Práctica 2. Distribuciones Comunes

Grado de Ingeniería Civil - Universidad de Cantabria

1. Distribuciones de probabilidad en R

En teoría hemos visto fórmulas analíticas para las funciones de probabilidad ($p_X(x)$) o densidad ($f_X(x)$) de varios modelos de distribuciones comunes. Hay distribuciones para las que la función de distribución ($F_X(x)$), a pesar de ser fácilmente calculable mediante un sumatorio o una integral, resulta demasiado laboriosa como para realizarla a mano o con una calculadora de bolsillo. El paquete básico de R proporciona funciones de probabilidad, densidad y distribución para todas las variables aleatorias comunes vistas en clase y alguna más.

La siguiente tabla muestra en negrita las distribuciones vistas en clase, junto con su nombre en R y los parámetros que se pueden pasar a las funciones.

Distribución	**Nombre en R**	**Argumentos**
beta	beta	shape1, shape2, ncp
binomial	binom	size, prob
Cauchy	cauchy	location, scale
chi-cuadrado	chisq	df, ncp
exponencial	exp	rate
F de Snedecor	f	df1, df2, ncp
gamma	gamma	shape, scale
geométrica	geom	prob
hipergeométrica	hyper	m, n, k
log-normal	lnorm	meanlog, sdlog
logística	logis	location, scale
binomial negativa	nbinom	size, prob
normal	norm	mean, sd
Poisson	pois	lambda
t de Student	t	df, ncp
uniforme	unif	min, max
Weibull	weibull	shape, scale
Wilcoxon	wilcox	m, n

Para cada una de estas variables aleatorias, R proporciona una función de probabilidad o de densidad, según corresponda, añadiendo el prefijo `d` en ambos casos. En el caso de la función de densidad, habrá que proporcionar, como mínimo, el valor de la variable para el que se quiere calcular la probabilidad o densidad de probabilidad. A continuación habrá que pasarle los parámetros de la distribución, de una forma muy parecida a la notación que usamos en clase. De la misma forma, mediante el prefijo `p`, obtenemos la función de distribución.

Como se observa en la tabla, los parámetros de entrada son distintos para cada distribución por lo que se recomienda consultar la ayuda en cada caso. Por ejemplo `?pbinom` nos indica el uso de esta función:

```
pbinom(q,size, prob)
```

donde `q` representa el valor de la variable aleatoria para el que se quiere calcular la función de distribución, `size` es el tamaño de la muestra (n) y `prob` es la probabilidad de éxito (p).

Además todas estas distribuciones están recogidas en el menú de R-Comander y se puede acceder a ellas a través del menú `Distribuciones` ⇒ `Distribuciones discretas/continuas`.

Para entender mejor como calcular las distintas opciones de **R** vamos a resolver un ejercicio tipo como los de los ejemplos resueltos en clase. De esta manera podemos usar R para comprobar los resultados que obtengamos aplicando las fórmulas vistas en clase.

Un lote de procesadores de ordenador contiene un 40 % de unidades defectuosas. Cuando se detecta el error ya han sido instalados todos los procesadores del lote en los ordenadores. Por esta razón el responsable del departamento de calidad decide realizar una revisión de algunos ordenadores que incorporan un procesador de dicho lote. Considerará que un ordenador es defectuoso si lo es el procesador.

1. Calcular la probabilidad de que menos de 4 equipos lleven instalado un procesador defectuoso de un total de 20 ordenadores revisados.
   La variable aleatoria X represente el número de ordenadores defectuosos y sigue una distribución B(n=20, p=0.4), luego la probabilidad que nos piden $P(X \leq 3) = F_{B(20,0,4)}(3)$
   ```
   > pbinom(3,20,0.4)
   [1] 0.01596116
   ```
   Nótese que, el primer argumento de la función es el valor de la variable aleatoria en el que queremos evaluar la función de distribución y, a continuación, los parámetros necesarios en el orden en que aparecen en la tabla anterior.

2. ¿Y la probabilidad de que más de 3 sean defectuosos?
   Nuestra variable aleatoria X, número de ordenadores defectuosos, sigue comportándose como una distribución B(n=20, p=0.4) y la probabilidad que debemos calcular en este caso será $P(X > 3) = 1 - F_{B(20,0,4)}(3)$ que vendrá dada por:
   ```
   > 1-pbinom(3,20,0.4)
   [1] 0.9840388
   ```

3. ¿Y la probabilidad de que 5 sean defectuosos?
   Nuestra variable aleatoria X, número de ordenadores defectuosos, sigue comportándose como una distribución B(n=20, p=0.4), sin embargo en este caso debemos calcular $P(X = 5) = p_{B(20,0,4)}(5)$ que vendrá dada por:
   ```
   > dbinom(5,20,0.4)
   [1] 0.07464702
   ```

4. Calcular la probabilidad de encontrar el primer ordenador defectuoso al examinar el séptimo equipo.
   En este caso la variable aleatoria X, número de ordenadores revisados hasta econtrar el primero defectuoso, sigue una distribución G(p=0.4) y la probabilidad que debemos calcular será $P(X = 7) = p_X(7)$ que calculada con la formulación fista en clase resulta ser 0.0186624.

   Como vereis en la ayuda de la distribución geométrica (`?dgeom`), **R** define de forma distinta la función de probabilidad de la distribución geométrica a como la hemos visto en clase, considerando la variable aleatoria como el número de fracasos (k) en lugar del número de intentos ($x = k + 1$):

$$p_K(k) = p(1 - p)^k, \quad k \geq 0.$$

De acuerdo con esta definición si k es la variable aleatoria número de fracasos, en **R** la probabilidad que nos piden calcular ($P(k = 6)$) vendrá dada por:

```
> dgeom(6,0.4)
```

5. Calcular la probabilidad de que el décimo ordenador revisado sea el cuarto no defectuoso.
En este caso la variable aleatoria X, será el número de ordenadores revisados hasta econtrar el décimo ordenador defectuoso que sigue una distribución BN(r=10,p=0.6) y la probabilidad que debemos calcular será $P(X = 10) = p_X(10)$ que calculada con la formulación vista en teoría resulta ser 0.0446.

   Si consultais la ayuda de la distribución Binomial Negativa que ofrece **R** (`?dnbinom`), observareis que esta distribución también considera como variable aleatoria k el número de fracasos antes de obtener un número r de éxitos (es decir, $x = k + r$). Esta variable sigue también una distribución binomial negativa cuya función de probabilidad es de la forma:

$$p_K(k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k, \quad k \geq 0$$

   De acuerdo con esta definición, en **R** la probabilidad que nos piden calcular vendrá dada por:

```
> dnbinom(6,4,0.6)
```

Para cada una de las variables aletorias indicadas, se puede obtener la función cuantil ($F_X^{-1}(p)$) mediante el prefijo `q`. Tomando como ejemplo el primer apartado calculado antes, en este caso para obtener el menor valor de c tal que $P(X \leq c) > 0{,}01596116$ dado que $X \sim B(20, 0{,}4)$:

```
> qbinom(0.01596116,20,0.4)
[1] 3
```

El primer cuartil de esa misma distribución vendría dado por:

```
> qbinom(0.25,20,0.4)
[1] 6
```

Nótese que, en este caso, el primer argumento es la probabilidad y a continuación se colocan los argumentos en el orden indicado en la tabla.

Por último, el prefijo `r` nos permite obtener una muesta aletoria de un cierto tamaño extraida de una variable aleatoria. Es decir, nos permite obtener valores concretos x de una variable aleatoria X. O, dicho de otra forma, simular un experimento que siga una cierta distribución. Con las funciones de distribución comunes podemos simular experimentos con variables contínuas y discretas con un campo de existencia infinito numerable. Por ejemplo una muestra aleatoria tomada de una distribución binomial $B(20, 0{,}5)$:

```
> rbinom(10, 20, 0.5)
 [1] 11 14  9 11  9  9 11 12 13 11
```

Estos serían 10 posibles resultados de un experimento binomial en el que se realizan 20 experimentos de Bernoulli con probabilidad de éxito 0.5 y se suma el número de éxitos en cada uno de ellos.

De la misma forma se pueden obtener muestras de cualquier otra distribución:

```
> rnorm(15)          # Muestra de tamaño 15 de N(0,1)
 [1]  1.31388200  0.22642715 -0.17006440  1.14187804  0.03307199 -0.74012294
 [7]  0.65993043  1.02097908  0.93444093  0.21143781 -0.79784992 -1.28328732

[13]  1.95319297 -0.90321425 -0.64991528
> rnorm(10, 165, 5)  # Muestra de tamaño 10 de N(165,5)
 [1] 155.6231 168.6797 167.6966 159.9683 153.0675 168.1637 164.1424 162.2332
 [9] 162.2115 175.4075
> rpois(10, 5)       # Muestra de tamaño 10 de Po(5)
 [1] 5 5 2 7 6 6 2 9 5 3
```

2. Representando funciones de distribución

Como ya se ha mencionado, todo lo anterior también se puede obtener utilizando los menús de R-commander. Éstos proporcionan, además, la opción de representar de forma sencilla la forma de las funciones de probabilidad, densidad y distribución. Podéis probar a utilizar las entradas del menú *Distribuciones* de R-commander.

Otra forma de dibujar las funciones de distribución (o cualquier otra función) en R es mediante la función `curve`. Por ejemplo, para dibujar la función de densidad normal tipificada entre -4 y 4, utilizaremos:

```
> curve(dnorm(x), -4, 4)
```

El nombre x que colocamos como primer argumento lo utiliza la función `curve` y lo interpreta como la variable independiente a representar. Podemos mejorar un poco el aspecto utilizando los argumentos habituales en funciónes de representación. Por ejemplo, etiquetar los ejes (con `xlab` e `ylab`), cambiar el color (`col`) o la anchura (`lwd`) de la línea o ponerle un título (`main`) al gráfico:

```
> curve(dnorm(x), -4, 4, col="red", lwd=3, ylab="Funcion de densidad N(0,1)")
```

Podemos añadir (`add=TRUE`) otra curva al mismo gráfico:

```
> curve(dt(x, 4), col="blue", lwd=3, add=TRUE)
```

En este caso hemos añadido una gráfica de la función de densidad de la variable t de Student con 4 grados de libertad. Podemos distinguirlas mediante una leyenda:

```
> legend(-4, 0.4, legend=c("N(0,1)", "t(4)"), col=c("red","blue"), lwd=3)
```

Los dos primeros argumentos de esta función dan la posición de la esquina superior izquierda de la leyenda.

Cambiando el valor de los grados de libertad, se puede ver como la variable t de Student tiende a la Normal tipificada a medida que aumentamos los grados de libertad.

Para dibujar una función de probabibilidad discreta podemos utilizar la función `plot`. Por ejemplo, para ver que la binomial se parece a la normal cuando n es suficientemente grande y la p no está próxima a 0 ni a 1:

```
> n <- 30
> p <- 0.5
> media <- n*p
> desv <- sqrt(n*p*(1-p))
> x.i <- 5:25
> plot(x.i, dbinom(x.i,n,p), type="h", lwd=4)
> curve(dnorm(x,media,desv), col="red", add=TRUE)
```

Se ha utilizado el argumento `type="h"` para que la función `plot` dibuje impulsos en cada valor en lugar de unir los puntos con líneas. Si se modifican los valores de `n` y `p` será necesario modificar los valores `x.i`

Practica tu mismo:

2.1 Dada una v.a. X binomial $B(200; 0{,}4)$. Se pide:

1. $P(X < 71)$ $\quad$ $P(X \geq 90)$ $\quad$ $F_X(100)$
2. $P(X \leq x) = 0{,}4$ $\quad$ $P(X \geq x) = 0{,}8$ $\quad$ $P(X = 60)$

2.2 Si lanzamos un dado 288 veces, calcúlese la probabilidad de obtener 5 ó 6 puntos más de 90 veces y menos de 120.
Comprobar el valor obtenido mediante la representación de la función de distribución asociada a este experimento.

2.3 Un ingeniero tiene una probabilidad 0.02 de cometer un error de consideración al realizar un proyecto. Determinar:

1. la probabilidad de cometer 2 errores si realiza 20 proyectos.
2. la probabilidad de cometer el primer error al realizar el sexto proyecto.
3. Si cometer 10 errores supone su expulsión de la empresa, calcular la probabilidad de ser expulsado tras realizar el proyecto número 15.

2.4 La centralita telefónica de un hotel recibe un número de llamadas por minuto que sigue una ley de Poisson con parámetro $\lambda = 0{,}5$. Determinar la probabilidad de que en un minuto al azar:

1. Se reciba una única llamada.
2. Se reciban un máximo de dos llamadas.
3. La centralita quede bloqueada, sabiendo que no puede realizar más de 3 conexiones por minuto.

2.5 Una v.a. X se distribuye uniformemente en el intervalo $(2,4)$. Se pide:

1. $P(X < 2{,}5)$ $P(X \geq 3{,}2)$ $P(2{,}2 < X < 3{,}5)$
2. $P(X \leq x) = 0{,}4$ $P(X \geq x) = 0{,}8$ $F_X(2{,}7)$

2.6 La estatura de 1000 estudiantes de bachillerato está distribuida normalmente con una media de 168cm y una desviación típica de 5cm. Calcúlese:

1. Entre qué valores en torno a la media se encontrará el 95 % de los alumnos.
2. Entre qué valores en torno a la media se encontrará el 50 % de los alumnos.
3. A partir de qué altura se encontrará el 15 % de los alumnos más cualificados.
4. A partir de qué altura se encontrá el 5 % de los alumnos más cualificados.