Első gyak

Monos Attila

2022-09-19

Valószínűség becslése, kísérletek

Dobjunk fel egy szabályos érmét néhányszor. Tegyük fel, hogy ez lett az eredmény:

$$F \quad I \quad I \quad F \quad I \quad F \quad F \quad F$$

Az érme természetesen véletlenszerűen működik; szeretnénk tudni, mekkora az esélye annak, hogy fejet dobunk, és mekkora az esélye annak, hogy írást dobunk. A dobások számától függően az alábbira juthatunk:

Dobások száma	1	2	3	4	5	6	7	8
Fej valószínűsége	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{2}$	$\frac{4}{7}$	$\frac{5}{8}$
Írás valószínűsége	1	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{2}$	$\frac{3}{5}$	$\frac{1}{2}$	$\frac{3}{7}$	$\frac{3}{8}$

A következőképp számítottuk ki ezeket: legyen k a sikeres kísérletek száma (pl. az, hogy fejet dobtunk), és n az összes kísérlet száma. Ekkor a sikeres kísérletek *relatív gyakorisága* $\frac{k}{n}$.

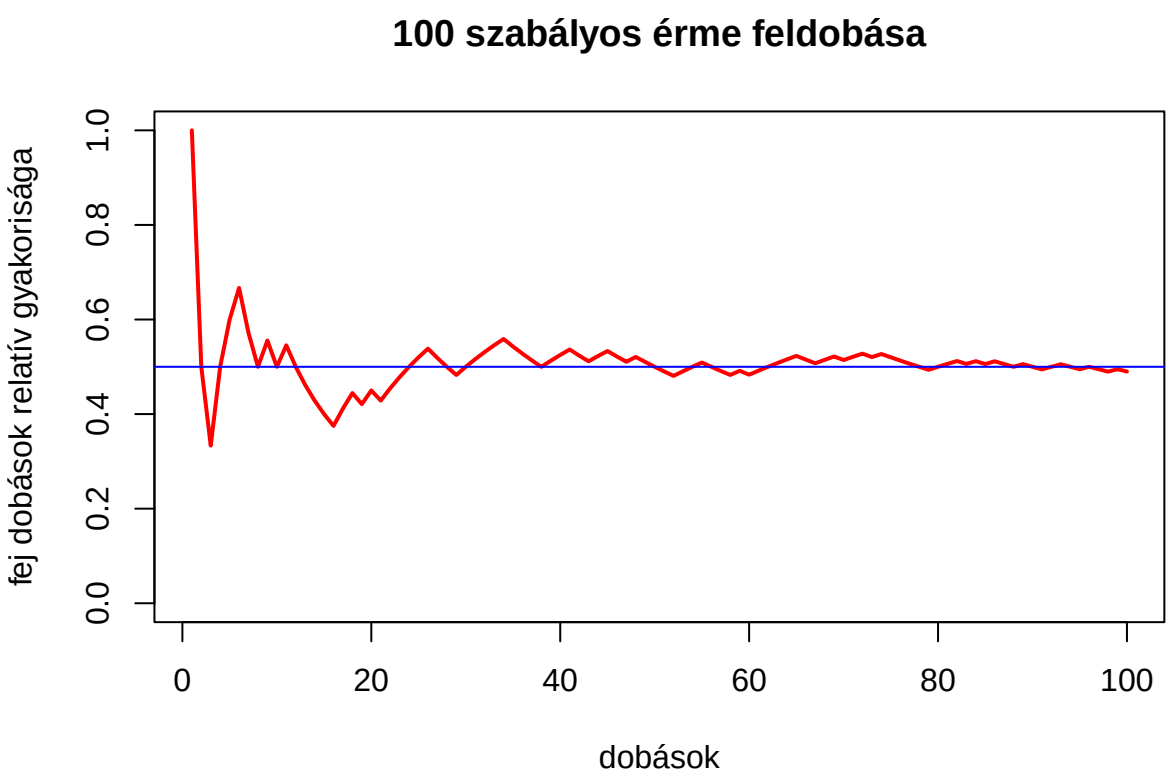
Mint tudjuk, mind a fej, mind az írás valószínűsége $\frac{1}{2}$. Buffon 4040 dobásból 2048 fejet kapott ($\frac{k}{n} = 0{,}5069$). Pearson 24000 dobásból 0,5005-ös relatív gyakoriságot kapott. Pongyolán mondva: ahogy $n \to \infty$, úgy tart a relatív gyakoriság a valószínűséghez.

```
n <- 100
erme <- c('F', 'I')
dobasok <- sample(erme, size = n, replace = TRUE, prob = c(1,1))

freqs <- table(dobasok)
freqs
```

```
## dobasok
##  F  I
## 49 51
```

```
fejgyak <- cumsum(dobasok == "F") / 1:n
plot( fejgyak,
      type = 'l', lwd = 2, col = 'red',
      ylim = c(0,1),
      ylab = "fej dobások relatív gyakorisága",
      xlab = "dobások",
      main = paste(n, "szabályos érme feldobása"))
abline(h = 1/2, col = 'blue')
```



Diszkrét valószínűségi változók

Valószínűségi változókról beszéltünk már – változók, amiknek az értéke a kísérlet kimenete. Egy valószínűségi változó diszkrét, ha véges sok, vagy megszámlálható végtelen sok értéket vehet fel.

Binomiális eloszlás

Dobjunk fel egyszerre négy egyforma érmét! A lehetséges kimenetelek:

4 fej	3 fej	2 fej	1 fej	0 fej
0 írás	1 írás	2 írás	3 írás	4 írás

Feladat: Mi a valószínűsége az egyes kimeneteleknek?

```
dbinom(0:4, size = 4, prob = 0.5)
```

```
## [1] 0.0625 0.2500 0.3750 0.2500 0.0625
```

Ezt a fajta jelenséget Bernoulli általánosította, ez a Bernoulli-kísérlet: két lehetséges kimenetel van, ezek valószínűsége p és $1-p$ (ahol $p \in (0,1)$ természetesen). Ezt a kísérletet egymástól függetlenül végrehajtjuk n alkalommal. Ekkor a lehetséges kimenetelek: $0, 1, \ldots, n$, vagyis ha X az a valószínűségi változó, mely leírja a sikeres kísérletek számát, akkor X diszkrét. Eloszlását vizsgálva:

$$P(X = k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

Kaptunk egy eseményteret, melynek $n+1$ eleme (eseménye) van, a fenti valószínűségekkel. Ez a *binomiális eloszlás*: $X \sim \text{Bin}(n, p)$.

Példaként nézzük meg az $n = 3$, $p = 0{,}3$ esetet.

```
dbinom(0, 3, 0.3) #Csak 1-re írja ki a valószínűséget
```

```
## [1] 0.343
```

```
dbinom(1, 3, 0.3)
```

```
## [1] 0.441
```

```
dbinom(2, 3, 0.3)
```

```
## [1] 0.189
```

```
dbinom(3, 3, 0.3)
```

```
## [1] 0.027
```

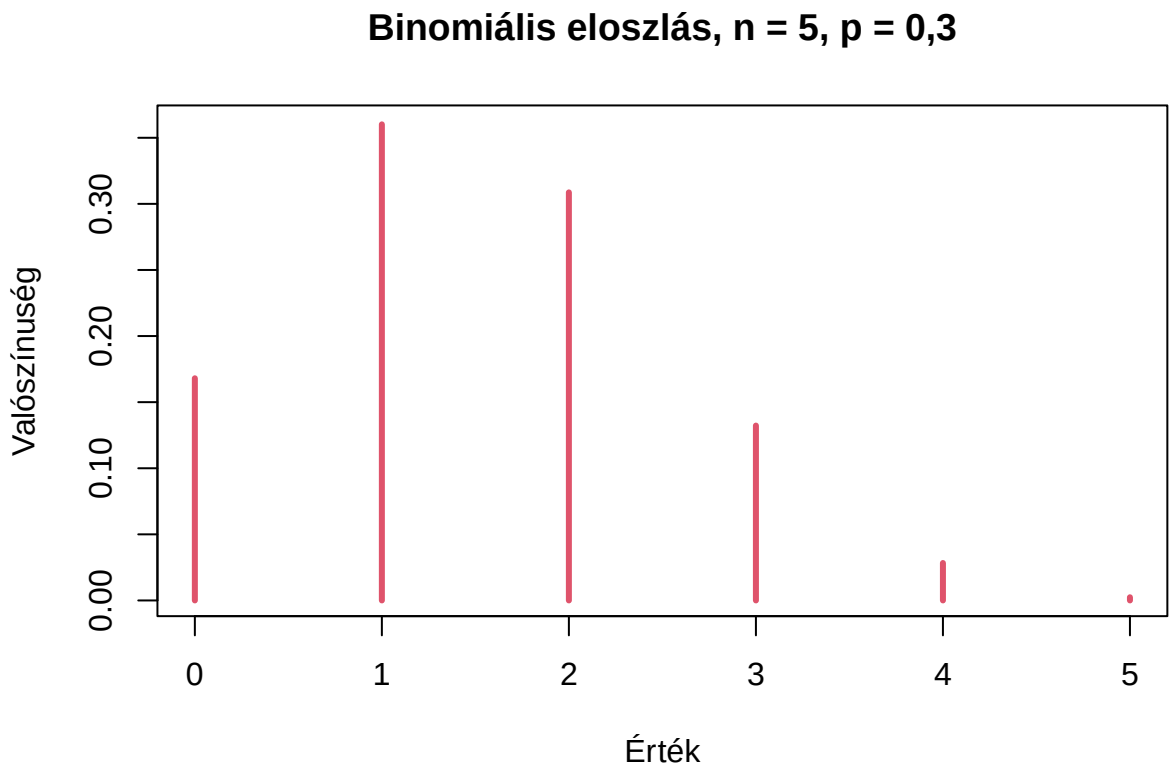
```
dbinom(c(1,3), 3, 0.3) #1-re és 3-ra írja ki a valószínűséget
```

```
## [1] 0.441 0.027
```

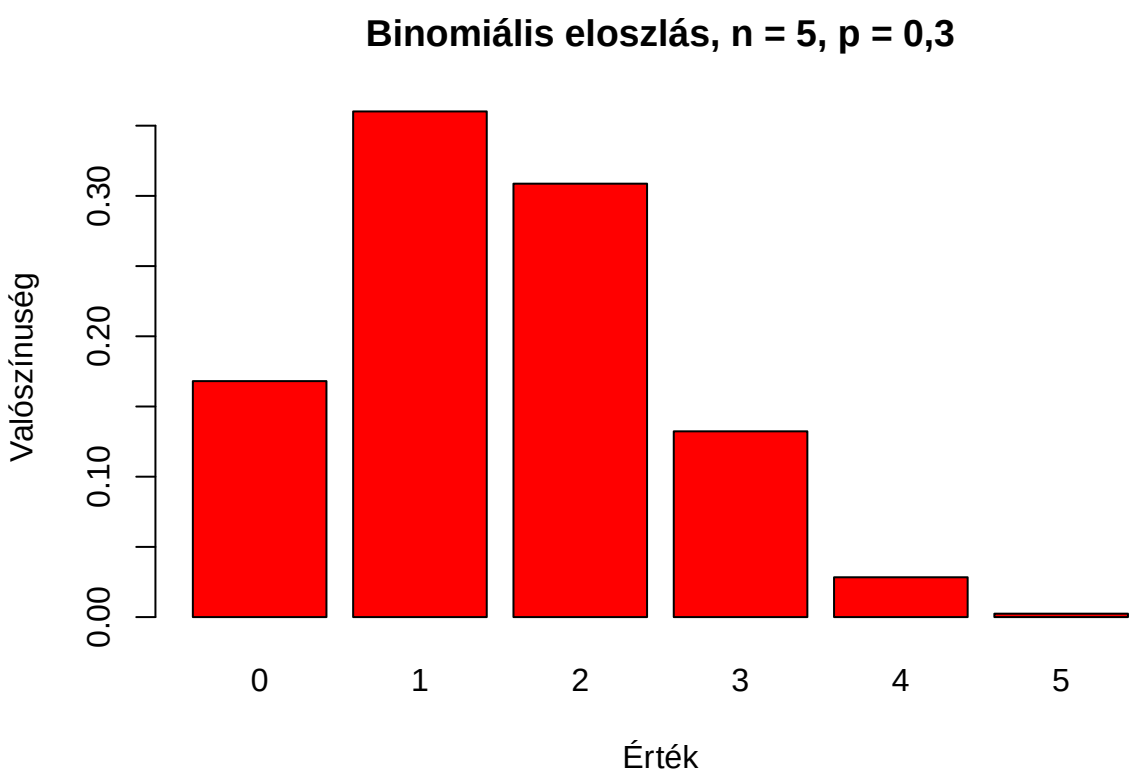
```
dbinom(1:3, 3, 0.3) #1-re, 2-re, 3-ra minden egész számra kiírja a valószínűséget
```

```
## [1] 0.441 0.189 0.027
```

```
plot(0:5, dbinom(0:5, 5, 0.3),
     type = "h", col = 2, lwd = 3,
     xlab = "Érték",
     ylab = "Valószínűség",
     main = paste("Binomiális eloszlás, n = 5, p = 0,3"))
```




```
barplot(dbinom(0:5, 5, 0.3),
        col = 'red',
        xlab = "Érték",
        ylab = "Valószínűség",
        main = "Binomiális eloszlás, n = 5, p = 0,3",
        names.arg = 0:5)
```



Eddig $P(X = k)$ értékeket tudunk számolni. Mi van, ha mondjuk azt szeretnénk kiszámolni, hogy $P(2 < X < 5)$? Ebben hogyan segíthet a `dbinom` parancs?

```
sum(dbinom(c(3,4), 5, 0.3))
```

```
## [1] 0.16065
```

Az eloszlásfüggvényt úgy definiáltuk, hogy $P(X \leq k)$ (Vagy $P(X < k)$, a két definíció lényegében ugyanaz). Ennek az értéke $k = 0, 1, \ldots, n$ esetén érdekes és fontos. Legyen most az egyszerűség kedvéért $n = 5$. Érdemes összevetni a `dbinom` és `pbinom` parancsok által visszaadott értékeket ugyanazon paraméterekre: ha csak egyedi valószínűségeket nézünk, akkor van egy csúcs valahol $n \cdot p$ körül, ha pedig eloszlásfüggvényt nézünk, akkor pedig monoton növekedik az oszlopok magassága:

```
pbinom(3, 5, 0.3)
```

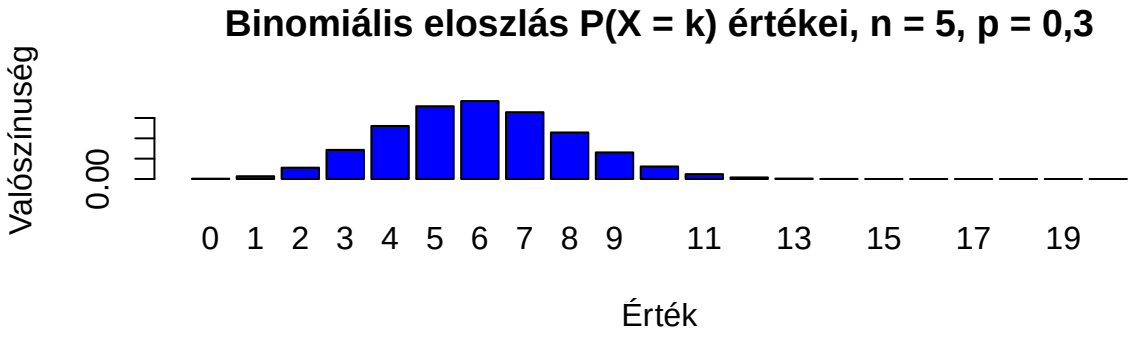
```
## [1] 0.96922
```

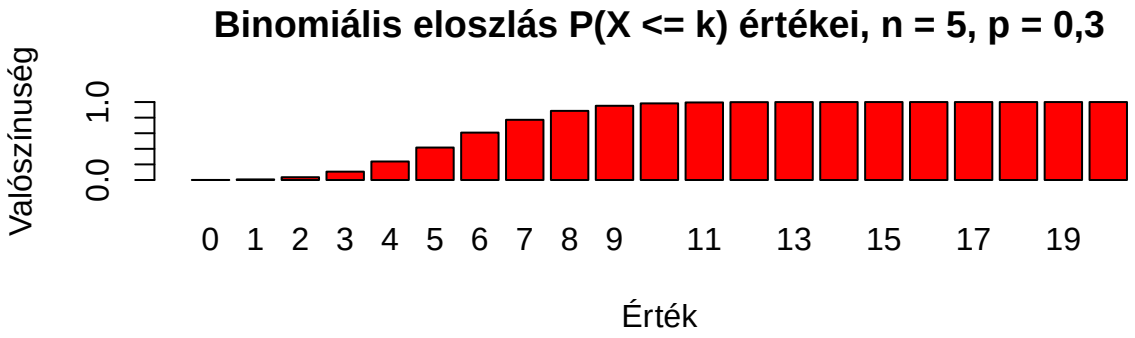
```
sum(dbinom(0:3, 5, 0.3))
```

```
## [1] 0.96922
```

```
par(mfcol=c(2,1)) #Két plotot összerak egy "táblázatba" -- 2 sor, 1 oszlop jelenleg
barplot(dbinom(0:20, 20, 0.3),
        col = 'blue',
```

```
        xlab = "Érték",
        ylab = "Valószínűség",
        main = "Binomiális eloszlás P(X = k) értékei, n = 5, p = 0,3",
        names.arg = 0:20)
barplot(pbinom(0:20, 20, 0.3),
        col = 'red',
        xlab = "Érték",
        ylab = "Valószínűség",
        main = "Binomiális eloszlás P(X <= k) értékei, n = 5, p = 0,3",
        names.arg = 0:20)
```





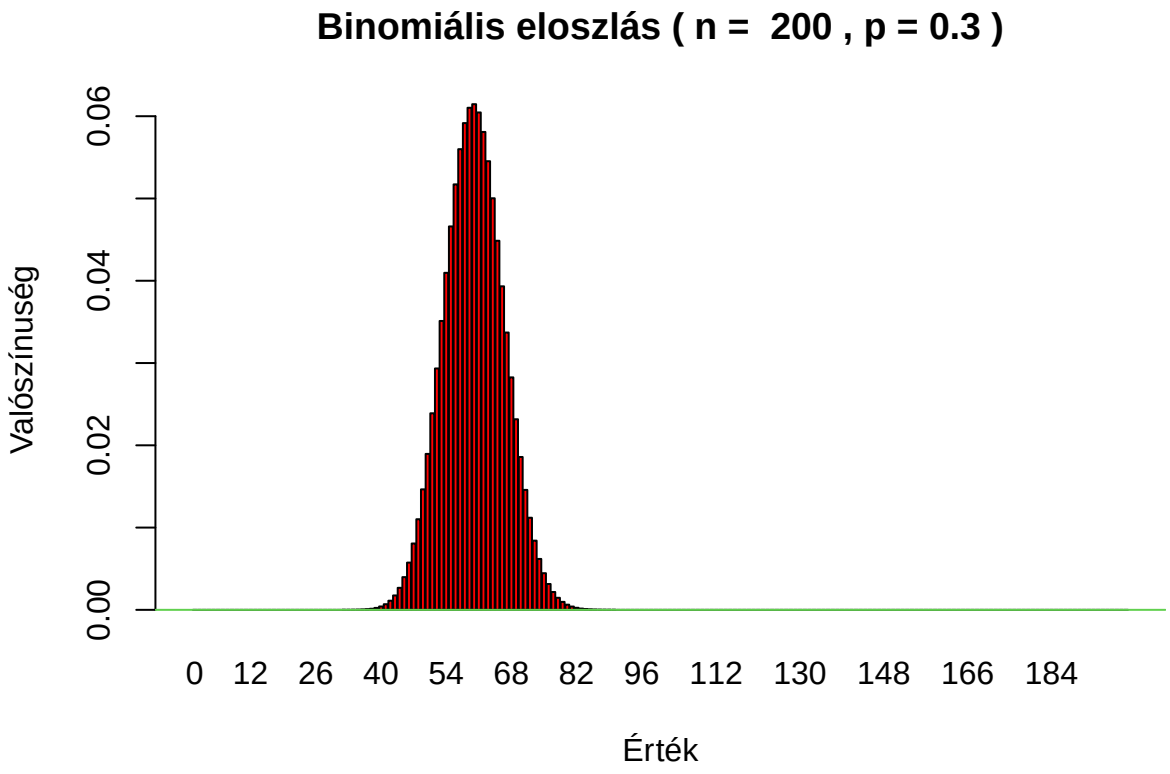
Ha ismerünk egy eloszlást, akkor abból könnyen tudunk értékeket generálni. Kísérletezgetéseknél ez hasznos lehet. Pl. úgy is meg lehet sejteni, hogy egy valószínűségi változó eloszlása pont binomiális, hogy "jól tudjuk szimulálni binomiális eloszlás szerint véletlen generált értékekkel". Persze ezután jön a matematika, hogy ez biztosan így van-e, de ez ettől még jó kiindulást adhat.

```
rbinom (10, 5, 0.3)
```

```
##  [1] 1 0 3 3 0 3 1 2 1 0
```

```
#Ezzel a kóddal lehet játszadozni, hogy hol lesz a "domb" különböző n és p értékekre
n = 200
p = 0.3
barplot(dbinom(0:n, n, p),
        col="red",
```

```
        xlab="Érték",
        ylab="Valószínűség",
        main= paste("Binomiális eloszlás ( n = ", n,", p =", p, ")"),
        names.arg = 0:n)
abline(h = 0, col = 3)
```



Indikátor eloszlás

Tekintsünk egy binomiális eloszlást, ahol a sikeres kísérlet valószínűsége p, és $n = 1$. Ekkor a binomiális eloszlás egy speciális esetét kapjuk, ami az *indikátor eloszlás*. Ezzel bináris eseményeket tudunk nagyon jól jellemezni; továbbá az indikátorok összegére való felbontás egy nagyon erős feladatmegoldó eszköz lehet, ha pl. várható értéket kell számolni. Az indikátor eloszlás egyszerűen leírható:

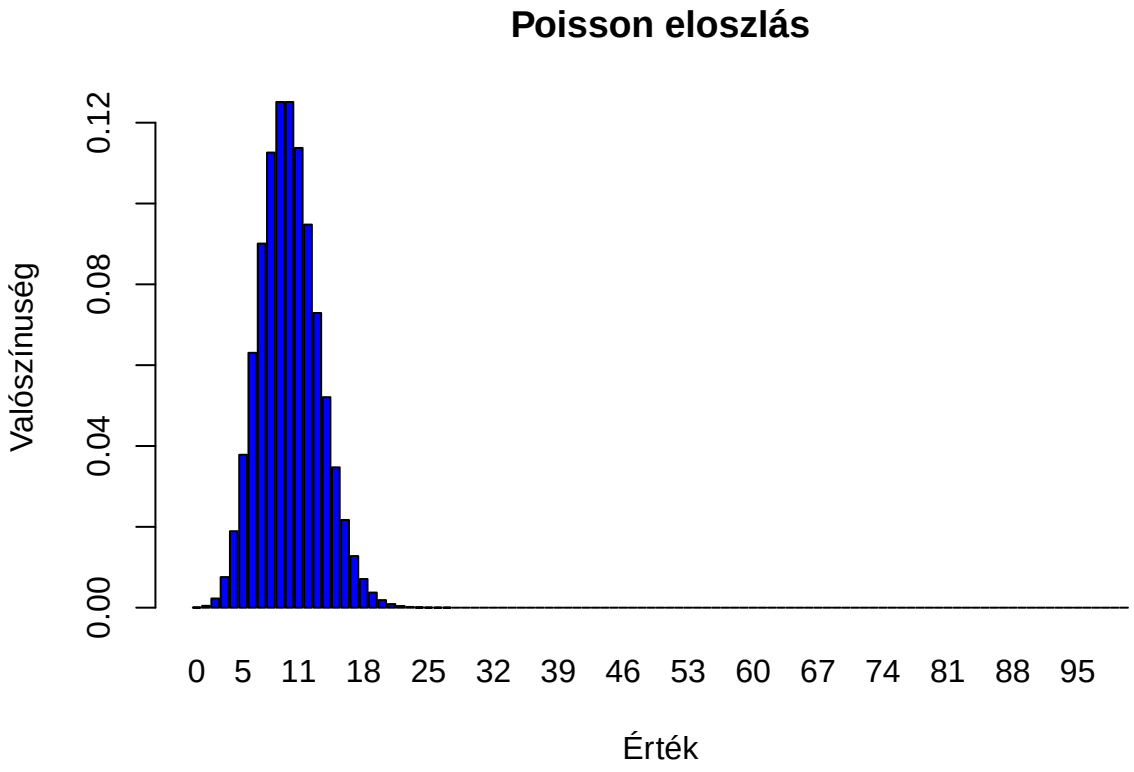
$$P(X = 1) = p, \quad P(X = 0) = 1 - p$$

Poisson eloszlás

Ha X *Poisson-eloszlású* ($X \sim \text{Pois}(\lambda)$), akkor $X = 0, 1, 2, \ldots$. Ne feledjük, ez még diszkrét! A Poisson-eloszlás egyetlen paramétere λ, mely bármilyen pozitív szám lehet. Ábrázoljuk a Poisson-eloszlást egy adott határig (hisz végtelen sok oszlopot nehéz rajzolni)!

```
lambda = 10
n = 100
barplot(dpois(0:n, lambda),
        col = "blue",
```

```
        xlab = "Érték",
        ylab = "Valószínűség",
        main = "Poisson eloszlás",
        names.arg = 0:n)
```



Egyébként a Poisson-eloszlásnak nincs meghatározó szabálya – nemes egyszerűséggel meg van adva, ha egy X valószínűségi változó Poisson. Hogyan tudjuk meg, hogy valami Poisson-eloszlású? Itt jön be a statisztika: tapasztalai eloszlást számítunk, és ha az nem tér el nagyon egy adott $\lambda > 0$ paraméterű Poisson-eloszlástól, akkor annak fogjuk tekinteni. Egyébként a Poisson-eloszlás az alábbi:

$$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \dots$$

Binomiális és Poisson eloszlások kapcsolata

Matematikailag megmutatható (sok-sok munkával, ettől eltekintünk), hogy ha n elég nagy, akkor a Pois(λ) és Bin($n, \frac{\lambda}{n}$) eloszlások "szinte ugyanolyanok". Ellenőrizzük ezt le!

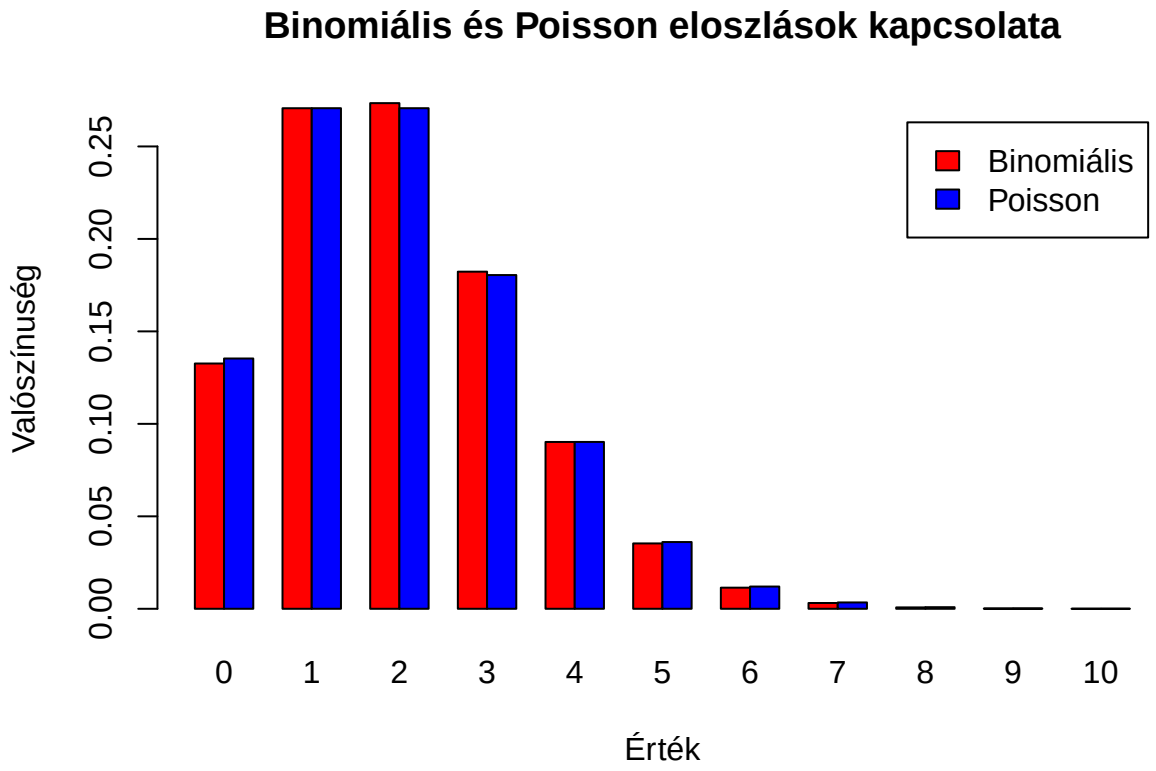
```
lim = 10
lambda = 2
n = 100
p = lambda/n

ranbin = dbinom(0:lim, n, p)
ranpoi = dpois(0:lim, lambda)

#Összeköti a két, azonos típusokból álló vektor elemeit.
```

```r
#Köze van a dataframe típushoz, lásd első laboros gyak első fele, és Rbevezetes.R állomány Zempléni Tan
together = rbind(ranbin[1:(lim + 1)], ranpoi)
rownames(together) = c("Binomiális", "Poisson")

barplot(together,
        col = c("red", "blue"),
        xlab = "Érték",
        ylab = "Valószínűség",
        main = "Binomiális és Poisson eloszlások kapcsolata",
        legend = rownames(together),
        beside = TRUE,
        names.arg = 0:lim)
```



Hipergeometriai eloszlás

A probléma a következő: van N db számítógépünk. Ebből M db-nak elég erős a processzora, de nem tudjuk, melyek azok. Így véletlenszerűen vásárolunk belőlük n darabot. A kérdés az, hogy mekkora eséllyel veszünk $0, 1, \ldots, n$ olyan gépet, mely elég erős?

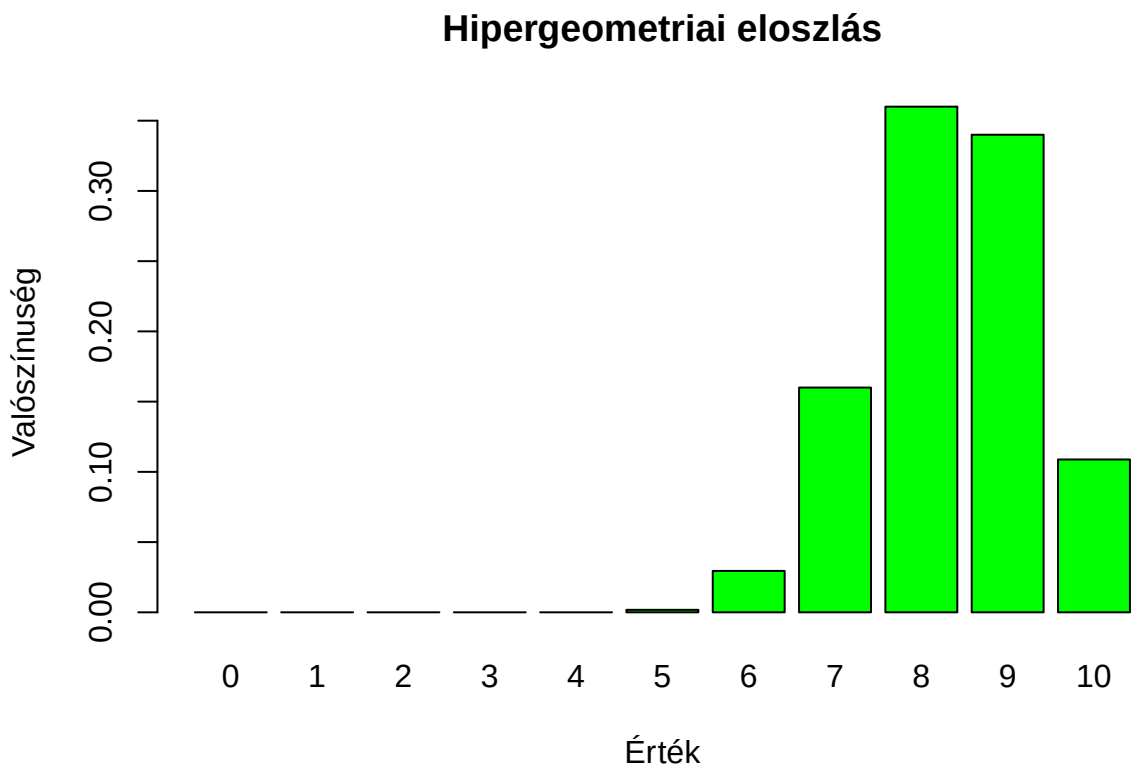
Általánosabban: van N golyónk, ebből M jelölt, és n darabot húzunk ki visszatevés nélkül. Jelölje X a kihúzott jelölt golyók számát. Ekkor $X = \text{HipGeo}(M, N, n)$

```r
N = 30
M = 25
n = 10
```

```r
dhyper(0:n, M, N - M, n)
```

```
##  [1] 0.000000000 0.000000000 0.000000000 0.000000000 0.000000000 0.001768347
##  [7] 0.029472443 0.159993263 0.359984843 0.339985685 0.108795419
```

```r
barplot(dhyper(0:n, M, N - M, n),
        col = "green",
        xlab = "Érték",
        ylab = "Valószínűség",
        main = "Hipergeometriai eloszlás",
        names.arg = 0:n)
```



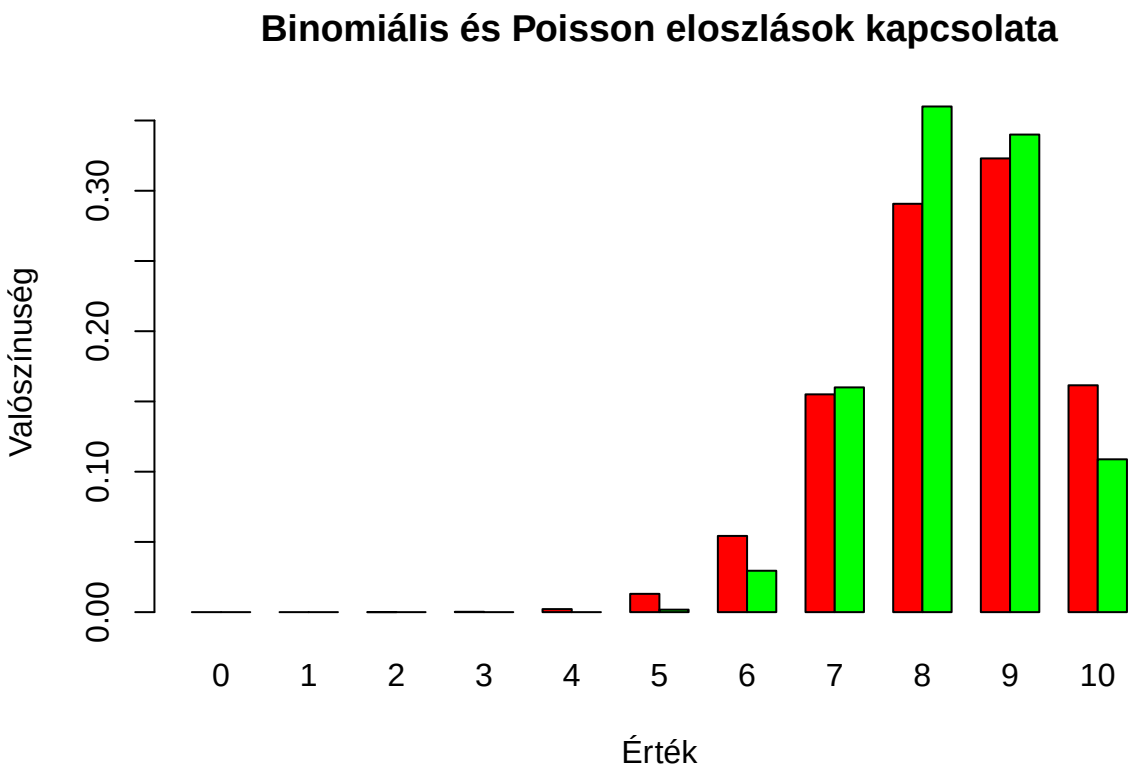

```r
N = 30
M = 25
n = 10
p = M/N

ranbin = dbinom(0:n, n, p)
ranhyp = dhyper(0:n, M, N-M, n)

together = rbind(ranbin[1:(lim + 1)], ranhyp)
rownames(together) = c("Binomiális", "Hipergeometrikus")

barplot(together,
```

```
        col = c("red", "green"),
        xlab = "Érték",
        ylab = "Valószínűség",
        main = "Binomiális és Poisson eloszlások kapcsolata",
        #legend = rownames(together),
        beside = TRUE,
        names.arg = 0:lim)
```



Feladatok

1. Tegyük fel, hogy az új internet-előfizetők véletelenszerűen választott 20%-a speciális kedvezményt kap. Mi a valószínűsége, hogy a 10 ismerősünk közül, akik most fizettek elő, legalább négyen részesülnek a kedvezményben?

   Megoldás:

   Jelölje X a kedvezményt kapott internetelőfizetők számát. Mindegyikük 0,2 eséllyel kap kedvezményt – azaz minden kísérlet 0,2 eséllyel sikeres. Két különböző előfizető esetén a kedvezmények sorsolása független, így tulajdonképp $n = 10$ független kísérletet hajtunk végre egymás után, mindegyük $p = 0{,}2$ valószínűséggel sikeres. Így $X \sim \text{Bin}(n, p)$.

   Az első megoldásban azt mondjuk, hogy X jó értékei $4, 5, \ldots, 10$. Ezek egymást kizáró eredmények, így $P(X = 4, \ldots, 10) = p(X = 4) + \ldots + P(X = 10)$.

   A második megoldásban egy alternatív megoldást adunk komplementer eseménnyel: az a rossz, ha legfeljebb 3-an kaptak kedvezményt, így ennek a valószínűségét kell 1-ből kivonni. Ehhez $P(X \leq 3)$-at kell kiszámolni.

2. Egy bükkösben a bükkmagoncok (bébi bükkfák?) négyzetméterenkénti száma Poisson-eloszlású, $\lambda = 2{,}5\text{db/m}^2$ paraméterrel. Mi a valószínűsége annak, hogy egy 1 m^2-es mintában

   (a) legfeljebb egy,

   (b) több, mint három magoncot találunk?

   Megoldás:

   Mint ahogy az előző feladatban is, jelölje X a négyzetméterenkénti bükkfamagoncok számát. Ekkor $X \sim \text{Pois}(2{,}5)$. Az első részfeladatban $P(X \geq 1) = 1 - P(X = 0)$-t kell kiszámolni. Itt természetesen a komplementer esemény jobban segít, hiszen a Poisson-eloszlásnak végtelen sok értéke lehet, azt meg nehéz összegezni.

   A második részfeladatban $P(X > 3) = 1 - P(X \leq 3)$ a keresett valószínűség.