Métodos Estadísticos y Simulación

Unidad 4: Distribuciones de probabilidad

Mg. J. Eduardo Gamboa U.

2025-10-19

Distribuciones de probabilidad

Las distribuciones de probabilidad son herramientas fundamentales en estadística y probabilidad, ya que permiten modelar fenómenos aleatorios y predecir comportamientos futuros con base en datos. A través de ellas, se asignan probabilidades a los posibles valores que puede tomar una variable aleatoria, ya sea discreta o continua.

Distribuciones de probabilidad discretas

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:

$$f(x) = P(X = x) = \begin{cases} \binom{n}{x}\pi^x(1-\pi)^{n-x}, & x \in \{0, 1, 2, \dots, n\} \\ 0, & \text{de otro modo} \end{cases}$$

Notación: $X \sim Bin(n, \pi)$

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

Estimador máximo verosímil

$$\hat{\pi} = \frac{x}{n}$$

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 $\rightarrow S \sim Bin(n = 200, \pi = 0.88)$

Aplicación

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:

$$f(x) = P(X = x) = \begin{cases} \binom{20}{x} 0.75^x \times 0.25^{20-x}, & x \in \{0, 1, 2, \dots, 20\} \\ 0, & \text{de otro modo} \end{cases}$$

La cantidad media de clientes que hacen una compra:

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

Su varianza:

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

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
```

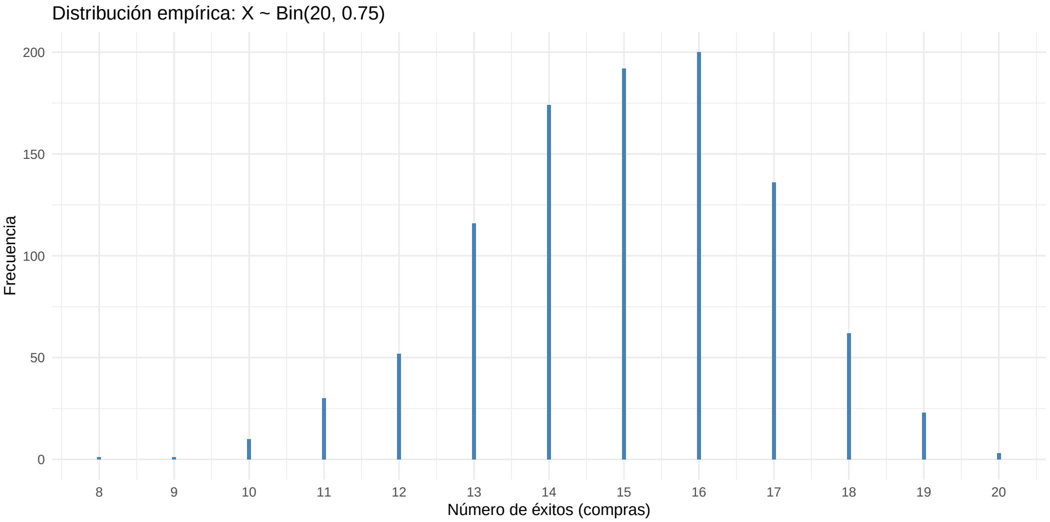
Esto significa que en al menos la mitad de las semanas, 15 o menos clientes realizan una compra.

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.


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:

$$f(x) = P(X = x) = \begin{cases} \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}, & \textbf{máx}(0, n + A - N) \leq x \leq \textbf{mín}(n, A) \\ 0, & \text{de otro modo} \end{cases}$$

Notación: $X \sim Hiper(N, n, A)$

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.

Estimador máximo verosímil

$$\hat{A} = \frac{N}{n}x$$

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
  $N \sim Hiper(N = 80, n = 10, A = 25)$

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 $\pi = 0.6$.

Aplicación

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:

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

es decir $x \in \{0, 1, 2\}$

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:

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

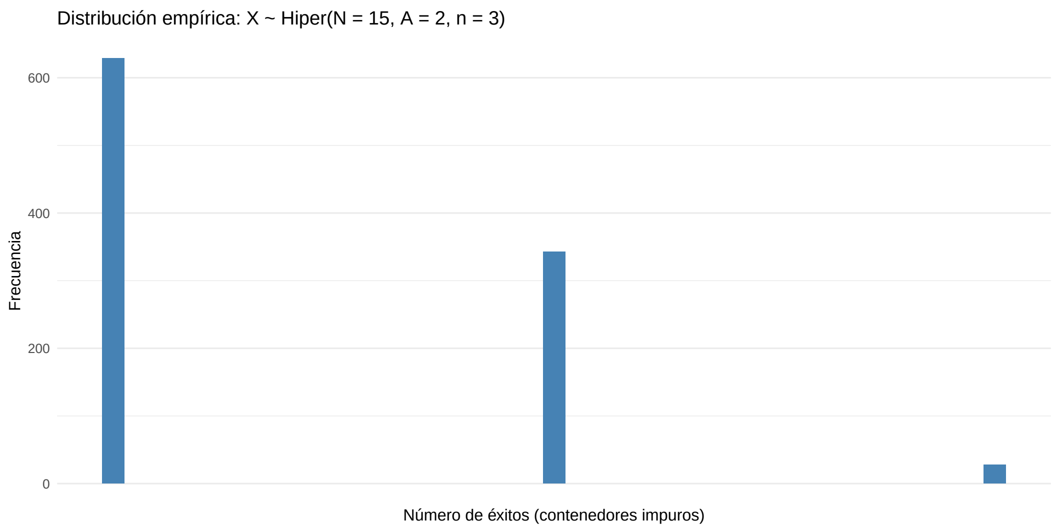
```
[1] 1
```

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:

```r
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
```


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

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

Varianca

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

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

Estimador máximo verosímil

$$\hat{\lambda} = \overline{x}$$

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 $A \sim Pois(6)$

Aplicación

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:

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

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, $X \sim Pois(1.8)$

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

```
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, $X \sim Pois(1.8)$

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

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

```
[1] 0.1087084
```

```r
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:

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

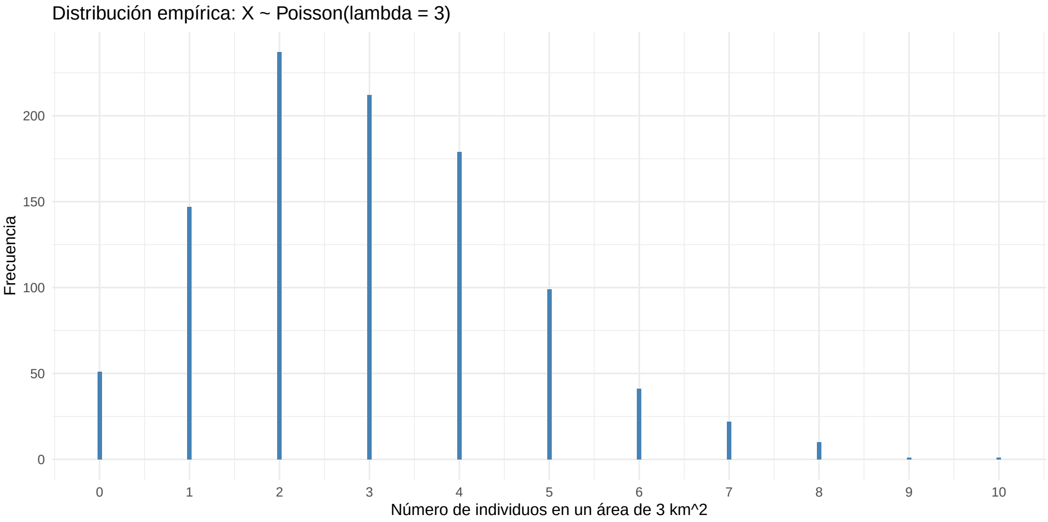
```
[1] 1
```

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
```


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 $Y \sim Pois(2)$

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

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

```
[1] 0.1805373
```

```r
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:

$$f(x) = P(X = x) = \begin{cases} (1-\pi)^x \pi, & x \in \{0, 1, 2, 3, ...\} \\ 0, & \text{de otro modo} \end{cases}$$

Notación: $X \sim Geom(\pi)$

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.

Estimador máximo verosímil

$$\hat{\pi} = \frac{n}{n + \sum_{i=1}^{n} X_i} = \frac{1}{1+\overline{x}}$$

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 $\rightarrow E \sim Geom(0.1)$

Aplicación

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:

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

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:

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

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

```
[1] 0.4437053
```

```
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
```

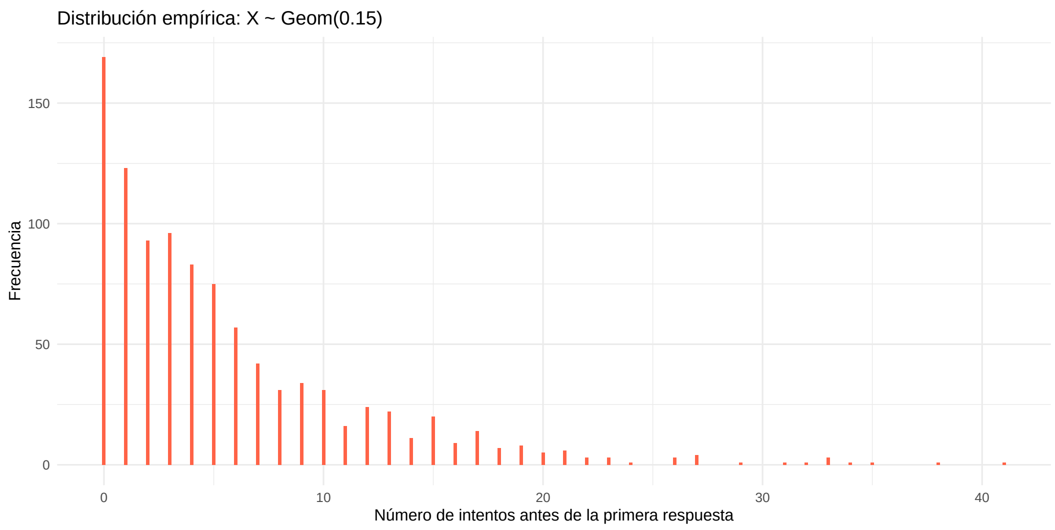
Esto significa que en al menos el 25% de los casos se tiene un intento fallido antes de recibir la respuesta por correo.

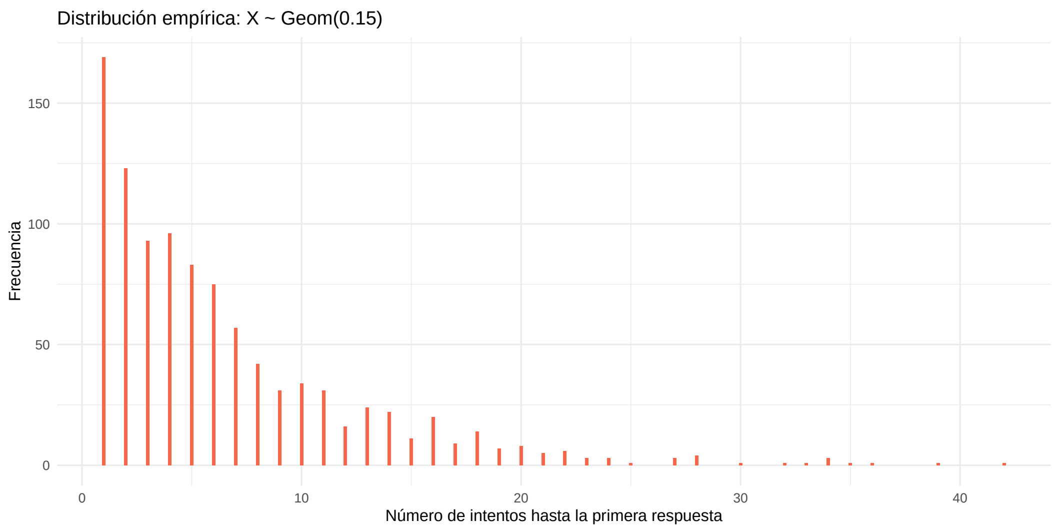
Simulación de una muestra aleatoria de la cantidad de correos enviados antes de recibir la respuesta.

```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.


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


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 $r = 1$)

Función de probabilidad

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

$$f(x) = P(X = x) = \begin{cases} \binom{x+r-1}{x}(1-\pi)^x \pi^r, & x \in \{0, 1, 2, 3, ...\} \\ 0, & \text{de otro modo} \end{cases}$$

Notación: $X \sim BinNeg(r, \pi)$

Media

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

Varianza

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

Estimador máximo verosímil

$$\hat{\pi} = \frac{nr}{nr + \sum_{i=1}^{n} X_i} = \frac{r}{r + \overline{x}}$$

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 $R \sim BinNeg(r = 6, \pi = 0.3)$

Aplicación

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:

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

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

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

```
[1] 16
```

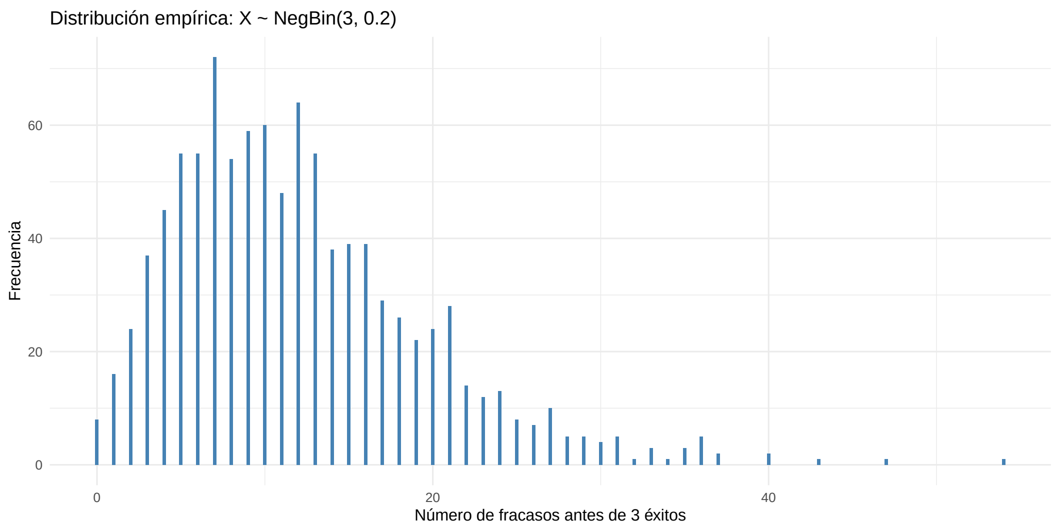
En al menos el 75% de los casos, el vendedor fallará como máximo 16 veces antes de cerrar existosamente 3 ventas.

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.



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

Otras distribuciones

- Distribución multinomial
- Distribución discreta uniforme
- Distribución Zipf
- Distribución Skellam
- Distribución Conway–Maxwell–Poisson

Distribuciones continuas

Distribución uniforme

- Distribución continua que modela situaciones donde todos los valores dentro de un intervalo son igualmente probables
- Representa la máxima ignorancia, incertidumbre o entropía: no se conoce ninguna concentración de probabilidad dentro del intervalo
- Comúnmente usada como modelo inicial en simulaciones, como componente de algoritmos estocásticos y como distribución a priori no informativa en inferencia bayesiana.
- La mayoría de los algoritmos de simulación comienzan generando variables $U \sim Unif(0,1)$, que luego se transforman para obtener otras distribuciones.

Función de densidad

Se define la V.A.C. X que toma valores en el intervalo $[a, b]$ donde $a < b$, cuya función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{de otro modo} \end{cases}$$

Notación: $X \sim Unif(a, b)$

Función de distribución acumulada

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & x \in [a, b] \\ 1, & x > b \end{cases}$$

Media

$$\mu_X = E(X) = \frac{a+b}{2} = Me$$

Varianza

$$\sigma_X^2 = V(X) = \frac{(b-a)^2}{12}$$

Asimetría

La distribución uniforme es simétrica.

Estimadores máximos verosímiles

$$\hat{a} = X_{(1)} \qquad \hat{b} = X_{(n)}$$

Ejemplos

- T $=$ Tiempo de espera para el bus si llega en cualquier momento dentro de un intervalo de 10 minutos $\rightarrow T \sim Unif(0, 10)$
- U $=$ Temperatura ambiente durante una hora establecida si se mantiene entre 18 y 22 grados Celsius $\rightarrow U \sim Unif(18, 22)$
- C $=$ Tiempo de carga de un archivo entre 30 y 50 segundos cuando no hay congestión $\rightarrow C \sim Unif(30, 50)$

Aplicación

En una planta embotelladora, una máquina realiza inspecciones automáticas de botellas a intervalos aleatorios entre 12 y 20 segundos. No hay preferencia por ningún valor dentro de ese rango.

Variable aleatoria

V.A. X = Tiempo (en segundos) hasta la próxima inspección, $X \sim Unif(12, 20)$

Función de densidad

$$f(x) = \frac{1}{8} \qquad x \in [12, 20]$$

Función de distribución acumulada

$$F(x) = \begin{cases} 0, & x < 12 \\ \frac{x-12}{8}, & x \in [12, 20] \\ 1, & x > 20 \end{cases}$$

El tiempo medio hasta la próxima inspección:

$$\mu_X = E(X) = \frac{12 + 20}{2} = 16$$

Su varianza:

$$\sigma_X^2 = V(X) = \frac{(20 - 12)^2}{12} = 5.333$$

Probabilidad de que una inspección ocurra entre 13 y 16 segundos:

$$P(13 \leq X \leq 16) = \int_{13}^{16} \frac{1}{8} dx = \frac{16-13}{8} = \frac{3}{8} = 0.375$$

```
[1] 0.375
```

$$P(13 \leq X \leq 16) = P(X \leq 16) - P(X \leq 13) = F(16) - F(13) = \frac{16-12}{8} - \frac{13-12}{8} = \frac{3}{8} = 0.375$$

```r
punif(16, min = 12, max = 20) - punif(13, min = 12, max = 20)
```

```
[1] 0.375
```

Probabilidad de que una inspección ocurra luego de los 17 segundos

$$P(X > 17) = 1 - P(X \leq 17) = 1 - F(17) = 1 - \frac{17 - 12}{8} = 0.375$$

```
punif(q = 17, min = 12, max = 20, lower.tail = FALSE)
```

```
[1] 0.375
```

Percentil 90 del tiempo de inspección:

$$F(x) = 0.90 \rightarrow \frac{x-12}{8} = 0.9 \rightarrow x = 8 \times 0.9 + 12 = 19.2$$

```r
qunif(p = 0.9, min = 12, max = 20)
```

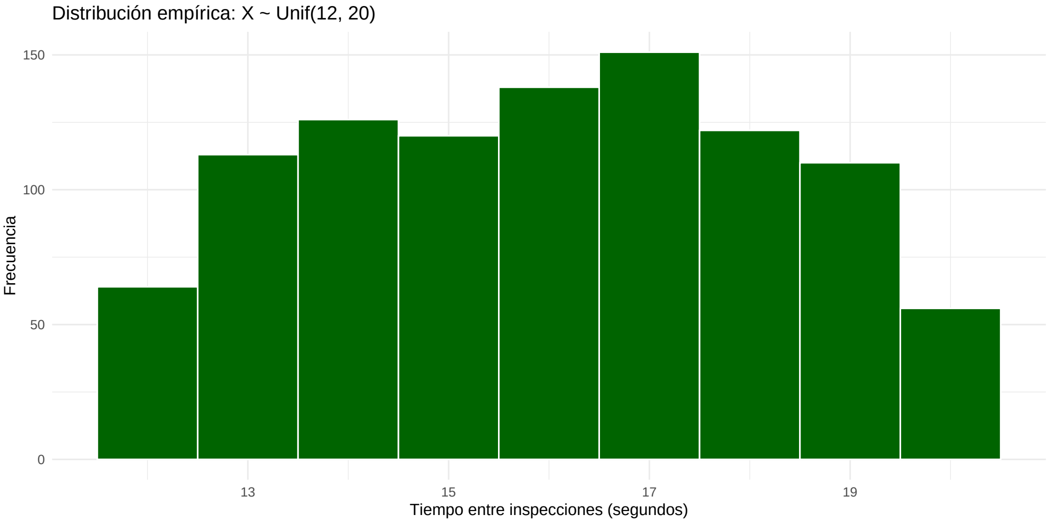
```
[1] 19.2
```

Esto significa que el 90% de las inspecciones ocurre hasta los 19.2 segundos.

Simulación de 7 tiempos de inspección aleatorios:

```r
set.seed(111)
runif(n = 7, min = 12, max = 20)
```

```
[1] 16.74385 17.81185 14.96338 16.11939 15.02131 15.34670 12.08526
```


Distribución empírica: X ~ Unif(12, 20)
Frecuencia
150
100
50
0
13
15
17
19
Tiempo entre inspecciones (segundos)

Distribución Exponencial

- Distribución continua que modela el tiempo entre dos eventos sucesivos en un proceso de Poisson.
- Se utiliza para describir tiempos de espera hasta que ocurra un evento, como la llegada de clientes, fallas de un sistema, o llamadas telefónicas.
- Es una distribución sin memoria: la probabilidad de que ocurra un evento en el futuro no depende de cuánto tiempo ha pasado.

Función de densidad

Se define la V.A.C. X como el tiempo entre eventos, con parámetro $\lambda > 0$, cuya función de densidad viene dada por:

$$f(x) = \begin{cases} \lambda \exp\left(-\lambda x\right), & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Función de distribución acumulada

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt = \begin{cases} 0, & x < 0 \\ 1 - \exp\left(-\lambda x\right), & x \geq 0 \end{cases}$$

Media

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

Varianza

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

Propiedad de falta de memoria

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

Es decir, el tiempo restante no depende de cuánto ya se ha esperado.

Estimador máximo verosímil

$$\hat{\lambda} = \frac{1}{\overline{x}}$$

Ejemplos

- T $=$ Tiempo hasta que llegue la próxima llamada a una central telefónica con $\lambda = 3 \rightarrow T \sim Exp(\lambda = 3)$
- C $=$ Tiempo entre llegadas de clientes en una tienda si se estima 1 cliente cada 4 minutos $\rightarrow T \sim Exp\left(\lambda = \frac{1}{4}\right)$
- S $=$ Tiempo de espera para cruzar una calle si el semáforo cambia al azar y en promedio lo hace cada 45 segundos. $\rightarrow S \sim Exp\left(\frac{1}{45}\right)$

Aplicación

Un cajero automático recibe, en promedio, un cliente cada 6 minutos. Se desea modelar el tiempo que pasa entre la salida de un cliente y la llegada del siguiente.

Variable aleatoria

X = Tiempo entre llegadas de clientes (en minutos), $X \sim Exp\left(\lambda = \frac{1}{6}\right)$

Función de densidad

$$f(x) = \frac{1}{6}\exp\left(-\frac{x}{6}\right) \qquad x \geq 0$$

Función de distribución acumulada

$$F(x) = \begin{cases} 0, & x < 0 \\ 1 - \exp\left(-\frac{x}{6}\right), & x \geq 0 \end{cases}$$

El tiempo medio entre cliente y cliente es:

$$\mu_X = E(X) = \frac{1}{1/6} = 6$$

Su varianza es:

$$\sigma_X^2 = V(X) = \frac{1}{(1/6)^2} = 36$$

La probabilidad de que el próximo cliente llegue antes de los 4 minutos:

$$P(X < 4) = F(4) = 1 - \exp\left(-\frac{4}{6}\right) = 0.48658$$

```r
pexp(q = 4, rate = 1/6)
```

```
[1] 0.4865829
```

La probabilidad de que el próximo cliente llegue luego de los 9 minutos:

$$P(X > 9) = 1 - F(9) = \exp\left(-\frac{9}{6}\right) = 0.2231302$$

```
pexp(q = 9, rate = 1/6, lower.tail = FALSE)
```

```
[1] 0.2231302
```

```
1 - pexp(q = 9, rate = 1/6)
```

```
[1] 0.2231302
```

La probabilidad de que el próximo cliente llegue entre los 3 y 7 minutos siguientes

$$P(3 < X < 7) = F(7) - F(3) = \left(1 - \exp\left(-\frac{7}{6}\right)\right) - \left(1 - \exp\left(-\frac{3}{6}\right)\right) = 0.2951274$$

```r
pexp(q = 7, rate = 1/6) - pexp(q = 3, rate = 1/6)
```

```
[1] 0.2951274
```

El percentil 33 del tiempo entre llegadas

```
qexp(p = 0.33, rate = 1/6)
```

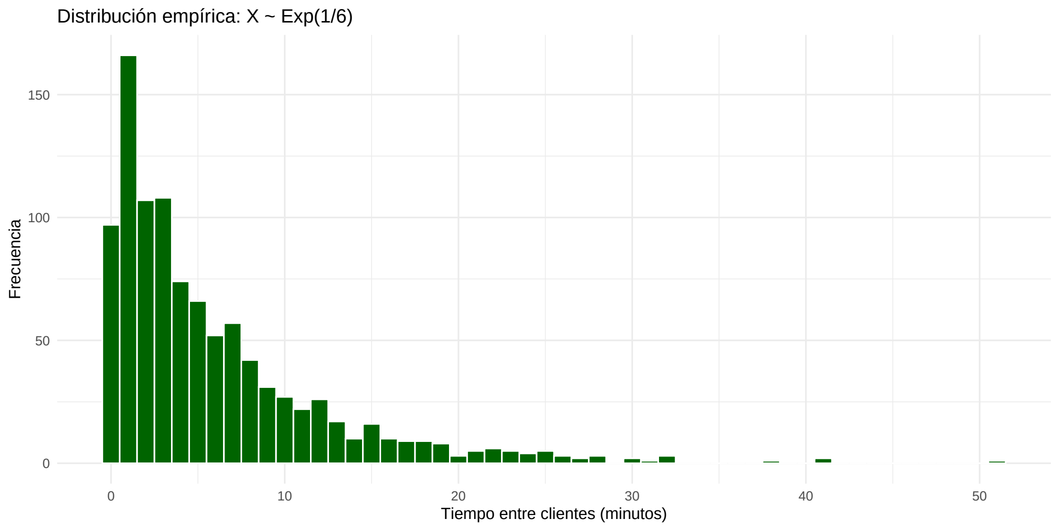
```
[1] 2.402865
```

Significa que el 33% de los tiempos entre clientes que llegan al cajero es como máximo de 2.40 minutos (2 minutos y 24 segundos).

Una muestra aleatoria de 5 tiempos de llegada

```r
set.seed(753)
rexp(n = 5, rate = 1/6)
```

```
[1]  0.04490393  2.09205825  5.62120365 10.65256807  7.35046952
```


Distribución empírica: X ~ Exp(1/6)
Frecuencia
150
100
50
0
0
10
20
30
40
50
Tiempo entre clientes (minutos)

Distribución Gamma

- Distribución continua que generaliza la exponencial, modelando el tiempo hasta que ocurren α eventos en un proceso de Poisson.
- Ampliamente utilizada en teoría de colas, fiabilidad, meteorología, biomedicina, y modelamiento de tiempos de espera acumulados.
- Si $\alpha = 1$, la distribución Gamma se reduce a la Exponencial

Función de densidad

Se define la V.A.C. X con distribución Gamma con parámetros $\alpha > 0$ y $\beta > 0$, cuya función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Media

$$\mu_X = E(X) = \alpha\beta$$

Varianza

$$\sigma_X^2 = V(X) = \alpha\beta^2$$

Asimetría

Asimétrica a la derecha (menos conforme α crece)

Estimador máximo verosímil

$$\hat{\beta} = \frac{\overline{x}}{\hat{\alpha}}$$

Sin embargo, no existe una fórmula cerrada para estimar α por máxima verosimilitud.

Ejemlos

- T: Tiempo que transcurre hasta que llegan 3 estudiantes a una clase, si cada llegada en promedio cada 2 minutos $\rightarrow T \sim Gamma(\alpha = 3, \beta = 2)$
- F: tiempo total hasta 5 fallas en un sistema de producción, con tiempo medio de 1.5 días entre fallas $\rightarrow F \sim Gamma(\alpha = 5, \beta = 1.5)$

Aplicación

En una central de monitoreo de emergencias, se sabe que, en promedio, ocurre un evento cada 10 minutos. Queremos modelar el tiempo hasta que ocurran 3 eventos consecutivos.

Variable aleatoria X =Tiempo que transcurre hasta 3 eventos

$$X \sim Gamma(\alpha = 3, \beta = 10)$$

Función de densidad

$$f(x) = \begin{cases} \frac{1}{\Gamma(3)10^3} x^2 \exp\left(-\frac{x}{10}\right), & x \geq 0 \\ 0, & x < 0 \end{cases}$$

El tiempo medio que transcurre hasta 3 eventos es:

$$\mu_X = E(X) = 3 \times 10 = 30$$

Su varianza:

$$\sigma 2_X = V(X) = 3 \times 10^2 = 300$$

Probabilidad de que ocurran los 3 eventos antes de 25 minutos:

$$P(X < 25) = 0.4562$$

```r
pgamma(q = 25, shape = 3, scale = 10)
```

```
[1] 0.4561869
```

Percentil 11 del tiempo que transcurre hasta que sucedan 3 eventos

```r
qgamma(p = 0.11, shape = 3, scale = 10)
```

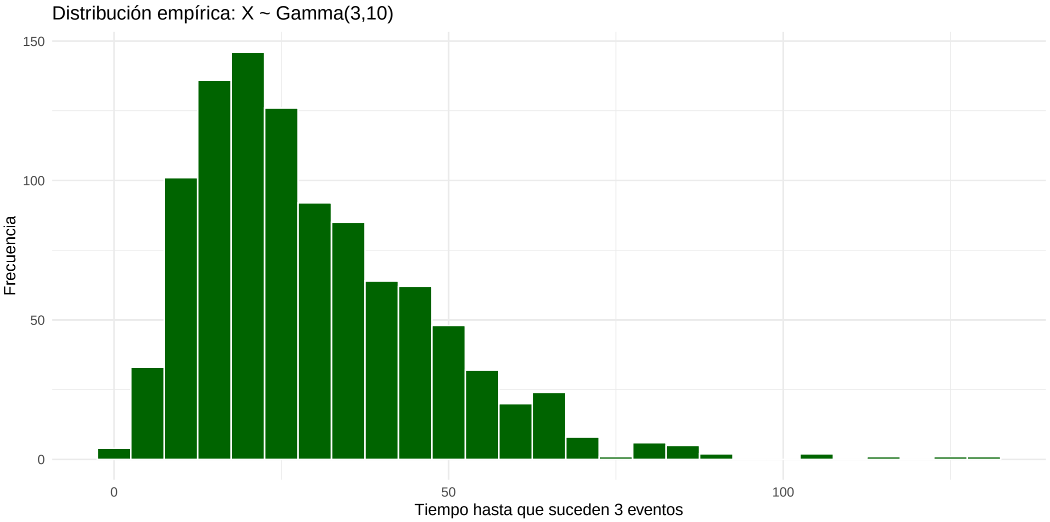
```
[1] 11.50693
```

El 11% de los tiempos que transcurren hasta que sucedan 3 eventos es menor o igual a 11.51 minutos.

Muestra aleatoria de 7 tiempos que transcurren hasta que sucedan 3 eventos:

```r
set.seed(852)
rgamma(n = 7, shape = 3, scale = 10)
```

```
[1] 25.997993 57.500235 22.171333 25.254647 16.790934  9.310327 47.792846
```



Distribución empírica: X ~ Gamma(3,10)
Frecuencia
150
100
50
0
0
50
100
Tiempo hasta que suceden 3 eventos

Distribución Normal

- Distribución continua ampliamente utilizada en estadística y ciencia de datos, conocida como la campana de Gauss.
- Su importancia se debe al Teorema Central del Límite, que establece que la suma (o media) de muchas variables independientes tiende a seguir una distribución normal, aun si las variables originales no lo sean.

Función de densidad

Se define la V.A. X con distribución Normal con parámetros μ y σ^2, cuya función de densidad viene dada por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \qquad x \in (-\infty, \infty)$$

Notación: $X \sim N(\mu, \sigma^2)$

Media

$$\mu_x = E(X) = \mu = Me$$

Varianza

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

Estimadores máximo verosímiles

$$\hat{\mu} = \overline{x} \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Estandarización

Para comparar valores provenientes de distintas normales:

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

Ejemplos

- P = Presión arterial sistólica en adultos sanos, en mm Hg
  $P \sim N(\mu = 120, \sigma^2 = 225)$
- T = Tiempo de resolución de un examen, en minutos $T \sim N(\mu = 90, \sigma^2 = 100)$
- E = Estatura de un estudiante de pregrado, en centímetros
  $E \sim N(\mu = 168, \sigma^2 = 90.25)$

Aplicación

La duración de la batería de cierto modelo de laptop sigue una distribución normal con media de 6 horas y desviación estándar de 1.2 horas.

Variable aleatoria

X = Duración de la batería, en horas.

$$X \sim N(6, 1.44)$$

Función de densidad

$$f(x) = \frac{1}{1.2\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-6}{1.2}\right)^2\right) \qquad x \in (-\infty, \infty)$$

Probabilidad de que una batería dure menos de 7 horas

$$P(X < 7) = 0.7977$$

```r
pnorm(q = 7, mean = 6, sd = 1.2)
```

```
[1] 0.7976716
```

Probabilidad de que una batería dure más de 3 horas y 15 minutos

$$P(X > 3.25) = 1 - P(X < 3.25) = 0.9813$$

```
pnorm(q = 3.5, mean = 6, sd = 1.2, lower.tail = F)
```

```
[1] 0.9813896
```

```
1-pnorm(q = 3.5, mean = 6, sd = 1.2)
```

```
[1] 0.9813896
```

Probabilidad de que la duración esté entre 7.5 y 9 horas:

$$P(7.5 < X < 9) = 0.099$$

```
pnorm(9, mean = 6, sd = 1.2) - pnorm(7.5, mean = 6, sd = 1.2)
```

```
[1] 0.09944011
```

Percentil 27 de la duración de batería

```r
qnorm(p = 0.27, mean = 6, sd = 1.2)
```

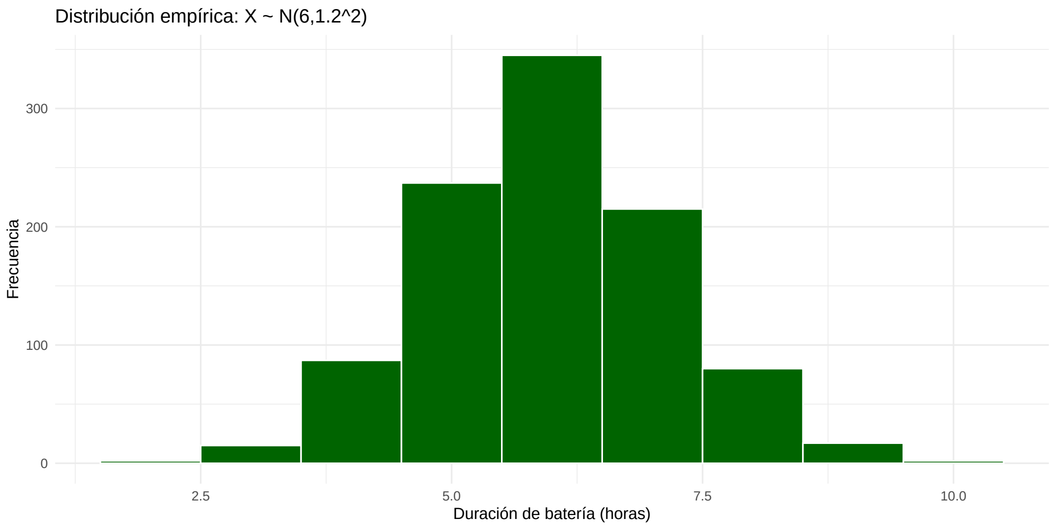
```
[1] 5.264624
```

Esto significa que el 27% de las baterías dura como máximo 5.26 horas.

Simulación de 6 tiempos de duración de batería

```r
set.seed(555)
rnorm(n = 6, mean = 6, sd = 1.2)
```

```
[1] 5.604168 6.604376 6.449243 8.266344 3.864117 7.062757
```


Distribución empírica: X ~ N(6,1.2^2)
Frecuencia
300
200
100
0
2.5
5.0
7.5
10.0
Duración de batería (horas)

Aproximaciones a la Normal

Aproximación de la distribución Binomial a la Normal

Si $X \sim Bin(n, \pi)$, entonces $X \approx N(\mu = n\pi, \sigma^2 = n\pi(1-\pi))$ siempre que $n\pi \geq 5$ y $n(1-\pi) \geq 5$. Sin embargo se debe usar la corrección por continuidad.

$$P(X = a) = P(a - 0.5 \leq X \leq a + 0.5)$$

$$P(X \leq a) = P(X \leq a + 0.5)$$

Por ejemplo, si $X \sim Bin(80, 0.3)$ entonces $X \approx N(\mu = 24, \sigma^2 = 16.8)$

$$P(X = 8) = 0.0626 \approx 0.0604$$

```
dbinom(x = 20, size = 80, prob = 0.3)
```

```
[1] 0.06262327
```

```
pnorm(q=20.5, mean=24, sd=sqrt(16.8)) - pnorm(q=19.5, mean=24, sd=sqrt(16.8
```

```
[1] 0.06044993
```

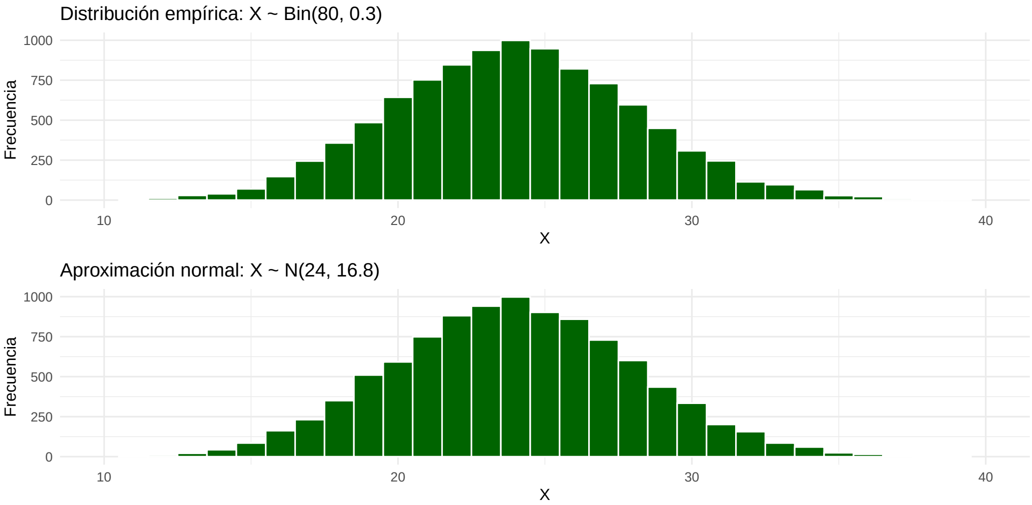
$$P(X \leq 15) = 0.016 \approx 0.019$$

```
pbinom(q = 15, size = 80, prob = 0.3)
```

```
[1] 0.01606465
```

```
pnorm(q = 15.5, mean = 24, sd = sqrt(16.8))
```

```
[1] 0.01904952
```


Distribución empírica: X ~ Bin(80, 0.3)
Frecuencia
1000
750
500
250
0
10
20
30
40
X
Aproximación normal: X ~ N(24, 16.8)
Frecuencia
1000
750
500
250
0
10
20
30
40
X

Aproximación de la distribución Hipergeométrica a la Normal

Si $X \sim Hiper(N, n, A)$, entonces $X \approx N\left(\mu = n\frac{A}{N}, \sigma^2 = n\frac{A}{N}\left(1 - \frac{A}{N}\right)\left(\frac{N-n}{N-1}\right)\right)$ siempre que $n\frac{A}{N} \geq 5$. Sin embargo se debe usar la corrección por continuidad.

$$P(X = a) = P(a - 0.5 \leq X \leq a + 0.5)$$

$$P(X \leq a) = P(X \leq a + 0.5)$$

Por ejemplo, si $X \sim Hiper(N = 250, n = 60, A = 30)$ entonces $X \approx N(\mu = 7.2, \sigma^2 = 4.835)$. Luego, $P(X = 10) = 0.0775 \approx 0.0811$.

```r
dhyper(x = 10, m = 30, n = 220, k = 60)
```

```
[1] 0.07745748
```

```r
pnorm(q=10.5, mean=7.2, sd=sqrt(4.835)) - pnorm(q=9.5, mean=7.2, sd=sqrt(4
```

```
[1] 0.08107484
```

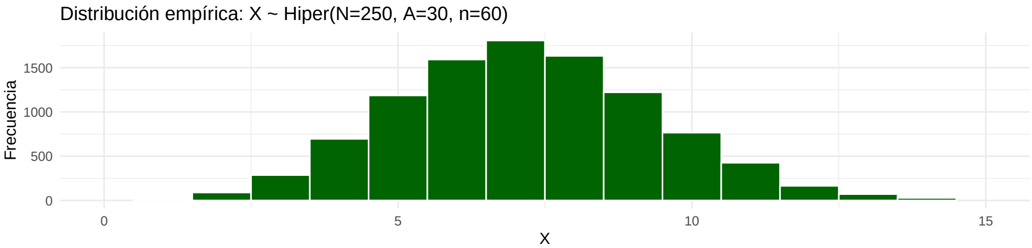
$$P(X \leq 10) = 0.9299 \approx 0.9333$$

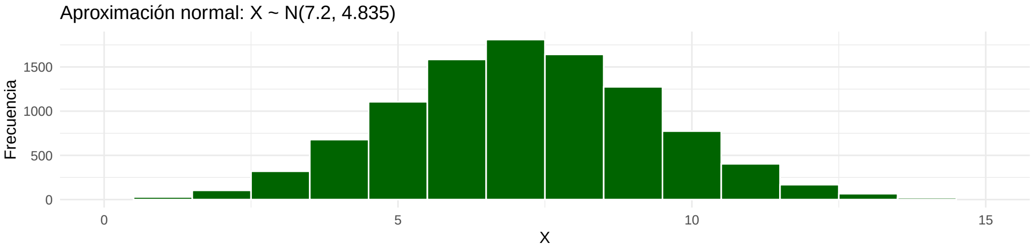
```r
phyper(q = 10, m = 30, n = 220, k = 60)
```

```
[1] 0.9299182
```

```r
pnorm(q = 10.5, mean=7.2, sd=sqrt(4.835))
```

```
[1] 0.9332932
```


Distribución empírica: X ~ Hiper(N=250, A=30, n=60)
Frecuencia
1500
1000
500
0
0
5
10
15
X


Aproximación normal: X ~ N(7.2, 4.835)
Frecuencia
1500
1000
500
0
0
5
10
15
X

Aproximación de la distribución Poisson a la Normal

Si $X \sim Pois(\lambda)$, entonces $X \approx N(\mu = \lambda, \sigma^2 = \lambda)$ siempre que $\lambda \geq 15$. Sin embargo se debe usar la corrección por continuidad.

$$P(X = a) = P(a - 0.5 \leq X \leq a + 0.5)$$

$$P(X \leq a) = P(X \leq a + 0.5)$$

Por ejemplo, si $X \sim Pois(50)$ entonces $X \approx N(\mu = 50, \sigma^2 = 50)$

$$P(X = 55) = 0.0422 \approx 0.0439$$

```
dpois(x = 55, lambda = 50)
```

```
[1] 0.04216435
```

```
pnorm(q=55.5, mean=50, sd=sqrt(50)) - pnorm(q=54.5, mean=50, sd=sqrt(50))
```

```
[1] 0.04392082
```

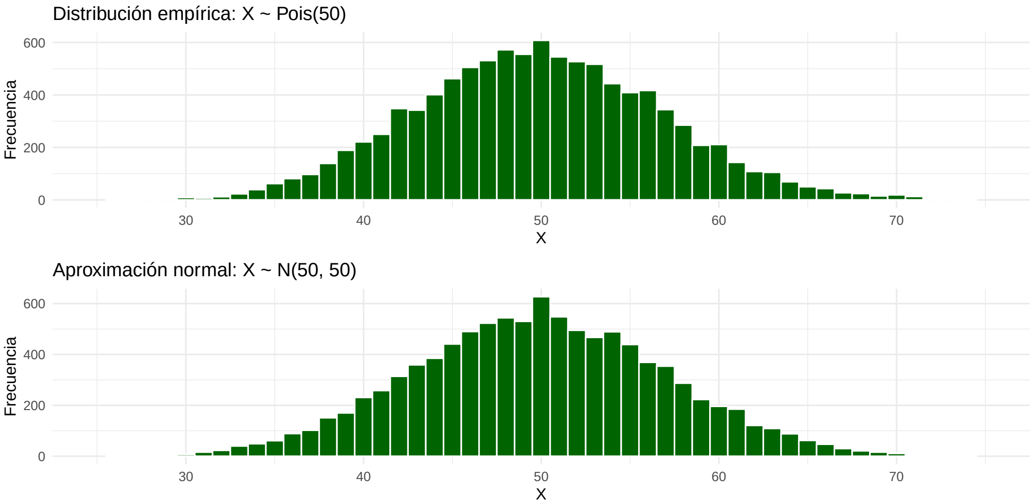
$$P(X \leq 60) = 0.9278 \approx 0.9312$$

```
ppois(q = 60, lambda = 50)
```

```
[1] 0.9278398
```

```
pnorm(q = 60.5, mean = 50, sd = sqrt(50))
```

```
[1] 0.9312181
```


Distribución empírica: X ~ Pois(50)
Frecuencia
600
400
200
0
30
40
50
60
70
X
Aproximación normal: X ~ N(50, 50)
Frecuencia
600
400
200
0
30
40
50
60
70
X

Distribución muestral de la media

Sea $X_1, ..., X_n$ una muestra de una distribución con media μ y varianza σ^2. Entonces, si el muestreo es con reemplazo de una población finita, o con o sin reemplazo de una población infinita:

- Caso 1: Si es una población con distribución normal: Entonces para cualquier n, $\overline{X}$ se distribuye como una Normal con media μ y varianza $\frac{\sigma^2}{n}$
- Caso 2. Si la población no proviene de una distribución normal. Entonces, para $n \geq 30$ suficientemente grande y por el El teorema del límite central, $\overline{X}$ se distribuye aproximadamnte como una Normal con media μ y varianza $\frac{\sigma^2}{n}$

Por lo tanto:

$$E(\overline{X}) = \mu_{\overline{X}} = \mu \qquad Var(\overline{X}) = \sigma^2_{\overline{X}} = \frac{\sigma^2}{n} \qquad \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Una clínica registra el tiempo de atención médica de sus pacientes como una variable que sigue una distribución normal con media 18 minutos y desviación estándar 6 minutos. Se toma una muestra aleatoria de $n=9$ pacientes. ¿Cuál es la probabilidad de que el tiempo promedio de atención sea mayor a 20 minutos?

En una ciudad, se estudian los ingresos diarios de trabajadores independientes. Se sabe que los ingresos son altamente asimétricos a la derecha, con algunos trabajadores que ganan mucho más que el promedio. Estudios previos han estimado que la media poblacional es de 120 soles y la desviación estándar es de 90 soles.Un investigador toma una muestra aleatoria de $n = 40$ trabajadores, ¿cuál es la probabilidad de que el ingreso promedio de la muestra esté entre 100 y 140 soles?

Distribución de una proporción muestral

Para n grande (mayor a 50) y utilizando el Teorema del Límite Central, se puede asegurar que la proporción muestral p, tiene una distribución aproximadamente normal con media $\mu_p = \pi$ y varianza $\frac{\sigma^2}{n} = \frac{\pi(1-\pi)}{n}$. Se estandariza como $Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$

Una encuesta nacional indica que aproximadamente el 62% de los adultos leen al menos un libro al mes. Un investigador realiza una encuesta a $n = 150$ adultos seleccionados al azar en una región del país. ¿Cuál es la probabilidad de que menos del 58% de los encuestados en esta muestra lean al menos un libro al mes?

Distribución t de Student

- Distribución continua utilizada principalmente para inferencia estadística cuando se trabaja con muestras pequeñas y la varianza poblacional es desconocida.
- Fue introducida por William Sealy Gosset bajo el seudónimo “Student”.
- A medida que los grados de libertad aumentan, la distribución t se aproxima a la distribución normal estándar.
- Sus colas son más pesadas que la distribución Normal.

Función de densidad

Se define la V.A. X con distribución Gamma con parámetros α y β, cuya función de densidad viene dada por:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad x \in (-\infty, \infty)$$

Media

$$\mu_X = E(X) = 0 \qquad \text{para } \nu > 1$$

Varianza

$$\sigma_X^2 = V(X) = \frac{\nu}{\nu - 2} \qquad \text{para } \nu > 2$$

Asimetría

La distribución t de Student es simétrica

Aplicaciones

- Pruebas de hipótesis para una o dos medias: test t de Student
- Construcción de intervalos de confianza para una o dos medias

Probabilidad de que $t_{(10)}$ sea menor que -2

$$P(t_{(10)} < -2) = 0.0367$$

```r
pt(q = -2, df = 10)
```

```
[1] 0.03669402
```

Probabilidad de que $t_{(100)}$ sea menor que -1

$$P(t_{(100)} < -1) = 0.16 \approx P(Z < -1)$$

```r
pt(q = -1, df = 100)
```

```
[1] 0.1598621
```

```r
pnorm(q = -1)
```

```
[1] 0.1586553
```

Percentil 49 de una distribución t de Student con 25 grados de libertad

```r
qt(p = 0.48, df = 25)
```

```
[1] -0.0506588
```

$$t_{(0.48,25)} = -0.051$$

Percentil 49 de una distribución t de Student con 90 grados de libertad

```r
qt(p = 0.49, df = 90)
```

```
[1] -0.02513868
```

```r
qnorm(p = 0.49)
```

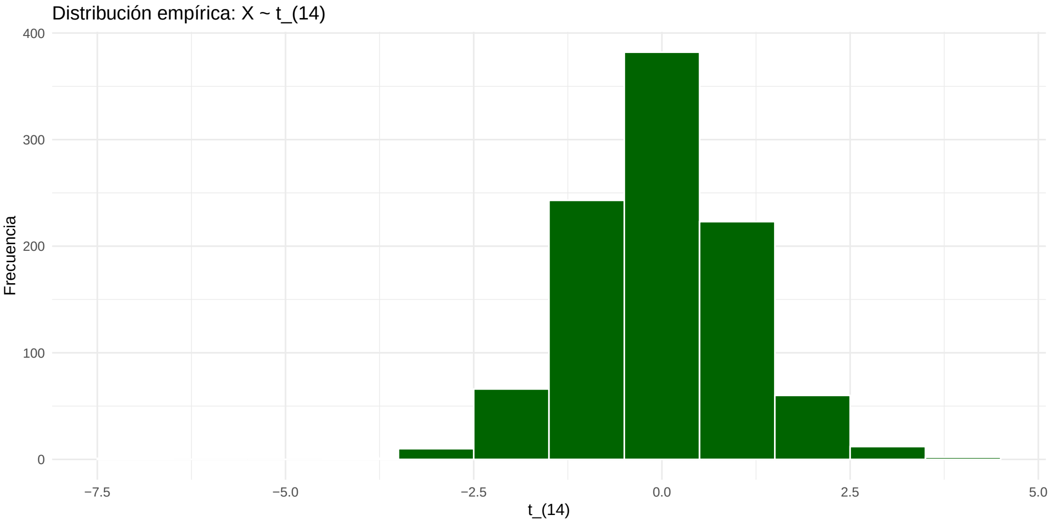
```
[1] -0.02506891
```

$$t_{(0.49,90)} = -0.0251 \approx Z_{0.49}$$

Simulación de una muestra aleatoria de 6 valores de $t_{(14)}$

```r
set.seed(2)
rt(n = 6, df = 14)
```

```
[1] -0.8982104  2.1171181  0.1514459 -0.1790589 -0.1331223  1.1038227
```



Distribución empírica: X ~ t_(14)
Frecuencia
400
300
200
100
0
−7.5
−5.0
−2.5
0.0
2.5
5.0
t_(14)

Aproximación de la t de Student a la Normal

Si $X \sim t(\nu)$, entonces $X \approx N(0,1)$ siempre que $\nu > 30$.
Por ejemplo, si $X \sim t(140)$ entonces $X \approx N(\mu = 0, \sigma^2 = 1)$

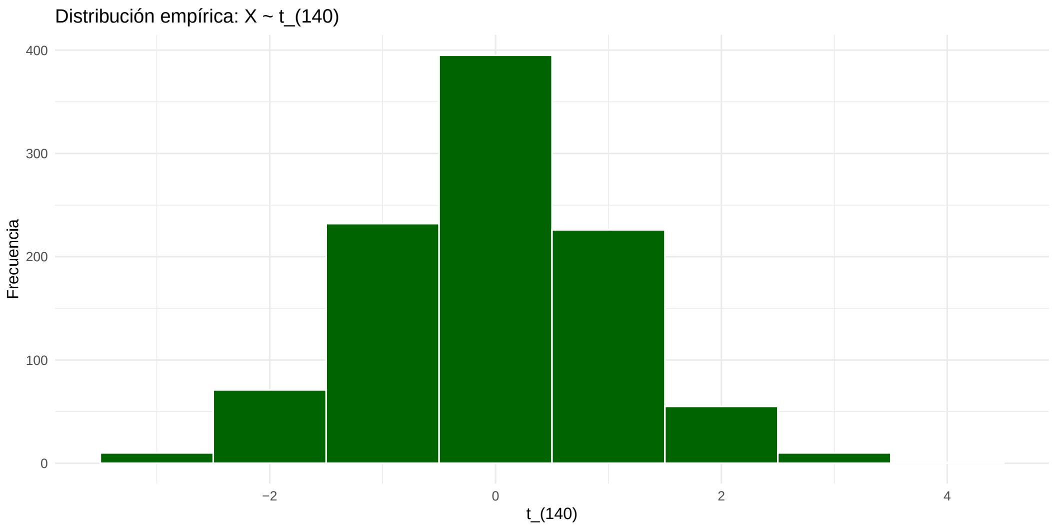
$$P(X \leq 2) = 0.9763 \approx 0.9772$$

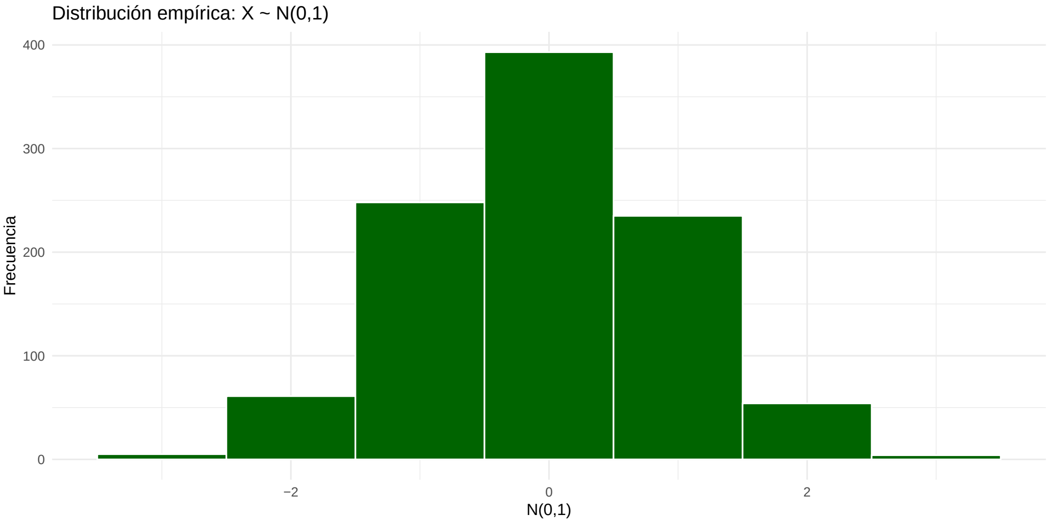
```r
pt(q = 2, df = 140)
```

```
[1] 0.9762826
```

```r
pnorm(q = 2, mean = 0, sd = 1)
```

```
[1] 0.9772499
```


Distribución empírica: X ~ t_(140)
Frecuencia
400
300
200
100
0
−2
0
2
4
t_(140)


Distribución empírica: X ~ N(0,1)
400
300
200
100
0
Frecuencia
−2
0
2
N(0,1)

Distribución Chi cuadrado

- Distribución continua utilizada en estadística inferencial, especialmente en pruebas de independencia, bondad de ajuste y estimación de varianza
- Se define como la distribución de la suma de los cuadrados de variables normales estándares independientes.
- A medida que los grados de libertad k aumentan, la distribución χ^2 se vuelve más simétrica y se aproxima a una normal.

Función de densidad

Se define la V.A. X con distribución Chi cuadrado con grado de libertad k, cuya función de densidad viene dada por:

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)}\, x^{(k/2-1)} e^{-x/2}, \quad x \geq 0$$

Notación: $X \sim \chi^2_{(k)}$

Media

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

Varianza

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

Simetría

Asimétrica a la derecha, especialmente para pocos grados de libertad

Aplicaciones

- Construcción de intervalos de confianza para la varianza de una variable aleatoria Normal
- Prueba de hipótesis para la varianza poblacional
- Prueba de bondad de ajuste (Chi-cuadrado de Pearson)
- Prueba de independencia en tablas de contingencia

Probabilidad de que $\chi^2_{(17)}$ sea menor a 13

$$P(\chi^2_{(17)} < 13) = 0.2638$$

```r
pchisq(q = 13, df = 17)
```

```
[1] 0.263814
```

Percentil 95 de una v.a. $\chi^2_{(2)}$

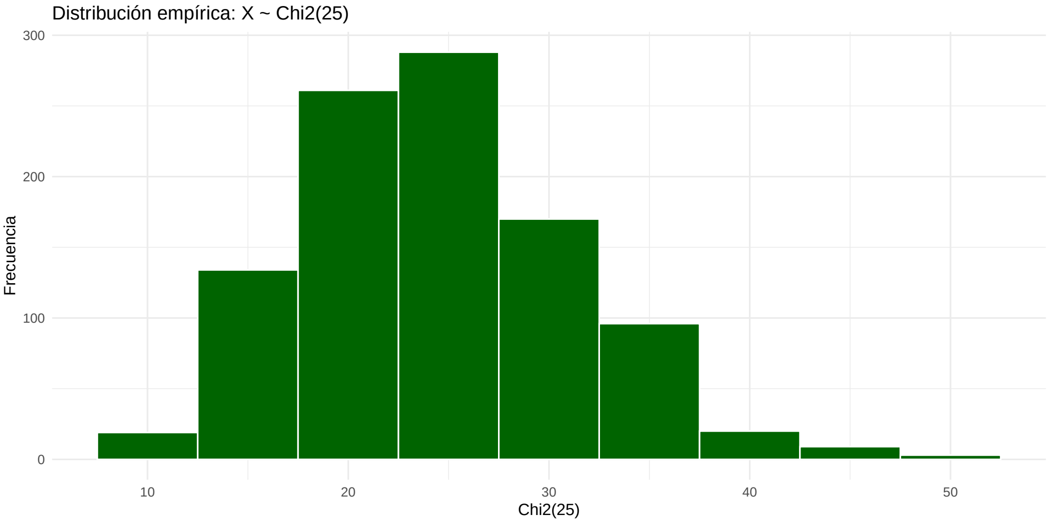
```r
qchisq(p = 0.95, df = 2)
```

```
[1] 5.991465
```

Muestra aleatoria de 6 observaciones de una variable aleatoria Chi cuadrado con 50 grados de libertad

```r
rchisq(n = 6, df = 50)
```

```
[1] 59.91733 58.19301 56.35884 51.14383 33.92111 45.16962
```


Distribución empírica: X ~ Chi2(25)
Frecuencia
300
200
100
0
10
20
30
40
50
Chi2(25)

Aproximación de la distribución Chi cuadrado a la Normal

Si $X \sim \chi^2(\nu)$, entonces $X \approx N(\mu = \nu, \sigma^2 = 2\nu)$ siempre que $\nu > 30$.
Por ejemplo, si $X \sim \chi^2(100)$ entonces $X \approx N(\mu = 100, \sigma^2 = 200)$

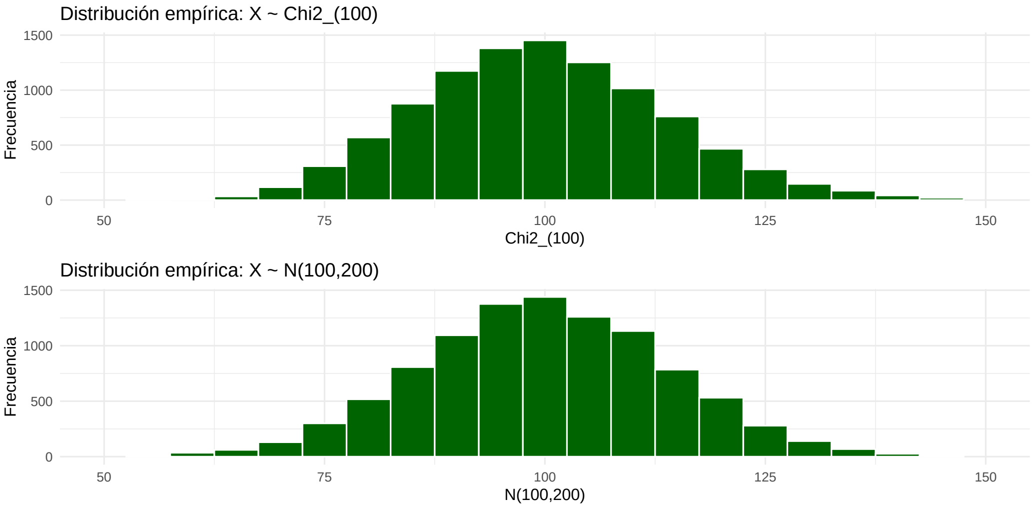
$$P(X \leq 110) = 0.7678 \approx 0.7602$$

```r
pchisq(q = 110, df = 100)
```

```
[1] 0.7677952
```

```r
pnorm(q = 110, mean = 100, sd = sqrt(200))
```

```
[1] 0.7602499
```


Distribución empírica: X ~ Chi2_(100)
Frecuencia
1500
1000
500
0
50
75
100
125
150
Chi2_(100)
Distribución empírica: X ~ N(100,200)
Frecuencia
1500
1000
500
0
50
75
100
125
150
N(100,200)

Distribución F

- Distribución continua que surge como razón de dos variables chi-cuadrado independientes, cada una dividida por sus respectivos grados de libertad.
- Es ampliamente utilizada en la comparación de varianzas, análisis de varianza (ANOVA), y en modelos de regresión.

Función de densidad

Se define la V.A. X con distribución F con parámetros d_1 y d_2, cuya función de densidad viene dada por:

$$f(x) = \frac{\Gamma\left(\frac{d_1+d_2}{2}\right)}{\Gamma\left(\frac{d_1}{2}\right)\Gamma\left(\frac{d_2}{2}\right)}\left(\frac{d_1}{d_2}\right)^{d_1/2}\frac{x^{(d_1/2-1)}}{\left(1+\frac{d_1}{d_2}x\right)^{(d_1+d_2)/2}}, \quad x > 0$$

Media

$$\mu_X = E(X) = \frac{d_2}{d_2 - 2}, \quad \text{si } d_2 > 2$$

Varianza

$$\sigma_X^2 = V(X) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 2)^2(d_2 - 4)}, \qquad \text{si } d_2 > 4$$

Asimetría

Asimétrica a la derecha; se aproxima a la normal cuando los grados de libertad son grandes

Aplicaciones

- razón de dos chi-cuadrado con a y b grados de libertad, $F(a, b)$
- cociente de cuadrados medios en el ANOVA

Probabilidad de que $F_{(12,19)}$ sea mayor a 1:

$$P(F_{(12,19)} > 1) = 0.4842$$

```r
pf(q = 1, df1 = 12, df2 = 19, lower.tail = F)
```

```
[1] 0.4842026
```

```r
1 - pf(q = 1, df1 = 12, df2 = 19)
```

```
[1] 0.4842026
```

Mediana de $F_{(18,2)}$

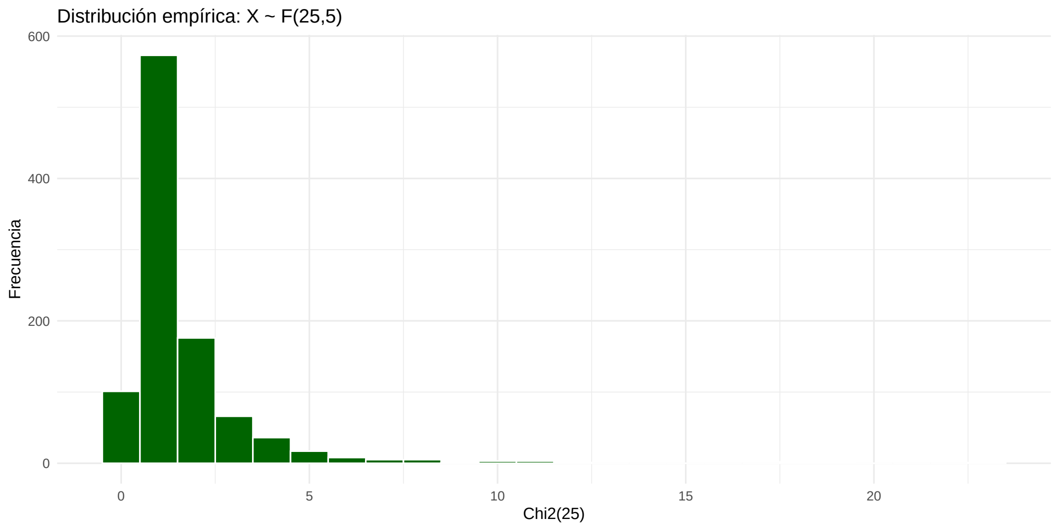
```r
qf(p = 0.5, df1 = 18, df2 = 2)
```

```
[1] 1.387853
```

Simulación de 6 valores $F_{(72,8)}$

```r
rf(n = 6, df1 = 72, df2 = 8)
```

```
[1] 1.7309381 1.8701737 1.8679992 0.9578774 0.5418074 1.1480291
```


Distribución empírica: X ~ F(25,5)
600
400
200
0
Frecuencia
0
5
10
15
20
Chi2(25)

Otras distribuciones continuas

- Distribución Beta
- Distribución Log-Normal
- Distribución Weibull
- Distribución Cauchy
- Distribución Pareto
- Distribución Rayleigh