## Cvičení z matematické statistiky 07

Poslední úprava dokumentu: 19. března 2024.

## Maxwellovo a normální rozdělení, QQ diagram, šikmost a špičatost

Do své složky `J:/matstat` si zkopírujte skript `cv04.R` z adresáře `V:/turcicova`. Tento skript následně otevřete v RStudiu a nastavte pracovní adresář pomocí `Session` → `Set Working Directory` → `To Source File Location`.

S teorií k tomuto cvičení vám může pomoci text *Náhodná veličina a její rozdělení* (dále pod zkratkou *NVR*), který najdete na mých stránkách, popřípadě si příslušné pojmy připomeňte z přednášky.

### Rozcvička

Jaká je pravděpodobnost, že z deseti hodů kostkou nám padně alespoň třikrát šestka?

## 1 Maxwellovo rozdělení

Hustota Maxwellova rozdělení s parametrem $a > 0$ má tvar

$$\begin{aligned} f(x) &= \frac{2}{a^3\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2a^2}}, \quad \text{pro } x \geq 0, \\ &= 0, \quad \text{pro } x < 0. \end{aligned} \tag{1}$$

Žádný explicitní vzorec pro distribuční funkci Maxwellova rozdělení bohužel neexistuje, lze ji zapsat pouze pomocí integrálu z hustoty, tj.

$$P(X \leq z) = \int_{-\infty}^{z} f(x)dx = \int_{-\infty}^{z} \frac{2}{a^3\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2a^2}} dx,$$

a tudíž medián i kvantily lze spočíst jen numericky. Střední hodnota je rovna

$$\mathrm{E}\, X = \sqrt{\frac{8}{\pi}} a.$$

Někdy se toto rozdělení nazývá též Maxwell-Boltzmannovo.

1) **Příklad - ideální plyn:** Při snaze o fyzikální popis souboru mnoha částic se ukázalo, že rozdělení jejich rychlostí v ideálním plynu o teplotě $T$ lze velmi dobře popsat funkcí:

$$g(v) = 4\pi \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-\frac{mv^2}{2kT}} v^2, \tag{2}$$

což je hustota Maxwellova rozdělení výše s parametrem $a = \sqrt{kT/m}$, kde $T$ značí teplotu v Kelvinech, $m$ hmotnost částice v kilogramech a $k$ je Boltzmannova konstanta.

Předpokládejme, že máme nádobu s 10 moly $CO_2$ o teplotě 300 K. V případě $CO_2$ je hodnota parametru $a$ rovna

$$a = \sqrt{kT/m} \doteq \sqrt{189T}.$$

✧ Vykreslete si hustotu rozdělní (2) pro různé hodnoty teploty $T$ z intervalu 200 až 500 K.

```
install.packages("shotGroups")
library(shotGroups)
T1 <- 200;   a <- sqrt(189*T1)
curve(dMaxwell(x,a),from = 0, to = 1000, xlab="v [m/s]", ylab= "f(v)",
ylim = c(0,0.003), main = "Maxwellovo rozdělení")
T2 <- 500;   a <- sqrt(189*T2)
curve(dMaxwell(x,a),col = 3, add= TRUE)
legend("topright", legend=c(paste(T1,"K"),paste(T2,"K")),col=c(1,3),lty=c(1,1))
```

Nejspíš jste zpozorovali, že při vyšší teplotě se výška kopečku sníží (hodnoty na ose $y$ klesnou). Je to proto, že celková plocha pod hustotou musí být vždy rovna jedné, a tudíž je nutno tuto plochu i při větším „roztažení“ křivky zachovat. To tedy znamená, že při vyšší teplotě je menší pravděpodobnost, že částice bude cestovat menší rychlostí a naopak jsou upřednostňovány vyšší rychlosti. Při poklesu teploty je naopak velmi pravděpodobné, že částice bude cestovat nízkou rychlostí.

✧ Spočtěte nejpravděpodobnější rychlost $v_p$ a zakreslete si ji do grafu hustoty.

Nejpravděpodobnější rychlost je ta, kde funkce $g(x)$ nabývá maxima, to najdeme jednoduše z podmínky $g'(v) = 0$. Derivace funkce $g$ se rovná

$$g'(v) = 4\pi \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-\frac{mv^2}{2kT}} \left[2v - \frac{mv^3}{kT}\right].$$

Položíme-li tento výraz roven 0, zjistíme, že

$$v_p = \sqrt{\frac{2kT}{m}} = \sqrt{2}a.$$

V případě $CO_2$ o teplotě 300 K je $v_p$ rovna 336.73 m/s.

```
T <- 300
a <- sqrt(189*T)
curve(dMaxwell(x,a),from = 0, to = 800, xlab="v [m/s]", ylab= "f(v)",
main = paste(T,"K"), xaxt="n")
axis(side=1, at=c(0,200,600,800), labels = TRUE)
vp <- sqrt(2)*a
abline(v=vp, lty=2)
axis(side=1, at=vp, labels = expression("v"[p]))
```

✧ Spočtěte střední rychlost $\mathrm{E}\, v$ a přidejte ji do grafu.

Střední rychlost bychom spočítali jako

$$\mathrm{E}\, v = \int_0^\infty v g(v) dv = \ldots = \sqrt{\frac{8}{\pi}}a = \sqrt{\frac{8kT}{\pi m}} = \frac{2}{\sqrt{\pi}} v_p.$$

V případě $CO_2$ o teplotě 300 K je $\mathrm{E}\, v$ rovna 379.96 m/s.

Jelikož $\mathrm{E}\, v > v_p$, tak vidíme, že ačkoli se většina částic pohybuje rychlostí $v_p$, je tam mnoho částic které se pohybují výrazně rychleji.

```
Ev <- 2/sqrt(pi)*vp
abline(v=Ev, lty=3)
axis(side=1, at=Ev, labels = "E(v)")
```

✧ Spočtěte střední kvadratickou rychlost $\mathrm{E}\, v^2$ a přidejte její odmocninu do grafu.

Tato rychlost vlastně představuje 2. moment Maxwellova rozdělení a spočteme ji (při integrování by se využily polární souřadnice) jako

$$\mathrm{E}\, v^2 = \int_0^\infty v g(v) dv = \ldots = \frac{3kT}{m}.$$

Tato rychlost je v praxi velmi důležitá, neboť pomocí ní lze určit střední kinetickou energii částice (molekuly): $\mathrm{E}(E_k) = \frac{1}{2}m\,\mathrm{E}(v^2)$. Aby byla tato rychlost porovnatelná se střední a nejpravděpodobnější rychlostí, zavádí se její odmocnina (*root-mean-square velocity*)

$$v_{rms} = \sqrt{\frac{3kT}{m}} = \sqrt{\frac{3}{2}}v_p.$$

V případě $CO_2$ o teplotě 300 K je $v_{rms}$ rovna 412.41 m/s.

```
v_rms <- sqrt(3/2)*vp
abline(v=v_rms, lty=4)
axis(side=1, at=v_rms, labels = expression("v"[rms]))
```

*Poznámka: Všimněte si, že poměry rychlostí jsou*

$$\frac{v_{rms}}{v_p} = \sqrt{\frac{3}{2}}, \quad \frac{v_{rms}}{\mathrm{E}\,v} = \sqrt{\frac{3\pi}{8}},$$

*a tedy jsou konstantní pro každý plyn bez ohledu na teplotu!*

✧ Jaká je pravděpodobnost nalezení molekuly $CO_2$, která se pohybuje rychlostí mezi 500 a 520 $m/s$? Jaký počet částic v naší nádobě se touto rychlostí pohybuje?

Pravděpodobnost, že rychlost částice nabyde hodnoty z intervalu $(500, 520)$ $m/s$ je rovna ploše pod křivkou hustoty v tomto intervalu, tj.

$$P(v \in (500, 520)) = \int_{500}^{520} g(v)dv = P(v < 520) - P(v < 500),$$

což lze v R snadno spočítat pomocí příkazu

```
a <- sqrt(189*300)
pMaxwell(520,a) - pMaxwell(500,a)
```

Výsledek jsou asi 3 %. Pokud bychom spočítali pravděpodobnost, že částice cestuje rychlostí $(v_p - 10, v_p + 10)$, dostali bychom hodnotu skoro 5 %. To, že je tato pravděpodobnost větší, je patrné už z tvaru hustoty rozdělení.

Abychom zjistili počet molekul v naší nádobě, které cestují rychlostí 500–520 $m/s$, musíme nejprve zjistit celkový počet částic v nádobě. Ten je roven

$$N = n \cdot N_A = 10 \cdot 6.023 \cdot 10^{23} = 6.023 \cdot 10^{24},$$

a tudíž počet molekul pohybujících se rychlostí v rozmezí 500–520 $m/s$ je $0.03 \cdot N \doteq 1.8 \cdot 10^{23}$.

## 2 Normální rozdělení

Mnohé metody, s nimiž se ještě potkáme, předpokládají u kvantitativních dat, že odpovídají normálnímu (Gaussovu) rozdělení. Toto rozdělení odvodil Abraham de Moivre v roce 1733, ale později dostalo název po německém matematikovi C. F. Gaussovi.

Základní charakteristiky normálního rozdělení jste viděli na přednášce, ale můžete si je připomenout v textu *NVR* na str. 10.

Hustota normálního rozdělení (tzv. Gaussova křivka, slangově nazývaná „gaussovka“) je funkce daná předpisem

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

kde $\mu$ a $\sigma^2$ jsou střední hodnota a rozptyl, kteréžto jsou parametry normálního rozdělení. Zajímavé je, že tyto parametry spolu nejsou žádným způsobem provázané. To u jiných rozdělení nenajdeme.

Vykreslete si „gaussovku“ pro různé hodnoty střední hodnoty a rozptylu (resp. směrodatné odchylky) a pozorujte vliv těchto parametrů na její tvar.


Obrázek 1: Německý matematik a fyzik Carl Friedrich Gauss (1777 - 1855) na staré německé marce.

```
strhod = 1       # stredni hodnota, zkousejte ruzne hodnoty
smerodch = 3      # smerodatna odchylka, zkousejte ruzne hodnoty
curve(dnorm(x,strhod,smerodch), from = -30, to = 30, ylab = "f(x)")
curve(dnorm(x,10,7), col=2, add=TRUE)  # pro porovnani pridame dalsi hustotu
```

✧ Jak jste jistě zpozorovali, střední hodnota normálního rozdělení udává střed hustoty (tj. střed „kopečku“). Rozptyl pak udává, jak bude křivka „roztažená“ do stran. Možná jste si také všimli, že při větší hodnotě rozptylu se výška kopečku sníží (hodnoty na ose $y$ klesnou). Je to proto, že celková plocha pod hustotou musí být vždy rovna jedné, a tudíž je nutno tuto plochu i při větším rozptylu (a tedy „roztaženější“ křivce) zachovat.

✧ Jaké jsou obecné vlastnosti hustoty?

✧ V jakém rozmezí leží 50 % hodnot? V jakém **nejkratším** rozmezí leží 50 % všech hodnot?

```
q25 <- qnorm(0.25, strhod, smerodch)
q75 <- qnorm(0.75, strhod, smerodch)
round(c(q25, q75), 3)
curve(dnorm(x,strhod,smerodch), from = -20, to = 20, ylab = "f(x)")
lines(c(q25,q75),c(0,0),col="green",lwd=3)
xx <- seq(q25,q75,length.out=101)
polygon(c(q25,xx,q75),c(0,dnorm(xx,strhod,smerodch),0),col="lightgreen")
```

✧ Dále si vykreslete tvar příslušné distribuční funkce. Jaké vlastnosti má obecně distribuční funkce a jaká je její spojitost s hustotou?

```
curve(pnorm(x,strhod,smerodch), from = -20, to = 20, ylab = "F(x)")
```

**1) Příklad - IQ**: Hodnota IQ koeficientu je tabelována tak, aby bylo jeho rozdělení normální se střední hodnotou 100 a směrodatnou odchylkou 15 (tímto tiše předpokládáme, že každý člověk je charakterizován hodnotou IQ určenou s neomezenou přesností).

Označme si veličinu představující IQ náhodně vybraného člověka jako $X$. Pak $X \sim N(100, 15^2)$. (Druhým parametrem normálního rozdělení je tradičně rozptyl, nikoli směrodatná odchylka, proto musíme 15 umocnit na druhou).

✧ Určeme pravděpodobnost, že IQ náhodně vybraného člověka je nižší než 120.

– Zajímá nás tedy $P(X \leq 120)$, což je hodnota distribuční funkce normálního rozdělení $N(100, 15^2)$ v bodě 120. Žádný explicitní vzorec pro distribuční funkci normálního rozdělení bohužel neexistuje, lze ji zapsat pouze pomocí integrálu z hustoty, tj.

$$P(X \leq z) = \int_{-\infty}^{z} f(x)dx = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx,$$

kde $f(x)$ je hustota $N(\mu, \sigma^2)$ (viz vzorec (3)). Pro nás tedy

$$P(X \leq 120) = \int_{-\infty}^{120} \frac{1}{\sqrt{2\pi 15^2}} e^{-\frac{(x-100)^2}{2 \cdot 15^2}} dx.$$

– Naštěstí tento integrál nemusíme počítat ručně, ale spočítá ho za nás R:

```
pnorm(120, 100, 15)
```

✧ Určeme pravděpodobnost, že IQ náhodně vybraného člověka je vyšší než 110.

– Zde je dotaz na $P(X > 110)$, což můžeme opět spočítat jako:

$$P(X > 110) = 1 - P(X \leq 110),$$

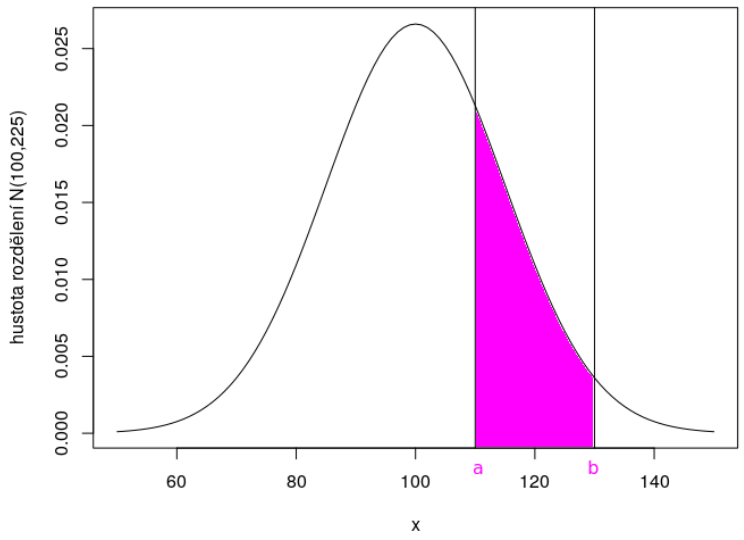
– což v R spočteme jako

```
1 - pnorm(110, 100, 15)
```

✧ Určete pravděpodobnost, že IQ náhodně vybraného člověka leží v intervalu (110, 130).

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

$$P(X \in (a, b)) = P(a < X < b) = \int_{a}^{b} f(x)dx.$$




– Místo počítání integrálu si ale můžeme rozmyslet, že

$$P(a < X < b) = P(X < b) - P(X < a) = F(b) - F(a).$$

– V našem případě je $a = 110$ a $b = 130$, a tudíž v R výsledek spočteme jako

```
pnorm(130, 100, 15) - pnorm(110, 100, 15)
```

✧ Určete pravděpodobnost, že IQ náhodně vybraného člověka je rovno 120.

✧ Jaké nejvyšší IQ musí člověk mít, aby patřil k 10 % nejméně chytrých v populaci?

```
qnorm(0.1, 100, 15)
```

✧ Jaké nejmenší IQ musí člověk mít, aby patřil k 5 % nejchytřejších v populaci?

```
qnorm(0.95, 100, 15)
```

✧ V jakém nejkratším intervalu se nachází IQ 95 % populace? Vykreslete si tyto meze do grafu hustoty.

```
curve(dnorm(x,100,15),from=60,to=140,lwd=2)   # graf hustoty
abline(h=0,col="grey")
# vykresleni prislusne plochy
lines(qnorm(c(0.025,0.975),100,15),c(0,0),lwd=3,col="green")
xx <- seq(qnorm(0.025,100,15),qnorm(0.975,100,15),length.out=101)
polygon(c(qnorm(0.025,100,15),xx,qnorm(0.975,100,15)),
c(0,dnorm(xx,100,15),0),col="lightgreen")
```

## 3 QQ diagram

Shodu rozdělení dat s předpokládaným rozdělením lze graficky nejlépe posoudit pomocí kvantil-kvantilového grafu (angl. *QQ plot*, což je zkratka pro *quantile-quantile plot*). Chceme-li tedy posoudit, zda by naše data mohla pocházet z normálního rozdělení, je nejlepším obrázek právě QQ plot, v případě normálního rozdělení nazývaný *normální diagram*.

V normálním diagramu jsou na ose $x$ teoretické kvantily normálního rozdělení $N(0,1)$ (ty jsou známy) a na ose $y$ jsou výběrové kvantily vypočítané z příslušných dat (výběrové kvantily jsme počítali na předchozích cvičeních pomocí příkazu `quantile`). Počet vykreslených kvantilů je roven počtu našich dat. Pocházejí-li naše data z normálního rozdělení (ne nutně jen $N(0,1)$), měly by si teoretické a výběrové kvantily odpovídat a body v normálním diagramu by měly ležet na přímce. Pokud data z normálního rozdělení nepocházejí, budou se body v normálním diagramu kolem přímky různě vlnit, popř. utíkat na koncích.

Nenechte se zmást tím, že výběrové kvantily v normálním diagramu porovnáváme s teoretickými kvantily zrovna $N(0,1)$ rozdělení. Pokud by naše data pocházela z normálního rozdělení s nějakou jinou střední hodnotou a jiným rozptylem (tj. obecně $N(\mu,\sigma^2)$), bude mít přímka z bodů akorát jiný sklon, ale její tvar zůstane zachován.

Posouzení normality na základě normálního diagramu je samozřejmě zcela subjektivní. Později budeme normální rozdělení dat rigorózně testovat pomocí **Shapiro-Wilkova testu**.

✧ Nakreslete **normální diagram** pro náhodný výběr z normálního rozdělení:

```
vyber <- rnorm(25, 100, 15)
qqnorm(vyber)
qqline(vyber)
```

✧ Podívejme se, jak by tento diagram vypadal pro výběr například z exponenciálního rozdělení

```
vyber2 <- rexp(1000,50)
qqnorm(vyber2)
qqline(vyber2)
```

## 4 Šikmost a špičatost

Význam pojmů hustota, šikmost a špičatost, které pro nás budou nyní důležité, se můžete podrobněji připomenout v textu *Náhodná veličina a rozdělení*, který najdete na mých stránkách mezi

materiály k jednomu z předchozích cvičení.

Teoretická šikmost normálního rozdělení je 0 (neboť jeho hustota je symetrická), tedy data pocházející z normálního rozdělení by měla mít hodnotu výběrové špičatosti (kterou budeme počítat níže) kolem nuly.

Teoretická špičatost normálního rozdělení je rovna 0, tedy u dat pocházejících z normálního rozdělení by se výběrová špičatost (vypočtená z dat) měla pohybovat kolem této hodnoty.

Spočtěme **výběrovou šikmost a špičatost** pro náš výběr. Pro připomenutí:

$$a_3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^3 \qquad \text{(výběrová šikmost)},$$

$$a_4 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^4 - 3 \quad \text{(výběrová špičatost)}.$$

✧ Veličiny

$$z_i = \frac{x_i - \bar{x}}{s_x} \quad (i = 1, \ldots, n)$$

se někdy nazývají z-skóry a v R je lze získat pomocí funkce `scale`. Výběrovou šikmost a špičatost proto snadno spočteme pomocí následujících příkazů:

```
mean(scale(vyber)^3)          # sikmost
mean(scale(vyber)^4) - 3      # spicatost
```

✧ Má-li veličina $X$ rozdělení $N(\mu, \sigma^2)$, jaké rozdělení má veličina $Z = \frac{X-\mu}{\sigma}$?

✧ Nakreslete si histogram vypočtených z-skórů a porovnejte ho s hustotou rozdělení $N(0, 1)$.

```
hist(scale(vyber), prob=TRUE)
curve(dnorm(x), col="red", add=TRUE)
```

## 5 Náhodné vektory

**2) Příklad - dvojrozměrné normální rozdělení**: Vygenerujte si náhodný výběr o rozsahu 100 z dvojrozměrného normálního rozdělení $N_2(\binom{0}{3}, V)$, kde varianční matice $V$ má tvar

$$V = \begin{pmatrix} 2 & 1.3 \\ 1.3 & 5 \end{pmatrix}.$$

```
install.packages("mnormt")
library(mnormt)
(V <- matrix(c(2, 1.3, 1.3, 5), nrow=2))
(Mu <- c(0, 3))
bvn <- rmnorm(100, Mu, V)
```

✧ Vykreslete si hustotu rozdělení $N_2(\binom{0}{3}, V)$.

```
x <- seq(-4, 6, 0.1)
y <- seq(-4, 9, 0.1)
f <- function(x, y) dmnorm(cbind(x, y), Mu, V)
z <- outer(x, y, f)
persp(x, y, z, theta=-20, phi=25, expand=0.6, ticktype='detailed')
```

✧ Nagenerované náhodné vektory (náš výběr bvn) vykreslete do bodového grafu a porovnejte s grafem vrstenic.

```
contour(x, y, z)   # graf vrstenic
plot(bvn[,1],bvn[,2],xlim=c(-4,6),ylim=c(-4,9))   # bodovy graf (scatter plot)
```

## 6 Konec práce

Než zavřete všechna okna, nezapomeňte si uložit poslední změny ve skriptovém souboru:
File ➥ Save
nebo klávesovou skratkou `Ctrl-s`.