5. DISTRIBUCIONES

R contiene un amplio conjunto de tablas estadísticas; para cada distribución hay comandos que permiten calcular la función de distribución, $F(x) = P(X = x)$; la función de distribución inversa; la función de densidad, y la generación de números pseudoaleatorios de la distribución. El nombre que toma cada distribución en R y los argumentos (parámetros) que cada distribución posee se muestran a continuación.

5.1 DISTRIBUCIONES DISCRETAS

Tabla 5. Distribuciones discretas

Distribución	Fdp	Nombre en R	Argumentos adicionales
Binomial	$\binom{n}{x} p^x q^{n-x}$	binom	size, prob
Geométrica	$q^{x-1}p$	geom	prob
Hipergeométrica	$\dfrac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$	hyper	m,n,k
Binomial Negativa	$\binom{x-1}{k-1} q^{x-k} p^x$	nbinom	size, prob
Poisson	$\dfrac{\lambda^x e^{-\lambda}}{x!}$	pois	lambda

5.2 DISTRIBUCIONES CONTINUAS

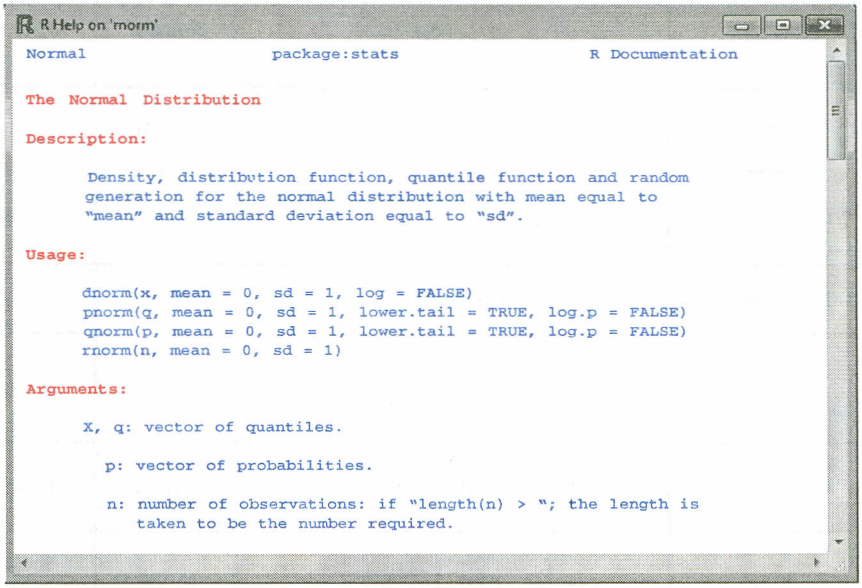
Tabla 6. Distribuciones continuas

Distribución	Fdp	Nombre en R	Argumentos adicionales
Beta	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	beta	shape1, shape2, ncp

(sigue...)

Distribución	Fdp	Nombre en R	Argumentos adicionales
Cauchy	$(\pi\lambda)^{-1}\left\{1+\left[\frac{x-\theta}{\lambda}\right]^2\right\}^{-1}$	cauchy	location, scale
Ji cuadrado	$\frac{1}{\Gamma(v/_2)2^{v/2}}\, x(v/_2{}^{-1})exp(^{-x}/_2)$	chisq	df,ncp
Exponencial	$\frac{1}{\theta}e^{-x}/\theta$	exp	rate
F	$\frac{\Gamma\left(\frac{v_1+v_2}{2}\right)v_1^{v_1/2}v_2^{v_2/2}}{\Gamma(^{v_1}/_2)\Gamma(^{v_2}/_2)}\, f^{(v_1-2)/2}(v_2+v_1f)^{(v_1+v_2)/2}$	f	df1,df2,ncp
Gamma	$\frac{1}{\Gamma(\alpha)\theta^\alpha}\, x^{\alpha-1}exp(^{-x}/\theta)$	gamma	Shape, scale
Log normal	$\frac{1}{\sqrt{2\pi}\sigma x}exp^{-\left(\frac{log(x)-\mu}{2\sigma^2}\right)^2}$	lnorm	meanlog, sdlog
Logística	$\frac{1}{s}exp\left(\frac{x-m}{s}\right)\left(1+exp\left(\frac{x-m}{s}\right)\right)^{-2}$	logis	location, scale
Normal	$\frac{1}{\sqrt{2\pi}\sigma}exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	norm	mean, sd
t de student	$\frac{\Gamma\left[(v+1)/2\right]}{\sqrt{\pi v}\,\Gamma(v/2)}\left[1+(t^2/v)\right]^{-(v-1)/2}$	t	df, ncp
Uniforme	$\frac{1}{b-a}$	unif	min, max
Weibull	$\frac{\alpha}{\theta^\alpha}x^{\alpha-1}\, exp^{[-(x/\theta)^\alpha]}$	weibull	shape,scale

Con los nombres que recibe cada distribución en R se construyen varias funciones que permiten calcular medidas de interés con respecto a cada una de ellas; con esto se tiene que si al nombre de la distribución se le precede por la letra **"d"** se obtiene la función de densidad, si se precede por la letra **"p"** se obtiene la función de distribución, si se precede de la letra **"q"** se obtiene la función de distribución inversa y si se precede de la letra **"r"** se genera números pseudoaleatorios, Si se requiere conocer las características de los argumentos de cada función, se consulta la ayuda interactiva; por ejemplo, si se tiene alguna duda sobre los argumentos utilizados para generar números aleatorios de la distribución Normal o calcular probabilidades de esta distribución, en la consola de R se escribe el comando **help(rnorm)**, con lo que se obtiene:



```
Normal                    package:stats                    R Documentation

The Normal Distribution

Description:

     Density, distribution function, quantile function and random
     generation for the normal distribution with mean equal to
     "mean" and standard deviation equal to "sd".

Usage:

     dnorm(x, mean = 0, sd = 1, log = FALSE)
     pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
     qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
     rnorm(n, mean = 0, sd = 1)

Arguments:

     x, q: vector of quantiles.

       p: vector of probabilities.

       n: number of observations: if "length(n) > "; the length is
          taken to be the number required.
```

Imagen 88. Ayuda interactiva para la distribución Normal

Algunos ejemplos en los cuales se pueden usar las diferentes funciones son:

Ejemplo: Sea X una variable aleatoria que representa el número de máquinas que una compañía produce en un mes; por experiencia, la compañía sabe que la probabilidad de producir una máquina defectuosa es de p = 0.16; considere que el mes pasado la compañía produjo 6 máquinas. Determine cuál es la probabilidad de que de estas 6 máquinas producidas 2 sean defectuosas.

```
>
> dbinom(2,6,0.16)
[1] 0.1911826
>
```

Imagen 89. Salida R distribución binomial

Calcular la probabilidad de que la empresa fabricó a lo sumo 2 máquinas defectuosas.

```
>
> pbimom(2,6,0.16)
[1] 0.9439641
>
```

Imagen 90. Salida R distribución binomial

Para calcular los valores de la función de probabilidad se define un vector con el posible número de máquinas defectuosas (espacio muestral) y se procede como se indica en la siguiente imagen:

```
>
> # Máquinas defectuosas
> x=c(0,1,2,3,4,5,6)
>
> # tamaño de la muestra
> size=6
>
> # Probabilidad de éxito
> prob=0.16
>
> # Probabilidad de éxito
> fdist=dbinom(x,size,prob)
>
> cbind(x,fdist)
     x        fdist
[1,] 0 3.512980e-01
[2,] 1 4.014835e-01
[3,] 2 1.911826e-01
[4,] 3 4.855431e-02
[5,] 4 6.936330e-03
[6,] 5 5.284823e-04
[7,] 6 1.677722e-05
>
```

Imagen 91. Salida R distribución binomial

Para calcular la función de distribución se emplea:

```
>
> # Función de distribución
> # Nº máquinas defectuosas
>
> Fdist=pbinom(x,size,prob)
>
> cbind(x,Fdist)
     x     Fdist
[1,] 0 0.3512980
[2,] 1 0.7527815
[3,] 2 0.9439641
[4,] 3 0.9925184
[5,] 4 0.9994547
[6,] 5 0.9999832
[7,] 6 1.0000000
>
```

Imagen 92. Salida R distribución binomial

Ejemplo: Se seleccionan 10 personas para un trabajo, de 20 ingenieros con doctorado; además, dentro del grupo de opcionados hay 5 ingenieros que fueron catalogados como los mejores. ¿Cuál es la probabilidad de que en el grupo de los 10 ingenieros seleccionados se incluyan cuatro de los cinco mejores.

```
>
> # Nº a calcular la probabilidad
> x=4
>
> # Nº que tienen la característica
> # "Ingeniero bueno"
> m=5
>
> # Nº que no tienen la característica
> n=20-5
>
> # Tamaño de la muestra
> k=10
>
> # Probabilidad
> dhyper(x,m,n,k)
[1] 0.1354489
>
```

Imagen 93. Salida R distribución hipergeométrica

Ejemplo: Un gramo de material radioactivo emite partículas alfa a una tasa constante de 3 partículas por segundo; hallar P(x=0), P(x=1), ..., P(x=7) y hallar la función de distribución.

```
>
> # Posibles valores de x
> x=c(0,1,2,3,4,5,6,7)
>
> # lambda
> lambda = 3
>
> # probabilidades
> fden=(dpois(x,lambda))
>
> # Función de distribución
> Fdist=(ppois(x, lambda))
>
> cdind(x,fden,Fdist)
     x       fden      Fdist
[1,] 0 0.04978707 0.04978707
[2,] 1 0.14936121 0.19914827
[3,] 2 0.22404181 0.42319008
[4,] 3 0.22404181 0.64723189
[5,] 4 0.16803136 0,81526324
[6,] 5 0.10081881 0.91608206
[7,] 6 0.05040941 0.96649146
[8,] 7 0.02160403 0.98809550
> .
```

Imagen 94. Salida R distribución Poisson

Ejemplo: Considere el lanzamiento de un dado hasta la primera aparición de un as; **X** es la variable aleatoria que cuenta el número de fallas hasta obtener un as. Determinar la probabilidad de que se necesiten 3 lanzamientos para obtener un as; determine la función de distribución para x = (0, 1, 2, 3, 4). En el lanzamiento de un dado cada número tiene 1/6, entonces la probabilidad de que salga un as en un lanzamiento es de 0.16.

```
>
> # Valores de x
> # Número de fallas
> x=c(0,1,2,3,4)
>
> # Probabilidades de un as
> prob=0.16
>
> # probabilidades
> fden=(dgeom(x,prob))
>
> # Función de distribución
> Fdist=(dgeom(x,prob))
>
> cdind(x,fden,Fdist)
     x       fden     Fdist
[1,] 0 0.16000000 0.1600000
[2,] 1 0.13440000 0.2944000
[3,] 2 0.11289600 0.4072960
[4,] 3 0.09483264 0.5021286
[5,] 4 0.07965942 0.5817881
>
```

Imagen 95. Salida R distribución geométrica

Se observa que la probabilidad de que se necesiten tres lanzamientos de un dado para obtener un as, es decir, fallar en los dos primeros lanzamientos, es de 0.1128.

Ejemplo: Considere el lanzamiento de un dado hasta obtener cierta cantidad de ases; **X** es el número de fallas hasta obtener un determinado número de ases y **size** es el número de éxitos. Determinar la probabilidad de que se necesiten 5 lanzamientos para obtener 3 ases.

```
>
> # Número de fallas
> x = 2
>
> # Número de éxitos
> size = 3
>
> # probabilidades de un as
> prob = (1/6)
>
> # Probabilidades
> dnbinom(x,size,prob)
[1] 0 0.01929012
>
```

Imagen 96. Salida R distribución binomial negativa

Considere que se necesitaron los siguientes lanzamientos de los datos para obtener los 3 ases especificados (3, 4, 5, 6, 7, 8, 9); construir la función de distribución.

```
>
> # Número de fallas
>  x=c(0,1,2,3,4,5,6)
>
> # Número de éxitos
> size = 3
>
> # Probabilidades de un as
> prob = (1/6)
>
> # Función de distribución
>  Fdist=pnbinom(x,size,prob)
>
>  cdind(x,Fdist)
     x        Fdist
[1,] 0  0.00462963
[2,] 1  0.01620370
[3,] 2  0.03549383
[4,] 3  0.06228567
[5,] 4  0.09577546
[6,] 5  0.13484689
[7,] 6  0.17825959
>
```

Imagen 97. Salida R distribución binomial negativa

Debe tenerse en cuenta que si se realizan tres lanzamientos y se obtienen tres ases, el número de fallas es cero, por tanto, la probabilidad es de 0.00462963; como aparece en la tabla de la imagen anterior, la columna denominada x representa la columna de las fallas.

Ejemplo: Una fábrica de alimentos empaca productos cuyos pesos están distribuidos normalmente con media de 450 gramos y desviación estándar de 20 gramos. Determine la probabilidad de que un paquete escogido al azar pese exactamente 470 gramos, $P(X = 470) = ?$

```
>
> # P ( X = 470 )
> dnorm(470, mean=450, sd=20)
[1,] 0.01209854
>
```

Imagen 98. Salida R distribución Normal

Encuentre la probabilidad de que un paquete escogido al azar pese entre 425 y 486 gramos. $P(425 < X < 486) = ?$

$$P(425 < X < 486) = P(X < 486) - P(X < 425)$$

```
>
> # P ( X < 486 )
> P1 = pnorm(486, mean=450, sd=20)
> P1
[1] 0.9649697
>
> # P ( X < 425 )
> P2 = pnorm(425, mean=450, sd=20)
> P2
[1] 0.1056498
>
> # P ( X < 486 ) - P( X < 425)
> P1 - P2
[1] 0.8584199
>
```

Imagen 99. Salida R distribución Normal

Ejemplo: Encontrar la probabilidad de que una muestra aleatoria de 5 observaciones, de una población normal con varianza $\sigma^2 = 1$, tenga una $S^2 \geq 0.265$. La variable aleatoria asociada con este ejemplo es la siguiente:

$$X^2 = \frac{S^2 (n - 1)}{\sigma^2}$$

Ecuación 8. Distribución ji-cuadrado

Con los datos del ejemplo se calcula el valor para esta variable así:

$$X^2 = \frac{S^2 (n - 1)}{\sigma^2} = \frac{0.265 (5 - 1)}{1}$$

Ecuación 9. Distribución ji-cuadrado

Con lo que se tiene que calcular la probabilidad $P(x^2 > 1.06)$ o análogamente $1 - P(x^2 < 1.06)$, mediante la distribución ji cuadrado.

```
>
> # Grados de libertad (n-1)
> df=(5-1)
>
> # P(ji >= 1.06)
> pchisq(1.06,df,lower.tail=FALSE)
[1] 0.9005656
>
> # 1 - P(ji >= 1.06)
> 1 - pchisq(1.06,df)
[1] 0.9005656
>
```

Imagen 100. Salida R distribución ji-cuadrado

Se tiene una probabilidad de 0.90 de que el valor de la varianza muestral sea mayor o igual a 1.06.

Ejemplo: El gerente de una fábrica de cierto tipo de alimentos asegura que el peso promedio del producto que elabora es de 165.285 g. El encargado de control de calidad toma una muestra de 16 paquetes del producto y los pesa. Los resultados fueron los siguientes: 165, 158, 153, 162, 171, 175, 173, 169, 166, 170, 164, 177, 148, 167, 152, 149. Encuentre la probabilidad de $\bar{x} < 163.6875$. La distribución muestral asociada con este ejemplo es la siguiente:

$$t = \frac{X - \mu}{S / \sqrt{n}}$$

Ecuación 10. Distribución t

con v = n - 1 grados de libertad, con esto se tiene lo siguiente:

```
>
> # vector con los elementos de la muestra
>
> X=c(165,158,153,162,171,175,173,169,
+      166,170,164,177,148,167,152,149)
>
> # Tamaño muestral
> n = length(X)
> n
[1] 16
>
> # Media muestral
> M = mean(X)
> M
[1] 163.6875
>
> # Desviación estándar muestral
> s = sd(X)
> s
[1] 9.235574
>
> # Valor para la v.a.
> t= (M - 165.285) / (s / sqrt(n))
> t
[1] -0.6918898
```

Imagen 101. Salida R distribución t

Con lo que se tiene que calcular la probabilidad P(t = -0.69189), mediante la distribución t,

```
>
> # Grados de libertad n-1
> df = 16 - 1
>
> # Probabilidades
> pt(-0.691, df)
[1] 0.2500601
>
```

Imagen 102. Salida R distribución t

La probabilidad de que la media sea menor de 163.6875 es de 0.250.

Ejemplo: En una prueba sobre efectividad de dos tipos distintos de píldoras para dormir, A y B, se utilizarán dos grupos independientes de personas con insomnio. A un grupo de 40 personas se le administrará la píldora A y al otro grupo, de 60, se le administrará la B, y se registrará el número de horas de sueño de cada individuo participante en el estudio. Si se supone que el número de horas de sueño de quienes usan cada tipo de píldora se distribuye normalmente con $\sigma^2_1 = \sigma^2_2$, determinar la probabilidad:

$$P\left(\frac{S_1^2}{S_2^2} < 1.93\right)$$

Ecuación 11. Distribución F

La distribución muestral asociada con el ejemplo anterior es la siguiente:

$$F = \frac{S_1^2\,\sigma_1^2}{S_2^2\,\sigma_2^2}$$

Ecuación 12. Distribución F

Como para el caso se tiene que $\sigma^2_1 = \sigma^2_2$, entonces, 1.93 es un valor de una distribución F con 39 grados de libertad en el numerador y 59 grados de libertad en el denominador, el cual se puede calcular de la siguiente forma:

```
>
> # tamaño de muestra A
> n1 = 40
> # tamaño de muestra B
> n2 = 60
>
> # Grados de libertad numerador A
> df1 = 40-1
> # Grados de libertad denominador B
> df2 = 60-1
>
> # valor a probar
> x = 1.93
>
> # Probabilidad
> pf( x, df1, df2)
[1] 0.9890417
>
```

Imagen 103. Salida R distribución F

Se tiene una probabilidad del 98.9% de que el cociente entre las varianzas del número de horas de sueño de cada individuo participante en el estudio con las píldoras A y B es menor a 1.93.

CONSIDERACIÓN

En ocasiones se tiene un conjunto de datos del cual se desconoce su distribución, pero se desea inferir con estos datos acerca de una población; así que se hace necesario estudiar la distribución que sigue el conjunto de datos; este estudio se puede realizar de varias formas. Una es mediante un análisis descriptivo, utilizando comandos para calcular medidas de posición, dispersión y de forma summary(), skeweness(), kurtosis(), sd() y range(), entre otras. Otra, mediante el uso de gráficas que permiten observar el comportamiento de los datos *stem(), boxplot(), hist().* También es posible graficar la distribución empírica acumulada de los datos con el comando **plot(ecdf(vector de datos))**. Si se desea saber si los datos se ajustan a la distribución Normal es posible realizar un diagrama cuantil-cuantil (Q–Q) con el comando **qqnorm(vector de datos)** o realizar una prueba de hipótesis para comprobar la normalidad; las pruebas serán descritas con mayor profundidad en la sección de inferencia estadística.

El comando **moment**(x, **order=**, **central=**, **absolute=**) permite calcular un vector de datos con los momentos muestrales de un orden específico; los argumentos utilizados son: **x** es el vector que contiene el conjunto de datos, **order=** se refiere al momento que se desea calcular, **central=** se relaciona con un valor lógico (si es TRUE, calcula el momento central; si es FALSE, calcula los momentos respecto al origen) y **absolute=** se relaciona con un valor lógico (si es TRUE, determina el momento aplicándole el valor absoluto a cada uno de los componentes del vector). Para poder trabajar con este comando es necesario cargar previamente el paquete **moments**. Por ejemplo, el momento de orden uno alrededor del origen coincide con la media, a continuación se muestra cómo calcular este primer momento en R.

```
>
> X=c(29,78,48,29,30,44,72,73,46,82,84,71,75,84,45,45)
>
> # Momento de orden uno alrededor del origen
> moment(X,order=1,central=FALSE,absolute=TRUE)
[1] 58.4375
>
> # Media del vector X
> mean(X)
[1] 58.4375
```

Imagen 104. Salida R momentos

Además, con el comando

all.moments(x,order.max=,central=,Absolute=) es posible calcular varios momentos a la vez en un conjunto de datos; los argumentos son los mismos utilizados en el comando **moment** y se le adiciona el argumento **orden.max=**, que fija hasta qué momento se desea calcular; teniendo en cuenta el vector X utilizado en el ejemplo anterior, se calculan los momentos con un orden máximo de cuatro.

```
>
> all.moments(X, order.max = 4, central = FALSE, absolute = TRUE)
 [1] 1.000000e+00 5.843750e+01 3.817937e+03 2.690324e+05 1.988036e+07
>
```

Imagen 105. Salida R momentos

5.3 EJERCICIOS

5.3.1 Un jugador de tejo revienta mecha dos veces de cada cinco que lanza. Si en una partida dicho jugador lanza 15 veces,

a. Calcule la probabilidad de que reviente tres veces mecha.

b. Calcule la probabilidad de que reviente por lo menos una vez.

c. Calcule la probabilidad de que el jugador reviente más de diez veces durante el juego.

5.3.2 El número de vehículos que llegan a un peaje durante una hora sigue una distribución de Poisson de parámetro $\lambda = 15$.

a. Calcule la probabilidad de que sólo llegue un vehículo.

b. Calcule la probabilidad de que lleguen más de 5 vehículos.

5.3.3 Sea Z una variable aleatoria normal estándar, es decir, $Z \sim N(0, 1)$; calcule las siguientes probabilidades:

a. $P(Z \leq 1.64)$ b. $P(Z \geq 0.5)$
c. $P(Z \leq -1)$ d. $P(-1.2 \leq Z \leq 2)$
e. $P(Z \geq -0.2)$ f. $P(Z \geq 1.96)$

5.3.4 Un estudio realizado en una empresa para determinar el coeficiente intelectual de los empleados arrojó que dicha medida se distribuye normal con parámetros $\mu = 99$ y $\sigma = 6$.

a. Calcule la probabilidad de que un empleado seleccionado al azar tenga un coeficiente intelectual por debajo de 90.

b. Calcule el porcentaje de empleados cuyo coeficiente intelectual está comprendido entre 97 y 108.

5.3.5 Genere una muestra de tamaño 20 de las siguientes poblaciones:

a. Distribución Binomial con parámetro $p = 0.35$

b. Distribución Poisson con parámetro $\lambda = 23$

c. Distribución Normal con parámetros $\mu = 25$ y $\sigma = 3.3$