Ciencia de Datos para Políticas Públicas

Nivelación

Pablo Aguirre Hormann
03/06/2021

Nivelación

Objetivos

- Repasar conceptos que probablemente han visto antes, pero de forma aplicada y con ejemplos en `R`
- Generar un lenguaje común

En el mundo de la Estadística/Econometría/Machine Learning se habla generalmente de **modelos**.

Queremos ver conceptos que nos ayuden a entender mejor ese "mundo" (y su nomenclatura) al revisarlos en la 2da parte del módulo 2.

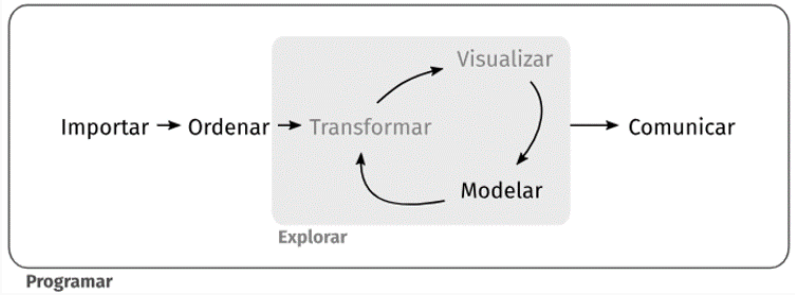
¿Para qué?

En particular

```
Call:
lm(formula = ROLL ~ UNEM + HGRAD + INC, data = datavar)

Residuals:
     Min       1Q   Median       3Q      Max
-1148.840  -489.712    -1.876   387.400  1425.753

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.153e+03  1.053e+03  -8.691 5.02e-09 ***
UNEM         4.501e+02  1.182e+02   3.809 0.000807 ***
HGRAD        4.065e-01  7.602e-02   5.347 1.52e-05 ***
INC          4.275e+00  4.947e-01   8.642 5.59e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 670.4 on 25 degrees of freedom
Multiple R-squared: 0.9621, Adjusted R-squared: 0.9576
F-statistic: 211.5 on 3 and 25 DF,  p-value: < 2.2e-16
```

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

En general



¿Qué veremos?

- Conceptos estadísticos
  - Conjuntos
  - Probabilidad
  - Variables discretas y continuas
  - Medidas de tendencia central y dispersión
  - Teorema del límite central
  - Distribuciones conjuntas
  - Distribuciones: Normal, χ^2, t, F
- `R`

Conjuntos

Proceso aleatorio

En un proceso aleatorio hay más de una posibilidad de resultado y la predicción de algún resultado particular es difícil:

Clásicos ejemplos:

- tirar una moneda
- tirar un dado

Proceso aleatorio

Espacio muestral

Es el conjunto de todos los posibles resultados de un proceso aleatorio.

- Dados: $\mathrm{S} = \{1, 2, 3, 4, 5, 6\}$
- Moneda: $\mathrm{S} = \{\mathrm{C}, \mathrm{S}\}$

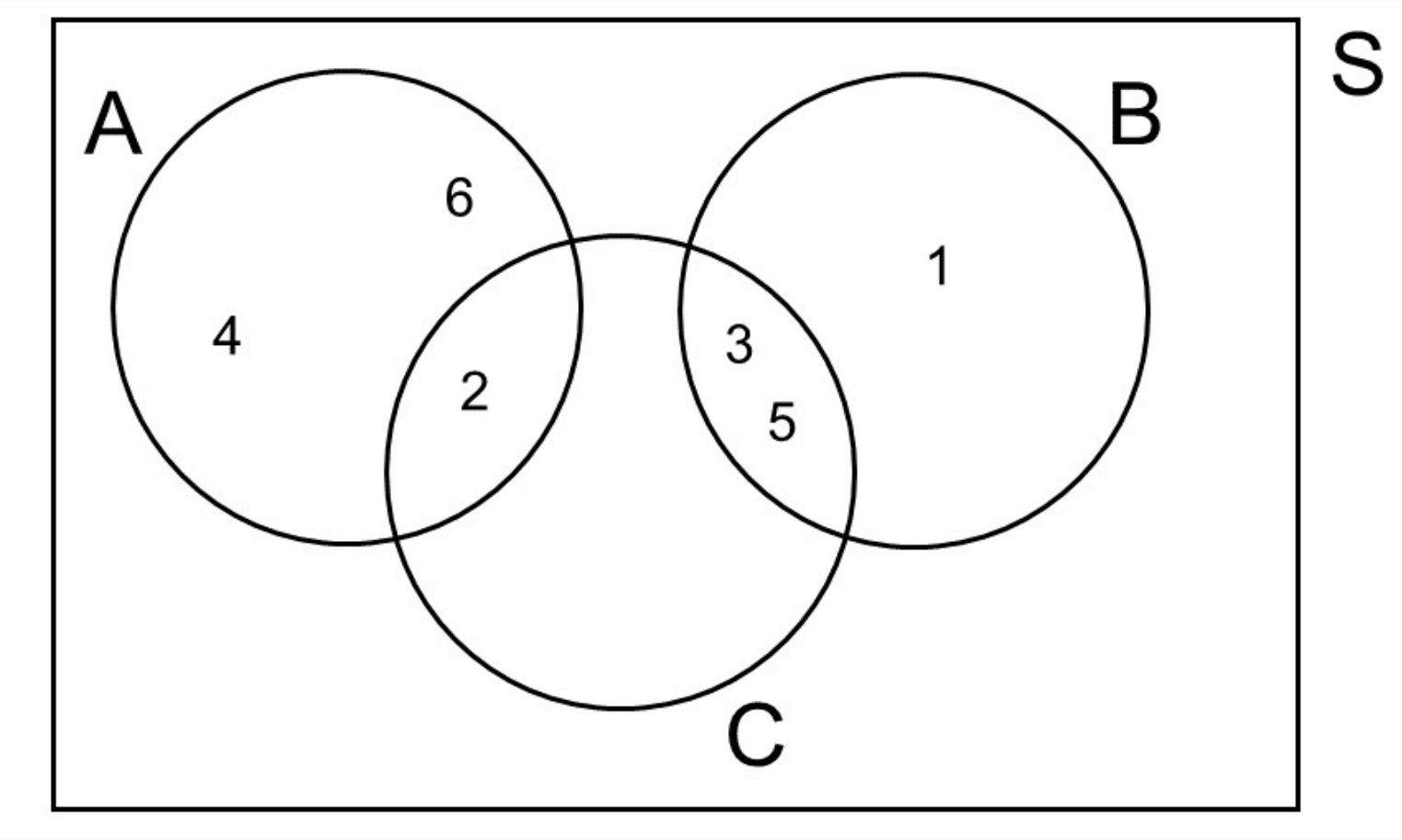
Evento

Es un subconjunto del espacio muestral.

Para el caso del dado, digamos que:

- A representa el evento de que al tirar un dado el resultado es un número par: $\mathrm{A} = \{2, 4, 6\}$
- B representa el evento de que al tirar un dado el resultado es un número impar: $\mathrm{B} = \{1, 3, 5\}$
- C representa el evento de que al tirar un dado el resultado es un número primo: $\mathrm{C} = \{2, 3, 5\}$

Diagrama de Venn



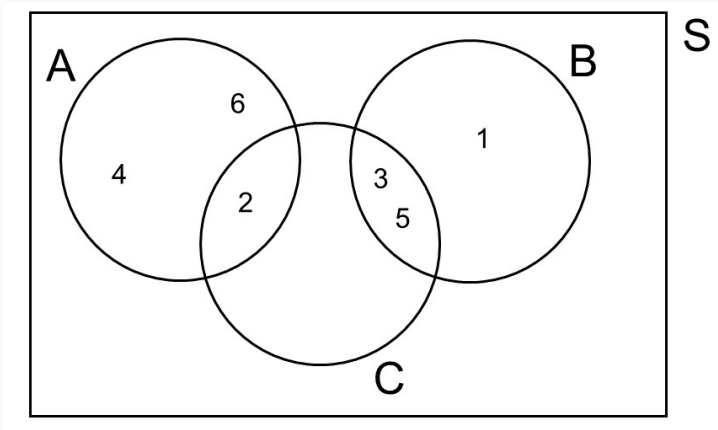
Complemento de un evento

Definición: el complemento de un evento son todos los resultados en el espacio muestral que no son el evento mismo.

Ejemplo: el complemento del evento C es el conjunto de resultados de tirar un dado que no corresponden a números primos.

Notación: C^c

$C^c : \{1, 4, 6\}$



Complemento de un evento

Definir tres *vectores* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$

```
A ← c(2,4,6)
B ← c(1,3,5)
C ← c(2,3,5)
```

Definir un vector que contiene los elementos de $\mathbf{A}$ y $\mathbf{B}$

```
AyB ← union(A, B)
```

Ver que elementos que están en $\mathbf{A}$ o $\mathbf{B}$, no están en $\mathbf{C}$

```
setdiff(AyB, C)
```

```
## [1] 4 6 1
```

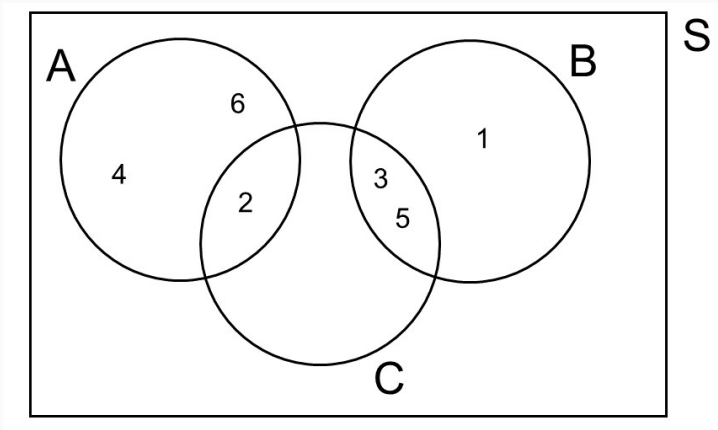
Para ver que hace una función (por ejemplo `setdiff()`) usar `?setdiff`

Eventos mutuamente excluyentes

Definición: los eventos mutuamente excluyentes son eventos que no pueden ocurrir al mismo tiempo.

Ejemplo: el evento A y el evento B son mutuamente excluyentes porque el resultado de tirar un dado no puede ser par e impar al mismo tiempo.

Los eventos A y C no son mutuamente excluyentes solamente porque 2 es tanto par como primo.



Eventos mutuamente excluyentes

¿Existe al menos un elemento de A igual a un elemento en C?

```
intersect(A, C)
```

```
## [1] 2
```

A y C **no** son mutuamente excluyentes

¿Existe al menos un elemento de B igual a un elemento en C?

```
intersect(B, C)
```

```
## [1] 3 5
```

B y C **no** son mutuamente excluyentes

¿Existe al menos un elemento de A igual a un elemento en B?

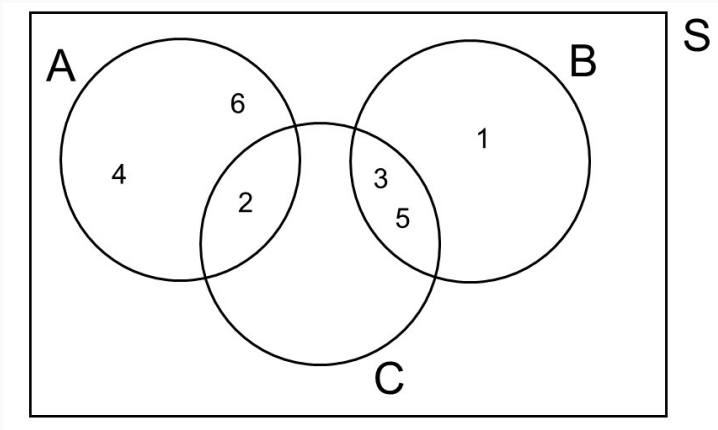
```
intersect(A, B)
```

```
## numeric(0)
```

A y B son mutuamente excluyentes

Notación de conjuntos

Descripción	Notación	Lectura	Elementos
Unión	$A \cup C$	A o C	$\{2, 3, 4, 5, 6\}$
Intersección	$A \cap C$	A y C	$\{2\}$



Notación de conjuntos

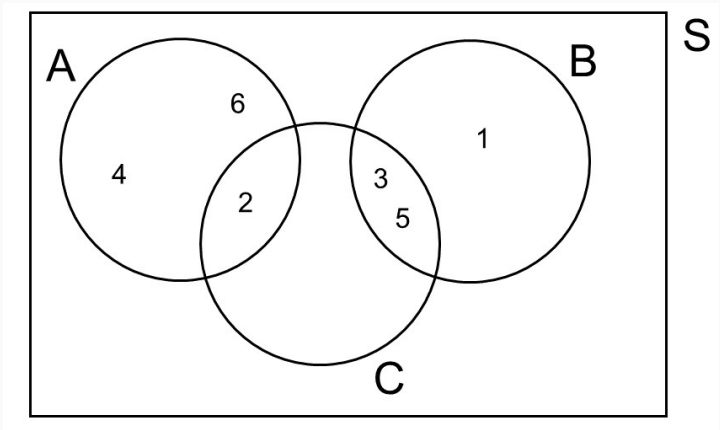
Descripción	Notación	Lectura	Elementos
Unión	$A \cup C$	A o C	$\{2, 3, 4, 5, 6\}$
Intersección	$A \cap C$	A y C	$\{2\}$

```
union(A, C)
```

```
## [1] 2 4 6 3 5
```

```
intersect(A, C)
```

```
## [1] 2
```



Probabilidades

Definamos probabilidad

Definición "frecuentista"

La probabilidad de un resultado está definida por la **proporción de veces** que ese resultado es observado a través de un **alto número de repeticiones de procesos aleatorios**.

Asuman que repetimos el proceso aleatorio de tirar una moneda y que registramos X, el número de veces que sale *cara* en n tiradas de moneda. Entonces:

$$\mathrm{P(Cara)} = \lim_{\mathrm{n}\to\infty} \frac{\mathrm{X}}{\mathrm{n}}$$

$$\mathrm{P(Cara)} = \frac{1}{2}$$

En la práctica

Definamos un vector *moneda* con los posibles eventos y simulemos una sola tirada de esa moneda.

```
moneda ← c("cara", "sello")
sample(moneda, 1)
```

```
## [1] "cara"
```

Repitamos 5 veces este proceso

```
tirar_moneda ← sample(moneda, 5, replace = TRUE)
table(tirar_moneda)
```

```
## tirar_moneda
##  cara sello
##     3     2
```

Dados estos resultados, vemos que $\mathrm{P}(\mathrm{Cara}) = \frac{3}{5} = 60\%$

En la práctica

Definamos un vector *moneda* con los posibles eventos y simulemos una sola tirada de esa moneda.

```
moneda ← c("cara", "sello")
sample(moneda, 1)
```

```
## [1] "cara"
```

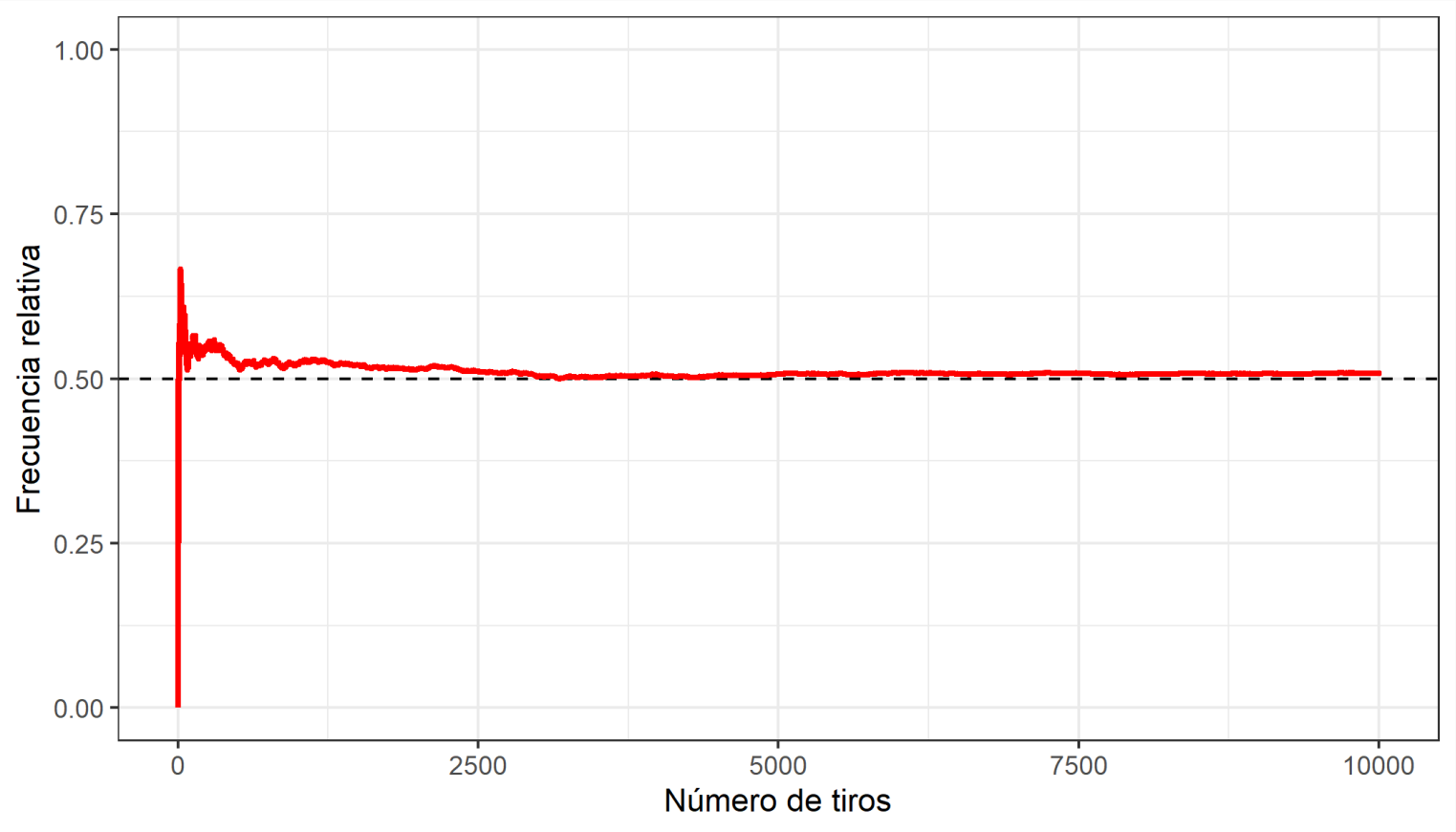
Repitamos 5.000 veces este proceso

```
tirar_moneda ← sample(moneda, 5000, replace = TRUE)
table(tirar_moneda)
```

```
## tirar_moneda
##  cara sello
##  2449  2551
```

En este caso vemos que $\mathrm{P}(\mathrm{Cara}) = \frac{2448}{5000} = 48,96\%$

En la práctica



Definición "bayesiana"

Al definir una probabilidad, además de considerar el número de veces que un resultado ocurre, la aproximación *bayesiana* considera información previa (*prior*) que se tenga sobre ese resultado.

No profundizaremos en este tema pero lo dejo como antecedente.

Cosas a saber sobre las probabilidades

- $0 \leq P(A) \leq 1$: La probabilidad de cualquier evento se encuentra entre cero y uno
- $P(S) = 1$: El conjunto de eventos es igual a 1

Probabilidad de eventos complementarios

$P(A) + P(A^c) = 1$. Por ejemplo, si sabemos que la **probabilidad de que llueva es 0.1**, entonces sabemos también que la **probabilidad de que no llueva es 0.9**.

Regla de Independencia y Multiplicación

Dos procesos son independientes si el saber el resultado de uno no afecta el resultado de otro. Por ejemplo, tirar dos monedas al aire.

Si el evento A y B son independientes, entonces: $P(A \cap B) = P(A) \times P(B)$

La probabilidad de sacar dos *caras* en dos tiradas de una misma moneda es: $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

Probabilidad Conjunta/Marginal/Condicional

GSS 2018

La *General Social Survey* (GSS) es una encuesta sociológica que se hace en Estados Unidos desde 1972. Es una encuesta muy completa que entrega información sobre distintos aspectos sobre los residentes del país.

		Cree en la vida después de la muerte		
		Sí	No	Total
Curso Universitario de Cs.	Sí	375	75	450
	No	485	115	600
	Total	860	190	1050

Eventos

Digamos que B representa el evento de aleatoriamente seleccionar una persona de esta muestra que **cree en la vida después de la muerte**.

Digamos que C representa el evento de aleatoriamente seleccionar una persona de esta muestra que **ha tomado un curso universitario de ciencias**.

Probabilidad marginal

		Cree en la vida después de la muerte		
		Sí	No	Total
Curso Universitario de Cs.	Sí	375	75	450
	No	485	115	600
	Total	860	190	1050

$\mathrm{P}(\mathrm{B})$ representa una *probabilidad marginal*. Así también $\mathrm{P}(\mathrm{C})$, $\mathrm{P}(\mathrm{B}^{\mathrm{C}})$, y $\mathrm{P}(\mathrm{C}^{\mathrm{C}})$.

Para calcular estas probabilidad solo debemos usar los valores en los márgenes de la tabla (por eso el nombre).

$$\mathrm{P}(\mathrm{B}) = \frac{860}{1050}$$

$$\mathrm{P}(\mathrm{C}) = \frac{450}{1050}$$

Probabilidad conjunta

		Cree en la vida después de la muerte		
		Sí	No	Total
Curso Universitario de Cs.	Sí	375	75	450
	No	485	115	600
	Total	860	190	1050

Noten que los eventos B y C **no son mutualmente excluyentes**. Una persona seleccionada aleatoriamente puede creer en la vida después de la muerte y podría haber tomado un curso universitario de ciencias. $B \cap C \neq \emptyset$

$$P(B \cap C) = \frac{375}{1050}$$

Noten también que $P(B \cap C) = P(C \cap B)$. El orden no importa.

Regla de adición

		Cree en la vida después de la muerte		
		Sí	No	Total
Curso Universitario de Cs.	Sí	375	75	450
	No	485	115	600
	Total	860	190	1050

$$P(B \cup C) = P(B) + P(C) - P(B \cap C)$$

$$P(B \cup C) = \frac{860}{1050} + \frac{450}{1050} - \frac{375}{1050} = \frac{935}{1050}$$

$$P(B \cup C) = P(B^c) + P(C^c) + P(B \cap C)$$

$$P(B \cup C) = \frac{485}{1050} + \frac{75}{1050} + \frac{375}{1050} = \frac{935}{1050}$$

Probabilidad condicional

		Cree en la vida después de la muerte		
		Sí	No	Total
Curso Universitario de Cs.	Sí	375	75	450
	No	485	115	600
	Total	860	190	1050

$\mathrm{P(B|C)}$ representa una *probabilidad condicional* ("Probabilidad de B dado C"). Así también $\mathrm{P(B^C|C)}$, $\mathrm{P(C|B)}$, y $\mathrm{P(C|B^C)}$.

Para calcular estas probabilidades nos enfocamos en las filas o columnas de la *información dada* o la condición. En otras palabras, reducimos el espacio muestral a esta información dada o condición.

La probabilidad de que una persona seleccionada aleatoriamente **crea en la vida después de la muerta DADO que ha tomado un curso universitario de ciencias**:

$$\mathrm{P(B|C)} = \frac{375}{450}$$

El orden acá si importa: $\mathrm{P(B|C)} \neq \mathrm{P(C|B)}$

Probabilidad condicional

Teorema de Bayes

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Nuevamente, no profundizaremos en este tema. Pero bueno tenerlo en el radar.

En el módulo tres hablarán un poco más sobre este tema (*Naive Bayes*)

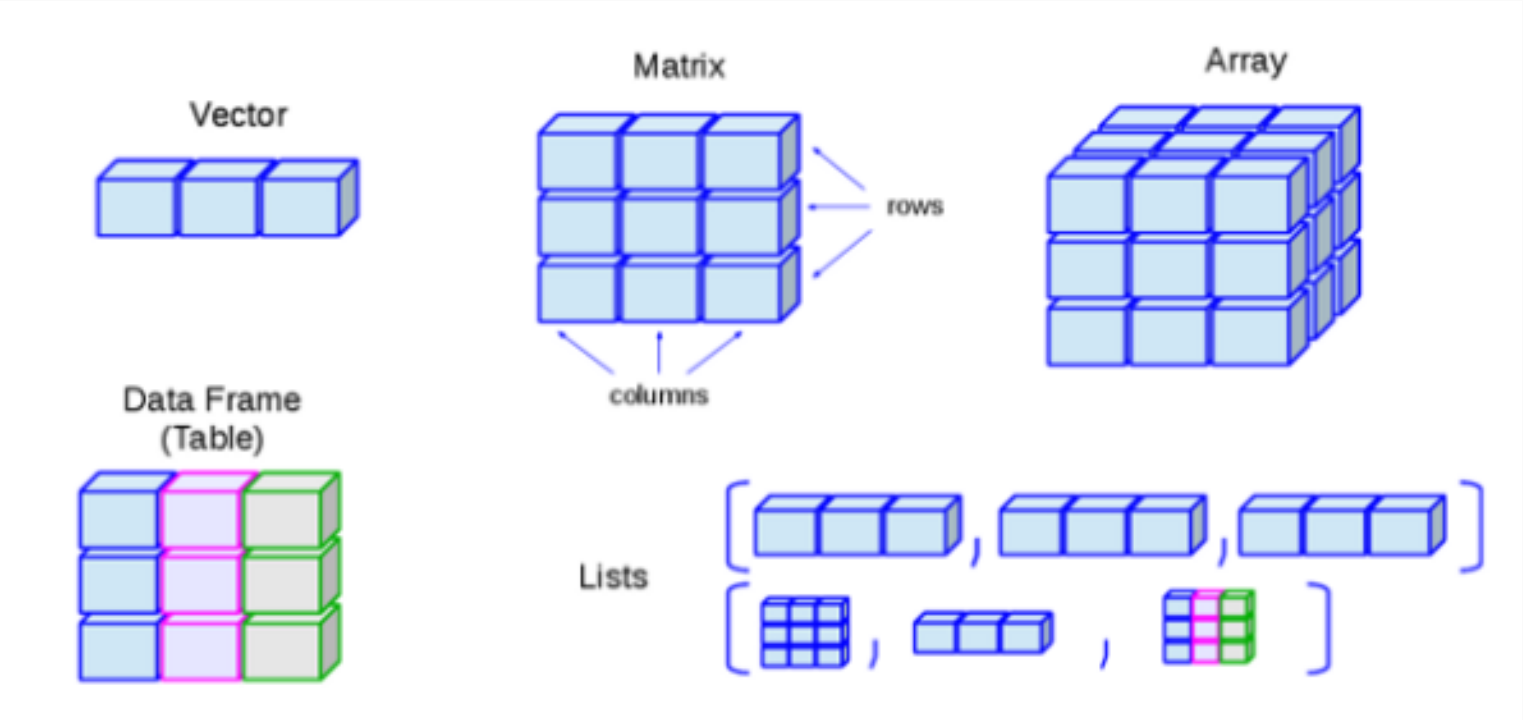
Ejercicio

Caso hipotético

		País		
		Narnia	Wakanda	Total
¿Tiene COVID?	Sí	40	150	190
	No	210	350	560
	Total	250	500	750

`EjercicioProbabilidadesDisc.R`

Tengan en cuenta



Respuestas

		País		
		Narnia	Wakanda	Total
¿Tiene COVID?	Sí	40	150	190
	No	210	350	560
	Total	250	500	750

P(COV ID No)

```
tabla[2,3]/tabla[3,3]
```

```
## [1] 0.7466667
```

P(Narnia ∩ COV ID No)

```
tabla[2,1]/tabla[3,3]
```

```
## [1] 0.28
```

P(COV ID Si|Narnia)

```
tabla[1,1]/tabla[3,1]
```

```
## [1] 0.16
```

P(Wakanda|COV ID No)

```
tabla[2,2]/tabla[2,3]
```

```
## [1] 0.625
```

Variables aleatorias discretas

Variables aleatorias discretas

Una función que asigna un resultado numérico a un proceso aleatorio se conoce como **variable aleatoria**.

Para una **variable aleatoria discreta**, estos resultados numéricos toman valores aislados (entre dos números no existe ninguno intermedio) y los posibles valores son finitos.

Usaremos letras mayúsculas, como X, Y , y Z, para representar variables aleatorias y minúsculas (x, y, y z) para indicador resultados observados.

Variable aleatorias discretas

Proceso aleatorio: Tirar 3 veces una misma moneda.

Espacio muestral: $\{CCC, CCS, CSC, SCC, CSS, SCS, SSC, SSS\}$

Digamos que X es una variable aleatoria discreta que representa el número total de *caras* en 3 tiros de moneda. ¿Cuál es $P(X = 2)$?

$P(X = 2) = \frac{3}{8} = 0.375$

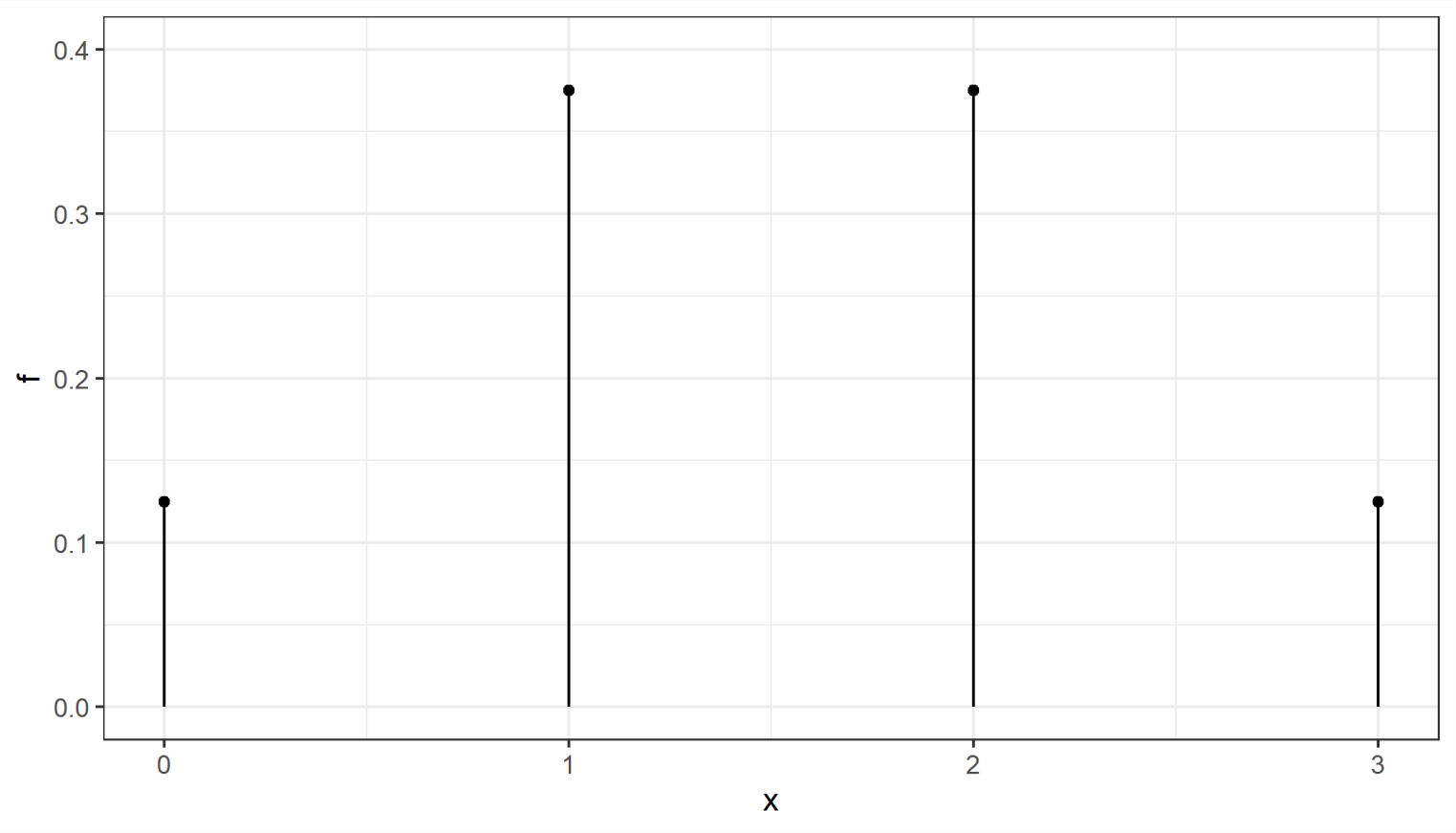
x	0	1	2	3
$P(X = x) = f(x)$	f(0)=0,125	f(1)=0,375	f(2)=0,375	f(3)=0,125

```
resultados ← replicate(expr = sum(sample(moneda, 3, replace = TRUE) == "cara"), n = 10000)
table(resultados)/10000
```

```
## resultados
##      0      1      2      3
## 0.1242 0.3705 0.3765 0.1288
```

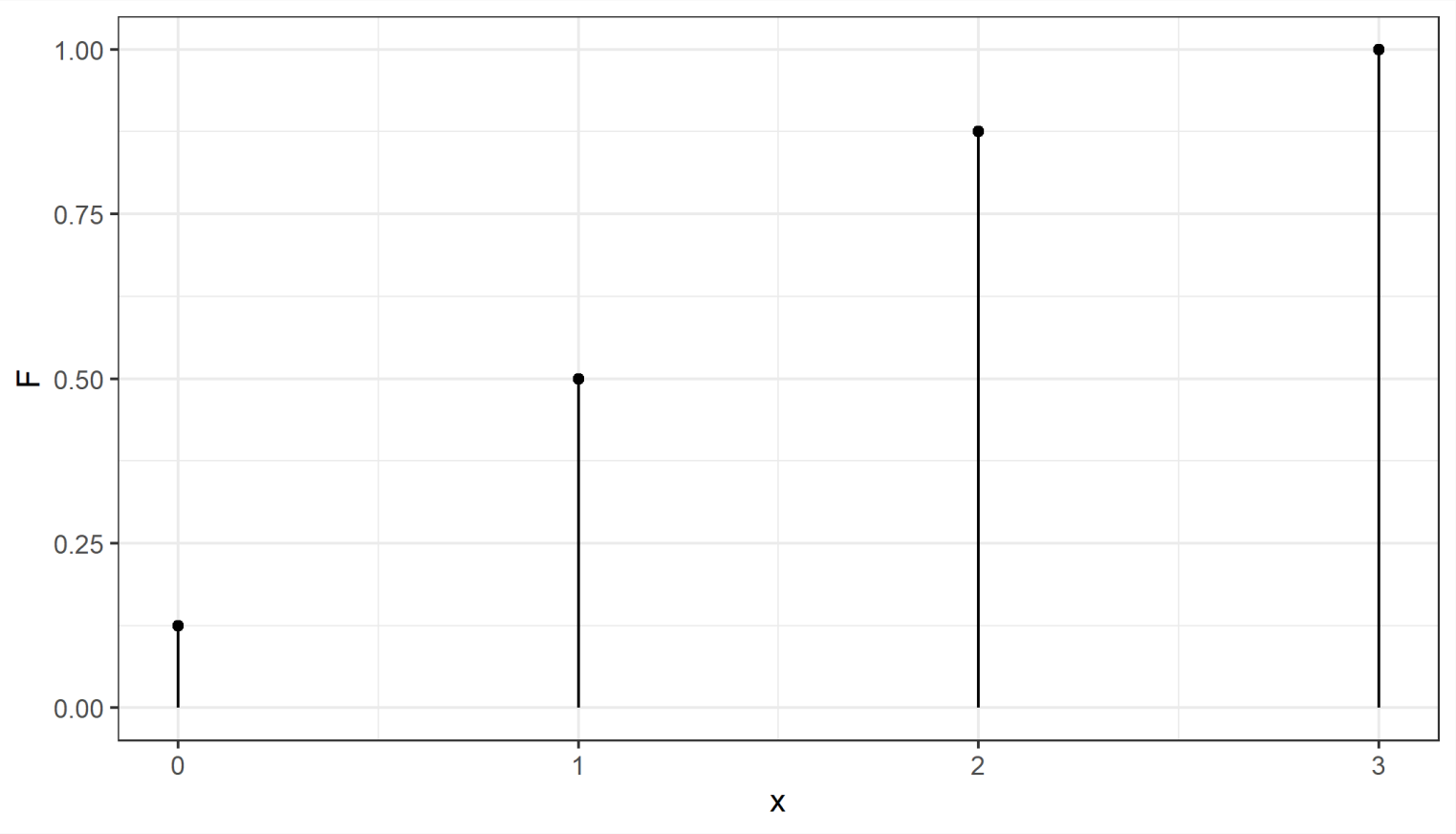
Función de Masa de Probabilidad

Para una variable aleatoria discreta, X, la distribución de todos los valores posibles de x puede ser graficada con una Función de Masa de Probabilidad (Probability Mass Function - **pmf**)



Función de Distribución Acumulada

	x	0	1	2	3
pmf	$P(X = x) = f(x)$	$f(0)=0,125$	$f(1)=0,375$	$f(2)=0,375$	$f(3)=0,125$
cdf	$P(X <= x) = F(x)$	$F(0)=0,125$	$F(1)=0,5$	$F(2)=0,875$	$F(3)=1$



Medidas de tendencia central

$$\{1, 2, 3, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Esperanza o Valor Esperado

Valor promedio/medio (ponderado).

```
mean(c(1,2,3,3,4,5,6,7,8,9,10,11,12))
```

```
## [1] 6.230769
```

Mediana

Valor (x) que divide las observaciones en dos grupos de igual tamaño.

```
median(c(1,2,3,3,4,5,6,7,8,9,10,11,12))
```

```
## [1] 6
```

Moda

Valor con la mayor probabilidad de ocurrencia (o mayor valor de la función de densidad)

```
table(c(1,2,3,3,4,5,6,7,8,9,10,11,12))
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12
##  1  1  2  1  1  1  1  1  1  1  1  1
```

Esperanza/Valor Esperado

Si repetimos muchas veces este proceso aleatorio de tirar una moneda tres veces, ¿cuantas *caras* esperamos ver en promedio?

x	0	1	2	3
$P(X = x) = f(x)$	0,125	0,375	0,375	0,125

$$E(X) = \sum [x * f(x)]$$

$$E(X) = (0 \times 0.125) + (1 \times 0.375) + (2 \times 0.375) + (3 \times 0.125)$$

$$E(X) = \mu = 1.5$$

```
mean(resultados)
```

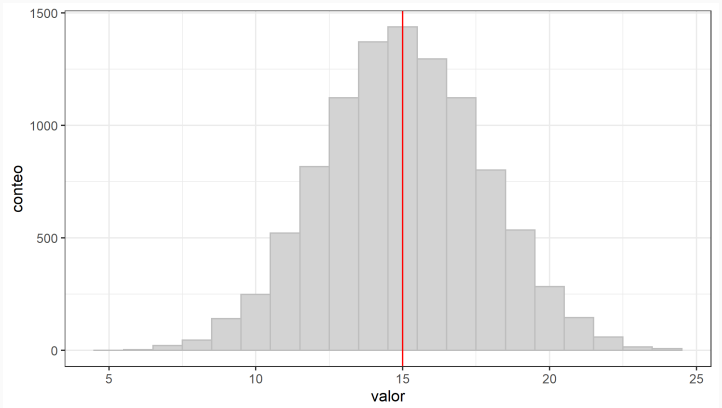
```
## [1] 1.5099
```

¿Cómo interpretamos que al tirar tres monedas el valor esperado sean 1.5 caras?

Interpretación en la práctica

Podemos decir que $\mu = 1.5$ al tirar tres monedas es lo mismo que esperar obtener 15 caras al repetir este proceso 10 veces. De hecho:

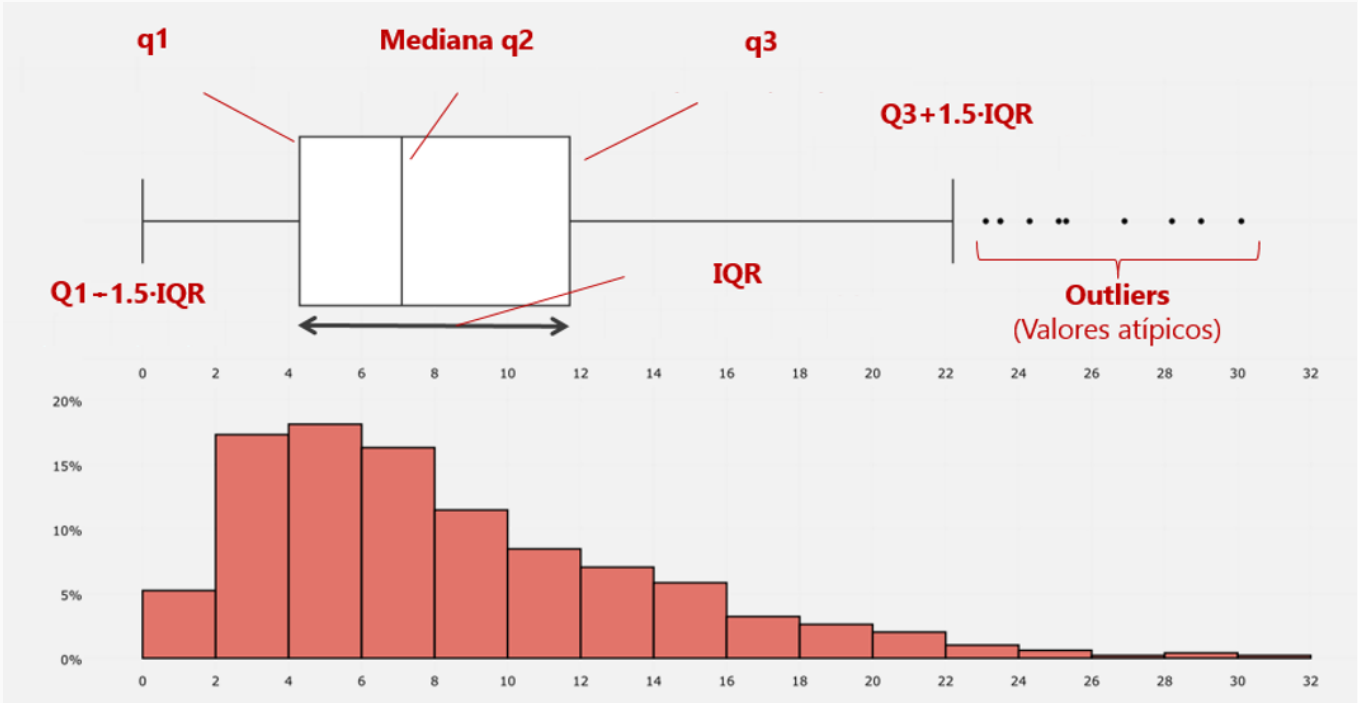
```
resultado_final ← vector()
for (i in 1:10000){
  resultado_por_proceso ← vector()
  for (j in 1:10){
    resultado_por_proceso[j] ← sum(sample(moneda,
  }
  resultado_final[i] ← sum(resultado_por_proceso)
}
```



Si simulamos 10.000 escenarios donde tiramos 10 veces 3 monedas y contamos el número de *caras*, obtenemos que el valor más probable (esperado) es 15.

Medidas de dispersión

- Varianza
- Desviación estándar
- Coeficiente de variación
- Rango intercuartil



Varianza

Probablemente han visto algo como esto:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i}^{n}(x_i - \mu)^2$$

Otra forma de verlo (una generalización):

$$\mathrm{Var}(X) = E[(X - \mu)^2]$$

$$\mathrm{Var}(X) = E[(X - E(X))^2]$$

$$\mathrm{Var}(X) = \sigma^2 = E(X^2) - [E(X)]^2$$

Varianza

x	0	1	2	3
P(X = x) = f(x)	0,125	0,375	0,375	0,125

$$Var(X) = E(X^2) - E[(X)]^2 = E(X^2) - 1.5^2$$

$$E(X^2) = \sum [x^2 * f(x)]$$

$$E(X^2) = (0^2 \times 0.125) + (1^2 \times 0.375) + (2^2 \times 0.375) + (3^2 \times 0.125)$$

$$E(X^2) = 3$$

$$Var(X) = 3 - 1.5^2 = 3 - 2.25 = 0.75 = \sigma^2$$

```
var(resultados)
```

```
## [1] 0.7559776
```

Desviación estándar: $\sigma = \sqrt{0.75} = 0.8660254$

Algo interesante...

La variable aleatoria, X, correspondiente al resultado de tirar tres monedas **sigue una distribución binomial** que depende de, n, número de intentos de un proceso aleatorio y, p, la probabilidad de éxito.

n y p los denominamos como **parámetros de la distribución**.

$$X \sim \text{Binomial}(n, p)$$

$S = \{0, 1, \ldots, n\}$

$P(X = x) = f(x) = \binom{n}{k} p^x (1-p)^{n-x}$

$E(X) = np$

$Var(X) = np(1-p)$

```
dbinom(x = 3, size = 3, prob = 0.5) #pmf
```

```
## [1] 0.125
```

```
pbinom(q = 3, size = 3, prob = 0.5) #cdf
```

```
## [1] 1
```

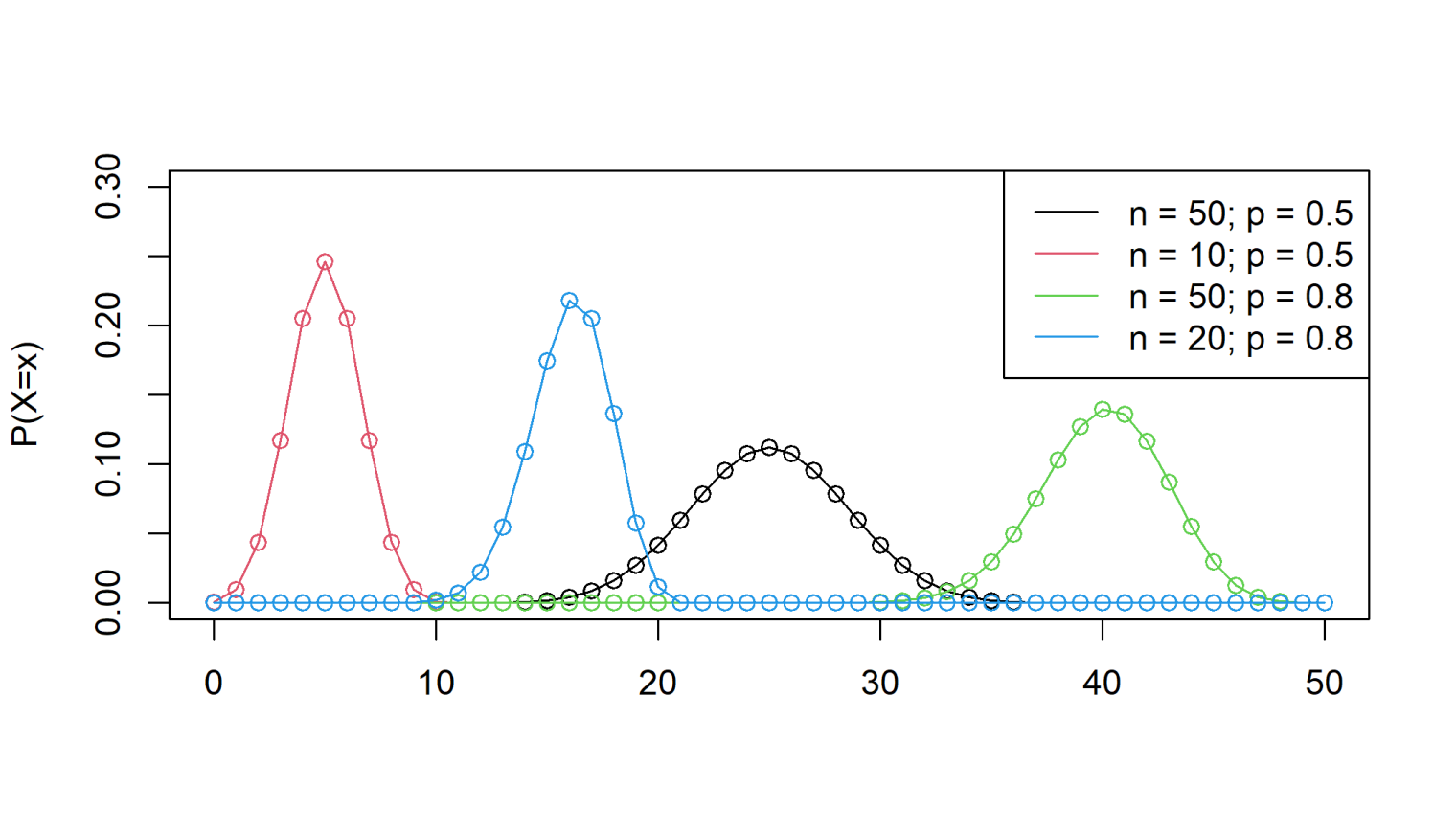
$E(X) = 3 \times 0.5 = 1.5$

$Var(X) = 3 \times 0.5 \times (1 - 0.5) = 0.75$

Algo interesante...

La variable aleatoria, X, correspondiente al resultado de tirar tres monedas **sigue una distribución binomial** que depende de, n, número de intentos de un proceso aleatorio y, p, la probabilidad de éxito.

n y p los denominamos como **parámetros de la distribución**.



Ejercicio

Suma de tirar dos dados

		Dado 1					
		1	2	3	4	5	6
Dado 2	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

`EjercicioDatos.R`

Suma de tirar dos dados

		Dado 1					
		1	2	3	4	5	6
Dado 2	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

$S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

¿Cuál es la *pmf* y *cdf*?

		2	3	4	5	6	7	8	9	10	11	12
pmf	$P(X = x) = f(x)$											
cdf	$P(X <= x) = F(x)$											

Calcule $E(X)$, $Var(X)$, y σ

Respuestas

¿Cuál es la *pmf* y *cdf*?

		2	3	4	5	6	7	8	9	10	11	12
pmf	$P(X = x) = f(x)$	0.028	0.056	0.083	0.111	0.139	0.167	0.139	0.111	0.083	0.056	0.028
cdf	$P(X <= x) = F(x)$	0.028	0.083	0.167	0.278	0.417	0.583	0.722	0.833	0.917	0.972	1

```
pmf ← table(combinaciones$suma_dados)/nrow(combinaciones)
pmf
```

```
##
##          2          3          4          5          6          7          8
## 0.02777778 0.05555556 0.08333333 0.11111111 0.13888889 0.16666667 0.13888889
##          9         10         11         12
## 0.11111111 0.08333333 0.05555556 0.02777778
```

Respuestas

¿Cuál es la *pmf* y *cdf*?

		2	3	4	5	6	7	8	9	10	11	12
pmf	$P(X = x) = f(x)$	0.028	0.056	0.083	0.111	0.139	0.167	0.139	0.111	0.083	0.056	0.028
cdf	$P(X <= x) = F(x)$	0.028	0.083	0.167	0.278	0.417	0.583	0.722	0.833	0.917	0.972	1

```
cdf ← cumsum(pmf)
cdf
```

```
##          2          3          4          5          6          7          8
## 0.02777778 0.08333333 0.16666667 0.27777778 0.41666667 0.58333333 0.72222222
##          9         10         11         12
## 0.83333333 0.91666667 0.97222222 1.00000000
```

Respuestas

Calcule $E(X)$, $Var(X)$, y σ

Desde las formulas

```
espacio_muestral ← unique(combinaciones$suma_dados
```

```
(e_x ← sum(espacio_muestral*pmf))
## [1] 7
(varianza_dados ← (sum(espacio_muestral^2*pmf))-(s
## [1] 5.833333
sqrt(varianza_dados)
## [1] 2.415229
```

Simular

```
guardar_suma ← replicate(
  expr = sum(sample(1:6, 2,replace = TRUE)),
  n = 10000)

mean(guardar_suma)
## [1] 6.9982
var(guardar_suma)
## [1] 5.902587
sd(guardar_suma)
## [1] 2.429524
```

Otras distribuciones discretas

Distribución Geométrica: X es el número de *fracasos* antes de observar el primer *éxito* en procesos independientes.

- $X \sim Geom(p)$

- $f(x) = (1-p)^x p$
- $E(X) = \frac{1-p}{p}$
- $Var(X) = \frac{1-p}{p^2}$

- `dgeom(x, prob)` (pmf) y `pgeom(q, prob)` (cdf)

Distribución Poisson: X es el número de ocurrencias de un evento dentro de un espacio o tiempo **finito**.

- $X \sim Poisson(\lambda)$

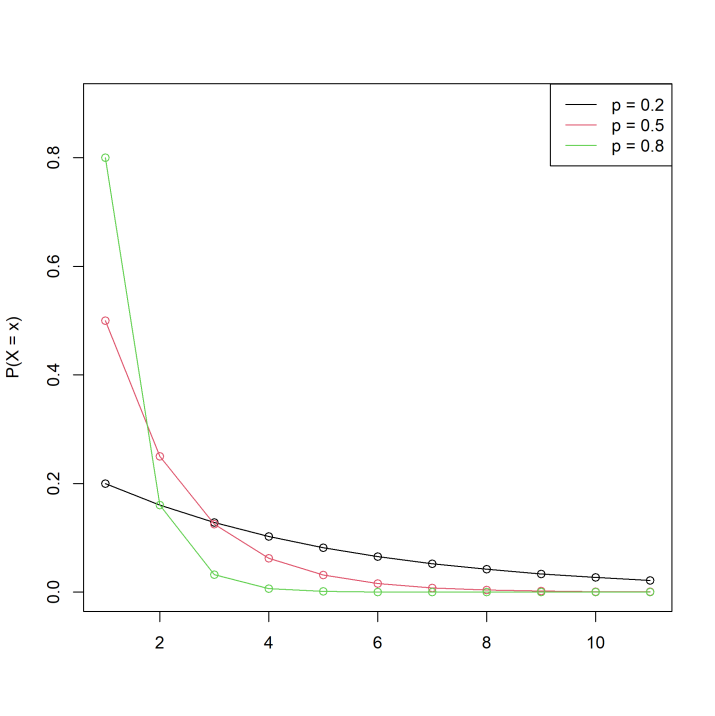
- $f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$
- $E(X) = Var(X) = \lambda$

- `dpois(x, lambda)` (pmf) y `ppois(lambda)` (cdf)

Otras distribuciones discretas

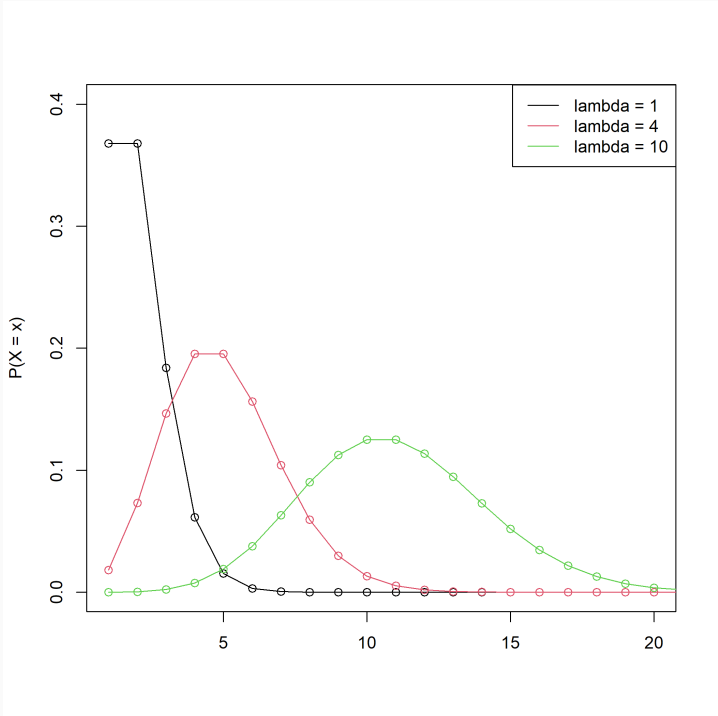
Distribución Geométrica: X es el número de *fracasos* antes de observar el primer *éxito* en procesos independientes.

- $X \sim Geom(p)$



Distribución Poisson: X es el número de ocurrencias de un evento dentro de un espacio o tiempo **finito**.

- $X \sim Poisson(\lambda)$



Variables aleatorias continuas

Variables aleatorias continuas

A diferencia de las variables discretas (que pueden tomar un número finito de valores), las variables continuas pueden tomar un **"número infinito de valores posibles"**.

El concepto de número infinito de valores posibles es una **abstracción matemática** ya que en la práctica todas las mediciones que hagamos consisten en un número finito (ingreso de una persona, longitud de un objeto, tiempo transcurrido).

La utilidad de considerar variables como continuas es que hace el **"manejo matemático" más fácil**.

En este caso no hablamos de distribución de función de masa de probabilidad (*pmf*) sino que de **función de densidad de probabilidad** (*pdf* en inglés).

La función de distribución acumulada (*cdf*) para una variable continua se define de forma similar al caso discreto ($\mathrm{P}(\mathrm{X} <= \mathrm{x})$).

Función de densidad de probabilidad

La función de densidad de probabilidad muestra la posibilidad relativa de la variable aleatoria continua en el espacio muestral. Esta debe cumplir que:

Para todo valor posible, la probabilidad tiene que ser mayor o igual a cero.

$$f(x) \geq 0 \; \forall \; x \in S_X$$

La suma de todos los casos posibles (área bajo la curva) debe ser igual a uno.

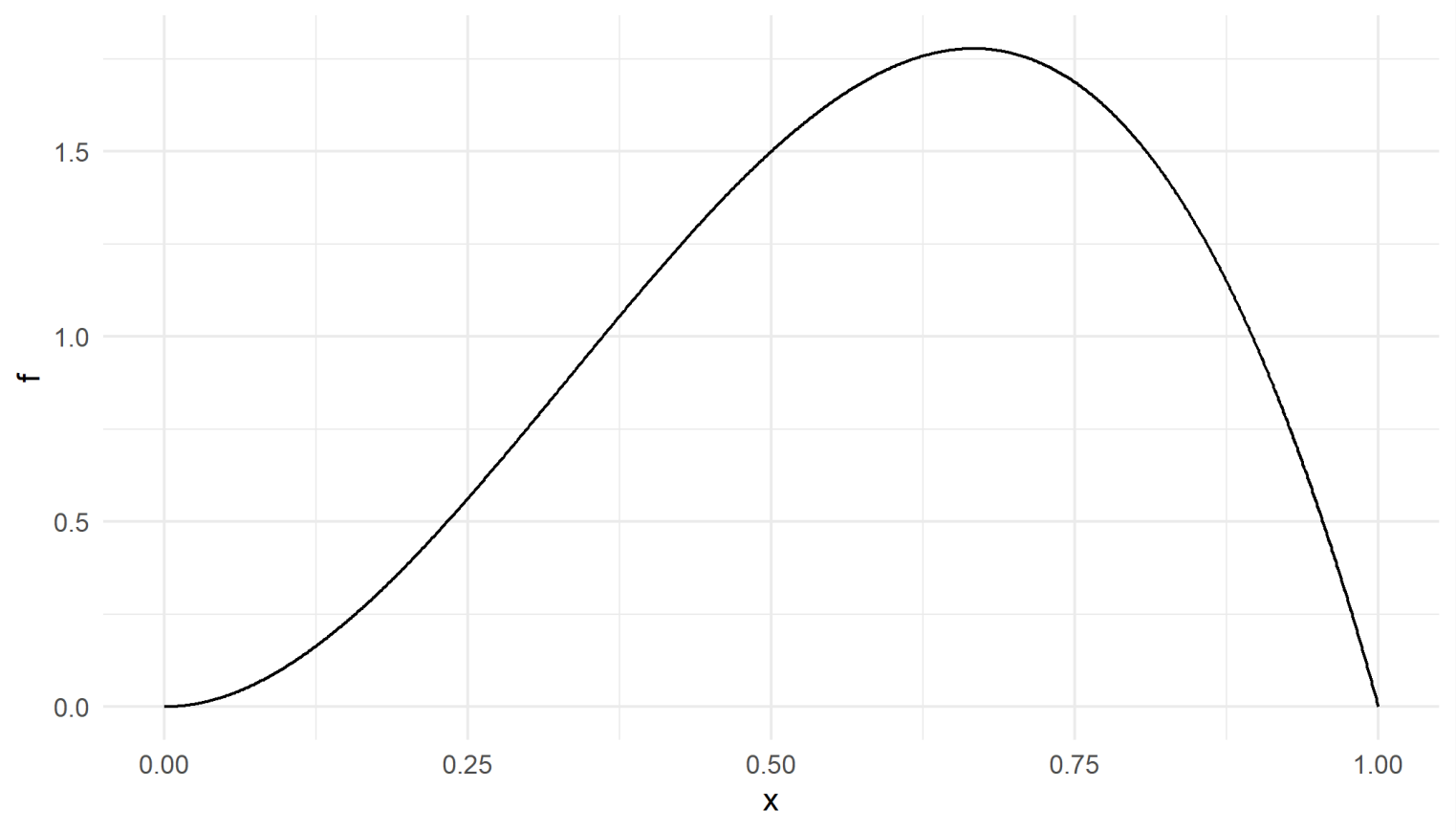
$$\int_{x \in S_X} f(x)dx = 1$$

Todo evento corresponde a una porción del área bajo la curva.

$$P(X \in B) = \int_{X \in B} f(x)$$

Ejemplo pdf

$$f(x) = 12x^2(1-x) \text{ si } 0 \leq x \leq 1$$



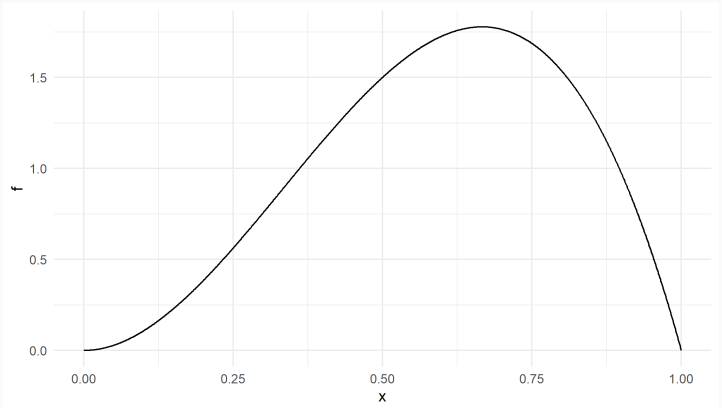
¿ f(x) mayor o igual a 0?

$f(x) = 12x^2(1-x)$

$x^2 \geq 0$

$(1-x) \geq 0$

$f(x) \geq 0 \; \forall \; x \in S_X$ **OK**



$f(x) = 12x^2(1-x)$ si $0 \leq x \leq 1$

Área bajo la curva = 1

$$\int_0^1 12(x^2)(1-x)dx$$

$$12\int_0^1 (x^2 - x^3)dx$$

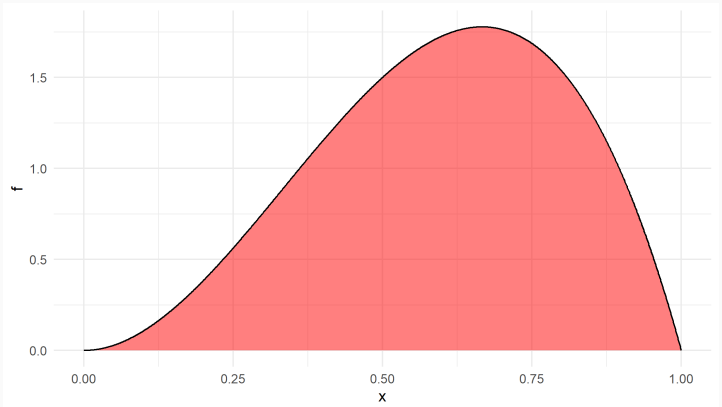
$$12[\frac{x^3}{3} - \frac{x^4}{4}\Big|_0^1]$$

$$12(\frac{1^3}{3} - \frac{1^4}{4}) - 12(\frac{0^3}{3} - \frac{0^4}{4}) = \frac{12}{3} - \frac{12}{4} = 4 - 3 = 1$$

```
f_ejemplo ← function(x){12*(x^2)*(1-x)}
integrate(f_ejemplo, 0, 1)
```

```
## 1 with absolute error < 1.1e-14
```

$\int_{x \in S_X} f(x)dx = 1$ **OK**



$f(x) = 12x^2(1-x)$ si $0 \leq x \leq 1$

Probabilidad = porción del área

$$P(0.25 < X < 0.5) = \int_{0.25}^{0.5} 12(x^2)(1-x)dx$$

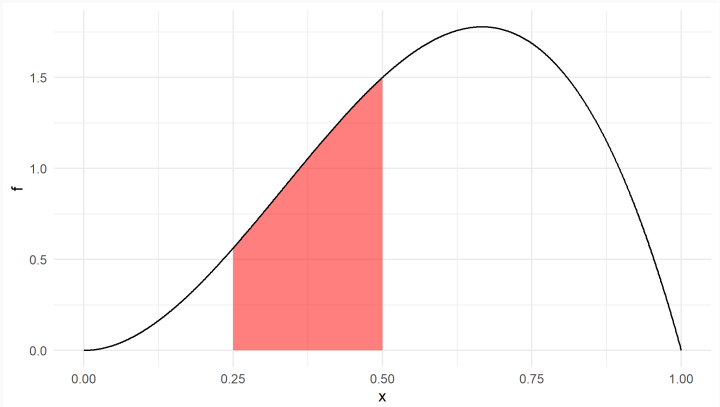
$$12\int_{0.25}^{0.5}(x^2 - x^3)dx$$

$$12[\frac{x^3}{3} - \frac{x^4}{4}\Big|_{0.25}^{0.5}] = 0.2617188$$

```
integrate(f_ejemplo, 0.25, 0.5)
```

```
## 0.2617188 with absolute error < 2.9e-15
```

$$P(X \in B) = \int_{X \in B} f(x)$$ **OK**



$$f(x) = 12x^2(1-x) \text{ si } 0 \leq x \leq 1$$

Probabilidad de un $x = 0$

$$P(X = x_i) = 0$$

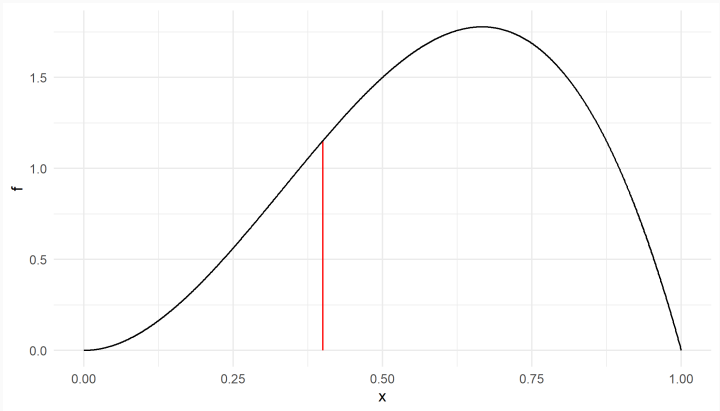
$$P(X = 0.4) = \int_{0.4}^{0.4} 12(x^2)(1-x)dx$$

$$12\int_{0.4}^{0.4}(x^2 - x^3)dx$$

$$12[\frac{x^3}{3} - \frac{x^4}{4}\Big|_{0.4}^{0.4}] = 0$$

```
integrate(f_ejemplo, 0.4, 0.4)
```

```
## 0 with absolute error < 0
```



$$f(x) = 12x^2(1-x) \text{ si } 0 \leq x \leq 1$$

CDF

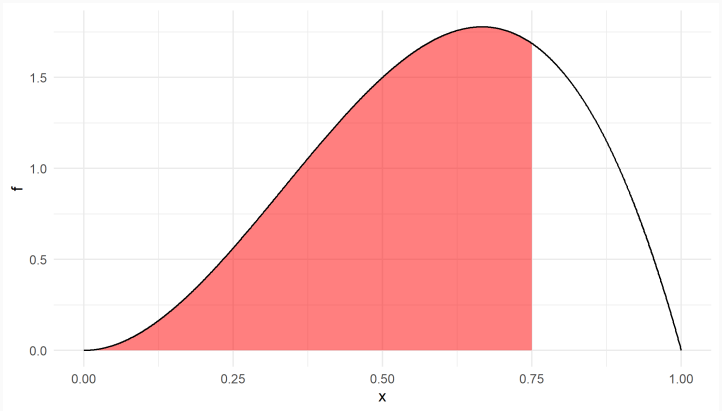
$$P(X \leq x) = F(x) = \int_1^x f(x)dx$$

$$F(0.75) = \int_0^{0.75} 12(x^2)(1-x)dx$$

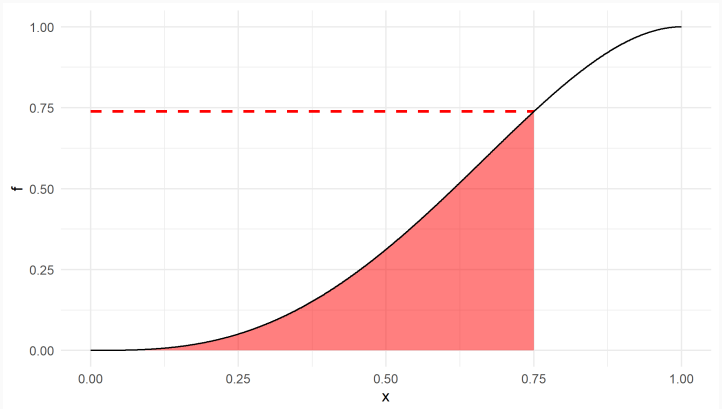
$$12\int_0^{0.75}(x^2 - x^3)dx$$

$$12[\frac{x^3}{3} - \frac{x^4}{4}\Big|_0^{0.75}] = 0.7382812$$

$$F'(x) = f(x)$$



$$f(x) = 12x^2(1-x) \text{ si } 0 \leq x \leq 1$$



$$F(x) = 4x^3 - 3x^4 \text{ si } 0 \leq x \leq 1$$

Valor esperado

$$E(X) = \int_{x \in S_X} (x * f(x))dx$$

$$\int_0^1 (x)(12)(x^2)(1-x)dx$$

$$12 \int_0^1 (x^4 - x^5)dx$$

$$12[\frac{x^4}{4} - \frac{x^5}{5}\Big|_0^1] = 0.6$$

```
f_ejemplo_e ← function(x){x*(12*(x^2)*(1-x))}
integrate(f_ejemplo_e, 0, 1)
```

```
## 0.6 with absolute error < 6.7e-15
```

Varianza

$$Var(X) = E(X^2) - E(X)^2$$

$$E(X^2) = \int_0^1 x^2(12)(x^2)(1-x)dx$$

$$E(X^2) = 12\int_0^1 (x^4 - x^5)dx$$

$$12[\frac{x^5}{5} - \frac{x^6}{6}\Big|_0^1] = 0.4$$

$$Var(X) = 0.4 - 0.6^2 = 0.04$$

```
f_ejemplo_var ← function(x){x^2*(12*(x^2)*(1-x))}
integrate(f_ejemplo_var, 0, 1)[[1]] - (integrate(f_
```

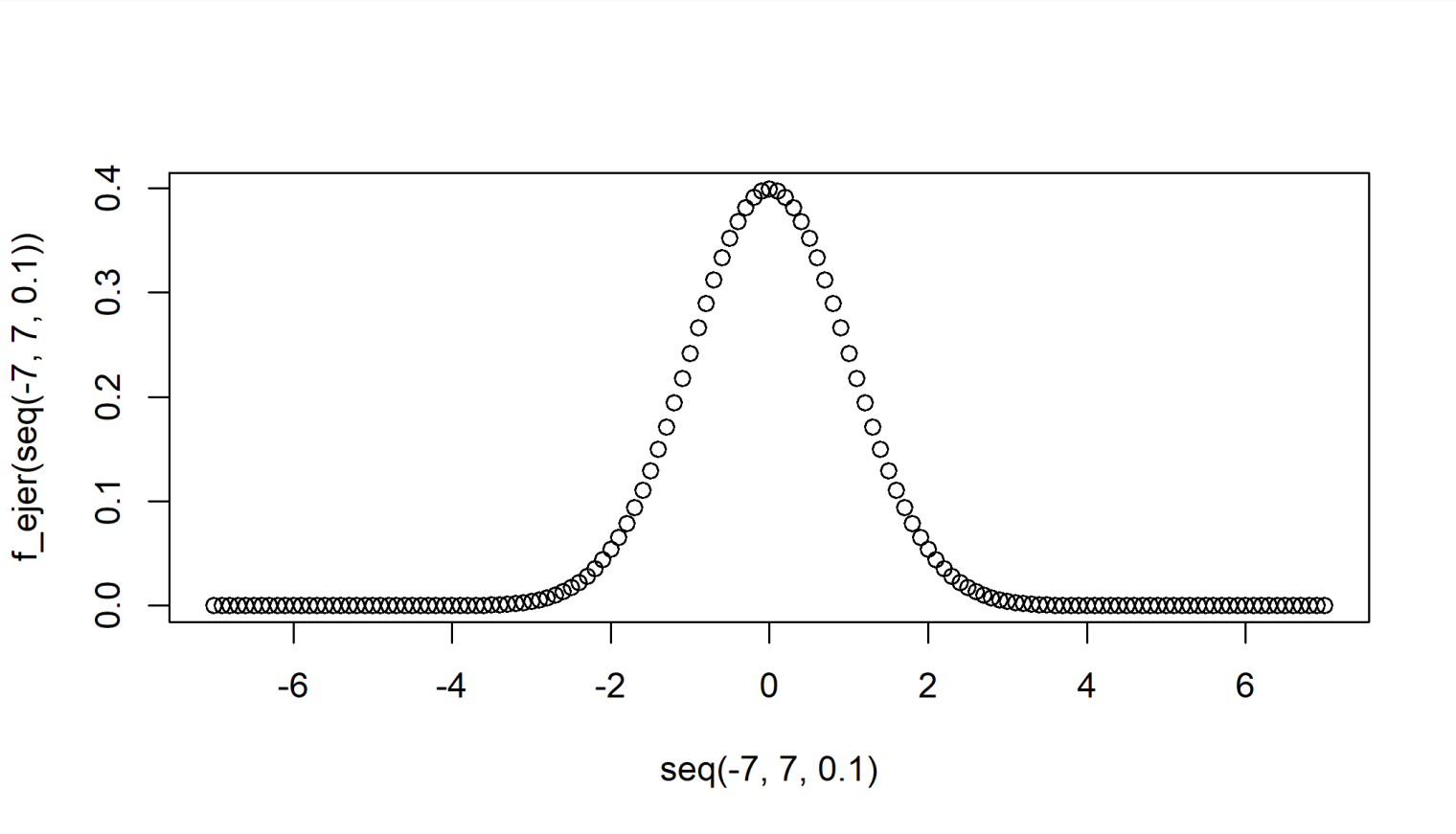
```
## [1] 0.04
```

Ejercicio

Ejercicio

`EjercicioProbabilidadesCont.R`

Respuestas



Distribución normal (o gaussiana)

Distribución normal

Esta es una distribución continua con características muy particulares (y de mucha utilidad).

$$X \sim N(\mu, \sigma^2)$$

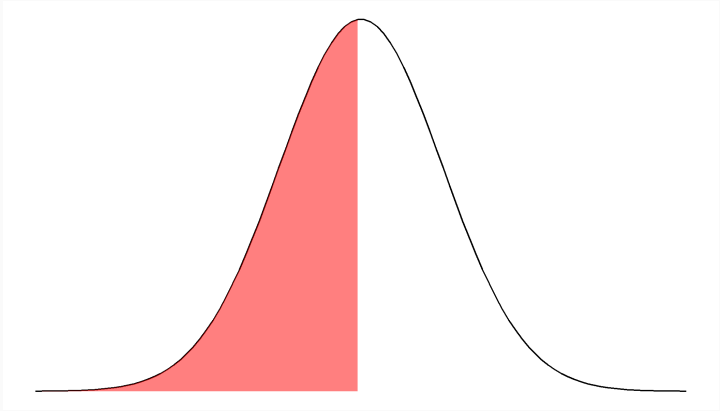
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Promedio = Mediana = Moda

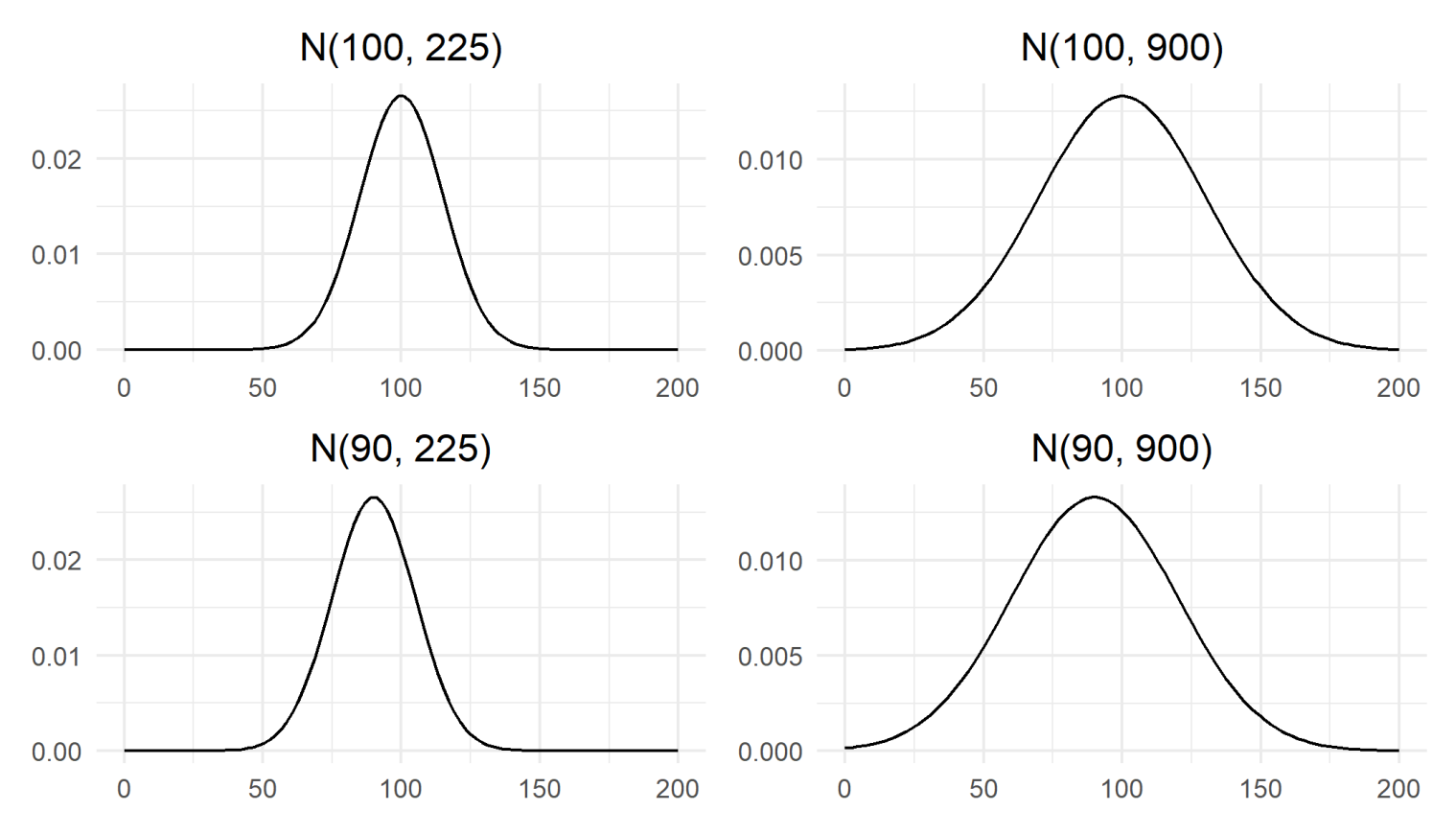
$$P(X < \mu) = 0.5$$



Parámetros de la distribución

$$X \sim N(\mu, \sigma^2)$$

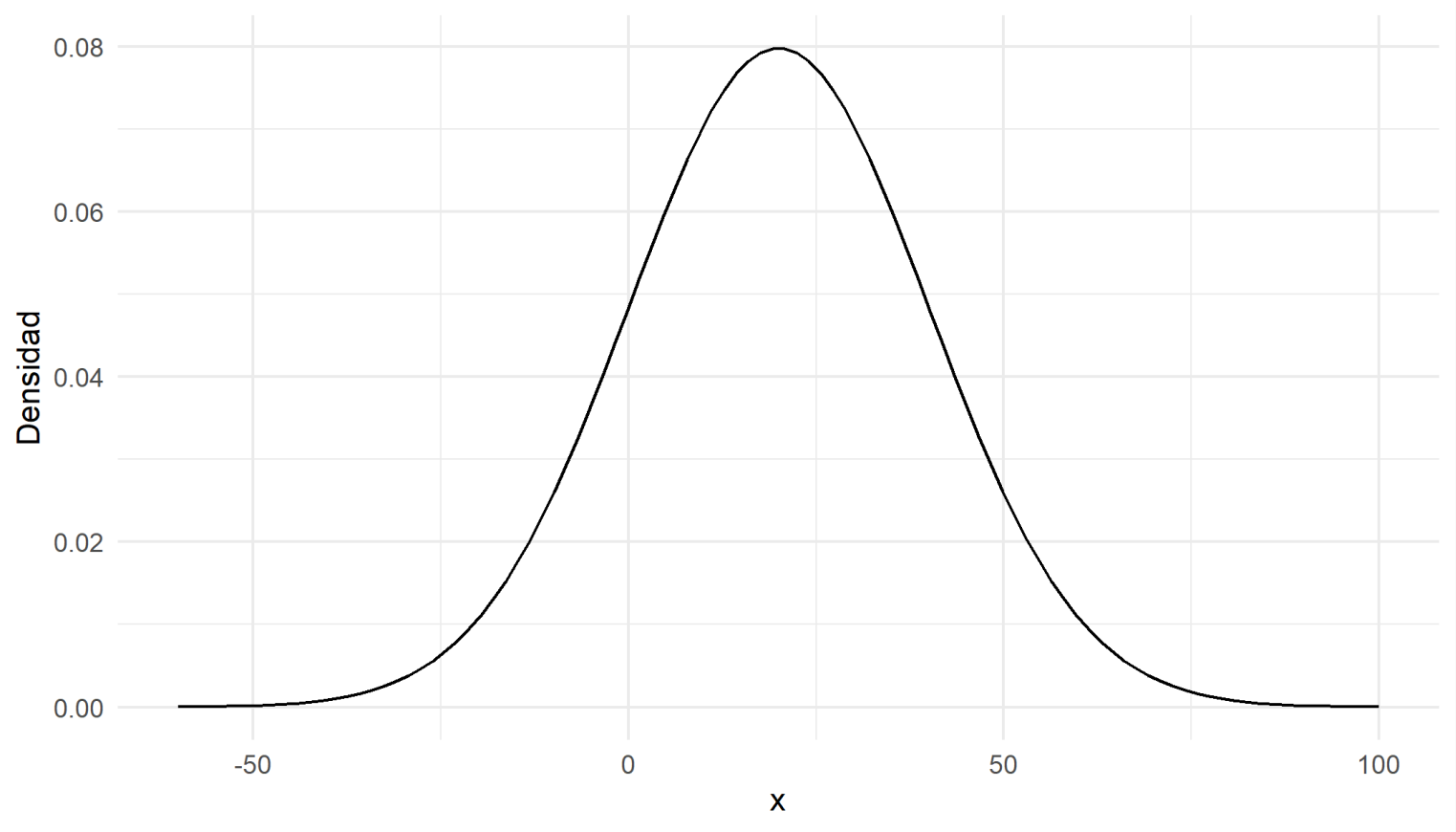
- μ: centro de la distribución
- σ^2: dispersión de la distribución



Ejemplo práctico

Digamos que las notas de una prueba con un máximo de 100 puntos se distribuye de forma normal con media 80 y desviación estándar de 5.

$$X \sim N(80, 5^2)$$

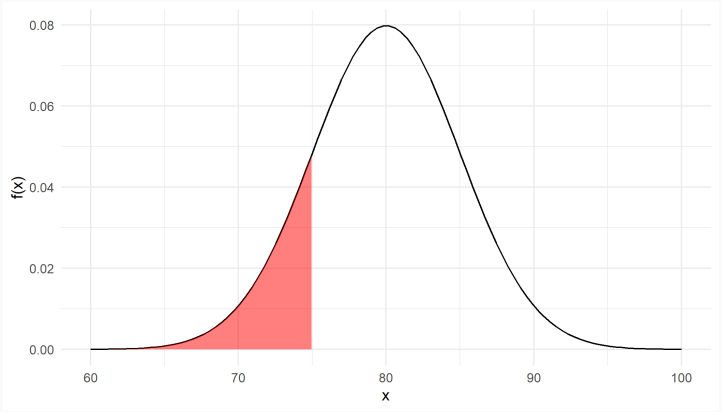


Calcular probabilidades

¿Cuál es la probabilidad de que un estudiante tenga menos de 75 puntos?

```
pnorm(75, mean = 80, sd = 5)
```

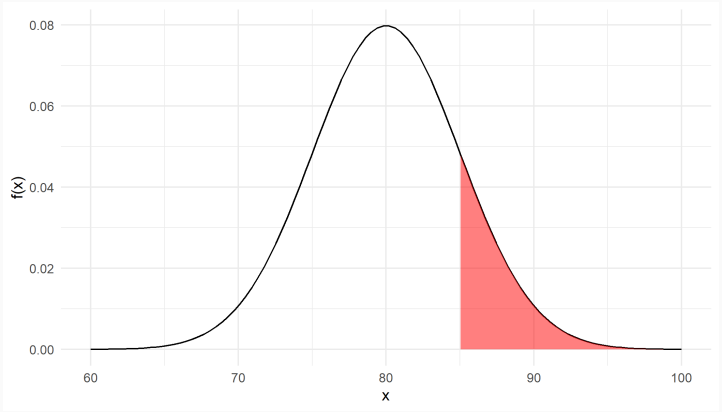
```
## [1] 0.1586553
```



Si un estudiante saca 75 puntos, el percentil al que pertenece es el 15.86%

Calcular probabilidades

¿Cuál es la probabilidad de que un estudiante tenga más de 85 puntos?



Calcular probabilidades

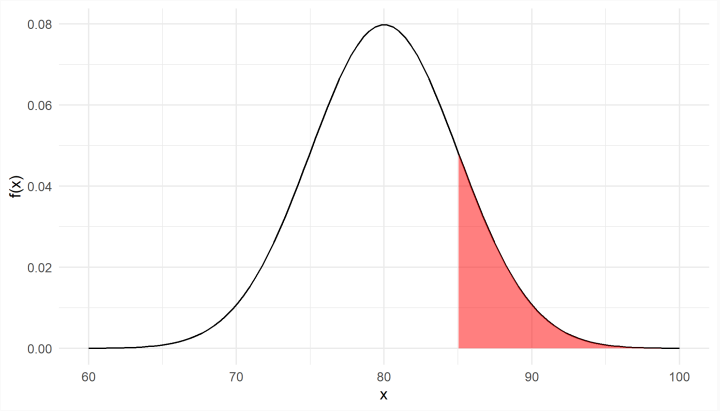
¿Cuál es la probabilidad de que un estudiante tenga más de 85 puntos?

```
1 - pnorm(85, mean = 80, sd = 5)
```

```
## [1] 0.1586553
```

```
pnorm(85, mean = 80, sd = 5, lower.tail = FALSE)
```

```
## [1] 0.1586553
```



N° de σ de la media

Un/a estudiantetuvo 85 puntos en la prueba. ¿A cuántas desviaciones estándar está desde la media?

```
(85-80)/5
```

```
## [1] 1
```

Un/a estudiante tuvo 77.5 puntos en la prueba. ¿A cuántas desviaciones está desde la media?

```
(77.5-80)/5
```

```
## [1] -0.5
```

Z-score

$$Z = \frac{X - \mu}{\sigma}$$

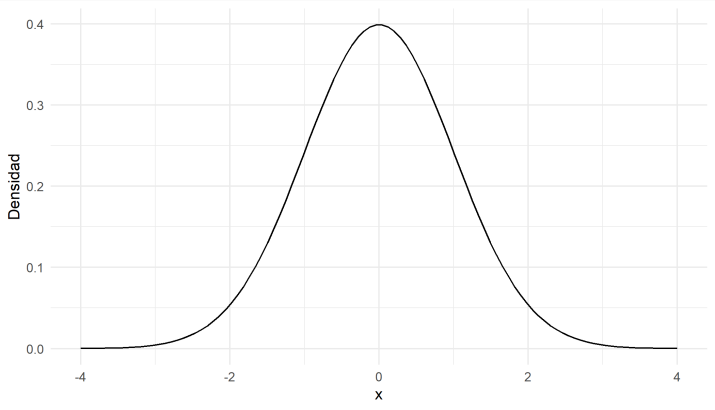
Podemos calcular *Z-scores* sin importar la distribución de los datos (no tienen que distribuir normalmente). Este solo muestra a cuantas desviaciones estándar se encuentra un valor de la media.

Una propiedad interesante de las distribuciones normales es que cualquier combinación lineal de variables aleatorias normales será también normal. Sabiendo esto, resulta que el *Z-Score* nos permite convertir cualquier distribución normal en una **distribución normal estándar** y así aprovechar las propiedades de esta.

Distribución normal estándar

$$X \sim N(0, 1)$$

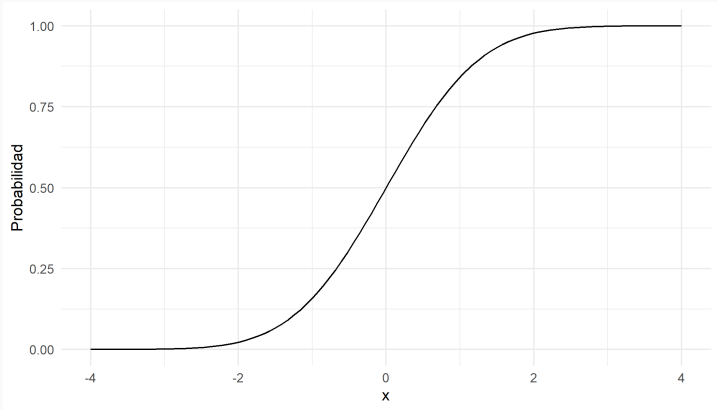
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} = \phi(x)$$




```
dnorm(0, mean = 0, sd = 1)
```

```
## [1] 0.3989423
```

$$F(x) = P(X < x) = \int_{-\infty}^{x} \phi(x)dx = \Phi(x)$$

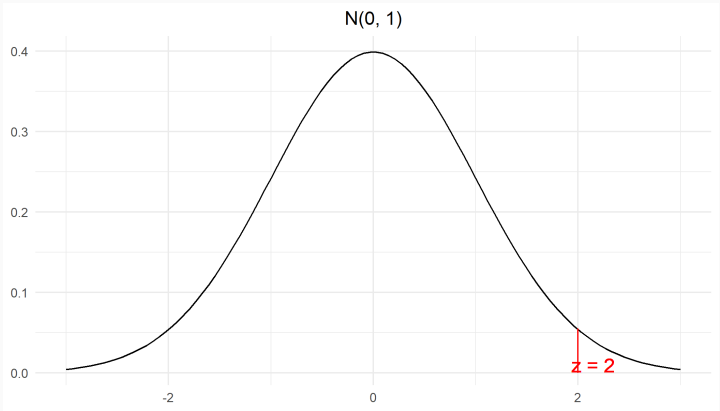



```
pnorm(0, mean = 0, sd = 1)
```

```
## [1] 0.5
```

Quizás recuerdan esto

Z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990




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

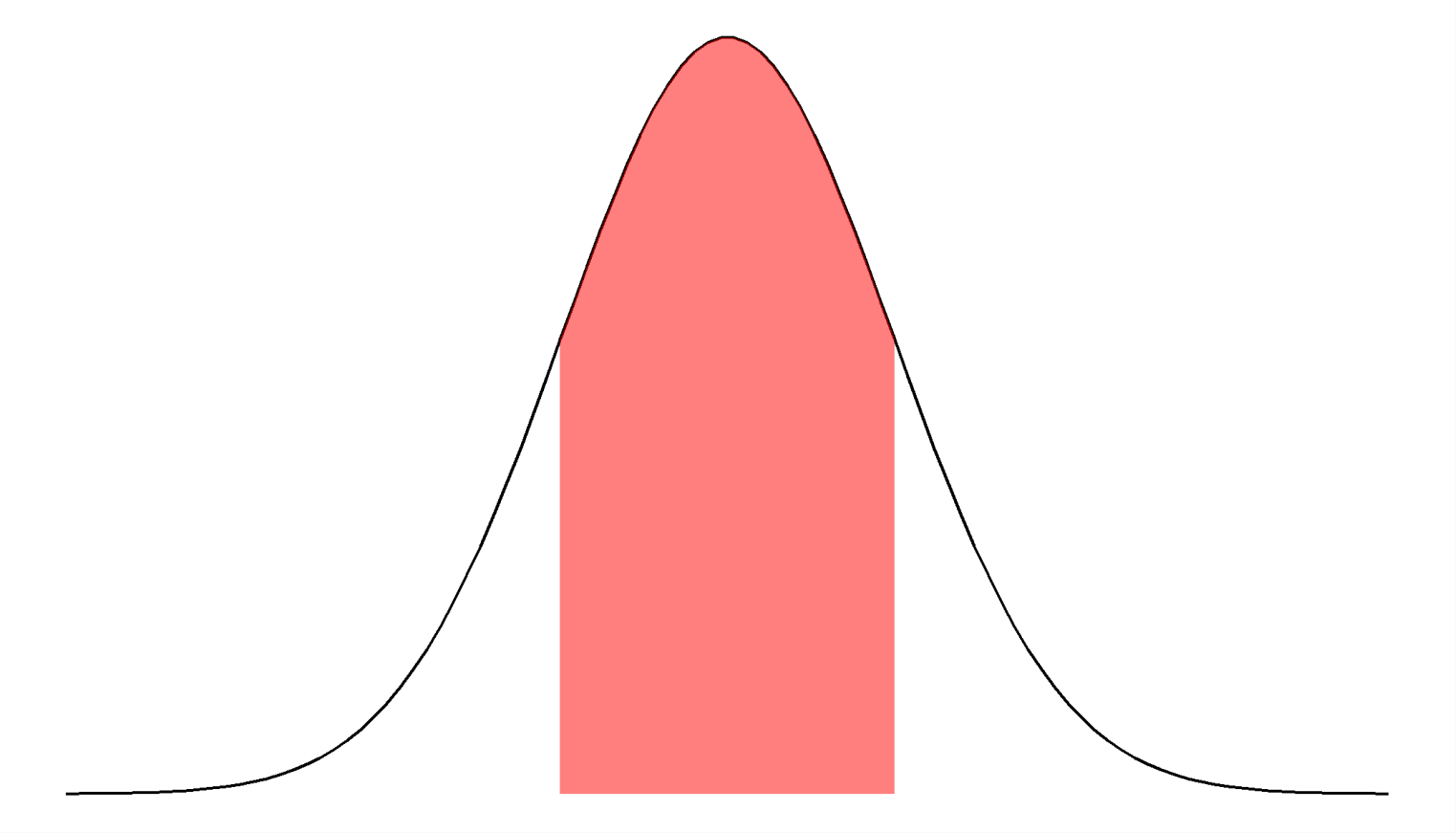
```
## [1] 0.9772499
```

Distribución normal estándar

Un **68.27%** de las observaciones están a **una** desviación estándar de la media.

```
pnorm(1, mean = 0, sd = 1) - pnorm(-1, mean = 0, sd = 1)
```

```
## [1] 0.6826895
```

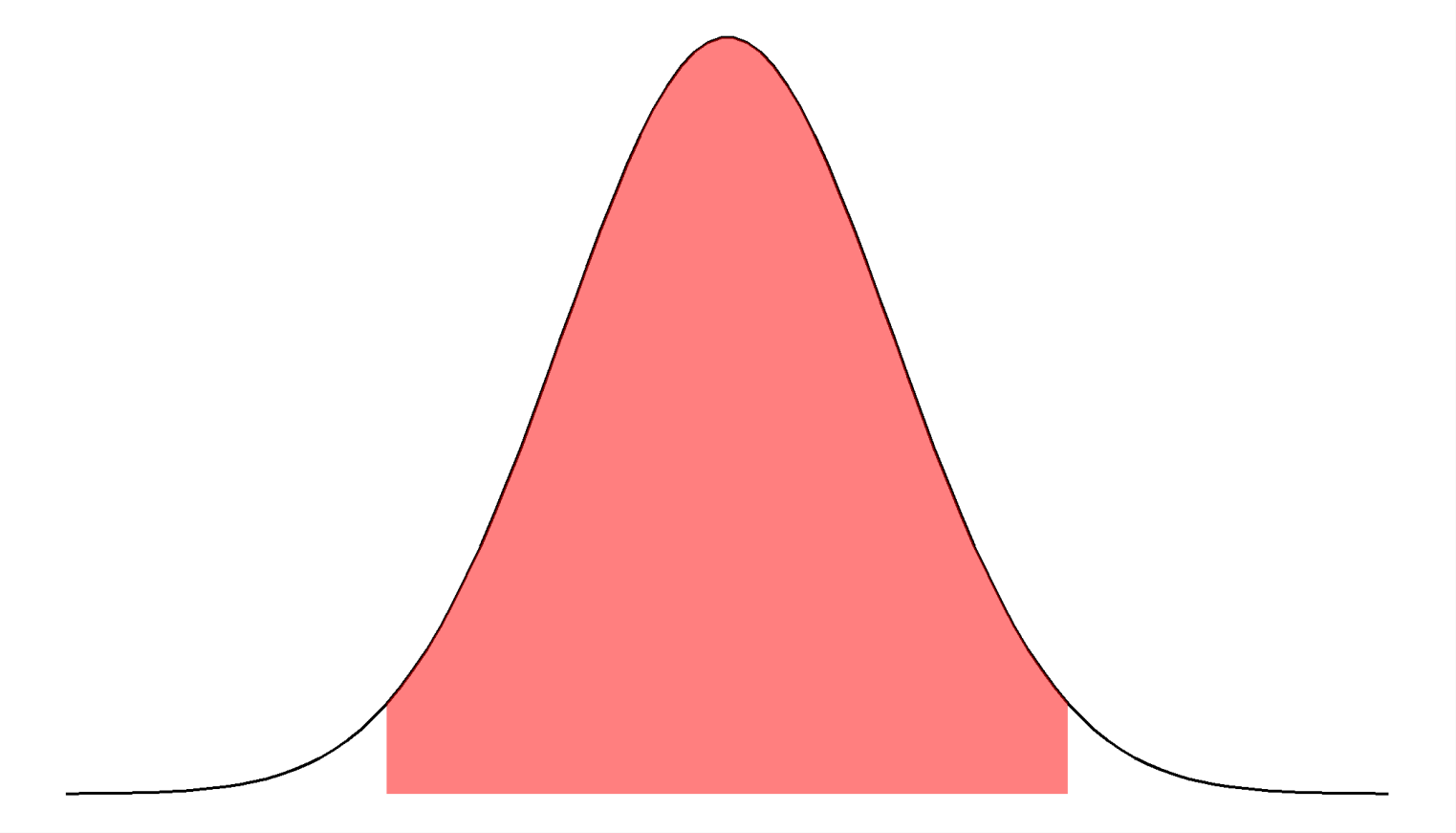


Distribución normal estándar

Un **95.45%** de las observaciones están a **dos** desviaciones estándar de la media.

```
pnorm(2, mean = 0, sd = 1) - pnorm(-2, mean = 0, sd = 1)
```

```
## [1] 0.9544997
```

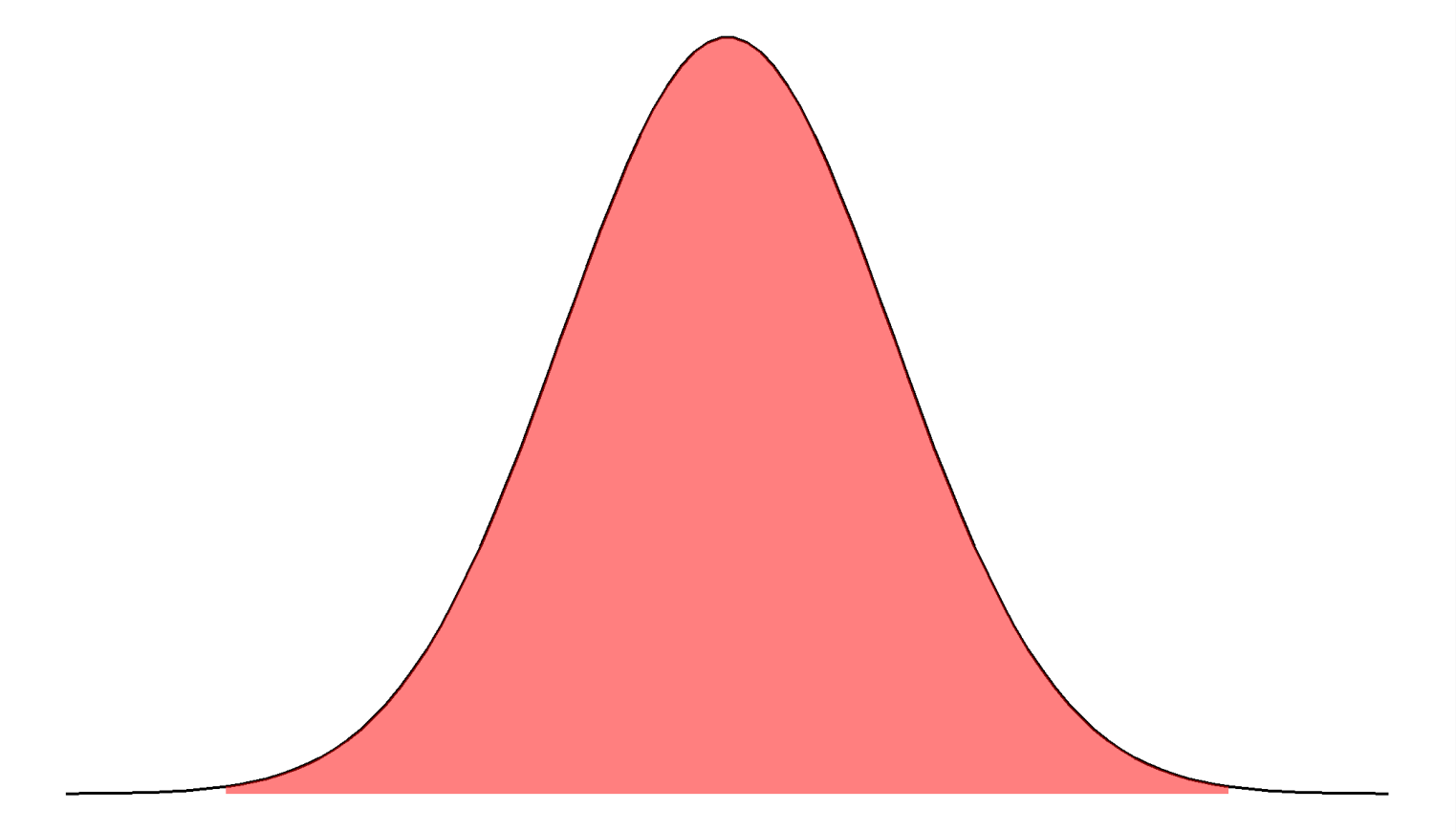


Distribución normal estándar

Un **99.73%** de las observaciones están a **tres** desviaciones estándar de la media.

```
pnorm(3, mean = 0, sd = 1) - pnorm(-3, mean = 0, sd = 1)
```

```
## [1] 0.9973002
```



Ejercicio

Ejercicio

`EjercicioDistNormal.R`

Respuestas

```r
# ¿Cuál es la probabilidad de que alguien mida menos de 1.6 metros?
pnorm(160, mean = 177, sd = 10)
## [1] 0.04456546

# ¿Cuál es la probabilidad de que alguien mida más de 1.8 metros?
1 - pnorm(180, 177, 10)
## [1] 0.3820886
pnorm(180, 177, 10, lower.tail = FALSE)
## [1] 0.3820886

# ¿Cuál es la probabilidad de que alguien mida entre 1.6 y 1.8 metros?
pnorm(180, 177, 10) - pnorm(160, 177, 10)
## [1] 0.573346
```

Otras distribuciones continuas

Distribución Uniforme Continua

- $X \sim U(a, b)$
- $f(x) = \frac{1}{b-a}$ si $a \leq x \leq b$
- $E(X) = \frac{b+a}{2}$
- $Var(X) = \frac{(b-a)^2}{12}$
- `dunif(x, min, max)` (pmf) y `punif(q, min, max)` (cdf)

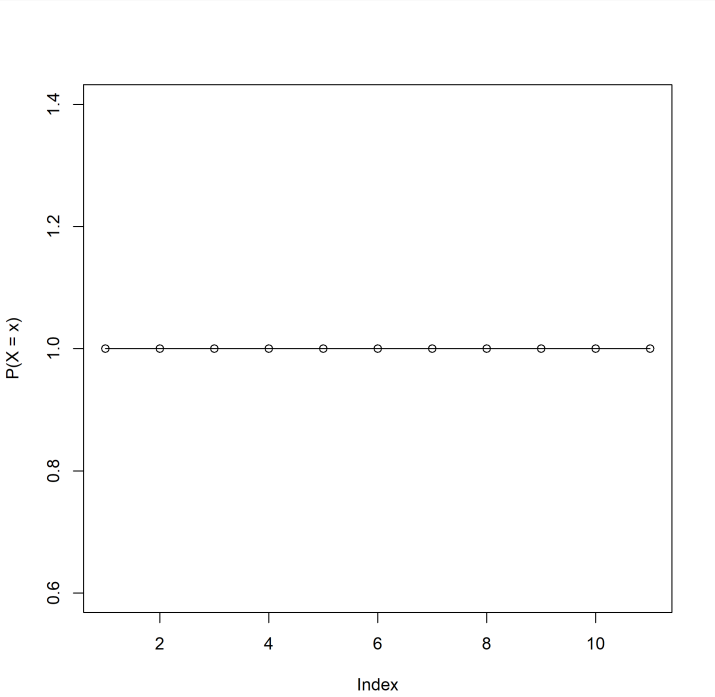
Distribución Exponencial

- $X \sim Exp(\lambda)$
- $f(x) = \lambda e^{-\lambda x}$ si $x \in [0, inf]$ y $\lambda > 0$
- $E(X) = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$
- `dexp(x, rate)` (pmf) y `pexp(x, rate)` (cdf)

Otras distribuciones continuas

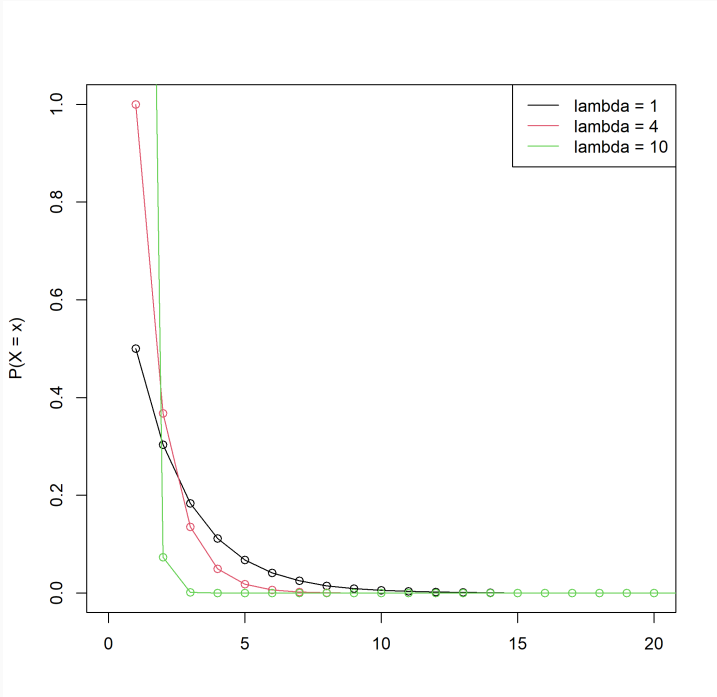
Distribución Uniforme Continua

- $X \sim U(a, b)$



Distribución Exponencial

- $X \sim Exp(\lambda)$



Teorema del Límite Central

Teorema del límite central

Si ciertas condiciones se cumplen, la distribución muestral se distribuirá de forma normal con media igual al parámetro poblacional. La desviación estándar será inversamente proporcional a la raiz cuadrada del tamaño muestral

Población vs Muestra

- **Población**: grupo bien definido de sujetos (por ejemplo, población de un país)
- **Muestra**: Subconjunto de individuos provenientes de una población que se obtienen a través de algún procedimiento de muestreo (ej. muestreo aleatorio simple)

Buscamos aprender algo de la **población** a través de una **muestra** (o varias). A esto lo denominamos **inferencia** (hablaremos más de esto en el módulo 2)

Censo

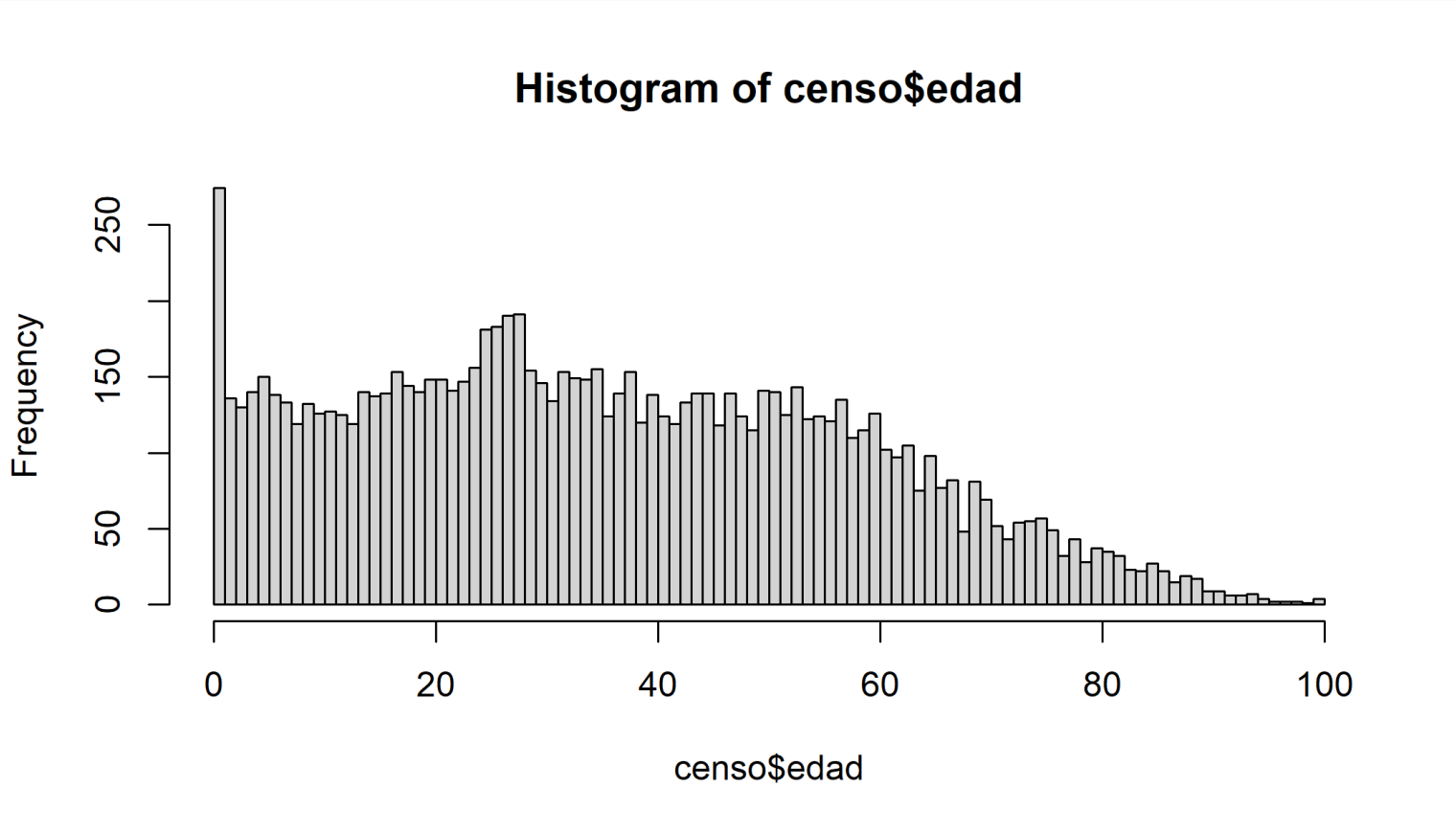
Asumamos que estos datos ($\mathrm{N} = 10.000$) corresponden al total del país.

```
censo ← read.csv("../datos/muestra_censo_2017.csv")
str(censo)
```

```
## 'data.frame':    10000 obs. of  4 variables:
##  $ ï..x        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ edad        : int  35 37 59 29 76 65 25 16 49 35 ...
##  $ sexo        : chr  "M" "H" "H" "M" ...
##  $ p_originario: chr  "no" "no" "no" "no" ...
```

Analicemos la variable `edad`.

Distribución de edad



Claramente no sigue una distribución normal

Sabemos μ y σ

Debido a que tenemos la información de todas las personas sabemos cual es el promedio y la desviación estándar de la población.

```
mean(censo$edad)
```

```
## [1] 36.0179
```

```
sd(censo$edad)
```

```
## [1] 22.14836
```

Pero practicamente nunca sabremos estos parámetros poblacionales, y lo que tenemos que hacer es sacar conclusiones desde **muestras** (hacer estimaciones).

Si sacamos una muestra desde una población y calculamos un parámetro, por ejemplo la media, llamamos a esto una **estimación puntual**.

Muestras

Una muestra de 100 observaciones

```
mean(sample(censo$edad, 100))
```

```
## [1] 40
```

Una segunda muestra de 100 observaciones

```
mean(sample(censo$edad, 100))
```

```
## [1] 37.89
```

Una tercera muestra de 100 observaciones

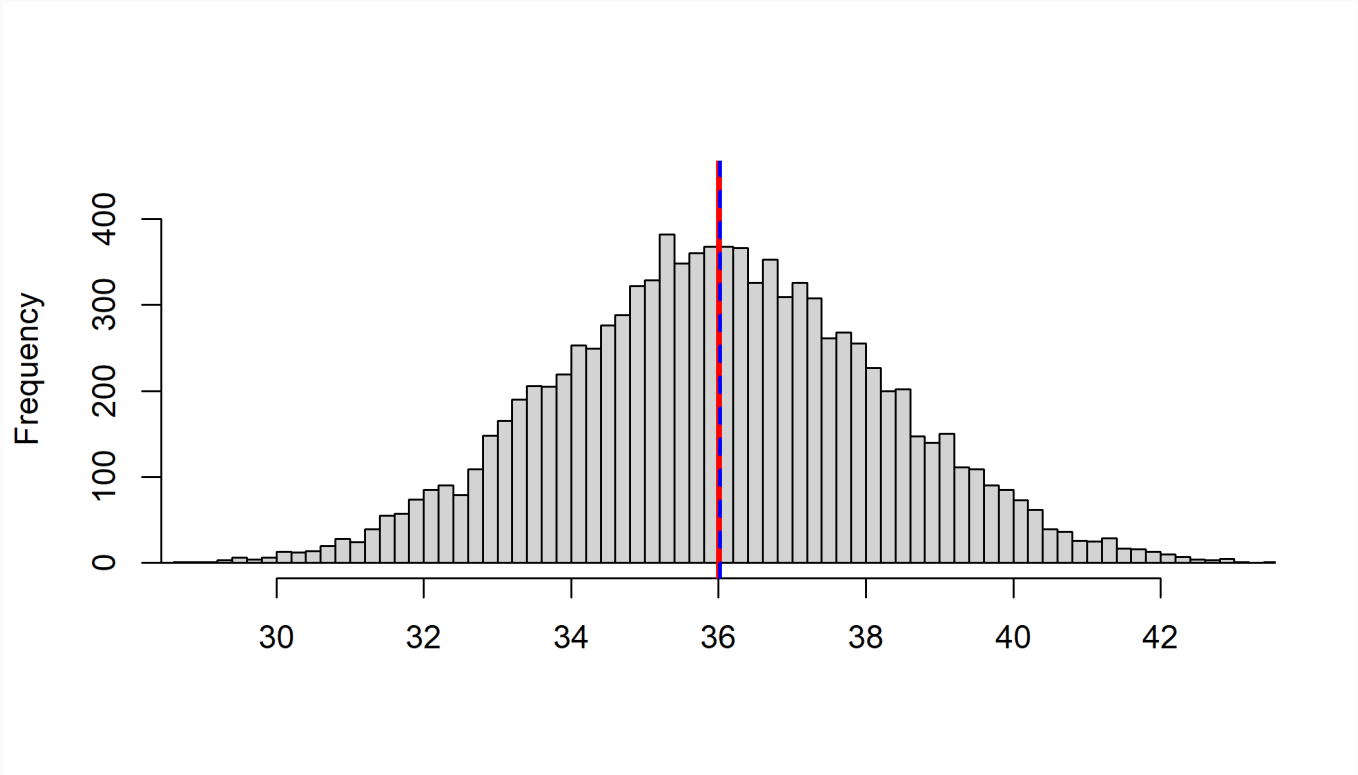
```
mean(sample(censo$edad, 100))
```

```
## [1] 40.05
```

¿Qué pasa si repetimos esto muchas veces?

10.000 muestras con $_{n}$=100

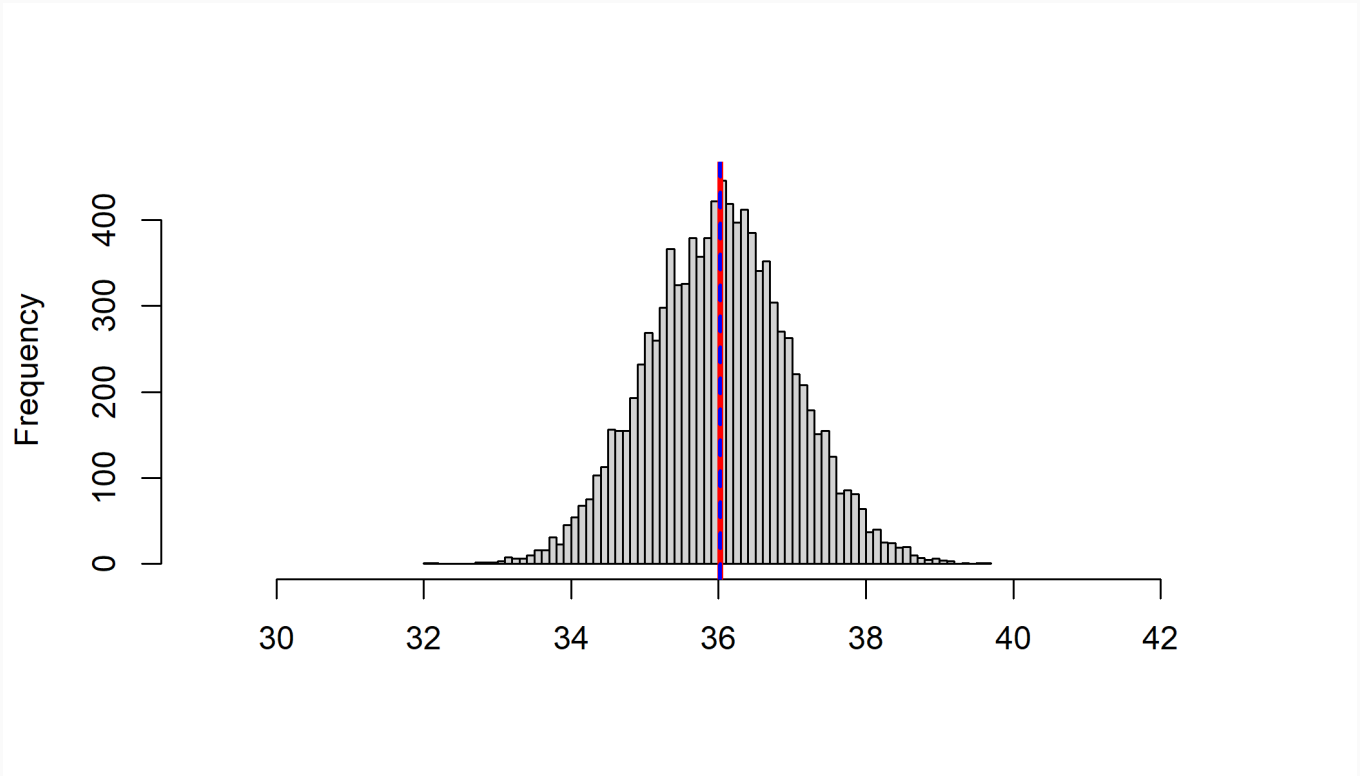
Sacamos 10.000 muestras de tamaño 100 y graficamos la distribución de cada uno de los promedios calculados:



¡La distribución de las estimaciones de cada muestra aproximan una distribución normal con media igual a la media poblacional!

10.000 muestras con $_{n}$=500

Sacamos 10.000 muestras de tamaño 500 y graficamos la distribución de cada uno de los promedios calculados:



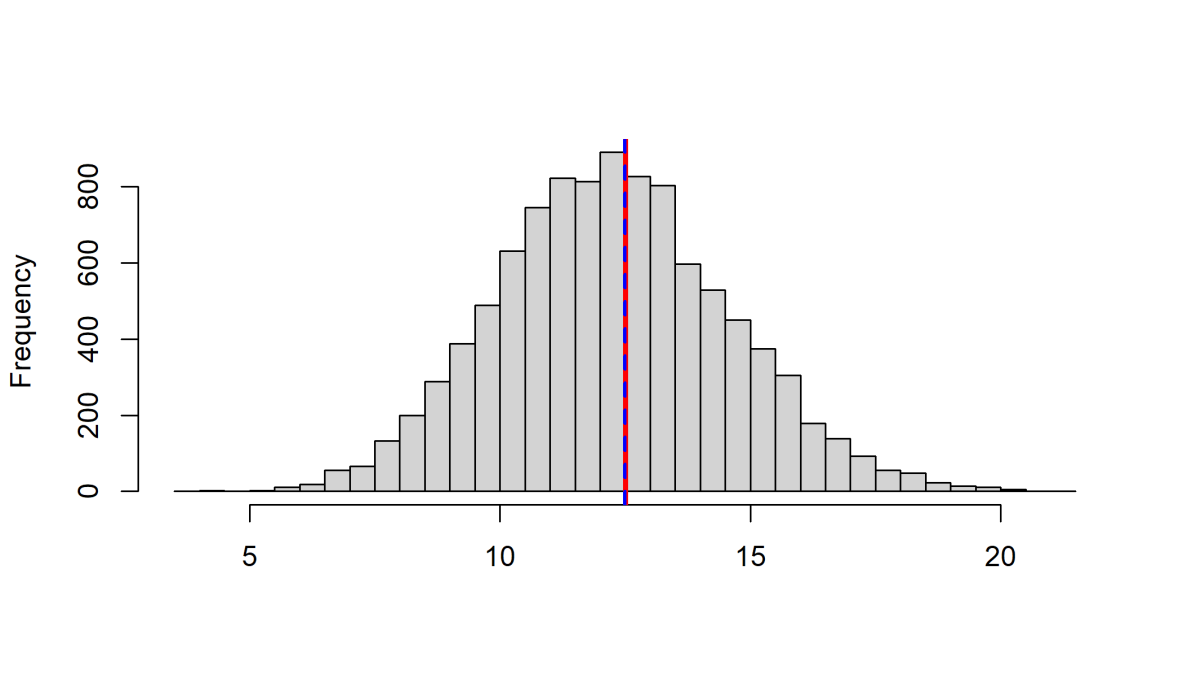
¡La distribución de las estimaciones de cada muestra aproximan una distribución normal con media igual a la media poblacional! **¡Y a mayor tamaño muestral menor la dispersión!**

Aplica también a proporciones

Proporción de personas que se identifican como parte de algún pueblo originario:

```
round(table(censo$p_originario)/10000, 3)*100
```

```
##
##   no   si
## 87.5 12.5
```



Teorema del límite central

Si ciertas condiciones se cumplen, la distribución muestral se distribuirá de forma normal con media igual al parámetro poblacional. La desviación estándar será inversamente proporcional a la raiz cuadrada del tamaño muestral

$$\bar{x} \sim \text{aproximadamente } N(\mu, \frac{\sigma^2}{n})$$

$$p \sim \text{aproximadamente } N(\pi, \frac{\pi(1-\pi)}{n})$$

Muy relevante a la hora de hacer inferencia a partir de muestras

Distribuciones conjuntas

Distribuciones conjuntas

Si X e Y son dos variables aleatorias, la **distribución conjunta** de X e Y permite calcular las probabilidades de eventos que involucren a ambas variables.

Por ejemplo, la probabilidad de que alguien mida entre 1.7 y 1.8 metros y que pese entre 60 y 80 kilogramos.

Desde una distribución conjunta podemos obtener **distribuciones marginales** (distribución de cada variable) y **distribuciones condicionales**.

Esperanzas condicionales

Si X e Y no son independientes, entonces saber algo de X me puede ayudar a predecir/explicar Y.

$\mathrm{E}(\mathrm{Y}|\mathrm{X})$ es una **función** que me dice para cada valor de X, la esperanza de Y de aquellos individuos con ese valor de X.

Relaciones entre variables

Covarianza

$$\mathrm{Cov}(X, Y) = \frac{\sum_1^n (X - \bar{X})(Y - \bar{Y})}{n}$$

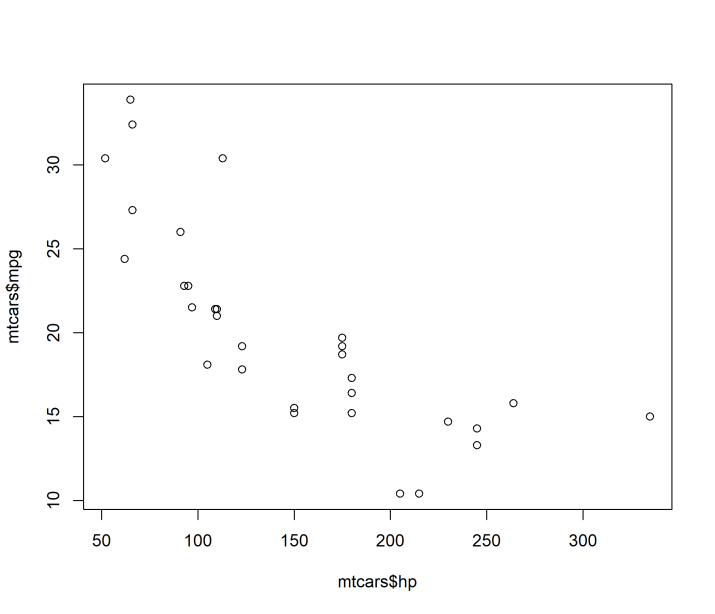
$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Correlación

Mismo signo que $\mathrm{Cov}(X, Y)$ pero adimensional (entre -1 y 1)

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{Var(X) Var(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Relaciones entre variables




```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Relaciones entre variables

Covarianza caballos de fuerza (`hp`)/millas por galón (`mpg`)

```
cov(mtcars$hp, mtcars$mpg)
```

```
## [1] -320.7321
```

Correlación caballos de fuerza (`hp`)/millas por galón (`mpg`)

```
cor(mtcars$hp, mtcars$mpg)
```

```
## [1] -0.7761684
```

```
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Distribuciones derivadas de la Normal

Distribución χ^2

La distribución χ^2 corresponde a la suma de M distribuciones normales estándar al cuadrado.

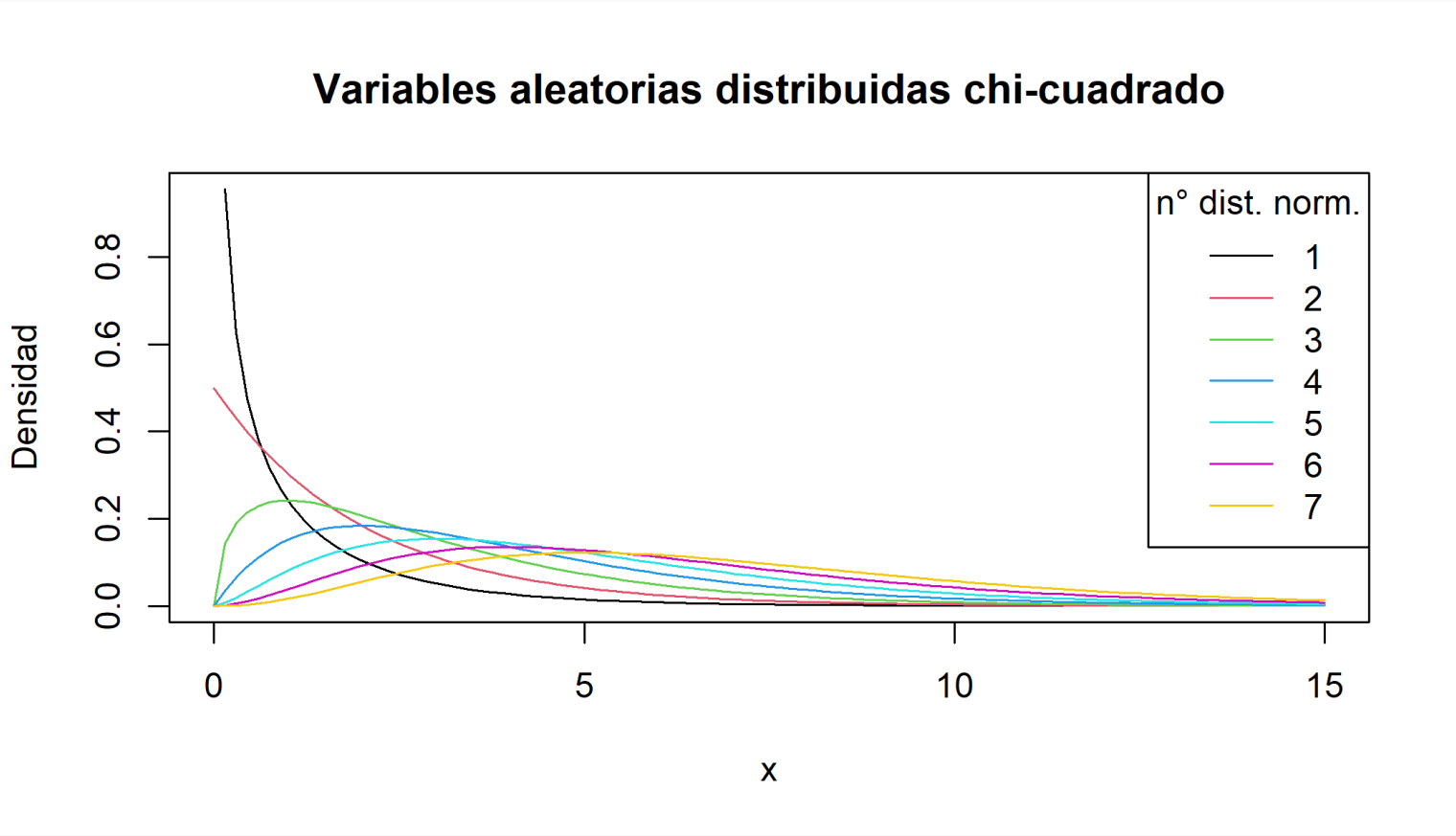
$$Z_1^2 + \ldots + Z_M^2 \sim \chi_M^2 \text{ con } Z_m \sim N(0, 1)$$

Para una variable X que siga una distribución χ_M^2:

- $E(X) = M$
- $Moda = M$
- $Var(X) = 2 \times M$

`dchisq(x, df)` (pmf) y `pchisq(q, df)` (cdf)

Distribución χ^2



Distribución t

Corresponde al ratio entre una normal y la raíz cuadrada de una chi-cuadrado.

Si Z es una distribución normal estándar, W es una variable aleatoria chi-cuadrado, χ^2_M, y ambas son independientes. Entonces:

$$\frac{Z}{\sqrt{Q/M}} = X \sim t_M$$

X es una variable aleatoria que sigue una distribución *t-student* (o solo t).

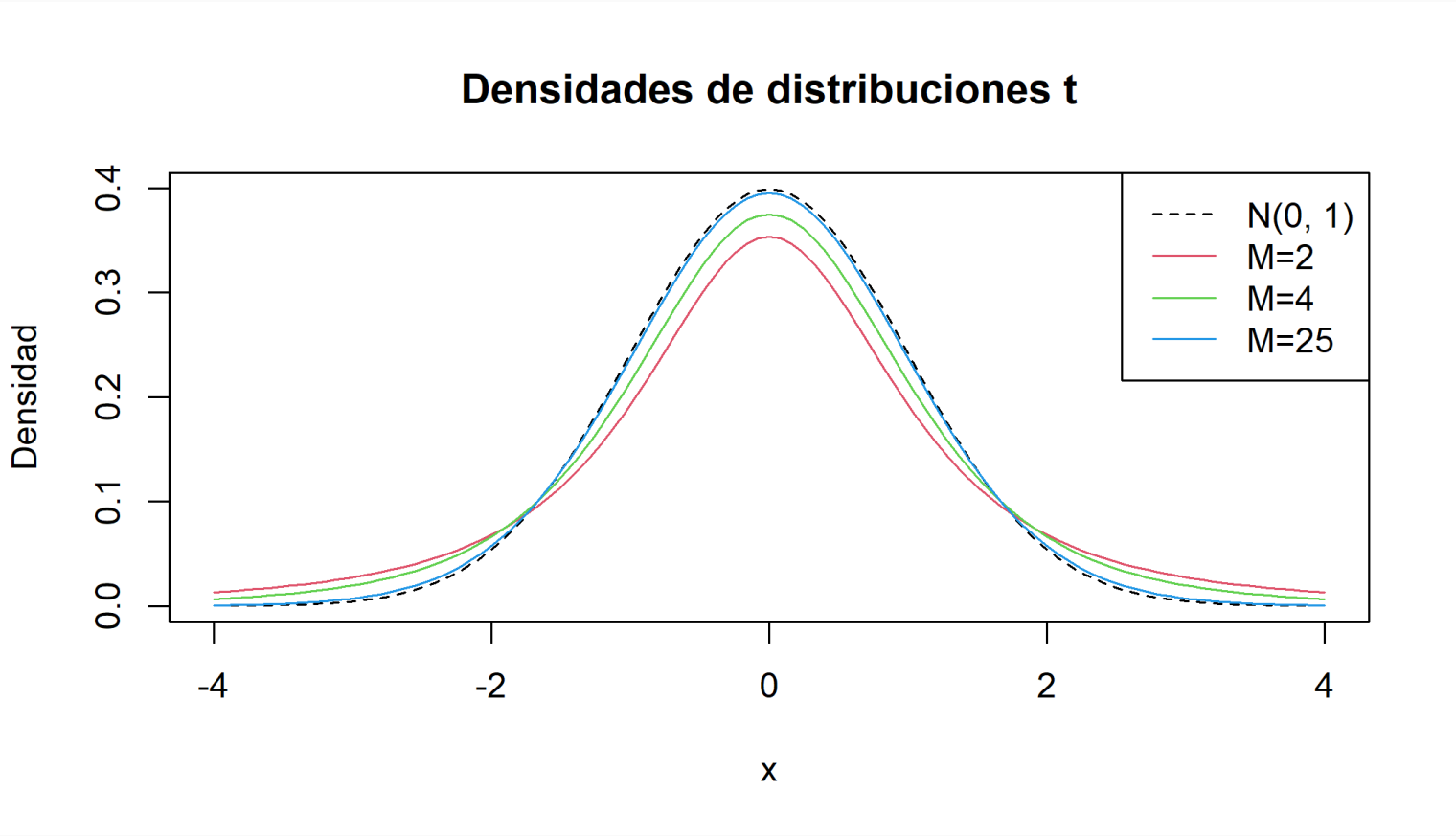
t_M es simétrica y con forma de campana similar a una distribución normal. De hecho, tiende a $N(0,1)$ cuando el número de observaciones es grande (se suele decir que mayor a 30).

Para una variable X que siga una distribución t_M:

- $E(X) = 0$
- $Var(X) = \frac{M}{M-2}$

`dt(x, df)` (pmf) y `pt(q, df)` (cdf)

Distribución t



Distribución F

Ratio de dos distribuciones chi-cuadrado.

Si W es una variable distribuida χ^2_{M} y V es una variable distribuida χ^2_{N} (con ambas variables independientes entre sí), entonces:

$$\frac{\mathrm{W}/\mathrm{M}}{\mathrm{V}/\mathrm{N}} = \mathrm{X} \sim \mathrm{F}_{\mathrm{M,N}}$$

X es una variable aleatoria que sigue una distribución $\mathrm{F}_{\mathrm{M,N}}$

`df(x, df1, df2)` (pmf) y `pf(q, df1, df2)` (cdf)

Lo que se viene

Módulo 2

- 3 clases intensivo de `R`:
  - Visualización
  - Manipulación
  - Transformación
- 3 clases de inferencia estadística (y `R`):
  - Pruebas de hipótesis
  - Regresiones lineales
  - Modelo logit/probit
  - Inferencia vs Predicción