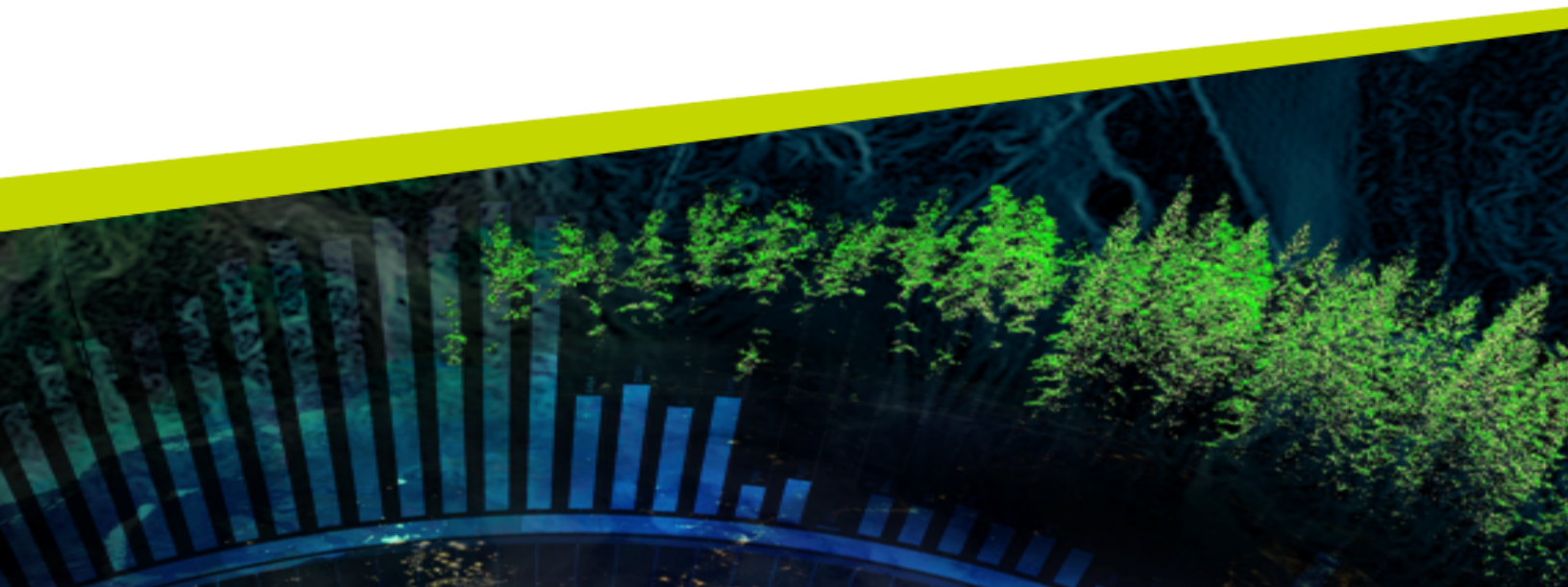
INFERENCIA Y MODELOS ESTADÍSTICOS

Jacqueline Köhler C. y José Luis Jara V.



CAPÍTULO 3. VARIABLES ALEATORIAS Y DISTRIBUCIONES DE PROBABILIDAD

Los conceptos que estudiaremos en este capítulo pueden resultar algo difíciles de entender, por lo que si necesitas más material, puedes consultar las fuentes en que se basa este capítulo: Diez y col. (2017, pp. 104-157) y Freund y Wilson (2003, pp. 104-106).

Definimos como **variable aleatoria** una variable o un proceso cuyo resultado sea numérico. Dichas variables se nombran con letras mayúsculas, y denotamos sus posibles valores por la letra minúscula correspondiente, acompañada de un subíndice. Las variables aleatorias tienen una **distribución de probabilidad**, la cual define la probabilidad de que ocurran los diferentes valores que dicha variable puede tomar.

3.1 VARIABLES ALEATORIAS

La definición de **variable aleatoria continua** es en realidad bastante sencilla: es una variable que puede tomar cualquiera de los infinitos valores posibles dentro de un intervalo.

Una **variable aleatoria discreta**, en cambio, solo puede tomar un conjunto finito de valores. Un ejemplo típico de variable aleatoria puede ser el lanzamiento de un dado. Si el dado está bien balanceado, tendremos igual probabilidad de obtener cualquiera de las caras. Pero es sabido que algunos tramposos fabrican dados adulterados para favorecer, por ejemplo, la obtención de valores 1 y 6. Una distribución aleatoria de la variable lanzamiento de un dado adulterado (X) podría ser la que se presenta en la tabla 3.1.

i	1	2	3	4	5	6	Total
x_i	1	2	3	4	5	6	-
$P(X = x_i)$	0.250	0.125	0.125	0.125	0.125	0.250	1.000

Tabla 3.1: distribución de probabilidad para el lanzamiento de un dado adulterado.

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.

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

En R, el paquete `DiscreteRV` permite trabajar con variables aleatorias discretas, como se ejemplifica en el script 3.1 (Cross, 2017).

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")
```

Para ayudarnos a entender mejor la noción de distribución de probabilidad, veamos la figura 3.1 (obtenida mediante el script 3.2). Ella nos muestra, de izquierda a derecha, las distribuciones de probabilidad para el puntaje total obtenido al lanzar 5, 10 y 20 dados, respectivamente.

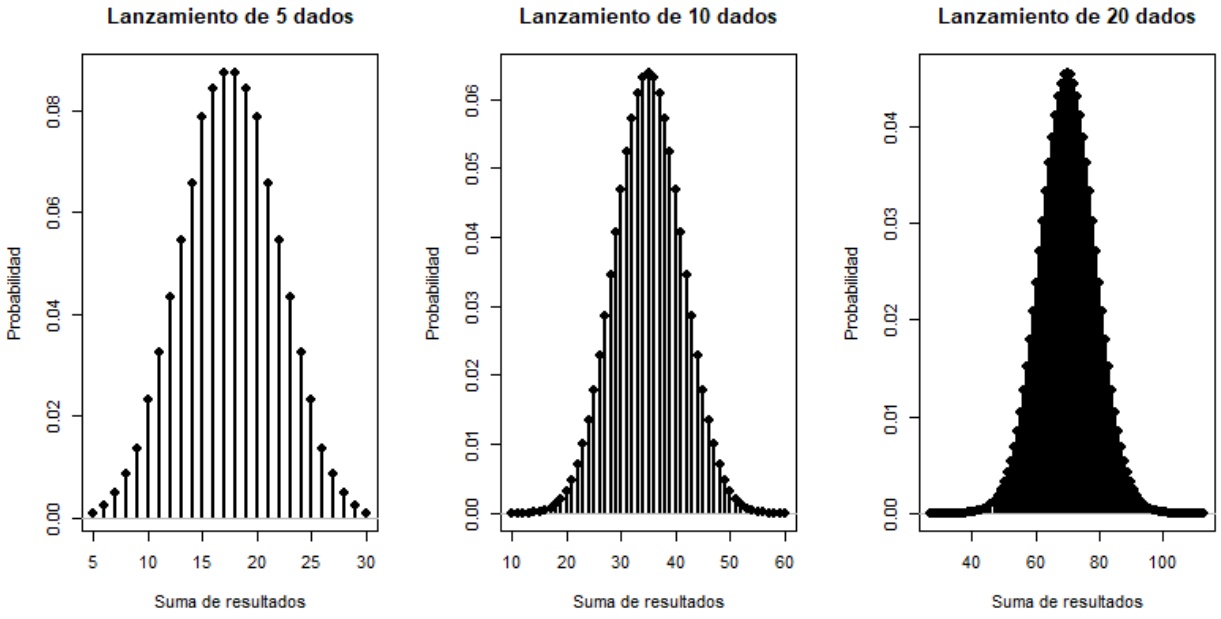


Figura 3.1: distribución de probabilidad para varios lanzamientos de un dado cargado.

Script 3.2: histogramas de variables aleatorias discretas en R.

```r
library(discreteRV)
library(ggpubr)

# Crear una variable discreta para representar el dado
# adulterado de la tabla 4.1.
resultados <- 1:6
```

```r
probabilidades = c(0.25, 0.125, 0.125, 0.125, 0.125, 0.25)
X <- RV(outcomes = resultados, probs = probabilidades)

# Crear vector con los resultados de 5 lanzamientos del dado.
lanzar_5 <- SofIID(X, n=5)

# Crear vector con los resultados de 10 lanzamientos del dado.
lanzar_10 <- SofIID(X, n=10)

# Crear vector con los resultados de 20 lanzamientos del dado.
lanzar_20 <- SofIID(X, n=20)

# Graficar los resultados.
par(mfrow=c(1, 3))

plot(lanzar_5,
     main="Lanzamiento de 5 dados",
     xlab="Suma de resultados",
     ylab="Probabilidad")

plot(lanzar_10,
     main="Lanzamiento de 10 dados",
     xlab="Suma de resultados",
     ylab="Probabilidad")

plot(lanzar_20,
     main="Lanzamiento de 20 dados",
     xlab="Suma de resultados",
     ylab="Probabilidad")
```

Conocer la distribución de probabilidad de una variable discreta nos ayuda a hacer estimaciones útiles. A modo de ejemplo, supongamos que un ingeniero de software debe crear un programa que resuelva un problema (siempre con instancias del mismo tamaño) con un tiempo de respuesta no mayor a 25 segundos. El histograma de la figura 3.2 muestra los tiempos de ejecución obtenidos para 500 pruebas de la solución propuesta, donde se observa que 30 de ellas tardaron en realidad más de 25 segundos, con un rango que va de 0 a 30 segundos. Así, podemos estimar la probabilidad de que el tiempo de ejecución sea mayor a 25 segundos dividiendo la cantidad de observaciones que cumplen este criterio por la cantidad total de instancias, como muestra la ecuación 3.3.

$$P(X > 25) = \frac{30}{500} = 0.06 \tag{3.3}$$

Frecuentemente resulta más adecuado expresar o modelar un fenómeno como una combinación de dos o más variables aleatorias. Por ejemplo, un jugador de baloncesto puede anotar canastas de uno, dos o tres puntos dependiendo de si encesta con un tiro libre, un lanzamiento desde dentro del área o desde fuera del área. Así, se tienen tres variables aleatorias:

1. X: Anotaciones por tiro libre.
2. Y: Anotaciones desde dentro del área.
3. Z: Anotaciones desde fuera del área.

Podemos representar el total de puntos anotados por el jugador como la suma de los puntos anotados de las tres formas posibles, lo que corresponde a una **combinación lineal** de las variables X, Y y Z. La fórmula general de una combinación lineal de n variables está dada por la ecuación 3.4, donde cada X_i corresponde a una variable aleatoria y cada c_i es una constante conocida.

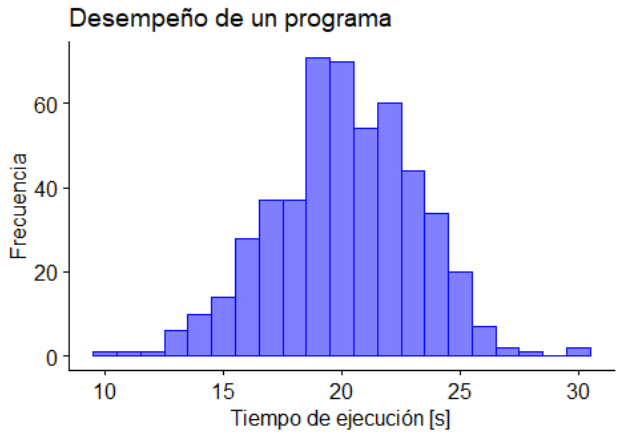


Figura 3.2: histograma para el desempeño del programa.

$$\sum_{i=1}^{n} c_i X_i \tag{3.4}$$

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.

```
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)
```

Al examinar con mayor detención los gráficos de la figura 3.1 podemos apreciar que, a medida que se efectúan más lanzamientos del dado, el histograma se asemeja cada vez más a una curva continua, la cual recibe el nombre de **función de densidad de probabilidad**, o simplemente **distribución** o **densidad**.

Las distribuciones tienen la propiedad de que el área total bajo la curva siempre es 1, lo que resulta muy útil al momento de calcular probabilidades, pues basta con calcular el área bajo la curva del segmento deseado. Volviendo al ejemplo del desempeño del programa, presentado en la figura 3.2, el tiempo de ejecución es en realidad una variable continua. Así, la probabilidad de que el tiempo de ejecución sea mayor a 25 segundos corresponde al área coloreada en el gráfico de la figura 3.3, con un valor de 0,048[2].

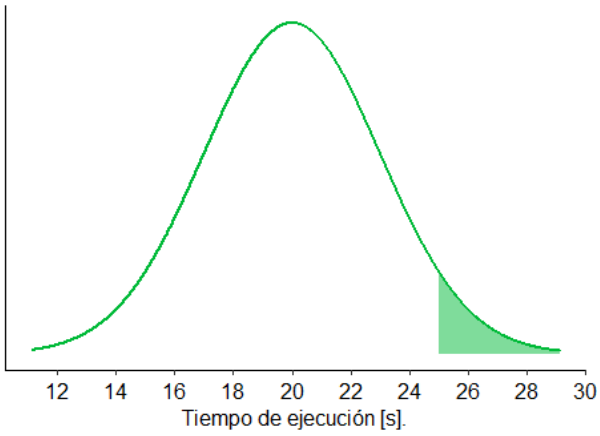


Figura 3.3: distribución para el desempeño del programa.

3.2 DISTRIBUCIONES CONTINUAS

Existen múltiples funciones de distribución continua que son de uso frecuente en estadística, las cuales se describen a continuación.

[2]El cálculo de esta probabilidad se aborda en el siguiente apartado

3.2.1 Distribución normal

También conocida como **distribución gaussiana**, la **distribución normal** es la más ampliamente empleada en estadística, pues muchas variables se acercan a esta distribución. Se caracteriza por ser unimodal y simétrica, con forma de campana. La figura 3.3 ejemplifica esta distribución.

La distribución normal se usa para modelar diversos fenómenos y podemos ajustarla mediante dos parámetros:

- μ: la media, que desplaza el centro de la curva a lo largo del eje x.
- σ: la desviación estándar, que modifica qué tan dispersos están los datos con respecto a la media.

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.

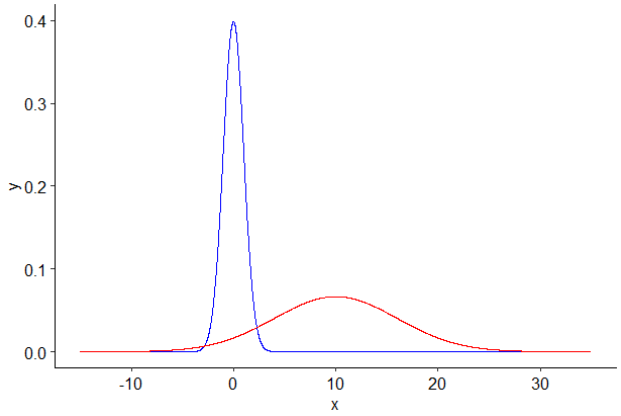


Figura 3.4: dos ejemplos superpuestos de distribución normal.

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)
```

Antes de continuar, fijémonos en las líneas 8 y 16 del script 3.4, donde se usa la función `dnorm(x, mean, sd)`. Esta función calcula la densidad de una distribución normal. Además de `dnorm()`, R nos ofrece otras funciones que también resultan de mucha ayuda:

- `pnorm(q, mean, sd, lower.tail)`: permite encontrar percentiles, los cuales corresponden a la **función de distribución acumulada** (es decir, la probabilidad de que la variable tome valores menores o iguales que un valor dado), a partir de las probabilidades.
- `qnorm(p, mean, sd, lower.tail)`: encuentra el percentil para las probabilidades dadas en `p`, por lo que es la función inversa de `pnorm()`.
- `rnorm(n, mean, sd)`: genera aleatoriamente `n` observaciones de la distribución normal especificada.

Los argumentos de esta familia de funciones son:

- `x, q`: vector de cuantiles (percentiles).
- `p`: vector de probabilidades.
- `mean`: media de la distribución normal.
- `sd`: desviación estándar de la distribución normal.
- `lower.tail`: valor lógico que señala cuál de los dos extremos o colas de la distribución emplear.
- `n`: tamaño del vector resultante.

Es importante señalar que, por defecto, `lower.tail` toma el valor verdadero, con lo que `pnorm()` y `qnorm()` operan con la cola inferior de la distribución. Si, en cambio, `lower.tail = FALSE`, dichas funciones operan con la cola superior (es decir, `pnorm()` nos entrega la probabilidad de que la variable tome valores mayores que un valor dado).

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.

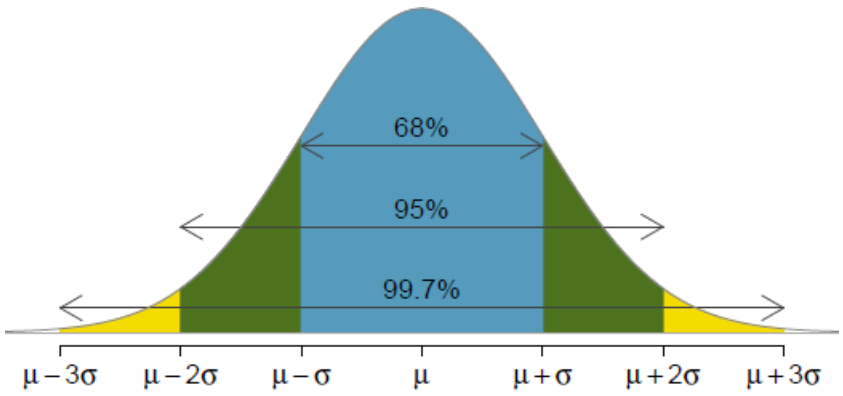


Figura 3.5: regla empírica de la distribución normal. Fuente: Diez y col. (2017, p. 136).

Muchas pruebas estadísticas operan bajo el supuesto de que los datos siguen una distribución normal. Como se insinuó en párrafos anteriores, la normalidad es siempre una aproximación, por lo que debemos verificar que el supuesto de una distribución normal sea aceptable. Una buena herramienta para ello es el **gráfico cuantil-cuantil**, también llamado **gráfico Q-Q**, que se muestra en la figura 3.6 y que podemos construir en R como muestra el script 3.5. En él podemos distinguir los siguientes elementos: un grupo de puntos, una recta y una región coloreada. Los puntos corresponden a las observaciones, mientras que la recta representa la distribución normal. En consecuencia, mientras más se asemeje el patrón que forman los puntos a la recta,

más parecida será la distribución a la normal. La banda coloreada establece el margen aceptable para suponer normalidad en el conjunto de datos. Así, para el conjunto de datos de la figura 3.6 sería imprudente aceptar el supuesto de normalidad.

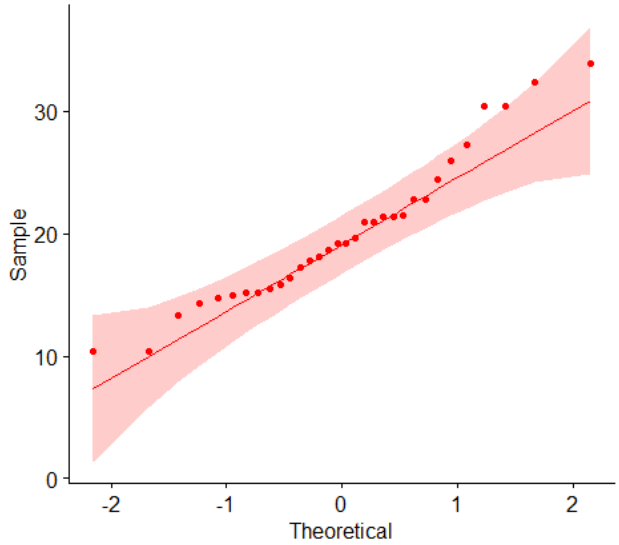


Figura 3.6: gráfico cuantil-cuantil.

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)
```

3.2.2 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.

3.2.3 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)`.

Donde:

- `x, q` son vectores de cuantiles (enteros no negativos).
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones.
- `df` son los grados de libertad.
- `lower.tail` es análogo al de la función `pnorm`.

La figura 3.7 muestra un ejemplo de la distribución χ^2.

3.2.4 Distribución t de Student

Ampliamente empleada cuando se trabaja con muestras pequeñas, la **distribución t de Student**, o simplemente **distribución t**, tiene, al igual que la distribución χ^2, los grados de libertad como único parámetro. A

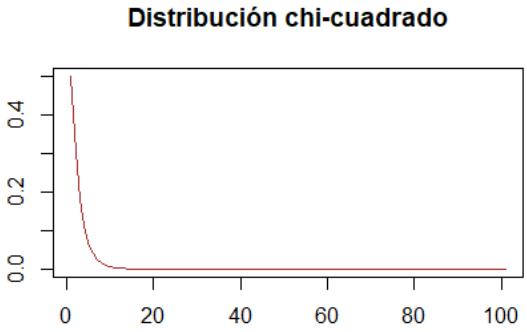


Figura 3.7: ejemplo de distribución χ^2 con 2 grados de libertad.

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.

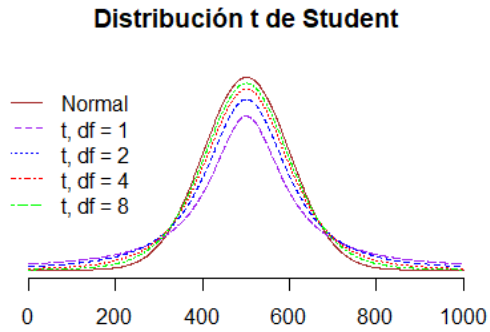


Figura 3.8: ejemplo de distribuciones t.

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.

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

Las funciones de R para esta distribución, cuyos argumentos son análogos a los que hemos visto para las distribuciones anteriores, son:

- `dt(x, df)`.
- `pt(q, df, lower.tail)`.

- `qt(p, df, lower.tail)`.
- `rt(n, df)`.

3.2.5 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.

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

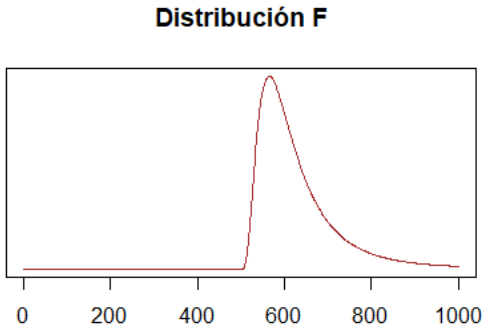


Figura 3.9: ejemplo de una distribución F.

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 $\nu_2 > 4$.

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

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

Las funciones de R para esta distribución son:

- `df(x, df1, df2)`.
- `pf(q, df1, df2, lower.tail)`.
- `qf(p, df1, df2, lower.tail)`.
- `rf(n, df1, df2)`.

Donde `df1` como `df2` corresponden a grados de libertad y los argumentos restantes son los mismos que ya hemos visto anteriormente.

3.3 DISTRIBUCIONES DISCRETAS

Al igual que con las variables aleatorias continuas, también existen diversas distribuciones discretas de uso frecuente en estadística.

3.3.1 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.

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

El paquete `extraDistr` de R ofrece 4 funciones, similares a las ya conocidas, para la distrubución de Bernoulli:

- `dbern(x, prob)`.
- `pbern(q, prob, lower.tail)`.
- `qbern(p, pro, lower.tail)`.
- `rbern(n, prob)`.

3.3.2 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.

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

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

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

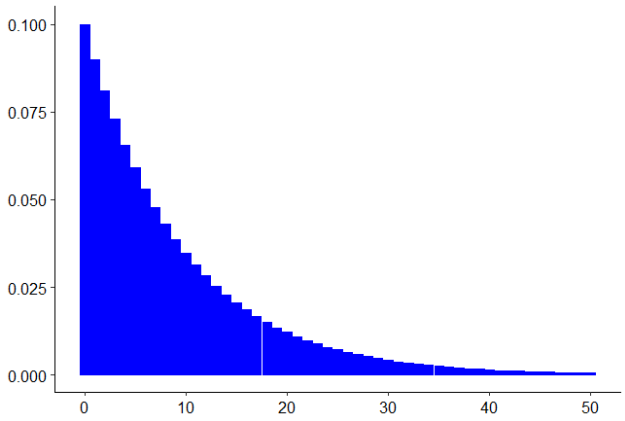


Figura 3.10: distribución geométrica para obtener un valor específico lanzando un dado de 20 caras.

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 en este caso es $p = 0.05$. El valor esperado, representado por la media, sería en este caso el que se presenta en la ecuación 3.18.

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

Una vez más, R ofrece funciones similares a las presentadas anteriormente para trabajar con distribuciones geométricas:

- `dgeom(x, prob)`.
- `pgeom(q, prob, lower.tail)`.
- `qgeom(p, prob, lower.tail)`.
- `rbern(n, prob)`.

3.3.3 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

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

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

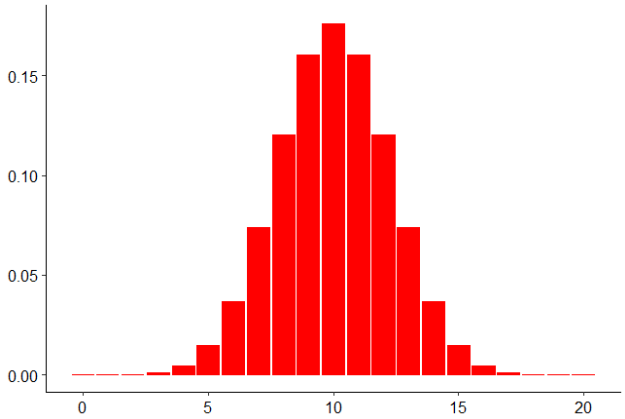


Figura 3.11: distribución binomial con $\mu = 400$ y $\sigma = 15.4019$.

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.

Las funciones que ofrece R para trabajar con esta distribución son:

- `dbinom(x, size, prob)`.
- `pbinom(x, size, prob)`.
- `qbinom(p, size, prob)`.
- `rbinom(n, size, prob)`.

Donde:

- `x` es un vector numérico.
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones.
- `size` corresponde al número de intentos.
- `prob` es la probabilidad de éxito de cada intento.

En la figura 3.11 podemos observar que, en cierto modo, la distribución binomial se asemeja a la distribución normal: ambas son simétricas, aunque la distribución binomial no tiene la forma de campana de la distribución normal. Esta similitud ofrece una importante ventaja, pues en ocasiones es posible usar la distribución normal para estimar probabilidades binomiales, evitando así el uso de la compleja fórmula de la ecuación 3.19. Formalmente, esta aproximación es válida cuando el tamaño de la muestra, n, es lo suficientemente grande para que tanto np como $n(1-p)$ sean mayores o iguales que 10. En este caso, los parámetros de la distribución normal aproximada son los mismos de la distribución binomial (ecuaciones 3.20 y 3.21).

3.3.4 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).

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

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

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

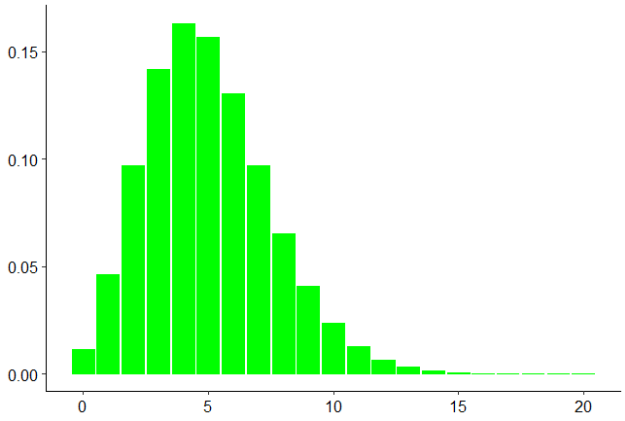


Figura 3.12: ejemplo de distribución binomial negativa.

Nuevamente, R dispone de cuatro funciones que permiten trabajar con esta distribución:

- `dnbinom(x, size, prob)`.
- `pnbinom(q, size, prob, lower.tail)`.
- `qnbinom(p, size, prob, lower.tail)`.
- `rnbinom(n, size, prob)`.

Donde:

[3]Traducción libre de los autores.

- `x, q` son vectores de cuantiles (enteros no negativos).
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones.
- `size` corresponde al número (no negativo) de intentos.
- `prob` es la probabilidad de éxito de cada intento.
- `lower.tail` es análogo al de la función `pnorm`.

3.3.5 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 $\sqrt{\lambda}$.

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

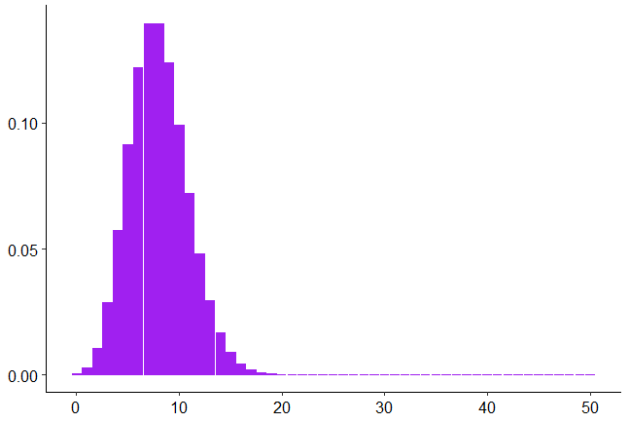


Figura 3.13: ejemplo de distribución de Poisson.

Las funciones de R para esta distribución son:

- `dpois(x, lambda)`.
- `ppois(q, lambda, lower.tail)`.
- `qqpois(p, lambda, lower.tail)`.
- `rpois(n, lambda)`.

Donde:

- `x, q` son vectores de cuantiles (enteros no negativos).
- `p` es un vector de probabilidades.
- `n` es la cantidad de observaciones.
- `lambda` es un vector no negativo de medias.
- `lower.tail` es análogo al de la función `pnorm`.

3.4 EJERCICIOS PROPUESTOS

1. Da un ejemplo de variable aleatoria (novedosa) que puedas observar en tus compañeros y que tenga una función de densidad de probabilidad discreta.
2. Para la variable anterior, ¿cuál sería el valor esperado? ¿Cuál sería la varianza? ¿Cómo te imaginas su función de densidad de probabilidad?
3. Lista tres nombres distintos con que también se llama a las funciones de densidad de probabilidad.
4. Si una variable aleatoria tiene una función de densidad de probabilidad con media igual a 30 y desviación estándar de 3, ¿por qué podría ocurrir que la probabilidad de que la variable tome el valor 30 sea nula, es decir, $P(X = 30) = 0$?
5. Según el Reporte Mensual de Empleo, las siguientes son las estadísticas (media $\pm$ desviación estándar) para las seis variables relevantes que se han estudiado en los últimos cinco años:
   a) Número de personas despedidas: 64.675 ± 8.321.
   b) Número de personas renunciadas: 118.543 ± 17.936.
   c) Número de personas jubiladas: 97.092 ± 11.147.
   d) Número de empleos creados: 24.715 ± 10.832.
   e) Número de personas contratadas: 301.345 ± 27.261.
   f) Número de personas entrando a la fuerza de trabajo: 26.444 ± 29.440.

   Con esta información, calcula la media y la desviación estándar de:
   a) Caída neta del empleo: $(d) - (a) - (b) - (c)$.
   b) Subida neta del empleo: $(e) - (a) - (b) - (c)$.
   c) Caída neta del desempleo: $(a) + (b) + (e) + (f)$.
   d) Vacancia del empleo: $(d) - (e)$.
6. ¿Qué significa que cierto valor de una variable aleatoria, usualmente con distribución normal, tenga $valor Z = 1,5$?
7. Según la regla empírica, ¿entre qué estaturas se podría encontrar al 95 % de los estudiantes varones del Departamento de Ingeniería Informática de la Universidad de Santiago de Chile, si esta variable sigue una distribución $N(\mu = 171, \sigma = 3)$?
8. ¿Qué información entrega un Gráfico Q-Q? ¿Para qué se usa?
9. La probabilidad de que un estudiante universitario chileno seleccionado al azar sea VIH positivo es 0,013. ¿Cuáles serían la media y la desviación estándar de esta variable?
10. En promedio, ¿a cuántos estudiantes universitarios se debería revisar hasta encontrar a un VIH positivo?
11. Si el Departamento de Salud de una Universidad chilena controla a 50 estudiantes por día durante una semana de clases (lunes a viernes), ¿cuál sería el número promedio de VIH positivos detectados cada día? ¿Con qué varianza?
12. Si la Universidad del ejercicio anterior dispone de 10 paquetes de tratamiento de VIH para estudiantes, ¿cómo podría saber a cuántos estudiantes debería examinar para poder asignarlos (suponiendo que todo estudiante VIH positivo acepta el tratamiento)?
13. Muestra un ejemplo novedoso de una variable aleatoria relacionada que podría seguir una distribución de Poisson.

REFERENCIAS

Cross, J. (2017). *Discrete Random Variables*.
Consultado el 9 de abril de 2021, desde https://rpubs.com/jcross/discreteRV

Devore, J. L. (2008). *Probabilidad y Estadística para Ingeniería y Ciencias* (7.ª ed.). CENAGE Learning.

Diez, D., Barr, C. D. & Çetinkaya-Rundel, M. (2017). *OpenIntro Statistics* (3.ª ed.).
https://www.openintro.org/book/os/.

Freund, R. J. & Wilson, W. J. (2003). *Statistical Methods* (2.ª ed.). Academic Press.