Universidad de Alcalá

Estadística

Práctica 6

Variables aleatorias y distribuciones continuas

Universidad de Alcalá

Contenido

1. Introducción

Con esta práctica se utiliza R y RStudio para calcular probabilidades de sucesos a partir de variables aleatorias continuas que tienen una distribución Normal o que se aproximan a una distribución Normal.

2. Variables continuas con distribución Normal

Una variable Normal es una variable continua (X) cuya función de probabilidad o densidad de probabilidad se ajusta a la siguiente expresión matemática, siendo μ la esperanza (media) y σ la desviación estándar de la variable:

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

Se representa como X:N(μ,σ) y se dice que X tiene una distribución Normal.

Por ejemplo, una variable X:N(5,1) que representa: "Calificación obtenida en una asignatura con una calificación media de 5 y una desviación estándar de 1".

En R se pueden utilizar las funciones indicadas en la siguiente tabla. Por simplicidad, para la esperanza "μ" usaremos "m" y para la desviación estándar usaremos "d". En la columna "Ejemplo de uso", se utiliza una variable N(5,1).

Función	Código R	Resultado	Ejemplo de uso

Densidad de probabilidad f(x)	`dnorm(x,m,d)`
	
	También puede usarse:
	
	`dnorm(x, mean=m, sd=d)`
		En teoría el resultado siempre debería ser 0, sólo debe usarse para dibujar la función de densidad.	`No debe usarse para calcular probabilidades, sólo para dibujar.`

Distribución de probabilidad F(x)=P{X≤x}= P{X<x}	`pnorm(x,m,p)`	Calcula la probabilidad de que una variable aleatoria Normal X:N(m,d) tenga un valor igual o menor que x.	`>pnorm(6,5,1)` `[1] 0.8413447`

Cuantil	`qnorm(q,n,p)`	Calcula el "menor" valor de una variable aleatoria Normal X:N(m,d) que cumple que la probabilidad de que la variable sea menor o igual a ese valor es igual a q.	`>qnorm(0.7,5,1)`
			`[1] 5.524401`
			
			Entonces se cumple: P{X≤5.524401} = 0.7

Función	Código R	Resultado	Ejemplo de uso
Generación aleatoria de muestras	`rnorm(m,n,p)`	Genera aleatoriamente un vector de m valores aleatorios de una variable Normal X:N(m,d).	`>rnorm(4,5,1)` `[1] 6.187306 4.550551 4.250897 4.2582126`
Esperanza E[X]	`m`		`>(E=5)` `[1] 5`
Varianza Var[X]	`d^2`		`>(Var=1^2)` `[1] 1`

Las funciones anteriores se pueden combinar con otras de generación de gráficos, para obtener los diagramas que se indican en la siguiente tabla.

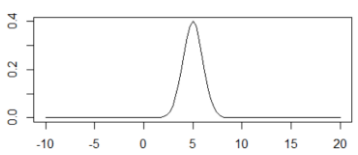
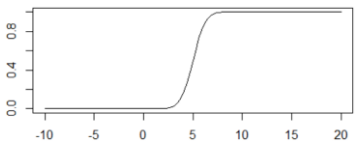
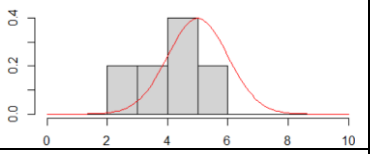
Diagrama	Código R	Comentarios	Ejemplo de uso
Función de densidad de probabilidad	`curve(dnorm(x,m,d),a,b)`	Dibuja el resultado de la función desde x=a hasta x=b	`>curve(dnorm(x,5,1), -10,20)` 
Función de distribución de probabilidad	`curve(pnorm(x,m,d),a,b)`	Dibuja el resultado de la función desde x=a hasta x=b	`>curve(pnorm(x,5,1), -10,20)` 

Diagrama	Código R	Comentarios	Ejemplo de uso

Histograma de una muestra de valores reales de la variable aleatoria	`hist(datos,breaks=a:b, freq=FALSE)`
	
	`curve(dnorm(x,m,d),a,b, add=TRUE)`
		Se puede representar primero el histograma con los datos de la variable y después la función de distribución de probabilidad Normal teórica, para ver si son similares.
		Los valores reales de la muestra están en "datos".
			`>datos=c(5,5,4,6,4,5,5,3,3,6)`
			
			`>hist(datos, breaks = 0:10 freq=FALSE)`
			
			`>curve(dnorm(x,5,1),0,10, add=TRUE, col="red")`
			 

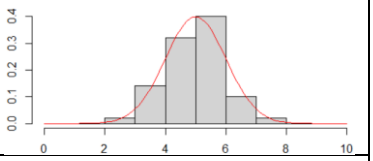
Histograma de una muestra de valores simulados de la variable aleatoria	`hist(rnorm(s,m,d), breaks=a:b, freq=FALSE)`
	
	`curve(dnorm(x,m,d),a,b, add=TRUE)`
		Se puede representar primero el histograma con los datos de la simulación de "s" datos y después la función de probabilidad Normal teórica, para ver si son similares.	`>hist(rnorm(100,5,1), breaks = 0:10, freq=FALSE)`
			
			`>curve(dnorm(x,5,1),0,10, add=TRUE, col="red")`
			 

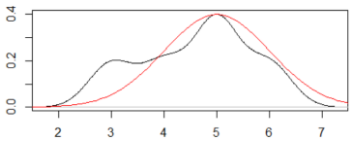
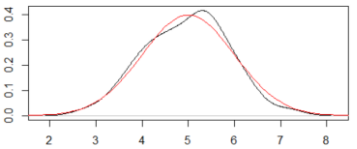
Diagrama de densidad a partir de una muestra de valores reales de la variable aleatoria	`plot(density(datos))`
	
	`curve(dnorm(x,m,d),a,b, add=TRUE)`
		Se utilliza la función density() de R.
		
		Los valores reales de la muestra están en "datos".
			`>datos=c(5,5,4,6,4,5,5,3,3,6)`
			
			`>plot(density(datos))`
			
			`>curve(dnorm(x,5,1),0,10, add=TRUE, col="red")`
			 

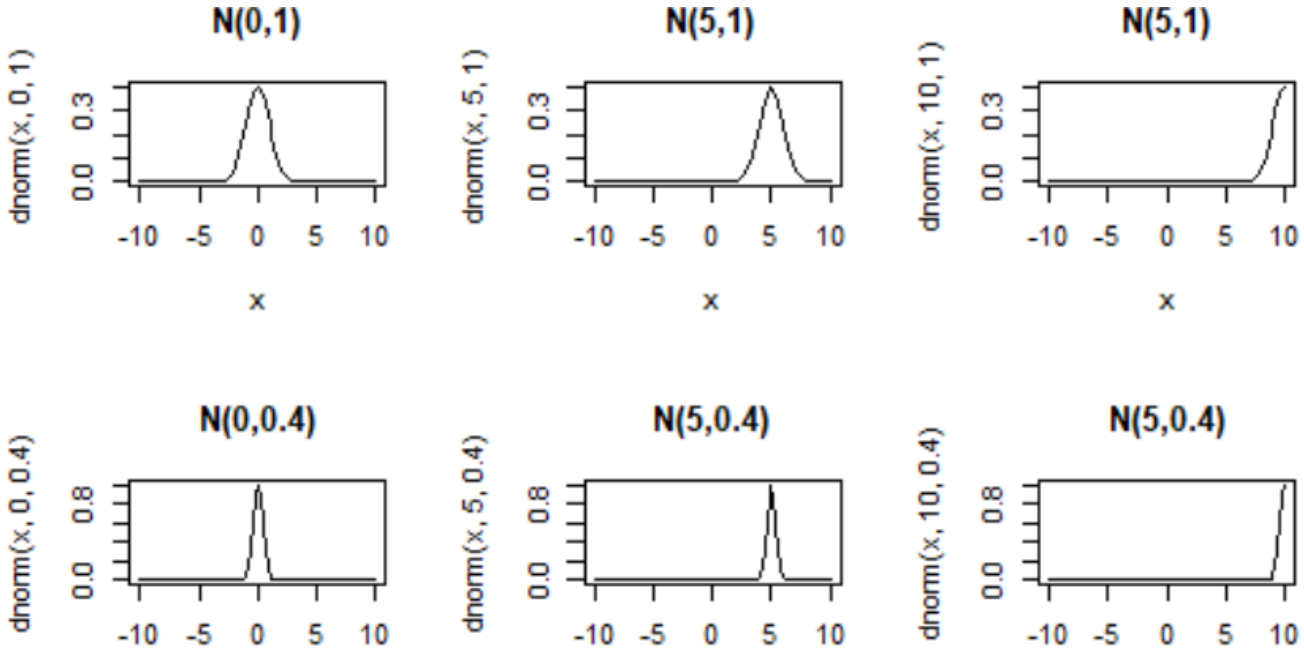
Diagrama de densidad a partir de una muestra de valores simulados de la variable aleatoria	`plot(density(rnorm(s,m,d)))`
	
	`curve(dnorm(x,m,d),a,b, add=TRUE)`
		Se utilliza la función density() de R.	`>plot(density(rnorm(100,5,1))`
			
			`>curve(dnorm(x,5,1),0,10, add=TRUE, col="red")`
			 

En la función `curve()` se indica `add=TRUE` para que se dibuje sobre el diagrama anterior.

Se puede comparar el efecto de modificar los parámetros m y d, mostrando simultáneamente varios gráficos.

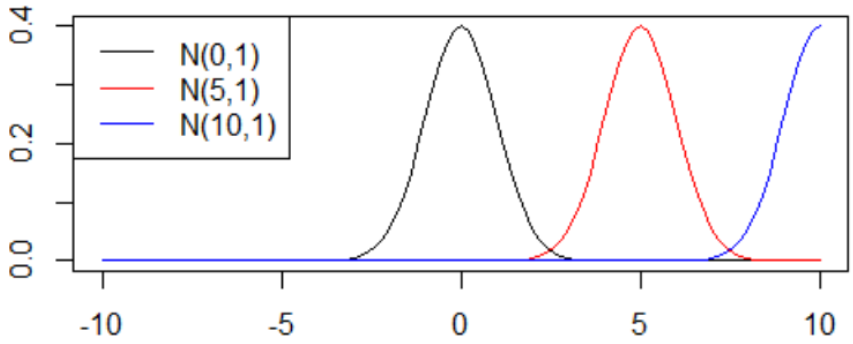
Por ejemplo, dividiendo la ventana gráfica en 2 filas y 3 columnas:

```
par(mfrow=c(2,3))
curve(dnorm(x,0,1),-10,10, main = "N(0,1)")
curve(dnorm(x,5,1),-10,10, main = "N(5,1)")
curve(dnorm(x,10,1),-10,10, main = "N(5,1)")
curve(dnorm(x,0,0.4),-10,10, main = "N(0,0.4)")
curve(dnorm(x,5,0.4),-10,10, main = "N(5,0.4)")
curve(dnorm(x,10,0.4),-10,10, main = "N(5,0.4)")
par(mfrow=c(1,1))
```



O se pueden dibujar varios diagramas en el mismo gráfico, el primero con plot y el resto con lines:

```
curve(dnorm(x,0,1),-10,10)
curve(dnorm(x,5,1),-10,10, add=TRUE, col="red")
curve(dnorm(x,10,1),-10,10, add=TRUE, col="blue")
legend("topleft",legend=c("N(0,1)","N(5,1)","N(10,1)"),
      col=c("black", "red", "blue"), bg="transparent", lty=1)
```



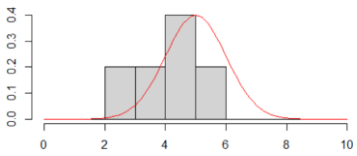
Se recomienda investigar diferentes formas de dibujar la leyenda en el gráfico.

2.1 Comprobar la normalidad de un conjunto de datos

Se puede comprobar si un conjunto de datos se adapta a una distribución Normal de varias formas:

1) Visualmente comparando el histograma de los datos con la función de densidad Normal
2) Visualmente comparando la función de densidad de los datos con la función de densidad Normal
3) Visualmente utilizando un diagrama de cuantiles (Q-Q plot)
4) Comprobando la regla empírica de la Normal: 68-95-99
5) Comprobando la asimetría y apuntamiento
6) Realizando un test o contraste de hipótesis

Método	Código R	Comentarios	Ejemplo de uso

Histograma datos vs Densidad Normal	`hist(x, freq=FALSE)`
	
	`curve(dnorm(x,m,d),a,b, add=TRUE)`
		x es el vector con la lista de valores de la variable aleatoria	`>x=c(5,5,4,6,4,5,5,3,3,6)`
			
			`>hist(x, freq=FALSE, breaks=0:10)`
			
			`>curve(dnorm(x,5,1),0,10, add=TRUE, col="red")`
			 

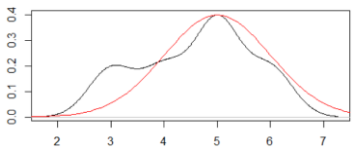
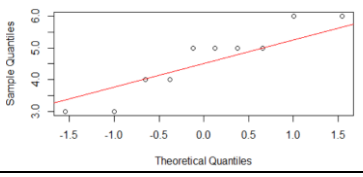
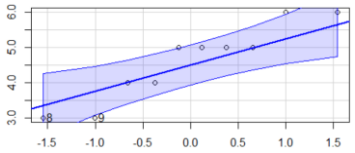
Densidad datos vs Densidad Normal	`x = datos`
	
	`plot(density(x))`
	
	`curve(dnorm(x,m,d),a,b, add=TRUE)`
		x es el vector con la lista de valores de la variable aleatoria	`>x=c(5,5,4,6,4,5,5,3,3,6)`
			`>plot(density(x))`
			`>curve(dnorm(x,5,1),0,10, add=TRUE, col="red")`
			 

Gráfico Q-Q (Cuantil-Cuantil)	`x = datos`
	
	`qqnorm(x)`
	
	`qqline(x)`
		Cuanto más próximos estén los puntos a la recta, más se ajustan los datos a una distribución Normal,	`>x=c(5,5,4,6,4,5,5,3,3,6)`
			`>qqnorm(x)`
			`>qqline(x,col="red")`
			 

Método	Código R	Comentarios	Ejemplo de uso

Diagrama Q-Q con región de confianza del 95%	`install.packages("car")`
	`library(car)`
	
	`qqPlot(x)`
		Cuantos más puntos queden dentro de la región, más se ajustan los datos a una distribución Normal,	`>x=c(5,5,4,6,4,5,5,3,3,6)`
			`>qqPlot(x)`
			 

Regla empírica 68-95-99	`n=length(x)`
	
	`length(x[x<m+d & x>m-d])/n`
	
	`length(x[x<m+2*d & x>m-2*d])/n`
	
	`length(x[x<m+3*d & x>m-3*d])/n`
		Regla empírica:
		- En m±d se encuentra el 68.27% de la distribución.
		- En m±2d se encuentra el 95.45% de la distribución
		- En m±3d se encuentra el 99.73% de la distribución
			`> x=c(5,5,4,6,4,5,5,3,3,6)`
			`> (n=length(x))`
			`[1] 10`
			`> (m=mean(x))`
			`[1] 4.6`
			`> (d=sd(x))`
			`[1] 1.074968`
			
			`> length(x[x<m+d & x>m-d])/n`
			`[1] 0.6`
			`> length(x[x<m+2*d & x>m-2*d])/n`
			`[1] 1`
			`> length(x[x<m+3*d & x>m-3*d])/n`
			`[1] 1`

Asimetría y Apuntamiento (Curtosis)	`install.packages("e1071")`
	`library(e1071)`
	
	`skewness(x)`
		Si[1] el valor absoluto del coeficiente de asimetría (skewness) es ≤1.5, y el del coeficiente de apuntamiento (curtosis) es ≤1.	`> x=c(5,5,4,6,4,5,5,3,3,6)`
			`> skewness(x)`
			`[1] 0.0133508`
			`> kurtosis(x)`
			`[1] -0.4657815`

Test de hipótesis de asimetría y curtosis de Jarque-Bera	`library("tseries")`
	
	`jarque.bera.test(x)`
		Si se obtiene un p-valor mayor que un valor tabulado[2] en función del tamaño de la muestra, se puede afirmar con un 95% de	`> x=c(5,5,4,6,4,5,5,3,3,6)`
			`> jarque.bera.test(x)`
			`p-value = 0.6983`

[1] Una distribución Normal tiene un coeficiente de asimetría y de apuntamiento de valor 0. No hay unanimidad entre los expertos sobre los límites alrededor de estos valores para considerar que unos datos son normales. Algunas propuestas, entre ellas la indicada en la tabla, pueden encontrarse en: https://rpubs.com/FabioScielzoOrtiz/Analisis_de_Normalidad_en_R
https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/Simon

[2] Tabla de posibles p-valores en función del tamaño de la muestra:
https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test

Método	Código R	Comentarios	Ejemplo de uso
		confianza que se ajusta a una distribución Normal.	

Test de hipótesis (Test de Shapiro-Wilk) Muestra pequeña[3]	`shapiro.test(x)`	Si se obtiene un p-value mayor que 0.05 se puede afirmar con un 95% de confianza que se ajusta a una distribución Normal.	`> x=c(5,5,4,6,4,5,5,3,3,6)`
			`> shapiro.test(x)`
			`Shapiro-Wilk normality test`
			`data: x`
			`W = 0. 89165,`
			`p-value = 0.177`

Test de hipótesis (Test de Kolmogorov-Smirnov) Muestra grande	`ks.test(x, pnorm, mean(x), sd(x))`	Si se obtiene un p-value mayor que 0.05 se puede afirmar con un 95% de confianza que se ajusta a una distribución Normal.	`> x=rnorm(100,5,2)`
			`> ks.test(x, pnorm, mean(x), sd(x))`
			`Asymptotic one-sample Kolmogorov-Smirnov test`
			`data: x`
			`D = 0.062082,`
			`p-value = 0.8356`

[3] No hay unanimidad respecto a lo que se considera muestra pequeña o grande. En general, para estas pruebas se suele suponer el límite en 50 datos. Aunque hay autores que defienden que el test de Shapiro-Wilk sólo debería realizarse para una muestra de como mínimo 30 datos.

2.2 Ejemplo: Calificaciones de estudiantes

ENUNCIADO

Mediante una encuesta, se saben las calificaciones de acceso a la universidad de 74 estudiantes de un curso de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá. Si se supone que tienen una distribución Normal, responder a las siguientes preguntas:

a) **Definir una variable aleatoria Normal para la calificación de un estudiante**
b) **Calcular la media o esperanza y la varianza de la variable aleatoria**
c) **Dibujar las funciones de densidad y distribución de la variable aleatoria**
d) **Cuál es la probabilidad de que un estudiante seleccionado al azar tenga una calificación menor de 7. Dibujar en el diagrama de la función de densidad el área correspondiente a la acumulación de la probabilidad obtenida.**
e) **Cuál es la probabilidad de que un estudiante seleccionado al azar tenga una calificación de entre 7 y 9. Dibujar en el diagrama de la función de densidad el área correspondiente a la acumulación de la probabilidad obtenida.**
f) **Cuál es la probabilidad de que un estudiante seleccionado al azar tenga una calificación mayor que 9. Dibujar en el diagrama de la función de densidad el área correspondiente a la acumulación de la probabilidad obtenida.**
g) **Cuál es el valor de los cuartiles de la variable aleatoria. Dibujar en el diagrama de la función de densidad el área correspondiente al segundo cuartil.**
h) **Si tenemos los datos de la encuesta original a los 74 estudiantes, comprobar si tomando las notas reales, el histograma se ajusta realmente a la distribución Normal.**
i) **Repetir el apartado anterior pero haciendo la comprobación utilizando el diagrama de la función de densidad generada con `density()`.**
j) **Repetir el apartado anterior pero haciendo la comprobación utilizando el diagrama Q-Q (Cuantil-Cuantil): sin y con región de confianza del 95%.**
k) **Repetir el apartado anterior pero haciendo la comprobación mediante la regla empírica 68-95-99 de la distribución Normal.**
l) **Repetir el apartado anterior haciendo la comprobación mediante los coeficientes de asimetría y apuntamiento.**
m) **Repetir el apartado anterior pero haciendo la comprobación mediante test de hipótesis.**
n) **Repetir los apartados h) a l) pero utilizando datos generados aleatoriamente con R, en lugar de los datos reales.**

SOLUCIÓN

a) Definir una variable aleatoria Normal para la calificación de un estudiante

X = Calificación de un estudiante.

La variable sigue una distribución Normal N(m,d), donde hay que calcular m y d.

La media es m, por lo que hay que acceder a las notas de los 74 estudiantes y calcular la media.

En el archivo encuesta.csv estaban los resultados de la encuesta.

```
> encuesta = read.csv2("encuesta.csv")
> (m = round(mean(encuesta$NOTA),2))
[1] 8.02
> (d = round(sd(encuesta$NOTA),2))
[1] 1.13
```

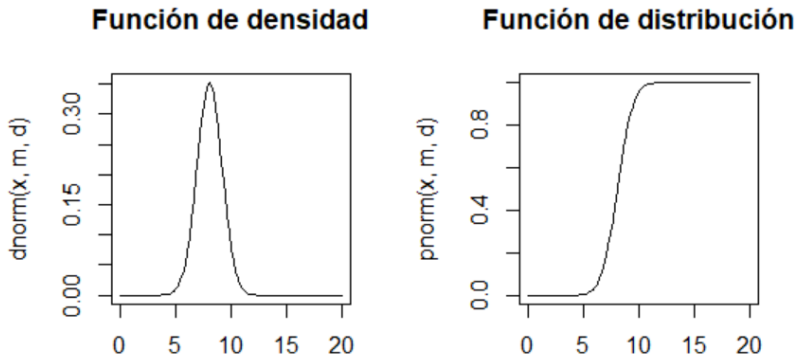
Por tanto la variable es X:N(8.02, 1.13).

b) Calcular la media o esperanza y la varianza de la variable aleatoria

```
> (Esperanza = m)
[1] 8.02
> (Varianza = round(d^2,2))
[1] 1.28
```

c) Dibujar las funciones de densidad y distribución.

```
> par(mfrow=c(1,2))
> curve(dnorm(x,m,d),0,20, main="Función de densidad")
> curve(pnorm(x,m,d),0,20, main="Función de distribución")
> par(mfrow=c(1,1))
```



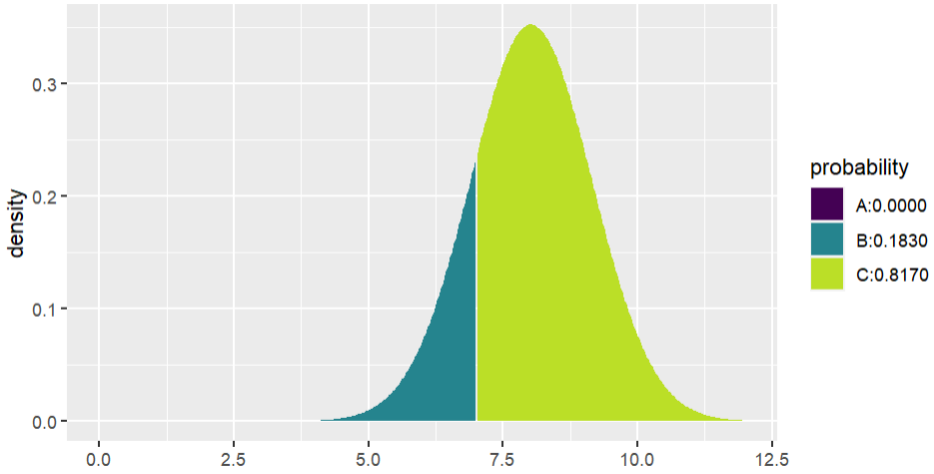
d) Cuál es la probabilidad de que un estudiante seleccionado al azar tenga una calificación menor de 7

Se trata de calcular P{X<7} que al ser una variable continua equivale a P{X≤7}. Se puede usar, por tanto la función de distribución `pnorm()`:

```
> (P.menor.7=pnorm(7,m,d))
[1] 0.1829814
```

Para dibujar el área correspondiente a la acumulación de la probabilidad obtenida, podemos ejecutar la función xnorm() del paquete mosaic:

```
> install.packages("mosaic")
> library(mosaic)
> xpnorm(c(0,7),m,d, xlim=c(0,12))
```



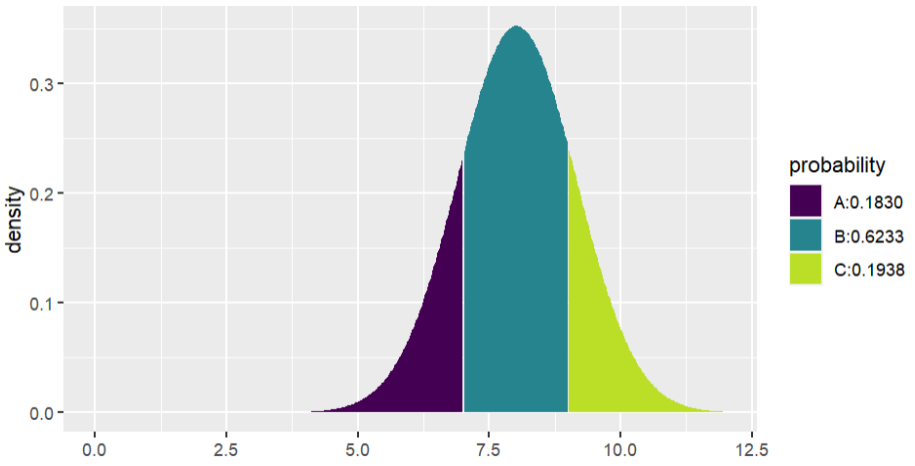
e) Cuál es la probabilidad de que un estudiante seleccionado al azar tenga una calificación de entre 7 y 9

Se trata de calcular P{X≤9} - P{X≤7}.

```
> (P.entre.7.y.9=pnorm(9,m,d)-pnorm(7,m,d))
[1] 0.6232614
```

Para dibujar el área correspondiente a la acumulación de la probabilidad obtenida, podemos ejecutar el siguiente comando:

```
> xpnorm(c(7,9),m,d, xlim=c(0,12))
```

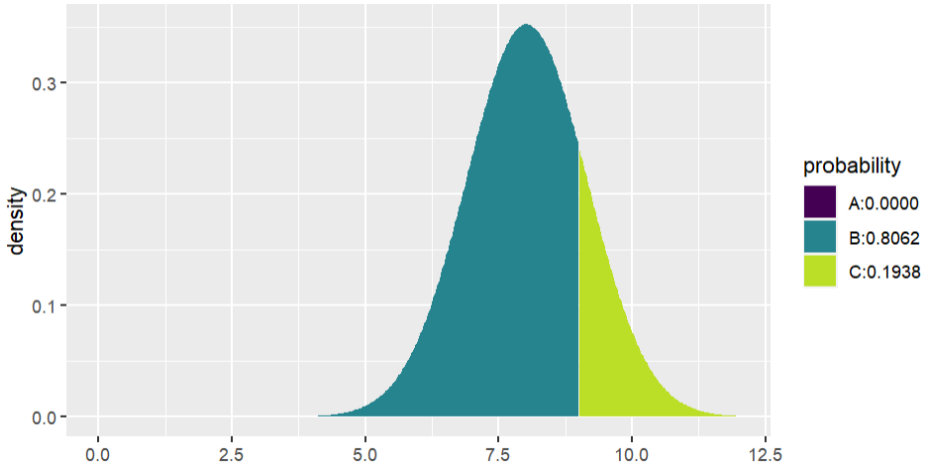


f) Cuál es la probabilidad de que un estudiante seleccionado al azar tenga una calificación superior a 9.

En ese caso P{X>9} = 1 – P{X<9}.

```
> (P.mayor.9=1-pnorm(9,m,d))
[1] 0.1937572
```

```
> 1-xpnorm(c(0,9),m,d, xlim=c(0,12))
```



g) Cuál es el valor de los cuartiles de la variable aleatoria (es decir los percentiles 25%, 50%, 75%, o cuantiles 0.25, 0.50, 0.75). Dibujar en el diagrama de la función de densidad el área correspondiente al segundo cuartil.

En lugar de ejecutar 4 veces la función `qnorm()`, una para cada cuartil, se puede pasar como primer argumento, un vector con los valores de los cuantiles correspondiente a cada cuartil:

```
> (cuartiles=round(qnorm(c(0.25,0.50,0.75),m,d),2))
[1] 7.26 8.02 8.78
```

Lo que indican estos resultados es que:

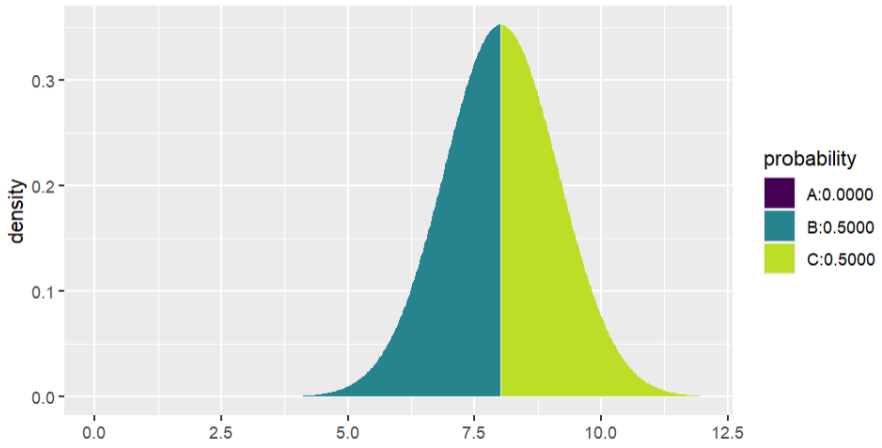
- P{X≤7.26} = 0.25
- P{X≤8.02} = 0.50
- P{X≤8.78} = 0.75

Puede comprobarse que es cierto con la función de distribución, por ejemplo, para el segundo cuartil:

```
> (segundo.cuartil=cuartiles[2])
[1] 8.02
> (p.segundo.cuartil=pnorm(segundo.cuartil,m,d))
[1] 0.5
```

Se puede mostrar gráficamente:

```
> xpnorm(c(0,segundo.cuartil),m,d, xlim=c(0,12))
```



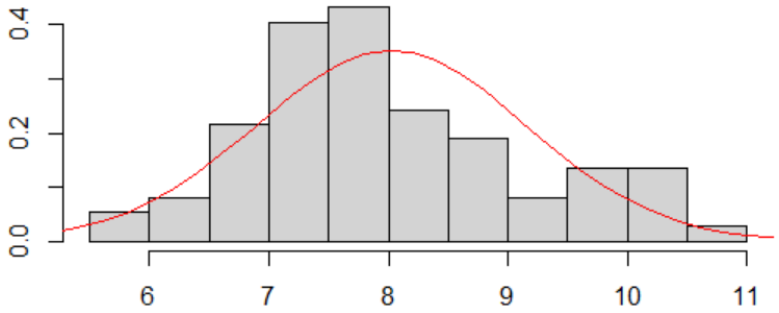
h) Si tenemos los datos de la encuesta original a los 74 estudiantes, comprobar si tomando las notas reales, el histograma se ajusta realmente a la distribución Normal

Las notas están en la columna NOTA:

```
> (notas.reales=encuesta$NOTA)
 [1]  8.50  7.10  8.63  8.62  8.20  8.70  7.21  7.63  8.40  8.30
[11]  8.20  9.10  9.79 10.11  8.02  7.31  7.50  8.71  8.34  9.21
[21]  7.80 10.30  7.99  6.90  7.80 10.00  8.59  7.00  8.05 10.80
[31]  7.99  8.55  7.34  6.75  9.56  7.42  6.94  7.21  7.68 10.28
[41]  7.86 10.26  7.27  5.80  7.30  7.14  8.60  7.50  8.00  7.54
[51]  7.29  7.83  6.75  9.81  6.80  6.44  6.65  7.80 10.27  7.60
[61]  7.87  7.00  7.08  7.48  8.07  5.82  6.50  9.90  7.50  6.50
[71]  9.46  8.00  7.80  7.65
```

Podemos crear un vector con esos valores y dibujar su histograma y superponer la curva de la función de densidad Normal.

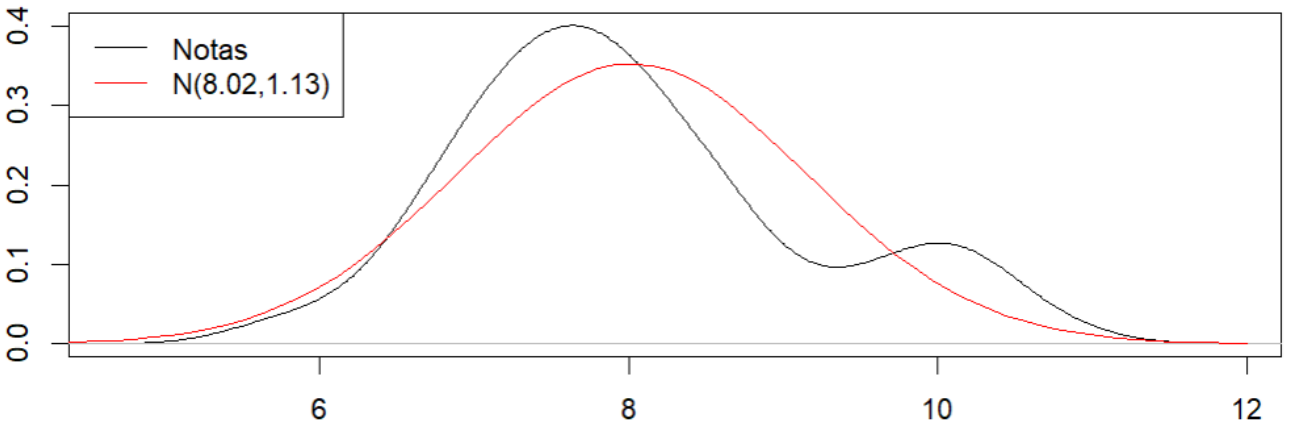
```
> hist(notas.reales, freq=FALSE)
> curve(dnorm(x,m,d),0,20, add = TRUE)
```



Puede comprobarse visualmente que el histograma con los datos reales se ajusta aproximadamente a la forma de la distribución Normal usada en este ejemplo.

i) Repetir el apartado anterior pero haciendo la comprobación utilizando el diagrama de la función de densidad generada con density().

```
> plot(density(notas.reales))
> curve(dnorm(x,m,d),0,12, add=TRUE, col="red")
> legend("topleft",legend=c("Notas","N(8.02,1.13)"), col=c("black", "red"), lty=1)
```

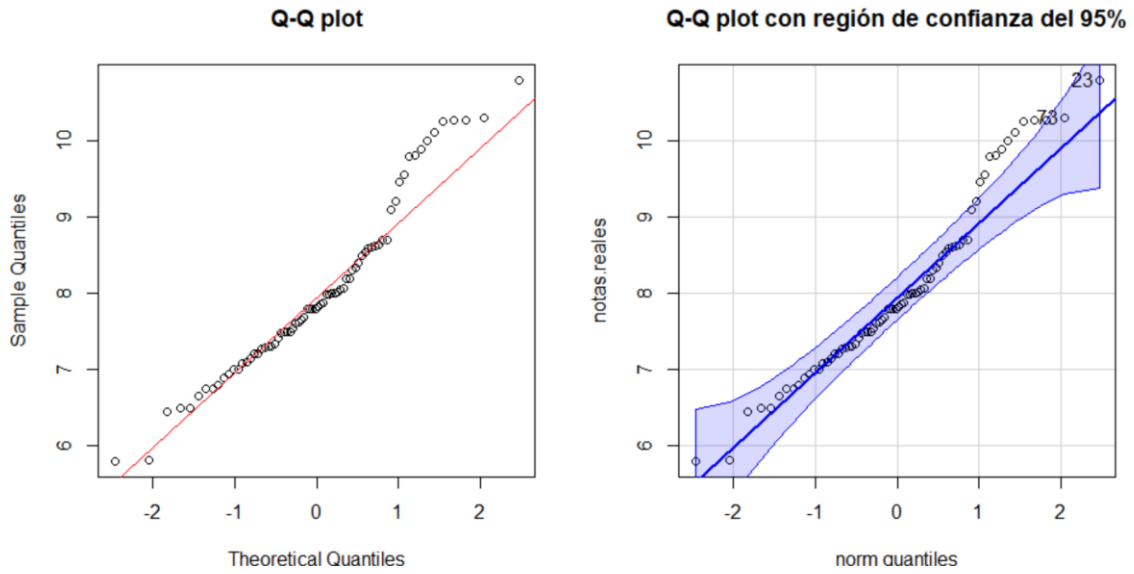


j) Repetir el apartado anterior, pero haciendo la comprobación utilizando el diagrama Q-Q (Cuantil-Cuantil) sin y con región de confianza del 95%.

En este apartado se va a usar la función `qqPlot()` del paquete "car", que hay que instalar previamente:

```
> install.packages("car")
```

```
> library(car)

> par(mfrow=c(1,2))
> qqnorm(notas.reales, main = "Q-Q plot")
> qqline(notas.reales, col="red")
> qqPlot(notas.reales, main = "Q-Q plot con región de confianza del 95%")
[1] 23 73
> par(mfrow=c(1,1))
```



Se observan puntos fuera de la región de confianza.

k) Repetir el apartado anterior, pero haciendo la comprobación mediante la regla empírica 68-95-99 de la distribución Normal

En una distribución Normal N(m,d) ,según la llamada “regla empírica” se cumple que en el intervalo:

- m±d se encuentra el 68.27% de la distribución
- m±2d se encuentra el 95.45% de la distribución
- m±3d se encuentra el 99.73% de la distribución

Se puede comprobar si se cumple en el caso de los datos de las notas reales:

```
> length(notas.reales[notas.reales<m+d & notas.reales>m-d])/74
[1] 0.7027027
> length(notas.reales[notas.reales<m+2*d & notas.reales>m-2*d])/74
[1] 0.972973
> length(notas.reales[notas.reales<m+3*d & notas.reales>m-3*d])/74
[1] 1
```

En todos los casos los valores obtenidos son parecidos a los de la regla empírica.

l) Repetir el apartado anterior haciendo la comprobación mediante los coeficientes de asimetría y apuntamiento.

```
> skewness(notas.reales)
[1] 0.5672277
> kurtosis(notas.reales)
[1] -0.2800343
```

Son valores muy cercanos a los de la Normal (0 en ambos casos) y están dentro del rango propuesto por algunos expertos, de un valor absoluto menor que 1.5 en la asimetría (skewness), y menor que 1 en el apuntamiento (kurtosis).

m) Repetir el apartado anterior pero haciendo la comprobación mediante test de hipótesis.

Se puede aplicar el test de Jarque-Bera.

```
> jarque.bera.test(notas.reales)

      Jarque Bera Test

data:  notas.reales
X-squared = 4.2631, df = 2, p-value = 0.1187
```

Según la tabla de p-valores para superar este test el p-valor debería ser mayor que 0.067 para 70 datos, y se comprueba que es superior.

Como el tamaño de la muestra es grande, también se puede aplicar el test de Test de Kolmogorov-Smirnov.

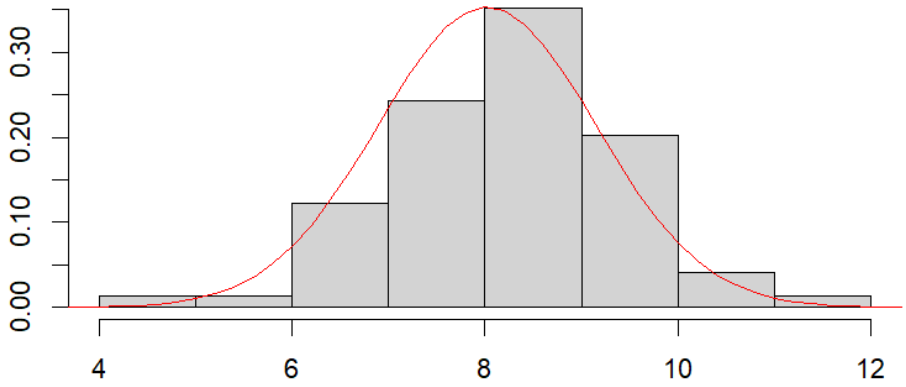
```
> ks.test(notas.reales, pnorm, mean(notas.reales), sd(notas.reales))

      Asymptotic one-sample Kolmogorov-Smirnov test

data:  notas.reales
D = 0.11842, p-value = 0.2505
alternative hypothesis: two-sided
```

El p-value es mayor que 0.05 por lo que se puede aceptar la hipótesis de que se ajusta a una distribución Normal.

m) Repetir los apartados h) a m) pero utilizando datos generados aleatoriamente con R, en lugar de los datos reales

Repetimos los comandos del apartado h), pero en este caso generamos aleatoriamente una muestra con 74 valores utilizando la función `rnorm()`.

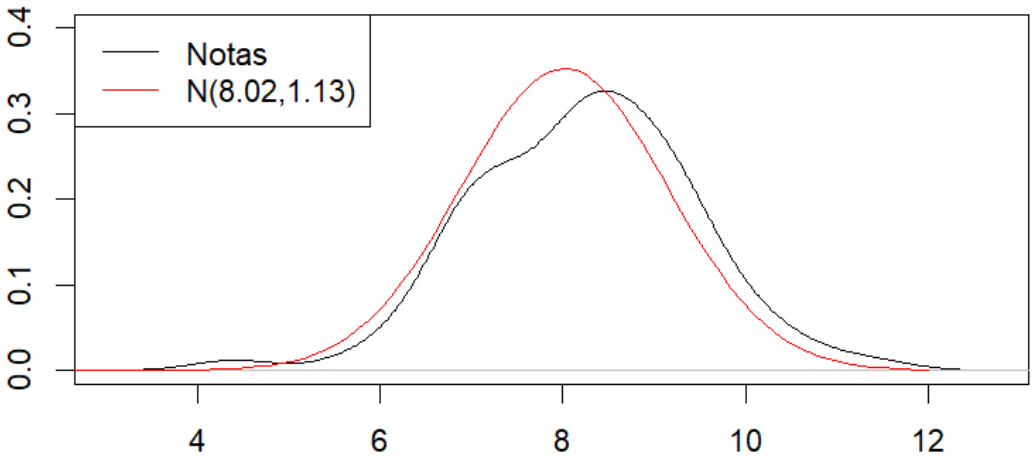
```
> notas.simuladas = rnorm(74,m,d)
> hist(notas.simuladas, freq=FALSE)
> curve(dnorm(x,m,d),0,20, add = TRUE)
```



Si se ejecuta la función `rnorm()` varias veces, en cada ocasión se obtendrá valores diferentes, pero en general puede comprobarse que se ajustan a la distribución Normal.

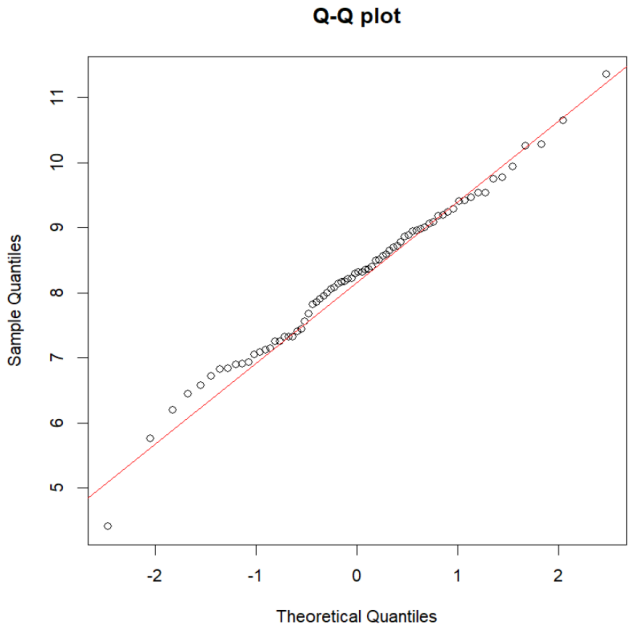
Usando la función de densidad como en i):

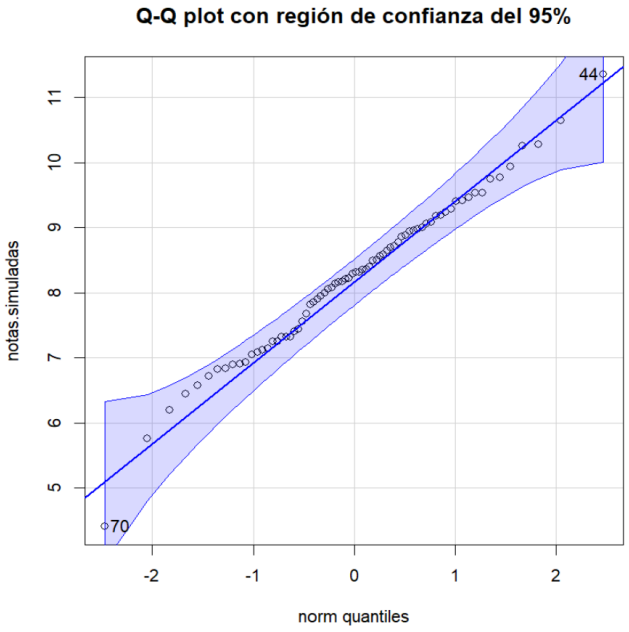
```
> plot(density(notas.simuladas))
> curve(dnorm(x,m,d),0,12, add=TRUE, col="red")
> legend("topleft",legend=c("Notas","N(8.02,1.13)"), col=c("black", "red"), lty=1)
```



Y con el gráfico Q-Q como en el j):

```
> par(mfrow=c(1,2))
> qqnorm(notas.simuladas, main = "Q-Q plot")
> qqline(notas.simuladas, col="red")
> qqPlot(notas.simuladas, main = "Q-Q plot con región de confianza del 95%")
[1] 15 59
> par(mfrow=c(1,1))
```





Comprobando la regla empírica 68-95-99:

```
> length(notas.simuladas[notas.simuladas<m+d & notas.simuladas>m-d])/74
[1] 0.6756757
> length(notas.simuladas[notas.simuladas<m+2*d & notas.simuladas>m-2*d])/74
[1] 0.9324324
> length(notas.simuladas[notas.simuladas<m+3*d & notas.simuladas>m-3*d])/74
[1] 0.9864865
```

Comprobando asimetría y apuntamiento:

```
> skewness(notas.simuladas)
[1] -0.2074239
> kurtosis(notas.simuladas)
[1] 0.5129791
```

Realizando el test de hipótesis de Jarque-Bera:

```
> jarque.bera.test(notas.simuladas)

        Jarque Bera Test

data:  notas.simuladas
X-squared = 1.6996, df = 2, p-value = 0.4275
```

Realizando el test de hipótesis de Kolmogorov-Smirnov:

```
> ks.test(notas.simuladas, pnorm, mean(notas.simuladas), sd(notas.simuladas))

        Exact one-sample Kolmogorov-Smirnov test
```

```
data:  notas.simuladas
D = 0.051189, p-value = 0.9848
alternative hypothesis: two-sided
```

Como p-value > 0.05 se puede afirmar con un 95% de confianza que se ajusta a una distribución Normal.

2.3 Aproximar una variable Binomial o de Poisson a una distribución Normal

1) Aproximar una variable Binomial a una distribución Normal

Si en una variable Binomial X:B(n,p) se cumple que n es grande[4], entonces se puede aproximar a una variable Normal X:N(m,d), con $m = np$ y $d = \sqrt{np(1-p)}$, y se puede calcular que la variable Binomial esté en el intervalo [a,b] calculando P{X≤b}-P{X<a} aplicando la Normal, pero en este caso sería P{X<b+0.5}-P{X<a-0.5}, porque hay que expandir el intervalo restando y sumando 0.5 en cada extremo debido a la conocida corrección de continuidad, debido a que la distribución Binomial es discreta y la Normal continua.

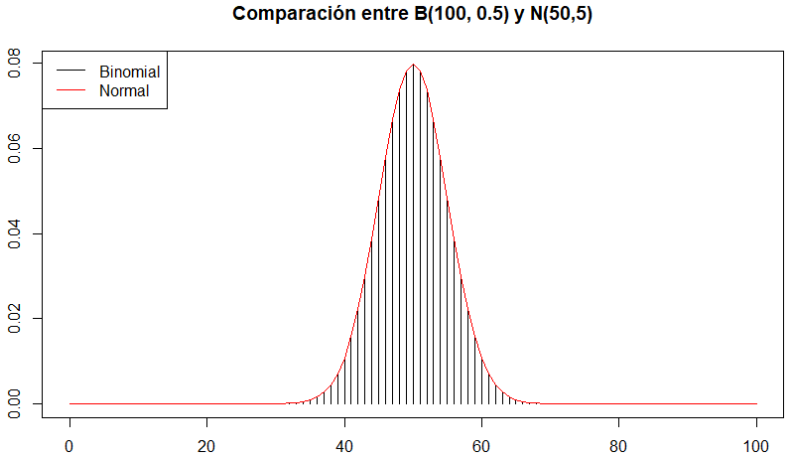
Se puede comprobar por ejemplo con B(100,0.5), aproximándola a una Normal N(50, 5).

```
> (n=100)
[1] 100
> (p=0.5)
[1] 0.5
> (m=n*p)
[1] 50
> (d=sqrt(n*p*(1-p)))
[1] 5
```

Podemos comparar gráficamente sus funciones de masa (Binomial) y densidad (Normal) de probabilidad:

```
> plot(0:n,dbinom(0:n,n,p), type="h", main="Comparación entre B(100, 0.5) y
N(50,5)")
> curve(dnorm(x,m,d),0,n, add = TRUE, col="red")
> legend("topleft", legend = c("Binomial", "Normal"), col = c("black", "red"),
lwd = 1)
```

[4] No hay unanimidad respecto a lo que se entiende por "grande" ni por la condición que debe cumplir p. Algunos autores establecen que la aproximación es buena cuando n≥30, np≥5 y n(1-p)≥5; otros autores establecen n>50 y p con un valor cercano a 0.5.



Y la probabilidad de que la variable Binomial esté, por ejemplo, en el intervalo [40,50], aplicando la función de distribución de la Normal:

```
> pnorm(50+0.5,m,d)-pnorm(40-0.5,m,d)
[1] 0.5219634
```

Que coincide en gran medida con el resultado que se obtendría aplicando la Binomial:

```
> pbinom(50,n,p)-pbinom(39,n,p)
[1] 0.5221945
```

2) Aproximar una variable de Poisson a una distribución Normal

Si en una variable de Poisson X:P(λ) se cumple que λ es grande[5], entonces se puede aproximar a una variable Normal X:N(m,d), con $m = \lambda$ y $d = \sqrt{\lambda}$, y se puede calcular que la variable de Poisson esté en el intervalo [a,b] calculando P{X≤b}-P{X<a} aplicando la Normal, pero en este caso sería P{X<b+0.5}-P{X<a-0.5}, porque hay que expandir el intervalo restando y sumando 0.5 en cada extremo debido a la conocida corrección de continuidad.

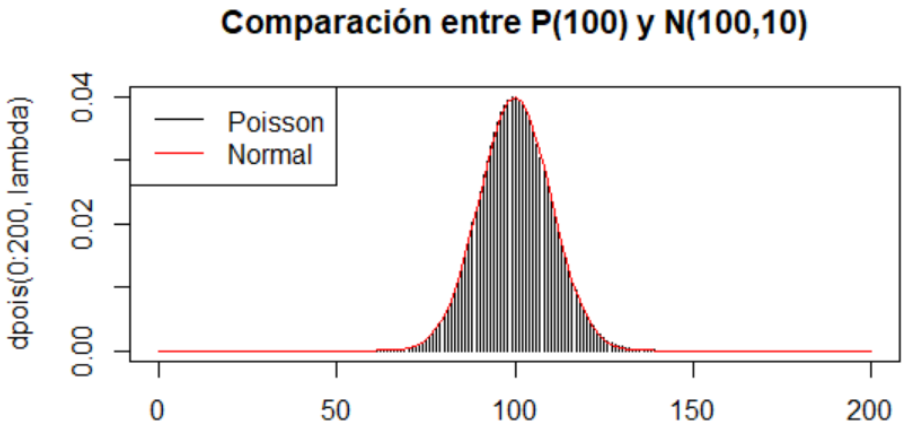
Se puede comprobar por ejemplo con P(100), aproximándola a una Normal N(100, 10).

```
> (lambda=100)
[1] 100
> (m=lambda)
[1] 100
> (d=sqrt(lambda))
[1] 10
```

[5] No hay unanimidad respecto a lo que se entiende por "grande". Algunos autores establecen que es suficiente λ>5, otros λ>10, otros λ>16.

Podemos comparar gráficamente sus funciones de masa (Poisson) y densidad (Normal) de probabilidad:

```
> plot(0:200,dpois(0:200,lambda), type="h", main="Comparación entre P(100) y
N(100,10)")
> curve(dnorm(x,m,d),0,200, add = TRUE, col="red")
> legend("topleft", legend = c("Poisson", "Normal"), col = c("black", "red"), lwd
= 1)
```



Y la probabilidad de que la variable de Poisson esté, por ejemplo, en el intervalo [80,90], aplicando la función de distribución de la Normal:

```
> pnorm(90+0.5,m,d)-pnorm(80-0.5,m,d)
[1] 0.1508739
```

Que coincide en gran medida con el resultado que se obtendría aplicando la de Poisson:

```
> ppois(90,lambda)-ppois(79,lambda)
[1] 0.1539338
```

3. Otras distribuciones continuas

En R se pueden utilizar otras distribuciones de probabilidad para variables aleatorias continuas, y para todas ellas existen las funciones de densidad, de distribución, de cálculo de cuantiles y de generación de valores aleatorios. Se trata de las siguientes:

Distribución	Función de densidad	Función de distribución	Cuantiles	Generación valores aleatorios
Beta	`dbeta()`	`pbeta()`	`qbeta()`	`rbeta()`
Cauchy	`dcauchy()`	`pcauchy()`	`qcauchy()`	`rcauchy()`
Chi-cuadrado	`dchisq()`	`pchisq()`	`qchisq()`	`rchisq()`
Exponencial	`dexp()`	`pexp()`	`qexp()`	`rexp()`
F	`df()`	`pf()`	`qf()`	`rf()`
Gamma	`dgamma()`	`pgamma()`	`qgamma()`	`rgamma()`
Gamma inversa	`dIGAMMA()`	`pIGAMMA()`	`qIGAMMA()`	`rIGAMMA()`
Log-normal	`dlnorm()`	`plnorm()`	`qlnorm()`	`rlnorm()`
T de Student	`dt()`	`pt()`	`qt()`	`rt()`
Uniforme	`dunif()`	`punif()`	`qunif()`	`runif()`
Weibull	`dweibull()`	`pweibull()`	`qweibull()`	`rweibull()`
Wilcoxon	`dwilcox()`	`pwilcox()`	`qwilcox()`	`rwilcox()`

Para ver la ayuda sobre las distribuciones en R se puede ejecutar el comando `help("Distributions")`.

Se pueden encontrar ejemplos en R en https://fhernanb.github.io/Manual-de-R/continuas.html. Y los fundamentos teóricos de éstas y otras distribuciones de variables contínuas en https://es.m.wikipedia.org/wiki/Distribución_de_probabilidad.

Se puede comprobar a qué distribución se aproxima más un conjunto de datos reales, usando la función `fitDist()` del paquete `gamlss`:

```
> install.packages("gamlss")
> library(gamlss)
```

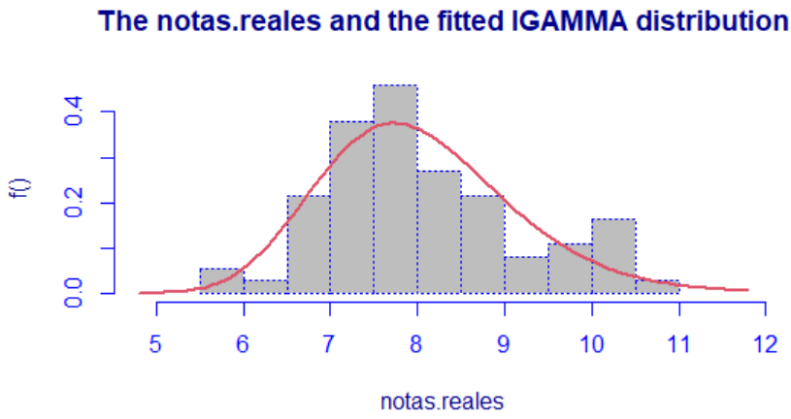
Por ejemplo, en el caso de los datos de las calificaciones de los 74 estudiantes del apartado 2.3, que estaban en el vector `notas.reales` sería:

```
> fitDist(notas.reales)
....
Family:  c("IGAMMA", "Inverse Gamma")
```

Al final del listado que se obtiene aparece que es la distribución “Gamma Inversa” a la que mejor se adapta el conjunto de las notas.

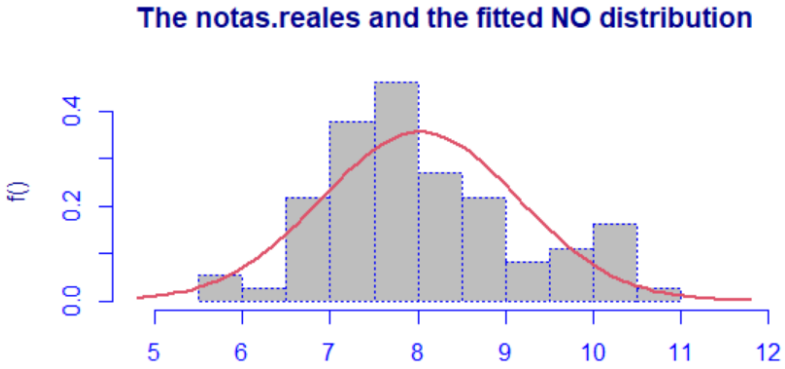
En el paquete `gamlss` también existe la función `histDist()` que se puede utilizar para dibujar el histograma de los datos junto a una función de densidad de la distribución, en este caso podemos indicar Gamma Inversa (IGAMMA):

```
> histDist(notas.reales, family=IGAMMA)

Family:  c("IGAMMA", "Inverse Gamma")
Fitting method: "nlminb"

Call:  gamlssML(formula = notas.reales, family = "IGAMMA")

Mu Coefficients:
[1]  2.045
Sigma Coefficients:
[1]  -1.993

 Degrees of Freedom for the fit: 2 Residual Deg. of Freedom   72
Global Deviance:     222.226
            AIC:     226.226
            SBC:     230.834
```



Podemos ver el mismo diagrama, pero comparando con la Normal (NO):

```
> histDist(notas.reales, family=NO)
```



3.1 Variables continuas con distribución general

Si dada una variable aleatoria continua de la que disponemos de un conjunto de valores, no encontramos ninguna distribución conocida (Normal, Beta, Gamma, etc.) a la que se ajuste, entonces se pueden utilizar las funciones `density()` y `ecdf()` para calcular empíricamente probabilidades relacionadas con variable.

Toda variable aleatoria continua X tiene asociada dos funciones:

- La función de densidad de probabilidad f(x): que representa la probabilidad de que la variable tenga un valor concreto: P{X=x}. En el caso de las variables continuas, esta función sólo se utiliza a efectos teóricos y de representación gráfica, porque la probabilidad de la variable en un punto no puede calcularse, es cero, y siempre hay que calcular la probabilidad en un intervalo. En el caso de R se puede asociar a la función `density()`.
- La función de distribución de probabilidad F(x), que representa la probabilidad de que la variable sea igual a inferior a un valor concreto: P{X≤x}. En una variable continua, al no tener sentido hablar de probabilidad de un valor, en este caso se cumple que P{X≤x}=P{X<x}. En el caso de R se puede asociar a la función `ecdf()`.

Si disponemos de una muestra de valores de la variable, con R se pueden manejar ambas funciones como se indica en la siguiente tabla.

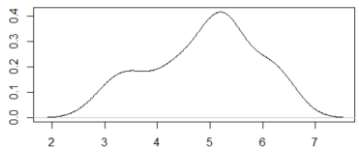
Función	Código R	Resultado	Ejemplo de uso

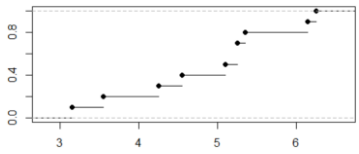
Densidad de probabilidad f(x)	`x = vector con la muestra de valores conocidos de la variable X`
	
	`f=density(x)`
		Al ser una variable continua, en teoría el resultado siempre debería ser 0, sólo debe usarse para dibujar la función de densidad.	`No debe usarse para calcular probabilidades, sólo para dibujar.`

Distribución de probabilidad F(x)=P{X≤x}= P{X<x}	`x = vector con la muestra de valores conocidos de la variable X`
	
	`F=ecdf(x)`
		Calcula la probabilidad de que la variable aleatoria X tenga un valor igual o menor que x.
		El nombre de la función proviene de “empirical cumulative distribution function”.
			`>x=c(5.1, 5.2, 4.5, 6.2, 4.2, 5.3, 5.2, 3.1, 3.5, 6.1)`
			
			`>F=ecdf(x)`
			
			`>F(4.8)`
			`[1] 0.4`

Las funciones anteriores se pueden combinar con otras de generación de gráficos, para obtener los diagramas que se indican en la siguiente tabla.

Diagrama	Código R	Ejemplo de uso	

Función de densidad de probabilidad	`plot(density(x))`	`>x=c(5.1, 5.2, 4.5, 6.2, 4.2, 5.3, 5.2, 3.1, 3.5, 6.1)`
		
		`>plot(density(x))`
			

Función de distribución de probabilidad	`plot(ecdf(x))`	`>x=c(5.1, 5.2, 4.5, 6.2, 4.2, 5.3, 5.2, 3.1, 3.5, 6.1)`
		
		`>plot(ecdf(x))`
			

4. Ejercicios propuestos

1) Mediante una encuesta cuyas respuestas están disponibles en el archivo encuesta.csv, y después de eliminar las respuestas de las personas que no contestaron alguna pregunta, se sabe el tiempo en minutos del viaje a la Escuela Politécnica de 74 estudiantes de un curso de la asignatura Estadística del Grado en Ingeniería en Sistemas de Información de la Universidad de Alcalá. Si se supone que tienen una distribución Normal, responder a las siguientes preguntas:
    a) Definir una variable aleatoria Normal para el tiempo del viaje de un estudiante
    b) Calcular la media o esperanza y la varianza de la variable aleatoria
    c) Dibujar las funciones de densidad y distribución de la variable aleatoria
    d) Cuál es la probabilidad de que un estudiante seleccionado al azar tarde menos de 1 hora en llegar a la escuela. Dibujar en el diagrama de la función de densidad el área correspondiente a la acumulación de la probabilidad obtenida.
    e) Cuál es la probabilidad de que un estudiante seleccionado al azar tarde entre 1 hora y 1 hora y media. Dibujar en el diagrama de la función de densidad el área correspondiente a la acumulación de la probabilidad obtenida.
    f) Cuál es la probabilidad de que un estudiante seleccionado al azar tarde más de 1 hora y media. Dibujar en el diagrama de la función de densidad el área correspondiente a la acumulación de la probabilidad obtenida.
    g) Cuál es el valor de los cuartiles de la variable aleatoria. Dibujar en el diagrama de la función de densidad el área correspondiente al segundo cuartil.
    h) Si tenemos los datos de la encuesta original a los 74 estudiantes, comprobar si tomando los tiempos reales, el histograma se ajusta realmente a la distribución Normal.
    i) Repetir el apartado anterior pero haciendo la comprobación utilizando el diagrama de la función de densidad generada con `density()`.
    j) Repetir el apartado anterior pero haciendo la comprobación utilizando el diagrama Q-Q (Cuantil-Cuantil) sin y con región de confianza del 95%.
    k) Repetir el apartado anterior pero haciendo la comprobación mediante la regla empírica 68-95-99 de la distribución Normal.
    l) Repetir el apartado anterior pero haciendo la comprobación mediante test de hipótesis.
    m) Repetir los apartados h) a l) pero utilizando datos generados aleatoriamente con R, en lugar de los datos reales.

2) Resolver el ejercicio 1) de la práctica P5 (parte 1) aproximando la distribución Binomial B(50,1/6) a una distribución Normal y comparar los resultados.
    a) Definir una variable aleatoria Normal aproximando a la Binomial
    b) Calcular la esperanza y varianza de la variable aleatoria
    c) Dibujar las funciones de probabilidad y distribución de la variable aleatoria
    d) Calcular la probabilidad de que nunca aparezca el lado con el valor 6.
    e) Calcular la probabilidad de que siempre aparezca el lado con el valor 6.
    f) Calcular la probabilidad de que el lado con el valor 6 aparezca menos de 10 veces.
    g) Calcular la probabilidad de que el lado con el valor 6 aparezca más de 10 veces.
    h) Calcular la probabilidad de que el lado con el valor 6 aparezca entre 20 y 40 veces.
    i) Calcular los cuantiles 0.50 y 0.90.
    j) Generar aleatoriamente 100 valores para la variable y comprobar si el histograma se ajusta realmente a la función de densidad de la Normal.

3) Resolver el ejercicio 3) de la práctica P5 (parte 1) aproximando la distribución de Poisson P(120) a una distribución Normal y comparar los resultados.
    a) Definir una variable Normal aproximando a la de Poisson para el número de visitas en una hora
    b) Calcular la esperanza y varianza de la variable aleatoria
    c) Dibujar las funciones de probabilidad y distribución de la variable aleatoria
    d) Cuál es la probabilidad de que haya 4 visitas en una hora
    e) Cuál es la probabilidad de que haya 4 o menos visitas en una hora
    f) Cuál es la probabilidad de que haya más de 4 visitas en una hora
    g) Cuál es la probabilidad de que no haya ninguna visita en una hora
    h) Cuál es la probabilidad de que haya entre 4 y 6 visitas en una hora
    i) Cuál es el valor de los cuartiles de la variable aleatoria
    j) Generar aleatoriamente 100 valores para la variable y comprobar si el histograma se ajusta realmente a la función de densidad de Poisson.
    k) Definir una nueva variable aleatoria para calcular la probabilidad de que en un día haya 100 visitas.
    l) Generar aleatoriamente 100 valores para la nueva variable y comprobar si el histograma se ajusta realmente a la función de densidad de Poisson.