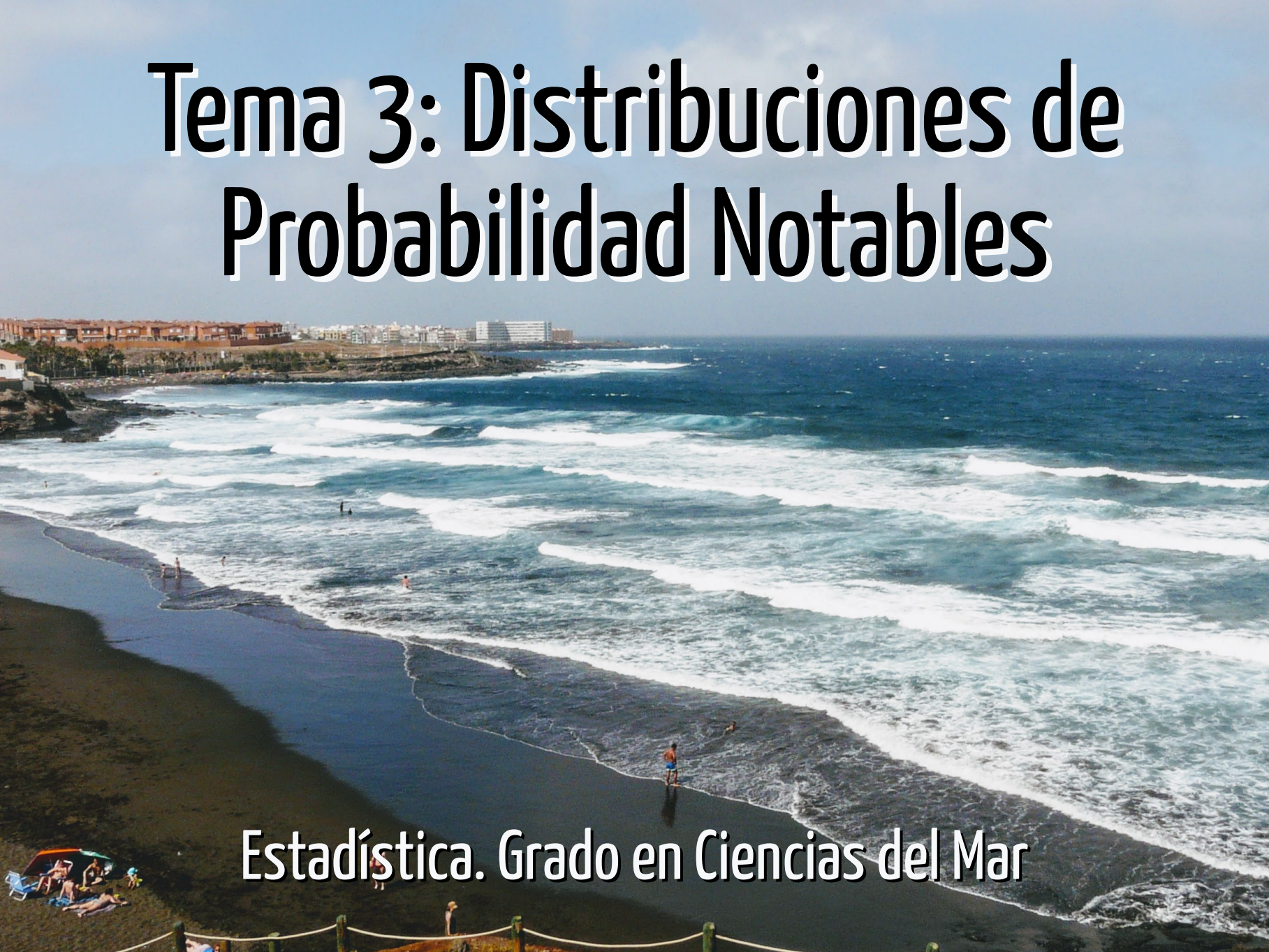
Tema 3: Distribuciones de Probabilidad Notables



Estadística. Grado en Ciencias del Mar

Distribuciones de probabilidad notables o "especiales"

En esta sección nos ocuparemos de algunas distribuciones de probabilidad que hemos llamado *"especiales"* simplemente en razón de su uso frecuente en las aplicaciones prácticas:

Distribuciones discretas

- Binomial
- Hipergeométrica
- De Poisson

Distribuciones continuas

- Exponencial
- Uniforme
- Normal

Distribución Binomial

Distribución Binomial

Jakob Bernoulli (1654–1705)

- Se llama **experimento de Bernoulli** a aquél que tiene sólo dos resultados posibles, 1 (éxito) y 0 (fracaso) con probabilidades respectivas p y $1-p$.

- Se llama **distribución binomial de parámetros n y p** y se denota $B(n,p)$, a la distribución de la variable:

X = *"Número de éxitos en n experimentos* ***independientes*** *de Bernoulli de parámetro p"*

Distribución Binomial

La función de probabilidad de la variable X con distribución $B(n, p)$ es de la forma:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \;\; k = 0, 1, \ldots, n$$

Distribución Binomial

Ejemplos

1. Se lanza 4 veces al aire una moneda equilibrada y se considera X= *"Número de caras en los cuatro lanzamientos"*. Entonces $X \approx B(4, \frac{1}{2})$

2. Se aplica un tratamiento a cada uno de 8 pacientes. El tratamiento puede tener éxito (cura al paciente) con probabilidad p o fracasar con probabilidad 0.3. Sea X el número de pacientes que se curan entre los ocho. Entonces $X \approx B(8, p)$

3. En cierta especie de peces el porcentaje de machos es del 40% y el de hembras el 60% restante. Durante un estudio se capturan al azar 50 ejemplares de esta especie. La variable X=*"Número de machos en esta muestra de peces"* sigue una distribución $B(50, 0.4)$

Deducción de la función de probabilidad de la distribución binomial.

En el ejemplo 2 anterior, sea X="*Número de pacientes que se curan entre los 8 tratados*" $\approx B(8,p)$. Calculemos $P(X=3)$

- Llamemos B al suceso *"el paciente se cura"* y B^c a su suceso contrario, siendo $P(B)=p$ y $P(B^c)=1-p$.
- Una de las formas en que se pueden curar sólo 3 pacientes es que sean los tres primeros:

$$B \cap B \cap B \cap B^c \cap B^c \cap B^c \cap B^c \cap B^c$$

- Como cada paciente se cura o no independientemente del resto, la probabilidad de que ocurra el suceso anterior es:

$$P(B \cap B \cap B \cap B^c \cap B^c \cap B^c \cap B^c \cap B^c) =$$

$$= P(B)P(B)P(B)P(B^c)P(B^c)P(B^c)P(B^c)P(B^c) =$$

$$= P(B)^3 \cdot P(B^c)^5 = p^3 \cdot (1-p)^5$$

Deducción de la función de probabilidad de la distribución binomial.

- El número de formas en que curarse 3 pacientes entre 8 coincide con el número de formas en que podemos elegir 3 posiciones entre 8, para colocar en ellas las B:

$$\binom{8}{3} = \frac{8!}{5! \cdot 3!}$$

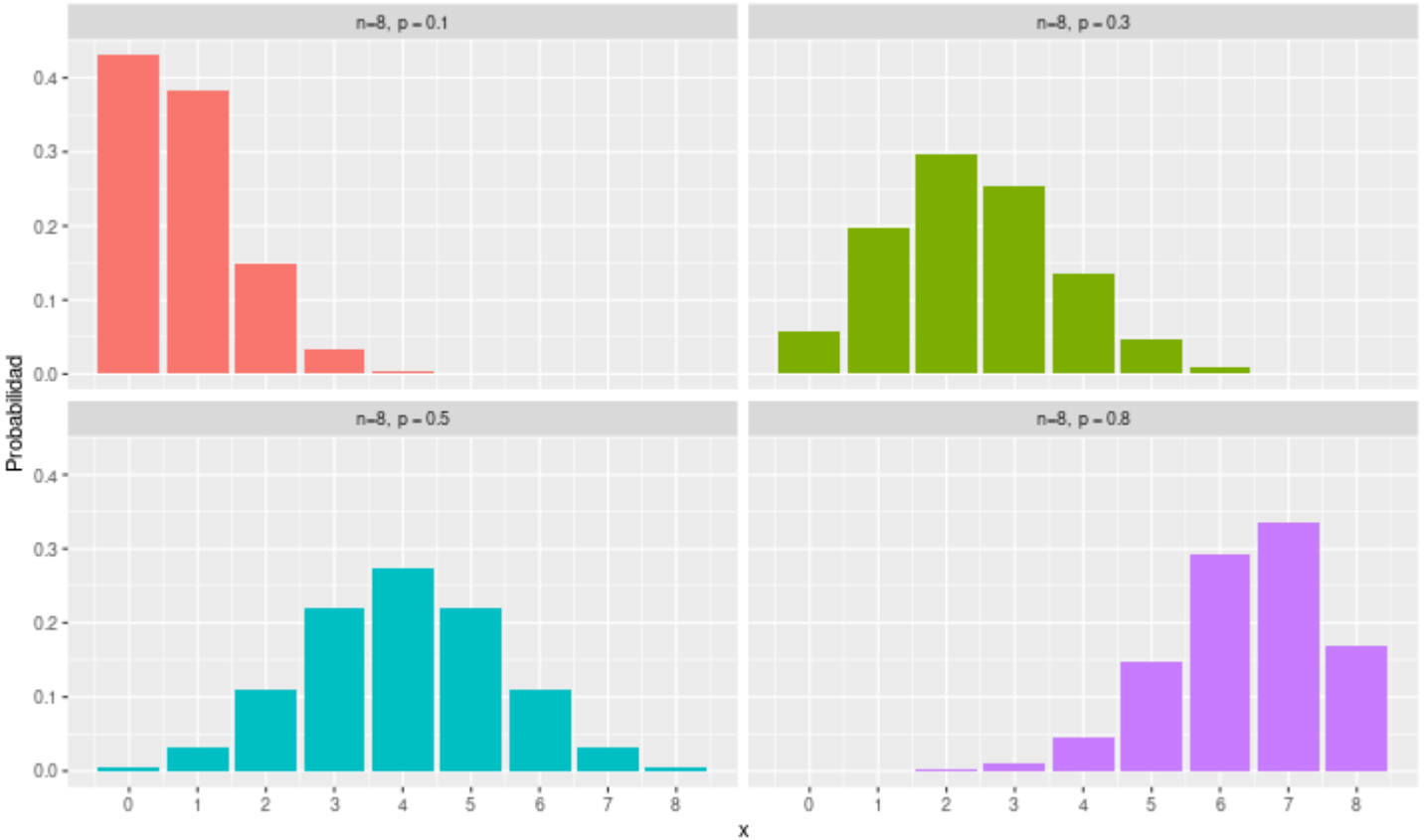
- La probabilidad total de que se curen 3 pacientes entre 8 será la suma de las probabilidades de cada una de estas $\binom{8}{3}$ formas. Como cada una de ellas tiene la misma probabilidad $p^3(1-p)^5$, la probabilidad total será:

$$P(X=3) = \binom{8}{3} p^3 (1-p)^5$$

- En general, si $X \approx B(n,p)$:

$$P\left(X=k\right) = \binom{n}{k} p^k (1-p)^{n-k}, \;\; k = 0, 1, \ldots$$

Representación gráfica de la distribución binomial para n=8 y varios valores de p



Esperanza y varianza de la distribución binomial.

$$E[X] = np$$

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

Distribución Hipergeométrica

Distribución Hipergeométrica

Christiaan Huygens (1629–1695)

- Supongamos que se dispone de una población finita de tamaño N, que está dividida en dos grupos: *éxitos* (con N_E elementos) y *fracasos* (con $N - N_E$ elementos).
- Se llama **Distribución Hipergeométrica** $H(n, N, N_E)$ a la distribución de probabilidad de la variable:

X = *"Número de éxitos obtenidos al extraer al azar y sin reemplazamiento n objetos de esta población"*

Distribución Hipergeométrica

La función de probabilidad de la distribución hipergeométrica $H(n, N, N_E)$ es de la forma:

$$P(X = k) = \frac{\binom{N_E}{k}\binom{N-N_E}{n-k}}{\binom{N}{n}}$$

Obviamente el valor de k está comprendido entre $\max\{0, n-(N-N_E)\}$ y $\min\{N_E, n\}$

La primera referencia histórica a la distribución hipergeométrica (aún sin ese nombre) aparece en el problema 4 del libro De Ratiociniis in Ludo Aleae (1657, p. 12) de Christiaan Huygens.

Deducción de la función de probabilidad de la distribución hipergeométrica

La función de probabilidad de la distribución hipergeométrica se sigue directamente de la regla de Laplace:

- Que ocurra el suceso $X = k$ significa que en la muestra de tamaño n hay k éxitos y $n - k$ fracasos.
- Los éxitos pueden ocurrir de $\binom{N_E}{k}$ formas y los fracasos de $\binom{N-N_E}{n-k}$ formas (no hay repeticiones y no importa el orden).
- Por tanto, hay $\binom{N_E}{k} \cdot \binom{N-N_E}{N-k}$ formas distintas en que pueden ocurrir k éxitos y $n - k$ fracasos (casos favorables).
- Hay $\binom{N}{n}$ formas de escoger n objetos de entre N (casos posibles).
- Por tanto, aplicando la regla de Laplace, $P(X = k) = \frac{\binom{N_E}{k}\binom{N-N_E}{n-k}}{\binom{N}{n}}$

Distribución Hipergeométrica

Ejemplo:

- En una pecera hay 10 peces de los cuales 6 son machos y 4 hembras. Se extraen al azar y sin reemplazamiento 5 peces. ¿Cuál es la probabilidad de que sean 3 machos y 2 hembras?

La variable *X = "Número de machos en la muestra de 5 peces"* es hipergeométrica $H(5, 10, 6)$

La probabilidad pedida es entonces:

$$P(X = 3) = \frac{\binom{6}{3}\binom{4}{2}}{\binom{10}{5}} = \frac{20 \cdot 6}{252} = 0.4762$$

Ejemplo:

- En las mismas condiciones, ¿cuál es la probabilidad de que los cinco peces extraídos sean machos?

$$P(X=5) = \frac{\binom{6}{5}\binom{4}{0}}{\binom{10}{5}} = \frac{6}{252} = 0.0238$$

- ¿Cuál es la probabilidad de que de los cinco peces extraídos sólo uno sea macho?

$$P(X=1) = \frac{\binom{6}{1}\binom{4}{4}}{\binom{10}{5}} = \frac{6}{252} = 0.0238$$

Esperanza y varianza de la distribución hipergeométrica

Llamando $p = \frac{N_E}{N}$:

$$E[X] = \frac{n \cdot N_E}{N} = np$$

$$Var(X) = np(1-p)\frac{(N-n)}{(N-1)}$$

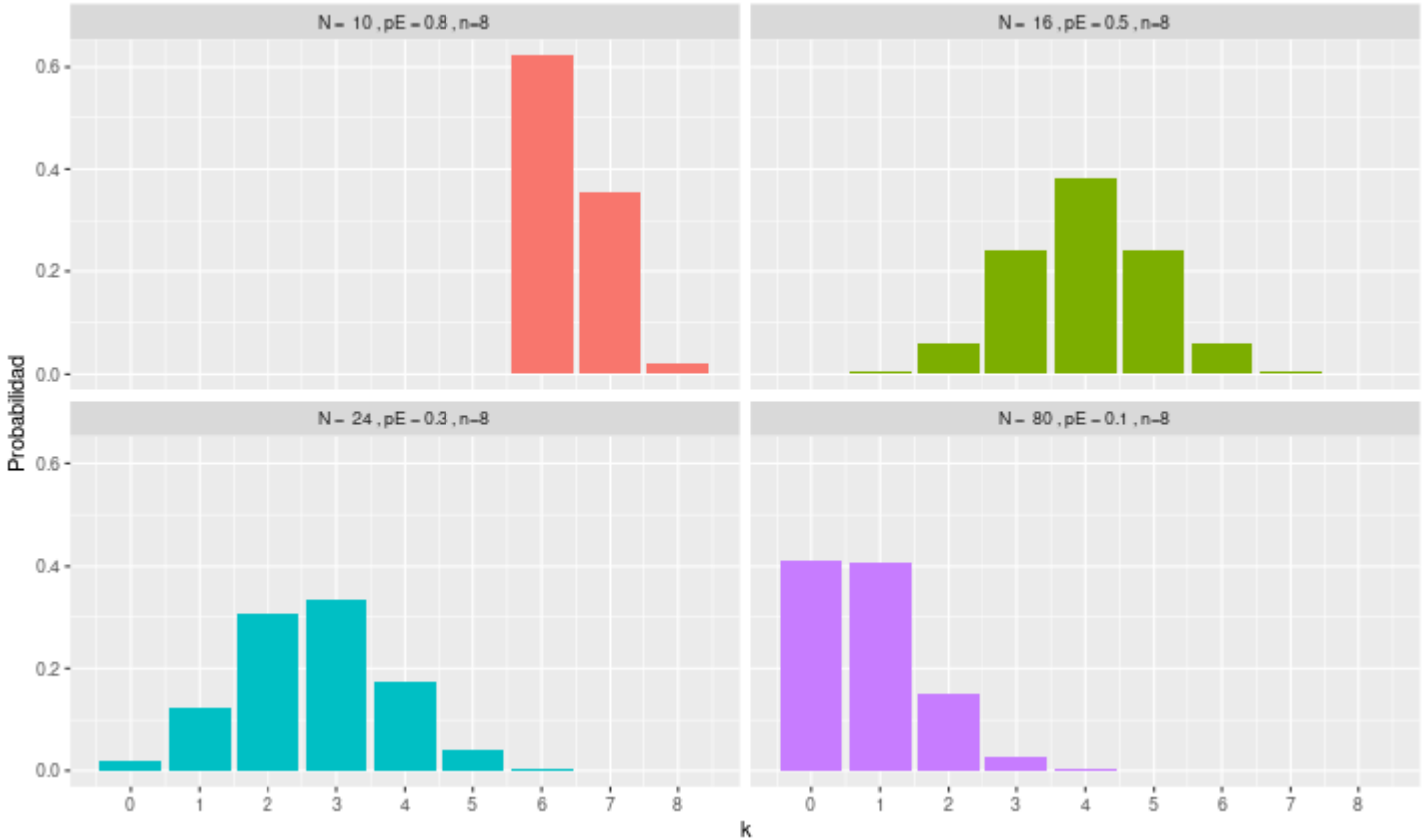
Convergencia de la distribución hipergeométrica a la binomial

- La diferencia fundamental entre las distribuciones binomial e hipergeométrica es que mientras en la primera hay reemplazamiento, en la segunda no.

- Cuando N es muy grande respecto de n, *aún cuando haya reemplazamiento*, es **muy difícil** que el mismo elemento sea extraído dos veces, por lo que la distribución binomial produce probabilidades muy similares a la hipergeométrica.

- La hipergeométrica y la binomial tienen la misma esperanza, y sus varianzas se aproximan cuando N es muy grande respecto de n.

Representación gráfica de la distribución binomial para $n = 8$ y varios valores de N y pE (probabilidad de éxito)



Distribución de Poisson

Distribución de Poisson

Siméon D. Poisson (1781–1840)

Aparece asociada a la variable aleatoria consistente en **contar** el número de ocurrencias de cierto proceso que se caracterizan por estar distribuidas:

- independientemente unas de otras
- de modo completamente aleatorio
- con tasa (densidad) constante
- en un medio continuo.

Distribución de Poisson

Ejemplos:

- El número de llamadas que llegan a una centralita telefónica durante un intervalo de tiempo.
- El número de tortugas que llegan para anidar a una playa durante un intervalo de tiempo.
- El número de partículas emitidas por un compuesto radiactivo durante un cierto periodo.
- El número de nidos de tortuga en una porción rectangular de una playa.
- El número de accidentes de tráfico que ocurren en una región durante un periodo determinado.

Distribución de Poisson

Sean:

- X_t = *"Número de eventos que ocurren en un periodo de duración t."*
- λ = *tasa media de ocurrencia por unidad de tiempo* de los eventos de interés (número medio de eventos por unidad de tiempo).

X_t sigue una **distribución de Poisson** de parámetro λ, y lo denotaremos $X_t \approx P(\lambda)$, si su función de probabilidad es de la forma:

$$P(X_t = k) = e^{-\lambda t} \frac{(\lambda t)^k}{k!}$$

Distribución de Poisson. Ejemplo:

Supongamos que el número de tortugas que llegan a desovar a una playa sigue una distribución de Poisson de parámetro λ= 3 tortugas por día.

- La probabilidad de que en 1 día lleguen 2 tortugas es:

$$P(X_1 = 2) = e^{-3}\frac{3^2}{3!} = 0.224$$

- La probabilidad de que en 1 día lleguen 3 o más tortugas es:

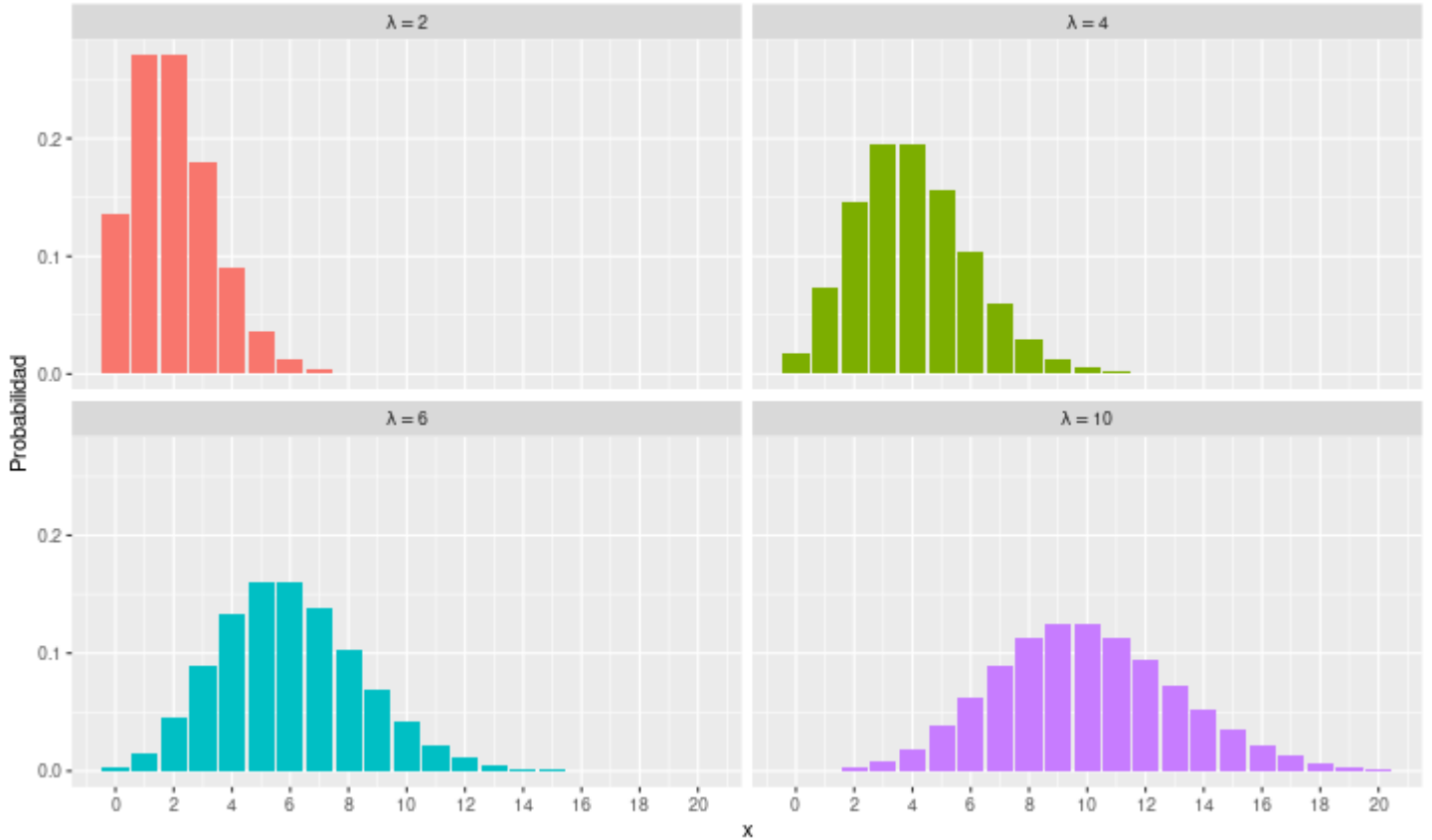
$$P\left(X_1 \geq 3\right) = 1 - P\left(X_1 < 3\right) = 1 - \left[P\left(X_1 = 0\right) + P\left(X_1 = 1\right) + P\left(X_1 = 2\right)\right] =$$

$$= 1 - \left(e^{-3}\frac{3^0}{0!} + e^{-3}\frac{3^1}{1!} + e^{-3}\frac{3^2}{3!}\right) = 0.58$$

- La probabilidad de que en 2 días lleguen 5 tortugas es:

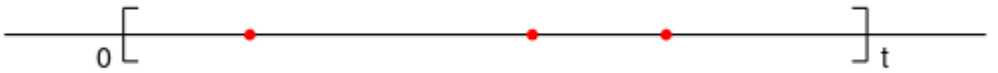
$$P\left(X_2 = 5\right) = e^{-3\cdot 2}\frac{(3\cdot 2)^5}{5!} = 0.1606$$

Representación gráfica de la distribución de Poisson para varios valores de λ

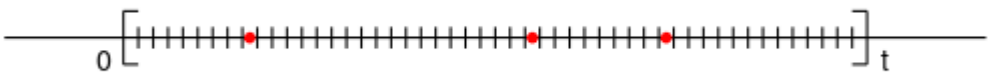


Deducción de la función de probabilidad de la distribución de Poisson

- Sea X una variable aleatoria con distribución de Poisson que cuenta el número de eventos que ocurren al azar e independientemente en un intervalo $[0, t]$ siendo λt el número esperado de eventos en ese intervalo.



- Dividimos el intervalo $[0, t]$ en n partes iguales; para un n lo suficientemente grande ($n \to \infty$), las partes en que lo hemos dividido llegan a hacerse tan pequeñas que en cada una de ellas puede ocurrir como máximo un evento, o ninguno.



- Como el número esperado de eventos en $[0, t]$ es λt, la probabilidad de que ocurra un evento en una de las n partes en que hemos dividido el intervalo puede aproximarse por $p = \frac{\lambda t}{n}$

Deducción de la función de probabilidad de la distribución de Poisson

- Dado que los distintos eventos ocurren de forma independiente, la probabilidad de que ocurran k eventos en las n partes en que hemos dividido el intervalo $[0, t]$ puede aproximarse por una binomial $B(n, p)$, y entonces:

$$P(X = k) = \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k}$$

$$\lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k} =$$

$$\lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{n^k} \frac{(\lambda t)^k}{k!} \left(1 - \frac{\lambda t}{n}\right)^{n-k} =$$

$$= \frac{(\lambda t)^k}{k!} \lim_{n\to\infty} 1 \cdot \left(1 - \frac{1}{n}\right) \cdot \dots \cdot \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda t}{n}\right)^n \left(1 - \frac{\lambda t}{n}\right)^{-k} =$$

$$\frac{(\lambda t)^k}{k!} 1 \cdot 1 \cdot \dots \cdot e^{-\lambda t} \cdot 1 = \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

Esperanza y Varianza de la distribución de Poisson

$$E[X] = \lambda t$$

$$Var(X) = \lambda t$$

Distribución exponencial

Distribución exponencial

- Esta distribución aparece asociada a fenómenos en los que la variable que se considera es la distancia entre eventos puntuales que se presentan en un medio continuo de acuerdo con una distribución de Poisson.
- Supongamos que $Y_t \approx P(\lambda t)$ cuenta el número de tortugas que llegan a una playa durante un intervalo de tiempo de duración t.
- Sea X el tiempo que falta hasta que llegue la siguiente tortuga.
- La probabilidad de que hasta la llegada de la próxima tortuga pase un tiempo mayor que t es igual a la probabilidad de que en un intervalo de longitud t no llegue ninguna tortuga:

$$P\left(X \leq t\right) = 1 - P\left(X > t\right) = 1 - P\left(Y_t = 0\right) = 1 - e^{-\lambda t}\frac{\lambda^0}{0!} = 1 - e^{-\lambda t}$$

- Si λ es el número medio de tortugas que llegan por unidad de tiempo, $\mu = \frac{1}{\lambda}$ es el tiempo medio que transcurre entre llegada y llegada.

Distribución exponencial

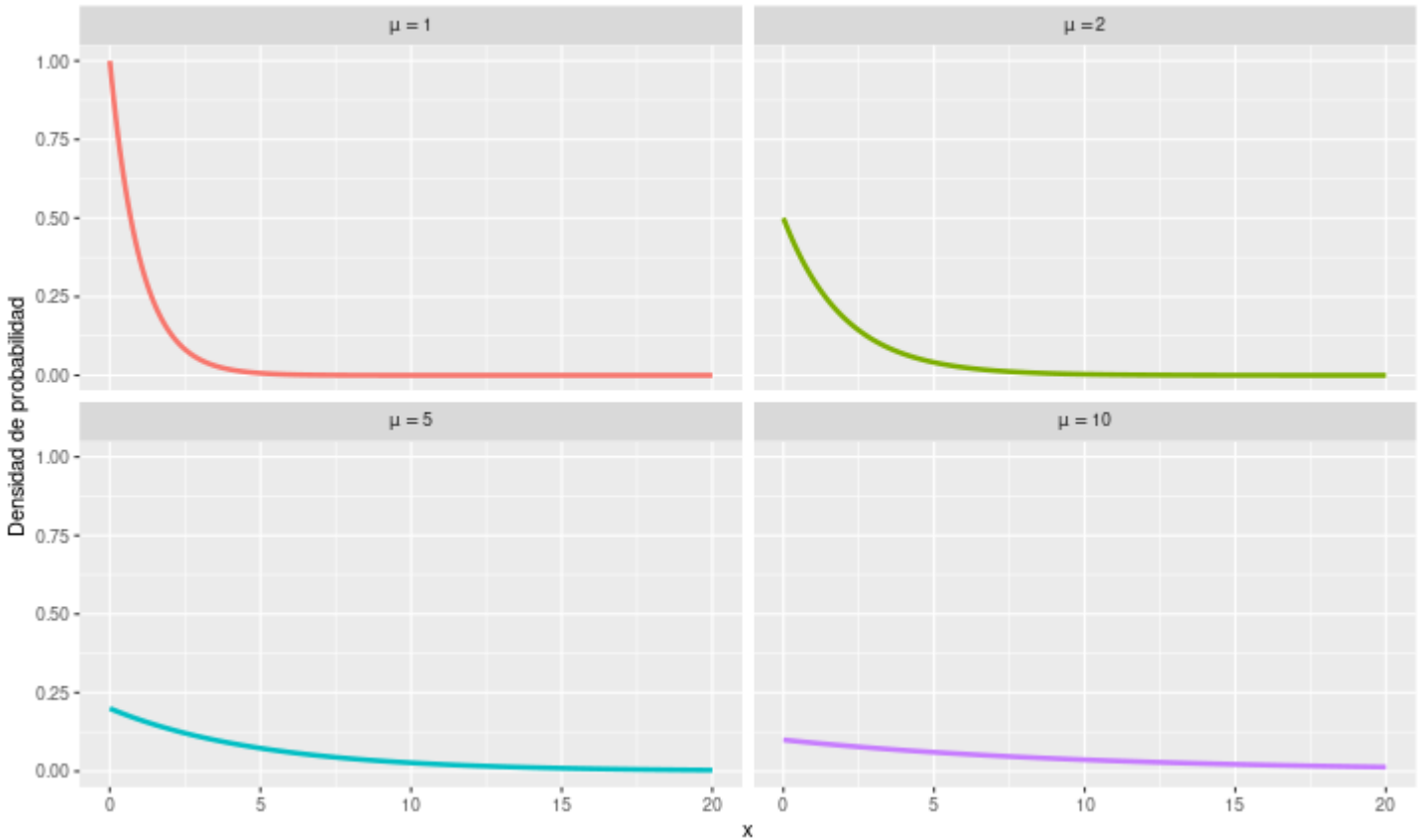
La variable aleatoria X se dice que sigue una **distribución exponencial** de parámetro μ, y se denota como $X \approx exp(\mu)$, si su función de distribución es de la forma:

$$F(t) = 1 - e^{-\frac{1}{\mu}t}$$

Su función de densidad es entonces:

$$f(t) = \frac{1}{\mu} e^{-\frac{1}{\mu}t}$$

Representación gráfica de la función de densidad de la distribución exponencial para varios valores de μ



Ejemplo:

El tiempo entre la caída de rayos durante una tormenta sigue una distribución exponencial de parámetro $\mu = 30\ segundos$

(1). ¿Cuál es la probabilidad de que el próximo rayo caiga antes de 20 segundos?

$$X \approx exp(30) \Rightarrow P\left(X \leq 20\right) = F\left(20\right) = 1 - e^{-\frac{20}{30}} = 0.4866$$

(2). ¿Cuál es la probabilidad de que el próximo rayo tarde más de 30 segundos en caer?

$$P\left(X > 30\right) = 1 - F\left(30\right) = e^{-\frac{30}{30}} = 0.3679$$

(3). ¿Cuál es la probabilidad de que el próximo rayo tarde entre 50 y 70 segundos en caer?

$$P\left(50 \leq X \leq 70\right) = \int_{50}^{70} \frac{1}{30} e^{-\frac{1}{30}x} dx = \left[-e^{-\frac{1}{30}x}\right]_{50}^{70} = e^{-\frac{50}{30}} - e^{-\frac{70}{30}} = 0.0919$$

Esperanza y varianza de la distribución exponencial

$$E[X] = \mu$$

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

Distribución uniforme

Distribución uniforme $U[a, b]$

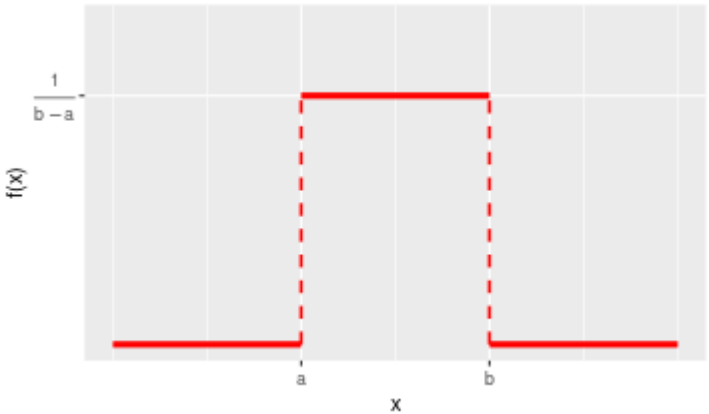
Esta es la distribución de la variable:

X = *"Resultado de elegir un valor al azar en el intervalo* $[a, b]$"

cuando la probabilidad se reparte de manera homogénea (uniforme) sobre el intervalo, esto es, no hay más probabilidad de observar valores en una zona que en otra del mismo.

Su función de densidad es de la forma:

$$f(x) = \frac{1}{b-a} \quad a \leq x \leq b$$



Ya hemos visto esta distribución en el ejemplo del lugar (aleatorio) donde se avería un coche en una carretera larga y recta.

Distribución uniforme: Función de distribución

$$F(x) = P\left(X \leq x\right) = \frac{x-a}{b-a} \quad x \in (a,b)$$

Distribución uniforme: Esperanza y varianza

Es fácil comprobar que:

$$E[X] = \frac{a+b}{2}$$

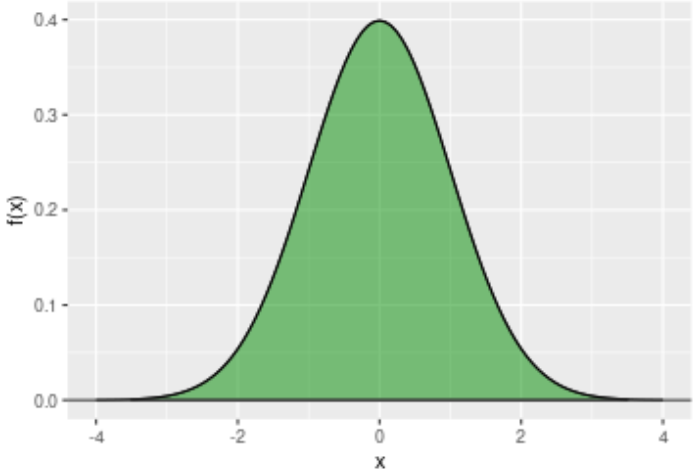
$$Var(X) = \frac{(b-a)^2}{12}$$

Distribución Normal o de Gauss

Distribución Normal o de Gauss

Carl Friedrich Gauss (1777-1855)

La distribución normal aparece asociada a variables aleatorias que se comportan de tal manera que lo más probable es observar valores en torno a la media; y a medida que los valores se alejan de la media, bien sea hacia arriba o hacia abajo, van siendo progresivamente más difíciles de observar:



Distribución Normal o de Gauss

Múltiples ejemplos de variables con esta distribución:

- Errores de Medida.
- Variables físicas: Temperatura y salinidad del mar en una zona concreta.
- Variables Biológicas: peso, talla, nivel de glucosa en sangre ...
- Consumo de energía diario en una ciudad o en una empresa.
- Kilometraje mensual recorrido por un vehículo arbitrario.
- ...

¿Por qué es tan ubicua la distribución normal?

Acción del teorema central del límite

Teorema central del límite

De forma intuitiva, el teorema central del límite afirma que dado un conjunto de variables aleatorias independientes, con cualquier distribución de probabilidad y suponiendo que su varianza sea finita, la suma de un número elevado de estas variables tiende a distribuirse de manera similar a una variable normal.

Distribución Normal o de Gauss

Una variable aleatoria X sigue una distribución Normal de parámetros μ (esperanza) y σ (desviación típica) y se denota como $X \approx N(\mu, \sigma)$, si su función de densidad de probabilidad es de la forma:

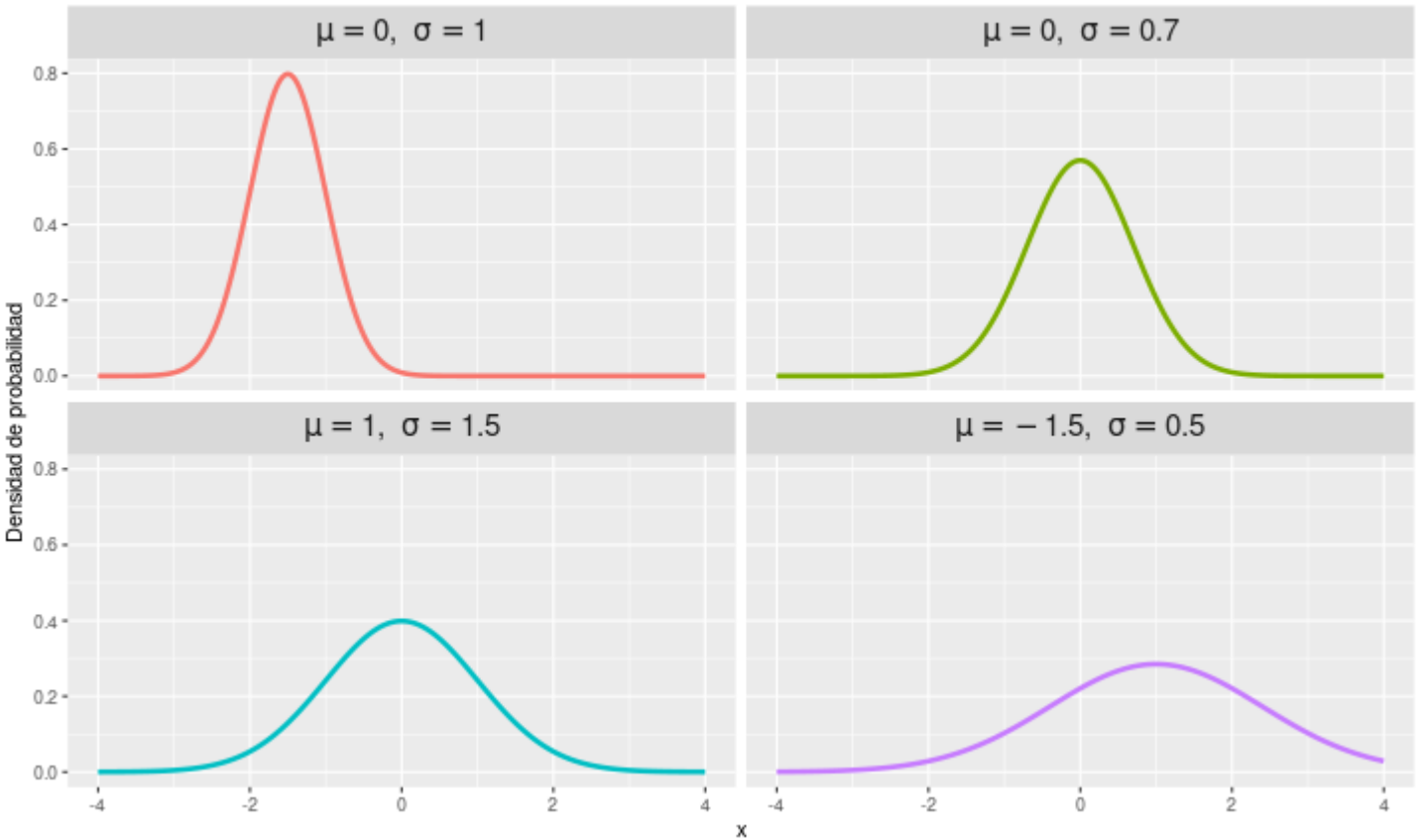
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \; x \in \mathbb{R}$$

Puede comprobarse que:

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \mu$$

$$Var[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \sigma$$

Representación gráfica de la función de densidad de la distribución normal para varios valores de μ y σ



Distribución normal: cálculo de probabilidades.

- La función de densidad de la distribución normal no puede integrarse de manera analítica, sino numérica. De ahí que para calcular probabilidades de la forma $P(X \leq x)$ o $P(a \leq X \leq b)$ haya que recurrir a calculadora, ordenador o tablas de referencia.
- El siguiente resultado permite convertir una variable $N(\mu, \sigma)$ en una $N(0, 1)$; de esta forma las probabilidades de la primera pueden calcularse a partir de las probabilidades de la segunda:

Teorema (de la tipificación): *Si una variable aleatoria X tiene distribución de probabilidad* $N(\mu, \sigma)$*, entonces,* $Z = \frac{(X-\mu)}{\sigma}$ *tiene distribución de probabilidad* $N(0, 1)$ *(normal estándar o tipificada). Por tanto:*

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$$

Distribución normal: uso de la tabla de la *N(0,1)*

- Esta tabla nos permite calcular probabilidades de la forma $P\left(Z > x\right)$ donde Z es una variable aleatoria con distribución $N\left(0,1\right)$ y x es un número de la forma $a.bc = a.b + 0.0c$. El valor de dicha probabilidad se encuentra en el cruce de la fila $a.b$ con la columna $0.0c$.
- Ejemplo: para calcular $P\left(Z > 1.23\right)$ se busca en el cruce de la fila 1.2 con la columa 0.03, donde se encuentra el valor 0.10935.
- En el caso de que se desee calcular la probabilidad de que Z sea mayor que un número negativo, se puede proceder aprovechando la circunstancia de que la función de densidad de Z es simétrica y por tanto:

$$P\left(Z > -x\right) = P\left(Z < x\right) = 1 - P(Z > x)$$

- Para calcular la probabilidad de un intervalo:

$$P(a \leq Z \leq b) = P(Z > a) - P(Z > b)$$

Distribución normal: uso de la tabla de la *N(0,1)*

- Sea $X \approx N(\mu = 8, \sigma = 3)$:

$$P(X \leq 10) = P\left(Z \leq \frac{10-8}{3}\right) =$$

$$= P(Z \leq 0.66) = 1 - P(Z > 0.66)$$

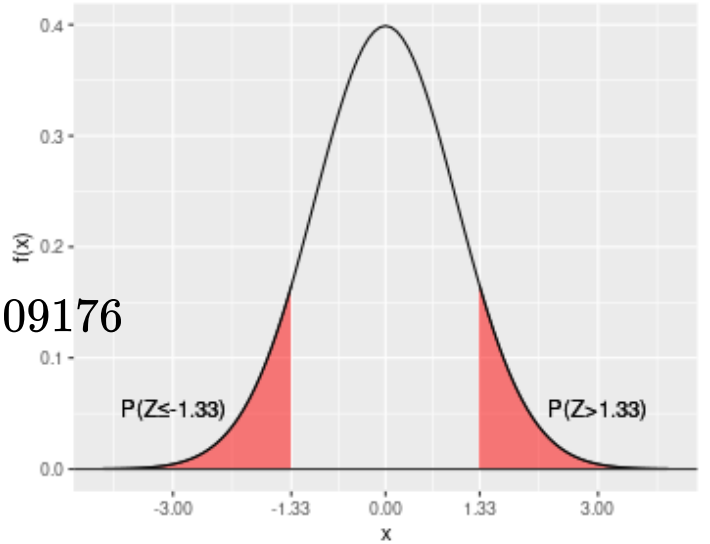
$$= 1 - 0.25463 = 0.74537$$

Distribución normal: uso de la tabla de la *N(0,1)*

- Sea $X \approx N(\mu = 8, \sigma = 3)$:

$$P(X \leq 4) = P\left(Z \leq \frac{4-8}{3}\right) =$$

$$= P(Z \leq -1.33) = P(Z > 1.33) = 0.09176$$



Propiedad reproductiva de la distribución Normal.

Dadas n variables aleatorias normales e independientes, tales que $X_i \approx N\left(\mu_i, \sigma_i\right),\ i = 1, \ldots, n$, su suma $\sum_{i=1}^{n} X_i$ sigue también una distribución normal, siendo:

$$\sum_{i=1}^{n} X_i \approx N\left(\sum_{i=1}^{n} \mu_i, \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

Propiedad reproductiva de la distribución Normal.

Como consecuencia de esta propiedad, en el caso particular de que $X_i \approx N(\mu, \sigma) \, \forall i = 1, \ldots, n$, aplicando las propiedades de la esperanza y la varianza, se tiene que:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

o, expresado de otra forma,

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \approx N(0, 1)$$

Ejemplo

El peso de las capturas diarias realizadas por un barco de pesca sigue una distribución normal de media 87 kg y desviación típica 16 kg. Durante un mes, el barco ha salido a pescar 20 días. ¿Cuál es la probabilidad de que el peso total de las capturas supere los 1800 kg?

Si $X \approx N(87, 16)$ entonces el peso total de las capturas sigue una distribución:

$$T = \sum_{i=1}^{20} X_i \approx N\left(20 \cdot 87, 16 \cdot \sqrt{20}\right) = N\left(1740, 71.55\right)$$

y por tanto:

$$P\left(T > 1800\right) = 1 - P\left(T \leq 1800\right) = 1 - \mathsf{pnorm}(1800, 1740, 71.55) = 0.20$$

Ejemplo

El peso de las capturas diarias realizadas por un barco de pesca sigue una distribución normal de media 87 kg y desviación típica 16 kg. Durante un mes, el barco ha salido a pescar 20 días, con probabilidad 0.95 ¿cuál será el peso mínimo de las capturas? También con probabilidad 0.95 ¿Cuál será el peso máximo?

Mínimo: es el percentil 5, pues:

$$P(X \geq m) = 0.95 \Rightarrow P(X \leq m) = 0.05 \Longrightarrow$$

$$\Longrightarrow m = \mathsf{qnorm}(0.05, 1740, 71.55) = 1622.311$$

Máximo: es el percentil 95, pues:

$$P(X \leq M) = 0.95 \Rightarrow M = \mathsf{qnorm}(0.95, 1740, 71.55) = 1857.69$$

Ejemplo

El peso de las capturas diarias realizadas por un barco de pesca sigue una distribución normal de media 87 kg y desviación típica 16 kg. Si se eligen al azar 10 días de pesca ¿Cuál es la probabilidad de que el peso medio de las capturas realizadas durante esos días esté entre 85 y 90 kg?

Como $X \approx N(87, 16)$, la captura media realizada durante 10 días seguirá una distribución:

$$\overline{X} = \frac{1}{10}\sum_{i=1}^{10} X_i \approx N\left(87, \frac{16}{\sqrt{10}}\right) = N\left(87, 5.06\right)$$

y por tanto:

$$P\left(85 \leq \overline{X} \leq 90\right) = P\left(\overline{X} \leq 90\right) - P\left(\overline{X} \leq 85\right) =$$

$$= \mathsf{pnorm}(90, 87, 5.06) - \mathsf{pnorm}(85, 87, 5.06) = 0.377$$

Ejemplo

Se ha observado que la desviación típica del peso de las capturas diarias realizadas por cada uno de los barcos que faenan en cierta región es $\sigma = 20$ kg. Se ha hecho un seguimiento de las capturas realizadas por el barco *Juanita II* durante los últimos 36 días de pesca. El peso total de las capturas durante ese periodo ha sido de 3532 kg, lo que da una captura media de $\overline{x} = 98.11$ kg por día. Suponiendo que el peso de las capturas diarias sigue una distribución $N(\mu, \sigma = 20)$, utiliza la información anterior para obtener un intervalo de confianza al 95% para el valor de μ.

Se sabe que:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N\left(0, 1\right)$$

Podemos utilizar R para encontrar los valores $z_{0.025}$ y $z_{0.975}$ tales que:

$$P\left(z_{0.025} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{0.975}\right) = 0.95$$

Ejemplo

Tenemos que $z_{0.025}$=`qnorm(0.025)`=-1.96 y $z_{0.975}$=`qnorm(0.975)`=1.96, y por tanto:

$$P\left(-1.96 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95$$

De aquí podemos despejar:

$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

$$P\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Ejemplo

Como $\sigma = 20$, $n = 36$ y $\overline{x} = 98.11$, sustituimos en la expresión anterior y se obtiene que con una confianza del 95%:

$$98.11 - 1.96\frac{20}{\sqrt{36}} \leq \mu \leq 98.11 + 1.96\frac{20}{\sqrt{36}}$$

$$98.11 - 6.53 \leq \mu \leq 98.11 + 6.53 \Rightarrow 91.58 \leq \mu \leq 104.64$$

Dicho de otra manera, la media $\overline{x}$ de 36 días nos dio el valor 98.11; podemos por tanto asumir que el peso medio μ de las capturas diarias *en general para todos los días, no solo para esos 36 días en particular*, es un valor parecido a 98.11.

Este valor es una **estimación** de μ. El procedimiento anterior (intervalo de confianza) nos permite evaluar el margen de error para dicha estimación: el verdadero valor de μ a lo mejor no es exactamente 98.11, pero con una confianza del 95% se diferencia de 98.11 como mucho en 6.53 kg, y está en el intervalo [91.58, 104.64]

Teorema Central del Límite.

El Teorema Central del Límite establece que dada una colección de variables aleatorias **independientes** $X_1, X_2, \ldots, X_n$ tales que $E[X_i] = \mu_i$ y $Var(X_i) = \sigma_i^2$, cuando $n \to \infty$ la distribución de probabilidad de la suma de estas variables es aproximadamente normal:

$$\sum_{i=1}^{n} X_i \approx N\left(\sum_{i=1}^{n} \mu_i,\ \sqrt{\sum_{i=1}^{n} \sigma_i^2}\right)$$

En el caso particular de que todas las X_i tengan la misma distribución, esto es, $E[X_i] = \mu$ y $Var(X_i) = \sigma^2\ \forall i$ se tiene que:

$$\sum_{i=1}^{n} X_i \approx N\left(n\mu,\ \sigma\sqrt{n}\right)$$

Aplicaciones del TCL: aproximación de la Binomial por la Normal:

Si X es una variable $B\left(n,p\right)$, su valor representa el número de éxitos en n experimentos independientes en cada uno de los cuales la probabilidad de éxito es p.

Si definimos el resultado de cada experimento como:

$$X_i = \begin{cases} 0 & 1-p \quad (\text{fracaso}) \\ 1 & p \qquad (\text{éxito}) \end{cases} \qquad (\text{Variable de Bernoulli})$$

podemos expresar la binomial como suma de variables de Bernoulli:

$$X = X_1 + X_2 + \cdots + X_n$$

Obsérvese que:

$$\mu_i = E[X_i] = 1 \cdot p + 0 \cdot (1-p) = p$$

$$\sigma_i^2 = Var(X_i) = E[X_i^2] - \left(E[X_i]\right)^2 = 1^2 \cdot p + 0^2 \cdot (1-p) - p^2 = p(1-p)$$

Aplicaciones del TCL: aproximación de la Binomial por la Normal

Entonces, si el valor de n es grande:

$$B(n,p) \approx X = \sum_{i=1}^{n} X_i \underset{n\to\infty}{\approx} N\left(n\mu,\ \sigma\sqrt{n}\right) = N\left(np, \sqrt{np(1-p)}\right)$$

- En general la aproximación es razonablemente buena cuando $n \geq 30$, $np \geq 5$ y $n(1-p) \geq 5$
- Como la normal es continua y la binomial es discreta, en el cálculo aproximado se considera que el valor (discreto) k es equivalente al intervalo $\left[k - \frac{1}{2}, k + \frac{1}{2}\right]$

Ejemplo:

Si $X \approx B(120, 0.35)$, se puede aproximar por

$$X_N \approx N(120 \cdot 0.35, \sqrt{120 \cdot 0.35 \cdot 0.65} = N(42, 5.2249)$$

- $P(X = 40) = 0.0716$ (Valor exacto)
- $P(X = 40) \cong P(39.5 < X_N < 40.5) = P(X_N < 40.5) - P(X_N < 39.5) = \\ = \ 0.387 - 0.3162 = 0.0709$ (Valor aproximado)
- $P(X \leq 40) = 0.3905$ (Valor exacto)
- $P(X \leq 40) \cong P(X_N \leq 40.5) = 0.387$ (Valor aproximado)
- $P(X \geq 40) = 0.6811$ (Valor exacto)
- $P(X \geq 40) \cong P(X_N \geq 39.5) = 0.6838$ (Valor aproximado)

Aproximación de la Binomial por la Normal

Supongamos que $X \approx B(n,p)$ es el número de éxitos en n pruebas independientes; si realizáramos efectivamente este experimento, $\hat{p} = \frac{X}{n}$ sería la **proporción observada** de éxitos en esas n pruebas. Si n es grande:

$$X \approx N\left(np, \sqrt{np\,(1-p)}\right)$$

y por tanto:

$$\hat{p} = \frac{X}{n} \approx N\left(\frac{np}{n}, \frac{\sqrt{np\,(1-p)}}{n}\right) = N\left(p, \sqrt{\frac{p\,(1-p)}{n}}\right)$$

De aquí se sigue que:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx N\,(0,1)$$

Aproximación de la Binomial por la Normal

Utilizando la distribución normal $N(0,1)$, podemos encontrar el valor $z_{\alpha/2}$ tal que:

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1-\alpha$$

De aquí se deduce que la diferencia entre el valor (desconocido) de p y el valor (observado) $\hat{p}$ en la muestra cumple:

$$P\left(|\hat{p}-p| \leq z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

Este resultado es aproximado, y sólo es válido si $n \geq 30$, $np \geq 5$ y $n(1-p) \geq 5$.

Aproximación de la Binomial por la Normal

La probabilidad anterior puede expresarse también como:

$$P\left(|p-\hat{p}| \leq z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

o lo que es lo mismo:

$$P\left(-z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}} \leq p-\hat{p} \leq z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

y de aquí:

$$P\left(\hat{p}-z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}} \leq p \leq \hat{p}+z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right) = 1-\alpha$$

esto es,

$$P\left(p \in \left[\hat{p}-z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}, \hat{p}+z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right]\right) = 1-\alpha$$

Aproximación de la Binomial por la Normal

Una vez realizado el experimento, $\hat{p}$ *no es una variable aleatoria, sino un valor fijo*. Podemos decir entonces que tenemos una confianza $1-\alpha$ en que el valor (desconocido) de p cae en el intervalo:

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{p\,(1-p)}{n}}\right]$$

Ahora bien, este intervalo es poco útil en la práctica, ya que sus extremos dependen de p, que es desconocido. Una opción es sustituirlo por el valor observado $\hat{p}$:

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right]$$

(*intervalo de Wald*) aunque, como es obvio, no podemos garantizar entonces que se consigue la confianza deseada $1-\alpha$. **Por ello no es recomendable utilizar este intervalo en la práctica**.

Aproximación de la Binomial por la Normal

En Agresti & Caffo, 2000 se señala como la aproximación del intervalo de Wald (al 95% de confianza) anterior mejora notablemente si se añaden 4 pseudo-observaciones a la muestra, 2 éxitos y 2 fracasos. De esta forma:

- n se sustituye por $\widetilde{n} = n + 4$
- El número de éxitos X se sustituye por $X + 2$. Por tanto la proporción observada de éxitos $\hat{p}$ se sustituye por $\tilde{p} = \frac{X+2}{n+4}$
- El intervalo ajustado para p (*intervalo de Agresti-Coull*) es entonces:

$$\left[\tilde{p} - z_{\alpha/2}\sqrt{\frac{\tilde{p}\left(1-\tilde{p}\right)}{\widetilde{n}}}, \tilde{p} + z_{\alpha/2}\sqrt{\frac{\tilde{p}\left(1-\tilde{p}\right)}{\widetilde{n}}}\right]$$

Ejemplo

Se desea conocer la proporción de hembras en una población de peces. Con este fin se obtiene una muestra de 200 peces elegidos aleatoriamente en esta población. En la muestra 140 peces son hembras. ¿Cuál es la proporción de hembras en esa población?

La proporción de hembras en la muestra es $\frac{140}{200} = 0.7 \cong 70\%$. Para evaluar el margen de error con que la proporción de hembras en la población se aproxima a este valor calculamos el intervalo de Agresti-Coull:

- $\widetilde{n} = 200 + 4$
- $\tilde{p} = \frac{142}{204} = 0.6961$
- $\left[\tilde{p} - z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\widetilde{n}}}, \tilde{p} + z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{\widetilde{n}}}\right] = [0.6961 \pm 1.96 \cdot 0.0322] =$

$$= [0.633, 0.759]$$

Por tanto a partir de estos datos podemos tener una confianza aproximada del 95% en que la proporción de hembras en la población es un valor comprendido entre el 63.3% y el 75.9%.