**Practica 3**

# Simulació de variables aleatòries

En aquesta pràctica veurem com simular variables aleatòries amb distribucions conegudes i no per tal de generar mostres, calcular probabilitats i representar gràficament les variables. Considerarem tan el cas discret com el cas absolutament continu. També veurem com comprovar uns dels més importants teoremes de l'anàlisis estocàstica.

## 1 Simulacions de variables conegudes

`R` té incorporades algunes funcions per tractar amb les distribucions conegudes. A la següent taula trobeu els nom que el programa fa servir per les variables aleatòries més utilitzades.

| Distribució | Sufix del nom | Paràmetres addicionals |
|---|---|---|
| binomial | `binom` | `size, prob` |
| geomètrica | `geom` | `prob` |
| hypergeomètrica | `hyper` | `m, n, k` |
| binomial negativa | `nbinom` | `size, prob` |
| Poisson | `pois` | `lambda` |
| uniforme | `unif` | `min, max` |
| exponencial | `exp` | `rate` |
| normal | `norm` | `mean, sd` |
| log-normal | `lnorm` | `meanlog, sdlog` |
| t de Student | `t` | `df, ncp` |
| Khi quadrat | `chisq` | `df, ncp` |
| gamma | `gamma` | `shape, scale` |
| beta | `beta` | `shape1, shape2, ncp` |
| Cauchy | `cauchy` | `location, scale` |

Taula 1: Noms en R de distribucions de probabilitat

Per cadascuna de les distribucions podem cridar quatre funcions diferents afegint al nom el prefix `r`, `d`, `p` o `q`:

`rnom(valors, paràmetres)` → Nombres aleatoris segons la distribució de probabilitat donada.
`dnom(valors, paràmetres)` → Funció de densitat (o massa) de probabilitat.
`pnom(valors, paràmetres)` → Funció de distribució de probabilitat (Fdd).
`qnom(valors, paràmetres)` → Funció quantila, és a dir, la (pseudo)inversa de la Fdd.

Aquestes funcions ens són útils per simular variables aleatòries, per dibuixar la funció de densitat (o massa) de probabilitat i la funció de distribució i per calcular probabilitats.

Anem a veure com treballar en `R` amb variables aleatòries conegudes. Veurem alguns exemples per $X \sim Binom(8, 0.4)$ i per $Y \sim N(40, 3)$ (Normal de mitjana 40 i desviació típica 3).

### Simulació de nombres aleatoris

Tan pel cas discret com pel cas absolutament continu, per simular una mostra d'una variable aleatòria coneguda d'una mida fixada utilitzem la funció `r` seguida del nom de la variable.

Per exemple, per simular una mostra de mida $n = 80$ de $X \sim Binom(8, 0.4)$ utilitzem

```
x <- rbinom(80, 8, 0.4)
```

Per simular una mostra de mida $n = 1000$ de $Y \sim N(40, 3)$ utilitzem:

```
y <- rnorm(1000, 40, 3)
```

### Representació de la funció de massa de probabilitat i de la funció de densitat

Per dibuixar la funció de massa de probabilitat i la funció de densitat cal definim primer un vector amb els valors de la variable i un vector amb les probabilitats.

Per la variables $X$:

```
x <- c(0:8)
prob <- dbinom(x,8,0.4)
plot(x, prob, type = "h", xlim = c(0,8), ylim = c(0,1))
```

Per la variables $Y$:

```
t <- seq(20, 60, by = 0.05)
y <- dnorm(t, 40, 3)
plot(t, y, type = "l")
```

Evidentment els límits de `xlim` cal ajustar-los als valors de la variable.

### Representació de la funció de distribució

Pel cas discret, per dibuixar la funció de distribució hem de seguir les següents instruccions:

```
acum <- cumsum(prob)
s <- stepfun(x, c(0,acum))
plot(s, verticals = FALSE)
```

Pel cas absolutament continu, cal utilitzar el prefix `p` seguit del nom de la variable. Per exemple, per dibuixar la distribució de $Y$, les instruccions que hem de seguir són:

```
t <- seq(20, 60, by = 0.05)
y <- pnorm(t, 40, 3)
plot(t, y, type = "l")
```

### Càlcul de probabilitats

Els prefixos `d`, `p` i `q` seguits del nom de la variable ens ajuden a calcular probabilitats concretes.

Pel cas discret:

- $P(X = 3)$ es calcula mitjançant

  ```
  dbinom(3, 8, 0.4)
  ```

- $P(X \leq 3)$ es calcula mitjançant

  ```
  pbinom(3, 8, 0.4)
  ```

Comproveu que $P(X \leq 3)$ també es pot calcular fent

```
dbinom(0, 8, 0.4) + dbinom(1, 8, 0.4) + dbinom(2, 8, 0.4)
    + dbinom(3, 8, 0.4)
```

- $P(X \geq 3) = 1 - P(X < 3) = 1 - P(X \leq 2)$ es calcula mitjançant

```
1 - pbinom(2, 8, 0.4)
```

- El prefix `q` permet trobar el valor de la variable en que la funció de distribució assoleix una probabilitat concreta. Per exemple, per saber el valor $k$ en que $P(X \leq k) = 0.59$ fem:

```
k <- qbinom(0.59, 8, 0.4)
```

Pel cas absolutament continu:

- $P(Y = 40)$: Recordeu que aquesta probabilitat sempre és zero ja que $Y$ és contínua.
- $P(Y \leq 45)$ es calcula mitjançant

```
pnorm(45, 40, 3)
```

- $P(Y \geq 35) = 1 - P(Y < 35)$ es calcula mitjançant

```
1 - pnorm(35, 40, 3)
```

- Per trobar $y$ tal que $P(Y \leq y) = 0.8$ fem

```
qnorm(0.8, 40, 3)
```

**Exercici 1.** Sigui $X$ una Poisson amb paràmetre $\lambda = 2$.

a) Dibuixeu la funció de probabilitat i la funció de distribució de $X$.

b) Calculeu les següents probabilitats

$$P(X = 6), \qquad P(X > 2), \qquad P(1 \leq X \leq 4).$$

c) Genereu 400 nombres aleatoris amb la mateixa distribució que $X$ i poseu-los en forma de matriu amb 40 files i 10 columnes. Calculeu la mitjana i la variància de cada fila. S'aproximen aquests valors als teòrics?
*Indicació: La funció* `apply` *ens permet fer càlculs paral·lels (per files o per columnes).*

**Exercici 2.** Considereu una variable aleatòria amb llei $U(-2, 2)$.

a) Dibuixeu-ne la funció de densitat i la funció de distribució.

b) Calculeu les probabilitats següents:

| | | |
|---|---|---|
| $P(X = 0.5)$ | $P(X \leq 0.7)$ | $P(X < 0.7)$ |
| $P(X \geq -1.2)$ | $P(X \leq -2)$ | $P(X < 2)$ |

# 2 Simulacions de variables no conegudes

## 2.1 Variables aleatòries discretes

Volem generar ara una mostra de mida $n$ d'una variable aleatòria $X$ discreta que pren valors $a_1, a_2, ..., a_n$ amb probabilitats $p_1, p_2, ..., p_n$ respectivament. És a dir, $P(X = a_i) = p_i$ per $i = 1, ..., n$.

El primer que fem és generar $n$ valors d'una variable uniforme $(0, 1)$: `runif(n)`, aquests valors que acabem de generar seran les probabilitats de manera que el segon pas serà convertir aquests valors en els valors $a_1, \ldots a_n$:

| **valors** | **convertir** |
|---|---|
| entre 0 i $p_1$ | $a_1$ |
| entre $p_1$ i $p_1 + p_2$ | $a_2$ |
| ... | |
| entre $p_1 + \ldots + p_{n-1}$ i 1 | $a_n$ |

El mètode per fer la simulació consisteix en:

- Obtenir nombres aleatoris $u_1, u_2, ..., u_n$

```
x <- runif(n)
```

- Fer la transformació següent: donat $u_i$, existirà un $k$ tal que $u_i \in (p_1 + ... + p_{k-1}, p_1 + ... + p_k]$. Aleshores, caldrà transformar $u_i$ en $a_k$.

```
x1 <- (x<=p_1)
x2 <- (x>p_1 & x<=p_1+p_2)
x3 <- (x>p_1+p_2 & x<=p_1+p_2+p_3)
...
xn <- (x>p_1+p_2+...+p_{n-1})
y <- a_1*x1+...+a_n*x_n
```

  La variable $y$ és la que contindrà els $n$ valors generats amb la funció de probabilitat donada.

**Exemple 1.** Volem generar una mostra de mida 10 d'una variable aleatòria $X$ tal que

$$P(X = 5) = \frac{1}{3}, \qquad P(X = 7) = \frac{1}{2}, \qquad P(X = 9) = \frac{1}{6}.$$

Seguint el mètode per la simulació, primer obtenim els nombres aleatoris:

```
u <- runif(10)
u
[1] 0.3405060 0.2556905 0.1379111 0.5610426 0.2238937
0.1841603 0.3919349 0.9773482 0.4329352 0.4750485
```

Després fem la transformació:

| **valors** | **convertir a** |
|---|---|
| entre 0 i $\frac{1}{3}$ | 5 |
| entre $\frac{1}{3}$ i $\frac{1}{3} + \frac{1}{2} = \frac{5}{6}$ | 7 |
| entre $\frac{5}{6}$ i 1 | 9 |

```
x1 <- (u<=1/3)
x1
[1] FALSE TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE FALSE

x2 <- (u>1/3 & u<=1/3+1/2)
x2
[1]  TRUE FALSE FALSE TRUE FALSE FALSE TRUE FALSE TRUE TRUE

x3 <- (u>1/3+1/2)
x3
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE

x <- 5*x1+7*x2+9*x3
x
[1] 7 5 5 7 5 5 7 9 7 7
```

El programari `R` té incorporada la funció `sample` que permet generar variables aleatòries discretes amb una sola instrucció.

**Continuació de l'Exemple 1.** Podem repetir l'exemple que acabem de resoldre utilitzant:

```
sample(c(5,7,9), 10, prob = c(1/3,1/2,1/6), replace = TRUE)
```

El càlcul de probabilitats i la representació de la funció de densitat i de la funció de distribució d'una variable no coneguda es poden obtenir fàcilment amb el vector de probabilitats i el corresponent vector de probabilitats acumulades.

**Exercici 3.** Considereu una variable aleatòria $X$ que té la funció de massa de probabilitat següent:

| Valors $x_i$ de $X$ | 3 | 5 | 7 | 9 |
|---|---|---|---|---|
| Probabilitats $p_i = P(X = x_i)$ | 0.1 | 0.3 | 0.2 | 0.4 |

a) Representeu la funció de massa de probabilitat i la corresponent funció de distribució.

b) Genereu una mostra de mida $N = 1000$ de nombres aleatoris amb la llei de probabilitat de $X$.

c) Calculeu les freqüències relatives de la mostra. Compareu-les amb les corresponents probabilitats teòriques.

d) Calculeu (teòricament) l'esperança i la variància de $X$. Després, calculeu la mitjana i variància empíriques de la mostra. Heu obtingut una bona aproximació?

## 2.2 Variables aleatòries absolutament contínues

### Simulació de nombres aleatoris

En el cas que vulguem simular una variable aleatòria absolutament contínua que no sigui cap de les conegudes, utilitzarem el **mètode de la funció inversa**, que es basa en el següent resultat:

**Teorema 1.** Sigui $X$ una variable aleatòria amb funció de distribució contínua i invertible $F_X$. Considerem $U \sim U(0,1)$. Aleshores, la variable aleatòria

$$V = F_X^{-1}(U)$$

té la mateixa distribució que la variable $X$.

En efecte,

$$F_V(x) = P(V \leq x) = P\left(F_X^{-1}(U) \leq x\right) = P\left(U \leq F_X(x)\right) = F_X(x).$$

El mètode per fer la simulació consisteix en:

- Obtenir nombres aleatoris $u_1, u_2, ..., u_n$:

```
u <- runif(n)
```

- Calcular $F_X^{-1}$
- Aplicar-la als nombres aleatoris i obtenir la mostra $x_1 = F_X^{-1}(u_1), ..., x_n = F_X^{-1}(u_n)$.

**Exemple 2.** Volem generar una mostra de mida 10 d'una variable aleatòria amb densitat

$$f(x) = 3x^2 \mathbb{1}_{(0,1)}(x).$$

Primer obtenim els nombres aleatoris:

```
u <- runif(10)
u
[1] 0.3513617 0.4313612 0.7389351 0.6115086 0.8457035
0.5707821 0.4215648 0.7860317 0.4976643 0.6077936
```

Després calculem la funció de distribució

$$F(x) = \begin{cases} 0, & x < 0, \\ \int_0^x 3y^2 dy = x^3, & 0 \leq x < 1, \\ 1, & 1 \leq x, \end{cases}$$

de manera que la inversa és $F^{-1}(u) = \sqrt[3]{u}$.
Finalment, apliquem $F^{-1}$ :

```
y <- u^{1/3}
y
[1] 0.7056426 0.7555798 0.9040701 0.8487912 0.9456695
0.8295135 0.7498161 0.9228831 0.7924627 0.8470689
```

Dibuixem l'histograma amb el dibuix de la funció de densitat.

```
hist(y, freq = FALSE)
z <- seq(0, 1, by = 0.05)
t <- 3*z^2
lines(z,t)
```

### Representació de la funció de densitat

Com en el cas de les variables conegudes, per dibuixar la funció de densitat cal definim primer un vector amb els valors de la variable.

Suposem que volem dibuixar la densitat

$$f(x) = \frac{3}{7}(x+1)^2 \mathbb{1}_{(0,1)}(x).$$

Com que $f(x) = 0$ si $x < 0$ o si $x > 1$, la successió de punts només cal agafar-la entre 0 i 1:

```
x <- seq(0, 1, by = 0.05)
f <- function(x){
        fx <- (3/7)*(x+1)^2
        fx
      }
plot(x, f(x), type = "l")
```

**Representació de la funció de distribució**

Recordeu que funció de distribució és la integral de la funció de densitat.

Suposem que volem dibuixar la funció de distribució de la llei que té per densitat la funció $f$ definida anteriorment. Integrant tenim:

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{1}{7}(x+1)^3 - \frac{1}{7} & \text{si } 0 < x < 1 \\ 1 & \text{si } x > 1 \end{cases}$$

Per dibuixar la funció de distribució utilitzem:

```
x <- seq(0, 1, by = 0.05)
F <- function(x){
        Fx <- (1/7)*(x+1)^3-1/7
        Fx
      }
plot(x, F(x), type = "l")
```

**Càlcul de probabilitats**

Pel càlcul de les probabilitats utilitzem la funció de distribució: $F(x) = P(X \leq x)$.

Per exemple, si la funció de distribució és la funció $F$ definida anteriorment:

- $P(X = 0.5)$: Aquesta probabilitat és zero.
- $P(X \leq 0.3)$ es calcula mitjançant

  ```
  F(0.3)
  ```

- $P(X \geq 0.8) = 1 - P(X < 0.8)$ es calcula mitjançant

  ```
  1 - F(0.8)
  ```

**Exercici 4.** Sigui $X$ una variable aleatòria absolutament contínua amb funció de densitat:

$$f(x) = \frac{5}{32}\, x^4, \quad 0 < x < 2.$$

a) Simuleu $N = 500$ valors de $X$ i dibuixeu l'histograma. Superposeu-hi el dibuix de la funció de densitat.

b) Calculeu (teòricament) l'esperança i la variància de $X$. Després calculeu la mitjana i variància empíriques de la llista de nombres aleatoris. Heu obtingut una bona aproximació?

# 3 Teorema central del límit

El Teorema central del límit ens diu que la distribució de la suma de moltes variables aleatòries independents amb la mateixa distribució, independentment de quina sigui aquesta, és aproximadament una llei normal. La versió general del teorema és:

**Teorema 2.** Sigui $X_1, X_2, \ldots, X_n$ una mostra aleatòria d'una distribució amb esperança $\mu$ i variància $\sigma^2$. Aleshores, quan $n$ és prou gran ($n > 30$),

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}},$$

es comporta com una distribució normal estàndard.

**Exercici 5.** Comproveu el teorema central del límit utilitzant unes variables amb distribució de Bernoulli. Seguiu els següents passos:

- Fixeu $N$ i $B$, per exemple $N = 200$ i $B = 400$, i genereu $N$ mostres de mida $B$, d'una distribució de Bernoulli amb paràmetre $p = 0.5$, en forma d'una matriu $A$, de dimensió $N \times B$.
- Calculeu la llista $a$ de les $N$ mitjanes de les $N$ files de $A$.
- Calculeu l'esperança teòrica $\mu$ i la variància teòrica $\sigma^2$ (en aquest cas de la llei *Be(p)*).
- Centreu i normalitzeu el vector de mitjanes. (Recordeu que cal restar la mitjana teòrica $\mu$ i dividir per la desviació teòrica dividida per l'arrel de la mida de la mostra $\sigma/\sqrt{n}$)
- Dibuxeu l'histograma del vector de mitjanes normalitzat amb la corba de la densitat de la normal.
- Que observeu? Cal augmentar el valor de $N$ per obtenir una aproximació millor?

## Problemes

1. Considereu una variable aleatòria $X$ que té la funció de massa de probabilitat següent:

$$\begin{aligned} P(X=2) &= \frac{1}{16} \\ P(X=k) &= \frac{k-3}{4k} \qquad k \in \{4,5,10,12,15,20\} \end{aligned}$$

   a) Representeu la funció de massa de probabilitat i la corresponent funció de distribució.

   b) Si sabem que la variable $X$ pren un valor més gran o igual a 5, quina és la probabilitat que el valor que agafa sigui menor o igual a 12?

   c) Calculeu l'esperança i la variància teòriques de $X$.

   d) Genereu una mostra de mida $n = 300$ de la variable aleatòria $X$ i calculeu-ne la mitjana i la variància empíriques. Heu obtingut una bona aproximació? Si considereu que no, augmenteu la mida de la mostra fins a trobar-ne una millor.

2. La distribució Gamma és una llei que depèn de dos paràmetres, $k$ i $\theta$. Si $k$ és un nombre sencer, aleshores la distribució $Gamma(k, \theta)$ representa la suma de $k$ variables aleatòries independents amb distribució Exponencial amb paràmetre $\theta$. Anem a veure un exemple d'aquesta última afirmació.

   a) Genereu una mostra de mida $n = 1000$ d'una variable aleatòria que sigui la suma de 5 Exponencials, totes amb paràmetre 2. Dibuixeu l'histograma i calculeu-ne mitjana i variància.

   b) Genereu una mostra de mida $n = 1000$ d'una variable aleatòria $Gamma(5, 2)$ i calculeu-ne mitjana i variància.

   c) Quin valors heu obtingut als dos apartats anteriors? S'aproximen la mitjana i la variància de les dos simulacions?

   d) Ara suposem que el temps de reparació d'una màquina segueix una distribució $Gamma(5, 2)$. El cost de la reparació és de 500 euros l'hora més el desplaçament de 300 euros. Quina és la probabilitat de que el cost de la reparació no superi els 2000 euros?

3. Considereu una variable aleatòria absolutament contínua $X$ amb funció de densitat:

$$f(x) = 3x^2 \mathbb{1}_{(0,1)}(x).$$

Utilitzeu aquesta distribució per comprovar el teorema central del límit seguint els mateixos passos de l'exercici de la pràctica.

4. Els coeficients d'intel·ligència d'un grup d'adults entre 20 i 34 anys tenen una distribució aproximadament normal de mitjana $\mu = 110$ i desviació típica $\sigma = 25$.

   a) Quin percentatge de persones entre 20 i 34 anys té coeficient més gran que 100?

   b) Quin percentatge de persones entre 20 i 34 anys té coeficient més petit que 150?

   c) Quin coeficient mínim tenen els adults entre 20 i 34 anys situats en el 25% que han obtingut millors resultats?

5. Considerem una sèrie d'experiments independents i idènticament distribuïts amb probabilitat $p$ d'èxit. La distribució Geomètrica és una llei coneguda que compta el nombre d'intents fallits fins a obtenir el primer èxit.

   a) Sigui $X$ una variable aleatòria amb distribució Geomètrica amb paràmetre $p = 0.3$. Dibuixeu-ne la funció de massa de probabilitat i la funció de distribució.

   b) Comproveu que la següent igualtat

$$P(X > m + n | X > m) = P(X > n - 1)$$

   és certa per valors de $m$ i $n$ entre 1 i 20.

   c) Utilitzant la distribució Geomètrica, feu 200 simulacions del nombre de llançaments necessaris d'un dau fins que surti un 5. Observeu que en aquest cas necessiteu una variable que compti els intents necessaris (no només els fallits) fins a obtenir el primer èxit.

   d) Sempre utilitzant la distribució Geomètrica, calculeu la probabilitat que el nombre de llançaments necessaris d'un dau fins que surti un 5 estigui entre 4 i 10 inclosos.