# Capítulo 3

# Distribuciones de Probabilidad Notables. Teorema Central del Límite.

## 1. Introducción

En este tema estudiaremos las distribuciones de probabilidad más habituales en las aplicaciones prácticas. En primer lugar veremos algunas distribuciones discretas –Bernoulli, binomial, hipergeométrica, geométrica y de Poisson–, y seguidamente algunas distribuciones continuas –uniforme, exponencial, gamma, Weibull y Normal–. De entre las distribuciones continuas destaca la normal ya que bajo determinadas condiciones aparece como límite de muchas variables. Estudiaremos tales condiciones y su interpretación, para finalmente ver las principales distribuciones de probabilidad que aparecen en la inferencia estadística cuando se toman muestras aleatorias de poblaciones que se distribuyen normalmente.

## 2. OBJETIVOS

Al finalizar este tema alumno deberá:

1. Conocer y saber calcular probabilidades asociadas a las distribuciones discretas notables, en particular, la binomial, la hipergeométrica y la de Poisson
2. Conocer y saber calcular probabilidades asociadas a las distribuciones continuas notables.
3. Entender el significado de los parámetros característicos de cada distribución, y como la elección adecuada de los valores de los parámetros permite modelar variables observadas en la naturaleza.

4. Conocer la distribución normal y su propiedad reproductiva. Utilizar la tabla de la distribución normal estándar. Entender y ser capaz de aplicar en situaciones prácticas el teorema central del límite.

5. Conocer las principales distribuciones que surgen en la inferencia estadística asociadas al muestreo (t de Student, chi-cuadrado y F de Fisher), así como manejar sus tablas.

6. Ser capaz de utilizar R para el cálculo de probabilidades en variables con las distribuciones vistas en este capítulo.

# 3. Principales distribuciones de probabilidad discretas.

## 3.1. Distribución Uniforme Discreta.

**Definición:** Una variable aleatoria $X$ que toma un número finito $n$ de valores $\{x_1, x_2, \ldots, x_n\}$ sigue una *distribución uniforme* si todos sus valores son equiprobables. Por tanto su función de probabilidad es de la forma:

$$f(x) = P(X = x) = \begin{cases} \frac{1}{n} & x \in \{x_1, x_2, \ldots, x_n\} \\ 0 & x \notin \{x_1, x_2, \ldots, x_n\} \end{cases}$$

**Esperanza y varianza:**

$$\mu = E[X] = \sum_{i=1}^{n} x_i p(X = x_i) = \sum_{i=1}^{n} x_i \cdot \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\sigma^2 = var(X) = \sum_{i=1}^{n} (x_i - \mu)^2 p(X = x_i) = \sum_{i=1}^{n} (x_i - \mu)^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

**Ejemplo:** Si $X$= "*Resultado obtenido al lanzar un dado equilibrado*":

$$\mu = E[X] = \sum_{i=1}^{k} p_i x_i = \frac{1}{6} \sum_{i=1}^{6} i = \frac{1}{6} \cdot 21 = 3,5$$

$$\sigma^2 = var[X] = \sum_{i=1}^{k} p_i (x_i - \mu)^2 = \frac{1}{6} \sum_{i=1}^{6} (i - 3,5)^2 = 2,91$$

## 3.2. Distribución de Bernoulli $Be\,(p)$

**Definición:** Una variable aleatoria $X$ sigue una distribución de Bernoulli, $Be\,(p)$, si sólo toma dos posibles valores: 1 ("*éxito*") ó 0 ("*fracaso*"), con probabilidades respectivas $p$ y $1-p$. Su función de probabilidad es, por tanto:

$$f(k) = P(X=k) = \begin{cases} p & k=1 \\ 1-p & k=0 \\ 0 & k \notin \{0,1\} \end{cases}$$

que podemos expresar también como:

$$f\,(k) = p^k\,(1-p)^{1-k}\,,\ k=0,1$$

**Esperanza y varianza:**

$$\mu = E[X] = \sum_{k\in\{0,1\}} k \cdot p\,(X=k) = 1\cdot p + 0\cdot(1-p) = p$$

$$\sigma^2 = var\,(X) = \sum_{k\in\{0,1\}} (k-\mu)^2\,P\,(X=k) = (1-p)^2 p + (0-p)^2(1-p) = p(1-p)$$

**Ejemplo:** Se realiza el experimento aleatorio consistente en lanzar una moneda equlibrada y se define la variable aleatoria:

$$X = \begin{cases} 0 & \text{si sale cara} \\ 1 & \text{si sale cruz} \end{cases}$$

Entonces

$$X \approx Be\left(\frac{1}{2}\right)$$

La función de probabilidad en este caso es:

$$P(X=1) = \frac{1}{2};\quad P(X=0) = 1 - \frac{1}{2} = \frac{1}{2}$$

y la media y varianza:

$$\mu = p = \frac{1}{2};\quad \sigma^2 = p(1-p) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

## 3.3. Distribución Binomial $B(n, p)$

**Definición:** Una variable aleatoria $X$ sigue una distribución *Binomial de parámetros $n$ y $p$* si representa el número de éxitos obtenidos al realizar $n$ repeticiones independientes de un experimento de Bernoulli, siendo $p$ la probabilidad de éxito en cada experimento.

Obviamente sólo son posibles entre 0 y $n$ éxitos. La función de probabilidad de esta variable es de la forma:

$$f(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \; k \in \{0, 1, 2, \ldots, n\}$$

La figura 1 muestra esta función de probabilidad para diversos valores de $n$ y $p$

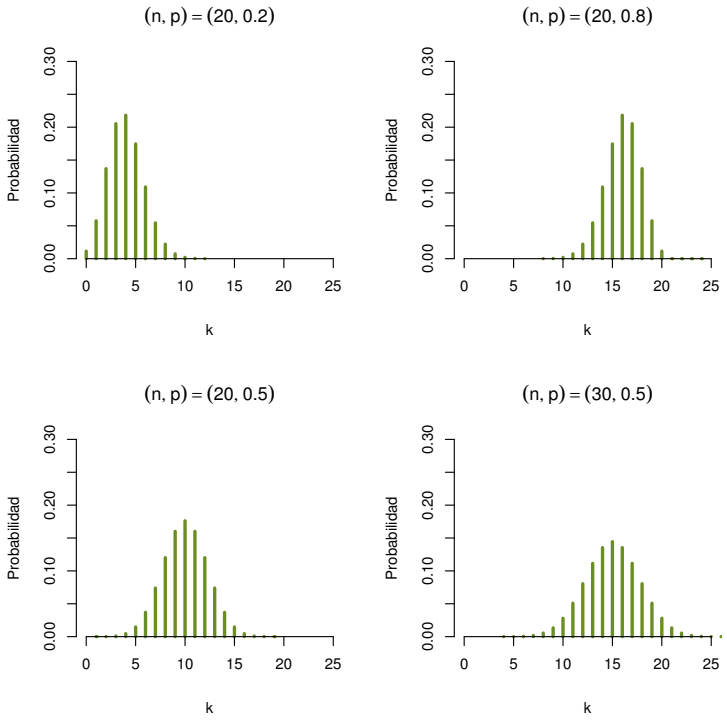


Figura 1: Función de probabilidad de la distribución binomial para diversos valores de $n$ y $p$. La altura de cada línea representa la $P(X = k)$.

**Esperanza y varianza:** Por definición, si $X \approx B(n, p)$ entonces $X = X_1 + X_2 + \cdots + X_k$,

siendo las $X_i$ variables de Bernoulli de parámetro $p$ independientes. Por tanto:

$$\begin{aligned}\mu = E[X] = E[X_1 + X_2 + \cdots + X_k] = E[X_1] + E[X_2] + \cdots + E[X_k] = \\ = p + p + \cdots + p = np \\ \sigma^2 = var(X) = var(X_1 + X_2 + \cdots + X_k) = var(X_1) + var(X_2) + \cdots + var(X_k) = \\ = p(1-p) + p(1-p) + \cdots + p(1-p) = np(1-p)\end{aligned}$$

**Ejemplo:** Se sabe que en la puesta de huevos de una tortuga, la probabilidad de que una cría sea macho es 0.30 y de que sea hembra es 0.70. El sexo de cada cría es independiente del resto. Se dispone de una puesta de 10 huevos y se considera la variable $X$="*Número de machos en la puesta*". ¿Cuál es la probabilidad de que $X = 5$?

De la descripción de esta variable se deduce que $X \approx B(10, 0{,}3)$. Por tanto:

$$P(X = 5) = \binom{10}{5} 0{,}3^5 (1 - 0{,}3)^{10-5} = 0{,}103$$

**Cálculo con R :** El programa R dispone de varias funciones para el cálculo de probabilidades asociadas a la distribución binomial. Concretamente, si $X \approx B(n, p)$, utilizando R podemos:

- Calcular el valor de la función de probabilidad*:* $P(X = k)$=`dbinom(k,n,p)`
- Calcular el valor de la función de distribución: $P(X \leq k)$=`pbinom(k,n,p)`
- Calcular los cuantiles: $q_\alpha = \text{mín}\{x : F(x) \geq \alpha\}$=`qbinom(`$\alpha$`,n,p)`
- Generar $m$ números aleatorios con distribución $B(n, p)$: `rbinom(m,n,p)`

**Ejemplo:** La siguiente sintaxis simula una muestra de 1000 valores de una distribución binomial de parámetros $n = 10$ y $p = 0{,}7$, y los representa en un diagrama de barras, junto a la representación gráfica de la función de probabilidad de la $B(10, 0{,}7)$ (figura 2). Asimismo se muestran las proporciones con que aparece cada valor $k$ en la muestra y su correspondiente probabilidad teórica $P(X = k) = \binom{10}{k} 0{,}7^k (1 - 0{,}7)^{10-k}$. Como puede apreciarse, con este valor de $n$, las probabilidades teóricas son muy similares a las proporciones muestrales observadas.

```
> n=10
> p=0.7
> muestra=rbinom(1000,n,p)
> probabilidades=dbinom(0:n,n,p)
> proporciones=prop.table(table(muestra))
> par(mfrow=c(1,2))
> plot(0:n,probabilidades,type="h",lwd=3,col="olivedrab",ylab="Probabilidad",xlab
> barplot(proporciones,xlab="k",ylab="Proporcion",main="(b)")
> prop=numeric(11);for(k in 0:10) prop[k+1]=length(which(muestra==k))/1000
> data.frame(k=0:10,Prob=round(probabilidades,3),Prop.obs=prop)

    k  Prob Prop.obs
1   0 0.000    0.000
2   1 0.000    0.000
3   2 0.001    0.003
4   3 0.009    0.008
5   4 0.037    0.033
6   5 0.103    0.097
7   6 0.200    0.207
8   7 0.267    0.256
9   8 0.233    0.236
10  9 0.121    0.116
11 10 0.028    0.044

>
```

## 3.4. Distribución Geométrica $Geo\,(p)$.

**Definición:** una variable aleatoria $X$ sigue una distribución *Geométrica de parámetro* $p$ si representa el número de experimentos de Bernoulli sucesivos e independientes que acaban en fracaso antes de que ocurra el primer éxito. Su función de probabilidad es por tanto:

$$f\,(k) = P\,(X = k) = (1-p)^k\,p,\ \ k = 0, 1, 2, \ldots$$

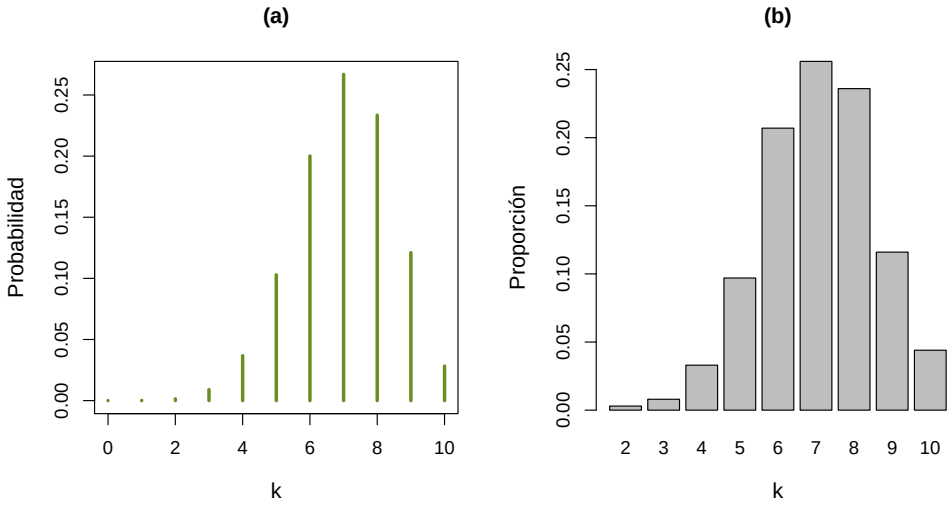


Figura 2: (a) Probabilidades correspondientes a la distribución $B\left(10,0{,}7\right)$ (b) Proporciones observadas en una muestra de tamaño $n = 1000$ de dicha distribución. Puede observarse la coincidencia entre ambas representaciones.

**Esperanza y varianza:**

$$\mu = E\left[X\right] = \sum_{k=0}^{\infty} k \cdot P\left(X = k\right) = \sum_{k=0}^{\infty} k\left(1-p\right)^{k} p = \frac{1-p}{p}$$

$$\sigma^2 = var\left(X\right) = \sum_{k=0}^{\infty} \left(k-\mu\right)^2 \cdot P\left(X = k\right) = \sum_{k=0}^{\infty} \left(k - \frac{1}{p}\right)^2 \left(1-p\right)^{k} p = \frac{1-p}{p^2}$$

**Ejemplo:** Sea $X$="*Número de lanzamientos de un dado equilibrado antes de que salga el primer 6*". Obviamente $X \approx Geo(\frac{1}{6})$. Así, por ejemplo, la probabilidad de que haya que lanzar el dado 9 veces antes del primer 6, sería:

$$P\left(X = 9\right) = \left(1 - \frac{1}{6}\right)^{9} \frac{1}{6} = 0{,}0323$$

El número esperado de veces que habría que lanzar el dado antes de que salga un 6 por primera vez sería $\mu = \frac{1-1/6}{1/6} = 5$

**Cálculo con R :** Si $X \approx Geo(p)$:

- Valor de la función de probabilidad: $P(X = k)$=`dgeom(k,p)`
- Valor de la función de distribución: $P(X \leq k)$=`pgeom(k,p)`
- Cuantiles: $q_\alpha = \text{mín}\{x : F(x) \geq \alpha\}$=`qgeom(`$\alpha$`,p)`
- Generación de $m$ números aleatorios con distribución $Geo(p)$: `rgeom(m,p)`

**Ejemplo:** Para calcular con R la probabilidad buscada en el ejemplo anterior ejecutamos:

```
> dgeom(9,1/6)

[1] 0.03230112

>
```

## 3.5. Distribución Hipergeométrica $H(n, N, N_E)$

**Definición:** Supongamos que se dispone de una población finita de tamaño $N$, que está dividida en dos grupos: $N_E$ "*éxitos*" y $N - N_E$ "*fracasos*". Una variable aleatoria $X$ sigue una distribución hipergeométrica si representa el número de éxitos obtenidos al extraer al azar y sin reemplazamiento $n$ objetos de esta población. La función de probabilidad de esta variable aleatoria es:

$$P(X = k) = \frac{\binom{N_E}{k}\binom{N-N_E}{n-k}}{\binom{N}{n}}, \quad x = \text{máx}\{0, n - (N - N_E)\}, \ldots, \text{mín}\{N_E, n\}$$

**Esperanza y varianza:** Si llamamos $p = \frac{N_E}{N}$ (probabilidad de éxito cuando se extrae un único objeto)

$$\mu = \frac{n \cdot N_E}{N} = np$$
$$\sigma_X^2 = \frac{N_E(N - N_E)\, n\,(N - n)}{N^2(N - 1)} = np(1 - p)\frac{(N - n)}{(N - 1)}$$

**Nota:** Es evidente que si el experimento donde surge la distribución hipergeométrica se realizara con reemplazamiento, la variable $X$ considerada tendría distribución binomial. Debe señalarse que, aún habiendo reemplazamiento, si $N$ es muy grande en comparación con $n$, resultaría muy difícil que un mismo objeto de la población fuera elegido aleatoriamente dos ó más veces, lo que es equivalente a que no haya reemplazamiento. Ello significa que la distribución hipergeométrica se va pareciendo cada vez más a la binomial a medida que $N$ crece con respecto a $n$. Puede observarse incluso en las expresiones de la esperanza y la varianza, que si $N$ se hace grande y $n$ es relativamente pequeño, se obtienen los mismos valores que en la binomial.

**Ejemplo:** De una urna en la que hay 10 bolas blancas y 5 bolas negras, se extraen 8 bolas sin reemplazamiento. ¿Cual es la probabilidad de que entre estas ocho haya 4 bolas negras? Si llamamos: $X =$ "*número de bolas negras en la muestra*" entonces $X \approx H\,(8, 15, 5)$ y:

$$P(X = 4) = \frac{\binom{5}{4}\binom{15-5}{8-4}}{\binom{15}{8}} = \frac{\binom{5}{4}\binom{10}{4}}{\binom{15}{8}} = 0{,}1632$$

**Cálculo con R :** la sintaxis a emplear con R para calcular probabilidades asociadas a la distribución geométrica es nuevamente similar a la ya vista en las distribuciones anteriores. Si $X \approx H\,(n, N, N_E)$ y llamamos $N_F = N - N_E$:

- Valor de la función de probabilidad: $P\,(X = k)$=`dhyper(k,NE,NF,n)`
- Valor de la función de distribución: $P\,(X \leq k)$=`phyper(k,NE,NF,n)`
- Cuantiles: $q_\alpha = \text{mín}\,\{x :\ F\,(x) \geq \alpha\}$=`qhyper(`$\alpha$`,NE,NF,n)`
- Generación de $m$ números aleatorios con esta distribución: `rhyper(m,ME,NF,n)`

Para obtener la probabilidad del ejemplo anterior utilizando R emplearíamos la función:

```
> dhyper(4,10,5,8)

[1] 0.1631702
```

**Aplicación a la estimación de un tamaño poblacional.** (**Método de captura - recaptura**) Una aplicación clásica de la distribución hipergeométrica al campo de las

ciencias biológicas es la siguiente: supongamos que se desea estimar aproximadamente el número de peces que hay en un lago. Para ello realizamos una captura inicial de $N_E$ peces (se capturan al azar, a lo largo de toda la extensión del lago), los marcamos y los devolvemos al agua. De esta forma ahora tenemos en el lago un total de $N$ peces ($N$ es desconocida) de los que $N_E$ *sabemos* que están marcados. Realizamos una segunda captura, ahora de $n$ peces y contamos cuántos hay marcados en esta recaptura. Obviamente el número de peces marcados en la recaptura sigue una distribución hipergeométrica $H(n, N, N_E)$ por lo que el número esperado de peces marcados en dicha recaptura es $n\frac{N_E}{N}$. Si en realidad se observaron $k$ peces marcados, igualamos ambas expresiones (esto es, suponemos que se captura exactamente lo que se esperaba capturar):

$$k = n\frac{N_E}{N}$$

de donde se obtiene el valor de $N$:

$$\hat{N} = n\frac{N_E}{k}$$

Obviamente este valor de $N$ es una aproximación, ya que la premisa de que lo que se esperaba pescar es lo que se pesca, no tiene que ser válida exactamente. Este es el punto de arranque para los diseños de muestreo más sofisticados que se emplean en la estimación de tamaños poblacionales.

## 3.6. Distribución de Poisson $P(\lambda)$

Las tortugas marinas suelen cavar sus nidos en la zona supramareal de playas fácilmente accesibles. Supongamos que en determinada playa se ha observado que las posiciones de los nidos se reparten completamente al azar en esa zona, con una densidad media de $\vartheta$ nidos por km$^2$. ¿Cómo podríamos calcular la probabilidad de que en una extensión de $S$ km$^2$ se encuentren $k$ nidos?

Por simplicidad supongamos que dicha región es rectangular, y que sobre la misma superponemos una malla tal como se muestra en la figura 3. La malla es lo suficientemente fina como para que en cada cuadrícula quepa como mucho un único nido. Las posiciones de los nidos se han marcado mediante puntos en el gráfico resultante. De esta forma el problema de determinar la probabilidad de que en esta zona haya $k$ nidos es equivalente a calcular la probabilidad de que $k$ cuadros de la malla estén ocupados por un nido. Si suponemos que en total la malla tiene $n$ cuadros, que la probabilidad de que un cuadro arbitrario esté ocupado

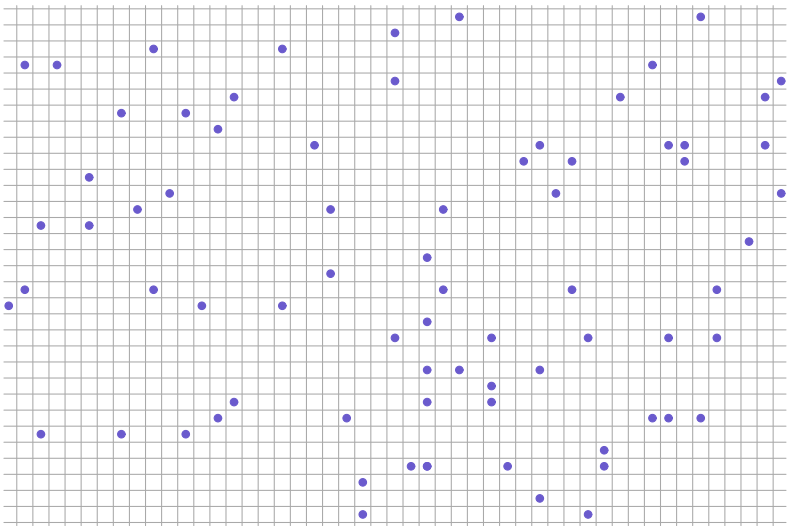

Figura 3: Región rectangular de superficie $S$ situada en la zona supramareal de una playa en la que hay nidos de tortuga. Sobre esta región se ha superpuesto una malla regular y se han marcado las posiciones de los nidos.

es $p$, y que los cuadros se ocupan independientemente unos de otros (esta última hipótesis es razonable si los nidos están repartidos completamente al azar, es decir, si no tienden a estar concentrados en un único sitio ni a estar sistemáticamente separados unos de otros), entonces la variable $X =$"*número de cuadros ocupados por nidos en la malla*" sigue una distribución binomial $B(n, p)$ donde:

- $n$ es un número muy grande (hay muchos cuadros en la malla).
- $p$ es un número muy pequeño (entre tantos cuadros, la probabilidad de que haya un nido en un cuadro concreto es minúscula).
- Como hay una densidad media de $\vartheta$ nidos por km$^2$ y la región estudiada mide $S$ km$^2$, el número esperado de nidos en la región es $\lambda = \vartheta S$. Como el valor esperado de la binomial es $n \cdot p$, debe ocurrir entonces que $n \cdot p{=}\lambda$ (de donde $p = \frac{\lambda}{n}$)

Así pues para calcular la probabilidad de $k$ nidos utilizando esta aproximación binomial

tendríamos:

$$
\begin{aligned}
P\left(X=k\right) &= \lim_{\substack{n\to\infty \\ p\to 0 \\ np=\lambda}} \binom{n}{k} p^k (1-p)^{n-k} = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} = \\
&= \lim_{n\to\infty} \frac{n(n-1)...(n-k+1)}{n^k} \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n-k} = \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} 1 \cdot \left(1-\frac{1}{n}\right) \cdot ... \cdot \left(1-\frac{k-1}{n}\right) \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} = \\
&= \frac{\lambda^k}{k!} e^{-\lambda}
\end{aligned}
$$

**Definición:** Una variable aleatoria discreta $X$ sigue una *distribución de Poisson* de parámetro $\lambda$, si su función de probabilidad es de la forma:

$$P(X=x) = \frac{\lambda^k}{k!} e^{-\lambda}; \;\; k = 0, 1, 2, 3, ....$$

siendo $\lambda$ un valor real positivo. La figura 4 muestra la forma de esta función de probabilidad para diversos valores de $\lambda$.

En el ejemplo anterior, el número de nidos de tortuga en una región de superficie $S$ sigue una distribución de Poisson de parámetro $\lambda = \vartheta S$, siendo $\vartheta$ el número medio de nidos por unidad de superficie.

En general, la distribución de Poisson constituye un modelo de probabilidad adecuado para aquellas variables aleatorias que cuentan el número de puntos que se encuentran en cierto espacio continuo, siempre y cuando estos puntos se encuentren repartidos completamente al azar. A modo de ejemplo podemos citar:

- Número de estrellas en cierta porción del firmamento (los puntos son las estrellas y el espacio continuo es la región estelar observada).
- Número de copépodos en un volumen de agua determinado (los puntos son los copépodos y el espacio continuo donde se encuentran es el volumen de agua).

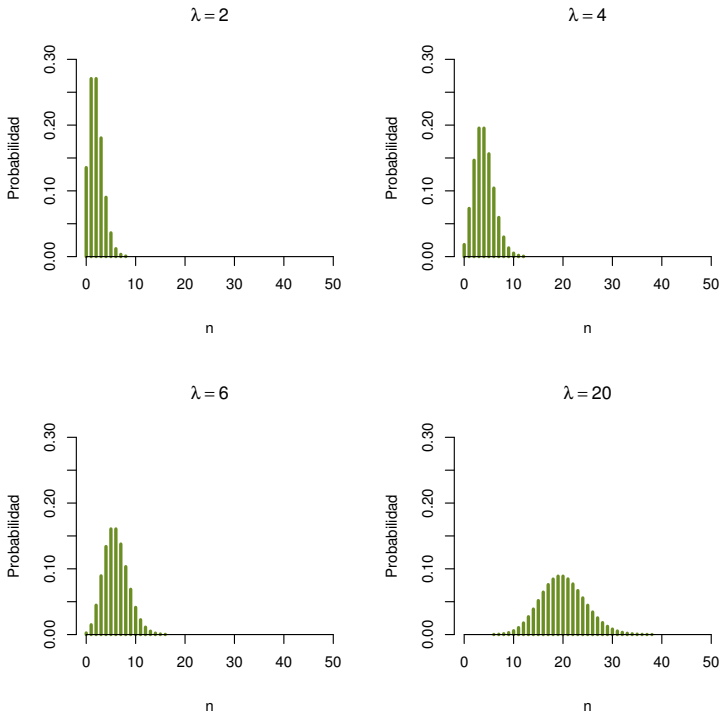


Figura 4: Función de Probabilidad de la distribución de Poisson para varios valores de $\lambda$. La altura de cada línea vertical representa la $P(X = k)$

- Número de llamadas telefónicas recibidas en una centralita a lo largo de un día (los puntos son los instantes en que se producen las llamadas, y el espacio continuo en que se sitúan estos puntos es el tiempo transcurrido entre las 0 y las 24 horas).

**Esperanza y varianza:** Puede probarse que:

$$E[X] = \sum_{k=0}^{\infty} kP(X = k) = \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} = \lambda$$

$$var(X) = E\left[X^2\right] - E[X]^2 = \sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!}e^{-\lambda} - \lambda^2 = \lambda$$

Este resultado era de esperar, ya que $X$ es el límite de una binomial cuya esperanza es $np = \lambda$ y cuya varianza es $np(1-p) = \lambda$ (ya que $np = \lambda$ y $p \to 0$, por lo que $(1-p) \cong 1$)

**Ejemplo:** Si la densidad de nidos de tortuga en una playa es de 0.01 nidos por $m^2$ (esto es, un nido cada 100 $m^2$), ¿cuál es la probabilidad de que una zona de 1000 $m^2$ de extensión haya 8 nidos?

En este ejemplo $\lambda = \vartheta S = 0{,}01 \cdot 1000 = 10$. Aplicando la distribución de Poisson:

$$P(X = 8) = \frac{10^8}{8!} e^{-10} = 0{,}113$$

**Cálculo con R :**

- Valor de la función de probabilidad: $P(X = k)$=`dpois(k,`$\lambda$`)`
- Valor de la función de distribución: $P(X \leq k)$=`ppois(k,`$\lambda$`)`
- Cuantiles: $q_\alpha = \text{mín}\{x : F(x) \geq \alpha\}$=`qpois(`$\alpha$`,`$\lambda$`)`
- Generación de $m$ números aleatorios con distribución $P(\lambda)$: `rpois(m,`$\lambda$`)`

**Continuación del ejemplo:** En el ejemplo anterior, si queremos calcular la probabilidad de que en una región de 1 km$^2$ de extensión haya más de 8 nidos:

$$P(X > 8) = 1 - P(x \leq 8) = 1 - \texttt{ppois(8,10)} = 1\text{-}0.333 = 0.667$$

La probabilidad de que en esa región haya entre 8 y 12 nidos puede hallarse como:

$$\begin{aligned} P(8 \leq X \leq 12) &= P(X \leq 12) - P(X \leq 7) = \\ &= \texttt{ppois(12,10)-ppois(7,10)} = \\ &= 0{,}792 - 0{,}22 = 0{,}572 \end{aligned}$$

**Aproximación de la distribución binomial:** Hemos obtenido la distribución de Poisson como límite de una binomial cuando $n \to \infty$ y $p \to 0$. La distribución de Poisson constituye en general una buena aproximación de la binomial $B(n, p)$ cuando $n > 20$ y $p < 0{,}05$, en cuyo caso $B(n, p) \cong P(\lambda)$, con $\lambda = n \cdot p$.

Para entender el sentido de esta aproximación consideremos el siguiente ejemplo: se sabe que el 1 % de los huevos de tortuga depositados en una playa son depredados por cangrejos. Si entre cuatro nidos totalizan 280 huevos, ¿cuál es la probabilidad de que ninguno sea depredado por cangrejos?.

Llamando $X$="Número de huevos depredados en los cuatro nidos", tendríamos que $X \approx B\,(280, 0{,}01)$. La probabilidad de que ningún huevo sea depredado sería:

$$P\,(X = 0) = (1 - 0{,}01)^{280} = 0{,}99^{280} = 0{,}05996$$

Muchas calculadoras no son capaces de realizar este cálculo (aquí lo hemos obtenido con R mediante `dbinom(0,280,0.01)`). La aproximación de Poisson nos indica que $X \approx B\,(280, 0{,}01) \cong P\,(280 \cdot 0{,}01) = P\,(2{,}8)$. Si utilizamos la distribución de Poisson para calcular la probabilidad pedida obtenemos

$$P\,(X = 0) = \frac{2{,}8^0}{0!} e^{-2{,}8} = e^{-2{,}8} = 0{,}06081$$

que se diferencia del verdadero valor en 0,00085, por lo que el error de aproximación es inferior a una milésima. Vemos, pues, que la aproximación mediante la distribución de Poisson funciona razonablemente bien, y es aconsejable su uso cuando no se dispone de medios informáticos avanzados.

**Aditividad de la distribución de Poisson.** Si dos variables aleatorias independientes $X_1$ y $X_2$ siguen sendas distribuciones de Poisson, $X_1 \approx P(\lambda_1)$ y $X_2 \approx P(\lambda_2)$, entonces $X_1 + X_2 \approx P(\lambda_1 + \lambda_2)$. En general, si $X_1, X_2, ..., X_n \approx P(\lambda)$ , y además son independientes, entonces $\sum_{i=1}^{n} X_i \approx P(n\lambda)$

# 4. Principales distribuciones de probabilidad continuas.

## 4.1. Distribución uniforme $U\,(a, b)$.

**Definición:** Una variable aleatoria $X$ sigue una *distribución uniforme* en el intervalo real $(a, b)$, si su función de densidad es constante sobre ese intervalo:

$$f\,(x) = \begin{cases} \frac{1}{b-a} & x \in (a, b) \\ 0 & x \notin (a, b) \end{cases}$$

En la práctica esta distribución corresponde a variables del tipo: $X$ = *"Resultado de elegir al azar un valor del intervalo (a,b)"* cuando la probabilidad de que el valor elegido caiga en un intervalo de amplitud $\ell$ dentro de $(a, b)$ es siempre la misma independientemente de la posición de dicho intervalo.

**Esperanza y varianza:**

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_a^b x \frac{1}{b-a} dx = \left[\frac{1}{b-a}\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{a+b}{2}$$

$$var(X) = E\left[X^2\right] - E[X]^2 = \int_a^b x^2 \frac{1}{b-a} dx - \left(\frac{a+b}{2}\right)^2 = \frac{1}{12}(b-a)^2$$

**Ejemplo:** la variable aleatoria $X =$ "*Distancia, medida desde el extremo inicial, a la que se rompe una cuerda homogénea de 1 metro cuando se tira con igual fuerza de ambos extremos*" que ya hemos visto en el capítulo anterior sigue una distribución $X \approx U(0,1)$.

**Cálculo con R :**

- Valor de la función de densidad $f(x)$=`dunif(x,a,b)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`punif(x,a,b)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qunif(`$\alpha$`,a,b)`
- Generación de $n$ números aleatorios con distribución $U(a,b)$: `runif(n,a,b)`

## 4.2. Distribución exponencial $\exp(\eta)$.

**Definición:** una variable aleatoria $X$ sigue una *distribución exponencial* de parámetro $\eta$ si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{1}{\eta} e^{-\frac{1}{\eta}x},\ x \geq 0$$

En la práctica, esta distribución aparece asociada a variables que miden la distancia entre sucesos puntuales que se dispersan completamente al azar en un medio continuo y cuyo número tiene, por tanto, distribución de Poisson. En efecto, supongamos por simplicidad que el medio continuo considerado es el tiempo y que estamos contando el número de eventos que ocurren hasta un instante $t$. Si el número de tales eventos sigue una distribución de Poisson, siendo $\lambda$ el número esperado de eventos por unidad de tiempo, ello significa que $\eta = \frac{1}{\lambda}$ es el tiempo esperado entre dos cualesquiera de tales

sucesos. Si llamamos $Y_t$="Número de sucesos ocurridos en un intervalo de duración $t$" entonces $Y_t \approx P(\lambda t) = P\left(\frac{1}{\eta}t\right)$. Si acaba de ocurrir uno de estos sucesos, y llamamos $X$ al tiempo que transcurre hasta que ocurre el siguiente, entonces:

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

(ya que $X \geq t$ significa que el siguiente suceso ocurre después de $t$, o lo que es lo mismo, que en un intervalo de duración $t$ no ha ocurrido ningún suceso, esto es $Y_t = 0$). Por tanto:

$$F(t) = P(X \leq t) = 1 - e^{-\frac{1}{\eta}t}$$

de donde:

$$f(t) = F'(t) = \frac{1}{\eta}e^{-\frac{1}{\eta}t},\ t \geq 0$$

La figura 5 muestra la forma de la distribución exponencial para varios valores del parámetro $\eta$.

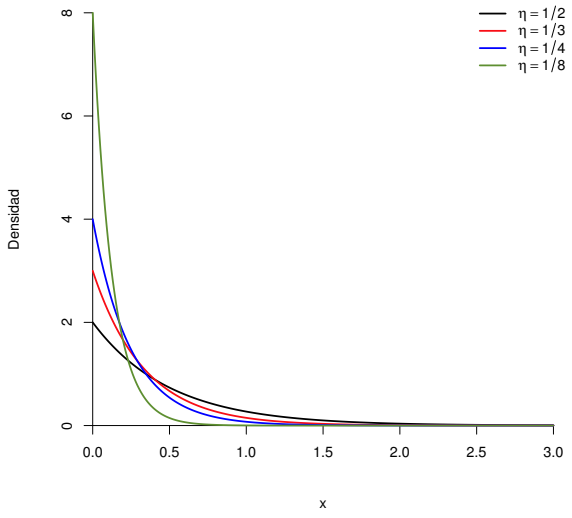


Figura 5: Función de densidad de la distribución exponencial para varios valores de $\eta$.

**Esperanza y varianza:**

$$E[X] = \int_0^\infty \frac{1}{\eta}xe^{-\frac{1}{\eta}x}dx = \eta$$

$$var(X) = E\left[X^2\right] - E[X]^2 = \int_0^\infty \frac{1}{\eta}x^2e^{-\frac{1}{\eta}x}dx - \left(\frac{1}{\eta}\right)^2 = \eta^2$$

**Ejemplo:** El tiempo que transcurre entre la caída de dos rayos sucesivos durante la fase central de una tormenta tropical sigue una distribución exponencial de parámetro 2.5 segundos. ¿Cuál es la probabilidad de que entre la caída de dos rayos sucesivos transcurran como mucho 3 segundos? ¿Cuál es el tiempo esperado que transcurre entre rayos sucesivos?

Sea X = "*Tiempo transcurrido entre dos rayos sucesivos*"$\approx exp(2{,}5)$. La probabilidad pedida es entonces:

$$P(X \leq 3) = 1 - e^{-\frac{1}{2,5}\cdot 3} = 1 - e^{-1,2} = 0{,}699$$

Dado que en una distribución exponencial el valor esperado coincide con su parámetro, el tiempo esperado entre rayos sucesivos es $E[X] = \eta = 2{,}5$ segundos.

**Cálculo con R :** Nótese que por defecto R espera recibir como parámetro el valor $1/\eta$ que recibe el nombre de *rate* (tasa).

- Valor de la función de densidad*:* $f(x)$=`dexp(x,`$1/\eta$`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pexp(x,`$1/\eta$`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qexp(`$\alpha$`,`$1/\eta$`)`
- Generación de $n$ números aleatorios con distribución $exp(\lambda)$: `rexp(n,`$1/\eta$`)`

Así, el cálculo de la probabilidad del ejemplo anterior en R sería:

$$P(X \leq 3) = \texttt{pexp(3,1/2.5)} = 0{,}699$$

**Falta de memoria de la distribución exponencial.** La distribución exponencial tiene una propiedad característica que suele denominarse "*falta de memoria*". Si $X$ es el tiempo entre dos ocurrencias consecutivas de un fenómeno, la *falta de memoria* significa que:

$$P(X \geq t + s \,|\, X \geq s) = P(X \geq t)$$

es decir, si desde la ocurrencia anterior ha transcurrido ya un tiempo $s$, la probabilidad de que aún falte un tiempo adicional $t$ hasta la próxima ocurrencia es independiente de $s$. Para entender este enunciado pensemos los siguientes ejemplos:

- Nos encontramos en una estación de metro esperando por el siguiente tren; la linea que esperamos es muy puntual y por término medio pasa un tren cada 10 minutos. Si el último tren pasó hace 9 minutos, podemos estar razonablemente seguros de que el tiempo que aún nos queda por esperar es del orden de 1 minuto. Podemos decir que el tiempo entre llegadas de trenes "*tiene memoria*": el tiempo transcurrido desde la última llegada nos informa sobre el tiempo que aún falta hasta la siguiente.
- En nuestra ciudad cae un premio grande de la lotería por término medio una vez cada 10 años. Si el último de estos premios cayó hace 9 años, eso no nos dice nada sobre cuantos años han de transcurrir aún hasta que vuelva a tocar un premio grande en la ciudad. El tiempo entre premios de la lotería *"no tiene memoria*": el tiempo transcurrido desde el último premio no da ninguna información sobre el tiempo que aún falta hasta el siguiente.

Es fácil comprobar la falta de memoria de la distribución exponencial:

$$P\left(X \geq t+s \,|\, X \geq s\right) = \frac{P\left(\{X \geq t+s\} \cap \{X \geq s\}\right)}{p\left(X \geq s\right)} = \\ = \frac{P\left(X \geq t+s\right)}{p\left(X \geq s\right)} = \frac{e^{-\frac{1}{\eta}(t+s)}}{e^{-\frac{1}{\eta}s}} = e^{-\frac{1}{\eta}t} = P\left(X \geq t\right)$$

Esta propiedad resulta útil para decidir si la distribución exponencial puede ser un buen modelo para el comportamiento de una variable de nuestro interés: podría serlo para el tiempo transcurrido entre premios de la lotería, pero desde luego no lo es para el tiempo entre trenes de una linea de metro.

## 4.3. Distribución de Weibull $W\left(\kappa, \eta\right)$.

**Definición:** Una variable aleatoria $X$ sigue una *distribución de Weibull* con parámetro de forma $\kappa$ y parámetro de escala $\eta$ si su función de distribución es de la forma:

$$F\left(x\right) = 1 - \exp\left(-\left(x/\eta\right)^{\kappa}\right), \; x \geq 0$$

Su función de densidad es:

$$f\left(x\right) = \frac{\kappa}{\eta}\left(\frac{x}{\eta}\right)^{\kappa-1} \exp\left(-\left(x/\eta\right)^{\kappa}\right), \; x \geq 0$$

En el caso particular de que $\kappa = 1$, la distribución de Weibull coincide con una exponencial de parámetro $\eta$.

La distribución de Weibull se utiliza con frecuencia para modelar el tiempo (aleatorio) que transcurre entre dos sucesos de interés, en particular cuando el tiempo transcurrido "*tiene memoria*" en el sentido apuntado más arriba. Así, por ejemplo, suele utilizarse:

- Para modelar la supervivencia: tiempo que sobreviven los enfermos con determinado tratamiento; tiempo que sobreviven las células en un cultivo; tiempo que dura un fenómeno meteorológico.
- Para modelar la fiabilidad: tiempo que dura un componente electrónico, mecánico, etc. en función de su edad y condiciones de uso.
- Para modelar tiempo entre eventos climatológicos: tiempo entre tormentas o ciclones, tiempo entre periodos fríos o cálidos.
- Para modelar tiempo entre determinados fenómenos geofísicos: tiempo entre réplicas de un terremoto, tiempo entre erupciones volcánicas.

Otras aplicaciones de la distribución de Weibull, dado el perfil de su función de densidad, son el modelado de la altura de ola, la velocidad de corriente marina o la velocidad del viento.

La figura 6 muestra la forma de la función de densidad de la distribución de Weibull para varios valores de $\kappa$ y $\eta$.

**Esperanza y varianza:**

$$\mu = E[X] = \int_0^\infty x f(x)\, dx = \eta \cdot \Gamma\left(1 + \frac{1}{\kappa}\right)$$
$$\sigma^2 = var(X) = \eta^2 \left[\Gamma\left(1 + \frac{2}{\kappa}\right) - \left(\Gamma\left(1 + \frac{1}{\kappa}\right)\right)^2\right]$$

siendo $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ la función gamma de Euler, que cumple las siguientes propiedades, útiles para el cálculo de sus valores:

1. $\Gamma(a) = (a-1)\Gamma(a-1)$
2. Si $n \in \mathbb{N}: \ \Gamma(n) = (n-1)!$

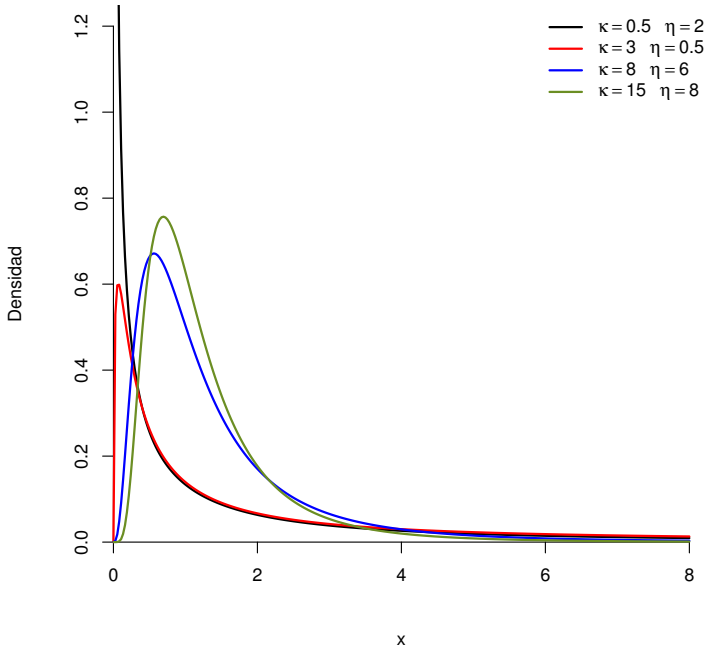


Figura 6: Función de densidad de la distribución de Weibull para varios valores de los parámetros $\kappa$ y $\eta$.

La función gamma de Euler se encuentra implementada en R : $\Gamma(a)$ =`gamma(a)`

**Cálculo con R :**

- Valor de la función de densidad: $f(x)$=`dweibull(x,`$\kappa$`,`$\eta$`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pweibull(x,`$\kappa$`,`$\eta$`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qweibull(`$\alpha$`,`$\kappa$`,`$\eta$`)`
- Generación de $n$ números aleatorios con distribución $exp(\lambda)$: `rweibull(n,`$\kappa$`,`$\eta$`)`

## 4.4. Distribución Gamma $\mathcal{G}(\kappa, \eta)$

**Definición:** Una variable aleatoria $X$ sigue una *distribución gamma* con parámetro de forma $\kappa$ y parámetro de escala $\eta$ si su función de densidad es de la forma:

$$f(x) = \frac{1}{\eta^\kappa \Gamma(\kappa)} x^{\kappa-1} \exp(-x/\eta) \ : \ x \geq 0$$

siendo $\Gamma(a)$ la función gamma de Euler. En el caso particular de que $\kappa = 1$, la distribución gamma se reduce a una exponencial de parámetro $\eta$.

En la práctica la distribución gamma suele utilizarse para modelar problemas como los ya descritos para la distribución de Weibull. La figura muestra la forma de la función de densidad de la distribución gamma para varios valores de sus parámetros.

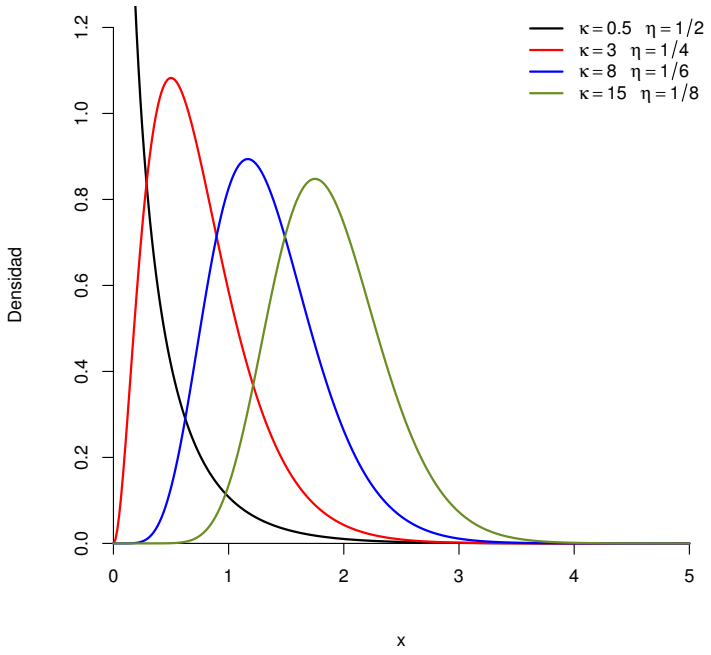


Figura 7: Función de densidad de la distribución Gamma para varios valores de $\kappa$ y $\eta$.

**Esperanza y varianza:**

$$\mu = E[X] = \kappa \cdot \eta$$
$$\sigma^2 = var(X) = \kappa \cdot \eta^2$$

**Cálculo con R :** la notación es similar a las distribuciones anteriores. Nótese que por defecto R espera recibir como parámetro el inverso del factor de escala $1/\eta$ que recibe el nombre de *rate* (tasa).

- Valor de la función de densidad*:* $f(x)$=`dgamma(x,`$\kappa$`,`$1/\eta$`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pgamma(x,`$\kappa$`,`$1/\eta$`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qgamma(`$\alpha$`,`$\kappa$`,`$1/\eta$`)`
- Generación de $n$ números aleatorios con distribución $exp(\lambda)$: `rgamma(n,`$\kappa$`,`$1/\eta$`)`

La siguiente proposición resulta de interés en las aplicaciones:

**Proposición.** *Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes y con distribución exponencial de parámetro $\eta$. Entonces $\sum_{i=1}^{n} X_i$ sigue una distribución gamma $\mathcal{G}(n, \eta)$.*

## 4.5. Distribución Normal $N(\mu, \sigma)$

**Definición:** Una variable aleatoria $X$ sigue una *distribución Normal* de parámetros $\mu$ (media) y $\sigma$ (desviación típica) si su función de densidad de probabilidad es de la forma:

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

Nótese que $f(x)$ es una función simétrica respecto a $x$, esto es $f(x) = f(-x)$. La figura 8 muestra la forma de esta función de densidad, que corresponde a la conocida *campana de Gauss.*

En la práctica, 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 que los valores cada vez más alejados de la media, bien sea hacia arriba o hacia abajo, van siendo progresivamente más difíciles de observar. Muchas variables biológicas se comportan aproximadamente de esta forma: la talla, el peso, la temperatura corporal, etc. También se comportan de esta manera los errores de medida. La distribución normal es una de las más frecuentes en la naturaleza, lo que se justifica de manera teorica por la acción del teorema central del límite, que veremos más adelante. Dicho de una manera intuitiva, este teorema indica que si una variable es el resultado de la suma de efectos de muchas otras variables independientes, la variable resultante tiene necesariamente distribución normal. Si se piensa que las variables que hemos citado –peso,

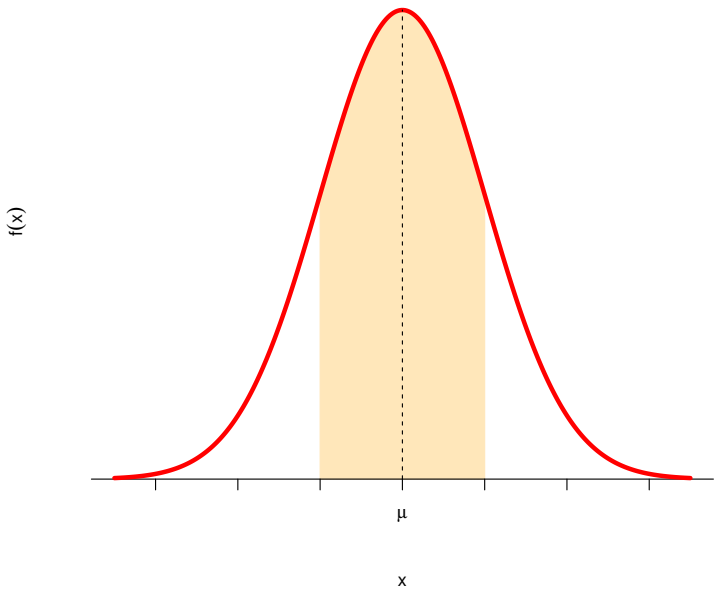


Figura 8: Función de densidad de la distribución normal. Está centrada en la media ($\mu$), valor en torno al cual se concentra la mayor parte de la probabilidad.

talla, errores de medida, ...– son precisamente el efecto de muchas pequeñas causas que operan independientemente se entiende por qué cabe esperar que su distribución de probabilidad sea precisamente normal.

La figura 9 muestra la forma de la función de densidad de la distribución normal con media $\mu = 0$ para varios valores de $\sigma$.

**Esperanza y varianza:** hemos definido la distribución normal precisamente a partir de sus esperanza y varianza. No obstante se puede comprobar resolviendo las integrales correspondientes, que tal como se ha definido la función de densidad $f(x)$ se verifica que:

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

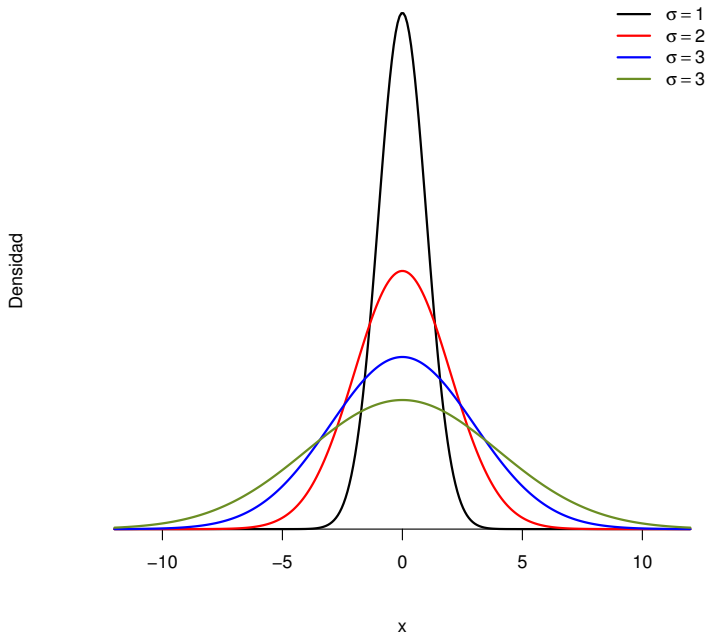


Figura 9: Función de densidad de la distribución normal de media $\mu = 0$ para varios valores de $\sigma$.

**Distribución normal tipificada:** El caso particular de la distribución normal con $\mu = 0$ y $\sigma = 1$ se conoce con el nombre de *distribución normal tipificada o estándar* $N(0,1)$. Si $Z \approx N(0,1)$ *denotaremos como* $\Phi(z) = P(Z \leq z)$.

Una de las dificultades prácticas que presenta la distribución normal es que su función de densidad no tiene una función primitiva, lo que significa que las probabilidades

$$P(X \leq x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2} du$$

deben calcularse numéricamente. Si bien R calcula esta probabilidad mediante la función `pnorm(x,`$\mu$`,`$\sigma$`)` (y existen muchos otros programas que lo hacen, así como la mayoría de las calculadoras científicas), es usual calcularla mediante el uso de tablas. El interés de la distribución normal tipificada es que es la única cuyas probabilidades se encuentran tabuladas.

**Uso de la tabla de la distribución normal tipificada.** Esta tabla sólo proporciona probabilidades de la forma $P(Z \geq z)$, siendo $Z \approx N(0,1)$, correspondientes al área sombreada en la figura 10. Para aprender a manejar esta tabla, supongamos que queremos

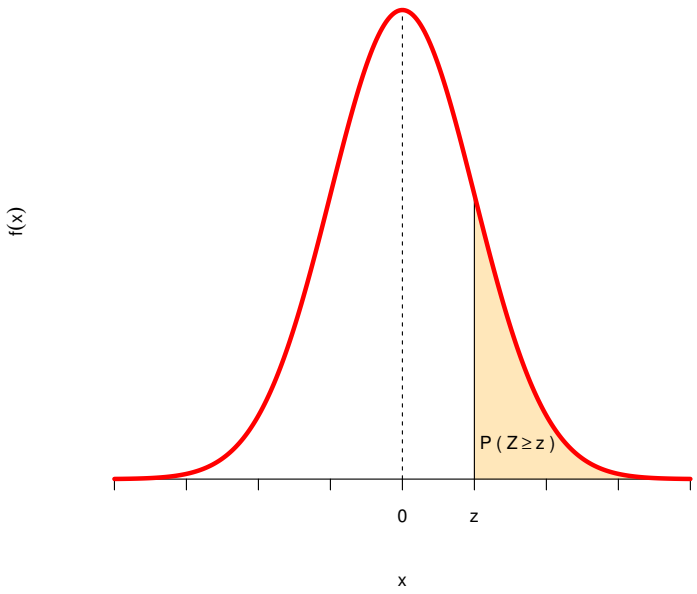


Figura 10: La tabla de la distribución $N(0,1)$ proporciona, para diversos valores de $z$, el valor de $P(Z \geq z)$, correspondiente al área sombreada.

calcular la probabilidad $P(Z \geq 2{,}16)$. Para ello simplemente separamos el número 2,16 en dos partes: una con la parte entera y las décimas (2,1), y otra con las centésimas (0,06). A continuación vamos a la tabla y buscamos el punto de cruce de la fila etiquetada como 2,1 y la columna etiquetada como 0,06, donde encontramos el valor 0,01539, que corresponde a la probabilidad buscada.

Si queremos calcular probabilidades de la forma $P(Z \leq z)$ simplemente utilizamos que $P(Z \leq z) = 1 - P(Z \geq z)$ y procedemos igual que antes. Si queremos calcular probabilidades para valores negativos de la variable basta tener en cuenta que la distribución normal es simétrica y por tanto que $P(Z \leq -z) = P(Z \geq z)$. Por último la tabla nos indica que si $z \geq 4$ entonces $P(Z \geq z) \cong 0$.

¿Cómo podemos utilizar esta tabla si queremos calcular probabilidades de una $N(\mu, \sigma)$ con $\mu \neq 0$ y $\sigma \neq 1$? En tal caso aplicaríamos el siguiente resultado:

***Proposición:*** *Si* $X \approx N(\mu, \sigma)$ *entonces* $Z = \frac{X-\mu}{\sigma} \approx N(0,1)$

El significado de esta proposición es fácil de entender: los valores de $Z$ se obtienen a partir de los de $X$ por *desplazamiento* (al restar $\mu$) y *cambio de escala* (al dividir por $\sigma$). Ninguna de estas transformaciones cambia la *forma* de la función

de densidad; por tanto $Z$ también debe seguir una distribución normal. Asimismo, la simple aplicación de las propiedades de la media y la varianza permite ver de inmediato que $E[Z] = \frac{1}{\sigma}E[X-\mu] = \frac{1}{\sigma}(E[X]-\mu) = \frac{1}{\sigma}(\mu-\mu) = 0$ y $var(Z) = \frac{1}{\sigma^2}var(X-\mu) = \frac{1}{\sigma^2}var(X) = \frac{1}{\sigma^2}\sigma^2 = 1$.

Para calcular entonces probabilidades de la forma $P(X \geq x)$ cuando $X \approx N(\mu, \sigma)$ con $\mu \neq 0$ y $\sigma \neq 1$ bastará con tener en cuenta que

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

y localizar el último valor directamente en la tabla. Así, por ejemplo, si $X \approx N(20, 4)$, para calcular $P(X \geq 25)$ procederíamos del siguiente modo:

$$P(X \geq 25) = P\left(\frac{X-20}{4} \geq \frac{25-20}{4}\right) = P\left(Z \geq \frac{5}{4}\right) = P(Z \geq 1{,}25) = 0{,}10565$$

donde hemos encontrado el valor 0,10565 en el cruce de la fila 1,2 con la columna 0,05 de la distribución normal estándar.

**Cuantiles de la $N(0,1)$ utilizando la tabla.** Un problema frecuente en la práctica es la determinación de cuantiles de la distribución $N(0,1)$. Recordemos que el cuantil $\alpha$ de una variable aleatoria $X$ es el valor $q_\alpha$ tal que $P(X \leq q_\alpha) = \alpha$. En el caso de la distribución normal estándar llamaremos $z_\alpha$ al cuantil $q_{1-\alpha}$; esto es, $z_\alpha$ es el valor tal que $P(Z \leq z_\alpha) = 1-\alpha$, o lo que es lo mismo, $P(Z > z_\alpha) = \alpha$.

Para calcular los cuantiles utilizando la tabla habremos de proceder a la inversa que para el cálculo de probabilidades; por ejemplo, supongamos que deseamos localizar el valor $z_{0,025}$ (es decir, el cuantil 0,975). Buscamos el valor 0,025 (o el que más se le aproxime) en el interior de la tabla; en este caso encontramos el 0,025 en el cruce de la fila 1,9 con la columna 0,06. Por tanto $z_{0,025} = 1{,}96$.

**Cálculo con R :**

- Valor de la función de densidad: $f(x)$=`dnorm(x,`$\mu$`,`$\sigma$`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pnorm(x,`$\mu$`,`$\sigma$`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qnorm(`$\alpha$`,`$\mu$`,`$\sigma$`)`
- Generación de $n$ números aleatorios con distribución $N(\mu, \sigma)$: `rnorm(n,`$\mu$`,`$\sigma$`)`

Podemos utilizar R para calcular las probabilidades que hemos visto en los ejemplos anteriores. En el caso particular de la normal estándar no es preciso especificar $\mu = 0$ y $\sigma = 1$. Así:

- $P(Z \geq 2{,}16)$=$1 - P(Z \leq 2{,}16)$ =`1-pnorm(2.16)`=0.01539
- si $X \approx N(20, 4)$, entonces $P(X \geq 25)$=`1-pnorm(25,20,4)`=0.10565

Asimismo, el cálculo de los cuantiles es muy simple con R :

- $z_{0{,}025} = q_{1-0{,}025} = q_{0{,}975}$ =`qnorm(0.975)`=1.96

Por último presentamos una importante propiedad de la distribución normal, que nos indica que la suma de variables normales sigue también una distribución normal. Esta propiedad tiene gran aplicación práctica, ya que muchas veces habrán de calcularse probabilidades de sumas de variables normales: peso total de los ejemplares de una muestra, ingresos totales de las sucursales de una empresa durante un día laboral, distancia total recorrida por un animal durante una migración,...

**Propiedad reproductiva de la distribución normal:** d*adas n variables aleatorias normales e independientes, tales que* $X_i \approx N(\mu_i, \sigma_i)$, $i = 1, \ldots, n$, *su suma* $\sum_{i=1}^{n} X_i$ *sigue también una distribución normal, siendo:*

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

Como consecuencia de esta propiedad, en el caso particular de que $X_i \approx N(\mu, \sigma)$ para $i = 1, \ldots, n$, 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,

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

## 4.6. Distribuciones de probabilidad asociadas al muestreo de variables con distribución normal.

En muchas ocasiones nos encontramos con problemas que se refieren a características globales de una variable evaluadas sobre una o varias poblaciones. Por ejemplo ¿la concentración media de cierto contaminante en una zona supera el umbral permitido por la legislación? ¿Es la velocidad media de desplazamiento en los individuos de una especie de delfín superior a la velocidad media en otra especie? ¿Se consigue mayor peso medio en los peces de una piscifactoría cuando se usa una dieta rica en hidratos de carbono o cuando se usa una rica en proteínas? ¿Se observa mayor variabilidad de talla en los machos o en las hembras de una especie? En estos ejemplos la pregunta a responder tiene que ver con los valores medios o las varianzas de estas variables en las poblaciones de interés. Ahora bien, en la práctica estos valores no se conocen, ya que no es posible acceder a todos los sujetos de la población.

Como veremos en el próximo capítulo, la única manera de responder a estas cuestiones consiste en adquirir información sobre las cantidades de interés a partir de una muestra aleatoria. Esto nos conduce a la siguiente cuestión: el valor medio de una variable en una población es único, pero como de una misma población es posible extraer muchas muestras distintas, habrá tantas medias muestrales como muestras sea posible extraer. Lo mismo puede decirse de la varianza. Si el problema es comparar dos poblaciones, pueden extraerse muchas muestras distintas de cada una y por tanto son posibles muchos valores distintos de la diferencia

entre las medias muestrales. Como *a priori,* antes de obtener la muestra (o muestras) es imposible predecir cuáles van a ser los valores resultantes de la media, la varianza o la diferencia de medias, en su caso, resulta que estas cantidades son *variables aleatorias.* Y si son variables aleatorias, debemos preguntarnos cuál es su distribución de probabilidad, ya que es precisamente mediante el uso de dicha distribución que podremos contestar a las preguntas planteadas más arriba.

En el caso particular de que la distribución de probabilidad de la variable de interés sea normal $N(\mu, \sigma)$, se conocen las distribuciones de probabilidad de algunas de las variables aleatorias que se presentan en el muestreo. Describimos a continuación dichas distribuciones y posponemos al próximo capítulo su aplicación concreta en los problemas de inferencia ligados al muestreo.

### 4.6.1. Distribución Chi-cuadrado $\chi^2_n$

**Definición:** Una variable aleatoria $X$ sigue una *distribución Chi-Cuadrado de Pearson* con $n$ grados de libertad ($\chi^2_n$) si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}},\ x \geq 0$$

Esta distribución es un caso particular de la gamma, concretamente la $\mathcal{G}\left(\frac{n}{2}, 2\right)$. La importancia práctica de esta distribución deriva de la siguiente propiedad, que constituye el fundamento de la inferencia sobre la varianza en variables con distribución normal.

**Proposición:** *Si* $Z_1, \ldots, Z_n$ *son* $n$ *variables aleatorias independientes con distribución* $N(0,1)$*, entonces*

$$X = Z_1^2 + \cdots + Z_n^2$$

*sigue una distribución* $\chi^2_n$*.*

**Esperanza y varianza:** si $X \approx \chi^2_n$:

$$\mu = E[X] = n$$
$$\sigma^2 = var(X) = 2n$$

La figura 11 muestra la densidad de la $\chi^2_n$ para varios valores de $n$.

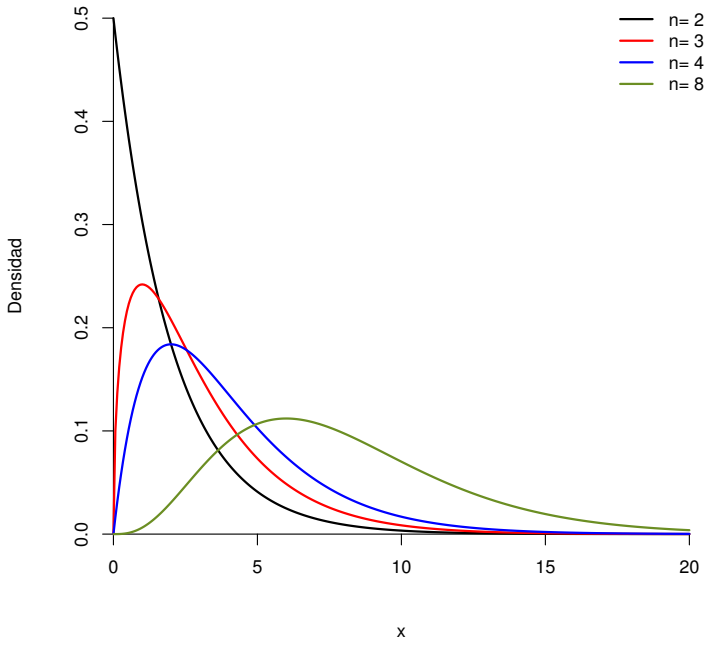


Figura 11: Función de densidad de la distribución $\chi^2_n$ para varios valores de $n$

**Cálculo con R :**

- Valor de la función de densidad*:* $f(x)$=`dchisq(x,n)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pchisq(x,n)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qchisq(`$\alpha$`,n)`
- Generación de $m$ números aleatorios con distribución $\chi^2_n$: `rchisq(m,n)`

### 4.6.2. Distribución *t* de Student $t_n$

**Definición:** Una variable aleatoria $X$ sigue una *distribución t de Student* con $n$ grados de libertad $(t_n)$ si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \; x \geq 0$$

Por ser una función cuadrática en $x$, la densidad de la $t$ de Student, al igual que ocurría con la normal, es simétrica respecto al eje de ordenadas, esto es, $f(x) = f(-x)$. En la figura 12 se muestra la forma de esta densidad para varios valores de $n$. Puede apreciarse la similitud de esta densidad con la normal. De hecho, para valores grandes de $n$ ambas funciones son prácticamente indistinguibles.

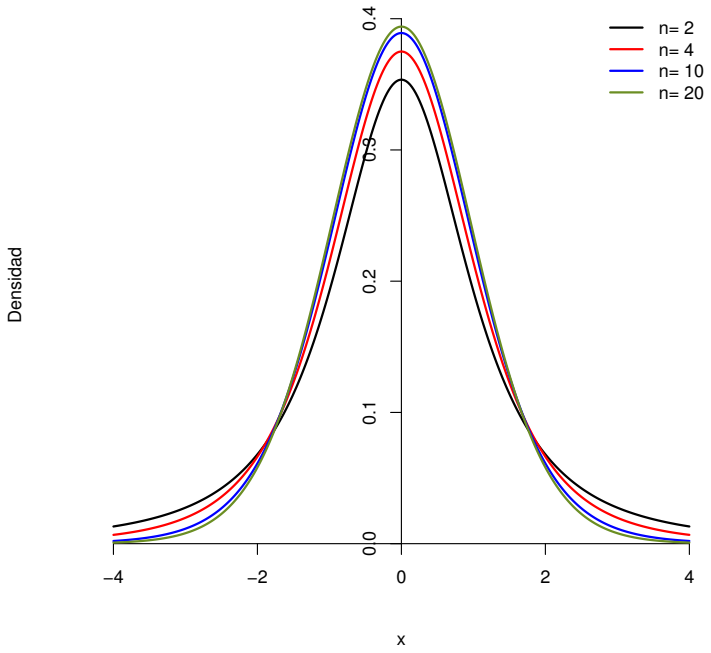


Figura 12: Función de densidad de la distribución $t$ de Student para varios valores de $n$.

El interés práctico de la distribución $t$ de Student deriva de la siguiente propiedad, que constituye el fundamento de la inferencia sobre la media en variables con distribución normal de varianza desconocida.

**Proposición:** *Sean* $Z \approx N(0,1)$ *e* $Y \approx \chi^2_n$ *dos variables aleatorias independientes. Entonces*:

$$T = \frac{Z}{\sqrt{Y/n}}$$

*sigue una distribución t de Student con n grados de libertad.*

**Esperanza y varianza:** Si $X \approx t_n$:

$$\mu = E[X] = 0 \quad (\text{Si } n > 1)$$
$$\sigma^2 = var(X) = \begin{cases} \infty & 1 < n \leq 2 \\ \frac{n}{n-2} & n > 2 \end{cases}$$

Para $n = 1$ no están definidas la media ni la varianza.

**Cálculo con R :**

- Valor de la función de densidad: $f(x)$=`dt(x,n)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pt(x,n)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qt(`$\alpha$`,n)`
- Generación de $m$ números aleatorios con distribución $t_n$: `rt(m,n)`

### 4.6.3. Distribución *F* de Fisher-Snedecor $F_{n_1,n_2}$.

**Definición:** Una variable aleatoria $X$ sigue una *distribución F de Fisher-Snedecor* con $n_1$ y $n_2$ grados de libertad ($F_{n_1,n_2}$) si su función de densidad de probabilidad es de la forma:

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{n_1/2} \frac{x^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}x\right)^{\frac{n_1+n_2}{2}}}, \; x \geq 0$$

En realidad, conocer la expresión de la función de densidad de la distribución $F$ de Fisher (al igual que la de la normal, la chi-cuadrado o la $t$ de Student) no nos sirve para calcular probabilidades directamente, ya que no admite primitiva, por lo deberán utilizarse métodos numéricos o tablas. El interés de esta distribución reside en su aplicación en la inferencia relacionada con la comparación de varianzas de variables con distribución normal, cuyo fundamento se encuentra en la siguiente propiedad.

**Proposición:** *Sean* $Y_1 \approx \chi^2_{n_1}$ *e* $Y_2 \approx \chi^2_{n_2}$ *dos variables aleatorias independientes. Entonces:*

$$X = \frac{Y_1/n_1}{Y_2/n_2}$$

*sigue una distribución de probabilidad* $F$ *de Fisher-Snedecor con* $n_1$ *y* $n_2$ *grados de libertad.*

De aquí se sigue también la siguiente propiedad de la distribución $F$:

$$X \approx F_{m,n} \Rightarrow \frac{1}{X} \approx F_{n,m}$$

**Esperanza y varianza:** Si $X \approx F_{n_1,n_2}$:

$$\mu = E[X] = \frac{n_2}{n_2 - 2}, \quad (\text{si } n_2 > 2)$$
$$\sigma^2 = var(X) = 2\left(\frac{n_2}{n_2 - 2}\right)^2 \frac{n_1 + n_2 - 2}{n_1(n_2 - 4)}, \quad (\text{Si } n_2 > 4)$$

La figura 13 muestra la forma de la función de densidad de la distribución $F$ para varios valores de $n_1$ y $n_2$.

**Cálculo con R :**

- Valor de la función de densidad: $f(x)$=`df(x,`$n_1$`,`$n_2$`)`
- Valor de la función de distribución: $F(x) = P(X \leq x)$=`pf(x,`$n_1$`,`$n_2$`)`
- Cuantil $q_\alpha = \{x : F(x) = \alpha\}$=`qf(`$\alpha$`,`$n_1$`,`$n_2$`)`
- Generación de $m$ números aleatorios con distribución $F_{n_1,n_2}$: `rf(m,`$n_1$`,`$n_2$`)`

## 4.7. Utilización de las tablas de la Chi-Cuadrado, $t$ de Student y $F$ de Fisher-Snedecor.

Como ya hemos señalado para el caso de la distribución normal, un problema que se presenta con frecuencia en la práctica es el cálculo de cuantiles de estas distribuciones. Para ello se

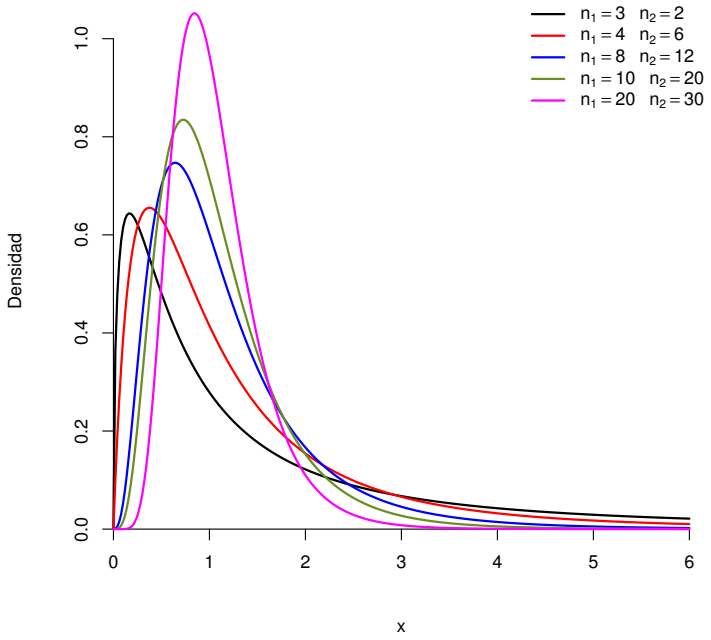


Figura 13: Función de densidad de la distribución $F$ para varios valores de $n_1$ y $n_2$.

dispone de tablas de fácil manejo, pero que no incluyen todos los posibles grados de libertad asociados a estas distribuciones (en algunos casos es preciso interpolar). Resulta recomendable en este caso utilizar R u otro software para el cálculo de estos cuantiles (algunas calculadoras lo implementan).

Llamaremos $\chi^2_{n,\alpha}$, $t_{n,\alpha}$ y $F_{n_1,n_2,\alpha}$ a los cuantiles $q_{1-\alpha}$ de las respectivas distribuciones con sus grados de libertad correspondientes. De esta forma:

- Si $X \approx \chi^2_n$, entonces $P\left(X \geq \chi^2_{n,\alpha}\right)=\alpha$
- Si $X \approx t_n$, entonces $P\left(X \geq t_{n,\alpha}\right)=\alpha$
- Si $X \approx F_{n_1,n_2}$ entonces $P\left(X \geq F_{n_1,n_2,\alpha}\right)=\alpha$

La figura 14 muestra la posición de estos cuantiles para cada distribución. El área sombreada es $\alpha$.

En las tablas de la $\chi^2_n$ y la $t_n$ los correspondientes valores de $\chi^2_{n,\alpha}$ y $t_{n,\alpha}$ se encuentran en el cruce de la fila $n$ y la columna $\alpha$. Los valores de $\alpha$ que figuran en la tabla son los de uso más frecuente. En el caso de la $F_{n_1,n_2}$ se dispone de una tabla para $\alpha = 0{,}025$ y otra para $\alpha = 0{,}05$ (en muchos libros, sobre todo los más antiguos pueden encontrarse tablas para otros

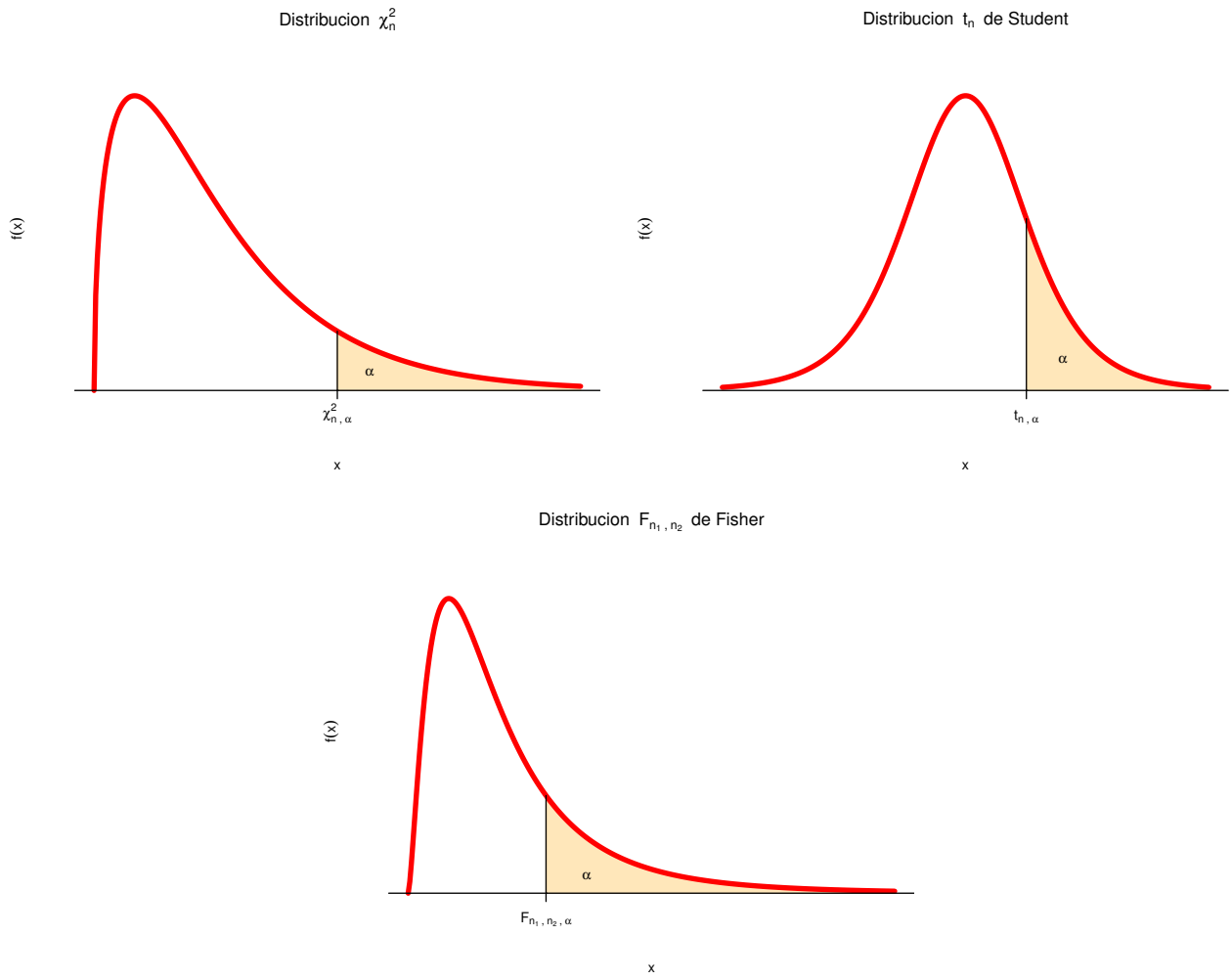


Figura 14: Posición de los cuantiles $q_{1-\alpha}$ de las distribuciones Chi-Cuadrado de Pearson, $t$ de Student y $F$ de Fisher-Snedecor. Estos cuantiles dejan a su derecha un área $\alpha$ (sombreada en las tres figuras).

valores de $\alpha$; hoy en día, con la ubicuidad de la informática, tales tablas en realidad resultan innecesarias). El valor $F_{n_1,n_2,\alpha}$ se localiza simplemente en el cruce de la fila $n_1$ con la columna $n_2$. A veces resulta de interés calcular $F_{n_1,n_2,1-\alpha}$ en cuyo caso se puede utilizar la propiedad siguiente:

$$F_{n_1,n_2,1-\alpha} = \frac{1}{F_{n_2,n_1,\alpha}}$$

Con R estos cuantiles se obtienen directamente como:

- $\chi^2_{n,\alpha}$=`qchisq(1-`$\alpha$`,n)`
- $t_{n,\alpha}$ =`qt(1-`$\alpha$`,n)`
- $F_{n_1,n_2,\alpha}$=`qf(1-`$\alpha$`,`$n_1$`,`$n_2$`)`

# 5. Teorema central del límite.

La propiedad reproductiva de la distribución normal, vista más arriba, nos indica que la suma de variables aleatorias independientes con distribución normal sigue también una distribución normal. El teorema central del límite va un poco más allá, estableciendo condiciones bajo las cuales la suma de variables aleatorias independientes *con distribución no necesariamente normal* sigue una distribución normal. Básicamente tales condiciones son dos: que las variables que se suman tengan todas la misma distribución, y que el número de sumandos sea grande. Estas condiciones se verifican en muchos casos de aplicación práctica; en particular, se cumplen cuando se realiza un muestreo de una variable $X$ con distribución no normal siempre que el número de observaciones sea suficientemente grande, ya que todas las observaciones $X_1, X_2, \ldots, X_n$ proceden de la misma distribución que $X$.

**Teorema Central del Límite** *Sea $X_1, \ldots, X_n$ una secuencia de variables aleatorias independientes y con la misma distribución de probabilidad, siendo $E[X_i] = \mu$ y $var(X_i) = \sigma^2$ (finita) para $i = 1, \ldots, n$. Entonces, para $n \to \infty$:*

$$P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq z\right) \to \Phi(z)$$

*siendo $\Phi(z)$ la función de distribución de la normal tipificada $N(0,1)$.*

**Nota:** El Teorema Central del Límite, tal como se ha enunciado requiere que todas las variables $X_i$ sean independientes y *tengan la misma distribución.* Existen otras versiones de este teorema, en las que se prueba que, bajo determinadas condiciones[1], si las $X_i$ son independientes *aunque tengan distribuciones de probabilidad diferentes*, su suma también tiene una distribución aproximadamente normal.

[1]Tales condiciones exigen la existencia de determinados momentos de las $X_i$, y que éstos no crezcan muy deprisa.

Nótese que:

- $E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right] = n\mu$
- $\text{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{var}\left(X_i\right) = n\sigma^2 \Rightarrow \text{sd}\left(\sum_{i=1}^{n} X_i\right) = \sigma\sqrt{n}$
- Por tanto, la conclusión del del teorema puede enunciarse diciendo que a medida que $n$ aumenta, la distribución de la suma *tipificada* $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$ se va aproximando a la $N\left(0,1\right)$.

Asimismo, si observamos que:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

el teorema central del límite puede expresarse también como:

$$P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z\right) \to \Phi\left(z\right) \quad \text{para } n \to \infty$$

o, dicho de otra forma, la distribución de probabilidad de la media aritmética *tipificada* $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ de una secuencia de $n$ variables independientes y con la misma distribución, de media $\mu$ y desviación típica $\sigma$, se va aproximando a la distribución normal $N\left(0,1\right)$ a medida que $n$ aumenta.

En la práctica, el efecto del teorema central del límite puede apreciarse frecuentemente para valores de $n$ que, si bien son grandes, distan mucho de $\infty$. En muchas ocasiones, con valores de $n$ del orden de entre 30 y 60 ya puede asumirse que, aproximadamente, $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \approx N\left(0,1\right)$ y $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N\left(0,1\right)$, o lo que es lo mismo, que aproximadamente $\sum_{i=1}^{n} X_i \approx N\left(n\mu,\, \sigma\sqrt{n}\right)$ y que $\overline{X} \approx N\left(\mu,\, \sigma/\sqrt{n}\right)$.

En la figura 15 puede apreciarse el significado de este teorema. Cada gráfica corresponde al histograma de 2.000 medias muestrales calculadas sobre muestras de tamaño respectivo 1, 10, 30 y 100 de una distribución exponencial de parámetro $\eta = 100$ (recuérdese que en la distribución exponencial el valor del parámetro coincide con su media). De esta forma cada histograma representa una aproximación a la función de densidad de la media muestral. La linea de trazos corresponde a la estimación de dicha densidad a partir de un suavizado del histograma. La linea roja corresponde a la densidad de una distribución normal cuya media coincide con la de la variable original.

Tal como se puede ver en los gráficos, cuanto mayor es el tamaño de la muestra sobre la que se calcula la media, tanto más se asemeja la distribución de la media a la distribución normal. Asimismo se observa que $E\left[\overline{X}\right]$ se aproxima a $\mu = 100$ y que a medida que $n$ aumenta, $var\left(\overline{X}\right)$ disminuye (de acuerdo con $var\left(\overline{X}\right) = \sigma/\sqrt{n}$).

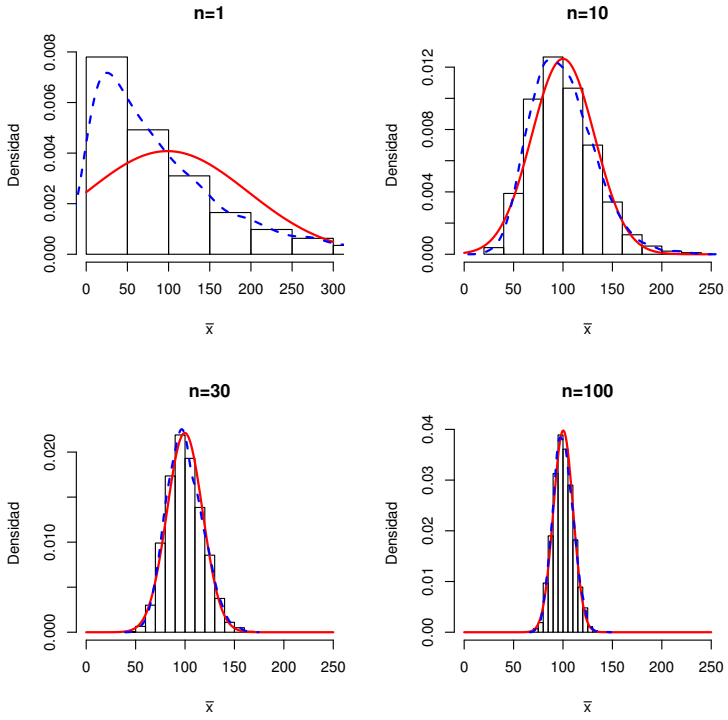


Figura 15: Ilustración del efecto del *Teorema Central del Límite*. A medida que aumenta el tamaño de la muestra ($n$), la distribución de la media aritmética va asemejándose cada vez más a la normal.

### Aproximación de la distribución binomial por la normal

Ya hemos visto en la sección 3.3 que si $X \approx B\left(n,p\right)$ entonces $X = X_1 + X_2 + \cdots + X_n$, siendo las $X_i$ variables de Bernoulli de parámetro $p$ independientes. De acuerdo con el teorema central del límite se tiene que, cuando $n \to \infty$:

$$\frac{X - np}{\sqrt{np\left(1-p\right)}} \approx N\left(0,1\right)$$

En general esta aproximación funciona bien cuando $np \geq 5$, si bien todavía puede mejorarse si se tiene en cuenta el hecho de que la distribución binomial es discreta y la normal es continua. En efecto, la distribución binomial sólo asigna probabilidades a los valores enteros

$0, 1, 2, \ldots, n$ mientras que la normal asignaría probabilidades a todo el rango continuo que contiene a estos valores. Para conseguir una mayor semejanza entre ambas asignaciones se considera que cada valor entero $k$ queda representado por el intervalo $(k - 0{,}5, k + 0{,}5)$. Este procedimiento recibe el nombre de *corrección por continuidad.* De esta forma, la aproximación de las probabilidades binomiales por el teorema central del límite se llevaría a cabo del siguiente modo:

$$\begin{aligned}
P(X = k) \cong & P(k - 0{,}5 \leq X \leq k + 0{,}5) \cong \\
& \cong P\left(\frac{k - 0{,}5 - np}{\sqrt{np(1-p)}} \leq \frac{X - np}{\sqrt{np(1-p)}} \leq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) = \\
& = P\left(\frac{k - 0{,}5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X \geq k) \cong & P(X \geq k - 0{,}5) \cong P\left(Z \geq \frac{k - 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X > k) \cong & P(X \geq k + 0{,}5) \cong P\left(Z > \geq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X \leq k) \cong & P(X \leq k + 0{,}5) \cong P\left(Z \leq \frac{k + 0{,}5 - np}{\sqrt{np(1-p)}}\right) \\
P(X < k) \cong & P(X \leq k - 0{,}5) \cong P\left(Z \leq \frac{k - 0{,}5 - np}{\sqrt{np(1-p)}}\right)
\end{aligned}$$

siendo $Z \approx N(0, 1)$

**Ejemplo:** Se dispone de 50 huevos de tortuga; la probabilidad de que un huevo dé lugar a un macho es 0.30. ¿Cuál es la probabilidad de que en total nazcan más de 16 machos?

Si $X$ es el número de machos, se tiene que $X \approx B(50, 0{,}3)$. La probabilidad pedida es

$$P(X > 16) \cong P(X \geq 16{,}5) \cong P\left(Z \geq \frac{16{,}5 - 50 \cdot 0{,}3}{\sqrt{50 \cdot 0{,}3 \cdot 0{,}7}}\right) = P(Z \geq 0{,}46) = 0{,}32276$$

(hemos utilizado la tabla de la $N(0, 1)$ para calcular la última probabilidad). Si utilizamos R para calcular esta probabilidad de manera exacta, obtenemos:

$$\begin{aligned}
P(X > 16) = \sum_{k=17}^{50} P(X = k) = \sum_{k=17}^{50} \binom{50}{k} 0{,}3^k (1 - 0{,}3)^{50-k} = \\
= \texttt{sum(dbinom(17:50,50,0.3))} = 0{,}31612
\end{aligned}$$

Como vemos el error de aproximación es de algo menos de 7 milésimas (0.00664).