresponde a los **grados de libertad**, usualmente representada por la letra griega ν, que son una estimación de la cantidad de observaciones empleadas para calcular un estimador. Otra forma de entender esta idea es como la cantidad de valores que pueden cambiar libremente en un conjunto de datos. Como ejemplo, supongamos que necesitamos una muestra de tres elementos cuya media sea 10. Una vez escogidos los primeros dos, solo queda una posibilidad para el tercero de modo que se cumpla con la media deseada. Así, solo los dos primeros valores pueden cambiar libremente, por lo que se tienen dos grados de libertad.

Esta distribución está relacionada con la ya conocida distribución Z, pues si sumamos los cuadrados de k variables aleatorias independientes que siguen una distribución Z, dicha suma sigue una distribución χ^2 con k grados de libertad:

$$\sum_{i=1}^{k} Z_i^2 \sim \chi^2(\nu = k) \tag{3.8}$$

La media de la distribución χ^2 es $\mu = \nu$, y su desviación estándar, $\sigma = 2\nu$.

Las funciones de R para esta distribución, similares a las descritas para la distribución normal, son:

- `dchisq(x, df)`.
- `pchisq(q, df, lower.tail)`.
- `qchisq(p, df, lower.tail)`.
- `rchisq(n, df)`.

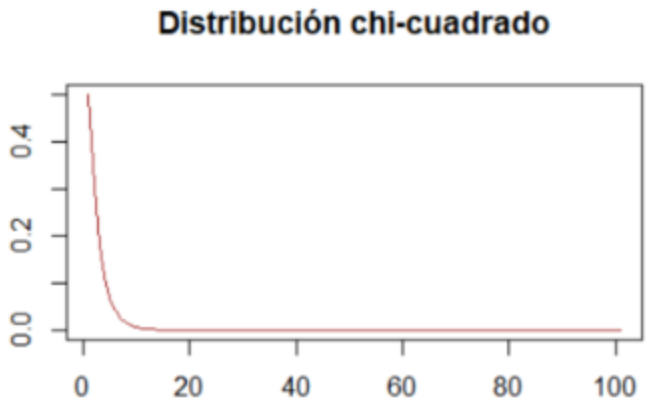

Figura 3.7: ejemplo de distribución χ^2 con 2 grados de libertad.

Distribución t de student

medida que los grados de libertad aumentan, esta distribución se asemeja cada vez más a la normal, aunque sus colas son más gruesas, como ilustra la figura 3.8.

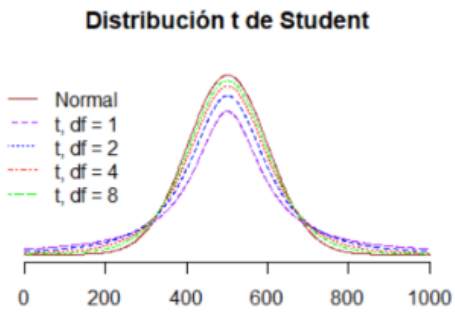

Figura 3.8: ejemplo de distribuciones t.

La distribución t se encuentra relacionada con las distribuciones vistas anteriormente de acuerdo a la ecuación 3.9, donde Z es una distribución normal estándar y $\chi^2(\nu)$ es una distribución χ^2 con ν grados de libertad.

$$Z\sqrt{\frac{v}{\chi^2(v)}} \sim t(\nu) \tag{3.9}$$

La media de la distribución t, para $\nu > 1$, es $\mu = 0$. Su desviación estándar, para $\nu > 2$, está dada por la ecuación 3.10.

$$\sigma = \sqrt{\frac{\nu}{\nu - 2}} \tag{3.10}$$

Funciones R

Distribución F

Otra distribución que usaremos a lo largo de este libro es la **distribución F**, ampliamente usada para comparar varianzas. La distribución F se relaciona con las anteriores de acuerdo a la ecuación 3.11, donde $\chi^2_1(\nu_1)$ y $\chi^2_2(\nu_2)$ son dos distribuciones χ^2 con ν_1 y ν_2 grados de libertad, respectivamente. Un ejemplo de una distribución F se puede encontrar en la figura 3.9.

$$\frac{\frac{X^2_1(\nu_1)}{\nu_1}}{\frac{X^2_2(\nu_2)}{\nu_2}} \sim F(\nu_1, \nu_2) \tag{3.11}$$

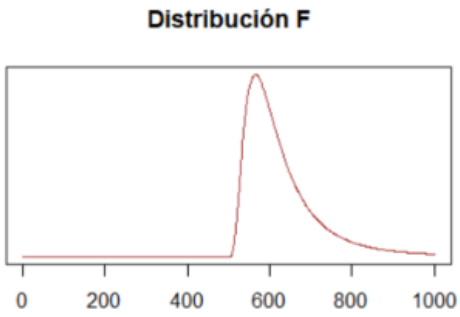

Figura 3.9: ejemplo de una distribución F.

Para $\nu_2 > 2$, la media de esta distribución está dada por la ecuación 3.12, y la desviación estándar corresponde a la ecuación 3.13 para $\nu_2 > 4$.

$$\mu = \frac{\nu_2}{\nu_2 - 2} \tag{3.12}$$

$$\sigma = \sqrt{\frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}} \tag{3.13}$$

Funciones en R

Distribución de Bernoulli

Una **variable aleatoria de Bernoulli** es aquella en que cada intento individual tiene solo dos resultados posibles: "éxito", que ocurre con una probabilidad p y se representa habitualmente con un 1, y "fracaso", que ocurre con probabilidad $q = 1 - p$ y suele representarse por un 0. La selección de qué resultado se considera como éxito o fracaso suele ser arbitraria. Para ilustrar esta idea, si dos personas lanzan una moneda al aire para sortear al ganador, cada una de ellas considerará una cara diferente de la moneda como un éxito.

Otro ejemplo que nos puede ayudar es el de lanzar varios dados de 20 caras, donde el éxito corresponda a obtener un 20 como resultado. Cada uno de ellos tiene una **probabilidad de éxito** (obtener 20) $p = 0.05$ y una **probabilidad de fracaso** (obtener otro valor) $q = 1 - p = 0.95$. Los lanzamientos de los dados son **independientes**, pues un dado no afecta a los demás.

Definimos la **proporción de la muestra** para una distribución de Bernoulli, $\hat{p}$, como la cantidad de éxitos dividida por la cantidad de intentos. Mientras mayor sea la cantidad de intentos, más cercano será el valor de $\hat{p}$ a la probabilidad real de éxito p.

Al igual que la distribución normal, la distribución de Bernoulli puede resumirse expresando su media ($\mu = p$) y su desviación estándar. Esta última está dada por la ecuación 3.14.

$$\sigma = \sqrt{p(1-p)} \tag{3.14}$$

Funciones en R

Distribución Geométrica

La **distribución geométrica** describe la cantidad de intentos que debemos realizar hasta obtener un éxito para variables de Bernoulli **independientes e idénticamente distribuidas**, es decir, que no se afectan unas a otras y cada una con igual probabilidad de éxito.

La probabilidad de obtener un éxito al n-ésimo intento está dada por la ecuación 3.15, donde podemos ver que las probabilidades en esta distribución decrecen exponencialmente rápido, como ilustra la figura 3.10. La media y la desviación estándar de la distribución geométrica están dadas, respectivamente, por las ecuaciones 3.16 y 3.17.

$$\Pr(\text{primer éxito al } n\text{-ésimo intento}) = (1-p)^{n-1}p \tag{3.15}$$

$$\mu = \frac{1}{p} \tag{3.16}$$

$$\sigma = \sqrt{\frac{1-p}{p^2}} \tag{3.17}$$

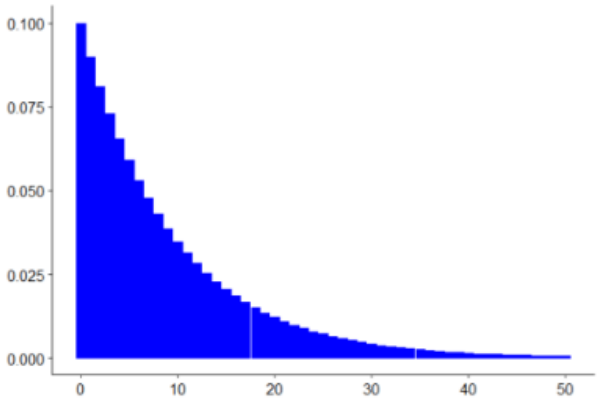

Figura 3.10: distribución geométrica para obtener un valor específico lanzando un dado de 20 caras.

Para entender mejor la utilidad de la distribución geométrica, consideremos la pregunta: ¿Cuántas veces tenemos que lanzar un dado de 20 caras para obtener un 1? Anteriormente vimos que la probabilidad de éxito para un dado es $p = 0.05$. El valor esperado, representado por la media, sería en este caso el que se presenta en la ecuación 3.18.

$$\mu = \frac{1}{p} = \frac{1}{0.05} = 20 \tag{3.18}$$

Funciones en R

Distribución de Poisson

Útil para estimar la cantidad de eventos en una población grande en un lapso de tiempo dado, por ejemplo, la cantidad de contagios de influenza entre los habitantes de Santiago en una semana, la **distribución de Poisson** (figura 3.13) tiene una función de probabilidad definida por la ecuación 3.25, donde λ es la tasa o cantidad de eventos que se espera observar en un lapso de tiempo dado y k puede tomar cualquier valor entero no negativo. La media de esta distribución está dada por λ y la desviación estándar, por $\sqrt{\lambda}$.

$$\Pr(\text{observar } k \text{ eventos}) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{3.25}$$

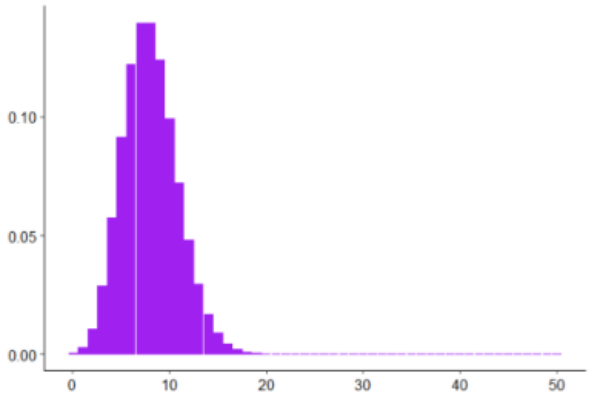

Figura 3.13: ejemplo de distribución de Poisson.

Funciones en R

Distribución Binomial

A diferencia de la distribución geométrica, la **distribución binomial** describe la probabilidad de tener exactamente k éxitos en n intentos independientes de Bernoulli con probabilidad de éxito p, cuya función de probabilidad está dada por la ecuación 3.19, donde:

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ corresponde a la cantidad de formas de obtener k éxitos en un total de n intentos.
- $p^k(1-p)^{n-k}$ es la probabilidad de tener un único éxito en solo una de las $\binom{n}{k}$ maneras posibles.

$$f(k,n,p) = \Pr(k;n,p) = \Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.19}$$

La media y la desviación estándar de la distribución binomial están dadas por las ecuaciones 3.20 y 3.21, respectivamente. Un ejemplo de esta distribución se presenta en la figura 3.11

$$\mu = np \tag{3.20}$$

$$\sigma = \sqrt{np(1-p)} \tag{3.21}$$

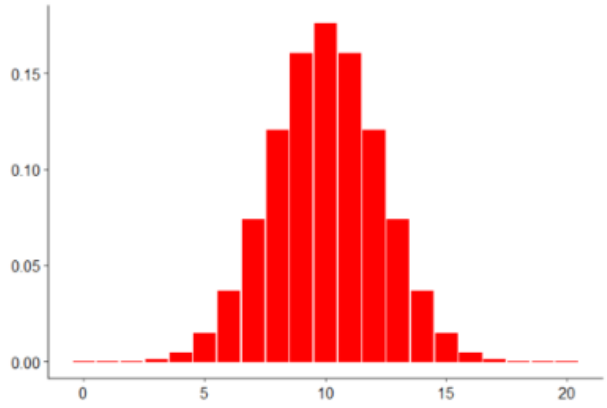

Figura 3.11: distribución binomial con $\mu = 400$ y $\sigma = 15.4019$.

Antes de decidir usar la distribución binomial, tenemos que verificar cuatro condiciones:

1. Los intentos son independientes.
2. La cantidad de intentos (n) es fija.
3. El resultado de cada intento puede ser clasificado como éxito o fracaso.
4. La probabilidad de éxito (p) es la misma para cada intento.

Funciones en R

Distribución binomial negativa

La **distribución binomial negativa** es algo más general que la binomial, pues describe la probabilidad de encontrar el k-ésimo éxito al n-ésimo intento. Como señalan Diez y col. (2017, p. 155), "en el caso binomial, en general se tiene una cantidad fija de intentos y se considera la cantidad de éxitos. En el caso binomial negativo, se examina cuántos intentos se necesitan para observar una cantidad fija de éxitos y se requiere que la última observación sea un éxito"[3].

Como adelanta la comparación anterior, antes de decidir usar la distribución binomial negativa tenemos que verificar cuatro condiciones:

1. Los intentos son independientes.
2. El resultado de cada intento puede ser clasificado como éxito o fracaso.
3. La probabilidad de éxito (p) es la misma para cada intento.
4. El último intento debe ser un éxito.

La función de probabilidad para esta distribución, ejemplificada en la figura 3.12, está dada por la ecuación 3.22. La varianza y la desviación estándar están dadas por las ecuaciones 3.23 y 3.24 (Devore, 2008, p. 120).

$$\Pr(k\text{-ésimo éxito al } n\text{-ésimo intento}) = \binom{n-1}{k-1} p^k (1-p)^{1-k} \tag{3.22}$$

$$\mu = \frac{k(1-p)}{p} \tag{3.23}$$

$$\sigma = \sqrt{\frac{k(1-p)}{p^2}} \tag{3.24}$$

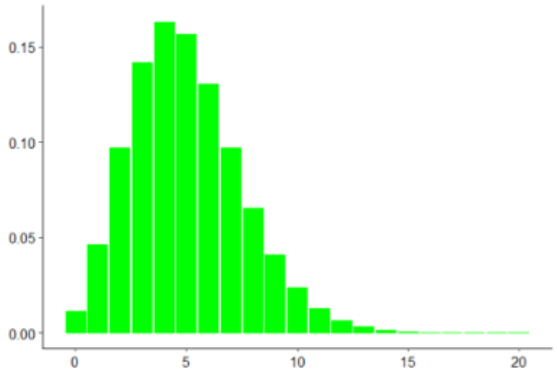

Figura 3.12: ejemplo de distribución binomial negativa.

Funciones en R