Binomické a Poissonovo rozdelenie

Cieľom tohto textu je ilustrovať priebeh binomického a Poissonovho rozdelenia, precvičiť si prácu s cyklom v matlabe a v rámci možností racionalizovať výpočtový proces.

Príklad 1:

Do školy v Hornej STUbni chodí 49 žiakov. Pravdepodobnosť absencie jedného študenta v ktorýkoľvek deň je 11%. Vypočítajme pravdepodobnosť, že v určitý deň bude chýbať *k* žiakov, pre *k* = 0, 1, 2, … , 49.

Pravdepodobnosť súčasnej absencie *k* žiakov je daná binomickým rozdelením:

$$p(k) = n! / (k! (n-k)!) * p^k * (1-p)^{(n-k)}$$

Tieto hodnoty v matlabe získame najjednoduchšie pomocou cyklu:

```
n = 49; p = 0.11;
for k = 0:n, bp(k+1) = factorial(n)/factorial(k)/factorial(n-k) * p^k * (1-p)^(n-k); end
sum(bp)
bar(0:n,bp)
```

Poznámky:

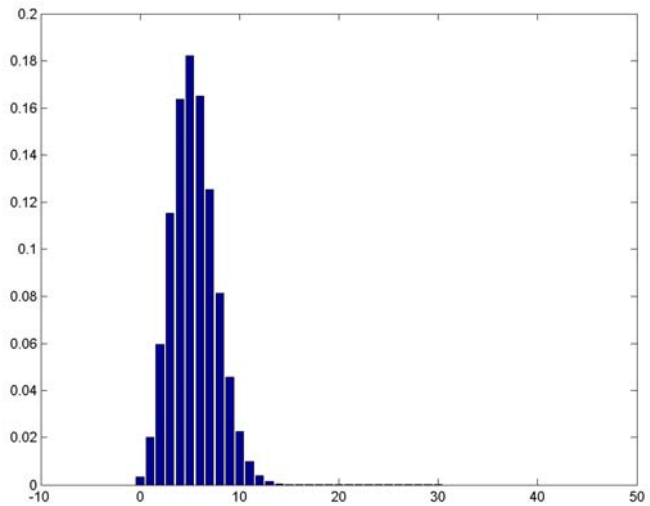
a) Spôsob výpočtu hodnôt bp nie je najvhodnejší a v ďalšom príklade ukážeme, prečo ho treba a ako ho možno vylepšiť.

b) Vektor pravdepodobností bp má dĺžku 50. Keďže Matlab indexuje zložky vektorov zásadne od jednotky vyššie, musíme písať bp(k+1) = … . Hodnotám p(0), p(1), … , p (49) teda zodpovedá bp(1), bp(2), … , bp(50).

c) Za predpisom pre výpočet bp(k+1) musí NUTNE nasledovať bodkočiarka. Ak ju tam nedáme, v každom kroku (je ich 50) Matlab vypíše všetky hodnoty vektora bp, ktoré už vypočítal, a zamorí nám obrazovku.

d) Správnosť výsledku skontrolujeme sčítaním hodnôt vektora bp. Výsledok by mal byť 1.

e) Napokon rozdelenie vykreslíme:



Príklad 2:

Do školy v Dolnej STUbni chodí 765 študentov. Kvôli problémom s dopravou je pravdepodobnosť absencie študenta v ktorýkoľvek deň 23% . Vypočítajme pravdepodobnosť, že v určitý deň bude chýbať *k* žiakov, pre *k* = 0, 1, 2, … , 765.

Opäť ide o výpočet hodnôt binomického rozdelenia pre n=765 a p=0.23. Ak by sme chceli postupovať podľa predošlého návodu, narazíme na problém – Matlab si neporadí s faktoriálom čísla 765. Najvyšší faktoriál, ktorý zvládne, je 170.

Vďaka tomu, že je Matlab rýchly, sme narazili iba na jeden problém, v skutočnosti sú však dva. Aj ten druhý však treba riešiť.

a) Prvý problém je vo veľkých hodnotách faktoriálov. Tu si treba uvedomiť, že my nepotrebujeme samostatné hodnoty všetkých troch faktoriálov, ale len výsledný podiel, ktorý nebude tak veľký. Výpočet treba prispôsobiť tak, aby sme sa týmto veľkým číslam vyhli.

b) Druhý problém spočíva v množstve zbytočných (zbytočne opakovaných) operácií. Napr. ak máme vypočítaný (k-1)!, nie je nutné počítať k! od začiatku, ale stačí vziať predošlý výsledok a násobiť ho číslom k. Počet operácií tak dosť podstatne klesne. Hoci je Matlab rýchly a pri jednoduchších výpočtoch nám nedá pocítiť neefektívnosť algoritmu, pri dlhších výpočtoch bude rozdiel citeľný. Avšak bez ohľadu na to, čo počítame, efektívnosť algoritmu by mala byť otázkou ako estetiky, tak aj elementárnej inžiniersko-programátorskej hrdosti.

Oba problémy môžeme naraz vyriešiť tak, že každý ďalší člen vektora hodnôt pravdepodobnostnej funkcie budeme počítať z predchádzajúceho člena. Ak platí:

$$p(k) = n! / (k!\ (n-k)!) * p^{k} * (1-p)^{(n-k)}$$
$$p(k-1) = n! / ((k-1)!\ (n-k+1)!) * p^{k-1} * (1-p)^{(n-k+1)}$$

potom

$$p(k) = p(k-1)*(n-k+1)/k*p/(1-p)$$

Cyklus pre výpočet bp bude vyzerať takto:

```
clear bp
n = 765; p = 0.23;
bp(1) = (1-p)^n;
for k=1:n, bp(k+1) = bp(k)*(n-k+1)/k*p/(1-p); end
sum(bp)
bar(0:n,bp), colormap([0.5 0 0])
```

Počet vykonaných operácií je teraz rádovo násobkom n, kým v predošlom prístupe si vyžadoval rádovo n^n operácií, čo je pre väčšie n likvidačná hodnota.

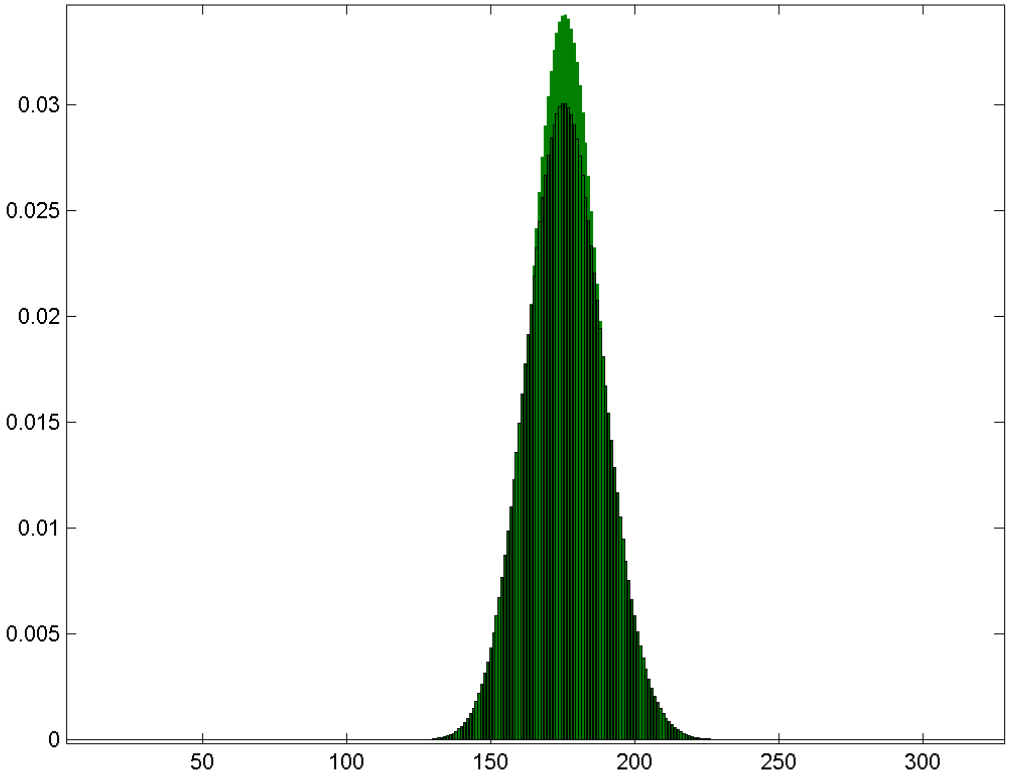
Príklad 3:

Pre väčšie n možno binomické rozdelenie úspešne aproximovať Poissonovým. Najmä ak chceme počítať pravdepodobnosť len pre niekoľko hodnôt *k*, použitie vzorca pre Poissonovo rozdelenie je jednoduchšie.

Zvoľme L=n*p. Hodnoty Poissonovho rozdelenia získame v duchu predošlých úvah opäť rekurentným spôsobom:

```
L=n*p;
pp(1) = exp(-L);
for k = 1:n, pp(k+1) = pp(k)*L/k; end
sum(pp)
hold on, bar(0:n,pp), colormap([0 0.5 0])
```

Ak si zväčšíme získaný obrázok, v detailných náhľadoch uvidíme, ako sa k sebe blížia obe rozdelenia:



Príklad 4:

Podľa štatistík je pravdepodobnosť pádu lietadla istej nemenovanej leteckej spoločnosti XY so zákazom vstupu na európske letiská p = 1/213 000.

a) Pri akom počte letov môže spoločnosť rátať s viac ako polovičnou pravdepodobnosťou, že si do svojich kroník zapíše nešťastie?

Na výpočet zvolíme postup od konca – pri akom počte letov sa pravdepodobnosť beznehodového fungovania dostane pod 50% ? Inými slovami, pre aké *k* platí $(1-p)^k \leq 0.5$?

```
>> log(0.5)/log(1-p)
                    ans =  1.4764e+005
```

Je to zhruba 148 tisíc letov.

b) Istý proeurópsky naladený minister navrhol, že po troch haváriách svojich lietadiel by každá spoločnosť mala prejsť prísnym akreditačným konaním. Aká je pravdepodobnosť, že v období, keď chce uskutočniť milión letov, by spoločnosti XY hrozilo takéto konanie?

Ide o pravdepodobnosť, že počas milión letov sa v prevádzke stane 3 alebo viac nehôd. Počítajme „odzadu“ – aká je pravdepodobnosť, že sa v tom čase stane 0, 1, 2 nehôd?

```
>> n=1000000; P012 = (1-p)^n + n*(1-p)^(n-1)*p + n*(n-1)/2*(1-p)^(n-2)*p^2

P012 =  0.1528
```

Pravdepodobnosť 3 alebo viacerých nehôd je potom

```
>> 1-P012
                ans =    0.8472
```

Spoločnosť po tomto zistení rozpredala časť strojov a naplánovala si obmedzený prevádzkový režim so 100 000 letmi. Má reálne šance vyhnúť sa problémom?

```
>> n=100000; P012 = (1-p)^n + n*(1-p)^(n-1)*p + n*(n-1)/2*(1-p)^(n-2)*p^2
>> 1-P012

P012 =    0.9878
ans  =    0.0122
```

Na takmer 99% sa problémom vyhnú. Chýbajúce percento poistili neobmedzenou ponukou letov zadarmo špeciálne pre pána ministra.

Úloha: Podľa štatistík je pravdepodobnosť pádu lietadla Boeing 737 spoločnosti British Airways jeden k stodesiatim miliónom (p = 1/110000000). Sformulujte podobné úlohy ako vyššie uvedené a), b) a riešte ich.

Príklad 5: Provokatér – hazardný hráč nám ponúka nasledovný podnik:
Pozýva nás na hru, kde máme v ľubovoľnom kole 40% šancu zvíťaziť. Avšak v prípade výhry dostaneme 600 Sk, kým v prípade prehry platíme len 350 Sk. Zaujímavé, no nezahrajte si…

a) Je rozumné sa do takejto hry púšťať?
b) Koľko kôl treba hrať, aby sa to oplatilo?

Ak je výhra dvojnásobná oproti prehre, stačí zvíťaziť najmenej v tretine hier, aby sme neboli stratoví. Pozrime sa na svoje šance pri rôznych počtoch n:

```
>> for n=1:20; k=ceil(n/3); p=0.4; q=p^n; s=0; for i=n:(-1):k, s=s+q;
       q=q/p*(1-p)*i/(n+1-i); end, [n,s], end
```

ans =

počet hier	*šance*
1.0000	0.4000
2.0000	0.6400
3.0000	0.7840
4.0000	0.5248
5.0000	0.6630
6.0000	0.7667
7.0000	0.5801
8.0000	0.6846
9.0000	0.7682
10.0000	0.6177
11.0000	0.7037
12.0000	0.7747
13.0000	0.6470
14.0000	0.7207
15.0000	0.7827
16.0000	0.6712
17.0000	0.7361
18.0000	0.7912
19.0000	0.6919
20.0000	0.7500

Šance dosť kolíšu, čo je dané aj skokmi pri zaokrúhľovaní n/3. Celkovo to však vyzerá všetko v náš prospech a treba hrať.

Úlohy: a) Ubezpečte sa, že šance nezačnú neskôr klesať. Berte pritom na vedomie, že pri vyšších hodnotách n začína mať Matlab problémy s presnosťou výpočtu.

b) Skúmajte situáciu pri iných hodnotách p a inom podieli výhry či prehry. Nezabudnite na prípad p=0.4, 600Sk výhra a 400Sk prehra, ani na p=0.5 a 500:500 a podobne.

Príklad 6:

a) Od pôrodu do prepustenia domov strávi žena s dieťaťom v priemere 4 dni[1] v pôrodnici. Podľa štatistík ročne prichádza do pôrodnice priemerne 1207 žien. Aká je stredná hodnota obsadenosti pôrodnice?

Pri daných číslach za rok potrebuje pôrodnica 3.5*1207=4828 lôžkodní. Denne je teda obsadených v priemere 4224.5/365 = 13.2274 lôžok.[2]

b) Na základe uvedených čísel ministerstvo rozhodlo o redukcii počtu lôžok na 18. Aká je pravdepodobnosť, že to v niektorý deň nebude stačiť?

Počítajme s Poissonovým rozdelením pravdepodobnosti, kde L=13.2274. Aby sme sa vyhli nekonečným súčtom, počítajme pravdepodobnosť toho, že prítomných žien bude najviac 18.

```
>> q=exp(-L); s=0; for i=0:18, s=s+q; q=q*L/(i+1); end, s

        s = 0.9207
```

Ak pri každom dni je pravdepodobnosť, že kapacity postačia, 92.07%, počas roku treba rátať s 7.93% dní, tj. asi 29 dňami, keď to stačiť nebude.

c) Koľko lôžok je potrebných na to, aby pravdepodobnosť zlyhania kapacít bola najviac 1% ?

Opäť počítame odzadu, teda hľadáme taký počet lôžok, aby pravdepodobnosť ich dostatku bola aspoň 99% :

```
>> q=exp(-L); s=0; i=0; while s<0.99, s=s+q; i=i+1; q=q*L/i; end, [i-1,s]

        ans =    22.0000    0.9908
```

Pri počte 22 lôžok sa dá počítať s tým, že len necelé percento dní, v priemere teda 3.3621 dňa, bude kritických. Dajme tomu, že ministerstvo pokorené brilantnou štatistickou analýzou súhlasilo a zvýšilo počet na 22.

d) Prípadnú preplnenosť pôrodnice treba podľa inštrukcií zhora riešiť presunom na vedľajšie oddelenie. Pre jednoduchosť predpokladajme, že obe oddelenie majú zhodné *parametre vyťaženosti* (myslia sa tým čísla dosádzané do Poissonovho vzorca). Aká je pravdepodobnosť, že sa týmto spôsobom podarí predísť kapacitným kolíziám?

Počítajme pravdepodobnosti toho, že „susedia“ majú práve obsadených 0, 1, 2, 3, … 21 miest (výsledok nazveme S) a teda voľných **aspoň** 22, 21, …, 1 miest (nazveme to Sc):

```
>> q=exp(-L); S=[]; for i=0:22, S=[S q]; q=q*L/(i+1); end
>> Sc=cumsum(S)
```

Ak je v pôrodnici do 22 žien, nie je problém. Ak ich je 22+k, treba ich k presunúť a na to treba, aby vedľa bolo aspoň k (tj. k alebo viac) miest voľných.

---

[1] Pre jednoduchosť berme do úvahy dni ako nedeliteľné celky.
[2] Každé lôžko matky má samozrejme priradenú kolísku pre dieťa.

Do vektora P uložme pravdepodobnosti toho, že v pôrodnici bude k=1:22 žien navyše (o počte nad 22 neuvažujeme, lebo na to ani celkom prázdne susedné oddelenie nebude stačiť):

```
>> q=exp(-L); P=[]; for i=0:44, P=[P q]; q=q*L/(i+1); end
>> P=P(24:45);
```

Ak je teda k žien „navyše“, čo je udalosť s pravdepodobnosťou P(k), potrebujeme mať u susedov aspoň k voľných lôžok, čo má pravdepodobnosť Sc(23-k). Pravdepodobnosť toho, že vznikne na pôrodnici pretlak a zároveň bude vedľa dostatok miesta, je

```
>> pkv=P*(flipud(Sc'))

pkv =    0.0089
```

Spolu s pravdepodobnosťou, že k pretlaku vôbec nepríde, je to 0.9996 pravdepodobnosť, že v pôrodnici nenastane neriešiteľná situácia. Inými slovami, s neriešiteľnou situáciou treba rátať v priemere tak raz za 7.6 roka.