Cvičení z matematické statistiky 03

Poslední úprava dokumentu: 13. března 2024.

Rozdělení náhodných veličin - vybraná diskrétní a spojitá rozdělení

Spusťte RStudio, otevřete si nový skript, uložte si ho a nastavte pracovní adresář na složku `matstat`.

S teorií k tomuto cvičení vám může pomoci text *Náhodná veličina a její rozdělení*, který najdete na mých stránkách (`https://web.natur.cuni.cz/uamvt/turcm6am`) mezi materiály ke 3. cvičení, popřípadě si příslušné pojmy připomeňte z přednášky.

Rozcvička

Vytvořte si dva vektory stejné délky (např. o 7 hodnotách) a vykreslete bodový diagram (tzv. scatter plot).

1 Binomické rozdělení

Připomeňme si interpretaci binomického rozdělení. Máme posloupnost n nezávislých dílčích pokusů, v nichž náhodný jev *zdar* nastává s pravděpodobností p. Náhodná veličina s binomickým rozdělením s parametry n, p je dána celkovým počtem zdarů v takové posloupnosti. Platí:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad \mathrm{E}\, X = np, \qquad \operatorname{var} X = np(1-p). \tag{1}$$

1) **Příklad s krajtou**: Předpokládejme, že se z vajíčka krajty vylíhne živý jedinec s pravděpodobností 0.5 a že tato pravděpodobnost nijak nezávisí na tom, co se vylíhne z ostatních vajíček. V následujících úkolech nás bude zajímat pravděpodobnost, s jakou se z 10 vajíček vylíhne určitý počet živých krajt.

(a) Rozmyslete si, že rozdělení veličiny udávající počet vylíhnutých krajt je binomické.

✧ Veličinou, která nás zajímá je počet úspěchů (= živých krajt) z 10 pokusů (= vajíček). To je přesně veličina s binomickým rozdělením a parametry $n = 10$ a $p = 0.5$. Označíme-li si tuto veličinu jako X, pak můžeme symbolicky psát $X \sim Bi(10, 0.5)$.

(b) Jaká je střední hodnota počtu vylíhnutých krajt?

✧ Dle vzorečku (1) je to $\mathrm{E}\, X = np = 10 \cdot 0.5 = 5$.

(c) S jakou pravděpodobností se z 10 vajíček vylíhne **právě** 5 živých krajt?

✧ Zajímá nás $P(X = 5)$. Dle vzorečku (1) vidíme, že

$$P(X = 5) = \binom{10}{5} (0.5)^5 (1-0.5)^{(10-5)} = \binom{10}{5} (0.5)^5 (0.5)^5,$$

✧ což v R můžeme jednoduše spočítat pomocí

```
choose(10,5)*0.5^5*(1-0.5)^(10-5)    # nebo:
dbinom(5, 10, 0.5)
```

(d) S jakou pravděpodobností se z 10 vajíček vylíhne **nejvýše** 5 živých krajt?

✧ Tady je dotaz na $P(X \leq 5)$, což je hodnota distribuční funkce binomického rozdělení v bodě 5. Z teorie víme, že

$$P(X \leq 5) = \sum_{k=0}^{5} P(X = k) = \sum_{k=0}^{5} \binom{10}{k} (0.5)^k (1 - 0.5)^{(10-k)},$$

✧ což v R vypočteme jako

```
pbinom(5, 10, 0.5)
```

(e) S jakou pravděpodobností se z 10 vajíček vylíhne **alespoň** 5 živých krajt?

✧ Zde je dotaz na $P(X \geq 5)$, což můžeme jednoduše převést na předchozí případ pomocí

$$P(X \geq 5) = 1 - P(X \leq 4).$$

✧ V R to pak vypočteme jako

```
1 - pbinom(4, 10, 0.5)
```

(f) Nechte si graficky znázornit pravděpodobnosti jednotlivých hodnot a vypsat číselné hodnoty. Dále si nechte vykreslit distribuční funkci počtu vylíhnutých krajt a připomeňte si, jaké vlastnosti má distribuční funkce diskrétního rozdělení.

✧ Všechny pravděpodobnosti $P(X = k)$ pro $k = 0, 1, \ldots, 10$ lze vypsat a vykreslit pomocí příkazů:

```
kk <- 0:10 # možné hodnoty
dbinom(kk,10,1/2) # psti možných hodnot
cbind("k"=kk,"pst"=round(dbinom(kk,10,1/2),6))
plot(kk,dbinom(kk,10,1/2),pch=16)
for (k in kk) lines(c(k,k),c(0,dbinom(k,10,1/2)))
abline(h=0,col="grey")
```

✧ Distribuční funkci vykreslíme příkazem:

```
plot(stepfun(kk,c(0,pbinom(kk,10,1/2))),pch=16,verticals=FALSE,
        main="Distribuční funkce")
abline(h=0:1,col="grey")
```

(g) Předpokládejme, že v dobrých podmínkách se živá krajta vylíhne s pravděpodobností 0.7.

✧ Jak se změnilo rozdělení počtu krajt?

```
cbind(kk,round(dbinom(kk,10,1/2),6),round(dbinom(kk,10,0.7),6))
plot(kk,dbinom(kk,10,1/2),pch=16,ylab="P(X = kk)", ylim=c(0,0.3))
for (k in kk) lines(c(k,k),c(0,dbinom(k,10,1/2)))
points(kk,dbinom(kk,10,0.7),pch=16,col=2)
for (k in kk) lines(c(k,k),c(0,dbinom(k,10,0.7)),col="red",lty="dotted")
```

✧ Jaká je nyní nejpravděpodobnější hodnota? Jaký je očekávaný počet vylíhnutých krajt?

✧ Nejpravděpodobnější hodnota se v tomto případě rovná střední hodnotě $\mathrm{E}\, X = 10 \cdot 0.7 = 7$, ale nemusí to tak být vždy.

✧ Jaká je nyní pravděpodobnost, že se vylíhne nejvýše 5 živých krajt?

```
pbinom(5, 10, 0.7)
```

2) Příklad - kostka: Házíme 20-krát symetrickou kostkou. Zajímá nás jaká je pravděpodobnost, že z těchto 20 hodů padně určitý počet šestek. Rozmyslete si, že veličina X představující počet šestek ve 20 hodech má binomické rozdělení $Bi(20, \frac{1}{6})$.

- ✧ S jakou pravděpodobností padnou právě 4 šestky?
- ✧ S jakou pravděpodobností padnou nejvýše 4 šestky?
- ✧ S jakou pravděpodobností padne alespoň 5 šestek?

2 Poissonovo rozdělení

Na rozdíl od binomického rozdělení zde není konečný počet dílčích pokusů, v nichž bychom zjišťovali, zda nastal či nenastal sledovaný náhodný jev (*zdar*). Zajímá nás počet výskytů sledovaného náhodného jevu během daného časového intervalu či na dané ploše „jednotkové“ velikosti apod., přičemž počty výskytů v různých časových intervalech (plochách) jsou nezávislé. Možné hodnoty nejsou shora nijak ohraničené, i když jsou velké hodnoty málo pravděpodobné. Střední hodnota i rozptyl jsou stejné, rovné jedinému parametru $\lambda > 0$. Platí tedy:

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad \mathrm{E}\, X = \lambda, \qquad \mathrm{var}\, X = \lambda. \tag{2}$$

3) Příklad - studenti: Předpokládejme, že se počet studentů, kteří dorazí na přednášku v intervalu 0-10 minut před jejím začátkem, řídí Poissonovým rozdělením s parametrem $\lambda = 15$.

- ✧ Připomeňte si, jak vypadá Poissonovo rozdělení. Nechte si znázornit pravděpodobnosti jednotlivých hodnot pro tento příklad.

```
lambda <- 15
kk <- 0:50        # omezíme se na tyto možné hodnoty
dpois(kk,lambda)
plot(kk,dpois(kk,lambda),pch=16)
arrows(x0=kk,y0=0,x1=kk,y1=dpois(kk,lambda),length=0)
abline(h=0,col="grey")
```

- ✧ Který počet studentů je nejpravděpodobnější?
  - ✧ Nejpravděpodobnější je počet 14 a 15 studentů.
- ✧ Spočtěte pravděpodobnost, že na přednášce bude právě 10 studentů.
  - Výpočet proveďte „ručně“ pomocí vzorečku.

    ```
    lambda^10/factorial(10)*exp(-lambda)
    ```

  - K výpočtu použijte funkci dpois(k,lambda) pro dané k a λ.

    ```
    dpois(10,lambda)
    ```

- ✧ Určete pravděpodobnost, s jakou na přednášku nikdo nedorazí.

```
dpois(0,lambda)
```

- ✧ Jaká je pravděpodobnost, že dorazí maximálně pět studentů?

```
ppois(5,lambda)
```

- ✧ S jakou pravděpodobností nebude stačit posluchárna s kapacitou 25 míst (ve smyslu, že někdo nebude mít místo k sezení).

```
1-ppois(25,lambda)
```

4) Příklad - pixely: Z deseti miliónů pixelů obrazovky jsou v průměru dva vadné. Uvažujte obrazovku o rozměru 1280 × 1024 pixelů. Pro modelování počtu vadných pixelů zde lze použít buď binomické rozdělení, nebo Poissonovo, kterým lze binomické rozdělení při malých hodnotách p dobře aproximovat.

- ✧ Porovnejte graficky rozdělení $Bi(n,p)$ a $Po(np)$ pro různá n a malá p. Uvažujte například $n = 20, p = 1/10$, dále $n = 50, p = 1/25$ a nakonec $n = 100, p = 1/50$. Uvidíte, že rozdělení $Bi(n,p)$ lze pro malá p velmi dobře aproximovat pomocí $Po(np)$.

```
par(mfrow=c(2,1))   # umozni mit dva obrazky pod sebou
n <- 50; p <- 1/25
kk <- 0:n
plot(kk,dbinom(kk,n,p),pch=16,ylab="Binom")
arrows(x0=kk,y0=0,x1=kk,y1=dbinom(kk,n,p),length=0)
abline(h=0,col="grey")
plot(kk,dpois(kk,n*p),pch=16, ylab="Poiss")
arrows(x0=kk,y0=0,x1=kk,y1=dpois(kk,n*p),length=0)
abline(h=0,col="grey")
par(mfrow=c(1,1))   # aby byl v grafickem okne opet pouze 1 obrazek
```

- ✧ Porovnejte pravděpodobnosti hodnot 0, 1, ..., 5, když uvažujete binomické rozdělení a jeho aproximaci rozdělením Poissonovým.

```
n <- 1280*1024
p <- 2/10000000
cbind("k"=0:5,"binom."=dbinom(0:5,n,p),"Pois"=dpois(0:5,n*p))
```

- ✧ Jaký je očekávaný počet vadných pixelů na obrazovce?
  Dle vzorečku (2) víme, že $\mathrm{E}\,X = \lambda$, přičemž v tomto případě je $\lambda = n \cdot p = 0.262$.

- ✧ Nechte si vykreslit pravděpodobnosti jednotlivých hodnot. Která hodnota je nejpravděpodobnější?

```
plot(0:5,dbinom(0:5,n,p),col="red",pch=16,xlab="k",
     ylim=c(0,max(dbinom(0:5,n,p))),ylab="dbinom(k,n,p)")
points(0:5,dpois(0:5,n*p),pch=3,col="blue")
legend("topright",legend=c("binomické","Poissonovo"),
        pch=c(16,3),col=c("red","blue"))
abline(h=0,col="grey")
```

- ✧ Jaká je pravděpodobnost, že na obrazovce bude alespoň jeden vadný pixel?

```
1-dpois(0,n*p)
```

- ✧ Jaká je pravděpodobnost, že na obrazovce bude více než jeden vadný pixel?

```
1-ppois(1,n*p)
```

3 Rovnoměrné rozdělení

Rovnoměrné rozdělení na intervalu (a,b) má hustotu tvaru

$$\begin{aligned} f(x) &= \frac{1}{b-a}, \quad \text{pro } x \in (a,b), \\ &= 0, \quad \text{pro } x \notin (a,b). \end{aligned}$$

a distribuční funkci

$$\begin{aligned} F(x) &= \frac{x-a}{b-a}, \quad \text{pro } x \in (a,b), \\ &= 0, \quad \text{pro } x \leq a, \\ &= 1, \quad \text{pro } x \geq b. \end{aligned}$$

Střední hodnota je $\mathrm{E}\,X = \frac{a+b}{2}$ a rozptyl je $\mathrm{var}\,X = \frac{(b-a)^2}{12}$. Medián je stejný jako střední hodnota.

5) **Příklad - metro**: Metro odjíždí každých pět minut. Přijdete v náhodný okamžik, takže lze předpokládat, že se doba čekání na metro řídí rovnoměrným rozdělením na intervalu $(0, 5)$.

✧ Vykreslete si graf hustoty daného rozdělení. Jaké jsou obecné vlastnosti hustoty?

```
curve(dunif(x,0,5),from=-1,to=6,lwd=2)  #  nepřesně!
curve(0*x+1/5,from=0,to=5,ylab="f(x)",xlim=c(-1,6),ylim=c(0,0.35),lwd=2)
lines(c(-1,0),c(0,0),lwd=2)
lines(c(5,6),c(0,0),lwd=2)
abline(h=0,v=0,col="grey")
```

✧ Nakreslete si graf distribuční funkce. Jaké vlastnosti má obecně distribuční funkce a jaká je spojitost mezi distribuční funkcí a hustotou?

```
curve(punif(x,0,5), from=-1, to=6,lwd=2, ylab = "F(x)")
```

✧ Uměli byste v obrázku hustoty (distr. funkce) vyznačit pravděpodobnost, že budete čekat méně než dvě minuty? Čemu je rovna tato pravděpodobnost?

Zajímá nás tedy $P(X \leq 2)$, což je hodnota distribuční funkce rovnoměrného rozdělení $R(0, 5)$ v bodě 2, tj. $F(2)$.

$$P(X \leq 2) = F(2) = \frac{2-0}{5-0} = \frac{2}{5},$$

což v R jednoduše spočítáme jako

```
punif(2,0,5)
```

✧ Jaká je pravděpodobnost, že budeme čekat déle než 4 minuty?

Zde je dotaz na $P(X > 4)$, což můžeme opět spočítat jako:

$$P(X > 4) = 1 - P(X \leq 4),$$

což v R spočteme jako

```
1-punif(4,0,5)
```

✧ Jaká je pravděpodobnost, že čekáním strávíme dobu z intervalu 1 až 5 minut?

Pravděpodobnost, že veličina X nabyde hodnoty z nějakého intervalu (c, d) je rovna ploše po křivkou hustoty v tomto intervalu. A tato plocha se vypočte jakožto určitý integrál z hustoty $f(x)$ přes interval (c, d):

$$P(X \in (c, d)) = P(c < X < d) = \int_c^d f(x)dx.$$

Místo počítání integrálu si ale můžeme rozmyslet, že

$$P(c < X < d) = P(X < d) - P(X < c),$$

tedy jde o rozdíl distribuční funkce v bodě c a v bodě d. V našem případě je $c = 1$ a $d = 5$. Rozdíl distribučních funkcí pak v R spočítáme jako

```
punif(5,0,5) - punif(1,0,5)
```

Možná vás zarazilo, že distribuční funkcí zde nazýváme i pravděpodobnost $P(X < z)$, místo tradičního $P(X \leq z)$. Pravda je, že u spojitých rozdělení je to jedno. Je to díky tomu, že u spojitých rozdělení je pravděpodobnost jednoho konkrétního bodu rovna nule, tj. $P(X = z) = 0$. A tudíž $P(X \leq z) = P(X = z) + P(X < z) = P(X < z)$.

✧ Určete pravděpodobnost, že budete čekat přesně 3 minuty.
Správná odpověď je

$$P(X = 3) = 0.$$

Někdo bude možná zkoušet: `dunif(3, 0, 5)`
ALE POZOR! Pro spojitá rozdělení je $\mathrm{P}(X = x) = 0$. Hustota je pro spojitá rozdělení sice jakousi „náhražkou“ za výraz $\mathrm{P}(X = x)$ u rozdělení diskrétních, avšak má význam pouze při určení pravděpodobnosti, že náhodná veličina X leží v nějakém intervalu (tato pravděpodobnost je pak rovna ploše pod křivkou hustoty v daném intervalu). Hustotu tedy nelze použít pro výpočet pravděpodobnosti, že se X rovná určité hodnotě, tj. $f(x) \neq P(X = x)$. Příkaz dunif(3, 0, 5) nám sice dá nějaké číslo (neboť hustota $f(x)$ je funkcí proměnné x, takže $f(3)$ je „nějaké číslo“), avšak toto číslo nemá význam pravděpodobnosti jevu „doba čekání = 3 minuty“, tj. $f(3) \neq P(\text{doba čekání} = 3)$.

✧ Vypočítejte medián a 10 % a 90 % kvantil. Co tyto hodnoty vyjadřují?

```
qunif(c(0.5,0.1,0.9),0,5)
```

4 Exponenciální rozdělení

Exponenciální rozdělení má hustotu danou předpisem

$$f(x) = \lambda e^{-\lambda x}, \quad \text{pro } x \geq 0, \tag{3}$$
$$= 0, \quad \text{pro } x < 0 \tag{4}$$

a distribuční funkci

$$F(x) = 1 - e^{-\lambda x}, \quad \text{pro } x \geq 0, \tag{5}$$
$$= 0, \quad \text{pro } x < 0. \tag{6}$$

Střední hodnota je $\mathrm{E}\,X = \frac{1}{\lambda}$ a rozptyl je $\mathrm{var}\,X = \frac{1}{\lambda^2}$.

6) Příklad - délka telefonního hovoru: Délka telefonního hovoru se řídí exponenciálním rozdělením se střední hodnotou 5 minut.

✧ Nechte si nakreslit hustotu tohoto rozdělení na intervalu (0, 25).

```
lambda <- 1/5
curve(dexp(x,lambda),0,25,xlab="x (minut)", ylab = "f(x)")
abline(h=0,v=0,col="grey")
```

✧ Jaká je pravděpodobnost, že hovor bude kratší než 3 minuty?

```
lines(c(0,3),c(0,0),lwd=3,col="red")
pexp(3,lambda)
```

✧ Jaká je pravděpodobnost, že hovor bude trvat více než 10 minut?

```
lines(c(10,25),c(0,0),lwd=3,col="green")
1 - pexp(10, lambda)
```

✧ Spočítejte medián (numericky) a srovnejte jej se vzorečkem z přednášky.

```
qexp(0.5, lambda)
log(2)/lambda
```

✧ Jaká je minimální délka 10 % nejdelších hovorů?

```
qexp(0.9, lambda)
```

5 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:

File ➥ **Save**

nebo klávesovou skratkou `Ctrl-s`.