PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA

Příklady

prof. RNDr. Daniela Jarušková, CSc.
RNDr. Martin Hála, CSc.

2016
České vysoké učení technické v Praze

(1) TITULNÍ LIST

ISBN 978-80-01-04828-3

(2) RUB TITULNÍHO LISTU

Předmluva

Skripta *Pravděpodobnost a matematická statistika 12 – příklady* jsou zamýšlena jako sbírka příkladů ke skriptu Matematická statistika od D. Jaruškové a jsou určena především studentům Stavební fakulty ČVUT. Našim přáním bylo vyhovět požadavku studentů, kteří chtěli lépe zvládnout předmět matematická statistika a poukazovali na to, že skripta Matematická statistika díky omezenému počtu stran neobsahují dostatečné množství příkladů.

Předkládaná skripta jsou rozdělena do kapitol: 1. Pravděpodobnost, 2. Diskrétní rozdělení, 3. Spojité rozdělení, 4. Náhodné vektory, 5. Centrální limitní věta, 6. Popisná statistika, 7. Teorie odhadu, 8. Testování hypotéz, 9. Analýza rozptylu, 10. Chí-kvadrát test, 11. Regrese a 12. Časové řady. Z předchozího výčtu je patrné, že čtenář najde i příklady k některým částem statistiky, které nejsou zahrnuty do skript Matematická statistika, ale které se na některých oborech či doktorandském studiu vyučují. Každá kapitola obsahuje dvě části - řešené a neřešené příklady. Mezi řešené příklady jsme zahrnuli kromě jednoduchých ukázkových příkladů i poněkud složitější příklady, které však více odpovídají problémům řešeným v praxi. Do neřešených příkladů jsme zařadili spíše jednodušší příklady nebo příklady, které jsou analogické řešeným problémům v první části. Najít dobré ukázkové příklady, zvlášť z matematické statistiky, není jednoduché. Často jsme sahali po příkladech z knih a skript našich kolegů z jiných univerzit a fakult. Kromě toho jsme použili i problémy, s kterými jsme se během své statistické praxe setkali. Reálné soubory dat však bývají často velmi rozsáhlé a zpracovat je bez použití počítačů je téměř nemožné. Při použití numericky náročnějších statistických postupů jsme používali program Matlab a jeho statistický toolbox. Stejný program jsme také používali k výpočtu kritických hodnot, které se mohou díky rozdílné numerické přesnosti lišit od kritických hodnot tabelovaných v Numerických tabulkách ke skriptu Matematická statistika či kritických hodnot z jiných statistických programů. Pokud si zvídavější čtenář bude chtít ověřit vlastním výpočtem některý z rozsáhlejších příkladů, rádi mu zpřístupníme příslušná data pro počítačové zpracování.

Očekáváme námitku, proč nejsou u neřešených příkladů uvedeny výsledky. Důvodů je více. Jednak hodláme neřešené příklady částečně využívat k zápočtům a zkouškám, přičemž na některých oborech zkoušející povoluje jakékoliv pomůcky. Podstatnějším důvodem je však podle našeho názoru skutečnost, že u většiny zejména statistických příkladů je obtížné, ne-li nemožné stručný a jednoznačný výsledek uvést. (Výsledek testu závisí na hladině významnosti, někdy lze užít i více srovnatelných postupů, výsledkem některých příkladů je grafický výstup, úvaha, důkaz, atd.) Domníváme se, že neuvedení výsledků neřešených příkladů je kompenzováno dostatečným (co do počtu stran dokonce významně převažujícím) množstvím řešených příkladů.

Věříme, že počítačová technika bude v budoucnu pro studenty dostupnější, což umožní řešit složitější statistické problémy. Doufáme, že zpracování reálných dat přispěje k pocitu, že matematická statistika je užitečná i zábavná věda.

Na závěr bychom rádi poděkovali Doc. RNDr. Gejzovi Dohnalovi, CSc. za pečlivé lektorování našich skript, RNDr. Jaromíru Antochovi, CSc. za významnou pomoc se sazbou a cenné rady a Dr. RNDr. Janě Noskové za cenné připomínky.

Obsah

1. Pravděpodobnost

Řešené příklady

Příklad 1.1.

Tři muži a šest žen se náhodně rozdělí na tři stejně početné skupiny. S jakou pravděpodobností bude v každé z nich jeden muž?

Řešení:

Můžeme si představit, že při rozdělování do skupin si každý vylosuje číslo od 1 do 9. Přitom ti, kteří si vylosovali čísla 1, 2, 3, vytvoří 1. skupinu, osoby mající čísla 4, 5, 6 vytvoří 2. skupinu a zbývající vytvoří 3. skupinu.

Elementárním jevem pro nás bude konkrétní (neuspořádaná) trojice čísel, vylosovaná muži. Tyto elementární jevy jsou zřejmě stejně pravděpodobné a je jich celkem $\binom{9}{3} = 84$. Elementární jevy příznivé zadání příkladu jsou takové trojice, z nichž jedno číslo je z množiny $\{1, 2, 3\}$, jedno číslo z množiny $\{4, 5, 6\}$ a jedno z množiny $\{7, 8, 9\}$. Těchto trojic je zřejmě $3 \cdot 3 \cdot 3 = 27$. Hledaná pravděpodobnost je tedy rovna

$$P = 27/84 = 0.3214.$$

Příklad 1.2.

Vybíráme z pěti vstupenek po 10 Kč, tří vstupenek po 30 Kč a dvou vstupenek po 50 Kč. Vylosujeme-li si náhodně 3 vstupenky, s jakou pravděpodobností bude jejich cena 70 Kč?

Řešení:

Při náhodném losování můžeme vybrat celkem $\binom{10}{3} = 120$ různých trojic (pořadí v trojici pro nás není důležité). Celková cena námi vybrané trojice bude 70 Kč v následujících případech:

a) vybereme 1 vstupenku za 50 Kč a 2 po 10 Kč, tento jev můžeme nakombinovat $2 \cdot \binom{5}{2} = 20$ způsoby,

b) vybereme 2 vstupenky po 30 Kč a 1 za 10 Kč, toho dosáhneme $\binom{3}{2} \cdot 5 = 15$ způsoby.

Pravděpodobnost našeho jevu tedy je

$$P = \frac{20 + 15}{120} = 0.2917.$$

Příklad 1.3.

Házíme-li pěti kostkami, může nastat jedna ze sedmi možností:

a) Všechny kostky ukazují různý počet oček.

b) Právě dvě kostky ukazují stejný počet a všechny ostatní jsou různé.

c) Právě tři kostky ukazují stejný počet a zbylé dvě jsou různé.

d) Právě čtyři kostky ukazují stejný počet oček.

e) Všech pět ukazuje stejný počet oček.

f) Právě dvě a dvě jsou stejné a poslední se liší.

g) Jedna dvojice je stejná a jedna trojice je stejná.

Jakou pravděpodobnost mají náhodné jevy a) - g)?

Řešení:

Spočtěme, kolik je příznivých výsledků odpovídajících jevům a) - g).

a) Na první kostce může padnout jednička, dvojka, ... šestka. Jestliže se má počet oček na druhé kostce lišit, pak možných výsledků na druhé kostce je pouze pět. Počet výsledků na třetí kostce lišících se od výsledku na první i druhé kostce je roven čtyřem atd. Všech příznivých výsledků je $6 \times 5 \times 4 \times 3 \times 2 = 720$.

b) V případě, že právě dvě kostky ukazují stejně a všechny ostatní se liší, musíme nejprve vybrat, které dvě kostky z pěti budou ukazovat stejně. Počet možností, jak vybrat dvojici, je $\binom{5}{2}$. Počet všech příznivých výsledků je $\binom{5}{2} \times 6 \times 5 \times 4 \times 3 = 3600$.

Obdobně spočteme počet příznivých výsledků pro ostatní případy:

c) tři stejné, ostatní různé: $\binom{5}{3} \times 6 \times 5 \times 4 = 1200$,

d) čtyři stejné, jedna různá: $\binom{5}{4} \times 6 \times 5 = 150$,

e) všechny stejné: $= 6$,

f) dvě a dvě stejné, jedna různá: $\binom{5}{2} \times \binom{3}{2} \times \frac{1}{2} \times 6 \times 5 \times 4 = 1800$ (zde součin $\binom{5}{2} \times \binom{3}{2} \times \frac{1}{2} = 15$ označuje počet všech možných seskupení dvou dvojic kostek, které ukazují stejně),

g) dvě stejné a tři stejné: $\binom{5}{2} \times 6 \times 5 = 300$.

Celkem všech možných výsledků je 7776.

Jestliže předpokládáme, že na všech kostkách padá jakákoliv strana se stejnou pravděpodobností 1/6, pak pravděpodobnosti náhodných jevů a) – g) dostaneme jako podíl počtu příznivých výsledků ku počtu všech možných výsledků.

Odpověď: Hledané pravděpodobnosti jsou postupně rovny: 720/7776, 3600/7776, 1200/7776, 150/7776, 6/7776, 1800/7776, 300/7776.

Příklad 1.4.

Dva hráči hrají sérii her o částku (sázku) C, přičemž tuto částku získá ten hráč, který jako první vyhraje k her. Pravděpodobnost výhry každé jednotlivé hry je pro oba hráče stejná (oba hráči jsou stejně dobří). Série her je předčasně ukončena ve chvíli, kdy jednomu hráči chybí do výhry jedna hra a druhému dvě hry. Jaké je spravedlivé rozdělení částky C mezi hráče?

Řešení:

Spravedlivé rozdělení částky odpovídá poměru, v jakém jsou pravděpodobnosti výhry jednotlivých hráčů, kdyby ve hře pokračovali. Při dalším pokračování hry by v první hře s pravděpodobností 1/2 vyhrál první hráč a s pravděpodobností 1/2 by vyhrál druhý hráč. Kdyby vyhrál druhý hráč, bylo by třeba sehrát ještě jednu hru. Tu by opět mohl s pravděpodobností 1/2 vyhrát první hráč a s pravděpodobností 1/2 druhý hráč. Odtud vyplývá, že pokud by hráči pokračovali ve hře až do konce, vyhrál by první hráč s pravděpodobností 3/4 a druhý s pravděpodobností 1/4.

Odpověď: Spravedlivé rozdělení částky C je v poměru 3:1 ve prospěch prvního hráče.

Příklad 1.5.

Bridž se hraje s 52 bridžovými kartami (13 piků, 13 srdcí, 13 kár, 13 trefů). Všechny karty se rozdají mezi čtyři hráče. Vždy dva a dva hrají spolu. Označme pro jednoduchost jednu dvojici: „sever–jih" a druhou: „východ–západ". Hra se hraje tak, že jeden hráč vynese kartu, poté přidá kartu hráč po levici a tak dále ve směru hodinových ručiček, až dají kartu všichni čtyři. Jeden zdvih tvoří tedy čtyři karty. Zdvih v beztrumfové hře získává ten hráč, jehož karta ve zdvihu byla ve vynesené barvě nejvyšší. Dále vynáší ten hráč, který získal předchozí zdvih. Barva se musí přiznávat, tj. má-li hráč kartu ve vynesené barvě, musí ji dát, nemusí však přebíjet, tj. dát kartu vyšší hodnoty. Celou hru tvoří 13 zdvihů, z nichž každý obsahuje 4 karty. Účelem hry je získat co nejvíce zdvihů. Po licitaci následuje sehrávka. Sehrávku hraje hlavní hráč (řekněme jih), který vylicitoval hru. Do prvního zdvihu však vždy vynáší hráč po levici hlavního hráče (západ), poté partner hlavního hráče (sever) položí karty na stůl a hře dále jen pasivně přihlíží. Hlavní hráč hraje se svými i partnerovými kartami, které jsou vyloženy na stole.

Předpokládejme, že první zdvih byl odehrán v trefové barvě, přičemž nejvyšší kartu dal jih a má tedy vynést do druhého zdvihu. Jih vidí, že v srdcové barvě leží v partnerových kartách (na stole) eso, dáma, junek, pětka a čtyřka. Sám má v ruce srdcovou desítku, devítku, osmičku, sedmičku a šestku. To znamená, že protihráči (východ a západ) mají 3 srdcové karty, a to krále, trojku a dvojku.

Jaká je pravděpodobnost, že hlavní hráč získá všechny srdcové zdvihy, jestliže při druhém výnosu vynese srdcové eso?

Řešení:

Hlavní hráč získá všechny srdcové zdvihy tehdy, jestliže je srdcový král u jednoho z protihráčů osamocen. Protihráči mají po odehrání prvního zdvihu dohromady 24 karet. Všech možných rozdělení těchto karet mezi ně je $\binom{24}{12}$. Příznivé výsledky jsou jednak takové, kdy má východ srdcového krále, ale nemá dvojku a trojku srdcovou. Takových možností je $\binom{21}{11}$, neboť $\binom{21}{11}$ je počet možností, jak dovybrat k srdcovému králi 11 nesrdcových karet. Dále jsou příznivé možnosti ty, kdy má východ srdcovou dvojku a trojku, ale nemá srdcového krále. Těch je $\binom{21}{10}$, neboť tolik je možností, jak k srdcové dvojce a trojce dovybrat 10 nesrdcových karet. Celkově je tedy příznivých výsledků $\binom{21}{11} + \binom{21}{10}$. Hledaná pravděpodobnost je rovna

$$\frac{\binom{21}{11} + \binom{21}{10}}{\binom{24}{12}} = 0.261.$$

Odpověď: Vynese-li hlavní hráč při druhém výnosu srdcové eso, získá všechny srdcové zdvihy s pravděpodobností 0.261.

Příklad 1.6.

V osudí je 5 černých koulí a 15 bílých koulí. Z osudí se náhodně vytáhne jedna koule. Poté se vrátí zpět, přidá se 20 koulí téže barvy, jakou měla vytažená koule, a tah se opakuje. Jaká je pravděpodobnost, že druhá vytažená koule bude černá?

Řešení:

Označme jevy:

A ... druhá vytažená koule je černá,
B ... první vytažená koule je černá,
$\bar{B}$ ... první vytažená koule je bílá.

Jev $\bar{B}$ je opačný k jevu B. Platí $P(B) = \frac{5}{20}$ a $P(\bar{B}) = \frac{15}{20}$. Dále platí pro podmíněné pravděpodobnosti $P(A|B) = \frac{25}{40}$ a $P(A|\bar{B}) = \frac{5}{40}$. Podle věty o úplné pravděpodobnosti

$$P(A) = P(A|B)\,P(B) + P(A|\bar{B})\,P(\bar{B}) = \frac{25}{40} \times \frac{5}{20} + \frac{5}{40} \times \frac{15}{20} = \frac{1}{4}.$$

Všimněte si, že pravděpodobnost vytažení černé koule v druhém tahu je stejná jako byla v prvním tahu. To bude platit i pro jiný poměr bílých a černých koulí a jiný počet koulí, které přidáváme.

Odpověď: Pravděpodobnost, že v druhém tahu vytáhneme černou kouli, je rovna 0.25.

Příklad 1.7.

Tři střelci, jejichž dlouhodobé úspěšnosti (tj. pravděpodobnosti zásahu) jsou po řadě 20%, 40% a 60%, současně vystřelili na terč. V terči byl poté zjištěn jeden zásah. S jakou pravděpodobností se strefil 3. střelec?

Řešení:

Uvažujme následující náhodné jevy:
P ... první střelec zasáhne terč,
D ... druhý střelec zasáhne terč,
T ... třetí střelec zasáhne terč,
Z ... při současném výstřelu všech střelců nastane právě jeden zásah.

Zajímá nás podmíněná pravděpodobnost $P(T|Z)$. Můžeme postupně vypočítat:

$$\begin{aligned} P(Z) &= P(P \cap \bar{D} \cap \bar{T}) + P(\bar{P} \cap D \cap \bar{T}) + P(\bar{P} \cap \bar{D} \cap T) = \\ &= 0.2 \times 0.6 \times 0.4 + 0.8 \times 0.4 \times 0.4 + 0.8 \times 0.6 \times 0.6 = 0.464, \\ P(T|Z) &= \frac{P(T \cap Z)}{P(Z)} = \frac{P(\bar{P} \cap \bar{D} \cap T)}{0.464} = 0.6207. \end{aligned}$$

Při výpočtu jsme mlčky předpokládali nezávislost mezi jevy P, D a T, což odpovídá rozumnému předpokladu, že střelci se při střelbě navzájem nijak neovlivňují.

Odpověď: Víme-li, že nastal jeden zásah, pak tento zásah pochází s pravděpodobností 0.6207 od třetího střelce.

Příklad 1.8.

Laboratoř, která provádí rozbory krve, potvrdí s pravděpodobností 95% existenci protilátek na virus určité nemoci, jestliže jí pacient skutečně trpí. Zároveň test určí jako pozitivní 1% osob, které však touto nemocí netrpí. Jestliže 0.5% populace trpí zmíněnou nemocí, jaká je pravděpodobnost, že určitá osoba, jejíž test byl pozitivní, skutečně onu nemoc má?

Řešení:

Pracujeme se čtyřmi náhodnými jevy:
P ... pacientův test je pozitivní,
N ... pacientův test je negativní,
D ... pacient trpí zjišťovanou nemocí,
H ... pacient netrpí zjišťovanou nemocí.

Zajímá nás podmíněná pravděpodobnost $P(D|P)$. Podle Bayesovy věty:

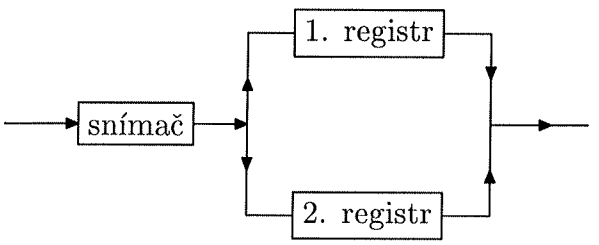
$$P(D|P) = \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|H)P(H)} = \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} = 0.323.$$

Skutečnost, že zjišťovaná pravděpodobnost je poměrně malá, je způsobena tím, že nemoc je málo rozšířená a mezi pozitivní se dostane omylem velké množství zdravých osob.

Odpověď: V takto prováděném testu je pravděpodobnost toho, že pacient skutečně nemocí trpí, rovna pouze 0.323.

Příklad 1.9.

Měřící zařízení sestává ze tří přístrojů – snímače a dvou registrů, které pracují zdvojeně tak, že se při poruše 1. registru automaticky zapojí 2. registr. Schéma zapojení je dáno následujícím obrázkem:



Pravděpodobnost poruchy snímače je 0.01, pravděpodobnost poruchy 1. registru je 0.05, resp. 2. registru 0.02. Určete pravděpodobnost, že snímací zařízení bude jako celek fungovat (za předpokladu, že poruchy jednotlivých zařízení vznikají nezávisle na sobě).

Řešení:

Označme si následující náhodné jevy:

S ... funguje snímač,
R_1 ... funguje 1. registr,
R_2 ... funguje 2. registr.

(Pruh pak označuje opačný jev, tj. $\bar{R}_1$ např. znamená jev, že 1. registr má poruchu.)

Zařízení bude zřejmě fungovat, bude-li fungovat snímač a alespoň jeden z obou registrů. Proto je hledaná pravděpodobnost rovna

$$P = P\{S \cap [R_1 \cup (\bar{R}_1 \cap R_2)]\} = P(S) \cdot [P(R_1) + P(\bar{R}_1) \cdot P(R_2)] =$$
$$= 0.99 \times (0.95 + 0.05 \times 0.98) = 0.989.$$

Odpověď: Pravděpodobnost fungování celého zařízení je přibližně 98.9%. Všimněme si, že tato pravděpodobnost je téměř stejná, jako pravděpodobnost fungování samotného snímače, i když jednotlivé registry nejsou příliš spolehlivé. Díky jejich zálohování se však pravděpodobnost poruchy celého zařízení silně zmenší.

Příklad 1.10.

Na základě úmrtnostních tabulek za rok 1996, publikovaných Českým statistickým úřadem, se odhaduje, že pravděpodobnosti úmrtí ve věku $x = 50, 51, \ldots, 59$ let u mužů

v České republice jsou čísla $q(x)$ uvedená v následující tabulce:

x	50	51	52	53	54
$q(x)$	0.008394	0.009557	0.010552	0.011321	0.012235

x	55	56	57	58	59
$q(x)$	0.013165	0.014106	0.015103	0.017251	0.018743

(Úmrtím ve věku x se rozumí situace, kdy osoba, která je v den svých x-tých narozenin naživu, zemře během následujícího roku, tj. nedožije se $(x+1)$-ních narozenin. Tedy např. číslo $q_{54} = 0.012235$ můžeme interpretovat tak, že z určitého velkého počtu 54-ti letých mužů se přibližně 1.22% nedožije 55 let.)

Odhadněte pravděpodobnost, že:
a) 50-ti letý muž se dožije 60 let,
b) 50-ti letý muž se dožije 55 let, ale nedožije se 60 let.

Řešení:

Je vhodné si uvědomit, že čísla $q(x)$ jsou vlastně podmíněné pravděpodobnosti, že určitý náhodně vybraný muž (z populace všech novorozenců) se nedožije věku $x+1$ let, víme-li, že byl naživu ve věku x let. Zavedeme-li si označení pro náhodné jevy:

Z_x ... osoba mužského pohlaví se dožije věku x, $(x = 1, 2, \dots)$

potom je

$$q(x) = P(\bar{Z}_{x+1}|Z_x) = \frac{P(\bar{Z}_{x+1} \cap Z_x)}{P(Z_x)},$$

$$p(x) = 1 - q(x) = P(Z_{x+1}|Z_x) = \frac{P(Z_{x+1} \cap Z_x)}{P(Z_x)} = \frac{P(Z_{x+1})}{P(Z_x)}.$$

Zde je $p(x)$ podmíněná pravděpodobnost, že se muž, který je naživu ve věku x let, dožije věku $x+1$ let. Poslední rovnost plyne z faktu, že $Z_{x+1} \subset Z_x$.

ad a) Podle pravidel pro počítání s podmíněnými pravděpodobnostmi a díky faktu, že $Z_{60} \subset Z_{59} \subset \cdots \subset Z_{50}$ platí:

$$P(Z_{60}|Z_{50}) = \frac{P(Z_{60})}{P(Z_{50})} = \frac{P(Z_{60})}{P(Z_{59})} \cdot \frac{P(Z_{59})}{P(Z_{58})} \cdots \frac{P(Z_{51})}{P(Z_{50})} = p(59) \cdot p(58) \cdots p(50) =$$
$$= 0.981257 \times 0.982749 \times \cdots \times 0.991606 = 0.8769.$$

ad b) Hledanou pravděpodobnost můžeme vypočítat takto:

$$P = P(\bar{Z}_{60} \cap Z_{55}|Z_{50}) = \frac{P(\bar{Z}_{60} \cap Z_{55})}{P(Z_{50})} = \frac{P(\bar{Z}_{60} \cap Z_{55})}{P(Z_{55})} \cdot \frac{P(Z_{55})}{P(Z_{50})} =$$
$$= [1 - P(Z_{60}|Z_{55})] \cdot P(Z_{55}|Z_{50}) = (1 - 0.92404) \times 0.94901 = 0.0721.$$

(Pravděpodobnosti $P(Z_{55}|Z_{50})$ a $P(Z_{60}|Z_{55})$ jsme vypočítali podobně jako v bodě a.)

Neřešené příklady

Příklad 1.11.

Jaká je pravděpodobnost, že náhodně vybrané dvojciferné číslo je dělitelné dvěma nebo pěti?

Příklad 1.12.

Krychle, která má všechny stěny obarveny, je rozřezána na tisíc krychliček o stejných rozměrech. Takto získané krychličky se pečlivě promíchají. Určete pravděpodobnost, že namátkou vybraná krychlička bude mít dvě stěny obarveny.

Příklad 1.13.

Tři muži sedí u baru. Aby mohli rozhodnout, kdo zaplatí útratu, každý z nich hází mincí. Muž, jehož mince ukáže jinou stranu než zbývajících dvou, platí útratu. Předpokládejme, že na všech mincích obě dvě strany padají se stejnou pravděpodobností. Jaká je pravděpodobnost, že k rozhodnutí, kdo bude platit, dojde již poté, co muži hodí poprvé?

Příklad 1.14.

Telefonní číslo je pěticiferné. Najděte pravděpodobnost toho, že všechny cifry budou různé. (První cifra nesmí být nula.)

Příklad 1.15.

Určete pravděpodobnost, že při hodu dvěma kostkami padne (padnou):
a) dvě šestky,
b) dvě stejná čísla,
c) součet větší než 7.

Příklad 1.16.

V losovací urně je celkem 12 koulí, z toho 5 zelených, 4 červené a 3 modré. Náhodně (najednou) vybereme 3 koule. S jakou pravděpodobností budou mít všechny různou barvu?

Příklad 1.17.

Náhodně vybereme trojici čísel z množiny $\{1, 2, \ldots, 10\}$ (vybíráme s vyloučením opakování). S jakou pravděpodobností lze z vybraných čísel sestavit první tři členy aritmetické posloupnosti?

Příklad 1.18.

Deset aut zaparkuje náhodně vedle sebe. Jaká je pravděpodobnost, že zvolená tři auta budou spolu sousedit?

Příklad 1.19.

Na šachovém turnaji hraje 20 hráčů, kteří jsou náhodně rozděleni do dvou skupin po deseti. Jaká je pravděpodobnost, že čtyři nejlepší budou hrát spolu?

Příklad 1.20.

Mějme úsečky o délkách 1, 3, 5, 7, 9. Jaká je pravděpodobnost, že ze třech náhodně vybraných úseček je možno sestrojit trojúhelník?

Příklad 1.21.

Při hře poker dostane každý pět karet. Jaká je pravděpodobnost, že dostanete do ruky pětici po sobě jdoucích karet, které však nejsou stejné barvy?

Příklad 1.22.

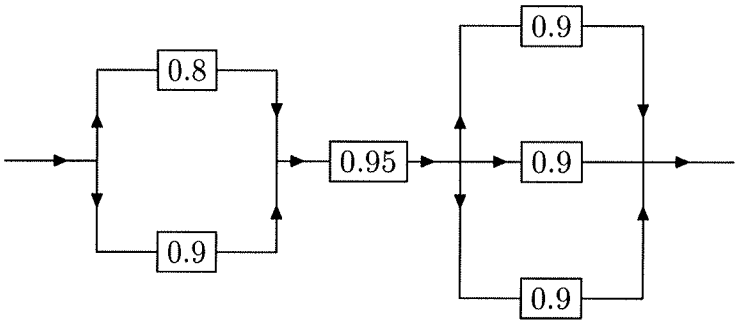
Z balíčku 32 karet vylosujeme kartu. A označuje náhodný jev, že karta je eso a B, že karta je červené barvy. Jsou jevy A a B nezávislé?

Příklad 1.23.

Ze zkušenosti víme, že se určité zařízení během hodiny porouchá s pravděpodobností 0.2, přičemž pravděpodobnost vzniku poruchy je v čase konstantní. Po dvou hodinách od uvedení do provozu bylo zjištěno, že zařízení nepracuje. S jakou pravděpodobností nastala porucha již během první hodiny?

Příklad 1.24.

Signalizační zařízení se skládá ze tří sériově zapojených okruhů, ve dvou z nich jsou paralelně zapojeny navzájem se zálohující prvky. Spolehlivosti jednotlivých prvků jsou přímo vyznačeny na schématu:



Určete pravděpodobnost, že signalizační zařízení bude mít poruchu (za předpokladu, že poruchy jednotlivých prvků vznikají nezávisle na sobě).

Příklad 1.25.

Určitý systém sestává ze dvou paralelně zapojených zařízení. Pravděpodobnosti výskytu poruchy na těchto zařízeních jsou 0.05 a 0.02, pravděpodobnost výpadku v dodávce elektřiny (na které závisí chod obou zařízení) je 0.04. Určete za předpokladu nezávislosti vzniku všech poruch, že celý systém bude fungovat, tj. že bude fungovat alespoň jedno z obou zařízení.

Příklad 1.26.

Sonda má dvě kamery, které mohou pracovat nezávisle na sobě. Každá z nich je vybavena pro případ poruchy korekčním mechanismem. Pravděpodobnost poruchy kamery je 0.1, pravděpodobnost úspěšné opravy případné poruchy pomocí korekčního mechanismu je 0.3. S jakou pravděpodobností se nepodaří ani jednou z kamer nic nafilmovat?

Příklad 1.27.

Na úseku AB má motocyklový závodník 12 překážek. Na každé z nich bude muset s pravděpodobností 0.1 zastavit. Pravděpodobnost toho, že od bodu B do cíle C dojede bez zastávky je rovna 0.7. Určete pravděpodobnost toho, že na úseku AC závodník nezastaví ani jednou.

Příklad 1.28.

Pravděpodobnost toho, že číslo losu bude mít stejné součty prvních tří a posledních tří cifer, je rovna 0.05525. Jaká je pravděpodobnost toho, že se takový los vyskytne mezi dvěma náhodně vybranými losy, jestliže tyto losy

a) mají po sobě jdoucí čísla,

b) byly získány nezávisle na sobě.

Příklad 1.29.

Petr si zakoupil dva losy ze dvou různých loterií. V první loterii je 150 000 losů, z kterých 50 000 vyhrává. V druhé loterii je 500 000 losů, z kterých 200 000 vyhrává. Jaká je pravděpodobnost, že vyhraje

a) jen los z první loterie,

b) jen jeden los,

c) oba losy,

d) aspoň jeden los,

e) jen los z druhé loterie,

f) nevyhraje ani jeden los?

Příklad 1.30.

Nechť je dáno 10 osudí, v každém je 10 koulí. V i-tém osudí je i černých a $10-i$ bílých koulí. Nejprve vybereme jedno osudí tak, aby každé mělo stejnou pravděpodobnost být vybrané. Poté vytáhneme jednu kouli z vybraného osudí opět tak, aby každá koule měla stejnou pravděpodobnost být vytažena. Jaká je pravděpodobnost, že vytažená koule bude černá?

Příklad 1.31.

Zamýšlíte koupit v autobazaru vůz jisté značky. Je ovšem známo, že 30% takových vozů má vadnou převodovku. Abyste získali více informací, najmete si mechanika, který je po projížďce schopen odhadnout stav vozu a jen s pravděpodobností 0.1 se zmýlí. Jaká je pravděpodobnost, že vůz, který zamýšlíte koupit, má vadnou převodovku

a) předtím, než si najmete mechanika,

b) jestliže mechanik předpoví, že vůz je dobrý?

Příklad 1.32.

Předpokládejte, že pravděpodobnost narození dvojčat stejného pohlaví je dvojnásobná proti téže pravděpodobnosti u dvojčat různého pohlaví. Dále předpokládejte, že pravděpodobnosti narození dvojčat různého pohlaví jsou pro obě pořadí stejná a že pravděpodobnost narození chlapce je 0.51, děvčete pak 0.49. Určete pravděpodobnost narození druhého chlapce, jestliže se jako první narodil chlapec.

Příklad 1.33.

Ke hledání ztraceného letadla bylo určeno 10 vrtulníků. Každého z nich lze k hledání letadla použít v jedné ze dvou oblastí, v nichž může být letadlo s pravděpodobnostmi 0.8 a 0.2. Každý vrtulník objeví letadlo, které je v prohledávané oblasti, s pravděpodobností 0.2. Jak je třeba rozdělit vrtulníky do těchto dvou oblastí, aby pravděpodobnost nalezení letadla byla maximální, jestliže každý vrtulník hledá nezávisle na ostatních? Určete, jaká bude pravděpodobnost nalezení letadla při optimálním rozdělení vrtulníků.

Příklad 1.34.

V losovací urně je 10 bílých a 5 černých koulí. Náhodně bez vracení vybíráme jednotlivé koule. Určete pravděpodobnost, že

a) jako první v pořadí vylosujeme bílou kouli,

b) jako druhou v pořadí vylosujeme bílou kouli, jestliže jsme poprvé vylosovali také bílou,

c) jako druhou v pořadí vylosujeme bílou kouli (aniž jsme si všimli, jakou barvu měla první vylosovaná koule).

Příklad 1.35.

Spojovacím kanálem je přenášen signál A, resp. B s pravděpodobností 0.84, resp. 0.16. Vzhledem k poruchám přenosu se 1/6 signálů A detekuje jako B, obdobně se 1/8 signálů B detekuje jako A. Určete pravděpodobnost, že určitý signál (aniž víme, jaký skutečně je) bude detekován na výstupu jako A. Určete pravděpodobnost, že signál, který byl detekován jako A, byl skutečně odeslán jako A.

Příklad 1.36.

Revizor ze zkušenosti ví, že zhruba v 26% tramvají při kontrole najde černého pasažéra. Kolik tramvají musí zkontrolovat, aby alespoň s 95% pravděpodobností našel alespoň jednoho černého pasažéra?

Příklad 1.37.

Účastník zapomněl poslední cifru telefonního čísla a rozhodl se, že jí bude postupně volit. S jakou pravděpodobností se dovolá nejpozději na čtvrtý pokus? (Předpokládáme, že při vytočení správného čísla se spojení uskuteční.)

Příklad 1.38.

V dílně pracují 3 stroje. První vyrobí 24%, druhý 36% a třetí 40% produkce dílny. První stroj vyrobí s pravděpodobností 0.02 zmetek, u druhého se toto stane s pravděpodobností 0.03 a u třetího s pravděpodobností 0.06. Vybereme-li náhodně dobrý výrobek, s jakou pravděpodobností byl vyroben třetím strojem?

Příklad 1.39.

Při výrobě 30% přístrojů byl použit zpřísněný technologický režim, zatímco při výrobě ostatních přístrojů standardní režim. Přitom pravděpodobnost bezporuchového chodu po dobu T je pro přístroj z první skupiny 0.97 a pro přístroj z druhé skupiny 0.82. Jaká je pravděpodobnost toho, že přístroj, který po dobu T bezporuchově pracoval, byl vyrobem ve zpřísněném režimu?

Příklad 1.40.

Předpokládejme hru bridž, která byla popsána v řešeném příkladu 1.5. V situaci, která je popsaná v příkladu 1.5, má hlavní hráč možnost hrát ještě jedním způsobem, kterému se říká impas.

Vynese z ruky malou kartu. Pokud dá hráč po jeho levici krále, dá ze stolu eso. Poté odehraje všechny srdcové karty, takže v srdcové barvě získá všech 5 zdvihů. Pokud hráč po jeho levici (západ) krále nedá, vynese hlavní hráč ze stolu dámu. Pokud má krále hráč po pravici (východ), hlavnímu hráči se nepodaří krále chytit, a tudíž neuhraje všech 5 srdcových zdvihů. Pokud krále má hráč po levici, pak mohou nastat dvě situace. Buď má hlavní hráč možnost vynést ze stolu malou kartu v jiné barvě a přebít ji velkou kartou z ruky. Poté impas v srdcích ještě jednou zopakuje. Pokud přechod do ruky nemá, musí vynést srdcové eso a doufat, že karty byly rozloženy v poměru 2:1.

Při hře zvané impas má hlavní hráč větší pravděpodobnost, že získá všechny srdcové zdvihy, než při prvním způsobu hry. Spočtěte pravděpodobnost, s jakou při tomto způsobu sehrávky získá všechny srdcové zdvihy, a obě pravděpodobnosti porovnejte.

2. Diskrétní rozdělení

Řešené příklady

Příklad 2.1.

Zkoušený přístroj je složen ze tří částí. Pravděpodobnost poruchy 1. části během zkoušky je 0.2, obdobně pravděpodobnost poruchy 2. části je 0.3 a pravděpodobnost poruchy 3. části je 0.4. Předpokládáme přitom, že poruchy jednotlivých částí vznikají nezávisle na sobě. Zkoušíme-li velké množství přístrojů, kolik porouchaných částí bude v průměru obsahovat vyzkoušený přístroj?

Řešení:

Průměrný počet porouchaných částí přístroje se bude pro velký rozsah zkoušených přístrojů blížit střední hodnotě náhodné veličiny X, která má rozdělení

$$\begin{aligned}
P(X=0) &= 0.8 \times 0.7 \times 0.6 = 0.336,\\
P(X=1) &= 0.2 \times 0.7 \times 0.6 + 0.8 \times 0.3 \times 0.6 + 0.8 \times 0.7 \times 0.4 = 0.452,\\
P(X=2) &= 0.2 \times 0.3 \times 0.6 + 0.8 \times 0.3 \times 0.4 + 0.2 \times 0.7 \times 0.4 = 0.188,\\
P(X=3) &= 0.2 \times 0.3 \times 0.4 = 0.024.
\end{aligned}$$

Střední hodnota veličiny X se rovná

$$\mathrm{E}\,X = 0 \times 0.336 + 1 \times 0.452 + 2 \times 0.188 + 3 \times 0.024 = 0.9.$$

Tentýž výsledek můžeme snadněji dostat následovně. Náhodná veličina X, která označuje počet porouchaných částí přístroje, se dá vyjádřit $X = X_1 + X_2 + X_3$, kde veličina X_i, $i = 1, 2, 3$ nabývá hodnoty 0, jestliže se i–tá část neporouchala, a nabývá hodnoty 1, jestliže se porouchala. Zřejmě $\mathrm{E}\,X_1 = 0 \times 0.8 + 1 \times 0.2 = 0.2$. Obdobně $\mathrm{E}\,X_2 = 0.3$ a $\mathrm{E}\,X_3 = 0.4$. Odtud $\mathrm{E}\,X = \mathrm{E}\,X_1 + \mathrm{E}\,X_2 + \mathrm{E}\,X_3 = 0.9$.

Odpověď: Průměrný počet porouchaných částí na jeden přístroj se rovná 0.9.

Příklad 2.2.

V urně máme 40 koulí, z toho 20 bílých a 20 černých. Hrajeme následující hru. Vytáhneme-li bílou kouli, získáme 1 Kč. Vytáhneme-li černou kouli, ztrácíme 1 Kč. Předpokládáme, že budeme tahat desetkrát. Které ze dvou pravidel je

a) výhodnější,

b) méně hazardní,

jestliže kouli po vytažení vracíme zpět nebo jestliže vytažené koule zpět nevracíme?

Řešení:

Označme

X_1 ... počet vytažených bílých koulí ve hře, ve které koule vracíme zpět,

X_2 ... počet vytažených bílých koulí ve hře, ve které koule zpět nevracíme,

Y_1 ... zisk ve hře, ve které koule vracíme zpět,

Y_2 ... zisk ve hře, ve které koule zpět nevracíme.

Zřejmě platí $Y_1 = X_1 - (10 - X_1) = 2X_1 - 10$ a $Y_2 = X_2 - (10 - X_2) = 2X_2 - 10$. Veličina X_1 je rozdělena podle binomického rozdělení s parametry $n = 10$ a $p = 1/2$. Veličina X_2 je rozdělena podle hypergeometrického rozdělení s parametry $N = 40$, $A = 20$, $n = 10$. Považujeme-li za výhodnější hru, ve které se v průměru získá více peněz, pak je rozhodovacím kritériem střední hodnota. Platí

$$\mathrm{E}\,Y_1 = 2\,\mathrm{E}\,X_1 - 10 = 2np - 10 = 2 \times 10 \times \frac{1}{2} - 10 = 0,$$
$$\mathrm{E}\,Y_2 = 2\,\mathrm{E}\,X_2 - 10 = 2n\frac{A}{N} - 10 = 2 \times 10 \times \frac{20}{40} - 10 = 0.$$

Za riskantnější zde můžeme považovat tu hru, pro níž je rozptyl zisku větší

$$\mathrm{Var}\,Y_1 = 4\,\mathrm{Var}\,X_1 = 4np(1-p) = 4 \times 10 \times \frac{1}{2} \times \frac{1}{2} = 10,$$
$$\mathrm{Var}\,Y_2 = 4\,\mathrm{Var}\,X_2 = 4n\frac{A}{N}\left(1 - \frac{A}{N}\right)\frac{N-n}{N-1} = 4 \times 10 \times \frac{20}{40} \times \left(1 - \frac{20}{40}\right) \times \frac{30}{39} \doteq 7.692.$$

Odpověď: Obě hry jsou tzv. spravedlivé, tj. průměr celkového zisku se pohybuje kolem nuly, ať už koule vracíme zpět nebo ne. První hra je však hazardnější, neboť rozptyl zisku je zde větší než v druhé hře.

Příklad 2.3.

Házíme dvěma kostkami. Najděte rozdělení náhodné veličiny označující maximum z počtu oček na dvou kostkách, jestliže víme, že minimum nabylo hodnoty 3.

Řešení:

Označme X maximum a Y minimum z počtu bodů na dvou kostkách. (Představme si pro jednoduchost, že jedna je červená a druhá modrá.) Veličina Y nabude hodnoty 3, jestliže buď na obou kostkách padne 3, nebo na červené padne 3 a na modré 4, 5 či 6, nebo na modré padne 3 a na červené 4, 5 či 6. Odtud vyplývá $P(Y = 3) = 7/36$. Jestliže víme, že Y nabyla hodnoty 3, pak veličina X může nabývat pouze jedné z hodnot 3, 4, 5, 6. Pravděpodobnost $P(X = 3 \cap Y = 3) = 1/36$, neboť jev $[X = 3 \cap Y = 3]$ znamená, že na obou kostkách padly trojky. Pravděpodobnost $P(X = 4 \cap Y = 3) = 2/36$, neboť jev $[X = 4 \cap Y = 3]$ znamená, že buď na červené kostce padla 3 a na modré 4, nebo naopak na červené 4 a na modré 3 atd.

Odpověď: Podmíněné rozdělení veličiny X za podmínky, že Y nabyla hodnoty 3, je rovno:

$$P(X = 3|Y = 3) = \frac{1/36}{7/36} = \frac{1}{7},$$
$$P(X = 4|Y = 3) = \frac{2/36}{7/36} = \frac{2}{7},$$
$$P(X = 5|Y = 3) = \frac{2/36}{7/36} = \frac{2}{7},$$
$$P(X = 6|Y = 3) = \frac{2/36}{7/36} = \frac{2}{7}.$$

Příklad 2.4.

Čistý roční výnos investiční společnosti byl v určitém roce 10%. Předpokládáme, že v každém následujícím roce bude roční výnos s pravděpodobností 1/3 stejný jako v roce bezprostředně minulém, s pravděpodobností 1/3 o 1% nižší a též s pravděpodobností 1/3 o 1% vyšší. Určete za tohoto předpokladu střední hodnotu a směrodatnou odchylku částky, kterou dostaneme po dvou letech, vložíme-li do dané investiční společnosti na počátku následujícího roku 200 000 Kč.

Řešení:

Pravděpodobnostní rozdělení splatné částky po dvou letech je zřejmě

$$S = \begin{cases} 200000(1+0.09)(1+0.08) = 235440 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.09)(1+0.09) = 237620 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.09)(1+0.10) = 239800 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.10)(1+0.09) = 239800 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.10)(1+0.10) = 242000 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.10)(1+0.11) = 244200 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.11)(1+0.10) = 244200 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.11)(1+0.11) = 246420 \text{ Kč} & \text{s pravděpodobností } 1/9, \\ 200000(1+0.11)(1+0.12) = 248640 \text{ Kč} & \text{s pravděpodobností } 1/9. \end{cases}$$

Pro střední hodnotu a rozptyl S platí:

$$\begin{aligned} \mathrm{E}\, S &= \tfrac{1}{9}(235440 + \cdots + 248640) = 242013.3\,, \\ \mathrm{Var}\, X &= \tfrac{1}{9}(235440^2 + \cdots + 248640^2) - 242014.3^2 = 16133600\,. \end{aligned}$$

Odpověď: Střední hodnota po 2 letech splatné částky je přibližně 242013 Kč, směrodatná odchylka je přibližně 4017 Kč. Všimněte si, že střední hodnota splatné částky je o něco málo větší než hodnota 242000 Kč, která odpovídá „průměrnému vývoji", tj. scénáři, kdy je v obou letech výnos 10%.

Příklad 2.5.

Padesátiletý muž si chce u určité pojišťovny pojistit případný dočasný důchod v roční výši 30000 Kč na dobu pěti let, odložený též o 5 let. Jaká je výše jednorázového nettopojistného, kalkuluje-li pojišťovna s úrokovou mírou ve výši 6% a používá-li úmrtnostní tabulky z roku 1996?

Řešení:

Vysvětlíme nejprve podrobněji použitou terminologii.

Zmiňovaným důchodem se rozumí, že muži bude poprvé po 5 letech od uzavření smlouvy pětkrát po sobě s roční periodou vyplácena částka 30000 Kč. Případností důchodu se rozumí, že každá splátka bude vyplacena, jen bude-li klient naživu, v opačném případě zbytek prostředků propadne bez náhrady pojišťovně. (Teoreticky se může stát, že propadne celé pojistné a nevyplatí se ani jedna splátka – totiž zemře-li klient před dosažením 55 let!) Jednorázovým nettopojistným se rozumí částka, kterou by měl klient

zaplatit pojišťovně při uzavření smlouvy, aby se jejím prostřednictvím pokryly právě všechny budoucí výdaje pojišťovny na splátky důchodu (se zohledněním časové hodnoty peněz a možnosti předčasného úmrtí klienta).

Uvedeme nyní důležitou úvahu.

Současná hodnota závazku pojišťovny zaplatit klientovi 30000 Kč např. za 5 let je při dané úrokové míře rovna $30000/(1+0.06)^5 = 22417.75$ Kč, pokud ovšem nepřihlížíme k možnosti, že se klient 55 let nemusí dožít. Přihlédneme-li k této možnosti, je střední hodnota závazku rovna

$$\frac{30000}{(1+0.06)^5} \cdot P(Z_{55}|Z_{50}) = 22417.75 \times 0.949009 = 21274.67.$$

Přitom jsme použili značení z příkladu 1.10, $P(Z_{55}|Z_{50})$ je pravděpodobnost, že se 50-ti letý muž dožije 55 let.

Interpretace posledního čísla je tato: pokud si pojišťovna dnes rezervuje 21274.67 Kč na každého 50-letého klienta, pak bude moci po 5 letech všem těmto dosud žijícím klientům vyplatit 30000 Kč, přičemž „v průměru ani nevydělá, ani neprodělá".

Hledané jednorázové nettopojistné A bude rovno střední hodnotě všech závazků pojišťovny spojených s výplatami splátek důchodu, tj.

$$\begin{aligned} A &= \sum_{k=5}^{9} \frac{30000}{(1+0.06)^k} \cdot P(Z_{50+k}|Z_{50}) = \\ &= 22417.75 \times 0.949009 + 21148.82 \times 0.936516 + 19951.71 \times 0.923305 + \\ &\quad + 18822.37 \times 0.909361 + 17756.95 \times 0.893673 = 92487.61\,. \end{aligned}$$

(Čísla $P(Z_{55}|Z_{50})$ až $P(Z_{59}|Z_{50})$ jsme vypočítali podle návodu z příkladu 1.10.)

Odpověď: Jednorázové nettopojistné pro pojištění daného důchodu je 92487.61 Kč. Poznamenejme však, že toto nettopojistné je v praxi podstatně menší (řádově o desítky procent), než tzv. bruttopojistné, tj. skutečná výše pojistného, vyžadovaná pojišťovnou. Do bruttopojistného se totiž promítají další náklady pojišťovny, jako jsou provize, správní náklady a též různé bezpečnostní přirážky. (Podrobněji viz Cipra (1995).)

Příklad 2.6.

V určité loterii bylo vydáno celkem 40000 losů, z nichž každý dvacátý vyhrává. Zakoupíme-li 50 losů, s jakou pravděpodobností alespoň na 5 vyhrajeme?

Řešení:

Náhodná veličina počtu vyhrávajících losů mezi zakoupenými losy se řídí hypergeometrickým rozdělením s parametry $N = 40000$, $A = 2000$ a $n = 50$. Jelikož parametr n je relativně velké číslo, N je velmi velké a podíl A/N je naopak velmi malý, je možno hypergeometrické rozdělení aproximovat binomickým rozdělením s parametry n a $p = A/N$ nebo dokonce Poissonovým rozdělením s parametrem $\lambda = n \cdot A/N = 2.5$. Označíme-li si jako X náhodnou veličinu počtu vyhrávajících losů mezi zakoupenými, pak přibližně platí:

$$P(X \geq 5) = 1 - P(X < 5) = 1 - \mathrm{e}^{-2.5}(1 + 2.5 + 2.5^2/2! + 2.5^3/3! + 2.5^4/4!) = 0.1088.$$

Odpověď: Při zakoupení 50 losů dané loterie vyhrajeme s pravděpodobností 0.1088 na pět nebo více než pět z nich.

Příklad 2.7.

Bridž se hraje s 52 bridžovými kartami, které se rozdají mezi čtyři hráče. Vždy dva a dva hráči hrají spolu. Při rozdání jste dostali do ruky dvě esa. Jaká je pravděpodobnost, že váš partner bude mít ony dvě zbývající esa?

Řešení:

Poté, co vy jste dostali 13 karet, zbývá na ostatní hráče 39 karet, v nichž jsou dvě esa. Váš partner dostane náhodně 13 karet ze zbývajících 39. Náhodná veličina X označující počet es v rukou vašeho partnera má hypergeometrické rozdělení s parametry $N = 39$, $A = 2$, $n = 13$. Pravděpodobnost, že váš partner dostane dvě zbývající esa, se rovná

$$P(X = 2) = \frac{\binom{2}{2}\binom{37}{11}}{\binom{39}{13}} \doteq 0.105.$$

Odpověď: Váš partner dostane zbývající dvě esa s pravděpodobností 0.105.

Příklad 2.8.

Hráč hry *Člověče, nezlob se* musí házet kostkou tak dlouho, dokud mu nepadne číslo 6, jinak nesmí nasadit figurku do hry. S jakou pravděpodobností figurku nasadí právě ve čtvrtém, resp. nejpozději ve čtvrtém kole? Jaká je střední hodnota a směrodatná odchylka počtu pokusů hráče o nasazení figurky do hry?

Řešení:

Uvažujme obecnější situaci, kdy nezávisle za stále stejných podmínek opakujeme určitý náhodný pokus (v principu není počet opakování pokusu omezen). Při každém opakování registrujeme, zda nastal určitý náhodný jev A, jehož pravděpodobnost výskytu při jedné realizaci pokusu je rovna stále stejnému číslu p $(0 < p < 1)$. Náhodná veličina X nechť označuje počet opakování pokusu před prvním výskytem jevu A. (Často se v této souvislosti mluví o počtu neúspěchů před prvním úspěchem.) Určíme rozdělení, střední hodnotu a rozptyl veličiny X. Poznamenejme, že jde o tzv. *geometrické rozdělení.*

Veličina X může nabývat hodnot $x = 0, 1, 2, \ldots$ a vzhledem k nezávislosti platí:

$$P(X = 0) = p, \quad P(X = 1) = p(1 - p), \quad \ldots, \quad P(X = x) = p(1 - p)^x.$$

Pro každé $p \in (0, 1)$ platí (jde o součet geometrické posloupnosti):

$$\sum_{x=0}^{\infty} p(1 - p)^x = 1.$$

Chápeme-li levou stranu jako funkci proměnné p (z tohoto pohledu jde o mocninnou řadu), můžeme ji podle této proměnné zderivovat a dostaneme:

$$\sum_{x=0}^{\infty}(1 - p)^x - \sum_{x=1}^{\infty} x \cdot p(1 - p)^{x-1} = 0,$$

po úpravě

$$\frac{1}{p} - \frac{\mathrm{E}\,X}{1-p} = 0 \quad \Longrightarrow \quad \mathrm{E}\,X = \frac{1-p}{p}.$$

Právě odvozený vztah říká, že pro všechna $p \in (0, 1)$ platí

$$\sum_{x=1}^{\infty} x \cdot p(1-p)^x = \frac{1-p}{p}.$$

Zderivujeme-li tuto rovnost opět podle proměnné p, dostaneme:

$$\sum_{x=1}^{\infty} x(1-p)^x - \sum_{x=1}^{\infty} x^2 \cdot p(1-p)^{x-1} = -\frac{1}{p^2},$$

po úpravě

$$\frac{1}{p}\,\mathrm{E}\,X - \frac{\mathrm{E}\,X^2}{1-p} = -\frac{1}{p^2} \quad \Longrightarrow \quad \mathrm{E}\,X^2 = (1-p)(2-p)/p^2$$

a proto

$$\mathrm{Var}\,X = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2 = \frac{(1-p)(2-p)}{p^2} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}.$$

Vraťme se k našemu příkladu:

Jevem A je padnutí čísla 6, parametr p je (za předpokladu, že jde o spravedlivou kostku) roven 1/6 a náhodná veličina X znamená kolikrát hráči padne číslo menší než 6 před prvním výskytem šestky. Nasadit figurku právě ve 4. kole znamená, že $X = 3$, nasazení nejpozději ve čtvrtém kole odpovídá jevu $\{X \leq 3\}$. Konečně počet pokusů hráče je veličina o jedničku větší než veličina X.

Platí:

$$P(X = 3) = \tfrac{1}{6}(1 - \tfrac{1}{6})^3 = 0.0965, \qquad P(X \leq 3) = 1 - P(X > 3) = 1 - (\tfrac{5}{6})^4 = 0.5177,$$
$$\mathrm{E}\,X = \frac{1-(1/6)}{1/6} = 5, \qquad \mathrm{Var}\,X = \frac{1-(1/6)}{(1/6)^2} = 30, \qquad \mathrm{sd}\,X = 5.477\,.$$

Odpověď: Hráč si nasadí figurku do hry právě ve 4. kole s pravděpodobností 0.0965, nejpozději ve 4. kole se tak stane s pravděpodobností 0.5177. Počet pokusů o nasazení figurky do hry má střední hodnotu 6 a směrodatnou odchylku 5.477.

Příklad 2.9.

Pracovníci přicházejí k terminálu nezávisle na sobě v průměru dva za hodinu. Jestliže má počet příchodů Poissonovo rozdělení, jaká je pravděpodobnost příchodu více než dvou pracovníků během dvouhodinového intervalu?

Řešení:

V příkladu se mlčky předpokládá, že příchody pracovníků tvoří Poissonův proces s konstantní intenzitou. V takovém případě má náhodná veličina X označující počet pracovníků, kteří přijdou k terminálu během dvouhodinového intervalu, Poissonovo rozdělení s parametrem $\lambda = 4$. Odtud vyplývá

$$P(X > 2) = 1 - P(X = 0) - P(X = 1) - P(X = 2) = 1 - \mathrm{e}^{-4} - 4\mathrm{e}^{-4} - \frac{4^2\mathrm{e}^{-4}}{2!} \doteq 0.762.$$

Odpověď: Pravděpodobnost toho, že během dvouhodinového intervalu přijdou k terminálu více než dva pracovníci, je rovna 0.762.

Příklad 2.10.

Sekretářka udělá v průměru v jednom z desetitisíce úderů překlep. Předpokládejme, že stránka má 40 řádek a na každém řádku je 80 písmen. S jakou pravděpodobností sekretářka neudělá ve dvoustránkovém textu ani jeden překlep?

Řešení:

Označme náhodnou veličinou X počet překlepů ve dvoustránkovém textu. Veličina X má binomické rozdělení s parametry $n = 6400$ a $p = 0.0001$, neboť předpokládáme, že pravděpodobnost překlepu v každém úderu je stále stejná, nezávislá na výskytu překlepů v předchozím textu. Vzhledem k tomu, že p je malé a n velké, je možno binomické rozdělení aproximovat Poissonovým s parametrem $\lambda = np = 0.64$. Odtud

$$P(X = 0) = \mathrm{e}^{-0.64} \doteq 0.527.$$

Odpověď: Pravděpodobnost, že sekretářka neudělá ani jeden překlep ve dvoustránkovém textu, je přibližně rovna 0.527.

Neřešené příklady

Příklad 2.11.

Přístroj obsahuje 5 tranzistorů. Předpokládejme, že došlo k poruše přístroje způsobené jedním tranzistorem. Opravář prohlíží náhodně jeden po druhém bez vracení. Určete rozdělení náhodné veličiny, která označuje počet potřebných zkoušek pro nalezení vadného tranzistoru.

Příklad 2.12.

V televizním kvízu se pokládají postupně tři otázky: A, B, C. Za správnou odpověď na otázku A se získá 1000 Kč, přičemž pravděpodobnost správné odpovědi je 0.8. Za správnou odpověď na otázku B se získá 2000 Kč, přičemž pravděpodobnost správné odpovědi je 0.5, a za správnou odpověď na otázku C se získá 8000 Kč, přičemž pravděpodobnost správné odpovědi je 0.2. Tázaný má možnost vybrat si pořadí, v jakém chce na otázky odpovídat. Jakmile však udělá chybu, ve hře dál nepokračuje a dostane jen tolik peněz, kolik získal správnými odpověďmi na předchozí otázky. V jakém pořadí je optimální odpovídat? Vysvětlete.

Příklad 2.13.

Dva hráči košíkové házejí střídavě míč na koš tak dlouho, dokud jeden z nich koš nezasáhne. První hráč zasáhne koš s pravděpodobností 0.4, zatímco druhý s pravděpodobností 0.6. Určete rozdělení pravděpodobnosti počtu hodů, které provede první z nich.

Příklad 2.14.

Házíme dvěma kostkami. Najděte střední hodnotu náhodné veličiny, označující součin počtu oček na dvou kostkách.

Příklad 2.15.

S jakou pravděpodobností vyhrajeme v jednom tahu Sportky při 6 vsazených číslech první, druhé, třetí, resp. čtvrté pořadí? (Princip losování jednoho tahu Sportky je následující: z 49 čísel se náhodně vybere 6 základních a jedno dodatkové číslo. 1. pořadí znamená uhádnutí všech základních čísel, 2. pořadí vyhrajeme, uhádneme-li 5 základních a dodatkové číslo, 3. pořadí vyhrajeme při uhádnutí 5 základních a 4. pořadí znamená uhádnutí 4 základních čísel.)

Příklad 2.16.

Při zásahu jádra atomu určitého prvku dojde s pravděpodobností 0.1 k vyzáření jisté částice. Kolem jaké hodnoty a s jakou směrodatnou odchylkou bude kolísat celkový počet vyzářených částic po zásahu n jader?

Příklad 2.17.

V jednom mililitru určitého dokonale rozmíchaného roztoku se v průměru nachází 15 určitých mikroorganismů. Určete pravděpodobnost, že při náhodném odběru vzorku o objemu $\frac{1}{2}$ mililitru bude ve zkumavce méně než 5 těchto mikroorganismů. (Použijte Poissonovo rozdělení.)

Příklad 2.18.

Hodíme osmi mincemi. S jakou pravděpodobností se bude počet rubů a líců lišit více než o 3?

Příklad 2.19.

V krabici je 20 bílých a 2 červené kuličky. Postupně losujeme kuličky z krabice, přičemž vybranou kuličku před dalším losováním

a) vracíme, b) nevracíme

zpět do krabice. Kolikrát minimálně musíme losovat, aby pravděpodobnost, že vytáhneme alespoň jednu červenou kuličku nebyla menší než 1/2?

Příklad 2.20.

Automobil má na své trase 4 semafory. Každý z nich s pravděpodobností 0.7 přikazuje zastavit, s pravděpodobností 0.3 dává volno. Určete za předpokladu, že se semafory přepínají nezávisle na sobě, střední hodnotu počtu semaforů projetých do prvního zastavení.

Příklad 2.21.

Hráč hraje s bankéřem následující hru. Hráč vylosuje (postupně bez vracení) z balíčku 32 karet dvě karty. Mají-li stejnou hodnotu, dostane hráč od bankéře 8 Kč, v opačném případě zaplatí 1 Kč. (Připomeňme, že jde o klasickou karetní soupravu, kde jsou 4 barvy a od každé je 8 karet různých hodnot.) Pro koho je hra výhodná a proč? (Návod: Najděte rozdělení a střední hodnotu zisku hráče v jedné hře.)

Příklad 2.22.

Hodinová dopravní intenzita na určitém místě dálnice v určitou denní dobu je 300 vozidel. S jakou pravděpodobností projede tímto místem během jedné minuty více než

6 vozidel? (Předpokládejte, že průjezdy automobilů v příslušné denní době tvoří Poissonův proces s konstantní intenzitou.)

Příklad 2.23.

Kamarád Vás pošle do sklepa, abyste donesl(a) 4 lahvová piva – z toho dvě desítky a dvě dvanáctky. Nevíte, kde rozsvítit, proto vezmete z basy poslepu 4 láhve. S jakou pravděpodobností jste vyhověl(a), víte-li, že v base bylo celkem 10 desítek a 6 dvanáctek?

Příklad 2.24.

Hodíme-li desetkrát kostkou, s jakou pravděpodobností padne právě 8 sudých čísel?

Příklad 2.25.

V krabici bylo 15 tenisových míčků, z nich 9 dosud nepoužitých. Hráči si pro první zápas náhodně vybrali 3 míčky, po skončení zápasu je vrátili do krabice. Pokud si z této krabice náhodně vybereme další trojici míčků, s jakou pravděpodobností budou všechny nepoužité?

Příklad 2.26.

Na začátku určitého kalendářního roku víme, že inflace za minulý rok dosáhla 8.6%. Pro následující roky předpokládáme tento scénář: V každém roce bude inflace s pravděpodobností 1/4 stejná jako v roce bezprostředně minulém, s pravděpodobností 1/2 klesne o $\frac{1}{2}$% a s pravděpodobností 1/4 klesne dokonce o 1%. S jakou pravděpodobností v tomto modelu klesne inflace po 3 letech pod 7.5%? Jaká bude střední hodnota a směrodatná odchylka kupní síly současných 1000 Kč po těchto 3 letech? (Pro vysvětlení: Je-li v k po sobě jdoucích letech inflace rovna postupně hodnotám $i_1, i_2, \ldots, i_k$, potom kupní síla částky C, kterou máme k dispozici na počátku 1. roku, klesne na konci k-tého roku na hodnotu $C/[(1+i_1)(1+i_2)\cdots(1+i_k)]$. V uvedeném vzorci je samozřejmě nutno inflaci dosazovat jako desetinné číslo, tj. je-li např. $i_1 = 8\%$, potom $(1 + i_1) = 1.08$, apod.)

Příklad 2.27.

Padesátiletý muž – podnikatel si chce půjčit od banky částku 2 000 000 Kč na dobu 10 let při roční úrokové míře 13%. Banka však požaduje mimo jiné pojištění tohoto úvěru u své dceřinné pojišťovny. Tj. podnikatel musí uzavřít dočasné pojištění pro případ smrti na zúročenou vypůjčenou částku a na dobu 10 let. Jaká je hodnota jednorázového nettopojistného, kalkuluje-li pojišťovna s technickou úrokovou mírou 4% a používá-li úmrtnostní tabulky z roku 1996 – viz příklad 1.10? (Při pojištění úvěru pojišťovna přebírá zodpovědnost za splacení úvěru bance včetně úroků v případě smrti pojištěného během smluvené doby. Pokud se pojištěný dožije konce smluvené doby – v našem případě 60 let – pojištění zanikne bez náhrady.)

Příklad 2.28.

Do turnaje se přihlásilo 20 hráčů, kteří se zcela náhodně rozlosovali do dvou stejně početných skupin. S jakou pravděpodobností budou

a) dva nejlepší hráči hrát v různých skupinách,

b) čtyři nejlepší hráči hrát po dvou v různých skupinách?

Příklad 2.29.

Student se ke zkoušce má naučit 60 otázek. Z nedostatku času se naučil jen 40. U zkoušky si vylosuje 3 otázky. S jakou pravděpodobností bude alespoň dvě umět? S jakou pravděpodobností nebude umět ani jednu?

Příklad 2.30.

Velkoobchod přebírá dodávky akumulátorů tak, že z každé dodávky náhodně vybere a zkontroluje 5%. Celou dodávku pak vrátí výrobci jako nevyhovující, obsahuje-li kontrolovaný vzorek více než 6% nevyhovujících akumulátorů. S jakou pravděpodobností bude dodávka 600 akumulátorů, obsahující 30 nevyhovujících, převzatá jako vyhovující?

Příklad 2.31.

Střílíme na cíl, dokud jej nazasáhneme. Jaká je střední hodnota a směrodatná odchylka počtu výstřelů, jestliže se pravděpodobnost zásahu cíle při jednom výstřelu nemění a je rovna 8%?

Příklad 2.32.

V určité dodávce jsou 3 promile vadných výrobků. Určete pravděpodobnost, že při kontrole 100 náhodně vybraných výrobků zjistíme více než jeden zmetek.

Příklad 2.33.

Určitý stroj vyrobí 2 součástky za minutu. Za jednu osmihodinovou směnu je průměrně 38 součástek mimo povolenou toleranci. Jaká je pravděpodobnost, že ze série 5 součástek budou 2 nebo více mimo toleranci?

Příklad 2.34.

Student v testu odpovídá na celkem 20 otázek. U každé z nich je 5 nabídnutých odpovědí (a, b, c, d, e), z nichž je vždy právě jedna správná. Student si je u 6 otázek jistý správnou odpovědí, u zbývajících otázek naopak odpovědi zaškrtává zcela náhodně. S jakou pravděpodobností zodpoví správně alespoň na polovinu ze všech otázek?

Příklad 2.35.

Televizní přijímač od určitého výrobce má v průměru 10 poruch za 10000 hodin provozu. S jakou pravděpodobností nastane alespoň jedna porucha během 200 hodin provozu?

Příklad 2.36.

Pravděpodobnost výroby zmetku určitým strojem je 0.01, při automatické kontrole výroby se stroj při detekování zmetku vždy zastaví. Jaká je střední hodnota a směrodatná odchylka počtu dobrých výrobků vyrobených mezi dvěma zmetky?

Příklad 2.37.

Ve městě, které má 150 000 obyvatel, bydlí v určité ulici 250 osob. Za účelem sociologického průzkumu bylo náhodně z registru všech obyvatel tohoto města vybráno 300 osob. S jakou pravděpodobností jsou mezi vybranými dva nebo více lidí ze zmiňované ulice?

3. Spojité rozdělení

Řešené příklady

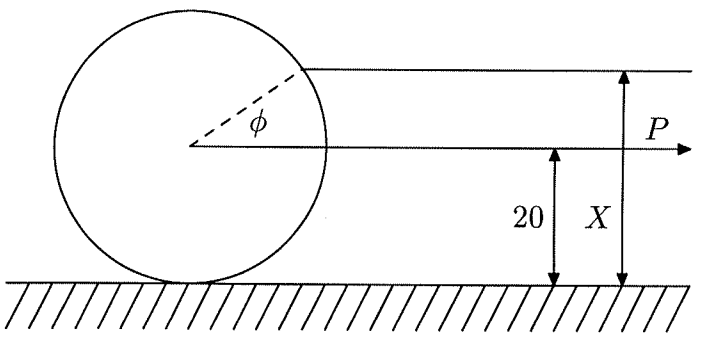
Příklad 3.1.

Ruské kolo se otáčí rovnoměrným pohybem kolem osy, která je umístěna 20 metrů nad nástupní plošinou. Díky poruše se kolo v určitém okamžiku zastaví. Určete rozdělení náhodné veličiny X, která označuje, v jaké výšce nad plošinou budeme v okamžiku zastavení. S jakou pravděpodobností bude naše kabina výš než 30 m nad nástupní plošinou? Spočtěte také střední hodnotu a směrodatnou odchylku veličiny X.

Řešení:

Veličina X je zřejmě spojitě rozdělená a nabývá pouze hodnot v intervalu $\langle 0, 40\rangle$. Hledejme nejprve distribuční funkci veličiny X:

$$F(x) = P(X < x).$$



Na obrázku je schématicky znázorněn řez ruským kolem rovinou kolmou k ose rotace. Označíme-li si jako ϕ orientovaný úhel, který svírá rameno naší kabiny s polopřímkou P rovnoběžnou s nástupní plošinou, pak tento úhel může nabývat v okamžiku zastavení libovolné hodnoty z intervalu $\langle -\frac{1}{2}\pi, \frac{3}{2}\pi\rangle$.

Vzhledem k tomu, že k zastavení dojde náhodně, je ϕ zřejmě spojitá náhodná veličina s rovnoměrným rozdělením na tomto intervalu (každý úhel ϕ má „stejnou možnost se vyskytnout"). Platí vztah

$$X = 20 + 20\sin\phi,$$

a proto

$$F(x) = P(20 + 20\sin\phi < x) = P\left(\sin\phi < \frac{x}{20} - 1\right).$$

Poslední nerovnost nastává (omezujeme se na $\phi \in \langle -\frac{1}{2}\pi, \frac{3}{2}\pi\rangle$ a $x \in \langle 0, 40\rangle$), jestliže

$$-\frac{\pi}{2} \le \phi < \arcsin\left(\frac{x}{20} - 1\right) \qquad \text{nebo} \qquad \pi - \arcsin\left(\frac{x}{20} - 1\right) < \phi \le \frac{3\pi}{2}.$$

Ze symetrie a z předpokladu rovnoměrného rozdělení plyne, že pravděpodobnost obou těchto jevů je stejná, a tedy

$$F(x) = 2P\left[-\frac{\pi}{2} \le \phi < \arcsin\left(\frac{x}{20} - 1\right)\right] = 2 \cdot \frac{\arcsin\left(\frac{x}{20} - 1\right) + \frac{\pi}{2}}{2\pi} =$$
$$= \frac{1}{2} + \frac{1}{\pi}\arcsin\left(\frac{x}{20} - 1\right) \qquad \text{pro } x \in \langle 0, 40\rangle.$$

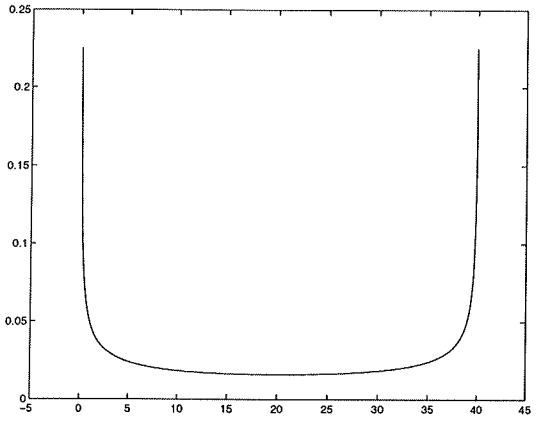
(Využili jsme skutečnost, že náhodná veličina ϕ má na intervalu $\langle -\frac{1}{2}\pi, \frac{3}{2}\pi\rangle$ konstantní hustotu rovnou $\frac{1}{2\pi}$.)

Celou distribuční funkci můžeme zapsat:

$$F(x) = \begin{cases} 0, & \text{pro } x < 0; \\ \frac{1}{2} + \frac{1}{\pi}\arcsin\left(\frac{x}{20} - 1\right), & \text{pro } x \in \langle 0, 40\rangle; \\ 1, & \text{pro } x > 40. \end{cases}$$

Dosazením do nalezené distribuční funkce najdeme hledanou pravděpodobnost:

$$P(X > 30) = 1 - F(30) = \frac{1}{2} - \frac{1}{\pi}\arcsin\left(\frac{30}{20} - 1\right) = \frac{1}{2} - \frac{1}{\pi}\cdot\frac{\pi}{6} = \frac{1}{3}.$$



Zderivováním distribuční funkce dostaneme po úpravě hustotu (viz obrázek)

$$f(x) = \begin{cases} \frac{1}{\pi}\cdot\frac{1}{\sqrt{40x - x^2}}, & \text{pro } x \in (0, 40); \\ 0, & \text{jindy.} \end{cases}$$

Střední hodnota veličiny X musí být vzhledem k symetrii hustoty rovna 20. Není těžké toto ověřit:

$$\mathrm{E}\,X = \int_0^{40} \frac{x}{\pi\sqrt{40x - x^2}}\,\mathrm{d}x = \frac{1}{20\pi}\int_0^{40} \frac{x}{\sqrt{1 - \left(\frac{x-20}{20}\right)^2}}\,\mathrm{d}x =$$

$$= \frac{1}{\pi}\int_{-1}^{1} \frac{20t + 20}{\sqrt{1 - t^2}}\,\mathrm{d}t = \left[-\frac{20}{\pi}\sqrt{1 - t^2} + \frac{20}{\pi}\arcsin t\right]_{-1}^{1} = 20.$$

(použili jsme operaci „doplnění na čtverec“ a substituci $\frac{x-20}{20} = t$.)

Rozptyl vypočteme podobně:

$$\mathrm{Var}\,X = \int_0^{40} \frac{(x-20)^2}{\pi\sqrt{40x - x^2}}\,\mathrm{d}x = \frac{1}{20\pi}\int_0^{40} \frac{(x-20)^2}{\sqrt{1 - \left(\frac{x-20}{20}\right)^2}}\,\mathrm{d}x =$$

$$= \frac{1}{\pi}\int_{-1}^{1} \frac{400t^2}{\sqrt{1 - t^2}}\,\mathrm{d}t = \left[-\frac{200}{\pi}t\sqrt{1 - t^2} + \frac{200}{\pi}\arcsin t\right]_{-1}^{1} = 200.$$

Odpověď: Rozdělení náhodné veličiny X, označující výšku kabiny v okamžiku zastavení, je spojité s hustotou $f(x)$ a distribuční funkcí $F(x)$ (funkční předpisy viz řešení). Střední hodnota této veličiny je rovna 20 m, směrodatná odchylka je $\sqrt{300} \doteq 17.32$ m. Pravděpodobnost, že kabina bude v okamžiku zastavení výše než 30 m, je rovna 1/3.

Poznamenejme na závěr, že rozdělení z tohoto příkladu se někdy nazývá *rozdělení arku-sinu*, tento název pochází od jeho distribuční funkce.

Příklad 3.2.

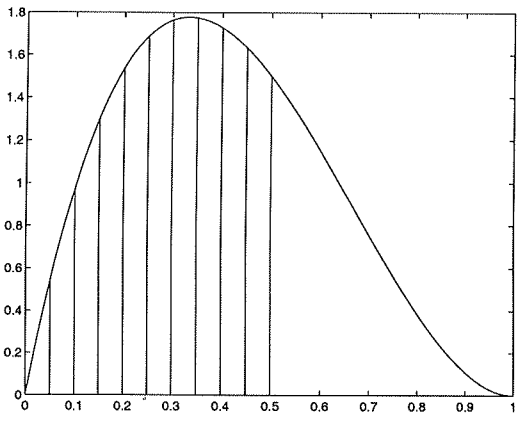
Tramvaj jezdí po dráze délky L. Pravděpodobnost toho, že cestující nastoupí do tramvaje v bodě x, je úměrná $x(L-x)^2$. Určete, kolik procent cestujících nastoupí v průměru dříve než v polovině trasy.

Řešení:

Náhodnou veličinou X označme vzdálenost místa, kde cestující nastoupí, od konečné, ze které tramvaj vyjíždí. Veličina X má diskrétní rozdělení, neboť zastávek je na trase jen konečný počet. Zastávky jsou však velmi hustě rozmístěné, a proto lze diskrétní rozdělení aproximovat spojitým:

$$P(X \in (a,b)) = \int_a^b f(x)\,\mathrm{d}x,$$

a tedy pravděpodobnost toho, že cestující nastoupí ve vzdálenosti x od začátku trasy, tj. $P(X \in (x, x+\mathrm{d}x))$, je přibližně rovna $f(x)\,\mathrm{d}x$.



Ze zadání plyne, že

$$f(x) = C\,x\,(L-x)^2.$$

Z podmínky $\int_0^L f(x)\,\mathrm{d}x = 1$ dostaneme výpočtem, že konstantu C je třeba volit rovnou $12/L^4$. Na obrázku je graf hustoty při volbě $L = 1$. Všimněte si, že hustota není symetrická kolem středu trasy. Tvar hustototy odpovídá zkušenosti, že před polovinou trasy nastupuje více cestujících než za polovinou trasy a že před konečnou jich nastupuje již velmi málo.

Podíl cestujících, kteří v průměru nastupují do tramvaje dříve než v polovině trasy, odpovídá pravděpodobnosti toho, že náhodně vybraný cestující nastoupí do tramvaje dříve než v polovině trasy:

$$P(X \in (0, L/2)) = \int_0^{L/2} \frac{12}{L^4} x(L-x)^2\,\mathrm{d}x = \frac{11}{16} = 0.6875.$$

Odpověď: Podíl cestujích, kteří nastoupí do tramvaje dříve než v polovině trasy, tvoří 68.75%.

Příklad 3.3.

Z místa A do místa B je třeba přenést signál nabývající dvou hodnot $+1$ a -1. Vzhledem k tomu, že je však signál zašuměn, přijímá místo čistého signálu příjemce signál + šum. Šum je náhodná veličina, jejíž rozdělení má hustotu:

$$f(x) = \begin{cases} \frac{1}{2}\mathrm{e}^{-x}, & \text{pro } x > 0; \\ \frac{1}{2}\mathrm{e}^{x}, & \text{pro } x \le 0. \end{cases}$$

Původní signál je v místě B detekován tak, že pokud je přijatý signál větší než -0.3, detekuje se 1, pokud je menší než -0.3, detekuje se -1. Stanovte pravděpodobnost chyby, jestliže víte, že v původním signálu bylo obsaženo 65 % jedniček a 35 % minus jedniček. Jaká je optimální rozhodovací mez pro tento případ?

Řešení:

Hustota pravděpodobnosti přijatého signálu (tj. zaslaného signálu + šumu) má v případě, že byla zaslána jednička, tvar:

$$f_1(x) = \begin{cases} \frac{1}{2}\mathrm{e}^{-(x-1)}, & \text{pro } x > 1; \\ \frac{1}{2}\mathrm{e}^{x-1}, & \text{pro } x \leq 1. \end{cases}$$

a v případě, že byla zaslána minus jednička, tvar:

$$f_{-1}(x) = \begin{cases} \frac{1}{2}\mathrm{e}^{-(x+1)}, & \text{pro } x > -1; \\ \frac{1}{2}\mathrm{e}^{x+1}, & \text{pro } x \leq -1. \end{cases}$$

Označíme-li při dané rozhodovací mezi d
$p_1(d)$ ... pravděpodobnost špatné interpretace, jestliže je posílána $+1$,
$p_{-1}(d)$ ... pravděpodobnost špatné interpretace, jestliže je posílána -1,
pak

$$p_1(d) = \begin{cases} \displaystyle\int_{-\infty}^{d} \frac{1}{2}\mathrm{e}^{x-1}\,\mathrm{d}x = \frac{1}{2}\mathrm{e}^{d-1}, & \text{pro } d \leq 1, \\ \displaystyle\int_{-\infty}^{1} \frac{1}{2}\mathrm{e}^{x-1}\,\mathrm{d}x + \int_{1}^{d} \frac{1}{2}\mathrm{e}^{-(x-1)}\,\mathrm{d}x = 1 - \frac{1}{2}\mathrm{e}^{-d+1}, & \text{pro } d > 1; \end{cases}$$

$$p_{-1}(d) = \begin{cases} \displaystyle\int_{d}^{\infty} \frac{1}{2}\mathrm{e}^{-(x+1)}\,\mathrm{d}x = \frac{1}{2}\mathrm{e}^{-d-1}, & \text{pro } d \geq -1, \\ \displaystyle\int_{d}^{-1} \frac{1}{2}\mathrm{e}^{x+1}\,\mathrm{d}x + \int_{-1}^{\infty} \frac{1}{2}\mathrm{e}^{-(x+1)}\,\mathrm{d}x = 1 - \frac{1}{2}\mathrm{e}^{d+1}, & \text{pro } d < -1. \end{cases}$$

Pravděpodobnost špatné interpretace $P(d)$ v případě, že je v signálu obsaženo 65% jedniček a 35% minus jedniček, je rovna

$$P(d) = \begin{cases} \frac{1}{2}\mathrm{e}^{d-1} \times 0.65 + (1 - \frac{1}{2}\mathrm{e}^{d+1}) \times 0.35, & \text{pro } d < -1; \\ \frac{1}{2}\mathrm{e}^{d-1} \times 0.65 + \frac{1}{2}\mathrm{e}^{-d-1} \times 0.35, & \text{pro } -1 \leq d \leq 1; \\ (1 - \frac{1}{2}\mathrm{e}^{-d+1}) \times 0.65 + \frac{1}{2}\mathrm{e}^{-d-1} \times 0.35, & \text{pro } d > 1. \end{cases}$$

Pro $d = -0.3$ je pravděpodobnost špatné interpretace signálu rovna 0.175475.

Chceme-li najít optimální mez d, je třeba minimalizovat $P(d)$ vzhledem k d. Funkce $P(d)$ je (jak se můžete snadnou úpravou přesvědčit) na intervalu $(-\infty, -1)$ klesající a

na intervalu $(1, \infty)$ rostoucí. Minimum je třeba hledat na intervalu $(-1, 1)$.

$$\begin{aligned} P'(d) = \frac{1}{2}\mathrm{e}^{d-1} \times 0.65 - \frac{1}{2}\mathrm{e}^{-d-1} \times 0.35 = 0, \\ \mathrm{e}^{d} \times 0.65 - \mathrm{e}^{-d} \times 0.35 = 0, \\ \mathrm{e}^{2d} = \frac{0.35}{0.65}, \\ d = -0.3095. \end{aligned}$$

Odpověď: Optimální mez pro rozhodování je -0.3095. Při rozhodovací mezi -0.3 budeme špatně interpretovat přibližně 17.6 % zaslaných znaků.

Příklad 3.4.

Doba životnosti X_A přístroje A (dána v rocích) má rozdělení s hustotou

$$f(x) = \begin{cases} \frac{2}{(x+2)^2}, & \text{pro } x \geq 0; \\ 0, & \text{pro } x < 0; \end{cases}$$

zatímco doba životnosti X_B přístroje B (dána v rocích) má rozdělení s hustotou

$$g(x) = \begin{cases} \frac{128}{(x+8)^3}, & \text{pro } x \geq 0; \\ 0, & \text{pro } x < 0. \end{cases}$$

a) Ověřte, zda jsou $f(x)$ a $g(x)$ opravdu hustoty.

b) Předpokládejme, že kritérium pro výběr přístroje je pravděpodobnost, s jakou přístroj přežije funkční dobu T. V kterém případě, tj. pro jakou délku funkční doby T, je lepší zakoupit přístroj A a v kterém případě přístroj B?

Řešení:

a) Nezáporná funkce $f(x)$ je hustota, jestliže $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$. Totéž platí i pro funkci $g(x)$. Ověřme tento požadavek:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = \int_{0}^{\infty} \frac{2}{(x+2)^2}\,\mathrm{d}x = \left[\frac{-2}{(x+2)}\right]_0^{\infty} = 1, \\ \int_{-\infty}^{\infty} g(x)\,\mathrm{d}x = \int_{0}^{\infty} \frac{128}{(x+8)^3}\,\mathrm{d}x = \left[\frac{-64}{(x+8)^2}\right]_0^{\infty} = 1. \end{aligned}$$

b) Pravděpodobnost, s jakou přístroj A přežije T, je rovna

$$P(X_A > T) = \int_{T}^{\infty} \frac{2}{(x+2)^2}\,\mathrm{d}x = \left[\frac{-2}{(x+2)}\right]_T^{\infty} = \frac{2}{T+2}$$

a pravděpodobnost, že přístroj B přežije dobu T, je rovna

$$P(X_B > T) = \int_{T}^{\infty} \frac{128}{(x+8)^3}\,\mathrm{d}x = \left[\frac{-64}{(x+8)^2}\right]_T^{\infty} = \frac{64}{(T+8)^2}.$$

Podle našeho rozhodovacího pravidla vybereme přístroj A, jestliže

$$\begin{aligned} \frac{2}{T+2} &> \frac{64}{(T+8)^2}, \\ (T+8)^2 &> 32(T+2), \\ T^2+16T+64 &> 32T+64, \\ T^2-16T &> 0, \\ T &> 16. \end{aligned}$$

Vybereme přístroj B, jestliže

$$\begin{aligned} \frac{2}{T+2} &< \frac{64}{(T+8)^2}, \\ 0 < T &< 16. \end{aligned}$$

Odpověď: Jestliže funkční doba má být delší než 16 let, je lépe zakoupit přístroj A, neboť pravděpodobnost, že přístroj bude fungovat ještě v této době, je větší než odpovídající pravděpodobnost pro přístroj B. Jestliže funkční doba má být kratší než 16 let, je lépe vybrat přístroj B. Poznamenejme, že pravděpodobnost přežití lze statisticky interpretovat jako podíl těch přístrojů z celkového množství, který bude fungovat ještě v čase T.

Příklad 3.5.

Rozdělení spojité náhodné veličiny X je dáno hustotou:

$$f(x) = c \cdot \mathrm{e}^{-|x|}, \qquad x \in \mathbf{R}.$$

Určete konstantu c, vypočtěte střední hodnotu, rozptyl, šikmost a špičatost veličiny X.

Řešení:

Snadno vypočteme, že

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 2c \int_{0}^{\infty} \mathrm{e}^{-x}\,\mathrm{d}x = 2c.$$

Odtud plyne, že $c = \frac{1}{2}$. Dále dostaneme:

$$\begin{aligned} \mathrm{E}\,X &= \frac{1}{2}\int_{-\infty}^{\infty} x\mathrm{e}^{-|x|}\,\mathrm{d}x = 0, \\ \mathrm{Var}\,X &= \frac{1}{2}\int_{-\infty}^{\infty} x^2\mathrm{e}^{-|x|}\,\mathrm{d}x = \int_{0}^{\infty} x^2\mathrm{e}^{-x}\,\mathrm{d}x = \left[-(x^2+2x+2)\mathrm{e}^{-x}\right]_0^{\infty} = 2, \\ \alpha &= \frac{\frac{1}{2}\int_{-\infty}^{\infty} x^3\mathrm{e}^{-|x|}\,\mathrm{d}x}{2^{(3/2)}} = 0. \end{aligned}$$

Špičatost spojité náhodné veličiny je definována

$$\beta = \frac{\int_{-\infty}^{\infty}(x-\mathrm{E}\,X)^4\,\mathrm{d}x}{(\mathrm{Var}\,X)^2} - 3.$$

(Zlomek v definici špičatosti je samozřejmě vždy nezáporný, pro zajímavost uveďme, že jeho hodnota je pro normální rozdělení s libovolnými parametry μ a σ^2 rovna číslu 3, tedy špičatost normálního rozdělení je rovna nule. Pomocí znaménka špičatosti tedy měříme, je-li dané rozdělení „špičatější" či „méně špičaté" než normální rozdělení.)

V našem příkladě vypočteme:

$$\int_0^\infty x^4 \mathrm{e}^{-x}\,\mathrm{d}x = \left[-(x^4+4x^3+12x^2+24x+24)\mathrm{e}^{-x}\right]_0^\infty = 24,$$

a proto

$$\beta = \frac{2\cdot\int_0^\infty \frac{1}{2}x^4\mathrm{e}^{-x}\,\mathrm{d}x}{2^2} = \frac{24}{4} - 3 = 3.$$

Odpověď: Střední hodnota a šikmost veličiny X jsou nulové, rozptyl je roven 2 a špičatost je rovna 3. Naše rozdělení je tedy „špičatější" než normální rozdělení (načrtněte si graf hustoty!).

Příklad 3.6.

Životnost žárovky X má exponenciální rozdělení se střední hodnotou $\mathrm{E}\,X = 400$ hod. S jakou pravděpodobností bude žárovka svítit dalších 100 hodin, jestliže už svítila 600 hodin?

Řešení:

Pravděpodobnost, že žárovka bude svítit déle než x hodin je rovna

$$P(X > x) = \mathrm{e}^{-x/\delta},$$

kde $\delta = \mathrm{E}\,X = 400$. Pravděpodobnost toho, že žárovka bude svítit ještě dalších 100 hodin (tj. její životnost bude delší než 700 hodin), jestliže již svítila déle než 600 hodin je rovna

$$P(X > 700|X > 600) = \frac{P(X > 700 \cap X > 600)}{P(X > 600)} = \frac{P(X > 700)}{P(X > 600)} = \frac{\mathrm{e}^{-700/400}}{\mathrm{e}^{-600/400}}$$

$$= \mathrm{e}^{-100/400} = 0.779.$$

Všimněme si, že pro náhodnou veličinu X, která má exponenciální rozdělení platí, že podmíněná pravděpodobnost toho, že veličina X nabude hodnoty větší než $x+y$ za podmínky, že již nabyla hodnoty větší než y, je rovna nepodmíněné pravděpodobnosti jevu, že veličina X nabude hodnoty větší než x. V praxi to znamená, že součástky, jejichž životnost má exponenciální rozdělení, nemá význam preventivně vyměňovat. Nové součástky mají totiž stejnou pravděpodobnost, že přežijí libovolnou dobu T, jako staré.

Odpověď: Pravděpodobnost, že žárovka bude svítit dalších 100 hodin, jestliže již svítila 600 hodin, je rovna 0.779.

Příklad 3.7.

Uvažujte náhodnou veličinu X s rozdělením z příkladu 3.5 a náhodnou veličinu Y, mající normální rozdělení se stejnou střední hodnotou a stejným rozptylem, jako veličina X. Vypočtěte a porovnejte pravděpodobnosti $P(|X| > a)$ a $P(|Y| > a)$, kde a volte postupně 2, 4, 6, 8.

Řešení:

Vzhledem k sudosti hustoty $f_X(x)$ náhodné veličiny X platí pro libovolné $a > 0$

$$P(|X| > a) = 2 \cdot \int_a^\infty f_X(x)\,\mathrm{d}x = \int_a^\infty \mathrm{e}^{-x}\,\mathrm{d}x = \mathrm{e}^{-a}.$$

Dosazením zadaných hodnot parametru a dostaneme:

$$P(|X| > 2) = 0.13534, \qquad P(|X| > 4) = 0.01832,$$
$$P(|X| > 6) = 0.00248, \qquad P(|X| > 8) = 0.00034.$$

Mají-li být střední hodnoty i rozptyly veličin X a Y stejné, musí pro parametry normálního rozdělení veličiny Y platit:

$$\mu = 0, \qquad \sigma^2 = 2.$$

Opět pro libovolné $a > 0$ platí:

$$P(|Y| > a) = 2[1 - P(Y \leq a)] = 2[1 - \Phi(a/\sqrt{2})].$$

Po dosazení zadaných hodnot parametru a a vyhledání (vypočtení) hodnot distribuční funkce normovaného normálního rozdělení Φ dostaneme:

$$P(|Y| > 2) = 0.15730, \qquad P(|Y| > 4) = 0.00468,$$
$$P(|Y| > 6) = 0.00002, \qquad P(|Y| > 8) = 0.00000.$$

Porovnáním vypočtených pravděpodobností vidíme, že rozdělení z příkladu 3.5 má „těžší chvosty“, než odpovídající normální rozdělení (tj. s větší pravděpodobností se mohou vyskytnout odlehlé hodnoty). Tato skutečnost jenom jiným způsobem dokresluje fakt, že rozdělení veličiny X je špičatější než rozdělení veličiny Y, jak jsme ukázali v příkladu 3.5.

Příklad 3.8.

Tvrdí se, že otázky v testu jsou dobře vybrány tehdy, jestliže počet bodů, které zkoušení získají, má přibližně normální rozdělení $N(\mu, \sigma^2)$. Zkoušející se rozhodl využít znalosti počtu bodů v jednotlivých testech k odhadu μ a σ^2 a poté známkovat podle následujícího principu:

známkou „1“, jestliže počet bodů $\geq \mu + \sigma$,
známkou „2“, jestliže $\mu \leq$ počet bodů $< \mu + \sigma$,
známkou „3“, jestliže $\mu - \sigma \leq$ počet bodů $< \mu$,
známkou „4“, jestliže $\mu - 2\sigma \leq$ počet bodů $< \mu - \sigma$,
známkou „5“, jestliže počet bodů $< \mu - 2\sigma$.

Jaký podíl zkoušených získá při tomto hodnocení jedničku, dvojku, ... , pětku?

Řešení:

Nechť X je náhodná veličina označující počet získaných bodů. Podle předpokladu má normální rozdělení $N(\mu, \sigma^2)$.

$$P(X \geq \mu + \sigma) = 1 - \Phi\left(\frac{\mu + \sigma - \mu}{\sigma}\right) = 1 - \Phi(1) = 0.1587,$$

$$P(\mu \leq X < \mu + \sigma) = \Phi\left(\frac{\mu + \sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - \mu}{\sigma}\right) =$$
$$= \Phi(1) - \Phi(0) = 0.8413 - 0.5 = 0.3413,$$

$$P(\mu - \sigma \leq X < \mu) = \Phi\left(\frac{\mu - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - \sigma - \mu}{\sigma}\right) = \Phi(0) - \Phi(-1) = 0.3413,$$

$$P(\mu - 2\sigma \leq X < \mu - \sigma) = \Phi\left(\frac{\mu - \sigma - \mu}{\sigma}\right) - \Phi\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right) =$$
$$= \Phi(-1) - \Phi(-2) = 0.1359,$$

$$P(X < \mu - 2\sigma) = \Phi\left(\frac{\mu - 2\sigma - \mu}{\sigma}\right) = \Phi(-2) = 0.0228.$$

Odpověď: Pokud zkoušející bude hodnotit velké množství testů, pak jedničku získá přibližně 15.87 %, dvojku 34.13 %, trojku 34.13 %, čtyřku 13.59 % a pětku 2.28 % zkoušených.

Příklad 3.9.

Firma získá z každého prodaného výrobku 100 Kč. Za výměnu během záruční lhůty zaplatí 300 Kč. Životnost výrobku v letech má normální rozdělení $N(3, 1)$. Jakou záruční dobu v měsících má firma stanovit, aby střední (průměrný) zisk byl alespoň 60 Kč?

Řešení:

Nechť X označuje zisk z výrobku. Jeho rozdělení je dáno následující tabulkou, kde p_x je pravděpodobnost, s jakou výrobek přežije záruční dobu x bez poruchy.

i	100	-200
$P(X = i)$	p_x	$(1 - p_x)$

Doba životnosti součástky T má $N(3, 1)$ rozdělení. Odtud plyne

$$p_x = P(T > x) = 1 - \Phi(x - 3).$$

Střední (průměrný) zisk při záruční době x je roven

$$\mathrm{E}\,X = 100p_x - 200(1 - p_x) = 300p_x - 200 = 300(1 - \Phi(x - 3)) - 200 = 100 - 300\Phi(x - 3).$$

Nyní je třeba stanovit záruční dobu x tak, aby $\mathrm{E}\,X \geq 60$:

$$\begin{aligned} 100 - 300\,\Phi(x-3) &\geq 60, \\ \Phi(x-3) &\leq 0.1333, \\ x - 3 &\leq -1.11, \\ x &\leq 1.89. \end{aligned}$$

Chceme-li vyjádřit záruční dobu v měsících, je třeba hodnotu x vynásobit dvanácti.

Odpověď: Nejdelší záruční lhůtou, při které je ještě průměrný zisk větší než 60 Kč, je 22 měsíců.

Příklad 3.10.

Výroba hřídelek je nastavena tak, aby jejich poloměr byl roven $\mu = 32$ mm. Ze zkušenosti se ví, že skutečný poloměr vyrobených hřídelek kolísá kolem této hodnoty jako normálně rozdělená náhodná veličina se směrodatnou odchylkou $\sigma = 0.5$ mm. Automatický kontrolní přístroj hlásí chybu (a zastaví výrobu), jakmile se poloměr hřídelky odchýlí od 32 mm o ± 1.5 mm. Předpokládejme, že vinou špatného postupu obsluhujícího personálu se nastavení soustruhu změnilo tak, že poloměr hřídelky kolísá místo kolem hodnoty $\mu = 32$ mm kolem hodnoty $\mu = 33$ mm. Nechť náhodná veličina označuje počet hřídelek, který soustruh vysoustruží, než automatické kontrolní zařízení zastaví provoz. Určete její rozdělení.

Řešení:

Náhodná veličina X označující poloměr vysoustružené hřídelky poté, co došlo ke změně, má normální rozdělení $N(\mu = 33, \sigma^2 = 0.25)$. Pravděpodobnost, že náhodně vybraná hřídelka bude ležet mimo toleranční meze 32 ± 1.5 mm, je rovna

$$p = 1 - P(X \in (30.5, 33.5)) = 1 - (\Phi(1) - \Phi(-5)) = 0.1587.$$

Náhodná veličina Y označující počet hřídelek, který se vysoustruží, než automatické kontrolní zařízení zastaví výrobu, má diskrétní rozdělení nabývající hodnot $1, 2, 3, \ldots$. Automatické zařízení zastaví výrobu při kontrole i–té hřídelky, jestliže poloměry prvních $i-1$ hřídelek byly v tolerančních mezích, zatímco poloměr i–té hřídelky leží mimo ně. Odtud

$$P(Y = i) = (1-p)^{i-1}p = 0.8413^{i-1} \times 0.1587.$$

Odpověď: Náhodná veličina Y označující počet vysoustružených hřídelek do doby, než automatický kontrolní přístroj zjistí chybu, nabývá hodnot $i = 1, 2, \ldots$ s pravděpodobnostmi $P(Y = i) = 0.8413^{i-1} \times 0.15874$. Přitom je $Y = X + 1$, kde X označuje počet po sobě vysoustružených hřídelek, jejichž rozměry byly v tolerančních mezích. Jak plyne z předchozích úvah, tato veličina X má geometrické rozdělení s parametrem p, tj. $P(X = i) = 0.8413^{i} \times 0.15874$ pro $i = 1, 2, \ldots$.

Příklad 3.11.

Časový odstup mezi dvěma po sobě jedoucími vozidly se často modeluje pomocí logaritmicko-normálního rozdělení. Na základě měření odhadujeme, že na určitém místě dálnice v určitou denní dobu je střední hodnota tohoto odstupu 4.5 s a směrodatná

odchylka je 3.2 s. Předpokládáme též, že časový odstup nikdy neklesne pod 0.4 s. Určete z těchto předpokladů parametry teoretického modelu, odhadněte pomocí tohoto modelu, v kolika procentech měření bude časový odstup

a) menší než 2.5 s, b) větší než 10 s.

Řešení:

Použijeme logaritmicko-normální rozdělení $\mathrm{LN}(\mu, \sigma^2, x_0)$, kde $x_0 = 0.4$. Platí vztahy

$$\begin{aligned} \mathrm{E}\,X &= x_0 + \mathrm{e}^{\mu+(\sigma^2/2)}, \\ \mathrm{sd}\,X &= \mathrm{e}^{\mu+(\sigma^2/2)} \cdot \sqrt{\mathrm{e}^{\sigma^2} - 1}. \end{aligned}$$

Z této soustavy rovnic můžeme jednoznačně vypočítat, známe-li x_0, $\mathrm{E}\,X$ a $\mathrm{sd}\,X$ (resp. $\mathrm{Var}\,X$), zbývající parametry:

$$\begin{aligned} \mu &= \ln \frac{(\mathrm{E}\,X - x_0)^2}{\sqrt{\mathrm{Var}\,X + (\mathrm{E}\,X - x_0)^2}}, \\ \sigma^2 &= \ln\left(1 + \frac{\mathrm{Var}\,X}{(\mathrm{E}\,X - x_0)^2}\right). \end{aligned}$$

Po dosazení $x_0 = 0.4$, $\mathrm{E}\,X = 4.5$ a $\mathrm{Var}\,X = 3.2^2 = 10.24$ dostaneme:

$$\mu = 1.1731, \qquad \sigma = 0.68972.$$

ad a) Hledanou pravděpodobnost vypočteme takto:

$$P(X < 2.5) = \Phi\left(\frac{\ln(2.5 - 0.4) - 1.1731}{0.68972}\right) = \Phi(-0.6251) = 0.2659.$$

ad b) Podobně vypočteme

$$P(X > 10) = 1 - P(X \leq 10) = 1 - \Phi(1.5784) = 0.0573.$$

Odpověď: Časový odstup bychom modelovali logaritmicko-normálním rozdělením s parametry $x_0 = 0.4$, $\mu = 1.1731$ a $\sigma = 0.68972$. Při použití tohoto modelu odhadneme, že ve 26.6% měření bude odstup menší než 2.5 s a ve 5.7% bude větší než 10 s.

Příklad 3.12.

Konstrukce měla být původně navržena na návrhovou hodnotu, jíž byl 0.5% (dolní) kvantil pevnosti betonu. Je známo, že pevnost použitého betonu má dvouparametrické logaritmicko-normální rozdělení se střední hodnotou $\mathrm{E}\,X = 3$ a směrodatnou odchylkou $\mathrm{sd}\,X = 0.63$. Statik povolil zvýšení návrhové hodnoty na 1% dolní kvantil. Jak se konkrétně změní návrhová hodnota?

Řešení:

Nejprve spočtěme parametry μ a σ dvouparametrického logaritmicko-normálního rozdělení. Pro střední hodnotu $\mathrm{E}\,X$ a směrodatnou odchylku $\mathrm{sd}\,X$ platí

$$\mathrm{E}\,X = \mathrm{e}^{\mu+\frac{\sigma^2}{2}} = 3,$$
$$\mathrm{sd}\,X = \mathrm{e}^{\mu+\frac{\sigma^2}{2}}\sqrt{\mathrm{e}^{\sigma^2}-1} = 0.63.$$

Odtud $\sigma = 0.207738$ a $\mu = 1.077$. Dále platí, že náhodná veličina X se řídí $LN(\mu, \sigma^2)$ právě tehdy, jestliže $\ln X$ se řídí $N(\mu, \sigma^2)$, a tedy $(\ln X - \mu)/\sigma$ řídí standardním normálním rozdělením $N(0, 1)$. Odtud plyne vztah mezi z_α označující $100 \cdot \alpha\,\%$ (dolní) kvantil $LN(\mu, \sigma^2)$ a v_α označující $100 \cdot \alpha\,\%$ (dolní) kvantil standardního normálního rozdělení $N(0, 1)$. Platí

$$P(X < z_\alpha) = P\left(\frac{\ln X - \mu}{\sigma} < \frac{\ln z_\alpha - \mu}{\sigma}\right) = \Phi\left(\frac{\ln z_\alpha - \mu}{\sigma}\right) = \alpha,$$

a tedy $\frac{\ln z_\alpha - \mu}{\sigma} = v_\alpha$, nebo jinak $z_\alpha = \mathrm{e}^{\mu + v_\alpha \sigma}$. Dolní 0.5% kvantil standardního normálního rozdělení $v_{0.005} = -2.576$ a dolní 1% kvantil standardního normálního rozdělení $v_{0.01} = -2.326$. Odtud $z_{0.005} = \mathrm{e}^{\mu - 2.576\sigma} = 1.7192$ a $z_{0.01} = \mathrm{e}^{\mu - 2.326\sigma} = 1.810857$.

Odpověď: Původně byla konstrukce navržena na návrhovou hodnotu 1.719. Později se použila návrhová hodnota 1.811.

Neřešené příklady

Příklad 3.13.

Struna dlouhá 1 m je zcela náhodně přestřižena na dvě části. S jakou pravděpodobností je poměr délky delší části ku délce kratší části větší než 3 : 1?

Příklad 3.14.

Spojitá náhodná veličina má tzv. *trojúhelníkové rozdělení* na intervalu $\langle -a, a\rangle$, tvoří-li graf její hustoty rovnoramenný trojúhelník se základnou $\langle -a, a\rangle$ (přičemž tato hustota je nulová vně $\langle -a, a\rangle$). Zapište matematicky tuto hustotu, vypočtěte střední hodnotu, rozptyl, šikmost a špičatost této veličiny.

Příklad 3.15.

Víme, že spojitá náhodná veličina má rovnoměrné rozdělení se střední hodnotou 4 a rozptylem 3. Určete toto rozdělení.

Příklad 3.16.

Předpokládejme, že spojité náhodné veličiny X, Y mají hustoty:

$$f_X(x) = \begin{cases} a/(x+1)^3, & \text{pro } x \geq 0; \\ 0, & \text{jindy.} \end{cases} \qquad f_Y(x) = \begin{cases} b/(x+1)^4, & \text{pro } x \geq 0; \\ 0, & \text{jindy.} \end{cases}$$

Určete konstanty a, b, vypočtěte a porovnejte střední hodnoty a směrodatné odchylky veličin X a Y.

Příklad 3.17.

Jaká je špičatost rovnoměrného rozdělení na intervalu $\langle 0, 1\rangle$? Jaká je špičatost rovnoměrného rozdělení na libovolném intervalu $\langle \alpha, \beta\rangle$?

Příklad 3.18.

Z technických údajů dvou dávkovačů lze zjistit, že odchylka X_1 v dávkách 1. dávkovače má hustotu

$$f_1(x) = \begin{cases} \frac{3}{4}(1-x^2), & \text{pro } x \in (-1,1); \\ 0, & \text{pro } x \notin (-1,1) \end{cases}$$

a odchylka X_2 v dávkách 2. dávkovače má hustotu

$$f_2(x) = \begin{cases} \frac{5}{8}(1-x^4), & \text{pro } x \in (-1,1); \\ 0, & \text{pro } x \notin (-1,1). \end{cases}$$

Rozhodněte, který dávkovač je lepší.

Příklad 3.19.

Doba do vybití baterie se řídí exponenciálním rozdělením.

a) Jaká je střední doba do vybití, víme-li, že 4000 hodin přežije 1% těchto baterií?

b) Je-li střední doba do vybití 3150 hodin, kolik procent těchto baterií přežije 4000 hodin?

Příklad 3.20.

Chybu při měření určité veličiny modelujeme normálním rozdělením s nulovou střední hodnotou a s rozptylem 1.5. Určete interval (souměrný podle počátku), ve kterém se bude chyba nacházet v 90% měření.

Příklad 3.21.

Chyba při měření velikosti úhlu určitým geodetickým přístrojem má normální rozdělení s nulovou střední hodnotou a se směrodatnou odchylkou 0.0117 gradu. Určete pravděpodobnost, že

a) změřená velikost úhlu překračuje skutečnou velikost více než o 0.005 gradu;

b) změřená velikost úhlu se liší od skutečné velikosti nejvýše o 0.01 gradu.

Příklad 3.22.

Obsah nečistot v odpadních vodách je popsán normálním rozdělením se střední hodnotou 0.18 a směrodatnou odchylkou 0.03. Vypočtěte

a) procento zkoušek, při kterých obsah nečistot překročí hodnotu 0.24,

b) hodnotu obsahu nečistot, která bude překročena v 1% zkoušek.

Příklad 3.23.

Maximální roční průtok určité řeky se řídí Pearsonovým rozdělením typu III se střední hodnotou 88 $\mathrm{m^3 s^{-1}}$, rozptylem 850 a šikmostí 0.55. Určete přibližně pomocí tabulek

a) pravděpodobnost, že v následujícím roce maximální průtok nepřekročí 150 $\mathrm{m^3 s^{-1}}$,

b) pravděpodobnost, že rozdíl mezi skutečným maximálním průtokem a jeho střední hodnotou v absolutní hodnotě nepřekročí 30 $\mathrm{m^3 s^{-1}}$,

c) hranici tzv. stoleté vody, tj. kritický průtok, který je překročen v průměru jednou za 100 let.

Příklad 3.24.

Konstrukce má být navržena na 5 % dolní kvantil pevnosti betonu, o které je známo, že její střední hodnota je rovna 30 MPa a směrodatná odchylka 7 MPa. Na jakou pevnost má být konstrukce navržena v případě, že předpokládáme, že pevnost betonu má dvouparametrické logaritmicko-normální rozdělení, a na jakou v případě, že předpokládáme, že má normální rozdělení?

Příklad 3.25.

Odhadujeme, že střední doba životnosti určitého přístroje je 115 dnů a směrodatnou odchylku odhadujeme na 70 dnů. S jakou pravděpodobností bude životnost náhodně vybraného přístroje mezi 100 a 150 dny, použijeme-li jako teoretický model normální rozdělení, resp. použijeme-li dvouparametrické logaritmicko-normální rozdělení?

4. NÁHODNÉ VEKTORY

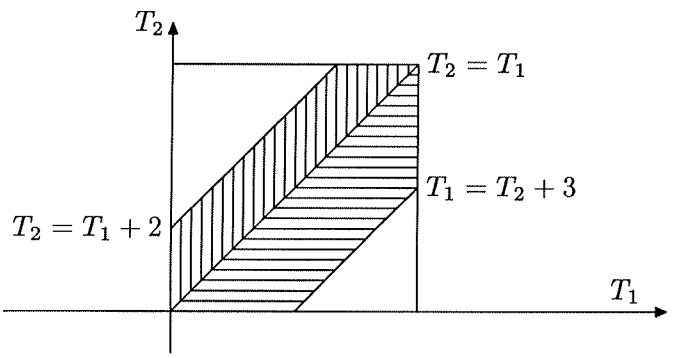
Řešené příklady

Příklad 4.1.

Dva vlaky mají přijet k určité rampě, sloužící k vykládce zboží. Každý z vlaků může přijet nezávisle na druhém kdykoliv během jednoho dne, každý čas příjezdu považujeme za stejně možný. S jakou pravděpodobností bude muset jeden z vlaků čekat na uvolnění rampy, jestliže dobu vykládky zboží u prvního vlaku odhadujeme na 2 hodiny a u druhého vlaku na 3 hodiny?

Řešení:

Čas příjezdu prvního vlaku označme T_1, čas příjezdu druhého vlaku označme T_2. Jde o nezávislé náhodné veličiny, obě s rovnoměrným rozdělením na intervalu $\langle 0, 24\rangle$. Náhodný vektor (T_1, T_2) má tedy rovnoměrné rozdělení na čtverci $\langle 0, 24\rangle \times \langle 0, 24\rangle$, tj. jeho hustota je konstantní a rovná $1/24^2$ uvnitř tohoto čtverce a nulová jinde.



První vlak bude čekat na uvolnění rampy, pokud bude $T_2 \leq T_1 < T_2 + 3$, druhý vlak bude čekat v případě, že $T_1 \leq T_2 < T_1 + 2$. Oba tyto náhodné jevy odpovídají vyšrafované ploše na obrázku. Hledaná pravděpodobnost je potom rovna objemu tělesa, jehož základnou je vyšrafovaná plocha a které má konstantní výšku, rovnou hustotě.

Tuto pravděpodobnost P vypočteme pohodlněji pomocí pravděpodobnosti doplňkového jevu:

$$P = 1 - \tfrac{1}{576}[\tfrac{1}{2}(24-3)^2 + \tfrac{1}{2}(24-2)^2] = 0.1970\,.$$

Příklad 4.2.

Svislý pilíř kruhového průřezu o poloměru 1 m je zatěžován silou působící kolmo k rovině průřezu. Působiště síly je náhodné, naše intuitivní představa je, že náhodný vektor souřadnic tohoto působiště má spojité rozdělení na průřezu pilíře, přičemž velikost hustoty je maximální ve středu a klesá úměrně s druhou mocninou vzdálenosti od středu průřezu až na nulovou hodnotu na okraji.

a) Určete hustotu takto popsaného vektoru působiště, vypočtěte marginální hustoty, střední hodnoty a rozptyly jednotlivých souřadnic působiště. Zjistěte, jsou-li tyto souřadnice nekorelované, resp. nezávislé náhodné veličiny.

b) Určete, s jakou pravděpodobností bude excentricita síly (tj. vzdálenost působiště od osy pilíře) větší než 50 cm. Jaká je střední hodnota této excentricity?

Řešení:

Označíme-li vektor souřadnic působiště síly jako (X, Y), pak má jeho hustota tvar

$$f(x,y) = \begin{cases} C - k(x^2+y^2), & \text{pro } x^2+y^2 \leq 1; \\ 0, & \text{jindy.} \end{cases}$$

kde C a k jsou nějaké kladné konstanty.

Z podmínky, že hustota má být nulová na okraji průřezu ($f(x,y) = 0$ pro $x^2+y^2 = 1$) plyne, že tyto konstanty jsou si rovny, tedy $C = k$.

Aby funkce $f(x,y)$ byla hustotou pravděpodobnosti, musí platit

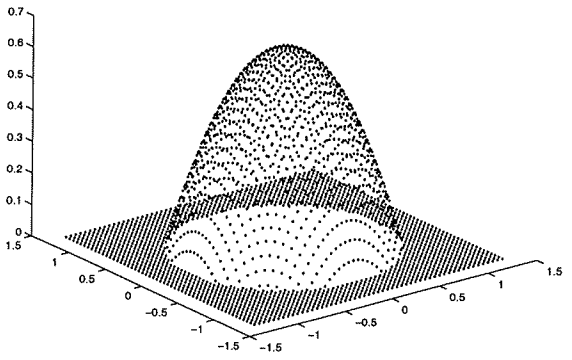
$$1 = \iint_{x^2+y^2\le 1} k(1-x^2-y^2)\,\mathrm{d}x\,\mathrm{d}y = 2\pi k \int_0^1 (1-\varrho^2)\varrho\,\mathrm{d}\varrho = \frac{1}{2}\pi k\,,$$

odkud

$$k = 2/\pi\,.$$

(Při výpočtu dvojného integrálu jsme použili substituci do polárních souřadnic.)

Odtud plyne, že sdružená hustota náhodného vektoru (X, Y) je rovna:



$$f(x,y) = \begin{cases} \frac{2}{\pi}(1-x^2-y^2), & \text{pro } x^2+y^2 \le 1; \\ 0, & \text{jindy.} \end{cases}$$

(Na obrázku je trojrozměrný graf této hustoty, jde o část rotačního paraboloidu.)

Počítejme marginální hustotu náhodné veličiny X (tj. první souřadnice vektoru působiště). Pro $|x| > 1$ je tato hustota evidentně nulová, pro $|x| \le 1$ je

$$f_1(x) = \int_{-\sqrt{1-x^2}}^{\sqrt{1-x^2}} \frac{2}{\pi}(1-x^2-y^2)\,\mathrm{d}y = \frac{8}{3\pi}(1-x^2)^{3/2}\,.$$

Analogicky dostaneme marginální hustotu veličiny Y:

$$f_2(y) = \begin{cases} \frac{8}{3\pi}(1-y^2)^{3/2}, & \text{pro } |y| \le 1; \\ 0, & \text{jindy.} \end{cases}$$

Vzhledem k tomu, že očividně neplatí pro všechna $(x,y) \in \mathbf{R}^2$ rovnost

$$f(x,y) = f_1(x)\cdot f_2(y)\,,$$

nejsou náhodné veličiny X, Y nezávislé.

Střední hodnoty veličin X a Y jsou nulové (ověřte sami!), vypočteme rozptyly, kovarianci a korelaci:

$$\operatorname{Var} X = \operatorname{Var} Y = \int_{-1}^{1} x^2 \cdot \frac{8}{3\pi}(1-x^2)^{3/2}\,\mathrm{d}x = \frac{1}{6}\,,$$

$$\operatorname{cov}(X,Y) = \iint_{x^2+y^2\le 1} xy \cdot \frac{2}{\pi}(1-x^2-y^2)\,\mathrm{d}x\,\mathrm{d}y = 0\,,$$

$$\operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{sd} X \cdot \operatorname{sd} Y} = 0.$$

Označme si R excentricitu působiště síly, tj.

$$R = \sqrt{X^2 + Y^2}\,.$$

Pro distribuční funkci veličiny R platí:

$$F_R(r) = P(R < r) = \begin{cases} \iint_{x^2+y^2\le r^2} \frac{2}{\pi}(1 - x^2 - y^2)\,\mathrm{d}x\,\mathrm{d}y = 2r^2 - r^4, & \text{pro } r \in \langle 0, 1\rangle; \\ 0, & \text{jindy.} \end{cases}$$

Hustotu R dostaneme zderivováním:

$$f_R(r) = \begin{cases} F_R'(r) = 4(r - r^3), & \text{pro } r \in \langle 0, 1\rangle; \\ 0, & \text{jindy.} \end{cases}$$

Nyní můžeme vypočítat střední hodnotu excentricity:

$$\mathrm{E}\,R = \int_0^1 r \cdot 4(r - r^3)\,\mathrm{d}r = \frac{8}{15} = 0.5333\ \mathrm{m}\,,$$

a hledanou pravděpodobnost:

$$P(R > 0.5) = 1 - F_R(0.5) = 0.5625\,.$$

Příklad 4.3.

Je třeba nalézt vzdálenost dvou míst A a C. Měření však nelze provést přímo a provádí se tak, že se měří v prvém kroku vzdálenost od A do B a v druhém kroku od B do C. Při měření vznikají náhodné chyby. Náhodné chyby přístroje, který je používán na měření vzdálenosti AB, mají směrodatnou odchylku $\sigma_1 = 1$ cm, zatímco směrodatná odchylka náhodných chyb přístroje, kterým se měří vzdálenost BC, se rovná $\sigma_2 = 3$ cm. Oba přístroje měří bez systematické chyby. Aby byl omezen vliv náhodných chyb, bylo rozhodnuto měření opakovat a za odhad konečné vzdálenosti bodů A a C vzít součet průměrů $\bar{x}_{AB} + \bar{x}_{BC}$. Cena jednoho měření je 1 000 Kč a celkově je k dispozici 10 000 Kč. Rozhodněte a matematicky odůvodněte, kolikrát by bylo optimální měřit vzdálenost AB a kolikrát BC tak, aby se vystačilo s finančními prostředky.

Řešení:

Abychom vystačili s finančními prostředky můžeme provést celkem 10 měření. Jestliže n krát změříme vzdálenost bodů A a B, budeme vzdálenost bodů B a C měřit $10-n$ krát. Optimální počet n je takový, při němž je rozptyl odhadu $\mathrm{Var}(\bar{x}_{AB} + \bar{x}_{BC}) = \frac{1}{n} + \frac{9}{10-n}$ nejmenší. Spočteme-li rozptyl pro všechny hodnoty $n = 1, \ldots, 9$, pak zjistíme, že rozptyl je nejmenší pro $n = 3$ a rovná se 1.619.

Odpověď: Optimální je měřit třikrát vzdálenost bodů AB a sedmkrát vzdálenost BC.

Příklad 4.4.

V dílně pracuje 20 strojů, z toho 12 starých a 8 nových. Pravděpodobnost poruchy starého stroje během pracovní směny je 0.3, tatáž pravděpodobnost je pro nový

stroj 0.1. Určete střední hodnotu a směrodatnou odchylku počtu strojů, které se během směny porouchají.

Řešení:

Označme si

P ... celkový počet strojů, u kterých dojde k poruše,
P_1 ... počet starých strojů, u kterých dojde k poruše,
P_2 ... počet nových strojů, u kterých dojde k poruše.

Zřejmě platí $P = P_1 + P_2$, přičemž náhodné veličiny P_1 a P_2 jsou nezávislé, každá s binomickým rozdělením s parametry $n_1 = 12$, $p_1 = 0.3$, resp. $n_2 = 8$, $p_2 = 0.1$.

Proto

$$\begin{aligned} \mathrm{E}\,P &= \mathrm{E}\,P_1 + \mathrm{E}\,P_2 = n_1 p_1 + n_2 p_2 = 4.4\,, \\ \mathrm{Var}\,P &= \mathrm{Var}\,P_1 + \mathrm{Var}\,P_2 = n_1 p_1 (1-p_1) + n_2 p_2 (1-p_2) = 3.24\,, \\ \mathrm{sd}\,P &= \sqrt{\mathrm{Var}\,P} = 1.8\,. \end{aligned}$$

Odpověď: Střední hodnota počtu strojů, u kterých dojde k poruše, je 4.4, směrodatná odchylka je 1.8.

Příklad 4.5.

Dobu čekání na obsluhu v opravně obuvi v určitou denní dobu modelujeme exponenciálním rozdělením se střední hodnotou 1/2 hodiny. Zákazník odhaduje, že oprava jeho bot může trvat 5, 10 nebo 15 minut (podle závažnosti opravy, kterou sám neumí posoudit) – všechny varianty považuje za stejně možné. Jakou dobu si musí rezervovat, aby během ní s 90% pravděpodobností celou opravu vyřídil?

Řešení:

Označme si jako T celkovou dobu, kterou zákazník v opravně stráví. T je součtem doby C čekání na opravu a doby O vlastní opravy, přičemž C a O jsou zjevně nezávislé náhodné veličiny. O je diskrétní náhodná veličina, která může nabývat hodnot 5, 10, 15 s pravděpodobnostmi 1/3. C je spojitá náhodná veličina s exponenciálním rozdělením s parametrem $\delta = 30$ (střední dobu čekání je třeba vyjádřit ve stejných jednotkách, tj. v minutách). Pro hustotu a distribuční funkci veličiny C tedy platí:

$$f_C(x) = \begin{cases} \frac{1}{30}\mathrm{e}^{-x/30}, & \text{pro } x > 0; \\ 0, & \text{jindy.} \end{cases}$$

$$F_C(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \begin{cases} 1 - \mathrm{e}^{-x/30}, & \text{pro } x > 0; \\ 0, & \text{jindy.} \end{cases}$$

Pro $x > 15$ můžeme vypočítat:

$$\begin{aligned} P(T < x) &= P(T < x \wedge O = 5) + P(T < x \wedge O = 10) + P(T < x \wedge O = 15) = \\ &= P(C < x - 5) \cdot P(O = 5) + P(C < x - 10) \cdot P(O = 10) + \\ &\quad + P(C < x - 15) \cdot P(O = 15) = \\ &= \tfrac{1}{3}\left[(1 - \mathrm{e}^{-(x-5)/30}) + (1 - \mathrm{e}^{-(x-10)/30}) + (1 - \mathrm{e}^{-(x-15)/30})\right] = \\ &= 1 - \tfrac{1}{3}(\mathrm{e}^{1/6} + \mathrm{e}^{1/3} + \mathrm{e}^{1/2})\mathrm{e}^{-x/30}\,. \end{aligned}$$

Pro $10 < x \leq 15$ je třeba předchozí výpočet upravit, $P(C < x - 15)$ je totiž rovno 0 a proto dostaneme

$$P(T < x) = \tfrac{1}{3}\left[(1 - \mathrm{e}^{-(x-5)/30}) + (1 - \mathrm{e}^{-(x-10)/30})\right] = \tfrac{2}{3} - \tfrac{1}{3}(\mathrm{e}^{1/6} + \mathrm{e}^{1/3})\mathrm{e}^{-x/30}.$$

Obdobné úvahy platí pro $5 < x \leq 10$, pro $x \leq 5$ je zřejmě $P(T < x) = 0$. Distribuční funkci veličiny T lze souhrnně zapsat předpisem

$$F_T(x) = \begin{cases} 0, & \text{pro } x \leq 5; \\ \tfrac{1}{3}(1 - \mathrm{e}^{1/6}\mathrm{e}^{-x/30}), & \text{pro } 5 < x \leq 10; \\ \tfrac{2}{3} - \tfrac{1}{3}(\mathrm{e}^{1/6} + \mathrm{e}^{1/3})\mathrm{e}^{-x/30}, & \text{pro } 10 < x \leq 15; \\ 1 - \tfrac{1}{3}(\mathrm{e}^{1/6} + \mathrm{e}^{1/3} + \mathrm{e}^{1/2})\mathrm{e}^{-x/30}, & \text{pro } x > 15. \end{cases}$$

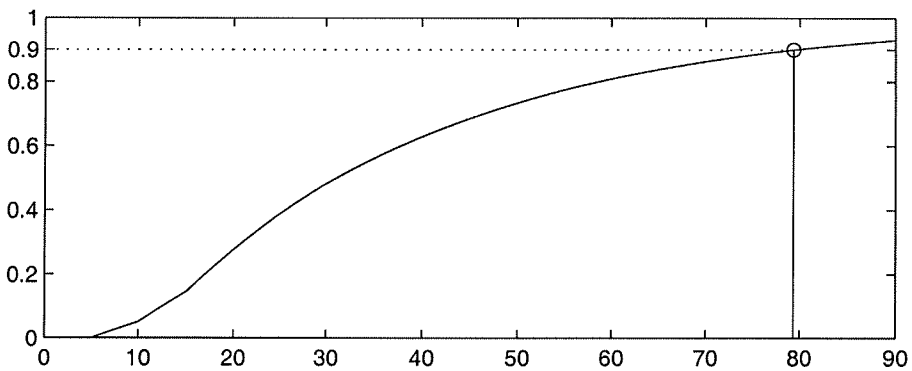
Hledáme takové x, pro které platí

$$F_T(x) = 0.9\,.$$

Z názoru je zřejmé, že musí být $x > 15$, tj. řešíme rovnici

$$1 - \tfrac{1}{3}(\mathrm{e}^{1/6} + \mathrm{e}^{1/3} + \mathrm{e}^{1/2})\mathrm{e}^{-x/30} = 0.9 \Longrightarrow$$
$$\Longrightarrow -x/30 = \ln 0.1 + \ln 3 - \ln(\mathrm{e}^{1/6} + \mathrm{e}^{1/3} + \mathrm{e}^{1/2}) \Longrightarrow x = 79.355.$$

Na obrázku je graf distribuční funkce F_T spolu s nalezeným bodem x.



Odpověď: Zákazník si musí rezervovat asi 1 hodinu a 19 minut.

Příklad 4.6.

Doba jednání referenta s občanem na určitém úřadě se řídí exponenciálním rozdělením se střední hodnotou 10 minut. Přijdeme-li do čekárny před referentovou kanceláří, ve které právě probíhá jednání, s jakou pravděpodobností budeme čekat

a) déle než 10 minut, není-li v čekárně kromě nás nikdo,

b) déle než 20 minut, je-li před námi ve frontě 1 člověk,

c) déle než 30 minut, jsou-li před námi 2 lidé?

Jaké je rozdělení, střední hodnota a směrodatná odchylka naší doby čekání v případech a), b), c)?

Řešení:

a) Naše doba čekání T_a je rovna zbývající době jednání referenta s občanem uvnitř. Tato zbývající doba má však opět exponenciální rozdělení se střední hodnotou 10 minut, bez ohledu na to, jak dlouho toto jednání již před naším příchodem probíhalo. Výše zmíněná zajímavá vlastnost exponenciálního rozdělení již byla komentována v řešení příkladu 3.6. (Uvážíme-li praktické zkušenosti s jednáním na úřadech, je tento teoretický model docela výstižný.)

Tedy

$$\mathrm{E}\,T_a = \mathrm{sd}\,T_a = 10, \qquad P(T_a > 10) = \int_{10}^{\infty} \frac{1}{10}\mathrm{e}^{-x/10}\,\mathrm{d}x = \mathrm{e}^{-10/10} = 0.3679\,.$$

b) Naše doba čekání T_b je v tomto případě rovna součtu doby T_1 zbývajícího jednání s prvním občanem a doby T_2 jednání s druhým občanem. Rozdělení obou veličin T_1 a T_2 je stejné (exponenciální s parametrem $\delta = 10$), přitom tyto veličiny můžeme považovat za nezávislé.

V dalším textu budeme hledat hustotu veličiny T_b. Provedeme však obecnější výpočet, jehož výsledek se nám může v budoucnu hodit.

Předpokládejme, že veličiny X a Y jsou nezávislé a spojité s hustotami $f(x)$ a $g(y)$. Odvodíme hustotu veličiny $Z = X + Y$.

Hustota náhodného vektoru (X, Y) je $h(x, y) = f(x) \cdot g(y)$. Počítejme distribuční funkci veličiny Z:

$$F_Z(z) = P(X + Y < z) = \iint_{x+y<z} f(x)g(y)\,\mathrm{d}x\,\mathrm{d}y =$$
$$= \int_{-\infty}^{\infty} f(x)\left(\int_{-\infty}^{z-x} g(y)\,\mathrm{d}y\right)\mathrm{d}x = \int_{-\infty}^{\infty} f(x)\cdot F_Y(z-x)\,\mathrm{d}x\,.$$

Hustotu veličiny Z nyní dostaneme zderivováním distribuční funkce:

$$f_Z(z) = \int_{-\infty}^{\infty} f(x)g(z-x)\,\mathrm{d}x\,.$$

Poznamenejme, že právě odvozené hustotě součtu dvou nezávislých spojitých náhodných veličin se říká *hustota konvoluce*. Použijeme nyní odvozený výsledek pro případ, kdy X a Y jsou nezávislé stejně rozdělené s exponenciálním rozdělením o parametru δ:

Jelikož

$$f(x) = \begin{cases} \frac{1}{\delta}\mathrm{e}^{-x/\delta}, & \text{pro } x > 0, \\ 0, & \text{pro } x \le 0, \end{cases} \qquad g(y) = \begin{cases} \frac{1}{\delta}\mathrm{e}^{-y/\delta}, & \text{pro } y > 0, \\ 0, & \text{pro } y \le 0, \end{cases}$$

platí

$$f_Z(z) = \begin{cases} \displaystyle\int_0^z \frac{1}{\delta}\mathrm{e}^{-x/\delta}\cdot\frac{1}{\delta}\mathrm{e}^{-(z-x)/\delta}\,\mathrm{d}x = \frac{1}{\delta^2}\int_0^z \mathrm{e}^{-z/\delta}\,\mathrm{d}x = \frac{z}{\delta^2}\mathrm{e}^{-z/\delta}, & \text{pro } z > 0; \\ 0, & \text{pro } z \le 0. \end{cases}$$

Dosadíme-li hodnotu parametru $\delta = 10$, dostaneme hustotu veličiny T_b:

$$f_{T_b}(z) = \begin{cases} \frac{z}{100}\mathrm{e}^{-z/10}, & \text{pro } z > 0; \\ 0, & \text{pro } z \leq 0. \end{cases}$$

Z toho už snadno dostaneme požadované výsledky:

$$\mathrm{E}\,T_b = \mathrm{E}\,T_1 + \mathrm{E}\,T_2 = 20, \quad \mathrm{Var}\,T_b = \mathrm{Var}\,T_1 + \mathrm{Var}\,T_2 = 200 \implies \mathrm{sd}\,T_b = 14.14\,,$$
$$P(T_b > 20) = \int_{20}^{\infty} \frac{x}{100}\mathrm{e}^{-x/10}\,\mathrm{d}x = \left[\mathrm{e}^{-x/10}\left(-\frac{x}{10} - 1\right)\right]_{20}^{\infty} =$$
$$= \mathrm{e}^{-20/10}\left(\frac{20}{10} + 1\right) = 0.4060\,.$$

c) V posledním případě je doba čekání T_c rovna součtu doby T_b a doby jednání T_3 se třetím občanem. Vzhledem k nezávislosti, kterou opět budeme předpokládat, je její hustota konvolucí hustot veličin T_b a T_3:

$$f_{T_c}(z) = \int_{-\infty}^{\infty} f_{T_b}(x) f_{T_3}(z - x)\,\mathrm{d}x =$$
$$= \begin{cases} \displaystyle\int_0^z \frac{x}{100}\mathrm{e}^{-x/10} \cdot \frac{1}{10}\mathrm{e}^{-(z-x)/10}\,\mathrm{d}x = \frac{z^2}{2000}\mathrm{e}^{-z/10}, & \text{pro } z > 0; \\ 0, & \text{pro } z \leq 0. \end{cases}$$

Proto

$$\mathrm{E}\,T_c = \mathrm{E}\,T_b + \mathrm{E}\,T_3 = 30, \quad \mathrm{Var}\,T_c = \mathrm{Var}\,T_b + \mathrm{Var}\,T_3 = 300 \implies \mathrm{sd}\,T_b = 17.32\,,$$
$$P(T_b > 30) = \int_{30}^{\infty} \frac{x^2}{2000}\mathrm{e}^{-x/10}\,\mathrm{d}x = \left[\mathrm{e}^{-x/10}\left(-\frac{x^2}{200} - \frac{x}{10} - 1\right)\right]_{30}^{\infty} =$$
$$= \mathrm{e}^{-30/10}\left(\frac{900}{200} + \frac{30}{10} + 1\right) = 0.4232\,.$$

Poznamenejme na závěr, že rozdělení, které vzniká součtem několika nezávislých exponenciálních rozdělení s týmž parametrem se nazývá *Erlangovo.*

Toto rozdělení má dva parametry:

δ ... parametr nezávislých exponenciálně rozdělených veličin, jejichž součtem Erlangovo rozdělení vzniká,

p ... přirozené číslo, označující počet těchto veličin.

My jsme vlastně již odvodili hustotu Erlangova rozdělení pro $p = 2$ a 3, indukcí lze snadno ukázat, že obecně je

$$f_{E(p,\delta)}(x) = \begin{cases} \frac{x^{p-1}}{\delta^p (p-1)!}\mathrm{e}^{-x/\delta}, & \text{pro } x > 0; \\ 0, & \text{pro } x \leq 0. \end{cases}$$

Všimněte si dále, že jsme vlastně v úlohách a), b), c) počítali pravděpodobnosti, že určitá náhodná veličina bude větší než její střední hodnota. Taková pravděpodobnost

by pro rozdělení se symetrickou hustotou byla rovna 0.5. V našem případě jsou tyto pravděpodobnosti menší než 0.5 a postupně rostou. To odpovídá faktu, že Erlangovo rozdělení má s rostoucí hodnotou parametru p stále menší šikmost.

Odpověď: Nebude-li před námi ve frontě nikdo, budeme čekat s pravděpodobností 0.3679 déle než 10 minut. Naše doba čekání má exponenciální rozdělení se střední hodnotou 10 minut, směrodatná odchylka je také 10 minut.

Bude-li před námi jeden člověk, budeme čekat s pravděpodobností 0.406 déle než 20 minut. Naše doba čekání bude mít rozdělení s hustotou f_{T_b} (viz řešení). Její střední hodnota bude 20 minut a směrodatná odchylka 14.14 minuty.

Budou-li před námi dva lidé, budeme čekat s pravděpodobností 0.4232 déle než 30 minut. Naše doba čekání bude mít rozdělení s hustotou f_{T_c} (viz řešení). Její střední hodnota bude 30 minut a směrodatná odchylka 17.23 minuty.

Příklad 4.7.

Dvě nezávislá měření se provedou přístrojem, který má směrodatnou odchylku 30 m a systematickou chybu +10 m. Jaká je pravděpodobnost toho, že chyby obou měření budou v absolutní hodnotě větší než 10 m, přičemž budou mít různá znaménka?

Řešení:

Označme X_1 chybu prvního měření a X_2 chybu druhého měření. Jedná se o chyby měření, a proto budeme předpokládat, že jsou obě normálně rozdělené. Chceme spočítat pravděpodobnost toho, že buď $X_1 > 10$ a současně $X_2 < -10$ nebo $X_1 < -10$ a současně $X_2 > 10$:

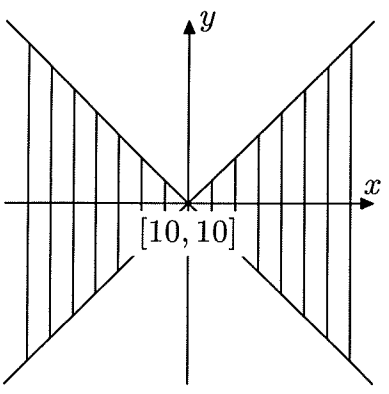
$$P(X_1 > 10 \cap X_2 < -10) + P(X_1 < -10 \cap X_2 > 10) =$$

$$2\,(1 - \Phi(0))\,\Phi\left(\frac{-20}{30}\right) = \Phi\left(\frac{-20}{30}\right) = 0.2525.$$

Při výpočtu jsme využili předpokladu, podle kterého jsou chyby měření nezávislé, a tedy pravděpodobnost průniku se rovná součinu pravděpodobností.

Odpověď: Pravděpodobnost toho, že chyby obou měření budou v absolutní hodnotě větší než 10 m, přičemž budou mít různá znaménka, je rovna přibližně 0.25.

Příklad 4.8.



Dvakrát po sobě vážíme závaží, jehož skutečná hmotnost je 10 g. Hodnoty X_1 a X_2, které vážením získáme, jsou díky chybám vážení náhodné veličiny, o kterých budeme předpokládat, že jsou nezávislé, stejně rozdělené s $N(\mu, \sigma^2)$ rozdělením, kde střední hodnota $\mu = 10$ g a směrodatná odchylka $\sigma = 0.2$ g.

a) Spočtěte pravděpodobnost, že výsledek druhého vážení je blíže 10 g než výsledek prvního vážení.

b) Spočtěte pravděpodobnost, že výsledek druhého vážení je menší než výsledek prvního vážení, ale ne o více než 0.2 g.

Řešení:

Náhodný vektor (X_1, X_2) má dvourozměrné normální rozdělení $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, kde $\mu_1 = 10$ g, $\mu_2 = 10$ g, $\sigma_1 = 0.2$ g, $\sigma_2 = 0.2$ g a $\rho = 0$. Hustota takového vektoru má tvar rotační plochy se středem v bodě $(10, 10)$.

a) Pravděpodobnost, že výsledek druhého vážení je blíže 10 g než výsledek prvního vážení, tj.

$$P(|X_2 - 10| < |X_1 - 10|),$$

odpovídá objemu tělesa s podstavou vyšrafovanou na obrázku a shora omezeného hustotou. Ze symetrie vyplývá, že hledaná pravděpodobnost je rovna 1/2.

b) Pravděpodobnost, že výsledek druhého vážení je menší než výsledek prvního vážení, ale ne o více než 0.2 g, je rovna:

$$P(X_1 - 0.2 < X_2 < X_1) = P(-0.2 < X_2 - X_1 < 0).$$

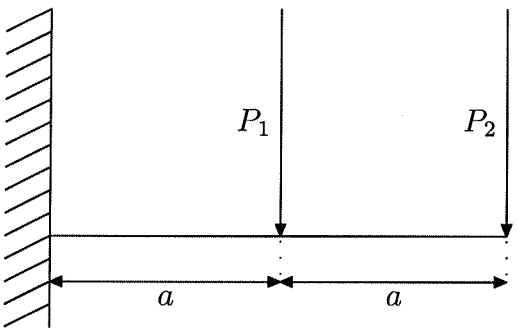
Náhodná veličina $Y = X_2 - X_1$ má normální rozdělení se střední hodnotou $\mu_Y = \mu_2 - \mu_1 = 0$ a směrodatnou odchylkou $\sigma_Y = \sqrt{\sigma_1^2 + \sigma_2^2} = 0.2\sqrt{2}$. Odtud

$$P(Y \in (-0.2, 0)) = \Phi\left(\frac{0}{0.2\sqrt{2}}\right) - \Phi\left(\frac{-0.2}{0.2\sqrt{2}}\right) = 0.5 - (1 - \Phi(0.7071)) \doteq 0.2602.$$

Odpověď: Pravděpodobnost, že výsledek druhého vážení je blíže 10 g než výsledek prvního vážení, je rovna 0.5. Pravděpodobnost, že výsledek druhého vážení je menší než výsledek prvního vážení, ale ne o více než 0.2 g, je rovna 0.26.

Příklad 4.9.

Nosník je zatěžován silami P_1 a P_2 měřenými v kN, které mohou být považovány za náhodné veličiny normálně rozdělené s vektorem středních hodnot μ a kovarianční maticí Σ:



$$\mu = \begin{pmatrix} 10 \\ 10 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 9 & 6 \\ 6 & 9 \end{pmatrix}.$$

Momentová kapacita nosníku je náhodná veličina B dána v kNm nezávislá na (P_1, P_2) se střední hodnotou $\mathrm{E}\,B = 250$ a rozptylem $\mathrm{Var}\,B = 900$.

Spočtěte, s jakou pravděpodobností dojde k poruše, jestliže $a = 4$ m.

Řešení:

Síly P_1 a P_2 jsou závislé, přičemž

$$\mathrm{cov}(P_1, P_2) = 6 \qquad \text{a} \qquad \mathrm{corr}(P_1, P_2) = \frac{\mathrm{cov}(P_1, P_2)}{\sqrt{\mathrm{Var}\,P_1\ \mathrm{Var}\,P_2}} = \frac{6}{3 \times 3} = \frac{2}{3}.$$

Ohybový moment

$$M = a\,P_1 + 2a\,P_2$$

má normální rozdělení se střední hodnotou $\mathrm{E}\,M = a\,\mathrm{E}\,P_1 + 2a\,\mathrm{E}\,P_2$ a rozptylem $\mathrm{Var}\,M = a^2\,\mathrm{Var}\,P_1 + (2a)^2\,\mathrm{Var}\,P_2 + 2 \times a \times 2a\,\mathrm{cov}(P_1, P_2)$. Pro $a = 4$ m platí $\mathrm{E}\,M = 120$ a $\mathrm{Var}\,M = 1104$. Rozdíl mezi momentem a momentovou kapacitou nosníku $R = M - B$ má normální rozdělení se střední hodnotou $\mathrm{E}\,R = \mathrm{E}\,M - \mathrm{E}\,B = -130$ a rozptylem $\mathrm{Var}\,R = \mathrm{Var}\,M + \mathrm{Var}\,B = 2004$. Nosník se poruší, jestliže je ohybový moment větší než momentová kapacita nosníku, tj. $R > 0$. Pravděpodobnost takového jevu se rovná:

$$P(R > 0) = 1 - \Phi\left(\frac{130}{\sqrt{2004}}\right) \doteq 0.0018.$$

Odpověď: Nosník se poruší s pravděpodobností 0.0018.

Příklad 4.10.

Uvažujeme dvě řeky. Denní průtoky první řeky v m^3/s jsou realizacemi náhodné veličiny X s logaritmicko-normálním rozdělením $LN(\mu_1 = 3, \sigma_1^2 = 0.5)$. Denní průtoky druhé řeky v m^3/s jsou realizacemi náhodné veličiny Y s logaritmicko-normálním rozdělením $LN(\mu_2 = 2.5, \sigma_2^2 = 1.5)$.

a) Jaký průměrný průtok má první řeka a jaký průměrný průtok má druhá řeka?

b) Považujeme-li průtoky v první a druhé řece za nezávislé, s jakou pravděpodobností bude průtok první řeky větší než průtok druhé? Kdy je možno průtoky v první a druhé řece považovat za nezávislé?

Řešení:

Z přísného matematického hlediska bychom měli v zadání úlohy předpokládat, že denní průtoky tvoří ergodickou stacionární posloupnost. (To se však v praxi obvykle mlčky předpokládá.) Průměrný průtok se pak při velkém počtu realizací blíží střední hodnotě, kde

$$\mathrm{E}\,X = \mathrm{e}^{\mu_1 + \sigma_1^2/2} = \mathrm{e}^{3+0.5/2} = 25.79\,\mathrm{m}^3/\mathrm{s},$$
$$\mathrm{E}\,Y = \mathrm{e}^{\mu_2 + \sigma_2^2/2} = \mathrm{e}^{2.5+1.5/2} = 25.79\,\mathrm{m}^3/\mathrm{s}.$$

Jestliže $X \sim LN(\mu_1, \sigma_1^2)$ a $Y \sim LN(\mu_2, \sigma_2^2)$, pak $\ln X \sim N(\mu_1, \sigma_1^2)$ a $\ln Y \sim N(\mu_2, \sigma_2^2)$. Považujeme-li veličiny X a Y nezávislé, pak rovněž veličiny $\ln X$ a $\ln Y$ jsou nezávislé a veličina $\ln X - \ln Y \sim N(0.5, 2)$. Odtud

$$P(X > Y) = P(\ln X - \ln Y > 0) = 1 - \Phi\left(\frac{-0.5}{\sqrt{2}}\right) = 1 - \Phi(-0.35355) \doteq 0.6382.$$

Odpověď: Průměrné střední denní průtoky jsou u obou řek přibližně stejné, rovné 25.79 m^3/s. S pravděpodobností 0.6382 je však průtok v první řece vyšší než v řece druhé. Tento výsledek je způsoben výrazně vyšším sešikmením druhého rozdělení. U druhé řeky se tedy zřejmě vyskytují občas velmi vysoké průtoky, tj. výrazně vpravo odlehlá pozorování, která zvyšují průměr. Většina průtoků je však nižších než u první řeky.

V aplikacích bychom mohli považovat průtoky dvou řek za nezávislé asi tehdy, jestliže je jejich povodí vzdálené, a tudíž množství srážek, které spadne v obou povodích, je možno považovat za nezávislé.

Neřešené příklady

Příklad 4.11.

Náhodný vektor (X, Y) má rovnoměrné rozdělení na kruhu $K = \{(x, y);\ x^2+y^2 \leq 1\}$ s hustotou:

$$f(x,y) = \begin{cases} \frac{1}{\pi}, & \text{pro } (x,y) \in K; \\ 0, & \text{pro } (x,y) \notin K. \end{cases}$$

a) Spočtěte marginální hustoty X a Y.
b) Zjistěte, zda veličiny X a Y jsou nezávislé.
c) Zjistěte, zda veličiny X a Y jsou nekorelované.

Příklad 4.12.

Nezávisle na sobě vygenerujeme dvě náhodná čísla s rovnoměrným rozdělením na intervalu $\langle 0, 5\rangle$. S jakou pravděpodobností
a) bude jejich součet menší než 2,
b) se budou lišit nejvýše o 1?

Příklad 4.13.

Na kružnici rovnoměrně náhodně a nezávisle na sobě vybereme tři body A, B, C. S jakou pravděpodobností bude trojúhelník ABC ostroúhlý?

Příklad 4.14.

V určitou denní dobu mají soupravy metra na všech trasách pravidelné časové odstupy 8 minut. Uvažujme cestujícího, který přijde v náhodném časovém okamžiku do stanice na trase A, dojede na přestupní stanici a dále pokračuje trasou B. S jakou pravděpodobností stráví čekáním na soupravy celkem více než 10 minut? (Předpokládejte, že jízdní řády obou tras na sobě nezávisí.)

Příklad 4.15.

Určete pravděpodobnost, že kvadratická rovnice

$$x^2 + px + q = 0$$

bude mít reálné kořeny, jestliže koeficienty p a q nezávisle vygenerujeme jako náhodná čísla, obě s rovnoměrným rozdělením na intervalu $\langle -1, 1\rangle$.

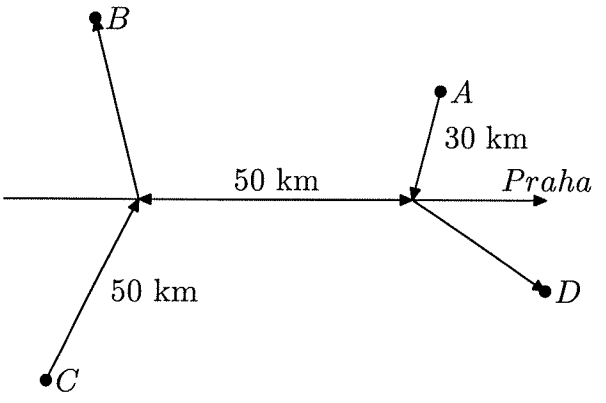
Příklad 4.16.

Dva lidé se domluvili, že se nezávisle na sobě někdy během 21 až 23 hodiny zastaví na určitém večírku. S jakou pravděpodobností se minou, pokud se každý z nich zdrží 45 minut? (Předpokládejte, že doby příchodu jsou rovnoměrně rozdělené na daném časovém intervalu.)

Příklad 4.17.

Na úsečce o délce l jsou náhodně umístěny body A a B, které rozdělí úsečku na tři menší úsečky. S jakou pravděpodobností se z těchto tří úseček dá sestavit trojúhelník?

Příklad 4.18.



Řidič Jarda vyjíždí denně z města A do města B rychlostí 60 km/hod. Řidič Pavel vyjíždí denně z města C do města D rychlostí 50 km/hod. Řidiči projíždějí část cesty dlouhé 50 km po státní silnici, ale v opačném směru. Řidič Jarda najíždí na státní silnici 30 km od města A, řidič Pavel 50 km od města C. S jakou pravděpodobností se oba řidiči na státní silnici potkají, jestliže oba vyjíždějí v libovolný čas mezi 10 a 14 hodinou?

Příklad 4.19.

Dokažte pomocí hustoty konvoluce, že mají-li dvě nezávislé náhodné veličiny rovnoměrné rozdělení na intervalu $\langle -a/2, a/2\rangle$, potom má jejich součet trojúhelníkové rozdělení na intervalu $\langle -a, a\rangle$ (viz příklad 3.14).

Příklad 4.20.

Střední doba obsluhy automobilu u stojanu čerpací stanice (do této doby je započten příjezd z fronty ke stojanu, vlastní čerpání, zaplacení a odjezd od stojanu) je 9 minut. S jakou pravděpodobností se za půl hodiny obslouží

a) více než 5 aut,
b) méně než 3 auta?

(Předpokládejte exponenciální rozdělení doby obsluhy a dále, že se fronta nikdy nevyprázdní.)

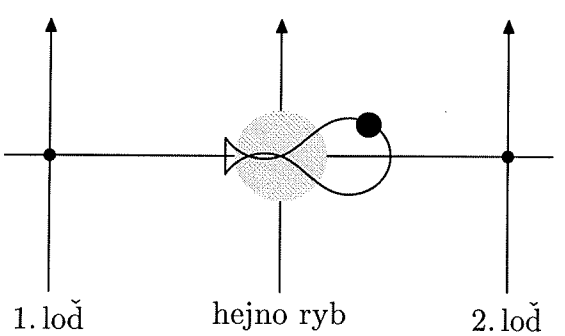
Příklad 4.21.

Délka skoků sportovce Petra měřená v cm má normální rozdělení $N(\mu_1, \sigma_1^2)$, kde $\mu_1 = 690$ a $\sigma_1 = 10$. Délka skoků sportovce Pavla měřená v cm má také normální rozdělení $N(\mu_2, \sigma_2^2)$, kde $\mu_2 = 705$ a $\sigma_2 = 15$. Na závody se kvalifikuje ten, kdo ze dvou skoků alespoň jednou skočí více než 700 cm.

a) S jakou pravděpodobností se oba kvalifikují na závody?

b) S jakou pravděpodobností se kvalifikuje Pavel, ale Petr ne?

Příklad 4.22.



Dvě rybářské lodi plují rovnoběžně vedle sebe. Šířka pásu v km, ve kterém se dá z lodi spatřit hejno ryb, je náhodná veličina, která má normální rozdělení $N(\mu, \sigma^2)$ s parametry $\mu = 3$ a $\sigma = 1.09712$. Pro obě lodi jsou tyto veličiny nezávislé. Jak daleko od sebe mají lodi plout, aby s pravděpodobností 0.5 bylo objeveno hejno ryb, které pluje uprostřed mezi nimi stejným směrem?

Příklad 4.23.

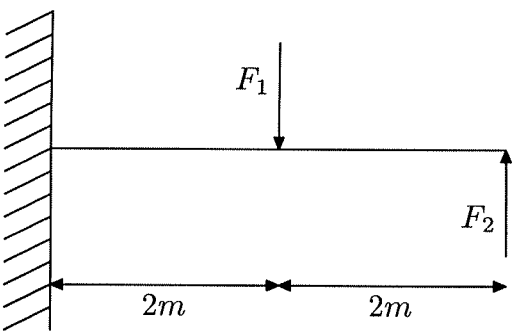
Zájemci o pilotování letadel musí projít fyzickým a dvěma psychologickými testy. Předpokládáme, že zisky bodů v jednotlivých testech mají přibližně normální rozdělení, přičemž průměrný (střední) počet bodů z fyzického testu je 100 a směrodatná odchylka je rovna 15. Průměrný (střední) počet bodů v prvním psychologickém testu je 80 a směrodatná odchylka je rovna 15, zatímco průměrný (střední) počet v druhém psychologickém testu je 120 a směrodatná odchylka je rovna 30. Fyzický test je nezávislý s oběma psychologickými testy, avšak psychologické testy jsou silně zkorelované s korelačním koeficientem $\rho = 0.7$. Celkový počet bodů se počítá tak, že se vezme průměr z psychologických testů a ten se zprůměruje s výsledkem z testu fyzického. Přijímací řízení splňují ti, jejichž výsledek je vyšší než 110 bodů.

Kolik procent zájemců splní podmínky přijímacího řízení?

Příklad 4.24.

Veličiny X_1, X_2 a X_3 pozorujeme s chybami majícími nulové střední hodnoty a směrodatné odchylky po řadě 4, 2 a 1 cm. S jakou pravděpodobností bude chyba aritmetického průměru z těchto veličin v absolutní hodnotě menší než 3 cm za předpokladu normálního rozdělení a nezávislosti těchto veličin?

Příklad 4.25.



Na vetknutý nosník působí dvě síly (viz obrázek) o náhodných velikostech. Předpokládáme, že vektor velikostí těchto sil má dvojrozměrné normální rozdělení, přičemž střední hodnoty sil F_1, F_2 jsou 10 kN, resp. 5 kN, směrodatné odchylky jsou 2, resp. 1 kN a korelační koeficient sil je 0.8. S jakou pravděpodobností nepřekročí ohybový moment ve vetknutí v absolutní hodnotě 2 kNm?

Příklad 4.26.

Plechový štítek má tvar obdélníka. Délky jeho stran měřené v cm X a Y jsou nezávislé náhodné veličiny, přičemž $X \sim LN(\mu_1, \sigma_1^2)$ a $Y \sim LN(\mu_2, \sigma_2^2)$, kde $\mu_1 = 1.386$ a $\sigma_1^2 = 0.000049$; $\mu_2 = 1.609$ a $\sigma_2^2 = 0.000121$.

a) Jaké jsou střední hodnoty a rozptyly rozměrů štítku?

b) S jakou pravděpodobností se liší obsah štítků od správné hodnoty $20\,\mathrm{cm}^2$ o více než $0.5\,\mathrm{cm}^2$?

5. Centrální limitní věta

Řešené příklady

Příklad 5.1.

Před volbami je v populaci státu 52% příznivců koaličních stran. Jaká je pravděpodobnost, že průzkum veřejného mínění o rozsahu $n = 1500$ ukáže nesprávně převahu opozice?

Řešení:

Označme X počet příznivců koalice ve výběru. Pokud byl výběr proveden zcela náhodně, pak X má binomické rozdělení s parametry $n = 1500$ a $p = 0.52$, které můžeme aproximovat normálním rozdělením o parametrech $\mu = np = 1500 \times 0.52 = 780$ a $\sigma^2 = np(1-p) = 1\,500 \times 0.52 \times 0.48 = 374.4$. Průzkum ukáže nesprávně převahu opozice, jestliže $X \leq 749$:

$$P(X \leq 749) = \Phi\left(\frac{749 + 0.5 - 780}{\sqrt{374.4}}\right) = \Phi(-1.576) \doteq 0.0575.$$

Poznamenejme, že jsme použili při výpočtu opravu na spojitost.

Odpověď: Průzkum o rozsahu $n = 1500$ ukáže nesprávně převahu opozice s pravděpodobností přibližně 0.0575.

Příklad 5.2.

Výletní člun má nosnost 5000 kg. Hmotnost (váha) cestujících je náhodná veličina se střední hodnotou 70 kg a směrodatnou odchylkou 20 kg. Kolik cestujících může člunem cestovat, aby pravděpodobnost přetížení člunu byla menší než 0.001?

Řešení:

Celková hmotnost X všech n cestujících má podle centrální limitní věty přibližně normální rozdělení se střední hodnotou $\mu = 70\,n$ a rozptylem $\sigma^2 = 400\,n$. Chceme najít takové maximální n, aby $P(X \geq 5000) < 0.001$.

$$\begin{aligned} P(X \geq 5000) = 1 - \Phi\left(\frac{5000 - 70\,n}{20\sqrt{n}}\right) &< 0.001, \\ \Phi\left(\frac{5000 - 70\,n}{20\sqrt{n}}\right) &> 0.999, \\ \frac{5000 - 70\,n}{20\sqrt{n}} &> 3.0902, \\ 70\,n + 61.804\sqrt{n} - 5000 &< 0, \\ n &\leq 64. \end{aligned}$$

Odpověď: Nejvyšší možná únosnost člunu za těchto podmínek je 64 osob.

Příklad 5.3.

Pro výpočet integrálu $I = \int_0^1 g(x)\,\mathrm{d}x$ lze použít metodu Monte Carlo. Výpočet probíhá tak, že generujeme náhodná čísla $\{X_i,\ i = 1, \ldots, n\}$ z rovnoměrného rozdělení na intervalu $\langle 0, 1\rangle$ a spočteme $\hat{I} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$. Jestliže počet vygenerovaných čísel n je dost velký, pak $\hat{I} \simeq \int_0^1 g(x)\,\mathrm{d}x$.

Spočtěte, s jakou pravděpodobností při výpočtu integrálu $\int_0^1 \sqrt{x}\,\mathrm{d}x$ metodou Monte Carlo bude chyba výsledku menší než 0.001, jestliže jsme pro výpočet použili 10000 vygenerovaných náhodných čísel, o kterých lze předpokládat, že jsou to realizace nezávislých náhodných veličin, řídících se $R(0, 1)$ rozdělením.

Řešení:

Nejprve si uvědomme, že distribuční funkce náhodné veličiny X, mající rovnoměrné rozdělení na intervalu $\langle 0, 1\rangle$, má tvar $F_X(x) = P(X < x)$:

$$F_X(x) = \begin{cases} 0, & \text{pro } x < 0; \\ x, & \text{pro } x \in \langle 0, 1\rangle; \\ 1, & \text{pro } x > 1. \end{cases}$$

Pro distribuční funkci $F_Y(y)$ náhodné veličiny $Y = \sqrt{X}$ platí pro $y \geq 0$:

$$F_Y(y) = P(Y < y) = P(\sqrt{X} < y) = P(X < y^2) = F_X(y^2).$$

Odtud

$$F_Y(y) = \begin{cases} 0, & \text{pro } y < 0; \\ y^2, & \text{pro } y \in \langle 0, 1\rangle; \\ 1, & \text{pro } y > 1; \end{cases}$$

a hustota $f_Y(y)$ náhodné veličiny Y má tvar

$$f_Y(y) = \begin{cases} 2y, & \text{pro } y \in \langle 0, 1\rangle, \\ 0, & \text{pro } y \notin \langle 0, 1\rangle. \end{cases}$$

Spočtěme střední hodnotu $\mathrm{E}\,Y$ a rozptyl $\mathrm{Var}\,Y$:

$$\mathrm{E}\,Y = \int_0^1 2y^2\,\mathrm{d}y = 2/3,$$

$$\mathrm{Var}\,Y = \int_0^1 2y^3\,\mathrm{d}y - (4/9) = 1/18.$$

Všimněme si, že $\mathrm{E}\,Y = I = \int_0^1 \sqrt{x}\,\mathrm{d}x$. Podle centrální limitní věty má průměr

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \sqrt{X_i} = \hat{I}$$

přibližně normální rozdělení se střední hodnotou $\frac{2}{3}$ a rozptylem rovným $\frac{1}{18n}$. Odtud

$$P\left(\left|\hat{I} - \frac{2}{3}\right| < 0.001\right) = \Phi\left(\frac{0.001}{\sqrt{\frac{1}{18}\frac{1}{10000}}}\right) - \Phi\left(-\frac{0.001}{\sqrt{\frac{1}{18}\frac{1}{10000}}}\right) \doteq 0.328.$$

Odpověď: Pravděpodobnost, že chyba výsledku při výpočtu integrálu $\int_0^1 \sqrt{x}\,\mathrm{d}x$ metodou Monte Carlo pomocí 10000 simulací bude menší než 0.001, je rovna přibližně 0.328.

Neřešené příklady

Příklad 5.4.

Házíme dokonale symetrickou kostkou, tj. kostkou, na které každá ze stran padá se stejnou pravděpodobností. S jakou pravděpodobností padne v 600 hodech více než 110 šestek?

Příklad 5.5.

Zdravotní stav pacientů trpících ekzémem se obvykle střídavě zlepšuje a zhoršuje. Kožní lékař, který provádí test nového léku, předepsal tento lék stu pacientům. Rozhodl se doporučit lék pro léčení ekzému, jestliže alespoň u 65% pacientů dojde ke zlepšení. S jakou pravděpodobností se zmýlí, tj. doporučí lék, i když je zkoušený lék bez žádoucího účinku?

Příklad 5.6.

Malé železné svorky používané v tavicí peci mají průměrnou hmotnost 6 g a směrodatnou odchylku 0.2 g. Byly vybrány náhodně dva vzorky, a to jeden o rozsahu $n = 100$ a druhý o rozsahu $n = 80$. Jaká je pravděpodobnost, že se průměry obou výběrů budou lišit o více než 0.02 g?

Příklad 5.7.

Podle tvrzení centrální limitní věty má součet většího počtu nezávislých stejně rozdělených náhodných veličin (s konečnou střední hodnotou a rozptylem) přibližně normální rozdělení. Této skutečnosti se využívá v jedné z metod pro simulaci náhodných čísel z normálního rozdělení. Metoda spočívá v tom, že se vygeneruje několik náhodných čísel z rovnoměrného rozdělení na intervalu $\langle 0, 1\rangle$ a ty se sečtou. Kolik čísel bychom měli sčítat, abychom bez normování směrodatnou odchylkou získali náhodná čísla z (přibližně) normálního rozdělení se směrodatnou odchylkou rovnou jedné?

6. Popisná statistika

Řešené příklady

Příklad 6.1.

Výsledky zkoušky z matematiky se dají shrnout do následující tabulky:

známka	1	2	3	4
počet studentů	5	37	85	23

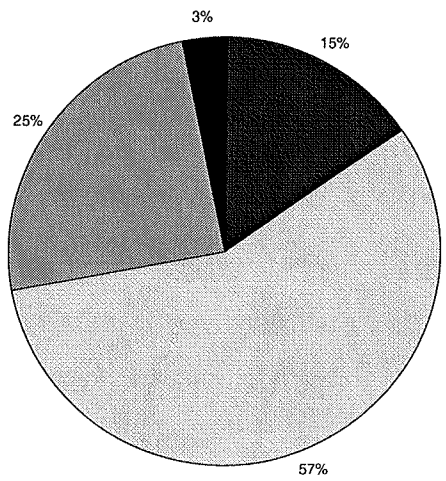
Spočtěte následující popisné statistiky : výběrový průměr, výběrový medián, výběrovou směrodatnou odchylku a výběrový koeficient šikmosti. Nakreslete koláčový graf (anglicky „pie chart“) a sloupcový graf (anglicky „bar chart“) pro zadaná data.

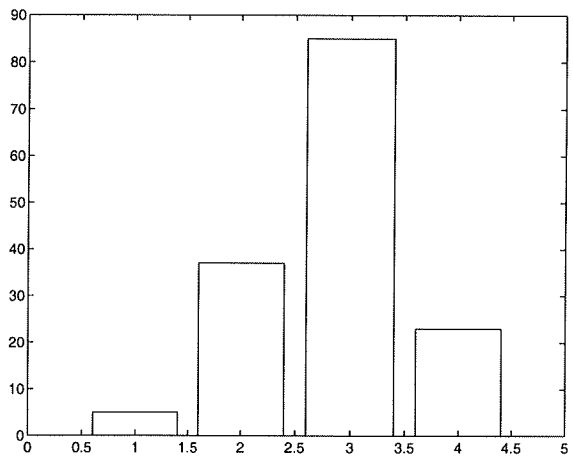
Řešení:

$$\begin{aligned}
&\text{Výběrový průměr} && \bar{x} = \frac{\sum \xi_i n_i}{n} && = 2.84\,,\\
&\text{výběrový medián} && \tilde{x} = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2} && = 3\,,\\
&\text{výběrová směrodatná odchylka} && \sigma_n = \sqrt{\frac{\sum (\xi_i - \bar{x})^2 n_i}{n}} && = 0.7126\,,\\
&\text{výběrový koeficient šikmosti} && A_3 = \frac{\frac{1}{n}\sum (\xi_i - \bar{x})^3 n_i}{\sigma_n^3} && = -0.3101\,.
\end{aligned}$$

Do shora uvedených vztahů jsme dosadili $\xi_1 = 1$, $\xi_2 = 2$, $\xi_3 = 3$, $\xi_4 = 4$ a $n_1 = 5$, $n_2 = 37$, $n_3 = 85$, $n_4 = 23$. Při výpočtu mediánu $x_{(n/2)}$ značí $n/2$ pořádkovou statistiku a $x_{((n/2)+1)}$ značí $(n/2)+1$ pořádkovou statistiku. Obě statistiky se zřejmě rovnají 3.

Na obrázcích vidíte koláčový graf a sloupcový graf pro data z našeho příkladu.





Příklad 6.2.

Z průměrných ročních průtoků v m^3/s Lužnice v Bechyni měřených v letech 1911–1988 byly spočteny tyto základní popisné charakteristiky:

výběrový průměr	$\bar{x} = 23.8675,$
výběrová směrodatná odchylka	$\sigma_n = 9.78344,$
výběrový koeficient šikmosti	$A_3 = 0.998701,$
výběrový koeficient špičatosti	$A_4 = 1.11259.$

Rozhodněte, zda je pro modelování průměrných ročních průtoků Lužnice vhodné normální rozdělení.

Řešení:

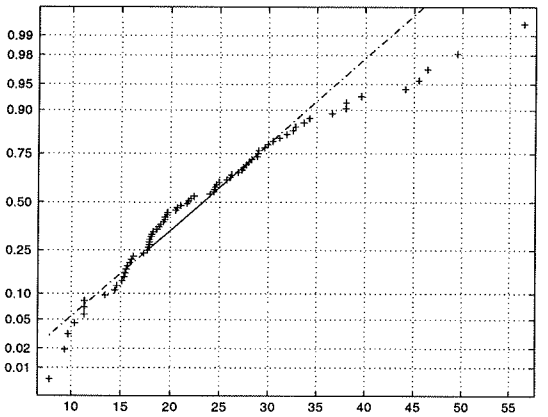
Pro zjišťování, zda určitý soubor dat je možné považovat za výběr z normálního rozdělení, se velmi často používá test založený na šikmosti a špičatosti. Koeficienty šikmosti i špičatosti normálního rozdělení jsou oba nulové. Pokud data pocházejí z normálního rozdělení, měl by být výběrový koeficient šikmosti a špičatosti malý, tj. tzv. standardizovaný koeficient šikmosti a standardizovaný koeficient špičatosti by měly být v absolutní hodnotě menší než 2. Pro velká n se standardizované výběrové koeficienty šikmosti a špičatosti počítají ze vztahů:

$$\alpha_3 = A_3\sqrt{n/6}, \quad \text{kde} \quad A_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{3/2}},$$

$$\alpha_4 = A_4\sqrt{n/24}, \quad \text{kde} \quad A_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{2}} - 3.$$

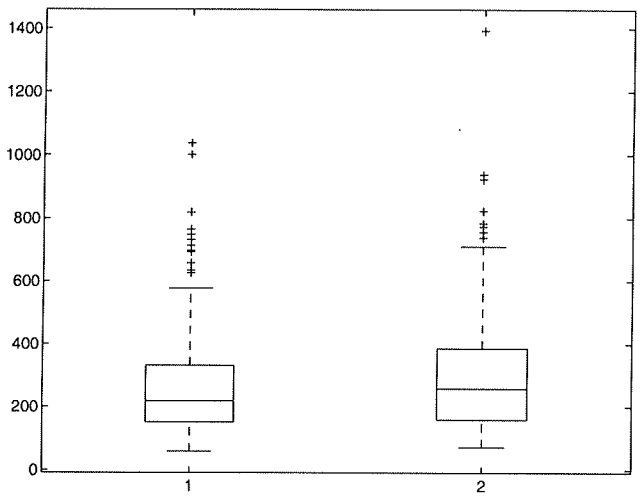
V našem případě $\alpha_3 \doteq 3.6$ a $\alpha_4 \doteq 2.0$. Odtud vyplývá, že normální rozdělení není příliš vhodný model.

K rychlému rozhodnutí, zda data mají nebo nemají normální rozdělení, se velmi často používá diagram normality (anglicky „normal plot“). Tento diagram zakresluje empirickou distribuční funkci daného náhodného výběru, přičemž nelineární dělení na svislé ose je voleno tak, aby pro data pocházející z normálního rozdělení byl tento graf přibližně lineární. Na obrázku je diagram normality pro roční průměrné průtoky Lužnice. Porovnáním znázorněných bodů s přímkou lze usuzovat na porušení předpokladu normality. Graf ukazuje zakřivení typické pro rozdělení s kladnou šikmostí.



Příklad 6.3.

Na obrázku vidíme graf, kterému se anglicky říká „box – whisker plot", pro dva soubory dat - prosincové a lednové průtoky v m^3/s Labe v letech 1851-1989, viz příklad 8.6. Porovnejte oba soubory.



Řešení:

Překlad anglického názvu grafu je „krabička s vousy". Horizontální úsečky v „krabičce" označují (odspodu) dolní kvartil, medián a horní kvartil. „Vousy" potom označují minimální a maximální pozorování. Případná odlehlá pozorování jsou zobrazena křížky. (Za odlehlé pozorování se zpravidla považuje číslo, ležící od okraje krabičky dále než je $1\frac{1}{2}$ násobek mezikvartilového rozpětí.)

Na první pohled vidíme, že rozložení obou souborů je vysoce asymetrické s kladnou šikmostí. Charakteristiky polohy (medián, dolní kvartil, horní kvartil) jsou mírně vyšší u druhého souboru. Také interkvartilové rozpětí, které slouží jako charakteristika rozptýlenosti, je pro druhý soubor větší. V druhém souboru vidíme jedno výrazně odlehlé pozorování dosahující téměř hodnoty 1400, což je řádově šestkrát větší hodnota než medián. Závěr tedy je, že lednové průtoky v letech 1851-1989 dosahovaly vyšších hodnot a více kolísaly než prosincové průtoky.

Příklad 6.4.

Tabulka na další stránce udává průměrné výnosy pšenice (ozimu) v kg z 10^4 m^2 v 20 oblastech Švédska $\{Y_i\}$, průměrné teploty vzduchu předchozí zimu (říjen-květen) v ^0C $\{X_{i1}\}$, průměrné teploty vzduchu za probíhající vegetační období (duben-září) v ^0C $\{X_{i2}\}$ a srážkový úhrn v mm ve vegetačním období ve třech meteorologických stanicích v oblasti $\{X_{i3}\}$. Data pocházejí z let 1913-1942. Spočtěte výběrovou korelační matici.

Řešení:

Výběrová korelační matice R spočtená z matice dat $U = \{(u_{i1}, u_{i2}, \ldots, u_{ik}), i = 1, \ldots, n\}$ je matice typu $k \times k$. Její prvek r_{mp} je výběrový korelační koeficient spočtený z $m-$tého a $p-$tého sloupce matice U :

$$r_{mp} = \frac{\sum_{i=1}^{n}(u_{im} - \bar{u}_m)(u_{ip} - \bar{u}_p)}{\sqrt{(\sum_{i=1}^{n}(u_{im} - \bar{u}_m)^2)(\sum_{i=1}^{n}(u_{ip} - \bar{u}_p)^2)}},$$

kde $\bar{u}_m = (1/n)\sum_{i=1}^{n} u_{im}$.

rok	úroda	teplota zima	teplota léto	srážky
1913	1990	2.7	12.8	230
1914	1950	3.1	13.7	268
1915	1630	1.9	12.0	188
1916	1720	1.3	11.7	315
1917	1560	1.0	12.7	180
1918	1680	1.6	12.0	261
1919	1980	2.3	12.2	216
1920	2180	1.7	12.8	346
1921	2370	3.1	13.1	131
1922	1790	1.1	11.8	256
1923	2400	1.6	11.2	327
1924	1410	0.1	11.8	320
1925	2570	3.7	13.2	382
1926	2180	1.1	12.5	279
1927	2150	2.5	12.2	351
1928	2530	0.8	10.5	324
1929	2100	0.8	10.9	196
1930	2330	3.6	12.4	381
1931	1850	1.6	10.7	273
1932	2230	1.9	12.5	289
1933	2510	2.2	11.9	338
1934	2600	3.0	13.5	267
1935	2480	3.2	12.3	372
1936	1940	2.8	12.3	357
1937	2770	2.1	13.5	358
1938	2570	3.3	12.9	202
1939	2510	3.8	13.4	311
1940	1420	-1.1	11.3	172
1941	810	-0.4	11.3	194
1942	1990	-2.4	11.2	261

Výběrová korelační matice je symetrická a na hlavní diagonále má jedničky.
Poznamenejme, že někdy nás zajímá výběrová kovarianční matice, jejíž prvky jsou výběrové kovariance. Výběrová kovarianční matice je symetrická a na její diagonále stojí výběrové rozptyly.

Odpověď: Výběrová korelační matice spočtená z dat o úrodě, zimní a letní teplotě a množství srážek má v našem případě tvar:

$$\begin{pmatrix} 1.0000 & 0.5934 & 0.4082 & 0.4600 \\ 0.5934 & 1.0000 & 0.6703 & 0.3184 \\ 0.4082 & 0.6703 & 1.0000 & 0.1072 \\ 0.4600 & 0.3184 & 0.1072 & 1.0000 \end{pmatrix}.$$

Neřešené příklady

Příklad 6.5.

Ve třídě jedné základní školy nemá žádného sourozence 8 dětí, 23 dětí má jednoho sourozence, 4 děti mají dva sourozence a jedno dítě má tři sourozence. Pro uvedená data načrtněte koláčový graf a sloupcový graf. Kolik je průměrný počet dětí v rodině žáků této třídy?

Příklad 6.6.

Při dopravním průzkumu byla sledována vytíženost vjezdu do určité křižovatky. Student, provádějící průzkum, si vždy při naskočení zeleného světla zapsal počet aut, čekajících ve frontě u semaforu. Jeho zapsané výsledky jsou:

6 3 1 5 3 2 3 5 7 1 2 8 8 1 6 1 8 5 5 8 5 4 7 2 5 6 3 4 2 8 4 4 5 5 4 3 3 4 9
6 2 1 5 2 3 5 3 5 7 2 5 8 2 4 2 4 3 5 6 4 6 9 3 2 1 2 6 3 5 3 5 3 7 6 3 7 5

Vytvořte tabulku skupinového rozdělení četností pro tato data, načrtněte sloupcový graf a vypočtěte následující výběrové statistiky: průměr, směrodatnou odchylku, medián, modus, rozpětí a šikmost.

Příklad 6.7.

V příkladu 8.6 jsou uvedeny prosincové a lednové průtoky Labe v Děčíně v letech 1851-1989. Porovnejte minimální a maximální prosincové a lednové průtoky. Porovnejte prosincové a lednové rozpětí.

Příklad 6.8.

Předpokládejme, že jsme pro n manželských párů zjistili výšku muže $\{X_i,\ i = 1,\dots,n\}$ a výšku ženy $\{Y_i,\ i = 1,\dots,n\}$. Ze zjištěných dat jsme spočítali výběrové průměry $\bar{x}$, $\bar{y}$, výběrové směrodatné odchylky ${}^x\sigma_n$, ${}^y\sigma_n$ a výběrový korelační koeficient r. Dokažte, že výběrový rozptyl ${}^z\sigma_n^2$ rozdílů výšek $\{Z_i = X_i - Y_i,\ i = 1,\dots,n\}$ splňuje:

$$ {}^z\sigma_n^2 = {}^x\sigma_n^2 + {}^y\sigma_n^2 - 2\,{}^x\sigma_n\,{}^y\sigma_n\,r. $$

Příklad 6.9.

Dokažte, že se při lineární transformaci dat nezmění absolutní hodnota korelačního koeficientu.

Příklad 6.10.

Následující tabulka udává skupinové rozdělení četností pro průměrné roční průtoky Lužnice v Bechyni měřené v letech 1911 – 1988. Spočtěte základní popisné statistiky, tj. výběrový průměr, výběrovou směrodatnou odchylku, výběrový koeficient šikmosti a špičatosti. Porovnejte charakteristiky spočtené na základě skupinového rozdělení četností s charakteristikami spočtenými z naměřených údajů v příkladě 6.2. Načrtněte histogram.

interval	6-12	12-18	18-24	24-30	30-36	36-42	42-48	48-54	54-60
četnost	7	17	18	20	7	4	3	1	1

Příklad 6.11.

Následující tabulka udává průměrné roční teploty v St. Peterburgu v letech 1860–1879 a 1960–1979. Spočtěte pro oba soubory základní popisné statistiky: průměr, medián, směrodatnou odchylku, rozpětí, interkvartilové rozpětí.

rok	teplota	rok	teplota
1860	4.9	1960	4.7
1861	4.8	1961	6.3
1862	2.3	1962	4.8
1863	6.9	1963	4.1
1864	4.7	1964	5.0
1865	4.7	1965	4.3
1866	5.5	1966	3.2
1867	3.3	1967	5.3
1868	4.9	1968	4.0
1869	6.0	1969	3.6
1870	4.0	1970	4.7
1871	3.3	1971	4.7
1872	6.0	1972	6.2
1873	5.1	1973	5.2
1874	5.4	1974	6.3
1875	2.5	1975	6.6
1876	4.0	1976	3.2
1877	4.0	1977	4.8
1878	5.9	1978	3.5
1879	4.8	1979	5.0

Příklad 6.12.

Najděte výběrovou šikmost a špičatost datových souborů z předchozího příkladu. Najděte také standardizovanou šikmost a špičatost.

7. Teorie odhadu

Řešené příklady

Příklad 7.1.

Při sledování doby do poruchy (v hodinách) určitého zařízení bylo získáno následujících osm údajů: 48, 16, 75, 29, 96, 67, 89, 22. Předpokládejme, že se jedná o výběr z exponenciálního rozdělení. Odhadněte střední (průměrnou) dobu životnosti zařízení. Odhadněte pravděpodobnost, že zařízení bude fungovat ještě po 100 hodinách.

Řešení:

Střední hodnota náhodné veličiny, která se řídí exponenciálním rozdělením s parametrem δ, je rovna právě tomuto parametru. Nejlepší nestranný odhad parametru δ se v tomto případě rovná maximálně věrohodnému odhadu a je roven výběrovému průměru $\hat{\delta} = \sum x_i/n = \bar{x} = 442/8 = 55.25$. Pravděpodobnost, že náhodná veličina X, která má exponenciální rozdělení s parametrem δ, přežije dobu t, je rovna $P(X > t) = \mathrm{e}^{-t/\delta}$ a je možné ji odhadnout pomocí $\mathrm{e}^{-t/\hat{\delta}}$. V našem případě hodnotou $\mathrm{e}^{-100/55.25} \doteq 0.164$.

Odpověď: Na základě zjištěných údajů odhadujeme, že střední doba životnosti zařízení je rovna 55.25 hodin a pravděpodobnost přežití 100 hodin je 0.164.

Příklad 7.2.

Dva studenti geodézie by rádi znali vzdálenost dvou bodů. Aby zmírnili vliv náhodných chyb měření, rozhodli se měření opakovat. Při měření používali dva stejně přesné měřící přístroje, z nichž ani jeden neměl systematickou chybu. Pavel provedl 10 měření, zatímco Petr provedl 15 měření. Oba dva ze svých měření spočítali aritmetický průměr a prohlásili ho za odhad skutečné vzdálenosti bodů. Když se opět sešli, nemohli se dohodnout, zda za konečný odhad mají prohlásit aritmetický průměr jejich dvou odhadů či zda mají raději vzít původní data a jako odhad uvažovat průměr ze všech 25 měření. Který postup je lepší a proč?

Řešení:

Označme $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$ Pavlův odhad a $\bar{Y} = \frac{1}{15}\sum_{j=1}^{15} Y_j$ Petrův odhad, kde $\{X_i, i = 1, \ldots, 10\}$ jsou Petrova měření a $\{Y_j, j = 1, \ldots, 15\}$ jsou Pavlova měření. Označme Z_1 první způsob odhadu:

$$Z_1 = \frac{\bar{X} + \bar{Y}}{2}$$

a Z_2 druhý způsob odhadu:

$$Z_2 = \frac{\sum_{i=1}^{10} X_i + \sum_{j=1}^{15} Y_j}{25}.$$

Nejprve připomeňme, že pro střední hodnotu a rozptyl lineární kombinace dvou nezávislých veličin platí $\mathrm{E}(aU + bV) = a\,\mathrm{E}U + b\,\mathrm{E}V$ a $\mathrm{Var}(aU + bV) = a^2\,\mathrm{Var}\,U + b^2\,\mathrm{Var}\,V$. Podobné vztahy se dají odvodit i pro více náhodných veličin. Odtud plyne, že průměr n náhodných veličin, které mají stejnou střední hodnotu μ a stejný rozptyl σ^2, má střední hodnotu μ a rozptyl σ^2/n.

Oba odhady Z_1 i Z_2 jsou nestranné. Lepší bude ten odhad, který bude mít menší rozptyl. Rozptyl prvního odhadu:

$$\operatorname{Var} Z_1 = \frac{\operatorname{Var}\bar{X} + \operatorname{Var}\bar{Y}}{4} = \frac{\frac{\sigma^2}{10} + \frac{\sigma^2}{15}}{4} = \frac{\sigma^2}{24}.$$

Rozptyl druhého odhadu:

$$\operatorname{Var} Z_2 = \frac{\sigma^2}{25}.$$

Všimněme si, že druhý odhad je vlastně vážený průměr s vahami 10/25 a 15/25, tj.

$$Z_2 = \frac{10}{25}\bar{X} + \frac{15}{25}\bar{Y}.$$

Odpověď: Oba odhady jsou nestranné. Menší rozptyl má odhad Z_2 (vážený průměr) a je proto lepším odhadem.

Příklad 7.3.

Agentura provádějící průzkum veřejného mínění plánuje šetření, na základě kterého chce odhadnout, kolik procent voličů podporuje současnou vládní koalici. Předpokládejme (v praxi tomu tak ovšem není), že jsou dotazovaní vybíráni zcela náhodně. Kolik dotazovaných by mělo být do výběru zařazeno, jestliže si vedení agentury přeje, aby se odhad získaný z výběru lišil od skutečného podílu příznivců koalice v celé populaci o méně než 3 %?

Řešení:

Pokud by pracovníci agentury chtěli mít svůj požadavek splněn stoprocentně, pak by se museli ptát všech voličů. Jakmile agentura vybere náhodně jen některé občany, může se jí (i když s malou pravděpodobností) stát, že se do výběru dostanou například jen příznivci koalice nebo naopak jen příznivci opozice a podobně. Znamená to, že požadavek, aby se odhad lišil od správné hodnoty o méně než 3 %, může být splněn jen s určitou spolehlivostí, například 90 %, 95 % nebo 99 %.

Počet příznivců koalice zařazených do výběru má přibližně binomické rozdělení s parametry n a p, kde parametr n odpovídá rozsahu výběru a parametr p podílu příznivců koalice mezi všemi voliči. $(1-\alpha)\,100\%$ interval spolehlivosti pro parametr p alternativního rozdělení má pro velká n tvar:

$$\hat{p} - u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad < \quad p \quad < \quad \hat{p} + u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

kde $\hat{p}$ je podíl příznivců koalice ve výběru. Odtud plyne, že

$$u_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.03.$$

Vzhledem k tomu, že $\hat{p}$ neznáme, uvažujme nejhorší možný případ, tj. případ, kdy interval bude nejdelší. Interval bude nejdelší pro $\hat{p} = 1/2$. Požadujeme-li například spolehlivost 95 %, pro kterou $u_{0.025} = 1.96$, pak musí rozsah výběru splňovat

$$1.96 \times \frac{1}{2} \times \frac{1}{\sqrt{n}} = 0.03 \quad \Longrightarrow \quad n = 1067.$$

Odpověď: Jestliže bude do náhodného výběru zařazeno 1067 osob, pak má agentura zajištěno se spolehlivostí minimálně rovnou 95 %, že se jejich odhad bude lišit o méně než 3 % od skutečného podílu příznivců koalice mezi všemi voliči.

Příklad 7.4.

500 naměřených hodnot bylo roztříděno do 11 intervalů délky 0.4:

⟨7.4-7.8)	⟨7.8-8.2)	⟨8.2-8.6)	⟨8.6-9.0)	⟨9.0-9.4)	⟨9.4-9.8)
5	21	43	71	103	111

⟨9.8-10.2)	⟨10.2-10.6)	⟨10.6-11)	⟨11-11.4)	⟨11.4-11.8)
83	39	15	8	1

Najděte 95% interval spolehlivosti pro střední hodnotu μ, předpokládáme-li, že data jsou realizací výběru z normálního rozdělení.

Řešení:

95% interval spolehlivosti pro střední hodnotu μ normálně rozdělené náhodné veličiny X má tvar:

$$\left(\bar{x} - t_{0.025}[n-1]\frac{s}{\sqrt{n}}, \bar{x} + t_{0.025}[n-1]\frac{s}{\sqrt{n}}\right),$$

kde $n = 500$ je počet naměřených dat, $\bar{x}$ je aritmetický průměr těchto dat, tj. $\bar{x} = \frac{\sum x_i}{n}$, a s je odhad směrodatné odchylky veličiny X, tj. $s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$. K výpočtu průměru a směrodatné odchylky však nemůžeme použít původní data, protože je bohužel neznáme. Můžeme je však spočítat alespoň přibližně, jestliže budeme předpokládat, že všechna data v daném intervalu nabývají prostřední hodnoty (třídního znaku). Jestliže intervaly nejsou příliš dlouhé, nepřesnost, které se dopustíme, nebude příliš velká. Při tomto zjednodušení $\bar{x} = 9.408$ a $s = 0.7267$.

Odpověď: Vzhledem k tomu, že 2.5 % horní kvantil t–rozdělení se přibližně rovná 1.96, bude mít hledaný 95 % interval spolehlivosti tvar: (9.344, 9.472).

Příklad 7.5.

Z 90 zkoušek meze kluzu konstrukční oceli z produkce určité ocelárny byl vypočten výběrový průměr $\bar{x} = 251.3384$ MPa a výběrový rozptyl $\sigma_n^2 = 319.4818$. Najděte 80% intervaly spolehlivosti pro střední hodnotu a směrodatnou odchylku meze kluzu (za předpokladu normality dat).

Řešení:

80% interval spolehlivosti pro střední hodnotu meze kluzu je dán vzorcem (viz předchozí příklad):

$$\left(\bar{x} - t_{0.1}[n-1]\frac{s}{\sqrt{n}}, \bar{x} + t_{0.1}[n-1]\frac{s}{\sqrt{n}}\right).$$

Po dosazení $n = 90$, $s = \sigma_{n-1} = \sqrt{\frac{n}{n-1}} \cdot \sigma_n = 17.9742$ a $t_{0.1}[89] = 1.2911$ dostaneme interval

$$(248.89, 253.78).$$

$100(1-\alpha)\%$ interval spolehlivosti pro parametr σ^2 normálního rozdělení (tj. pro rozptyl) má tvar

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}[n-1]}, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}[n-1]}\right).$$

Zde $\chi^2_{\frac{\alpha}{2}}[n-1]$ je jako obvykle horní $100\frac{\alpha}{2}\%$ kvantil χ^2 rozdělení s $(n-1)$ stupni volnosti. $\chi^2_{1-\frac{\alpha}{2}}[n-1]$ znamená též horní $100(1-\frac{\alpha}{2})\%$ kvantil a tedy současně dolní $100\frac{\alpha}{2}\%$ kvantil.

Po dosazení $s^2 = 323.0715$, $\chi^2_{0.9}[89] = 72.387$ a $\chi^2_{0.1}[89] = 106.469$ dostaneme interval pro σ^2:

$$(306.41, 379.22),$$

odkud získáme interval pro směrodatnou odchylku σ odmocněním obou mezí:

$$(16.43, 19.93).$$

Poznamenejme, že ve statistických tabulkách bývají kvantily χ^2-rozdělení uvedeny jen pro omezený počet stupňů volnosti. Je-li počet stupňů volnosti ν dostatečně velký, můžeme χ^2-rozdělení přibližně nahradit normálním rozdělením se střední hodnotou ν a rozptylem 2ν. Platí tedy

$$\chi^2_p[\nu] \doteq \nu + \sqrt{2\nu} \cdot u_p,$$

kde u_p je horní $100p\%$ kvantil normovaného normálního rozdělení. Při této aproximaci dostaneme přibližné vyjádření pro $100(1-\alpha)\%$ interval spolehlivosti pro parametr σ^2 normálního rozdělení:

$$\left(\frac{\sqrt{n-1}\, s^2}{\sqrt{n-1} + \sqrt{2}\, u_{\alpha/2}}, \frac{\sqrt{n-1}\, s^2}{\sqrt{n-1} - \sqrt{2}\, u_{\alpha/2}}\right).$$

V našem případě bychom po dosazení za $u_{\alpha/2} = u_{0.1} = 1.2816$ a po odmocnění mezí dostali přibližný interval pro směrodatnou odchylku:

$$(16.46, 20.00).$$

Na závěr uveďme, že zejména v případě rozptylu, resp. směrodatné odchylky, často hledáme jednostranný interval spolehlivosti, jehož levá mez je nulová. Chceme totiž najít horní mez pro rozptyl (směrodatnou odchylku), která není s danou pravděpodobností překročena. Takový $100(1-\alpha)\%$ interval spolehlivosti má pro σ^2 tvar:

$$\left(0, \frac{(n-1)s^2}{\chi^2_{1-\alpha}[n-1]}\right),$$

po dosazení

$$\text{pro } \sigma^2\text{:} \quad (0, 370.43) \quad \Longrightarrow \quad \text{pro } \sigma\text{:} \quad (0, 19.25).$$

Odpověď: 80% interval spolehlivosti pro střední hodnotu meze kluzu je přibližně $(248.9, 253.8)$, 80% interval pro směrodatnou odchylku je $(16.43, 19.93)$, případně jednostranný 80% interval pro směrodatnou odchylku je $(0, 19.25)$.

Příklad 7.6.

V jisté americké elektrikářské firmě se cena C za dojednanou práci počítá následovně:

$$C = 10 \times 16 \times y = 160 \times y,$$

kde y je počet dní potřebných pro vykonání práce. Předpokládá se totiž, že elektrikář pracuje 10 hodin denně a cena za hodinu práce je 16 \$. Je známo, že 99% interval spolehlivosti pro střední počet dní potřebných pro vykonání práce je 14.1 ∓ 0.375. Odůvodněte, proč 99% interval spolehlivosti pro střední cenu elektrikářské práce je 2256 ∓ 60 \$.

Řešení:

99% interval spolehlivosti pro střední hodnotu náhodné veličiny X, mající normální rozdělení, má tvar:

$$(\bar{x} - t_{0.005}[n-1]\frac{s_x}{\sqrt{n}}, \bar{x} + t_{0.005}[n-1]\frac{s_x}{\sqrt{n}}),$$

kde n je počet naměřených dat, $\bar{x}$ je aritmetický průměr těchto dat, tj. $\bar{x} = \frac{\sum x_i}{n}$, a s_x je odhad směrodatné odchylky veličiny X, tj. $s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$. Obdobně 99% interval spolehlivosti pro střední hodnotu veličiny $Y = k\,X$, kde k je nějaká konstanta, má tvar:

$$(\bar{y} - t_{0.005}[n-1]\frac{s_y}{\sqrt{n}}, \bar{y} + t_{0.005}[n-1]\frac{s_y}{\sqrt{n}}),$$

kde $\bar{y} = \frac{\sum y_i}{n}$ a $s_y = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$, přičemž $y_i = k\,x_i$.

Odpověď: Vzhledem k tomu, že $\bar{y} = k\,\bar{x}$ a $s_y = k\,s_x$, stačí pro výpočet intervalu spolehlivosti pro střední hodnotu veličiny $Y = k\,X$ vynásobit obě krajní meze intervalu spolehlivosti konstantou k. V našem případě se konstanta k rovná 160.

Příklad 7.7.

V tabulce jsou dány průměry a odhady směrodatných odchylek měsíčních srážkových úhrnů v mm měřených v Břevnově, spočtené z údajů z let 1921 – 1989.

	průměr	směr.odch.
leden	26.7246	12.9579
únor	25.4493	12.5634
březen	30.4058	15.8201
duben	38.3768	22.1141
květen	63.5507	31.8971
červen	72.2029	40.7234
červenec	76.2609	42.7832
srpen	73.3913	42.3189
září	42.4493	28.0711
říjen	39.2319	29.7416
listopad	30.0870	14.7596
prosinec	27.7971	16.5410

Najděte 90 % predikční intervaly pro měsíční srážkové úhrny v příštím roce. (Předpokládejte, že měsíční srážkové úhrny mají normální rozdělení.)

Řešení:

90 % predikční interval má tvar:

$$\left(\bar{x} - t_{0.05}[n-1]\ s\ \sqrt{1+\frac{1}{n}}, \bar{x} + t_{0.05}[n-1]\ s\ \sqrt{1+\frac{1}{n}}\right).$$

Pro lednovou hodnotu je například 90 % predikční interval: (4.9603,48.4889).

Neřešené příklady

Příklad 7.8.

Geigrův–Müllerův přístroj zaznamenával 5 hodin a 26 minut v intervalech délky 7.5 sekund počet vyzářených α-částic. Následující tabulka uvádí četnosti intervalů, ve kterých zaznamenal přístroj určitý počet α-částic. Předpokládejme, že počet α-částic vyzářených v 7.5 sekundových intervalech se řídí Poissonovým rozdělením. Odhadněte parametr λ.

počet částic	0	1	2	3	4	5	6	7	8	9	10	11	12
počet intervalů	57	203	383	525	532	408	273	139	45	27	10	4	2

Příklad 7.9.

V Čechách žije N_1 mužů a N_2 žen. Chceme odhadnout, kolik v průměru měsíčně spotřebují šampónu. Za tímto účelem vybereme n osob, z toho n_1 mužů a n_2 žen, a zeptáme se jich na jejich měsíční spotřebu šampónu. Získáme údaje o spotřebě vybraných mužů: $X_1, \ldots, X_{n_1}$ a údaje o spotřebě vybraných žen: $Y_1, \ldots, Y_{n_2}$. Předpokládejme, že spotřeby mužů jsou náhodné veličiny se střední hodnotou μ_1 a rozptylem σ^2 a spotřeby žen jsou náhodné veličiny se střední hodnotou μ_2 a rozptylem σ^2. Střední (průměrnou) spotřebu šampónu všech obyvatel, tj. hodnotu $N_1\,\mu_1 + N_2\,\mu_2$, odhadneme jako $N_1\,\bar{X} + N_2\,\bar{Y}$, kde $\bar{X} = (1/n_1)\sum_{i=1}^{n_1} X_i$ a $\bar{Y} = (1/n_2)\sum_{j=1}^{n_2} Y_j$. Dokažte, že tento odhad je optimální, jestliže $n_1 : n_2 = N_1 : N_2$.

Příklad 7.10.

Opakovaná měření rychlosti vody v potrubí dala výsledky v m/s: 4.22, 4.24, 4.27, 4.22, 4.23. Najděte 95 % interval spolehlivosti pro střední rychlost vody v potrubí za předpokladu, že naměřené hodnoty lze považovat za výběr z normálního rozdělení.

Příklad 7.11.

U 100 vzorků zeminy byl zjištěn obsah určité látky v mg. Data byla roztříděna do 14 intervalů délky 5 mg:

⟨15-20)	⟨20-25)	⟨25-30)	⟨30-35)	⟨35-40)	⟨40-45)	⟨45-50)
1	4	4	6	7	11	21

⟨50-55)	⟨55-60)	⟨60-65)	⟨65-70)	⟨70-75)	⟨75-80)	⟨80-85)
14	8	10	8	1	4	1

Předpokládejte, že data jsou náhodným výběrem z normálního rozdělení.

a) Najděte nestranné odhady střední hodnoty a rozptylu obsahu této látky v zemině.

b) Najděte 95% intervalový odhad pro tuto střední hodnotu.

c) Odhadněte, s jakou pravděpodobností by obsah této látky v dalším odebraném vzorku překročil hodnotu 72 mg.

Příklad 7.12.

Při zjištění obzvlášť důležité charakteristiky bylo provedeno 100 měření této charakteristiky a stanoven 99 % interval spolehlivosti. Kolik by stačilo udělat měření, jestliže by se podařilo zlepšit přesnost měření tak, aby směrodatná odchylka chyby měření klesla na polovinu původní hodnoty, jestliže si přejeme, aby délka 99 % intervalu spolehlivosti zůstala stejná?

Příklad 7.13.

Následující tabulka udává počet živě narozených dětí v letech 1980 – 1995 v České republice. Za předpokladu, že se pravděpodobnost narození chlapce nemění, najděte pro ni 99 % interval spolehlivosti.

rok	chlapci	děvčata
1980	79409	74392
1981	74063	70375
1982	72579	69159
1983	70719	66712
1984	70253	66688
1985	69662	66219
1986	68539	64817
1987	67305	63616

rok	chlapci	děvčata
1988	67830	64837
1989	65669	62687
1990	66970	63594
1991	66644	62710
1992	62701	59004
1993	62115	58910
1994	54704	51875
1995	49405	46692

Příklad 7.14.

Pro sto novomanželských párů v České republice byla zjišťována výška manžela i manželky. Průměr výšky muže byl 183.2 cm a směrodatná odchylka 4.4 cm, zatímco průměr výšky ženy 171.4 cm a směrodatná odchylka 3.5 cm. Výběrový korelační koeficient r se rovnal 0.38. Spočtěte 95% interval spolehlivosti pro střední (průměrný) rozdíl výšek novomanželů v České republice. Najděte 95% predikční interval pro rozdíl výšek novomanželů, kteří se mají nastěhovat do sousedního uvolněného bytu. Předpokládejte, že výška muže i ženy má přibližně normální rozdělení. (Použijte tvrzení z příkladu 6.8.)

8. Testování hypotéz

Řešené příklady

Příklad 8.1.

Hráč hry „Člověče, nezlob se“ by rád zjistil, zda pravděpodobnost padnutí šestky na jeho kostce je opravdu rovna 1/6. Rozhodl se, že kostkou hodí 600 krát. Jestliže se počet šestek n_6 bude pohybovat v mezích $82 \leq n_6 \leq 118$, bude muset připustit, že se o kostce nedá říci, že pravděpodobnost padnutí šestky není rovna 1/6. Jakmile však počet šestek bude ležet vně tohoto rozmezí, raději kostku ze hry vyřadí.

a) Pokud jeho postup chápeme jako test, jakou má hladinu významnosti?

b) Pokud pravděpodobnost padnutí šestky je pouze 0.15, s jakou pravděpodobností zjistí svým postupem nesrovnalost?

Řešení:

Hráč vlastně testuje nulovou hypotézu $H_0 : p = 1/6$ proti alternativě $A : p \neq 1/6$. Hladina významnosti testu odpovídá pravděpodobnosti, s jakou je zamítána nulová hypotéza, i když je pravdivá. V našem případě to znamená určit, jaká je pravděpodobnost, že na kostce, na které padá šestka s pravděpodobností 1/6, nebude počet šestek n_6 mezi 82 až 118. Počet šestek má zde binomické rozdělení s parametry $n = 600$ a $p = 1/6$, které můžeme podle centrální limitní věty aproximovat normálním rozdělením s parametry $\mu = np = 100$ a $\sigma^2 = np(1-p) = 600 \times (1/6) \times (5/6) \doteq 83.3333$. Odtud (s použitím korekce)

$$1 - P(82 \leq n_6 \leq 118) = 1 - \left(\Phi\left(\frac{118 + 0.5 - 100)}{\sqrt{83.3333}} \right) - \Phi\left(\frac{82 - 0.5 - 100)}{\sqrt{83.3333}} \right) \right) \doteq 0.0427.$$

Pokud je pravděpodobnost padnutí šestky 0.15, pak má počet šestek n_6 binomické rozdělení s parametry $n = 600$ a $p = 0.15$, které můžeme opět aproximovat normálním se střední hodnotou $\mu = np = 600 \times 0.15 = 90$ a rozptylem $\sigma^2 = np(1-p) = 600 \times \times 0.15 \times 0.85 = 76.5$. Odtud

$$1 - P(82 \leq n_6 \leq 118) = 1 - \left(\Phi\left(\frac{118 + 0.5 - 90)}{\sqrt{76.5}} \right) - \Phi\left(\frac{82 - 0.5 - 90)}{\sqrt{76.5}} \right) \right) \doteq 0.1661.$$

Odpověď: a) Hladina významnosti, která odpovídá hráčovu rozhodnutí, je rovna 4.3%. b) Pokud ve skutečnosti na hráčově kostce padá šestka s pravděpodobností 0.15, je vyřazena shora uvedeným rozhodovacím pravidlem s pravděpodobností 0.166.

Příklad 8.2.

Pro kontrolu správnosti nastavení měřícího přístroje bylo provedeno 10 měření zkušebního etalonu se správnou hodnotou $\mu_0 = 15.20$. Byly získány tyto výsledky: 15.23, 15.21, 15.19, 15.16, 15.26, 15.22, 15.23, 15.26, 15.23, 15.29. Lze považovat pozorované odchylky od správné hodnoty za náhodné chyby nebo je důvod k podezření na přítomnosti systematické chyby? (Předpokládáme, že jde o měření, při kterém náhodné chyby mají normální rozdělení.)

Řešení:

Jelikož je žádoucí odhalit jak zápornou, tak kladnou systematickou chybu, použijeme test hypotézy $H : \mu = \mu_0$ proti $A : \mu \neq \mu_0$. Spočtěme $\bar{x} = 15.228$ a $s = 0.03706$. Dosaďme do levé strany zamítacího pravidla

$$\frac{|\bar{x} - \mu_0|\sqrt{n}}{s} = 2.3893.$$

Zvolíme-li hladinu významnosti $\alpha = 0.05$, pak levá strana zamítacího pravidla je větší než 2.5 % horní kvantil $t-$rozdělení o 9 stupních volnosti $t_{0.025}[9] = 2.2622$.

Odpověď: Nulovou hypotézu $H_0 : \mu = \mu_0$ zamítáme, tj. usuzujeme na přítomnost systematické chyby.

Příklad 8.3.

Úkolem zkoušky je zjistit, jaká doba míchání stačí, aby směs byla dostatečně homogenní. Homogenitu lze charakterizovat rozptylem hodnot koncentrace určité látky ve směsi, které zjistíme v různých místech zásobníku. Předpokládá se, že koncentrace je náhodná veličina s rozdělením $N(\mu, \sigma^2)$ a dostatečnou homogenitu látky charakterizuje rozptyl $\sigma_0^2 = 0.003$ nebo menší. Chceme se dozvědět, zda míchání po dobu deseti minut zajistí postačující homogenitu. Po deseti minutách míchání bylo odebráno z různých míst nádoby $n = 10$ vzorků směsi a zjištěna v nich koncentrace látky. Z takto provedeného náhodného výběru byl vypočten výběrový rozptyl $\sigma_n^2 = 0.0022$. Jaký je možno na základě pokusu udělat závěr?

Řešení:

V rámci teorie testování hypotéz můžeme řešit problém tak, že testujeme nulovou hypotézu $H_0 : \sigma^2 \leq \sigma_0^2 = 0.003$ proti alternativě $A : \sigma^2 > \sigma_0^2 = 0.003$. Zvolíme-li hladinu významnosti $\alpha = 0.05$, má zamítací pravidlo tvar:

$$\frac{(n-1)\,s^2}{\sigma_0^2} > \chi^2_{0.05}[n-1],$$

kde $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$ a $\chi^2_{0.05}[n-1]$ je horní 5 % horní kvantil χ^2-rozdělení o $n-1$ stupních volnosti. Vzhledem k tomu, že platí $n\,\sigma_n^2 = (n-1)\,s^2 = \sum(x_i - \bar{x})^2$, nulovou hypotézu na 5 % hladině významnosti nezamítáme, neboť:

$$\frac{n\,\sigma_n^2}{\sigma_0^2} \doteq 7.333 < \chi^2_{0.05}[9] \doteq 16.919.$$

Velmi vysoká $p-$hodnota rovnající se 0.602 posiluje naši důvěru v platnost nulové hypotézy.

Odpověď: Z výsledku pokusu se zdá, že deset minut míchání je postačující doba k tomu, aby směs byla dobře promíchána.

Příklad 8.4.

Do laboratoře přivezli nové váhy. Laborant se obával, že váhy mohou mít systematickou chybu. Rozhodl se tedy, že jakmile bude průměrný rozdíl mezi naměřenou

hodnotou a skutečnou hmotností pro 10 závaží (jejichž správnou hmotnost znal) v absolutní hodnotě větší než 0.05 g, půjde váhy reklamovat jako vadné. Jeho měření dopadla následovně:

správná hodnota	2.0	2.0	2.0	3.0	3.0	3.0	4.0	4.0	5.0	5.0
naměřená hodnota	2.2	2.0	2.0	3.1	3.1	2.8	4.0	4.1	5.1	4.9

Jestliže bereme jeho rozhodnutí jako zamítací pravidlo pro testování hypotéz, jak vysoké hladině významnosti přibližně odpovídá? Bylo jeho rozhodnutí příliš shovívavé vůči výrobci či příliš přísné? V jakém případě byste šli váhy reklamovat vy a proč?

Řešení:

Jestliže předpokládáme, že přesnost vah se v rozmezí $2-5$ g nemění a že se v tomtéž rozmezí nebude měnit ani případná systematická chyba, pak rozdíl mezi naměřenou a správnou hodnotou je náhodná veličina X s normálním rozdělením $N(\mu, \sigma^2)$. Problém odhalení případné systematické chyby můžeme v rámci matematické statistiky řešit testováním nulové hypotézy $H_0 : \mu = 0$ proti oboustranné alternativě $A : \mu \neq 0$. Pravidlo pro zamítnutí nulové hypotézy na hladině α má tvar:

$$\frac{|\bar{x}|\sqrt{n}}{s} > t_{\alpha/2}[n-1],$$

kde $t_{\alpha/2}[n-1]$ označuje $100\,\alpha/2\,\%$ horní kvantil $t-$rozdělení o $n-1$ stupních volnosti. Spočtěme rozdíly mezi naměřenými a skutečnými hmotnostmi závaží: 0.2, 0.0, 0.0, 0.1, 0.1, -0.2, 0.0, 0.1, 0.1, -0.1. Průměr spočtený z těchto hodnot $\bar{x} = 0.03$ a odhad směrodatné odchylky $s = 0.11595$. Pokud bychom volili, jak je nejčastěji zvykem, hladinu významnosti $\alpha = 0.05$, pak bychom zamítli nulovou hypotézu, jestliže

$$|\bar{x}| > \frac{s\ t_{0.025}[9]}{\sqrt{n}} = 0.08295,$$

neboť horní 2.5% kvantil t-rozdělení o 9 stupních volnosti je roven 2.262.

Pokud chceme najít hladinu významnosti, která odpovídá laborantovu rozhodnutí, musíme najít takový horní kvantil $t_{\alpha/2}[9]$, aby platilo

$$\frac{s\ t_{\alpha/2}[9]}{\sqrt{n}} = 0.05.$$

Toto je splněno velmi přibližně pro $t_{0.1}[9]$.

Odpověď: Laborantovo rozhodnutí přibližně odpovídá 20% hladině významnosti. Laborantovo rozhodnutí je výrazně přísnější než rozhodnutí založené na testování hypotéz s 5% hladinou významnosti.

Příklad 8.5.

Byla navržena studie k ověření předpokladu, že muži mají v průměru vyšší diastolický tlak než ženy. U náhodně vybraných mužů a žen byly naměřeny následující hodnoty diastolického tlaku:

Muži: 76, 76, 74, 70, 80, 68, 90, 70, 90, 72, 76, 80, 68, 72, 96, 80.

Ženy: 76, 70, 82, 90, 68, 60, 62, 68, 80, 74, 60, 62, 72.

Jaký byl výsledek této studie?

Řešení:

V rámci matematické statistiky je možné výsledek studie zhodnotit pomocí metody testování hypotéz. Předpokládáme-li, že hodnoty diastolického tlaku mužů jsou výběrem z normálního rozdělení $N(\mu_M, \sigma_M^2)$ a žen výběrem z $N(\mu_Z, \sigma_Z^2)$, jedná se o testování nulové hypotézy, která tvrdí, že střední (průměrná) hodnota diastolického tlaku je pro obě pohlaví stejná, tj. $H_0 : \mu_M = \mu_Z$, proti alternativě, že je střední (průměrná) hodnota diastolického tlaku mužů vyšší než žen, tj. $A : \mu_M > \mu_Z$.

Vzhledem k tomu, že nevíme, zda je variabilita diastolického tlaku u obou pohlaví stejná, je třeba ještě pomocně otestovat nulovou hypotézu $H_0^p : \sigma_M^2 = \sigma_Z^2$ proti alternativě $A^p : \sigma_M^2 \neq \sigma_Z^2$. Pro oba problémy zvolme hladinu významnosti $\alpha = 0.05$.

Abychom mohli dosadit do jednotlivých zamítacích pravidel, spočtěme odhady středních hodnot a rozptylů na základě naměřených dat $\widehat{\mu_M} = \bar{x}_M \doteq 77.375$, $\widehat{\sigma_M^2} = s_M^2 \doteq \doteq 69.7167$ a $\widehat{\mu_Z} = \bar{x}_Z \doteq 71.0769$, $\widehat{\sigma_Z^2} = s_Z^2 \doteq 85.0769$. Nejprve dosaďme do zamítacího pravidla pro testování shodnosti rozptylů:

$$\frac{s_M^2}{s_Z^2} \doteq 1.220 < F_{0.025}[12, 15] \doteq 2.963.$$

Odtud vyplývá, že jsme na 5 % hladině významnosti nulovou hypotézu o shodnosti rozptylů nezamítli. Vzhledem k tomu, že $p-$hodnota příslušná k tomuto testu je rovna 0.353, zdá se, že se variabilita diastolického tlaku u obou pohlaví významně neliší. Pro testování shodnosti středních hodnot můžeme tedy použít zamítací pravidlo, které předpokládá shodu rozptylů:

$$\frac{\bar{x}_M - \bar{x}_Z}{\sqrt{(n-1)s_M^2 + (m-1)s_Z^2}} \sqrt{\frac{nm}{n+m}} \sqrt{n+m-2} \doteq 1.928 > t_{0.05}[27] \doteq 1.703.$$

Na 5 % hladině významnosti jsme zamítli nulovou hypotézu o shodnosti středního (průměrného) diastolického tlaku u obou pohlaví ve prospěch alternativy, která tvrdila, že střední diastolický tlak je vyšší u mužů než u žen. Podíváme-li se však na výsledek testu podrobněji, zjistíme, že $p-$hodnota je rovna přibližně 0.032. Znamená to, že pokud bychom volili 1 % hladinu významnosti, pak by nulová hypotéza zamítnuta nebyla. To může být způsobeno například malými rozsahy výběrů. V tomto případě bychom asi doporučili studii rozšířit na větší počet měření.

Odpověď: Pro zhodnocení výsledku studie jsme použili statistickou metodu testování hypotéz o shodnosti středního (průměrného) diastolického tlaku u obou pohlaví (za předpokladu normality dat). Na 5 % hladině významnosti byla nulová hypotéza, která tvrdila, že obě pohlaví mají v průměru stejný diastolický tlak, zamítnuta ve prospěch alternativy, která naopak tvrdila, že muži mají v průměru vyšší diastolický tlak než ženy. Přesto, že se zdá, že studie potvrdila předpoklad o vyšším průměrném diastolickém tlaku mužů, navrhovali bychom provést měření u většího počtu mužů a žen.

Příklad 8.6.

Od roku 1851 se pravidelně provádí měření průtoku řeky Labe v Děčíně. V tabulce jsou udány průměrné měsíční průtoky v m^3/s v prosinci a lednu v letech 1851 – 1989. Poznamenejme, že hydrologové pracují s takzvanými hydrologickými roky, které začínají již listopadem. Pro data v tabulce to znamená, že dvojice odpovídají bezprostředně za sebou jdoucím měsícům. Rozhodněte, zda prosincové a lednové průtoky kolísají kolem téže hodnoty a případný rozdíl mezi prosincovým a lednovým průměrem lze vysvětlit jen náhodnými odchylkami, nebo zda existuje statisticky významný rozdíl mezi středními (průměrnými) průtoky v prosinci a lednu. (Údaje jsou uvedeny po sloupcích.)

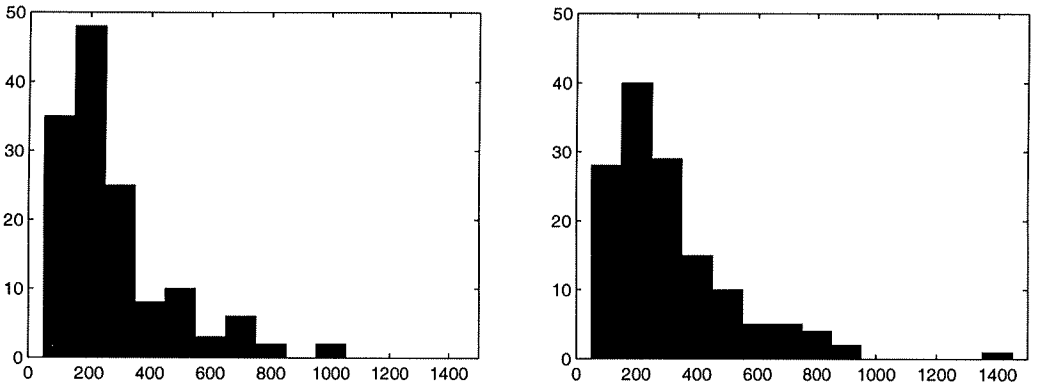
prosinec	leden	prosinec	leden	prosinec	leden	prosinec	leden
220.000	266.000	338.000	190.000	115.161	439.709	291.774	338.451
485.000	480.000	217.000	143.000	107.290	278.451	536.838	369.580
200.000	185.000	102.129	162.548	657.000	454.064	217.322	265.129
134.000	226.000	186.451	108.806	320.741	233.354	297.483	396.451
626.000	469.000	118.225	382.548	154.193	143.032	144.470	174.322
150.000	278.000	377.967	149.806	276.161	756.548	328.774	205.903
270.000	287.000	169.064	307.387	264.193	709.612	222.322	250.193
114.000	100.000	94.096	137.032	152.193	283.354	135.345	104.206
191.000	150.000	107.516	95.032	187.612	132.935	169.645	141.645
163.000	388.000	136.709	111.483	169.516	176.161	283.354	293.806
160.000	170.000	278.225	147.064	395.935	357.000	457.774	353.903
130.000	141.000	110.193	106.741	266.838	626.677	521.548	692.354
107.000	205.000	215.290	195.741	161.419	101.806	388.709	552.322
120.000	99.000	160.483	311.354	82.838	132.096	205.032	188.677
67.000	85.000	116.677	535.903	183.838	156.645	159.612	137.258
74.000	116.000	239.709	150.096	216.935	302.225	394.419	248.483
315.000	592.000	332.516	540.032	226.806	200.161	234.258	155.290
318.000	310.000	248.548	379.129	176.516	532.290	140.322	112.935
302.000	225.000	451.322	204.193	191.741	410.677	151.870	282.967
506.000	274.000	212.032	137.129	1001.612	228.322	1038.000	773.935
245.000	575.000	249.580	239.838	240.193	284.516	210.516	783.677
126.000	171.000	216.000	324.322	526.161	263.516	261.548	213.870
189.000	153.000	270.935	175.548	218.516	144.935	332.161	308.225
139.000	131.000	77.741	75.677	108.225	257.741	282.483	377.806
58.000	222.000	350.645	509.741	397.129	150.064	713.258	333.032
452.000	225.000	749.096	432.064	256.967	279.903	354.709	410.580
202.000	248.000	107.258	323.032	212.709	169.903	696.548	823.580
119.000	182.000	464.193	323.935	302.106	938.387	199.967	399.741
133.000	260.000	470.645	202.000	91.016	113.067	142.548	171.774
152.000	298.000	197.612	634.516	194.935	226.451	187.065	165.097
765.000	320.000	731.741	923.935	222.096	237.096	336.872	478.654
149.000	158.000	226.000	738.193	131.619	150.548	235.651	673.499
819.000	705.000	113.935	299.967	290.645	233.322	430.542	367.633
260.000	339.000	213.161	400.451	72.096	71.638	633.717	454.240
577.000	169.000	692.387	1393.161	300.838	490.774		

Řešení:

Nejprve uveďme základní popisné charakteristiky daných dvou souborů dat.

	prosinec	leden
průměr	277.63	311.94
směrodatná odchylka	188.65	209.87
koeficient šikmosti	1.72	1.85

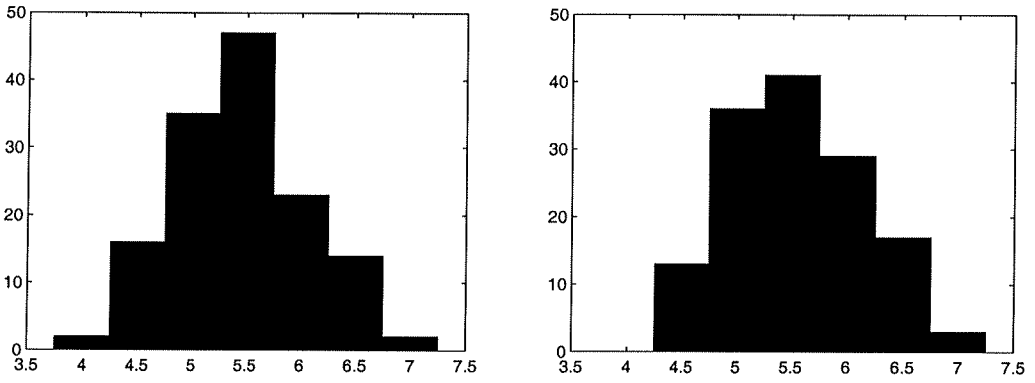
Prosincové a lednové průměry $\{x_i\}$ a $\{y_i\}$ nemohou být považovány za realizaci výběrů z normálního rozdělení, neboť mají vysoký koeficient šikmosti. Totéž potvrzují i jejich histogramy. Test založený na standardizované šikmosti zamítá předpoklad normality dat u obou souborů, neboť standardizovaná šikmost pro prosincové průměry je rovna přibližně 8.3 a pro lednové průměry je rovna 8.9. Obě hodnoty vysoko překračují 2.5% horní kvantil normálního rozdělení, který se rovná 1.96.



Často se tvrdí, že průměrné měsíční průtoky mají logaritmicko-normální rozdělení. Uveďme opět základní popisné charakteristiky spočtené ze zlogaritmovaných dat $\{\log x_i\}$ a $\{\log y_i\}$.

	prosinec	leden
průměr	5.4375	5.5561
směrodatná odchylka	0.6014	0.6003
koeficient šikmosti	0.283	0.261

Standardizovaný koeficient šikmosti spočítaný pro transformovaná data je roven přibližně 1.4 pro prosincová data a 1.3 pro lednová data. To znamená, že test založený na standardizované šikmosti nulovou hypotézu o normalitě dat nezamítl. Také histogramy ukazují, že se kladné sešikmení výrazně snížilo a že by zlogaritmovaná data mohla být považována za výběry z normálního rozdělení $N(\mu_p, \sigma_p^2)$, respektive $N(\mu_l, \sigma_l^2)$, a tedy původní data za výběry z logaritmicko-normálního rozdělení $LN(\mu_p, \sigma_p^2)$, respektive $LN(\mu_l, \sigma_l^2)$.

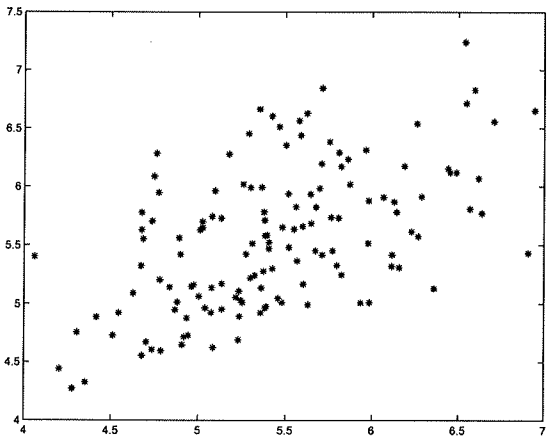


Jestliže $\mu_p = \mu_l$ a $\sigma_p^2 = \sigma_l^2$, pak také střední hodnota prosincových průtoků $e^{\mu_p + \sigma_p^2/2}$ se bude rovnat střední hodnotě lednových průtoků $e^{\mu_l + \sigma_l^2/2}$. Pro odhady parametrů platí:

$$\widehat{\mu_p} = \frac{1}{n}\sum \log x_i = 5.43755, \qquad \widehat{\sigma_p^2} = \frac{1}{n-1}\sum (\log x_i - \widehat{\mu_p})^2 = 0.364353$$
$$\widehat{\mu_l} = \frac{1}{n}\sum \log y_i = 5.55612, \qquad \widehat{\sigma_l^2} = \frac{1}{n-1}\sum (\log y_i - \widehat{\mu_l})^2 = 0.362968.$$

Zjednodušme si problém a předpokládejme, že $\sigma_p^2 = \sigma_l^2$. (Předpoklad je opodstatněn tím, že odhady $\widehat{\sigma_p^2}$ a $\widehat{\sigma_l^2}$ se příliš neliší.) Dále budeme testovat nulovou hypotézu H_0 : $\mu_p = \mu_l$ proti alternativě A : $\mu_p \neq \mu_l$.

Rozptylový graf (scatter plot) ukazuje (což také potvrzuje vysoký výběrový korelační koeficient spočtený z hodnot $\{\log x_i\}$ a $\{\log y_i\}$, který je roven 0.553), že veličiny $\{\log X_i\}$ a $\{\log Y_i\}$ nemůžeme považovat za nezávislé.



Závislost mezi prosincovými průtoky $\{X_i\}$ a lednovými průtoky $\{Y_i\}$ je způsobena tím, že pokud v určitém roce více prší, budou průtoky v prosinci i lednu vyšší, pokud méně prší, budou oba nižší. Závislost mezi průtoky se pak přenáší i na logaritmy těchto veličin, tj. na veličiny $\{\log X_i\}$ a $\{\log Y_i\}$. Proto místo původního dvouvýběrového problému použijeme párový test. Testujeme nulovou hypotézu H_0 : $\mu_p - \mu_l = 0$ proti alternativě A : $\mu_p - \mu_l \neq 0$. O náhodných veličinách $\{Z_i = \log X_i - \log Y_i\}$ budeme předpokládat, že mají normální rozdělení $N(\mu_z, \sigma_z^2)$, kde $\mu_z = \mu_p - \mu_l = -0.1189$. Spočtěme odhady $\widehat{\mu_z} = \bar{z} = -0.1186$ a $\widehat{\sigma_z^2} = s_z^2 = 0.3255$ a dosaďme do zamítacího pravidla:

$$\frac{|\bar{z}|}{s_z}\sqrt{n} \doteq 2.4504 > t_{0.025}[n-1] \doteq 1.9773.$$

Odpověď: Za předpokladu, že prosincové a lednové průtoky mají logaritmicko-normální rozdělení $LN(\mu_p, \sigma_p^2)$, resp. $LN(\mu_l, \sigma_l^2)$, se stejným parametrem $\sigma_p^2 = \sigma_l^2$, nulovou hypotézu o shodnosti parametrů μ_p a μ_l zamítáme. Znamená to, že prosincové a lednové průměry nejspíš kolísají kolem různých hodnot.

Příklad 8.7.

Od roku 1921 měří pražská meteorologická stanice v Břevnově srážky. V příkladu 7.7 jsou udány průměry a směrodatné odchylky měsíčních úhrnů v mm zjišťovaných v letech 1921 – 1989. Předpokládejte, že měsíční úhrny jsou výběrem z normálního rozdělení a otestujte, zda existuje statisticky významný rozdíl mezi únorovými a březnovými úhrny. Výběrový korelační koeficient r mezi únorovými a březnovými úhrny spočtenými z dat z let 1921 – 1989 je roven 0.2351.

Řešení:

Předpokládejme, že existuje kladná korelovanost mezi srážkovými úhrny měsíců, které jdou za sebou. Tento předpoklad může být odůvodněn tím, že fronta, která přináší déšť, obvykle trvá více dnů a tudíž může přesahovat z jednoho měsíce do druhého. Jestliže v daném roce jsou srážky v určitém měsíci větší, pak budou pravděpodobně větší i v měsíci následujícím. Jestliže předpokládáme kladnou korelovanost mezi únorovými a březnovými srážkami, tj. veličinami $\{X_i\}$ a $\{Y_i\}$, musíme použít párový test. Pro použití párového testu je třeba znát průměr $\bar{z}$ a odhad směrodatné odchylky s_z rozdílů mezi březnovým a únorovým úhrnem v témže roce. Platí

$$\begin{aligned}
\bar{z} &= \bar{y} - \bar{x} \doteq 30.4058 - 25.4493 = 4.9565, \\
s_z^2 &= \frac{1}{n-1}\sum((y_i - x_i) - (\bar{y} - \bar{x}))^2 = \\
&= \frac{1}{n-1}\left(\sum(x_i - \bar{x})^2 + \sum(y_i - \bar{y})^2 - 2\cdot\sum(x_i - \bar{x})(y_i - \bar{y})\right) = \\
&= s_x^2 + s_y^2 - 2\,r\,s_x s_y = \\
&= 12.5634^2 + 15.8201^2 - 2 \times 0.2351 \times 12.5634 \times 15.8201 = 314.6603,
\end{aligned}$$

kde $\bar{x}$ a s_x označují průměr a odhad směrodatné odchylky únorových úhrnů a obdobně $\bar{y}$ a s_y průměr a odhad směrodatné odchylky březnových úhrnů. Zvolme hladinu významnosti $\alpha = 0.05$ a dosaďme do zamítacího pravidla:

$$\frac{|\bar{z}|}{s_z}\sqrt{n} \doteq 2.321 > t_{0.025}[n-1] \doteq 1.995.$$

Odpověď: Na 5 % hladině významnosti jsme zamítli nulovou hypotézu, že neexistuje statisticky významný rozdíl mezi únorovými a březnovými srážkovými úhrny.

Neřešené příklady

Příklad 8.8.

U zatím používané technologie výroby kuliček do kuličkových ložisek je pravděpodobnost vyrobení zmetku 0.02. Výrobce by rád přešel na novou, úspornější technologii. Při jejím ověřování bylo vyrobeno 480 kuliček, z toho 13 zmetků. Otestujte, zda nová technologie významně nezhoršuje kvalitu produkce.

Příklad 8.9.

Sociální demokracie v parlamentních volbách roku 1996 získala 26.44% platných hlasů. V jednom z předvolebních průzkumů v červnu 1998 z 1839 dotázaných respondentů 415 uvedlo, že budou volit sociální demokracii, 200 bylo nerozhodnutých, 285 prohlásilo, že volit nebudou a zbylí uvedli jinou politickou stranu. Lze z výsledků tohoto průzkumu soudit, že preference sociální demokracie ve srovnání s rokem 1996 významně stouply?

Příklad 8.10.

Výroba superfosfátu je seřízena tak, aby dávala 15 kg kyseliny fosforečné na každých 100 kg hnojiva. V 10 vzorcích jsme zjistili obsah kyseliny fosforečné v %: 14.2, 15.7, 16.3, 13.8, 13.6, 14.7, 15.5, 14.0, 13.7, 13.9. Předpokládejte, že data jsou realizací výběru z normálního rozdělení. Máme důvod domnívat se, že průměrný obsah kyseliny fosforečné je nižší, než se požaduje?

Příklad 8.11.

Sedm kontrolních vážení dávek z dávkovače dalo výsledky v kg: 50.1, 49.4, 49.2, 49.4, 49.3, 49.9, 50.2. Správná hodnota dávky má být 50 kg. Nedochází k šizení zákazníků? Předpokládejte, že data jsou výběrem z normálního rozdělení.

Příklad 8.12.

Směrodatná odchylka automatického zařízení na plnění lahví piva nemá překročit 1 ml. Při kontrole byly zjištěny objemy 11 lahví v l: 0.4981, 0.5016, 0.5004, 0.4978, 0.4996, 0.5002, 0.4874, 0.4890, 0.4772, 0.5013, 0.4961. Předpokládejme, že data pocházejí z normálního rozdělení. Můžeme zařízení považovat za dostatečně přesné?

Příklad 8.13.

Má automatické zařízení z příkladu 8.12 systematickou chybu, víme-li, že správná hodnota plnění má být 0.5 l?

Příklad 8.14.

U 28 zkušebních vzorků betonu byla zjištěna pevnost v tlaku. Z dat bylo vypočten výběrový průměr $\bar{x} = 17.57$ MPa a odhad směrodatné odchylky $s = 2.23$ MPa. Jsme oprávněni za předpokladu normálního rozdělení tvrdit, že průměrná pevnost je významně menší než 18 MPa?

Příklad 8.15.

Množství oleje vyrobeného při rafinaci ropy je náhodná veličina, která se řídí přibližně normálním rozdělením. Obvyklou technologií se z určitého množství ropy získá

v průměru 59 litrů oleje. Z 10 vzorků ropy téhož množství se novou metodou získaly následující výtěžky oleje v l: 58, 60, 62, 60, 63, 57, 62, 64, 60, 58. Dává nová metoda v průměru významně vyšší výtěžek než obvyklá technologie?

Příklad 8.16.

Na celkové těžbě železné rudy se podílejí dva úseky. Otestujte, zda je kvalita rudy těžené na obou úsecích stejná. Kontrolní analýza poskytla následující výsledky (údaje jsou v %, předpokládáme opět normalitu):

1. úsek: 33.1, 34.7, 32.8, 36.5, 37.8, 33.8, 35.4, 37.2, 36.4, 35.8, 36.7, 37.8, 35.9.

2. úsek: 32.6, 31.8, 33.4, 33.5, 32.7, 34.0, 33.8, 32.8, 33.0, 33.9, 33.1, 32.9, 34.4, 34.1, 33.7, 33.5, 33.0, 32.7, 34.3, 33.8.

Příklad 8.17.

Bylo provedeno po 18 zkouškách pevnosti v tahu na vzorcích dvou různých kabelů. Souhrnné výsledky jsou:

kabel 1. druhu: $\bar{x} = 345.61$ MPa, $s_x^2 = 11.90$ MPa^2,

kabel 2. druhu: $\bar{y} = 340.50$ MPa, $s_y^2 = 35.91$ MPa^2.

Ověřte

a) zda lze považovat pozorovaný rozdíl ve variabilitě výsledků za náhodný (tj. zda je přijatelný předpoklad, že rozptyly pevnosti obou druhů kabelu jsou stejné),

b) zda střední hodnota pevnosti obou druhů kabelu je stejná.

Příklad 8.18.

Mírou seřízení dávkovače je směrodatná odchylka od požadované dávky. U dvou dávkovačů (A, B) byly zjištěny následující odchylky v kg od požadované dávky :

A	+0.5	+0.1	-0.3	-0.2	-0.7	+0.9	+0.1	-0.3	+0.3	+0.2	---
B	-0.3	+0.1	+0.2	-0.1	0	+0.3	+0.2	-0.2	-0.1	+0.3	-0.1

Předpokládejte, že hmotnosti dávek u obou dávkovačů jsou realizacemi normálně rozdělených náhodných veličin a otestujte, zda tyto dávkovače jsou stejně seřízeny.

Příklad 8.19.

Předpokládejte, že únorové a březnové úhrny měřené v témže roce v Břevnově (viz příklad 8.7) jsou nezávislé veličiny a použijte dvouvýběrový t-test. Bude nyní znít závěr jinak?

Příklad 8.20.

Pekařství uvažuje o zakoupení nové pece. Správné pečení vyžaduje, aby teplota byla během procesu konstantní. Byla provedena studie variability teploty během pečení u dvou typů pecí. Odhad rozptylu teploty před tím, než termostat opět zapálil plamen, byl u jednoho typu pece $\widehat{\sigma_1^2} = s_1^2 = 2.4$ (spočteno z 16 měření) a $\widehat{\sigma_2^2} = s_2^2 = 3.2$ (spočteno z 12 měření) pro druhý typ pece. Poskytuje tato informace důvod k závěru, že je rozdíl v rozptylu u těchto dvou typů pecí? Předpokládejte normální rozdělení teplot.

Příklad 8.21.

Deset vzorků bylo zpracováno vysokou teplotou. Měřila se jejich tvrdost před a po zpracování:

před zpracováním	3.15	2.98	3.00	2.75	3.21	3.33	2.95	2.81	3.26	2.88
po zpracování	3.21	2.99	3.11	2.91	3.22	3.28	3.09	3.00	3.28	2.99

Předpokládejte, že data jsou realizací výběru z dvojrozměrného normálního rozdělení a otestujte hypotézu, že zpracováním vysokou teplotou se zvýší tvrdost vzorků.

Příklad 8.22.

V cementárně byla provedena kontrola hmotností dávek v kg ze dvou dávkovačů.

1. dávkovač: 51.5, 47.0, 48.5, 53.0, 47.3, 48.1, 48.8, 49.2, 52.3, 47.1, 49.5, 46.3, 50.1.
2. dávkovač: 50.3, 50.7, 48.2, 50.1, 49.8, 48.9, 51.1, 48.9, 50.3.

Otestujte za předpokladu normality, zda můžeme rozptýlenost, resp. střední hodnotu dávek u obou dávkovačů považovat zhruba za stejnou.

Příklad 8.23.

U stejného stavebního materiálu vyráběného dvěma firmami (A, B) byla zjišťována vlhkost v %. Data byla roztříděna do společné tabulky skupinového rozdělení četností:

	⟨11-12)	⟨12-13)	⟨13-14)	⟨14-15)	⟨15-16)	⟨16-17)	⟨17-18)
četnost A	0	1	4	12	13	9	3
četnost B	2	1	6	13	9	4	0

Rozhodněte, zda se průměrná vlhkost materiálu od těchto dvou firem významně liší. Předpokládejte normální rozdělení vlhkosti.

Příklad 8.24.

U skupiny náhodně vybraných řidičů z povolání a u skupiny náhodně vybraných čerstvých absolventů autoškoly byla měřena rychlost reakce v setinách sekundy na světelný signál. Předpokládejme normální rozdělení reakční doby. Z údajů první skupiny, která měla 10 osob, bylo vypočteno $\bar{x} = 34$, $s_x^2 = 3.023$. Z údajů druhé skupiny o 20 osobách bylo vypočteno $\bar{y} = 42$, $s_y^2 = 5.921$.

a) Můžeme považovat variabilitu reakční doby v těchto dvou skupinách za stejnou?

b) Potvrzují data, že řidiči z povolání mají významně kratší reakční dobu, než začínající řidiči?

Příklad 8.25.

Krychelná pevnost betonu B 40 v MPa byla zkoušena na sadě 16 krychlí jednak určitou nedestruktivní metodou a poté destruktivně.

nedestruktivně	56.6	53.9	60.5	59.0	61.0	59.3	43.8	57.1
destruktivně	51.8	51.0	57.1	61.6	59.8	53.5	45.5	60.7

nedestruktivně	47.7	50.7	53.0	56.0	60.0	63.6	53.3	52.0
destruktivně	49.8	51.1	49.7	59.3	58.2	59.7	54.6	47.8

Předpokládejte, že data pocházejí z dvojrozměrného normálního rozdělení a otestujte, zda se změřená pevnost nedestruktivní metodou v průměru jen náhodně liší od výsledku zjištěného destruktivně (tj. zda jsou tyto dvě metody záměnné).

Příklad 8.26.

U 30 motorů určitého typu byly změřeny zkušební a provozní výkony v %:

zkušební	provozní	zkušební	provozní	zkušební	provozní
88.91	89.30	84.12	88.48	76.56	77.34
89.65	91.83	78.95	80.11	85.93	80.87
74.40	66.37	79.97	75.32	77.60	76.17
86.57	87.00	75.66	71.44	83.55	79.31
81.39	77.42	87.04	78.13	82.55	80.00
82.33	81.12	75.47	68.61	81.96	72.39
73.80	61.51	84.95	77.04	86.29	80.07
80.96	80.42	77.95	76.75	85.44	82.50
77.67	70.94	89.71	82.11	88.80	90.37
73.51	68.52	74.18	72.03	89.41	93.64

Předpokládejte, že data jsou výběrem z dvojrozměrného normálního rozdělení. Otestujte, zda můžeme zkušební a provozní výkony považovat v průměru za stejné. Výrobce tvrdí, že provozní výkon může být menší než zkušební výkon, v průměru však nejvýše o 1%. Otestujte, zda toto tvrzení můžeme považovat za pravdivé.

Příklad 8.27.

V devíti po sobě jdoucích dnech zjišťovali dva laboranti obsah dusíku v % v určitém plynu:

1. laborant	4	32	35	43	34	36	48	33	33
2. laborant	18	37	38	36	47	48	57	28	42

Předpokládejte, že data jsou výběrem z dvojrozměrného normálního rozdělení. Je statisticky významný rozdíl mezi průměrnými výsledky 1. a 2. laboranta?

Příklad 8.28.

Otestujte, zda únorové a březnové srážkové úhrny, měřené v břevnovské meteorologické stanici (viz příklad 8.7) jsou kladně zkorelované. Hladinu významnosti volte $\alpha = 0.05$.

Příklad 8.29.

Při výrobě asfaltu byl zjišťován bod měknutí ve stupních Celsia a penetrace v mm při 25°C:

bod měknutí	46.5	48	46.5	47.5	47	47.5	49	48	47.5	48.5
penetrace	9.9	9.1	9.8	9.4	9.8	9.5	8.4	8.8	9.6	8.7

Můžeme tyto dvě veličiny považovat za nezávislé? Předpokládejte, že data pocházejí z dvojrozměrného normálního rozdělení.

Příklad 8.30.

Na základě dat z příkladu 6.4 otestujte, zda je statisticky významná závislost mezi výnosy pšenice ozimu a průměrnou zimní teplotou, mezi výnosy pšenice a teplotou během vegetačního období a mezi výnosy pšenice a množstvím srážek během vegetačního období. Předpokládejte, že sdružené rozdělení všech zjišťovaných veličin je normální.

Příklad 8.31.

Otestujte, zda data z příkladu 6.4 vykazují statisticky významnou závislost mezi teplotou v zimním a letním období v témže roce. Čím by taková závislost mohla být případně vysvětlena? Předpokládejte stejně jako v předchozím příkladě normalitu.

9. Analýza rozptylu

Řešené příklady

Příklad 9.1.

Za účelem zkoumání vlivu ročního období na pevnost bylo zjištěno celkem 30 krychelných pevností betonu určité třídy v MPa (jaro a podzim byly vzhledem k podobným teplotním poměrům sloučeny do jedné skupiny).

Roční období:	jaro, podzim	léto	zima
Zjištěné pevnosti:	23.8	23.5	23.2
	24.2	23.0	23.5
	25.1	24.2	25.2
	27.6	25.6	25.8
	26.7	27.0	19.8
	26.0	27.5	20.2
	21.8	21.7	26.7
	23.6	20.6	20.2
	23.0	22.1	21.0
	21.5	21.1	20.5
Součet hodnot:	243.3	236.3	226.1
Součet čtverců:	5956.19	5636.37	5173.43

Otestujte, má-li roční období vliv na pevnost betonu. Pokud ano, určete významně se lišící dvojice ročních období.

Řešení:

Použijeme analýzu rozptylu – jednoduché třídění. Předpokládáme, že pevnosti zjištěné na jaře a na podzim tvoří náhodný výběr $X_{1,1}, \ldots, X_{1,10}$ z normálního rozdělení o parametrech $\mu + \alpha_1$ a σ^2, letní údaje tvoří náhodný výběr $X_{2,1}, \ldots, X_{2,10}$ z normálního rozdělení o parametrech $\mu + \alpha_2$ a σ^2 a konečně zimní údaje tvoří náhodný výběr $X_{3,1}, \ldots, X_{3,10}$ z normálního rozdělení o parametrech $\mu + \alpha_3$ a σ^2. Nulová hypotéza říká, že daný faktor (tj. roční období) nemá vliv na sledovanou veličinu (tj. pevnost). Alternativou je opačné tvrzení. Matematicky zapsáno:

Nulová hypotéza H_0: $\alpha_i = 0$ pro všechna $i = 1, 2, 3$.

Alternativa A: Alespoň jedno α_i je různé od nuly.

Z výše uvedených tří nezávislých výběrů spočítáme statistiky $X_{i.}$ a $\sum_{j=1}^{10} X_{ij}^2$ pro $i = 1, 2, 3$, viz poslední dva řádky v tabulce. Určíme dále celkový součet všech hodnot jako součet předposledního řádku tabulky: $X_{..} = 705.7$ a součet všech druhých mocnin jako součet posledního řádku tabulky: $\sum_{i=1}^{3} \sum_{j=1}^{10} X_{ij}^2 = 16765.99$.

V dalším kroku vypočteme celkový součet čtverců S_T, součet čtverců mezi úrovněmi daného faktoru S_A a reziduální součet čtverců S_e.

$$S_T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{..})^2 = \left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} X_{ij}^2\right) - \frac{1}{N}X_{..}^2\,,$$

$$S_A = \sum_{i=1}^{m} n_i(\bar{x}_{i.} - \bar{x}_{..})^2 = \left(\sum_{i=1}^{m}\frac{1}{n_i}X_{i.}^2\right) - \frac{1}{N}X_{..}^2\,,$$

$$S_e = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_{i.})^2 = S_T - S_A\,.$$

V našem příkladu je $m = 3$, $n_1 = n_2 = n_3 = 10$, $N = 30$ a dosazením do pravých stran vzorců dostáváme:

$$S_T = 16765.99 - \tfrac{1}{30} \times 705.7^2 = 165.5737\,,$$
$$S_A = (\tfrac{1}{10} \times 243.3^2 + \tfrac{1}{10} \times 236.3^2 + \tfrac{1}{10} \times 226.1^2) - \tfrac{1}{30} \times 705.7^2 = 14.9627\,,$$
$$S_e = 165.5737 - 14.9627 = 150.611\,.$$

Konečně vypočteme testovou statistiku

$$F = \frac{S_A/(m-1)}{S_e/(N-m)} = \frac{(N-m)\cdot S_A}{(m-1)\cdot S_e} = \frac{27 \times 14.9627}{2 \times 150.611} = 1.341\,,$$

která je výrazně menší než kvantil $F_{0.05}[2, 27] = 3.354$. (Odpovídající p-hodnota testu je 0.278.) Nulovou hypotézu tedy nemůžeme zamítnout.

Předchozí výpočty se dají shrnout do přehledné tabulky – uvádíme nejdříve její obecný tvar a potom dosazení:

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testovací statistika
faktor A	S_A	$m-1$	$S_A/(m-1)$	$\dfrac{S_A/(m-1)}{S_e/(N-m)}$
reziduální	S_e	$N-m$	$S_e/(N-m)$	---
celkový	S_T	$N-1$	---	---

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testovací statistika
roční období	14.9627	2	7.4813	1.341
reziduální	150.611	27	5.5782	---
celkový	165.5737	29	---	---

Odpověď: Nepodařilo se prokázat, že by pevnost betonu dané třídy závisela na ročním období.

Příklad 9.2.

Byly zkoumány výnosy sena v q/ha v závislosti na typu půdy a na způsobu hnojení. Každá kombinace byla realizována čtyřikrát, výnosy byly:

Normální půda, bez hnojení	28, 32, 30, 30,
normální půda, přírodní hnojivo	37, 36, 39, 36,
normální půda, umělé hnojivo	34, 38, 37, 36,
kyselá půda, bez hnojení	31, 27, 30, 29,
kyselá půda, přírodní hnojivo	34, 34, 30, 38,
kyselá půda, umělé hnojivo	42, 40, 41, 39.

Testujte, je-li výnos sena významně ovlivněn typem půdy, způsobem hnojení, popř. projevují-li se významně interakce obou faktorů.

Řešení:

Je třeba použít analýzu rozptylu – dvojné třídění s interakcemi. Vysvětlíme právě na tomto příkladu jednotlivé kroky i teoretické pozadí metody.

Data nejprve uspořádáme do přehlednější tabulky, přidáme další dva sloupce a řádky pro součty hodnot a součty čtverců (tyto pomocné hodnoty budeme potřebovat při dalších výpočtech).

Typ půdy	Způsob hnojení			Součet hodnot	Součet čtverců
	Bez hnojení	Přírodní hnojivo	Umělé hnojivo		
Normální	28, 32, 30, 30	37, 36, 39, 36	34, 38, 37, 36	413	14355
Kyselá	31, 27, 30, 29	34, 34, 30, 38	42, 40, 41, 39	415	14653
Součet hodnot	237	284	307	828	---
Součet čtverců	7039	10138	11831	---	29008

Dvojné třídění předpokládá, že máme data roztříděná pomocí dvou faktorů, z nichž první má m úrovní a druhý má l úrovní, do $m \cdot l$ nezávislých náhodných výběrů $(X_{111}, \ldots, X_{11n})$, $(X_{121}, \ldots, X_{12n})$, ..., $(X_{ml1}, \ldots, X_{mln})$. (V každém tomto výběru se předpokládá stejný počet pozorování, rovný n – uvažujeme jen tzv. vyvážené dvojné třídění.) Je-li $n = 1$, mluví se o dvojném třídění bez opakování, chceme-li však použít dvojné třídění s interakcemi, musíme mít $n > 1$.

V našem příkladu je prvním faktorem typ půdy, v experimentu rozlišujeme dvě úrovně – normální a kyselou půdu. Druhým faktorem je zřejmě způsob hnojení mající tři faktory. Tedy $m = 2$, $l = 3$ a $n = 4$.

Teoretickým předpokladem metody je, že všech $m \cdot l$ náhodných výběrů pochází z normálního rozdělení se stejným rozptylem, rovným neznámému parametru σ^2 a s obecně různými středními hodnotami. Důležitou otázkou pak je, jak tyto střední hodnoty závisí na jednotlivých faktorech. Obvykle se základní (nejobecnější) model zapisuje ve tvaru:

$$X_{ijk} \sim N(\mu + \alpha_i + \beta_j + \lambda_{ij}\,,\, \sigma^2), \qquad i = 1, \ldots, m, \quad j = 1, \ldots, l, \quad k = 1, \ldots, n.$$

Zde μ, α_i, β_j, λ_{ij} jsou neznámé parametry, vyjadřující právě vliv jednotlivých faktorů na střední hodnotu. Přitom α_i vyjadřuje vliv prvního faktoru, β_j vliv druhého faktoru a λ_{ij} vliv interakcí (spolupůsobení) obou faktorů. Pokud by např. platilo, že $\alpha_i = 0$ a $\lambda_{ij} = 0$ pro všechna i a j, znamenalo by to, že první faktor ani interakce faktorů nemají vliv na sledovanou veličinu.

Testují se nezávisle na sobě 3 hypotézy:

- Nulová hypotéza H_A: $\alpha_i = 0$ pro všechna $i = 1, \ldots, m$.

  Alternativa A: Alespoň jedno α_i je různé od nuly.

  Zamítací pravidlo: $F_A = \dfrac{S_A/(m-1)}{S_e/(N-ml)} > F_\alpha[m-1, N-ml]$.

- Nulová hypotéza H_B: $\beta_j = 0$ pro všechna $j = 1, \ldots, l$.

  Alternativa A: Alespoň jedno β_j je různé od nuly.

  Zamítací pravidlo: $F_B = \dfrac{S_B/(l-1)}{S_e/(N-ml)} > F_\alpha[l-1, N-ml]$.

- Nulová hypotéza H_{AB}: $\lambda_{ij} = 0$ pro všechna $i = 1, \ldots, m$, $j = 1, \ldots, l$.

  Alternativa A: Alespoň jedno λ_{ij} je různé od nuly.

  Zamítací pravidlo: $F_{AB} = \dfrac{S_{AB} \,/\, [(m-1)(l-1)]}{S_e/(N-ml)} > F_\alpha[(m-1)(l-1), N-ml]$.

Hypotéza H_A tvrdí, že první faktor (obvykle označovaný písmenem A, jeho úrovně tvoří řádky v tabulce analýzy rozptylu) nemá vliv na sledovanou veličinu. V našem příkladu jde tedy o hypotézu, že výnos sena není ovlivněn typem půdy. Alternativou je opačné tvrzení. Analogický význam má hypotéza H_B, která se vztahuje ke druhému faktoru (faktoru B tvořícímu sloupce tabulky). V našem příkladu tato hypotéza tvrdí, že výnos sena není ovlivněn způsobem hnojení.

Hypotéza H_{AB} tvrdí, že není žádný vliv interakcí obou faktorů. Příslušná alternativa pak říká, že se interakce projevují, tj. v našem příkladu by to znamenalo, že některé hnojivo mé specifické účinky na určitý druh půdy, projevující se významným zvýšením nebo snížením výnosu sena.

Pro dosazení do zamítacích pravidel je třeba vypočítat statistiky:

S_T ... celkový součet čtverců,
S_A ... součet čtverců mezi úrovněmi faktoru A,
S_B ... součet čtverců mezi úrovněmi faktoru B,
S_{AB} ... součet čtverců odpovídající interakcím mezi faktory A a B,
S_e ... reziduální součet čtverců.

Statistika S_A např. měří, jak daleko jsou řádkové průměry od celkového průměru. Veličina $s^2 = S_e/(N-ml)$ se nazývá reziduální rozptyl a je odhadem neznámého parametru σ^2. Hypotézu H_A zamítneme, je-li statistika S_A příliš velká, tj. liší-li se řádkové průměry příliš od celkového průměru. Přitom se ovšem musí vzít v úvahu i vlastní náhodné kolísání dat, vyjádřené právě reziduálním rozptylem. Podobné vysvětlení mají i další dva testy.

Uvedeme nyní vzorce pro výpočet potřebných statistik. První z vzorců na řádku je vždy názornější, druhý se pak používá k praktickému výpočtu.

$$S_T = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}(X_{ijk} - \bar{x}_{...})^2 = \left(\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n} X_{ijk}^2\right) - \frac{1}{N}X_{...}^2\,,$$

$$S_A = l \cdot n \sum_{i=1}^{m}(\bar{x}_{i..} - \bar{x}_{...})^2 = \left(\frac{1}{l \cdot n}\sum_{i=1}^{m} X_{i..}^2\right) - \frac{1}{N}X_{...}^2\,,$$

$$S_B = m \cdot n \sum_{j=1}^{l}(\bar{x}_{.j.} - \bar{x}_{...})^2 = \left(\frac{1}{m \cdot n}\sum_{j=1}^{l} X_{.j.}^2\right) - \frac{1}{N}X_{...}^2\,,$$

$$S_e = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n}(X_{ijk} - \bar{x}_{ij.})^2 = \left(\sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n} X_{ijk}^2\right) - \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{l} X_{ij.}^2\,,$$

$$S_{AB} = n\sum_{i=1}^{m}\sum_{j=1}^{l}(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2 = S_T - S_A - S_B - S_e\,.$$

Používáme podobné označení jako v jednoduchém třídění, zobecněné o další index. Tečka na místě určitého indexu znamená sčítání přes všechny hodnoty tohoto indexu. Pruh navíc znamená průměrování přes všechny hodnoty indexu, na jehož místě je tečka. N označuje celkový počet všech dat, tj. $N = m \cdot l \cdot n$.

Platí tedy např.:

$X_{i..} = \sum_{j=1}^{l}\sum_{k=1}^{n} X_{ijk}$ ... součet všech hodnot pro i-tou úroveň prvního faktoru, tj. součet hodnot pro první řádek tabulky,

$X_{...} = \sum_{i=1}^{m}\sum_{j=1}^{l}\sum_{k=1}^{n} X_{ijk}$ ... součet všech hodnot všech souborů,

$\bar{x}_{i..} = X_{i..}/(l \cdot n)$ ... průměr všech hodnot pro i-tou úroveň prvního faktoru, tj. průměr hodnot pro první řádek tabulky,

$\bar{x}_{...} = X_{...}/N$ ... průměr všech hodnot všech souborů, apod.

Použijeme-li součty hodnot a součty druhých mocnin, uvedené v tabulce na začátku řešení, snadno vypočteme:

$$S_T = 29008 - \tfrac{1}{24}828^2 = 442\,,$$
$$S_A = \tfrac{1}{3\cdot 4}(413^2 + 415^2) - \tfrac{1}{24}828^2 = 0.1667\,,$$
$$S_B = \tfrac{1}{2\cdot 4}(237^2 + 284^2 + 307^2) - \tfrac{1}{24}828^2 = 318.25\,.$$

Pro výpočet reziduálního součtu čtverců potřebujeme veličiny $X_{ij.}$, což jsou součty hodnot jednotlivých výběrů pro kombinace úrovní jednotlivých faktorů. Tyto veličiny uvádíme v následující tabulce:

120	148	145
117	136	162

Pomocí nich vypočteme:

$$S_e = 29008 - \tfrac{1}{4}(120^2 + 148^2 + \cdots + 162^2) = 68.5\,,$$
$$S_{AB} = 442 - 0.1667 - 418.25 - 68.5 = 55.0833\,.$$

Opět uvedeme tabulku analýzy rozptylu, přehledně zachycující již vypočtené veličiny i mechanismus dosazování do zamítacích pravidel pro naše tři hypotézy.

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testovací statistika
faktor A	S_A	$m-1$	$\dfrac{S_A}{m-1}$	$\dfrac{S_A/(m-1)}{S_e/(N-ml)}$
faktor B	S_B	$l-1$	$\dfrac{S_B}{l-1}$	$\dfrac{S_B/(l-1)}{S_e/(N-ml)}$
interakce	S_{AB}	$(m-1)(l-1)$	$\dfrac{S_{AB}}{(m-1)(l-1)}$	$\dfrac{S_{AB}/[(m-1)(l-1)]}{S_e/(N-ml)}$
reziduální	S_e	$N-ml$	$S_e/(N-ml)$	---
celkový	S_T	$N-1$	---	---

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testovací statistika
typ půdy	0.1667	1	0.1667	0.044
hnojení	318.25	2	159.125	41.814
interakce	55.0833	2	27.5417	7.237
reziduální	68.5	18	3.8056	---
celkový	442	23	---	---

Hodnotu testovací statistiky $F_A = 0.044$ je třeba porovnat s kvantilem $F_{0.05}[1, 18] = 4.414$, hodnoty statistik $F_B = 41.814$ a $F_{AB} = 7.237$ se obě porovnávají s kvantilem $F_{0.05}[2, 18] = 3.555$. Vidíme tedy, že typ půdy rozhodně nemá významný vliv na výnos, naopak vliv způsobu hnojení je velmi významný. Vliv interakcí je také významný, při porovnání statistiky F_{AB} s kvantilem $F_{0.01}[2, 18] = 6.013$ však vidíme, že na 1-procentní hladině významnosti již interakce významné nejsou.

Odpověď: Výnos sena není významně ovlivněn typem půdy, je velmi významně ovlivněn způsobem hnojení. Interakce typu půdy a způsobu hnojení se projevují též významně.

Příklad 9.3.

Rozhodněte na základě dat z příkladu 9.2, které způsoby hnojení půdy se významně liší ve výnosu sena.

Řešení:

V příkladu 9.2 se prokázal významný vliv faktoru B na výnosy, položená otázka je tedy přirozená.

Podle Scheffeho kritéria se významně liší ty hodnoty parametrů α_i a α_t (tj. významně se liší i-tá a t-tá úroveň faktoru A pro taková i a t), pro které platí nerovnost:

$$|\bar{x}_{i..} - \bar{x}_{t..}| > \sqrt{\frac{2(m-1)S_e}{ln(N-ml)} F_\alpha[m-1, N-ml]}.$$

Podobně platí, že se významně liší ty hodnoty parametrů β_j a β_s (tj. významně se liší j-tá a s-tá úroveň faktoru B pro taková j a s), pro které platí nerovnost:

$$|\bar{x}_{.j.} - \bar{x}_{.s.}| > \sqrt{\frac{2(l-1)S_e}{mn(N-ml)} F_\alpha[l-1, N-ml]}.$$

V našem příkladu jde o prokazování odlišností mezi úrovněmi faktoru B, pravá strana poslední nerovnosti je:

$$\sqrt{\frac{2 \times 2 \times 68.5}{2 \times 4 \times 18} \times F_{0.05}[2, 18]} = 2.6\,.$$

Vypočteme-li absolutní hodnoty rozdílů mezi sloupcovými průměry výnosů sena, dostaneme pro rozdíl mezi žádným a přírodním hnojivem hodnotu 5.875, pro rozdíl mezi přírodním a umělým hnojivem hodnotu 2.875 a pro rozdíl mezi žádným a umělým hnojivem hodnotu 8.75. Vidíme tedy, že se na 5-ti procentní hladině významnosti všechny způsoby hnojení významně liší.

Kdybychom zvolili hladinu významnosti $\alpha = 0.01$, pak by se rozdíly mezi sloupcovými průměry porovnávaly se změněnou pravou stranou zamítacího pravidla 3.383 a tedy rozdíl mezi přírodním a umělým hnojivem by se nepodařilo prokázat.

Odpověď: Na 5-ti procentní hladině významnosti se významně liší všechny způsoby hnojení, na 1-procentní hladině by se významný rozdíl mezi přírodním a umělým hnojivem nepodařilo prokázat.

Příklad 9.4.

Byl zkoušen nový postup výroby určité součástky. Prověřovaly se čtyři různé způsoby zpracování a tři různé druhy materiálu. U 12 výrobků, náhodně vybraných vždy pro jeden postup zpracování a pro jeden druh materiálu, byly zjištěny ukazatele jakosti, uvedené v následující tabulce:

Způsob zpracování	Druh materiálu			Součet hodnot	Součet čtverců
	I	II	III		
i	6.5	7	5.7	19.2	123.74
ii	6.9	6.3	4.9	18.1	111.31
iii	6.7	6.4	4.4	17.5	105.21
iv	6.1	6.7	4.9	17.7	106.11
Součet hodnot	26.2	26.4	19.9	72.5	---
Součet čtverců	171.96	174.54	99.87	---	446.37

Rozhodněte, je-li jakost výrobků významně ovlivněna způsobem zpracování, popř. druhem použitého materiálu.

Řešení:

Data, ze kterých vycházíme, jsou roztříděna pomocí dvou faktorů do $4 \cdot 3$ skupin, tentokrát je však v každé skupině jen jedno pozorování. Nemůžeme použít dvojné třídění s interakcemi mimo jiné i proto, že při aplikaci vzorců z příkladu 9.2 bychom dostali nesmyslné výsledky – např. počet stupňů volnosti $N-ml$ by byl roven nule, apod. Proto použijeme jednodušší verzi dvojného třídění – dvojné třídění bez interakcí. (Tato verze dvojného třídění se může použít i v případě, kdy je $n > 1$, jsme-li předem přesvědčeni, že by interakce neměly působit.)

Budeme postupovat stručněji.

Základní model se ve dvojném třídění bez interakcí zapisuje ve tvaru:

$$X_{ijk} \sim N(\mu + \alpha_i + \beta_j\,,\, \sigma^2), \qquad i = 1,\ldots,m, \quad j = 1,\ldots,l, \quad k = 1,\ldots,n.$$

Testují se nezávisle na sobě 2 hypotézy:

- Nulová hypotéza H_A: $\alpha_i = 0$ pro všechna $i = 1,\ldots,m$.
  Alternativa A: Alespoň jedno α_i je různé od nuly.
  Zamítací pravidlo: $F_A = \dfrac{S_A/(m-1)}{S_e/(N-m-l+1)} > F_\alpha[m-1, N-m-l+1].$
- Nulová hypotéza H_B: $\beta_j = 0$ pro všechna $j = 1,\ldots,l$.
  Alternativa A: Alespoň jedno β_j je různé od nuly.
  Zamítací pravidlo: $F_B = \dfrac{S_B/(l-1)}{S_e/(N-m-l+1)} > F_\alpha[l-1, N-m-l+1].$

Pro dosazení do zamítacích pravidel se nejprve vypočítají statistiky S_T, S_A, S_B podle naprosto stejných vzorců jako v dvojném třídění s interakcemi. Veličina S_{AB} se nepočítá a reziduální součet čtverců dostaneme jednoduše ze vztahu:

$$S_e = S_T - S_A - S_B.$$

Pokud je $n = 1$, lze samozřejmě všechny vzorce zjednodušit – třetí index je nadbytečný.

V našem příkladu máme:

$$S_T = 446.37 - \tfrac{1}{12}72.5^2 = 8.3492\,,$$
$$S_A = \tfrac{1}{3}(19.2^2 + 18.1^2 + 17.5^2 + 17.7^2) - \tfrac{1}{12}72.5^2 = 0.5758\,,$$
$$S_B = \tfrac{1}{4}(26.2^2 + 26.4^2 + 19.9^2) - \tfrac{1}{12}72.5^2 = 6.8317\,,$$
$$S_e = 8.3492 - 0.5758 - 6.8312 = 0.9417\,.$$

Zdroj variability	Součet čtverců	Počet stupňů volnosti	Podíl	Testovací statistika
způsob zpracování	0.5758	3	0.1919	1.223
druh materiálu	6.8317	2	3.4158	21.765
reziduální	0.9417	6	0.1569	---
celkový	8.3492	11	---	---

Hodnotu testovací statistiky $F_A = 1.223$ je třeba porovnat s kvantilem $F_{0.05}[3,6] = 4.757$, hodnotu statistiky $F_B = 21.765$ s kvantilem $F_{0.05}[2,6] = 5.143$.

Odpověď: Způsob zpracování kvalitu významně neovlivňuje, zatímco vliv druhu materiálu je velmi významný.

Příklad 9.5.

Rozhodněte na základě dat z příkladu 9.4, které druhy materiálu se významně liší ve výsledné jakosti výrobků.

Řešení:

Postup řešení se velmi podobá řešení příkladu 9.3, odlišnost je však ve stupních volnosti. Podle Scheffeho kritéria se v modelu bez interakcí významně liší ty hodnoty parametrů β_j a β_s, pro které platí nerovnost:

$$|\bar{x}_{.j.} - \bar{x}_{.s.}| > \sqrt{\frac{2(l-1)S_e}{mn(N-m-l+1)} F_\alpha[l-1, N-m-l+1]}.$$

V našem příkladu je pravá strana poslední nerovnosti:

$$\sqrt{\frac{2 \times 2 \times 0.9422}{4 \times 1 \times 6} \times F_{0.05}[2,6]} = 0.899\,.$$

Vypočteme-li absolutní hodnoty rozdílů mezi sloupcovými průměry ukazatelů jakosti, dostaneme:

$$\bar{x}_{.1} - \bar{x}_{.2} = 0.05, \qquad \bar{x}_{.1} - \bar{x}_{.3} = 1.575, \qquad \bar{x}_{.2} - \bar{x}_{.3} = 1.625.$$

Odpověď: Významně se liší první druh materiálu od třetího a druhý druh od třetího.

Neřešené příklady

Příklad 9.6.

U tří dávkovacích automatů bylo náhodně vybráno a zváženo po pěti dávkách (hmotnosti dávek jsou uvedeny v gramech):

A	49	52.5	48	53	50.5
B	45	50	47.5	51	43.5
C	55	52.5	54.5	52.5	49.5

Rozhodněte, zda jednotlivé automaty dávkují v průměru stejně, popř. které dvojice automatů se významně liší.

Příklad 9.7.

Ze tří výroben prefabrikátů bylo získáno celkem 20 výsledků zkoušek krychelné pevnosti v MPa:

A	25.8	27.6	33	26.2	28.7	31.1	---	---
B	28.7	26.9	34.6	37.3	25.1	28.7	---	---
C	28.6	37.1	25.1	33.8	37.2	26.4	32.5	29.4

Rozhodněte, zda průměrné pevnosti prefabrikátů z jednotlivých výroben můžeme považovat za stejné. V opačném případě označte významně se lišící dvojice výroben.

Příklad 9.8.

Studenty, kteří konali přijímací zkoušky na jistou vysokou školu, bylo možno rozdělit do čtyř skupin podle typu absolvované střední školy. Z výsledků přijímací zkoušky, ohodnocených u každého studenta v rozsahu 0 až 20 bodů, byly v každé skupině zjištěny údaje zpracované v následující tabulce:

Typ SŠ	Počet studentů	Součet hodnot	Součet čtverců
pražské gymnázium	108	1192	14121
jiné gymnázium	212	2120	23144
pražská SPŠ	198	1544	13424
jiná SPŠ	214	1608	13994

Rozhodněte, je-li výsledek u přijímací zkoušky závislý na absolvované střední škole. Pokud ano, označte významně se lišící typy škol.

Příklad 9.9.

Během jednoho pracovního dne byly u 3 směn a u 5 strojů zjišťovány počty zmetků:

Směna	Stroj				
	A	B	C	D	E
1.	4	11	14	9	21
2.	7	9	15	11	18
3.	7	6	12	7	19

Rozhodněte, zda je počet zmetků významně ovlivněn směnou či druhem stroje.

Příklad 9.10.

Ve dvou srovnatelných oblastech byl sledován prodej dvojím způsobem baleného zboží ve třech stejně dlouhých časových obdobích, lišících se způsobem reklamy. Výsledky jsou tržby za prodané zboží v tisících dolarů.

Způsob balení	Druh reklamy		
	žádná	v tisku	v tisku i v TV
v sáčku	1.1, 2.0	1.2, 0.7	6.9, 3.6
v krabičce	3.5, 3.8	4.5, 2.9	11.7, 5.4

Rozhodněte, zda způsob balení, druh reklamy, popř. jejich interakce významně ovlivňují prodej.

Příklad 9.11.

Byla zkoumána pevnost v ohybu železobetonových dílců. K dispozici je celkem 30 změřených pevností zkušebních dílců v MPa, rozlišených podle betonárky, dodávající beton pro jejich výrobu, do 2 skupin a podle výrobny do 3 skupin.

Betonárka	Číslo výrobny		
	I	II	III
A	17 21 19 15 23	18 20 24 22 19	22 17 23 21 25
B	17 22 18 23 20	19 25 24 22 19	25 19 28 24 23

Rozhodněte, zda betonárka, popř. výrobna významně ovlivňují pevnost.

Příklad 9.12.

Tři modifikace určitého typu vozu stejné značky jsou současně vyráběny ve dvou továrnách. U celkem 18 náhodně vybraných vozů byly zjištěny výkonové charakteristiky. V první tabulce jsou údaje z první továrny, ve druhé pak z druhé továrny:

Modifikace	1	2	3	1	2	3	1	2	3
výkon	33.3	34.5	37.4	33.4	34.8	36.8	32.9	33.8	37.6

Modifikace	1	2	3	1	2	3	1	2	3
výkon	32.6	33.4	36.6	32.5	33.7	37	33	33.9	36.7

Rozhodněte, zda se výkon vozu liší významně v závislosti na továrně, modifikaci typu, popř. zda se projevují významně interakce obou faktorů.

10. Chí-kvadrát test

Řešené příklady

Příklad 10.1.

Nesymetričnost tvaru kostky či její nevyváženost může způsobit, že hra s touto kostkou není spravedlivá, to znamená, že se pravděpodobnosti padnutí jednotlivých stran liší. Spravedlivost byla studována následujícím způsobem. Kostkou se opakovaně házelo a zaznamenával se počet jedniček, dvojek, ..., šestek. Výsledky hodů jsou uvedeny v následující tabulce.

počet oček	1	2	3	4	5	6
četnost výskytu	979	1002	1015	980	1040	984

Je možné prokázat na základě provedených hodů, že hra s touto kostkou je nespravedlivá?

Řešení:

Počet hodů kostkou $n = 6000$. Pokud je kostka spravedlivá, mělo by každé z čísel $1, \ldots, 6$ padat s pravděpodobností $p_i = 1/6$, $i = 1, \ldots, 6$. Otestujme pomocí testu χ^2 dobré shody, zda se empirické četnosti n_i, $i = 1, \ldots, 6$, statisticky významně liší od teoretických četností np_i, $i = 1, \ldots, 6$.

počet oček	1	2	3	4	5	6
n_i	979	1002	1015	980	1040	984
np_i	1000	1000	1000	1000	1000	1000
$\frac{(n_i - np_i)^2}{np_i}$	0.441	0.004	0.225	0.400	1.600	0.256

Testová statistika $\sum_{i=1}^{6} \frac{(n_i - np_i)^2}{np_i} = 2.926$. Horní 5% kvantil χ^2 rozdělení o 5 stupních volnosti je roven 11.0705.

Odpověď: Vzhledem k tomu, že testová statistika není větší než 5 % kritická hodnota, tj. shora uvedený 5 % horní kvantil χ^2 rozdělení, nulovou hypotézu o vyváženosti kostky nezamítáme. Pro zajímavost uveďme ještě $p-$hodnotu testu, která je rovna 0.7114. Velmi vysoká $p-$hodnota při poměrně velkém počtu hodů nás ještě více přesvědčuje o tom, že kostka může být považována za spravedlivou.

Příklad 10.2.

V lékařské literatuře se uvádí, že v Evropě má 42% obyvatelstva krevní skupinu A, 10% skupinu B, 4% skupinu AB a 44% skupinu 0. V jedné transfúzní stanici zjistili, že během určitého období bylo mezi 350 dobrovolnými dárci 140 dárců s krevní skupinou A, 41 se skupinou B, 10 se skupinou AB a 159 se skupinou 0. Je mezi těmito počty a evropskými údaji statisticky významný rozdíl?

Řešení:

Otestujme testem χ^2 dobré shody, zda si odpovídají teoretické četnosti np_i, $i = 1, \ldots, 4$, spočtené z pravděpodobností p_i, $i = 1, \ldots 4$, uváděných v lékařské literatuře, a četnosti n_i, $i = 1, \ldots, 4$, zjištěné transfúzní stanicí.

n_i	140	41	10	159
np_i	147	35	14	154
$\frac{(n_i - np_i)^2}{np_i}$	0.333	1.029	1.143	0.162

Testová statistika $\sum \frac{(n_i - np_i)^2}{np_i}$ nabývá hodnoty 2.6671. Jestliže testujeme na hladině významnosti $\alpha = 0.05$, pak je třeba hodnotu 2.6671 srovnat s 5 % horním kvantilem χ^2-rozdělení o 3 stupních volnosti, který se rovná $\chi^2_{0.05}[3] = 7.8147$.

Odpověď: Vzhledem k tomu, že testová statistika nepřekročila kritickou hodnotu 7.8147, nedá se říci, že by mezi počty zjištěnými transfúzní stanicí a údaji uváděnými v lékařské literatuře byl statisticky významný rozdíl. Naopak, $p-$hodnota shora uvedeného χ^2 testu dobré shody je rovna 0.4458, což naznačuje poměrně dobrou shodu.

Příklad 10.3.

Kvalitní generátor náhodných čísel má generovat čísla, která mohou být považována za realizace nezávislých stejně rozdělených náhodných veličin, které mají požadované rozdělení. Abychom prověřili kvalitu generátoru náhodných čísel ze standardního normálního rozdělení, vygenerovali jsme jím 1000 čísel. Interval $(-2.5, 2.5)$ jsme rozdělili na intervaly o délce 0.5 a vygenerovaná čísla jsme do nich zařadili. Kromě toho byla 4 vygenerovaná čísla menší než -2.5 a 5 vygenerovaných čísel bylo větších než 2.5. Četnosti jednotlivých intervalů byly následující:

$(-\infty,-2.5)$	(-2.5,-2.0)	(-2.0,-1.5)	(-1.5,-1.0)	(-1.0,-0.5)	(-0.5,0.0)
4	18	30	93	156	210

(0.0,0.5)	(0.5,1.0)	(1.0,1.5)	(1.5,2.0)	(2.0,2.5)	$(2.5,\infty)$
165	164	103	44	8	5

Otestujte testem χ^2 dobré shody, zda čísla mohou být považována za výběr ze standardního normálního rozdělení.

Řešení:

Tabulka udává pro jednotlivé intervaly (a_{i-1}, a_i) skutečné četnosti n_i, teoretické četnosti $np_i = 1000\,(\Phi(a_i) - \Phi(a_{i-1}))$ a výrazy $(n_i - np_i)^2/np_i$, které měří shodu mezi skutečnými a teoretickými četnostmi.

interval	n_i	np_i	$(n_i - np_i)^2/np_i$
(-∞,-2.5)	4	6.2	0.7862
(-2.5,-2.0)	18	16.5	0.1288
(-2.0,-1.5)	30	44.1	4.4851
(-1.5,-1.0)	93	91.8	0.0145
(-1.0,-0.5)	156	149.9	0.2498
(-0.5, 0.0)	210	191.5	1.7945
(0.0, 0.5)	165	191.5	3.6578
(0.5, 1.0)	164	149.9	1.3300
(1.0, 1.5)	103	91.8	1.3542
(1.5, 2.0)	44	44.1	0.0001
(2.0, 2.5)	8	16.5	4.4097
(2.5, ∞)	5	6.2	0.2356

Statistika $\chi^2 \doteq 18.45$ a je menší než 5 % horní kvantil χ^2 rozdělení o 11 stupních volnosti, který se rovná 19.68.

Odpověď: Testem χ^2 dobré shody se neprokázalo, že by generátor náhodných čísel ze standardního normálního rozdělení pracoval špatně.

Příklad 10.4.

Zkoušky životnosti τ provedené u 200 žárovek daly následující výsledky:

τ (hod.)	0-300	301-600	601-900	901-1200	1201-1500	1501-1800
n_i	53	41	30	22	16	12

τ (hod.)	1801-2100	2101-2400	2401-2700	2701-3000	3001-3300	3301-více
n_i	9	7	5	3	2	0

Pomocí testu χ^2 dobré shody ověřte, zda je exponenciální rozdělení vhodným modelem pro rozdělení doby životnosti žárovky.

Řešení:

Exponenciální rozdělení je rozdělení s hustotou

$$f(x;\delta) = \begin{cases} \frac{1}{\delta}\,\mathrm{e}^{-x/\delta}, & \text{pro } x > 0\,; \\ 0, & \text{pro } x \leq 0\,. \end{cases}$$

Pravděpodobnost, že se náhodná veličina X, která se řídí exponenciálním rozdělením s parametrem δ, realizuje uvnitř intervalu (a, b), je dána:

$$P\,(X \in (a,b)) = \int_a^b \frac{1}{\delta}\, e^{-x/\delta}\, dx = e^{-a/\delta} - e^{-b/\delta}.$$

Parametr δ má význam střední doby životnosti a odhadne se jako:

$$\hat{\delta} = \frac{\sum n_i \xi_i}{n} = 871.5,$$

kde ξ_i označuje třídní znak (prostřední hodnotu daného intervalu), n_i četnost i-té třídy a n počet pozorování.

Následující tabulka udává skutečné a teoretické četnosti pro jednotlivé třídy a shodu mezi nimi. Všimněte si, že dvě a dvě poslední třídy byly spojeny tak, aby teoretické četnosti všech tříd byly větší než 5.

τ	0-300	300-600	600-900	900-1200	1200-1500
n_i	53	41	30	22	16
np_i	58.2474	41.2836	29.2603	20.7386	14.6988
$\frac{(n_i-np_i)^2}{np_i}$	0.4727	0.0019	0.0187	0.0767	0.1152

τ	1500-1800	1800-2100	2100-2400	2400-3000	3000-*více*
n_i	12	9	7	8	2
np_i	10.4179	7.3838	5.2334	6.3382	6.3979
$\frac{(n_i-np_i)^2}{np_i}$	0.2403	0.3537	0.5963	0.4357	3.0231

Testová statistika $\sum \frac{(n_i-np_i)^2}{np_i} = 5.3345$, což je mnohem menší hodnota než 5 % horní kvantil χ^2 rozdělení o 8 stupních volnosti $\chi^2_{0.05}[8] = 15.5073$. Počet stupňů volnosti je roven 8, neboť jsme pozorování rozdělili do 10 tříd a odhadovali jsme jeden parametr.

Odpověď: Tvrzení, že životnost žárovek má exponenciální rozdělení, nemůžeme zamítnout. Naopak vezmeme-li do úvahy vysokou p–hodnotu rovnající se 0.7213, zdá se, že exponenciální rozdělení je dobrým modelem pro naměřená data.

Příklad 10.5.

Je známo, že pravděpodobnost narození chlapce je 0.515, zatímco narození dívky je 0.485. Tvrdívá se, že pohlaví dětí v rodině se dvěma dětmi jsou nezávislá. To znamená, že pravděpodobnost, že druhé dítě bude holčička (resp. kluk), je rovna 0.485 (resp. 0.515) nezávisle na tom, jakého pohlaví je prvně narozené dítě. V rámci demografického průzkumu bylo zjišťováno pohlaví dětí 100 000 matek. Do výběru byly zařazeny jen ženy, které dvakrát rodily, a to při každém porodu pouze jedno dítě. Výsledky byly následující:

2 chlapci	2 dívky	různé pohlaví
27 130	23 820	49 050

Otestujte, zda lze pohlaví dětí v rodině se dvěma dětmi považovat za nezávislá.

Řešení:

Pokud jsou pohlaví nezávislá, pak pravděpodobnost narození dvou chlapců je rovna $0.515 \times 0.515 = 0.265225$, pravděpodobnost narození dvou dívek $0.485 \times 0.485 = 0.235225$ a pravděpodobnost narození párečku $2 \times 0.515 \times 0.485 = 0.49955$. Dále použijme test χ^2 dobré shody. Skutečné četnosti a teoretické četnosti jsou uvedeny v následující tabulce:

	2 chlapci	2 dívky	různé pohlaví
skutečné četnosti	27 130	23 820	49 050
teoretické četnosti	26 522.5	23 522.5	49 955

Testová statistika je rovna 34.073, což je mnohem větší hodnota než 5 % horní kvantil χ^2-rozdělení o 2 stupních volnosti, který se rovná 5.921.

Odpověď: Hypotézu, která tvrdí, že pohlaví dětí v rodině je nezávislé, zamítáme na 5 % hladině významnosti. Velmi malá p–hodnota, rovná $4 \cdot 10^{-8}$, nás velmi výrazně přesvědčuje o tom, že nulová hypotéza neplatí.

Neřešené příklady

Příklad 10.6.

Ke zjištění, zda jsou všechny tři dopravní pruhy dálnice stejně vytíženy, bylo provedeno sčítání počtu vozidel, která projela za určitou dobu v jednotlivých pruzích. Výsledky jsou dány v následující tabulce:

levý	střední	pravý
357	294	321

Jak byste pomocí testu χ^2 zhodnotili zjištěné údaje?

Příklad 10.7.

V následující tabulce jsou uvedeny počty živě narozených dětí v Československu v jednotlivých měsících roku 1957. Otestujte, zda se děti rodily během roku rovnoměrně nebo zda byly v počtech narozených dětí během roku nenáhodné výkyvy. (Poznamenejme, že rok 1957 byl nepřestupný.)

měsíc	I	II	III	IV	V	VI
počet dětí	21182	19960	22787	22805	23120	21859

měsíc	VII	VIII	IX	X	XI	XII
počet dětí	21367	20357	20946	20037	18728	19592

Příklad 10.8.

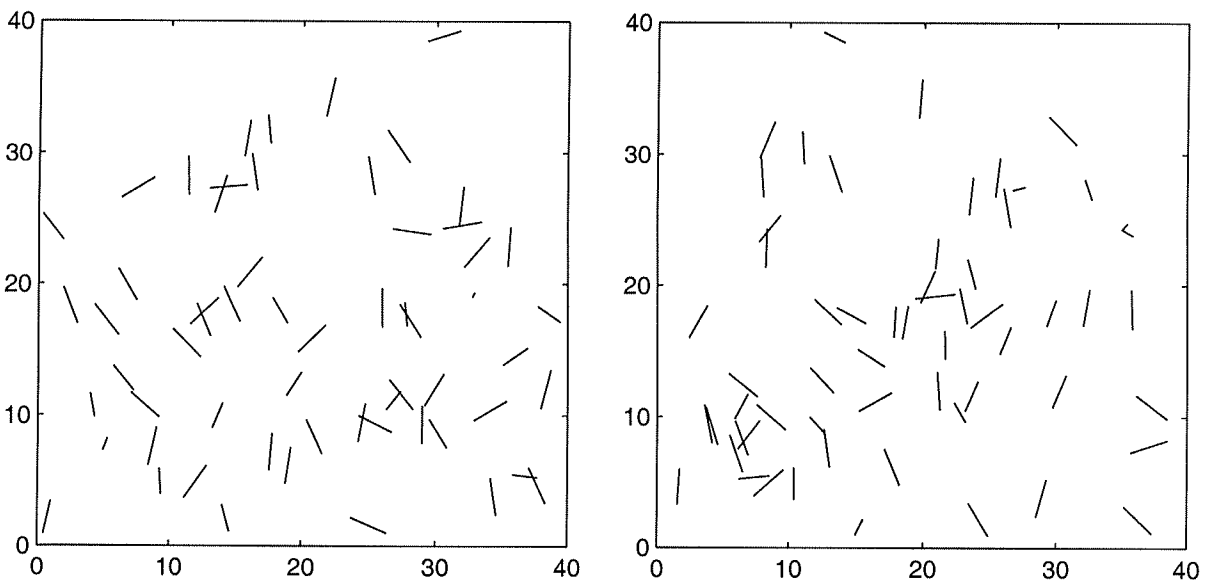
Mendelova teorie dědičnosti tvrdí, že četnosti kulatých a žlutých, hranatých a žlutých, kulatých a zelených a hranatých a zelených hrášků při křížení dvou druhů hrachu by měly být v poměru:

$$9 \ : \ 3 \ : \ 3 \ : \ 1.$$

Při ověřování své teorie Mendel získal četnosti: $315, 101, 108, 32$. Je možné na základě těchto dat zamítnout pomocí testu χ^2 Mendelovu teorii?

Příklad 10.9.

Za účelem zkoumání pevnosti železobetonu byly do záměsi betonu přimíchány ocelové tyčinky délky 3 cm. Z této záměsi byla po promíchání vyrobena zkušební krychle o hraně 40 cm. Máme k dispozici rentgenové snímky této krychle zepředu a z boku. Můžeme na základě těchto snímků soudit, že tyčinky byly v záměsi rovnoměrně rozmíchány?



Návod: Krychli je možné rozdělit vhodnými myšlenými řezy a použít test dobré shody. Je však třeba si uvědomit, že rozdělíme-li např. levý snímek na menší čtverce a porovnáme empirické a teoretické četnosti v těchto čtvercích, dostaneme informaci o tom, zda jsou tyčinky rovnoměrně rozmístěny právě a jen z tohoto pohledu, tj. zepředu. Totéž platí o testu dobré shody, provedenému pouze na základě pravého snímku. Nejsprávnější (ale poměrně pracné) je provést prostorové dělení např. na menší krychle, na základě porovnání obou snímků zařadit jednotlivé tyčinky do těchto krychlí a opět porovnat empirické a teoretické četnosti.

Příklad 10.10.

Střelec vypálil z pistole do 100 terčů po deseti výstřelech. Byl registrován počet zásahů v každém terči. Výsledky jsou uvedeny v následující tabulce:

počet zásahů	0	1	2	3	4	5	6	7	8	9	10
četnost	0	2	4	10	22	26	18	12	4	2	0

Ověřte pomocí testu χ^2 na hladině významnosti $\alpha = 0.05$, zda výsledky střelby odpovídají binomickému rozdělení.

Příklad 10.11.

V příkladě 7.8 je dáno skupinové rozdělení četností počtu vyzářených α-částic. Ověřte pomocí testu χ^2 dobré shody, zda se počet α-částic vyzářených během 7.5 sekundového intervalu opravdu řídí Poissonovým rozdělením.

Příklad 10.12.

Můžeme počty vozidel čekajících na zelenou v příkladu 6.6 považovat za výběr z Poissonova rozdělení? Ověřte pomocí testu χ^2 dobré shody.

Příklad 10.13.

Abychom ověřili, zda generátor náhodných čísel z rovnoměrného rozdělení na intervalu (0,1) opravdu generuje výběr z požadovaného rozdělení, vygenerovali jsme 1000 náhodných čísel. Interval (0,1) jsme rozdělili na 10 podintervalů o délce 0.1 a vygenerovaná čísla jsme do nich zařadili. Tabulka udává skupinové rozdělení četností:

(0,0.1)	(0.1,0.2)	(0.2,0.3)	(0.3,0.4)	(0.4,0.5)
99	91	92	97	99

(0.5,0.6)	(0.6,0.7)	(0.7,0.8)	(0.8,0.9)	(0.9,1)
108	103	102	110	99

Zjistěte, zda je možné na základě vygenerovaných náhodných čísel pomocí testu χ^2 dobré shody prokázat špatnou kvalitu generátoru.

Příklad 10.14.

Ověřte, zda předpoklad o normálním rozdělení obsahu látky v zemině pro data z příkladu 7.11 byl oprávněný. Použijte opět χ^2 test dobré shody.

Příklad 10.15.

V meteorologické stanici v Chebu byla v letech 1971 až 1980 zjišťována maximální měsíční rychlost větru v m/s. Z těchto $n = 120$ údajů byla sestavena tabulka skupinového rozdělení četností:

interval	⟨10.5-12.5)	⟨12.5-14.5)	⟨14.5-16.5)	⟨16.5-18.5)	⟨18.5-20.5)
četnost	5	14	23	20	16

⟨20.5-22.5)	⟨22.5-24.5)	⟨24.5-26.5)	⟨26.5-28.5)	⟨28.5-30.5)	⟨30.5-32.5)
13	11	7	6	3	2

Rozhodněte pomocí testu χ^2 dobré shody, zda tato data mohou být považována za výběr z dvouparametrického logaritmicko-normálního rozdělení.

Příklad 10.16.

Použijte data z příkladu 6.10 a pomocí testu χ^2 dobré shody rozhodněte, zda pro modelování ročních průměrných průtoků Lužnice v Bechyni je lepší normální nebo dvouparametrické logaritmicko-normální rozdělení.

Příklad 10.17.

Příklad 12.1 udává průměrné roční teploty měřené v Klementinu od roku 1807 do roku 1992. Otestujte testem χ^2 dobré shody, zda teploty v letech 1900 – 1992 mohou být považovány za výběr z normálního rozdělení. Zvolte mezi teplotou 7.3^0C a 12.3^0C nejprve deset a poté dvacet tříd. Třídy pospojujte tak, aby teoretická četnost každé třídy byla alespoň pět.

Příklad 10.18.

Tramvaj jezdí po dráze délky L. Při dopravním průzkumu byla zjišťována vzdálenost d mezi místem, v němž cestující nastoupí do tramvaje, od místa výjezdu tramvaje, vztažená k délce trasy L, tj. veličina $X = d/L$. Bylo zjištěno, že 114 cestujících nastoupilo dříve než v 1/4 trasy, 205 cestujích mezi 1/4 a 1/2 trasy, 115 cestujích mezi 1/2 a 3/4 trasy a 23 cestujích nastoupilo až za 3/4 trasy. Rozdělení veličiny X se obvykle modeluje rozdělením s hustotou

$$f(x;c) = \begin{cases} (c+1)(c+2)x(1-x)^c, & \text{pro } 0 \le x \le 1; \\ 0, & \text{jinde.} \end{cases}$$

Vyberte mezi rozděleními s hustotami $f(x;1.75)$, $f(x;2)$, $f(x;2.25)$, $f(x;2.5)$ rozdělení, které nejlépe odpovídá získaným datům. Jako kritérium použijte testovou statistiku, na které je založen test χ^2 dobré shody.

11. Regrese

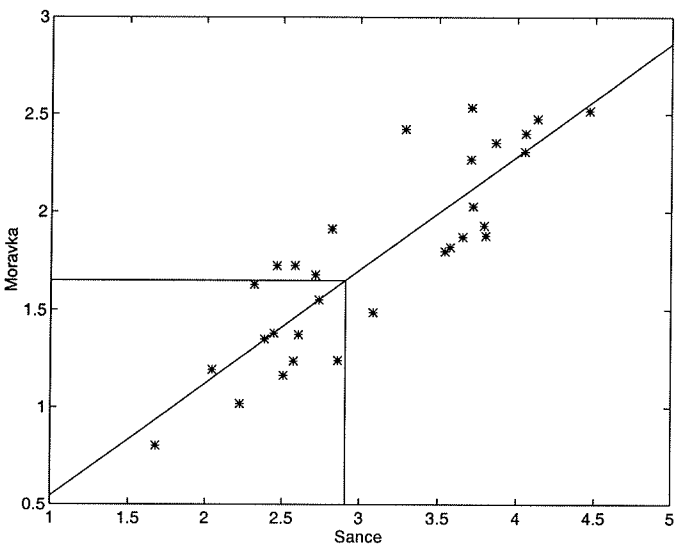
Řešené příklady

Příklad 11.1.

V letech 1931 - 1961 byly měřeny průtoky v profilu nádrže Šance na Ostravici a v profilu nádrže Morávka na Morávce. Roční průměry v m^3/s jsou dány v následující tabulce:

rok	Šance	Morávka	rok	Šance	Morávka
1931	4.130	2.476	1946	2.608	1.374
1932	2.386	1.352	1947	2.045	1.194
1933	2.576	1.238	1948	3.543	1.799
1934	2.466	1.725	1949	4.055	2.402
1935	3.576	1.820	1950	2.224	1.019
1936	2.822	1.913	1951	2.740	1.552
1937	3.863	2.354	1952	3.792	1.929
1938	3.706	2.268	1953	3.087	1.488
1939	3.710	2.534	1954	1.677	0.803
1940	4.049	2.308	1955	3.802	1.878
1941	4.466	2.517	1956	2.862	1.241
1942	2.584	1.726	1957	2.509	1.165
1943	2.318	1.631	1958	3.656	1.872
1944	3.721	2.028	1959	2.447	1.381
1945	3.290	2.423	1960	2.717	1.679

Spočtěte výběrový korelační koeficient. Předpokládejte, že v jednom z následujících let chybí hodnota průměrného ročního průtoku pro nádrž Morávka. V tomto roce činil průměrný roční průtok v profilu nádrže Šance na Ostravici 2.910 m^3/s. Jak bychom na základě lineární regrese odhadli hodnotu průměrného ročního průtoku v profilu nádrže Morávka?



Řešení:

Výběrový korelační koeficient $r = 0.871749$ je poměrně dost blízko hodnotě 1, což umožňuje za pomoci lineární regrese odhadnout chybějící hodnotu. Metodou nejmenších čtverců odhadneme posunutí a a směrnici b v lineární regresi. Odhad $\hat{a} = -0.035006$ a odhad $\hat{b} = 0.579481$. Optimální přímka vyjadřující vztah mezi průtoky v profilu nádrže Morávka a průtoky v profilu nádrže Šance je následující:

$$y = -0.035006 + 0.579481 \cdot x.$$

Dosadíme-li nyní $x = 2.91$, získáme hodnotu $y = 1.651$.

Odpověď: Na základě znalosti průtoků v profilu nádrže Morávka a Šance v letech 1931-1960, je možné odhadnout průměrný roční průtok v profilu nádrže Morávka v roce, kdy průtok v profilu Šance činil 2.91 m^3/s, hodnotou 1.651 m^3/s .

Příklad 11.2.

V ideálních podmínkách zvyšuje automobil rychlost v závislosti na čase t podle vztahu

$$v(t) = c_0 + c_1 t^2.$$

Působením nerovností terénu a jinými náhodnými vlivy je naměřená rychlost v_t v čase t rovna:

$$v_t = v(t) + e_t,$$

kde $\{e_t\}$ jsou náhodné chyby, o kterých budeme předpokládat, že jsou nezávislé stejně rozdělené s $N(0, \sigma^2)$ rozdělením. V půlsekundových intervalech byly měřeny rychlosti v km/hod zrychlujícího se automobilu. Výsledky měření jsou shrnuty v následující tabulce:

t	0	0.5	1	1.5	2	2.5	3
v_t	59.4048	62.6056	64.2801	68.4284	74.1422	80.8535	96.4670

Odhadněte bodově a intervalově, na kolik vzroste rychlost za následující 1 sekundu, pokud automobil bude zrychlovat i nadále stejným způsobem.

Řešení:

Rychlost auta roste lineárně v závislosti na nové proměnné x, která se rovná naměřenému času umocněnému na druhou, tj. $x = t^2$. Uvážíme-li ještě náhodné chyby, pak platí

$$v_i = c_0 + c_1 x_i + e_i, \quad i = 1, \ldots, 7.$$

Hodnoty $\{x_i, i = 1, \ldots, 7\}$ a $\{v_i, i = 1, \ldots, 7\}$ jsou uvedeny v tabulce:

x_i	0	0.25	1	2.25	4	6.25	9
v_i	59.4048	62.6056	64.2801	68.4284	74.1422	80.8535	96.4670

Jedná se tedy vlastně o lineární regresi s normálními chybami, kde c_0 a c_1 lze odhadnout metodou nejmenších čtverců:

$$\widehat{c_1} = \frac{\sum(x_i - \bar{x})v_i}{\sum(x_i - \bar{x})^2} \doteq 3.79357,$$
$$\widehat{c_0} = \bar{v} - \widehat{c_1}\bar{x} \doteq 59.9825.$$

Pro bodový odhad rychlosti v čase $t = 4$, tj. $x = 16$, platí

$$\widehat{v(4)} = 59.9825 + 3.79357 \times 16 \doteq 120.68.$$

Abychom našli predikční interval, je třeba odhadnout rozptyl náhodných chyb σ^2. Platí $\widehat{\sigma^2} = RSS/(n-2)$, kde residuální součet čtverců $RSS = (1-d)\sum(v_i - \bar{v})^2 = 17.975389$ ($d = r^2$ je koeficient determinace a r je korelační koeficient). Odtud $\hat{\sigma} = 1.89607$. Dále spočtěme výraz

$$\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i - \bar{x})^2}} = \sqrt{1 + \frac{1}{7} + \frac{(16 - 3.25)^2}{68.25}} = \sqrt{3.524725} = 1.877425.$$

$100\,(1-\alpha)\,\%$ predikční interval má tvar:

$$(120.68 - 1.877425 \times 1.89607\, t_{\alpha/2}[5],\ 120.68 + 1.877425 \times 1.89607\, t_{\alpha/2}[5]).$$

Odtud například 90 % predikční interval se rovná (113.51 , 127.85), neboť $t_{0.05}[5] = 2.015$.

Odpověď: Bude-li automobil zrychlovat i nadále stejným způsobem, odhadujeme, že za následující sekundu vzroste jeho rychlost na 120.68 km/hod. S 90% spolehlivostí se jeho skutečná rychlost bude pohybovat mezi 113.51 km/hod a 127.85 km/hod.

Příklad 11.3.

Ve dvou blízkých meteorologických stanicích A a B byla měřena každodenně teplota vzduchu. V jednom dni došlo ve stanici A k výpadku měření, a tudíž bylo třeba doplnit časovou řadu denních teplot pomocí naměřené hodnoty ze stanice B na základě metody lineární regrese. K použití lineární regrese opravňovalo zjištění, že výběrový korelační koeficient r mezi řadami měření ve stanici A a stanici B byl roven 0.9274. O naměřených údajích se dá předpokládat, že se řídí přibližně normálním rozdělením. Po určité době zjistili ve stanici B, že po celou dobu měl jejich měřící přístroj systematickou chybu, díky které měřil stále o $0.8\,^0$C více než byly správné hodnoty. Rozhodli se tedy snížit původně naměřená data o tuto hodnotu.

a) Je třeba přepočítat doplněnou hodnotu ve stanici A ?

b) Změní se původně spočítaný výběrový korelační koeficient r?

Matematicky zdůvodněte.

Řešení:

Použitím metody nejmenších čtverců je možno odhadnout koeficienty a a b v lineární regresi mezi naměřenými hodnotami ve stanici B, tj. hodnotami $\{X_i,\ i = 1, \ldots, n\}$, a odpovídajícími hodnotami naměřenými ve stanici A, tj. $\{Y_i,\ i = 1, \ldots, n\}$:

$$Y_i = a + b\,X_i$$

pomocí

$$\hat{b} = \frac{\sum(X_i - \bar{X})Y_i}{\sum(X_i - \bar{X})^2}, \quad \hat{a} = \bar{Y} - \hat{b}\,\bar{X}.$$

Označme indexem (N) nové hodnoty a indexem (ST) staré hodnoty. Zmenšíme-li naměřené hodnoty $\{X_i\}$ o 0.8, zmenší se i aritmetický průměr o tutéž hodnotu, tj. $\bar{X}^{(N)} = \bar{X}^{(ST)} - 0.8$. Odtud vyplývá, že odhad $\hat{b}$ zůstane stejný. Pro starý odhad koeficientu a platí

$$\hat{a}^{(ST)} = \bar{Y} - \hat{b}\,\bar{X}^{(ST)}$$

a pro nový platí

$$\hat{a}^{(N)} = \bar{Y} - \hat{b}\,\bar{X}^{(N)}.$$

Odtud zřejmě

$$\hat{a}^{(N)} = \hat{a}^{(ST)} + \hat{b}\,0.8.$$

Pro doplněnou hodnotu $\hat{Y}_{i_0}^{(N)}$ při použití nových dat platí

$$\hat{Y}_{i_0}^{(N)} = \hat{a}^{(N)} + \hat{b}\,X_{i_0}^{(N)} = \hat{a}^{(ST)} + \hat{b}\,0.8 + \hat{b}\,(X_{i_0}^{(ST)} - 0.8) = \hat{a}^{(ST)} + \hat{b}\,X_{i_0}^{(ST)} = \hat{Y}_{i_0}^{(ST)},$$

kde $\hat{Y}_{i_0}^{(ST)}$ je doplněná hodnota při použití původních dat. Znamená to, že doplněnou původní hodnotu není třeba měnit.

Výběrový korelační koeficient r je definován vztahem:

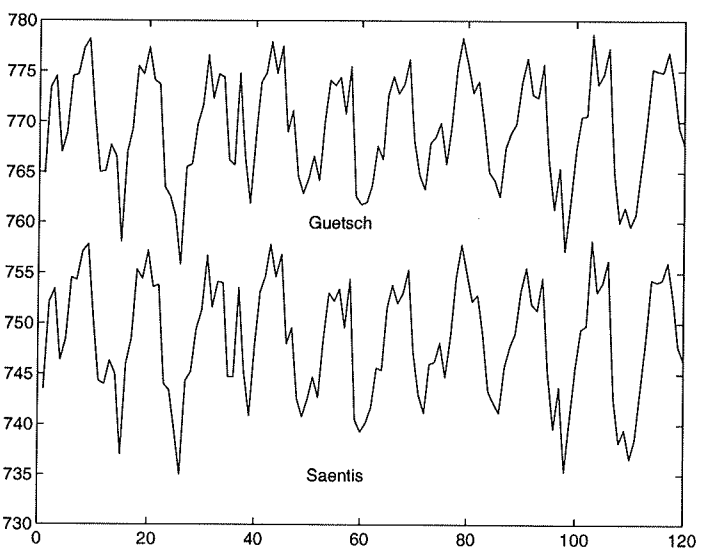
$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}}.$$

Vzhledem k tomu, že se rozdíly $\{(X_i - \bar{X}), i = 1, \ldots, n\}$ posunutím nezmění, zůstane také korelační koeficient stejný.

Odpověď: Doplněnou hodnotu není třeba přepočítávat a také výběrový korelační koeficient se nezmění.

Příklad 11.4.

Ve dvou švýcarských meteorologických stanicích Säntis a Gütsch jsou měřeny (kromě jiných meteorologických veličin) též výšky atmosferického tlaku v hPa. Obrázek znázorňuje průměrné měsíční hodnoty během 10 let.



Nechť veličiny $\{Y_i,\, i = 1,\ldots,120\}$ označují průměrné měsíční hodnoty naměřené ve stanici Säntis a veličiny $\{X_i,\, i = 1,\ldots,120\}$ označují průměrné měsíční hodnoty naměřené ve stanici Gütsch. Předpokládejme, že náhodné veličiny $\{Y_i,\, i = 1,\ldots,120\}$ za podmínky, že $\{X_i = x_i\}$, jsou nezávislé normálně rozdělené se střední hodnotou $\{a + b \cdot x_i\}$ a stejným rozptylem σ^2. Metodou nejmenších čtverců byly odhadnuty parametry a a b:

$$\hat{a} = -72.5939, \quad \hat{b} = 1.06693.$$

Dále byla odhadnuta směrodatná odchylka $\hat{\sigma} = \sqrt{RSS/(n-2)} = 0.536346$, kde RSS označuje residuální součet čtverců, a spočteny odhady směrodatných odchylek odhadů parametrů a a b:

$$s_{\hat{a}} = 6.96833, \quad s_{\hat{b}} = 0.00905526.$$

Otestujte nulovou hypotézu $H_0 : b = 1$ proti alternativě $A : b \neq 1$.

Řešení:

Poznamenejme, že nulová hypotéza H_0 znamená, že se v meteorologické stanici Gütsch měří (až na náhodné chyby měření) tatáž veličina jako ve stanici Säntis pouze posunutá díky různé nadmořské výšce obou stanic.

Předpoklad o podmíněném normálním rozdělení umožňuje použití $t-$testu, viz poznámka v článku 30, Jarušková (1996). Zamítací pravidlo má pro naši úlohu tvar:

$$\frac{|\hat{b} - 1|}{s_{\hat{b}}} > t_{\alpha/2}[n-2],$$

kde α je zvolená hladina významnosti. Zvolíme-li $\alpha = 0.05$, pak horní kvantil $t-$rozdělení $t_{\alpha/2}[n-2]$ je roven přibližně 1.98. Levá strana zamítacího pravidla je rovna přibližně 7.391.

Odpověď: Nulovou hypotézu, která tvrdí, že se časové řady ročních průměrných tlaků naměřených ve stanicích Säntis a Gütsch po odstranění náhodných chyb měření liší pouze tím, že jsou vůči sobě posunuty o konstantu, zamítáme.

Příklad 11.5.

Při hodnocení zkoušek na únavu lze popsat závislost počtu kmitů V do lomu na napětí x rovnicí

$$V = \delta \exp(\beta x)\, U,$$

kde U je náhodná chyba, o které se předpokládá, že má dvouparametrické logaritmicko-normální rozdělení $LN(\mu, \sigma^2)$, kde $\mu = 0$. Následující tabulka udává napětí $\{x_i\}$ v MPa a počet kmitů $\{v_i\}$ pro $i = 1,\ldots,16$:

x_i	560	560	560	560	580	580	580	580
$v_i \cdot 10^{-3}$	2845	3322	9411	14713	2597	4429	5523	6868

x_i	600	600	600	600	650	650	650	650
$v_i \cdot 10^{-3}$	554	1227	3446	3684	348	530	728	780

Najděte 95% interval spolehlivosti pro hodnotu $\xi = \delta \exp(\beta x)$ při napětí $x = 630$ MPa.

Řešení:

Zlogaritmováním předchozího vztahu získáme vztah mezi logaritmem počtu kmitů do lomu $Y = \ln V$ a napětím x v MPa:

$$Y = \alpha + \beta x + e,$$

kde $\alpha = \ln \delta$ a $e = \ln U$ je náhodná veličina normálně rozdělená s nulovou střední hodnotou a rozptylem rovným σ^2. Vzhledem k tomu, že zkoušky byly prováděny na různých vzorcích, je možné předpokládat, že chyby $\{e_i = \ln U_i, i - 1, \ldots, 16\}$ jsou nezávislé náhodné veličiny.

Metodou nejmenších čtverců odhadneme koeficienty α a β, přičemž $\hat{\beta} = -0.027294$ a $\hat{\alpha} = 30.945998$. Dále odhadneme rozptyl $\hat{\sigma^2} = RSS/14 = 0.407204$ a hodnotu logaritmu středního počtu kmitů do lomu při napětí $x = 630$:

$$\hat{Y}(630) = \hat{\alpha} + \hat{\beta} \cdot 630 = 13.750564.$$

95% interval spolehlivosti pro logaritmus počtu kmitů do lomu při napětí $x = 630$ MPa:

$$\hat{Y}(630) \mp t_{0.025}[n-2]\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}.$$

Dosazením získáme interval $(13.273, 14.228)$. Odlogaritmováním mezí tohoto intervalu dostaneme 95% interval spolehlivosti pro $\xi = \delta \exp(\beta x)$ při napětí $x = 630$ MPa.

Odpověď: 95% interval spolehlivosti pro $\xi = \delta \exp(\beta x)$ při napětí $x = 630$ MPa je následující:

$$(582 \cdot 10^3, 1500 \cdot 10^3).$$

Příklad 11.6.

Gumová podložka je zatěžována silou X a měří se její pokles Y. Ze zkušenosti víme, že vztah mezi X a Y je následující:

$$Y = a X + b X^2.$$

Díky náhodným chybám měření a dalším náhodným vlivům je však naměřený pokles podložky při určitém zatížení $\{x_i, i = 1, \ldots, n\}$ náhodná veličina $\{Y_i, i = 1, \ldots, n\}$, přičemž platí:

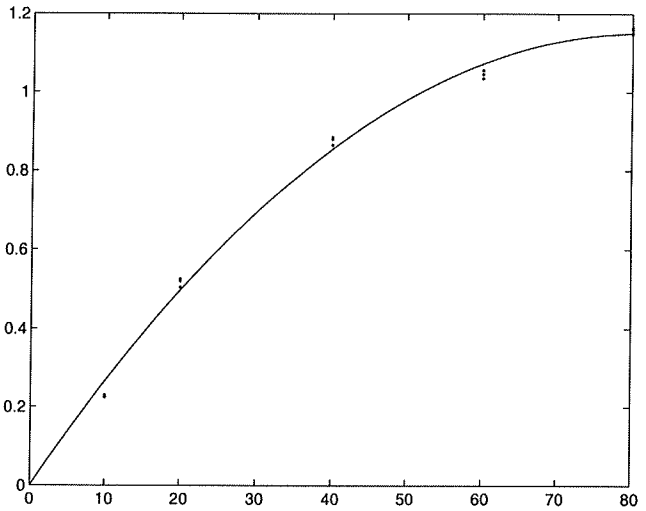
$$Y_i = a x_i + b x_i^2 + e_i, \quad i = 1, \ldots, n.$$

O náhodných chybách $\{e_i, i = 1, \ldots, n\}$ předpokládáme, že jsou to nezávislé stejně rozdělené náhodné veličiny s rozdělením $N(0, \sigma^2)$.

Podložku jsme zatížili postupně silami 10, 20, 40, 60 a 80 kN a při každém zatížení jsme provedli tři měření poklesu podložky. Výsledky jsou dány v tabulce:

x_i [kN]	10	20	40	60	80
y_i [mm]	0.225	0.505	0.880	1.045	1.150
	0.230	0.520	0.865	1.035	1.160
	0.225	0.525	0.885	1.055	1.165

Data včetně uvažovaného modelu jsou znázorněna na následujícím obrázku.



Odhadněte bodově a intervalově průměrný pokles podložky, zatížíme-li ji silou 30, 50 a 70 kN.

Řešení:

Pro zjištění závislosti mezi poklesem podložky Y a zatížením X jsme provedli $n = 15$ měření. Odhad koeficientů a a b metodou nejmenších čtverců je řešením soustavy normálních rovnic:

$$\hat{a} \sum_{i=1}^{n} x_i^2 + \hat{b} \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} x_i\, y_i$$
$$\hat{a} \sum_{i=1}^{n} x_i^3 + \hat{b} \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} x_i^2\, y_i.$$

Maticově se dá tatáž soustava zapsat také takto:

$$(X'X)\,\hat{\beta} = X'Y,$$

kde

$$X' = \begin{pmatrix} 10 & 10 & 10 & 20 & 20 & \dots & 80 \\ 100 & 100 & 100 & 400 & 400 & \dots & 6400 \end{pmatrix},$$

$$Y = (0.225 \quad 0.230 \quad 0.225 \quad 0.505 \quad 0.520 \quad \dots \quad 1.165)',$$

$$\hat{\beta} = (\hat{a}, \quad \hat{b})'.$$

Inverzní matice $(X'X)^{-1}$ k matici soustavy $X'X$ je zde až na chyby vzniklé zaokrouhlováním rovna

$$(X'X)^{-1} = \begin{pmatrix} 430.49729 \cdot 10^{-6} & -6.08699 \cdot 10^{-6} \\ -6.08699 \cdot 10^{-6} & 0.09195 \cdot 10^{-6} \end{pmatrix}$$

a vektor pravých stran

$$X'Y = \begin{pmatrix} 609.1 \\ 38422.0 \end{pmatrix}.$$

Vyřešením soustavy získáme odhady: $\hat{a} = 0.0283413$, $\hat{b} = -0.00017465$.

Bodově můžeme odhadnout průměrný pokles podložky při zatížení $x_0 = 30$ kN hodnotou:

$$\hat{Y}(30) = \hat{a} \cdot 30 + \hat{b} \cdot 30^2 = 0.0283413 \times 30 - 0.00017465 \times 900 \doteq 0.693 \quad [\text{mm}].$$

Obdobně průměrný pokles podložky při zatížení $x_0 = 50$ kN lze odhadnout hodnotou $\hat{Y}(50) = 0.980$ mm a průměrný pokles podložky při zatížení $x_0 = 70$ kN hodnotou $\hat{Y}(70) = 1.128$ mm.

Dále najděme intervaly spolehlivosti pro průměrný pokles podložky při hodnotách zatížení $x_0 = 30, 50, 70$ kN. Intervaly spolehlivosti nám dávají představu o tom, jak přesně jsme průměrné poklesy podložky při jednotlivých zatíženích odhadli. Abychom ohodnotili velikost náhodných chyb, musíme odhadnout směrodatnou odchylku σ. K odhadu směrodatné odchylky je třeba znát residuální součet čtverců RSS:

$$RSS = Y'Y - Y'X\hat{\beta} = 0.0104492393.$$

Pro odhad směrodatné odchylky platí:

$$\hat{\sigma} = \sqrt{RSS/(n-2)} = \sqrt{0.0104492393/13} = 0.028351.$$

Zaveďme pro zatížení x_0 vektor X_0 předpisem $X_0 = (x_0\,,\ x_0^2)$. $100\,(1-\alpha)\%$ interval spolehlivosti pro průměrný pokles podložky $\hat{Y}(x_0)$ při zatížení x_0 je dán následovně:

$$\left(\hat{Y}(x_0) \mp t_{\alpha/2}[n-2]\,\hat{\sigma}\,\sqrt{X_0\,(X'X)^{-1}\,X_0'}\right).$$

Odtud například plyne, že 99% interval spolehlivosti pro průměrný pokles podložky při zatížení $x_0 = 30$ kN je roven:

$$(0.693 \mp 3.012 \times 0.028351 \times 0.365007) = (0.693 - 0.031, 0.693 + 0.031),$$

neboť $t_{0.005}[13] = 3.012$. Obdobně 99% interval spolehlivosti pro průměrný pokles podložky při zatížení $x_0 = 50$ kN je roven $(0.980 - 0.031, 0.980 + 0.031)$ a pro $x_0 = 70$ kN je roven $(1.128 - 0.032, 1.128 + 0.032)$.

Poznamenejme, že odhady $\hat{a}$ a $\hat{b}$ jsou rovněž řešením jednodušší lineární soustavy:

$$\hat{a} \sum_{j=1}^{5} z_j^2 + \hat{b} \sum_{j=1}^{5} z_j^3 = \sum_{j=1}^{5} z_j\,\bar{y}_j,$$

$$\hat{a} \sum_{j=1}^{5} z_j^3 + \hat{b} \sum_{j=1}^{5} z_j^4 = \sum_{j=1}^{5} z_j^2\,\bar{y}_j,$$

kde $z_1 = 10$, $z_2 = 20$, $z_3 = 40$, $z_4 = 60$, $z_5 = 80$ a $\bar{y}_1$ je průměr prvního sloupce, tj. $\bar{y}_1 = (y_1 + y_2 + y_3)/3$, $\bar{y}_2$ je průměr druhého sloupce atd. Tvrzení je důsledkem toho, že

$$\sum_{i=1}^{15} x_i^2 = 3 \sum_{j=1}^{5} z_j^2, \qquad \sum_{i=1}^{15} x_i^3 = 3 \sum_{j=1}^{5} z_j^3 \qquad \text{a} \qquad \sum_{i=1}^{15} x_i^4 = 3 \sum_{j=1}^{5} z_j^4.$$

Dále

$$\sum_{i=1}^{15} x_i\, y_i = 3 \sum_{j=1}^{5} z_j\, \bar{y}_j, \qquad \sum_{i=1}^{15} x_i^2\, y_i = 3 \sum_{j=1}^{5} z_j^2\, \bar{y}_j.$$

Soustavy jsou až na násobek tří stejné, a tudíž mají i stejné řešení. Chceme-li bodově odhadnout průměrný pokles podložky při nějakém zatížení x_0, stačí uvažovat zjednodušený problém, kde data pro odhad lze shrnout do tabulky:

z_j [kN]	10	20	40	60	80
$\bar{y}_j$ [mm]	0.226667	0.516667	0.876667	1.045000	1.158333

Chceme-li však získat intervalový odhad, pak je třeba vzít v úvahu, že počet měření je roven 15, neboť se vzrůstajícím počtem měření vzrůstá přesnost odhadu, to znamená, že se intervaly spolehlivosti zmenšují.

Odpověď: Průměrný pokles podložky při zatížení $x_0 = 30$ kN jsme odhadli hodnotou 0.693 mm, při zatížení $x_0 = 50$ kN hodnotou 0.980 mm a při zatížení $x_0 = 70$ kN hodnotou 1.128 mm. 99% intervaly spolehlivosti pro průměrný pokles podložky při zatížení $x_0 = 30, 50, 70$ kN jsou postupně rovny:

$$(0.662, 0.724), \quad (0.949, 1.011), \quad (1.096, 1.160).$$

Příklad 11.7.

Podle geodetických zkušeností jsou relativní vlhkost H daná v %, tlak P v mmHg a teplota vzduchu T v ^{0}C veličinami, které ovlivňují výsledek měření trigonometrické výšky Y. Obvykle se předpokládá lineární model

$$Y_i = a + b\,H_i + c\,P_i + d\,T_i + e_i,$$

kde $\{e_i\}$ jsou náhodné chyby měření. Předpokládáme, že veličiny $\{Y_i\}$ při známých hodnotách $\{H_i = h_i\}$, $\{P_i = p_i\}$ a $\{T_i = t_i\}$ jsou nezávislé mající normální rozdělení se středními hodnotami $\{a + b\,h_i + c\,p_i + d\,t_i\}$ a stejným rozptylem σ_e^2. Při opakovaném měření jsme zjistili tyto hodnoty:

Y_i	H_i	P_i	T_i
135.8003	56	753	22.0
135.8000	56	753	23.0
135.7990	50	752	25.0
135.7998	45	751	26.0
135.7997	41	752	26.5
135.8008	40	751	24.0
135.8023	54	753	22.0
135.8017	65	753	20.0
135.8032	66	754	18.0
135.8055	68	751	17.0
135.8064	80	753	13.0
135.8053	72	753	17.0
135.8043	68	754	19.0

Zjistěte použitím metody testování hypotéz, zda je možné model zjednodušit.

Řešení:

Popsaný problém je problém lineární regrese se třemi vysvětlujícími proměnnými (regresory) - relativní vlhkostí H, tlakem P a teplotou T. Nejprve odhadněme metodou nejmenších čtverců koeficienty a, b, c, d. Odhady jsou řešením lineární soustavy rovnic:

$$(X'X)\hat{\beta} = X'Y,$$

kde

$$X = \begin{pmatrix} 1 & H_1 & P_1 & T_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & H_{13} & P_{13} & T_{13} \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{13} \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \\ \hat{d} \end{pmatrix}.$$

V našem případě jsou odhady následující: $\hat{a} = 136.029316$, $\hat{b} = -0.000032$, $\hat{c} = -0.00028$, $\hat{d} = -0.000712$. Odhadnutý model má tvar

$$\hat{Y} = 136.029316 - 0.000032\,H - 0.00028\,P - 0.000712\,T.$$

Odhadněme ještě směrodatnou odchylku chyb měření σ_e, která nám dává představu o tom, jak velké jsou chyby měření. Platí $\hat{\sigma_e} = \sqrt{RSS/(n-4)}$, kde $RSS = Y'Y - Y'X\hat{\beta} = 7.857219 \cdot 10^{-6}$. Odtud $\hat{\sigma_e} = 9.343577 \cdot 10^{-4}$.

Nyní otestujme hypotézy

$$\begin{aligned} H_b &: b = 0 \quad \text{proti} \quad A_b : b \neq 0, \\ H_c &: c = 0 \quad \text{proti} \quad A_c : c \neq 0, \\ H_d &: d = 0 \quad \text{proti} \quad A_d : d \neq 0. \end{aligned}$$

Pokud se nám některá nulová hypotéza nepodaří zamítnout, zdá se být rozumné odpovídající proměnnou vyloučit z modelu, protože přispívá jen zanedbatelně k vysvětlení

závisle proměnné Y. Zvolíme-li hladinu významnosti $\alpha = 0.05$, pak má zamítací pravidlo pro testování H_b proti A_b tvar:

$$\frac{|\hat{b}|}{s_{\hat{b}}} > t_{0.025}[9], \quad \text{kde} \quad s_{\hat{b}} = \hat{\sigma}_e\sqrt{v_{22}}.$$

Obdobně zamítací pravidlo pro testování H_c proti A_c má tvar:

$$\frac{|\hat{c}|}{s_{\hat{c}}} > t_{0.025}[9], \quad \text{kde} \quad s_{\hat{c}} = \hat{\sigma}_e\sqrt{v_{33}}$$

a pravidlo pro testování H_d proti A_d má tvar:

$$\frac{|\hat{d}|}{s_{\hat{d}}} > t_{0.025}[9], \quad \text{kde} \quad s_{\hat{d}} = \hat{\sigma}_e\sqrt{v_{44}}.$$

Hodnoty v_{22}, v_{33}, v_{44} jsou prvky diagonály symetrické matice $(X'X)^{-1}$. V našem případě

$$(X'X)^{-1} = \begin{pmatrix} 70415.3560 & 11.8627 & -95.1551 & 23.7658 \\ 11.8627 & 8.4619 \cdot 10^{-3} & -1.7060 \cdot 10^{-2} & 2.2915 \cdot 10^{-2} \\ -95.1551 & -1.7060 \cdot 10^{-2} & 1.2875 \cdot 10^{-1} & -3.5269 \cdot 10^{-2} \\ 23.7658 & 2.2915 \cdot 10^{-2} & -3.5269 \cdot 10^{-2} & 6.8418 \cdot 10^{-2} \end{pmatrix}$$

Odtud $v_{22} = 8.4619 \cdot 10^{-3}$, $v_{33} = 1.2875 \cdot 10^{-1}$, $v_{44} = 6.8418 \cdot 10^{-2}$, a tedy $s_{\hat{b}} = 0.000086$, $s_{\hat{c}} = 0.000335$, $s_{\hat{d}} = 0.000244$ a $|\hat{b}|/s_{\hat{b}} = 0.3710$, $|\hat{c}|/s_{\hat{c}} = 0.8338$, $|\hat{d}|/s_{\hat{d}} = 2.9113$. Kvantil $t_{0.025}[9] = 2.262$. Po dosazení do všech třech zamítacích pravidel zjišťujeme, že na hladině významnosti $\alpha = 0.05$ se nulová hypotéza H_d zamítá ve prospěch alternativy A_d. Příslušná p–hodnota je rovna 0.017. Nulové hypotézy H_b a H_c se nezamítají. Příslušné p–hodnoty jsou rovny $p_b = 0.7192$ a $p_c = 0.4260$. Zdá se tedy, že ani vlhkost ani tlak nemají vliv na naměřenou trigonometrickou výšku. Pro jistotu však zkusme nejprve z modelu vyloučit proměnnou s největší p–hodnotou, tj. v našem případě vlhkost. Takovému postupnému zpětnému vylučování se říká anglicky „stepwise backward selection".

Uvažujeme model:

$$Y_i = f + g\,P_i + h\,T_i + \epsilon_i, \quad i = 1, \ldots, n,$$

kde opět předpokládáme, že veličiny $\{\epsilon_i\}$ shrnují chyby měření i další náhodné vlivy. Uděláme-li podobný předpoklad o podmíněném normálním rozdělením veličin $\{Y_i, i = 1, \ldots, n\}$, poskytuje metoda nejmenších čtverců nejlepší odhady parametrů f, g a h:

$$\hat{f} = 136.074017, \quad \hat{g} = -0.000344, \quad \hat{h} = -0.000625$$

a odhad směrodatné odchylky $\hat{\sigma}_\epsilon = \sqrt{RSS/(n-3)} = 0.000893$. Obdobně jako v předchozím zjistíme $s_{\hat{f}} = 0.207139$, $s_{\hat{g}} = 0.000274$, $s_{\hat{h}} = 0.000071$. Dále testujme hypotézy

$$H_g : g = 0 \quad \text{proti} \quad A_g : g \neq 0,$$
$$H_h : h = 0 \quad \text{proti} \quad A_h : h \neq 0.$$

Dosazením do zamítacích pravidel

$$\frac{|\hat{g}|}{s_{\hat{g}}} > t_{0.025}[n-3], \quad \frac{|\hat{h}|}{s_{\hat{h}}} > t_{0.025}[n-3],$$

zjistíme, že na hladině $\alpha = 0.05$ nulovou hypotézu H_h zamítáme, ale H_g nikoliv. Odtud vyplývá, že proměnnou P (tlak) opravdu můžeme vyloučit z modelu. Zjednodušený model má tvar:

$$Y_i = \alpha + \beta T_i + v_i,\ i = 1, \ldots, n,$$

kde veličiny $\{v_i\}$ opět vyjadřují náhodné vlivy. Za předpokladu, že veličiny $\{Y_i\}$ při známých hodnotách teplot $\{T_i = t_i,\ i = 1, \ldots, n\}$ jsou nezávislé normálně rozdělené se stejným rozptylem σ_v^2 a středními hodnotami $\{\alpha + \beta t_i,\ i = 1, \ldots, n\}$, můžeme parametry opět odhadnout metodou nejmenších čtverců a odhady regresních parametrů budou mít opět normální rozdělení. Pro odhady regresních parametrů α a β platí $\hat{\alpha} = 135.814$ a $\hat{\beta} = -0.000585339$ a odhad směrodatné odchylky $\hat{\sigma_v} = 0.000916027$. Koeficient determinace $d = 87.93\%$.

Pro zajímavost ještě uveďme, že kdybychom uvažovali závislost trigonometrické výšky Y na relativní vlhkosti H:

$$Y_i = \gamma + \delta H_i + w_i,\ i = 1, \ldots, n,$$

pak by test nulové hypotézy $H_\delta : \delta = 0$ proti alternativě $A_\delta : \delta \neq 0$ nulovou hypotézu zamítl, přičemž příslušná p–hodnota by byla velmi nízká, rovná 0.00013. Koeficient determinace $d = 75.11\%$. Odtud by vyplývalo, že hodnota trigonometrické výšky závisí výrazně na hodnotě relativní vlhkosti. My jsme však relativní vlhkost z předchozího modelu vyloučili. Jak je to možné? Důvod spočívá v tom, že výběrový korelační koeficient mezi teplotou a relativní vlhkostí je velmi blízký hodnotě -1, konkrétně $r = -0.9557$. To znamená, že relativní vlhkost je téměř lineární funkcí teploty. Pro náš lineární model závislosti trigonometrické výšky Y na regresorech H, P a T to znamená, že teplota a relativní vlhkost mají na změřenou výšku téměř shodný vliv. Jinými slovy do regresního modelu by se měla zařadit buď teplota nebo relativní vlhkost, nikoliv však obě proměnné společně. Protože však teplota vysvětlovala naměřenou výšku přece jen o malinko lépe, rozhodli jsme vytvořit výsledný model na jejím podkladě. V teorii lineární regrese s více proměnnými se případu, kdy jsou sloupce matice X téměř lineárně závislé říká, kolinearita. Kolinearita se projevuje špatnou podmíněností matice $X'X$, a tudíž špatnou stabilitou řešení soustavy normálních rovnic.

Nakonec ještě poznamenejme, že kvůli přesnějšímu numerickému výpočtu odhadů koeficientů v lineární regresi s více proměnnými se doporučuje regresory standardizovat odečtením průměrů a vydělením příslušnými směrodatnými odchylkami. V našem případě by to znamenalo, že bychom vycházeli z modelu:

$$Y_i = \tilde{a} + \tilde{b}\frac{H_i - \bar{H}}{s_H} + \tilde{c}\frac{P_i - \bar{P}}{s_P} + \tilde{d}\frac{T_i - \bar{T}}{s_T} + e_i,$$

kde $\bar{H} = \sum H_i/n$, $s_H = \sqrt{\sum \left(H_i - \bar{H}\right)^2/n}$ a podobně pro zbývající regresory.

Odpověď: Původní model je možno zjednodušit na model obsahující pouze jednu nezávislou proměnnou (teplotu T):

$$Y_i = 135.814 - 0.000585339\,T_i + v_i, \quad i = 1, \ldots, n,$$

kde veličiny $\{v_i\}$ vyjadřují náhodné chyby.

Neřešené příklady

Příklad 11.8.

Pozorováním jsme získali hodnoty $\{x_i,\ i = 1 \ldots, 5\}$ a $\{y_i,\ i = 1, \ldots, 5\}$:

x_i	1	2	3	4	5
y_i	1.2875	1.9626	2.2643	1.4113	3.3631

Rozhodněte, zda je lepší model

$$y_i = a + b\,x_i, \quad i = 1, \ldots, 5$$

nebo

$$y_i = a + b\,x_i^2, \quad i = 1, \ldots, 5,$$

jestliže kritériem je součet čtverců residuí.

Příklad 11.9.

Naměřili jsme realizace (x_1, y_1), ..., (x_n, y_n) náhodných vektorů (X_1, Y_1), ..., (X_n, Y_n). Uvažujeme dvě závislosti

$$Y_i = a + b\,X_i, \quad i = 1, \ldots, n,$$
$$X_i = c + d\,Y_i, \quad i = 1, \ldots, n.$$

Pro oba typy závislostí použijeme metodu nejmenších čtverců k odhadu a a b, respektive c a d. Budou pro oba případy shodné koeficienty determinace? Budou pro oba případy shodné residuální součty čtverců?

Příklad 11.10.

Uvažujme lineární regresi mezi výškou člověka X měřenou v m a jeho váhou Y vyjádřenou v kg:

$$Y_i = a + b\,X_i + e_i, \quad i = 1, \ldots, n,$$

Veličiny $\{e_i,\ i = 1, \ldots, n\}$ vyjadřují náhodné chyby. Označme $\hat{a}$ a $\hat{b}$ odhady parametrů a a b získané metodou nejmenších čtverců a RSS_{yx} odpovídající residuální součet čtverců. Předpokládejme, že vstupní data je nutno přepočítat tak, že výška člověka U bude udána v cm, tj. $U_i = 100\,X_i, i = 1, \ldots, n$. Lineární regrese nyní bude mít tvar

$$Y_i = c + d\,U_i + v_i, \quad i = 1, \ldots, n,$$

kde $\{v_i,\ i = 1, \ldots, n\}$ vyjadřují náhodné chyby. Označme $\hat{c}$ a $\hat{d}$ odhady parametrů c a d metodou nejmenších čtverců a RSS_{yu} příslušný residuální součet čtverců. Jaký vztah bude mezi $\hat{c}$, $\hat{d}$ a $\hat{a}$, $\hat{b}$ a jaký mezi RSS_{yx} a RSS_{yu}?

Příklad 11.11.

Uvažujme lineární závislost mezi množstvím srážek $\{X_i\}$ udávaném v mm a průtoky v řece $\{Y_i\}$ měřené v l/s:

$$Y_i = a + b\,X_i + e_i, \quad i = 1, \ldots, n,$$

kde $\{e_i\}$ označují náhodné chyby. Předpokládejme, že $\hat{a}$ a $\hat{b}$ jsou odhady parametrů a a b metodou nejmenších čtverců a RSS_{yx} residuální součet čtverců. Jestliže průtoky $\{Y_i\}$ v l/s přepočteme na průtoky $\{U_i\}$ v m^3/s, změní se lineární vztah následovně:

$$U_i = c + d\,X_i + v_i, \quad i = 1, \ldots, n,$$

kde $\{v_i\}$ označují opět náhodné chyby. Jestliže $\hat{c}$ a $\hat{d}$ označují odhady parametrů c a d metodou nejmenších čtverců a RSS_{ux} příslušný residuální součet čtverců, jaký je vztah mezi $\hat{a}, \hat{b}$ a $\hat{c}, \hat{d}$ a jaký mezi RSS_{yx} a RSS_{ux}?

Příklad 11.12.

Předpokládejme, že prodloužení pružiny Δl závisí na zatěžovací síle F v rozmezí $\langle 0, 1\rangle$ lineárně:

$$\Delta l = k\,F.$$

Aby se při odhadu konstanty k omezil vliv náhodných chyb měření $\{e_i\}$, o kterých se předpokládá, že jsou to nezávislé náhodné veličiny splňující $\mathrm{E}\,e_i = 0, \mathrm{Var}\,e_i = \sigma^2$, měření se několikrát opakuje. Bylo rozhodnuto provést 10 měření. Získané hodnoty $(F_1, \Delta l_1), \ldots, (F_{10}, \Delta l_{10})$ splňují

$$\Delta l_i = k\,F_i + e_i, \quad i = 1, \ldots, 10.$$

Jak volit hodnoty $F_1, \ldots, F_{10} \in \langle 0, 1\rangle$, aby odhad $\hat{k}$ získaný metodou nejmenších čtverců byl nejlepší, tj. měl nejmenší rozptyl? Jaký bude mít při optimální volbě $\{F_i\}$ odhad $\hat{k}$ tvar?

Příklad 11.13.

Měděná trubka má délku $L_0 = 1000$ mm při teplotě $t_0 = 20\,^0\mathrm{C}$. Bylo měřeno, o kolik se tato trubka prodlouží, stoupne-li teplota o $\Delta t\,^0\mathrm{C}$.

Δt	10	20	30	40	50	60
ΔL	0.18	0.35	0.48	0.65	0.84	0.97

Je známo, že pro délkovou roztažnost platí vzorec

$$\Delta L = \alpha\,L_0\,\Delta t,$$

kde α je tzv. koeficient roztažnosti. Tento koeficient je třeba odhadnout na základě naměřených hodnot a stanovit pro něj 95% interval spolehlivosti.

Příklad 11.14.

Uvažujte data z příkladu 8.25 a předpokládejte, že pevnosti betonu zjištěné destruktivní zkouškou d_i jsou při daných hodnotách výsledků nedestruktivních zkoušek n_i nezávislé normálně rozdělené náhodné veličiny se středními hodnotami $a + b \cdot n_i$ a stejnými rozptyly σ^2. Odhadněte parametry a a b a otestujte, liší-li se a významně od nuly. Otestujte dále, liší-li se výsledky destruktivních a nedestruktivních zkoušek po odstranění náhodných chyb pouze posunutím. Najděte 80% predikční interval pro výsledek destruktivní zkoušky, zjistíme-li nedestruktivní metodou pevnost 58.5 MPa.

Příklad 11.15.

Při kontrolních měřeních rozměrů silikátových štítových dílců bylo náhodně vybráno 8 dílců vykazujících vesměs kladné odchylky v délce i výšce od normových hodnot:

odchylka délky [mm]	3	4	4	5	8	10	6	3
odchylka výšky [mm]	4	6	5	6	7	13	9	4

Najděte lineární regresní model závislosti odchylky výšky na odchylce délky. Otestujte hypotézu, že regresní přímka prochází počátkem. (Předpokládejte normální rozdělení chyb.)

Příklad 11.16.

Uvažujte data z příkladu 8.26 a předpokládejte, že provozní výkony motoru jsou přibližně lineární funkcí zkušebních výkonů a že chyby tohoto vyjádření jsou nezávislé normálně rozdělené náhodné veličiny se stejným rozptylem. Najděte tento lineární regresní model a otestujte, je-li možno přejít k jednoduššímu modelu, kde regresní přímka prochází počátkem. Pokud ano, najděte směrnici této regresní přímky.

Příklad 11.17.

Desetkrát bylo změřeno ojetí vzorku pneumatiky O a brzdná dráha s při jinak stejných podmínkách (typ automobilu, jeho rychlost, atd).

O [mm]	0.2	0.35	0.4	0.15	0.4	0.25	0.3	0.35	0.25	0.3
s [m]	8.7	8.9	9.2	8.6	9.1	8.7	8.8	9	8.6	8.9

Předpokládejte, že závislost brzdné dráhy na ojetí vzorku je přibližně lineární. Najděte příslušný regresní model, posuďte jeho kvalitu. Předpokládejte dále, že brzdné dráhy s_i při daném ojetí O_i jsou nezávislé normálně rozdělené náhodné veličiny se stejným rozptylem a najděte bodový a 80% intervalový odhad pro brzdnou dráhu při daném ojetí vzorku 0.32 mm.

Příklad 11.18.

V letech 1989–1994 došlo k výraznému zvýšení počtu automobilů v České republice a zároveň vzrostl i počet registrovaných automobilových nehod:

rok	počet aut	počet nehod
1989	2330755	79717
1990	2411297	94664
1991	2483222	101387
1992	2580297	125599
1993	2746995	152157
1994	2900000	156242

Předpokládejme, že vztah mezi počtem nehod N a počtem aut A je přibližně lineární. Odhadněte bodově a intervalově průměrný nárůst počtu nehod při zvýšení počtu automobilů o 100 000. Předpokládejte podobně jako v příkladu 11.4., že počty nehod $\{N_i\}$ při daných počtech aut $\{A_i = a_i\}$ tvoří nezávislé náhodné veličiny mající normální rozdělení se středními hodnotami $\{b_0 + b_1\, a_i\}$ a stejným rozptylem σ^2.

Příklad 11.19.

Závislost pevnosti betonu v tahu Y na době zrání T je přibližně vyjádřena funkcí

$$Y = a \cdot b^{1/T}.$$

U 10 vzorků betonu z jedné záměsi byla po různém počtu dnů T zjišťována pevnost v tahu Y v $\mathrm{kp} \cdot \mathrm{cm}^{-2}$:

T	2	2	3	3	7	7	28	28	28	28
Y	21.9	24.5	29.8	24.2	34.5	33.1	41.8	35.7	40.3	37.3

Zlogaritmováním výše uvedeného vztahu a následným použitím metody nejmenších čtverců najděte odhady parametrů a a b. Odhadněte pevnost betonu v tahu po 40 dnech.

Příklad 11.20.

Z fyziky je známo, že barometrický tlak P závisí na nadmořské výšce h přibližně vztahem

$$P \doteq \alpha \cdot \mathrm{e}^{\beta h}.$$

Za ustáleného počasí bylo v určité oblasti provedeno 6 měření tlaku v různých nadmořských výškách:

h [m]	0	270	840	1452	2116	3203
P [mmHg]	760	737	686	635	584	508

Předpokládejte, že platí

$$P_i = \alpha \cdot \mathrm{e}^{\beta h_i} \cdot U_i,$$

kde U_i jsou nezávislé náhodné chyby mající logaritmicko-normální rozdělení $LN(0, \sigma^2)$. Najděte za tohoto předpokladu 95% interval spolehlivosti pro parametr β.

Příklad 11.21.

Zatížená cihla pokrytá laminátem se v čase udaném v hodinách prohýbala takto:

čas [hod]	průhyb [mm]	čas [hod]	průhyb [mm]	čas [hod]	průhyb [mm]
0	0.650	198	1.375	407	1.495
2	0.900	215	1.415	414	1.510
23	1.100	222	1.415	431	1.515
28	1.150	242	1.420	438	1.515
47	1.225	246	1.425	503	1.525
71	1.275	263	1.430	527	1.530
78	1.285	270	1.430	575	1.540
95	1.285	335	1.465	599	1.545
100	1.285	342	1.465	671	1.555
167	1.350	359	1.485	695	1.555
174	1.350	383	1.485	719	1.565
191	1.375	390	1.495		

Předpokládejme, že závislost průhybu na čase může být vyjádřena následovně:

$$w(t) = 0.65 + a_1(1 - \exp(-t/b_1)) + a_2(1 - \exp(-t/b_2)),$$

kde $b_1 = 2.59$ a $b_2 = 211.04$. Odhadněte pomocí metody nejmenších čtverců maximální průhyb, jestliže se čas neomezeně zvětšuje.

Příklad 11.22.

Předpokládejme, že smyková únosnost perforované lišty $\{P_i\}$ závisí na ploše ocelové výztuže na 1 mm délky perforované lišty $\{a_i\}$ a průměrné cylindrické pevnosti betonu v tlaku $\{f_i\}$ v den zkoušky následovně:

$$P_i = \alpha\, a_i + \beta\, f_i + e_i, \quad i = 1, \ldots, n,$$

kde veličiny $\{e_i\}$ v sobě zahrnují další náhodné vlivy. Při zkouškách únosnosti byly naměřeny následující hodnoty:

P_i [N/mm]	a_i [mm^2/mm]	f_i [MPa]
656.7	0.2683	41.2
1010.5	0.6713	37.4
357.3	0.0000	36.4
256.4	0.0000	20.0
282.1	0.0000	26.6
306.4	0.0000	28.7
445.3	0.2685	18.2
836.2	0.2685	32.4
740.2	0.2685	29.4
927.4	0.6713	29.4
816.2	0.6713	14.3
525.6	0.0967	21.4

Metodou nejmenších čtverců odhadněte koeficienty α a β a předpovězte smykovou únosnost lišty, jestliže plocha ocelové výztuže na 1 mm délky perforované lišty činí 0.0967 mm^2/mm a pevnost betonu v tlaku v den zkoušky 20.0 MPa.

Příklad 11.23.

Během 61 let byly zjišťovány průměrné roční průtoky Labe $\{W_i,\ i = 1,\ldots,61\}$ v m^3/s. Zároveň je v těchto letech znám roční úhrn srážek $\{R_i,\ i = 1,\ldots,61\}$ v mm a průměrná roční teplota $\{T_i,\ i = 1,\ldots,61\}$ ve ^{0}C měřené v jedné meteorologické stanici v povodí Labe. Je známo, že průměrný průtok v těchto letech byl roven $\bar{W} = 11.957377$ m^3/s, průměrný roční úhrn srážek $\bar{R} = 691.147541$ mm a průměrná roční teplota $\bar{T} = 7.544262$ ^{0}C. Výběrová kovarianční matice udávající výběrové kovariance mezi ročními průměrnými průtoky, úhrny srážek a průměrnými ročními teplotami spočtená z těchto 61 dat je následující:

$$\begin{pmatrix} 9646.826153 & 6704.076393 & -22.200815 \\ 6704.076393 & 8580.099454 & -14.850139 \\ -22.200815 & -14.850139 & 0.467198 \end{pmatrix}.$$

Předpokládejme, že závislost průměrných ročních průtoků na ročních úhrnech srážek a průměrných teplotách je přibližně lineární:

$$W_i = b_0 + b_1 R_i + b_2 T_i, \quad i = 1,\ldots,61.$$

Předpokládejme podobně jako v příkladu 11.4., že podmíněné rozdělení ročních průtoků, známe-li hodnoty srážkových úhrnů a průměrných teplot, je normální se stejným rozptylem. Pomocí testování hypotéz metodou „stepwise backward selection" rozhodněte, zda model není možno zjednodušit vypuštěním některého z regresorů. Pro zjednodušení výpočtu odhadů parametrů doporučujeme použít model

$$W_i = a + b_1 (R_i - \bar{R}) + b_2 (T_i - \bar{T}), \quad i = 1,\ldots,61.$$

Příklad 11.24.

Při měření délek se někdy v geodézii používá odraženého paprsku. Paprsek se odráží na odrazce, jež se může otáčet kolem svislé osy (natočení) a vodorovné osy (sklon). Natočení a sklon se měří ve stupních. Následující tabulka ukazuje, jak naměřená délka při pokusných měřeních závisela na natočení a sklonu. Metodou „ stepwise backward selection" se pokuste najít optimální model ve tvaru maximálně polynomu druhého stupně ve dvou proměnných, který by vyjadřoval závislost délky na natočení a sklonu odrazky. Předpokládejte, že náhodné chyby mají normální rozdělení.

nato-čení	sklon	délka [m]
35	90	38.2920
40	90	38.2915
50	90	38.2900
60	90	38.2897
70	90	38.2895
80	90	38.2895
90	90	38.2885
100	90	38.2880
110	90	38.2880
120	90	38.2885
130	90	38.2895
38	80	38.2905
40	80	38.2895
50	80	38.2885
60	80	38.2880
70	80	38.2880
80	80	38.2880
90	80	38.2875
100	80	38.2875
110	80	38.2875
120	80	38.2876
130	80	38.2886
65	70	38.2885
70	70	38.2876

nato-čení	sklon	délka [m]
80	70	38.2874
90	70	38.2865
100	70	38.2860
110	70	38.2860
115	70	38.2865
118	70	38.2874
38	100	38.2935
40	100	38.2925
50	100	38.2914
60	100	38.2910
70	100	38.2905
80	100	38.2895
90	100	38.2894
100	100	38.2890
110	100	38.2890
120	100	38.2900
130	100	38.2915
63	110	38.2920
65	110	38.2915
70	110	38.2905
80	110	38.2905
90	110	38.2905
100	110	38.2895

nato-čení	sklon	délka [m]
110	110	38.2900
118	110	38.2910
78	120	38.2915
80	120	38.2914
90	120	38.2905
100	120	38.2905
105	120	38.2905
90	125	38.2905
78	60	38.2865
80	60	38.2865
90	60	38.2860
100	60	38.2855
105	60	38.2855
90	55	38.2860
85	55	38.2860
90	50	38.2860
90	60	38.2864
90	70	38.2866
90	80	38.2875
90	100	38.2894
90	110	38.2896
90	120	38.2905
90	125	38.2910

Příklad 11.25.

Při zkouškách akcelerace vozidla byly registrovány časové okamžiky t_i průjezdu vozidla i-tým kontrolním bodem. Tyto kontrolní body byly umístěny na nultém, desátém, dvacátém, ... až padesátém metru trasy.

i	0	1	2	3	4	5
t_i	1.1	2.4	3.3	4.1	4.6	5.1

Předpokládejte, že závislost ujeté dráhy na čase je přibližně kvadratická. Pomocí příslušného regresního modelu odhadněte polohu vozidla v čase $t = 6$ s.

Příklad 11.26.

V 10 vybraných prodejnách se zjišťovaly váhové úbytky masa v závislosti na počtu dnů uskladnění.

uskladněno dnů	1	2	2	3	4	6	1	5	3	3
úbytek váhy [%]	1.1	1.4	1.5	1.7	1.7	2.0	0.9	1.8	1.6	1.5

Předpokládejte, že poměrný váhový úbytek masa závisí přibližně kvadraticky na době uskladnění. Najděte tuto závislost. Předpokládejte dále, že náhodné chyby v modelu jsou nezávislé normálně rozdělené veličiny se stejným rozptylem a otestujte, není-li možno zjednodušit kvadratický model na lineární.

Příklad 11.27.

V příkladu 6.4 byly v letech 1913-1942 sledovány výnosy pšenice ozimu, průměrné teploty vzduchu v předcházejícím zimním období, teploty vzduchu v letním vegetačním období a množství srážek ve vegetačním období. Předpokládejme, že závislost průměrného výnosu pšenice na zimní a letní průměrné teplotě a úhrnu srážek ve vegetačním období je přibližně lineární. Předpokládejme podobně jako v příkladu 11.4, že podmíněné rozdělení výnosů, známe-li hodnoty průměrných letních a zimních teplot a srážkových úhrnů, je přibližně normální. Pomocí testování hypotéz metodou „stepwise backward selection“ rozhodněte, zda všechny regresory opravdu výnos pšenice ovlivňují.

12. Časové řady

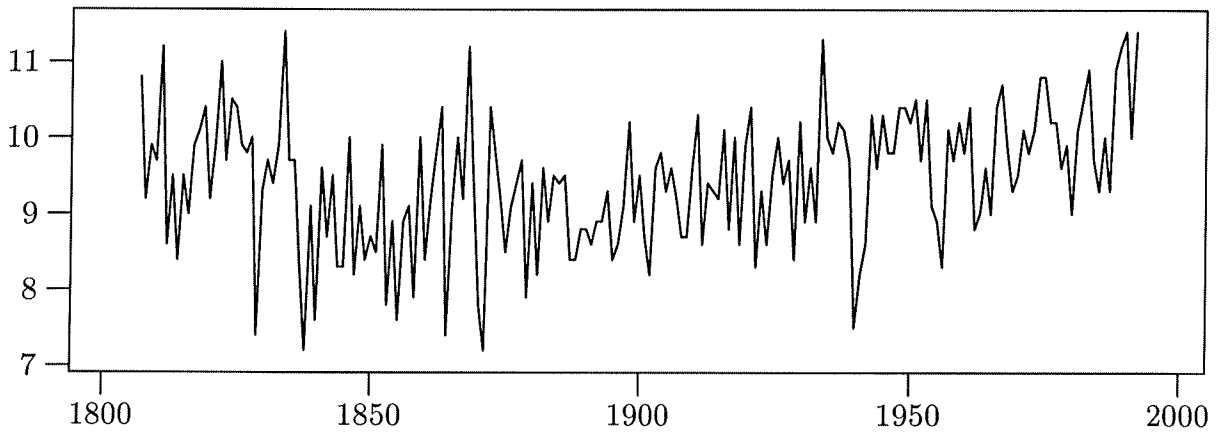
Řešené příklady

Příklad 12.1.

V následující tabulce jsou udány průměrné roční teploty měřené v Klementinu od roku 1807 do roku 1992.

1807	10.8	1838	7.2	1869	9.4	1900	9.5	1931	8.9	1962	8.8
1808	9.2	1839	9.1	1870	7.8	1901	8.7	1932	9.6	1963	9.0
1809	9.9	1840	7.6	1871	7.2	1902	8.2	1933	8.9	1964	9.6
1810	9.7	1841	9.6	1872	10.4	1903	9.6	1934	11.3	1965	9.0
1811	11.2	1842	8.7	1873	9.8	1904	9.8	1935	10.0	1966	10.4
1812	8.6	1843	9.5	1874	9.2	1905	9.3	1936	9.8	1967	10.7
1813	9.5	1844	8.3	1875	8.5	1906	9.6	1937	10.2	1968	9.9
1814	8.4	1845	8.3	1876	9.1	1907	9.2	1938	10.1	1969	9.3
1815	9.5	1846	10.0	1877	9.4	1908	8.7	1939	9.7	1970	9.5
1816	9.0	1847	8.2	1878	9.7	1909	8.7	1940	7.5	1971	10.1
1817	9.9	1848	9.1	1879	7.9	1910	9.6	1941	8.2	1972	9.8
1818	10.1	1849	8.4	1880	9.4	1911	10.3	1942	8.6	1973	10.1
1819	10.4	1850	8.7	1881	8.2	1912	8.6	1943	10.3	1974	10.8
1820	9.2	1851	8.5	1882	9.6	1913	9.4	1944	9.6	1975	10.8
1821	9.9	1852	9.9	1883	8.9	1914	9.3	1945	10.3	1976	10.2
1822	11.0	1853	7.8	1884	9.5	1915	9.2	1946	9.8	1977	10.2
1823	9.7	1854	8.9	1885	9.4	1916	10.1	1947	9.8	1978	9.6
1824	10.5	1855	7.6	1886	9.5	1917	8.8	1948	10.4	1979	9.9
1825	10.4	1856	8.9	1887	8.4	1918	10.0	1949	10.4	1980	9.0
1826	9.9	1857	9.1	1888	8.4	1919	8.6	1950	10.2	1981	10.1
1827	9.8	1858	7.9	1889	8.8	1920	9.9	1951	10.5	1982	10.5
1828	10.0	1859	10.0	1890	8.8	1921	10.4	1952	9.7	1983	10.9
1829	7.4	1860	8.4	1891	8.6	1922	8.3	1953	10.5	1984	9.7
1830	9.3	1861	9.2	1892	8.9	1923	9.3	1954	9.1	1985	9.3
1831	9.7	1862	9.8	1893	8.9	1924	8.6	1955	8.9	1986	10.0
1832	9.4	1863	10.4	1894	9.3	1925	9.5	1956	8.3	1987	9.3
1833	9.9	1864	7.4	1895	8.4	1926	10.0	1957	10.1	1988	10.9
1834	11.4	1865	9.1	1896	8.6	1927	9.4	1958	9.7	1989	11.2
1835	9.7	1866	10.0	1897	9.1	1928	9.7	1959	10.2	1990	11.4
1836	9.7	1867	9.2	1898	10.2	1929	8.4	1960	9.8	1991	10.0
1837	8.3	1868	11.2	1899	8.9	1930	10.2	1961	10.4	1992	11.4

Řada je rovněž znázorněna na následujícím obrázku.



Průměrné roční teploty v Klementinu v letech 1807 – 1992.

Za předpokladu, že průměrné roční teploty mohou být považovány za nezávislé náhodné veličiny s normálním rozdělením, otestujte, zda úsek řady $Y_1, \ldots, Y_{93}$ mezi léty 1900 až 1992 může být považován za stacionární nebo zda má statisticky významný kladný trend.

Řešení:

Uvažujeme model lineární regrese:

$$Y_i = a + b\,i + e_i, \quad i = 1, \ldots, 93.$$

Metodou nejmenších čtverců odhadneme a a b a směrodatnou odchylku náhodných chyb σ:

$$\hat{a} = 9.095, \quad \hat{b} = 0.0126436, \quad \hat{\sigma} = 0.700824.$$

Testujeme $H_b\colon b = 0$ proti $A_b\colon b > 0$. Na hladině významnosti $\alpha = 0.05$ nulovou hypotézu zamítáme, protože $\hat{b}/s_{\hat{b}} = 0.0126463/0.002707 \doteq 4.671 > t_{0.05}[91] \doteq 1.662$. Příslušná $p-$hodnota je rovna $5.2 \cdot 10^{-6}$.

Odpověď: Pokud předpokládáme, že teploty měřené v Klementinu jsou nezávislé náhodné veličiny, pak má řada statisticky významný trend dokonce na hladině významnosti $\alpha = 0.00001$. Teplota mezi léty 1900 – 1992 stoupala v průměru o $0.01\,^0$C za rok.

Příklad 12.2.

V poslední době počet narozených dětí v České republice výrazně klesá. Počet živě narozených dětí v letech 1980 – 1995 udává následující tabulka:

rok	počet	rok	počet
1980	153801	1988	132667
1981	144438	1989	128356
1982	141738	1990	130564
1983	137431	1991	129354
1984	136941	1992	121705
1985	135881	1993	121025
1986	133356	1994	106579
1987	130921	1995	96097

Otestujte, zda po roce 1990 klesá počet narozených dětí rychleji než před tímto rokem.

Řešení:

Pokud rok 1990 neměl na rychlost poklesu počtu narozených dětí žádný vliv, můžeme uvažovat lineární závislost počtu narozených dětí Y_t, $t = 1980, \ldots, 1995$ na čase:

$$Y_t = a + b \cdot (t - 1979) + e_t.$$

Pokud se však pokles počtu narozených dětí po roce 1990 ještě více zvýšil, uvažujme regresi ve tvaru:

$$Y_t = a + b \cdot (t - 1979) + c \cdot (t - 1990)_+ + e_t.$$

Poznamenejme, že kladná část d_+ nabývá hodnoty 0, pokud d je záporné, a hodnoty d, pokud d je větší nebo rovno nule. Druhý model znamená, že časová řada klesala před rokem 1990 se směrnicí b a po roce 1990 se směrnicí $b + c$, přičemž se předpokládá spojité chování řady v roce 1990. Rozhodnout, který z modelů časové řady je správný, můžeme pomocí testování hypotéz. Uvažujeme lineární regresi s více vysvětlujícími parametry $Y = X\beta + e$, kde

$$Y = \begin{pmatrix} Y_{1980} \\ \vdots \\ Y_{1995} \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 11 & 0 \\ 1 & 12 & 1 \\ 1 & 13 & 2 \\ \vdots & \vdots & \vdots \\ 1 & 16 & 5 \end{pmatrix}, \qquad \beta = \begin{pmatrix} a \\ b \\ c \end{pmatrix}.$$

Testujeme nulovou hypotézu $H_c : c = 0$ proti alternativě $A_c : c < 0$. Předpokládejme, že chyby $\{e_t\}$ jsou nezávislé a mají normální rozdělení. Metodou nejmenších čtverců odhadněme koeficienty a, b a c:

$$\hat{a} = 147522.7, \quad \hat{b} = -1631.576981, \quad \hat{c} = -3840.971096$$

a směrodatnou odchylku náhodných chyb $\hat{\sigma} = 4404.758758$. Matice

$$(X'X)^{-1} = \begin{pmatrix} 0.3948720 & -0.0487179 & 0.0871795 \\ -0.0487179 & 0.0076340 & -0.0172494 \\ 0.0871795 & -0.0172494 & 0.0634033 \end{pmatrix}.$$

Odtud $s_{\hat{c}} = \hat{\sigma}\sqrt{0.0634033} = 1109.11846$ a $\hat{c}/s_{\hat{c}} = -3.4631 < -t_{0.05}[13] = -2.1604$. Na hladině významnosti $\alpha = 0.05$ nulovou hypotézu H_c zamítáme ve prospěch alternativy A_c.

Odpověď: Vycházíme-li ze shora uvedených statistických modelů, zjišťujeme, že se rychlost poklesu počtu narozených dětí v České republice po roce 1990 statisticky významně zvýšila. Zatímco před rokem 1990 poklesl ročně počet narozených dětí v průměru přibližně o 1632, po roce 1990 o 5473.

Příklad 12.3.

Předpokládejme, že pozorujeme nestacionární časovou řadu $\{Y_t\}$, jejíž závislost na čase je dána vztahem:

$$Y_t = m(t) + e_t,$$

kde $\{e_t\}$ reprezentují náhodné chyby. Předpokládejme, že mají nulovou střední hodnotu a shodný rozptyl. Regresní funkci $\mathrm{E}\,Y_t = m(t)$ však neumíme parametrizovat. V tom případě ji můžeme alespoň odhadnout neparametricky. Mezi nejpoužívanější neparametrické odhady regresní funkce $m(t)$ patří jádrové odhady ve tvaru:

$$\hat{m}(t) = \frac{\sum_{i=1}^{n} Y_i\, K\left(\frac{t-i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{t-i}{h}\right)},$$

kde $K(\cdot)$ je tzv. jádro a h je vhodně zvolená konstanta. Volíme-li jádro $K(\cdot)$ ve tvaru (tzv. klouzavé okénko):

$$K(x) = 1 \quad \text{pro} \quad |x| \leq 1,$$
$$K(x) = 0 \quad \text{pro} \quad |x| > 1,$$

odhadujeme funkci $m(t)$ v bodě t klouzavým průměrem

$$\hat{m}(t) = \frac{\sum_{i=-h}^{h} Y_{t+i}}{2h+1} \quad \text{pro} \quad h+1 \leq t \leq n-h.$$

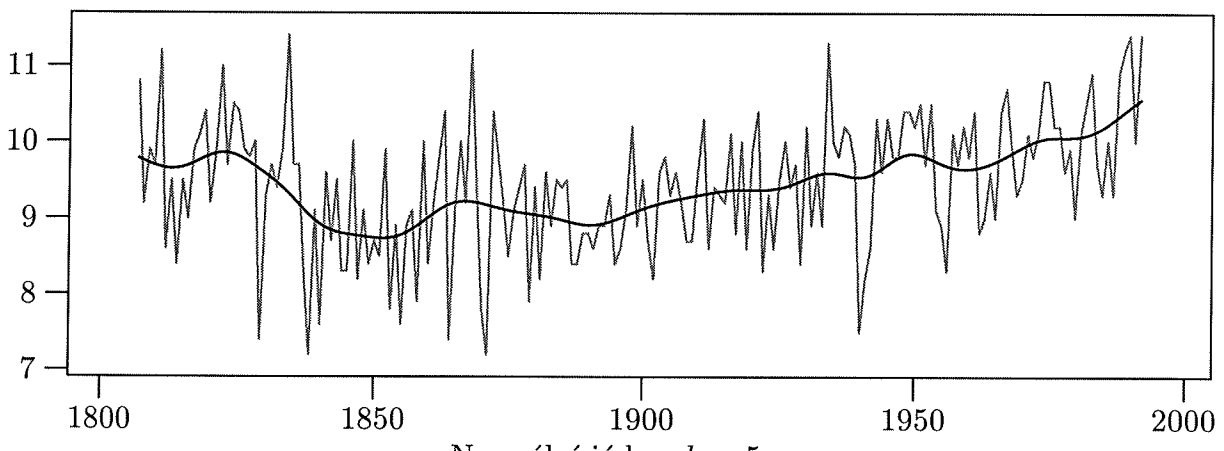
Velmi oblíbeným je též normální jádro

$$K(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}.$$

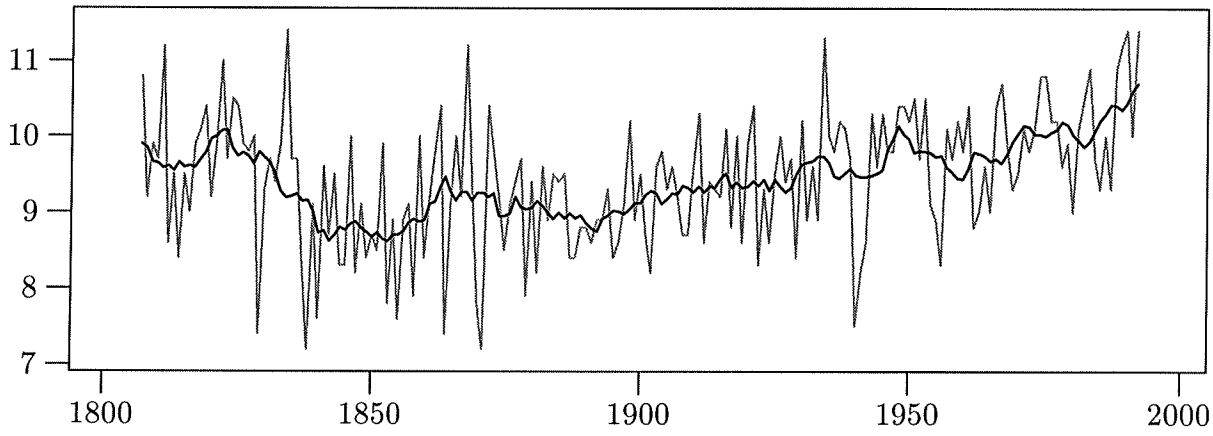
Poznamenejme, že jádrové odhady střední hodnoty časové řady patří mezi vyhlazovací (anglicky „smoothing“) metody.

Odhadněte pomocí jádrového odhadu regresní funkci $m(t)$ vyjadřující časový vývoj střední hodnoty průměrných ročních teplot měřených v Klementinu v letech 1807-1992 (viz příklad 12.1). Použijte normální jádro i klouzavé okénko, konstantu h volte v obou případech rovnu 5.

Řešení:



Normální jádro, $h = 5$.



Rovnoměrné jádro, $h = 5$.

Příklad 12.4.

V tabulce je zachycen vývoj kursu akcie Investiční banky na burze cenných papírů v Praze v jednotlivých burzovních dnech období od 13. července 1993 do 9. června 1994 (kursy v Kč jsou chronologicky seřazeny po sloupcích). Vyrovnejte tuto časovou řadu pomocí klouzavého průměru, klouzavého mediánu a metodou jednoduchého exponenciálního vyrovnávání. Pomocí akciogramu rozhodněte, kdy by bývalo bylo výhodné akcie tohoto cenného papíru koupit, resp. prodat.

4000	3000	4835	4220	5500	7400	6800	6100	6000	5995	5900
3600	2745	5800	4600	6050	7400	6120	6000	5900	5980	5310
3240	2200	6960	4600	6655	6660	5510	6000	5900	5900	4900
1800	2500	5570	4900	7320	7325	5700	5400	5900	5900	4500
2000	2700	5000	5000	7300	7995	6100	5940	5790	5900	4050
2200	2800	5500	5100	8000	7200	6000	6000	5900	5900	3655
2290	3360	4400	5100	8000	7050	5900	6000	6000	5900	3550
2500	4030	3520	5100	7200	6800	5900	6000	6000	5860	

Řešení:

Průběhy akciových kursů lze analyzovat nejrůznějšími způsoby. V tomto příkladu popíšeme nejpoužívanější metody tzv. technické analýzy. Tato analýza je založená na předpokladu, že dosavadní vývoj kursů umožňuje v rámci krátkodobého a střednědobého investování odhadnout budoucí vývoj.

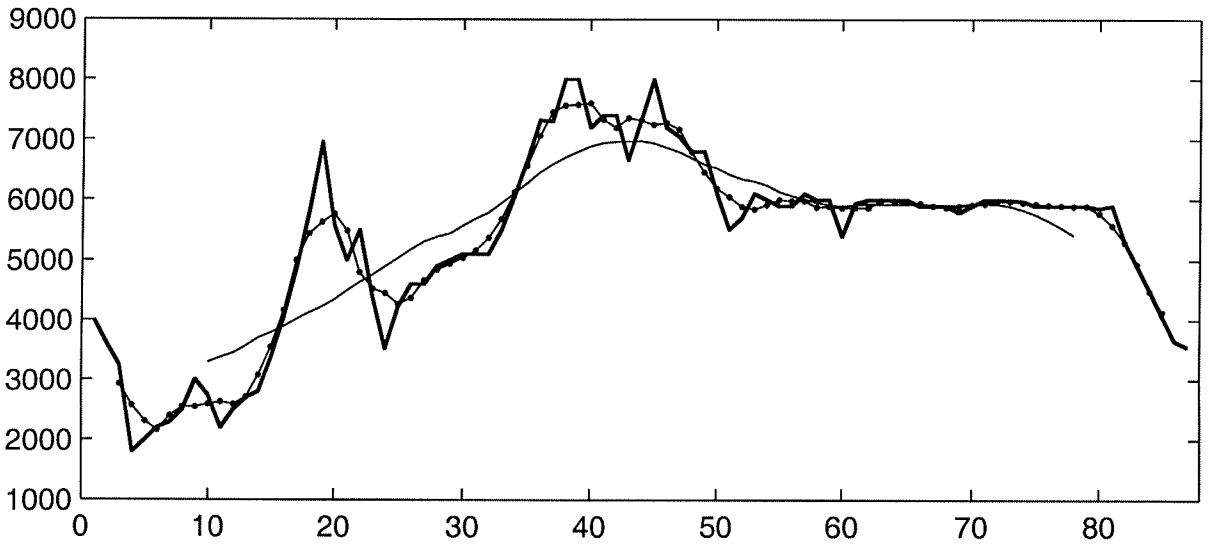
Naše časová řada kursů je i s daty burzovních dnů uvedena znovu v závěru řešení tohoto příkladu. Můžete si tam všimnout, že jednotlivá pozorování neměla stejné časové odstupy, v burzovních dnech byly značné nepravidelnosti. Toto je však zvykem ignorovat a skutečná data konání jednotlivých burzovních dnů nebudou hrát při naší analýze žádnou roli.

Vyrovnání časové řady klouzavými průměry bylo zmíněno v příkladu 12.3. Volba konstanty h, resp. volba délky klouzavého okénka $2h + 1$, přes které průměrujeme, je

subjektivní záležitostí. Zvolíme-li veliké h, dosáhneme opticky lepšího vyhlazení, ale můžeme potlačit některé podstatné složky původní časové řady. Podle zvolené délky klouzavého okénka se klouzavé průměry někdy dělí na krátkodobé, střednědobé a dlouhodobé. V našem příkladu zvolme pro ilustraci krátký průměr s $h = 2$ (tj. délku klouzavého okénka rovnu 5) a dlouhý průměr s $h = 9$ (tedy délku klouzavého okénka rovnu 19). Uveďme konkrétní výpočet pro $h = 2$:

$$\hat{y}_3 = \frac{1}{5}(y_1 + y_2 + y_3 + y_4 + y_5) = 2928, \quad \hat{y}_4 = \frac{1}{5}(y_2 + y_3 + y_4 + y_5 + y_6) = 2568, \quad \ldots .$$

Na prvním obrázku je kurs akcie vyznačen tlustou čarou, krátký průměr je vyznačen body spojenými obyčejnou čarou a dlouhý průměr obyčejnou čarou. Poznamenejme, že klasické jednoduché klouzavé průměry nevyrovnávají počáteční a koncový úsek časové řady.



Výše uvedená metoda k vyrovnání hodnoty y_n dané časové řady používá jak minulé hodnoty y_{n-i}, tak i hodnoty budoucí y_{n+i} (šlo o klouzavý průměr přes symetrické okénko). Toto není na závadu, pokud chceme danou časovou řadu jako celek rozumně vyhladit, interpolovat nějakou mezilehlou hodnotu, apod. Pokud však chceme předpovídat budoucí vývoj řady na základě znalosti jejího průběhu $\ldots, y_{n-2}, y_{n-1}, y_n$, chceme mít k dispozici metodu, která umožní vyrovnat hodnotu y_n bez použití budoucích členů řady. V dalším textu takové metody uvedeme, budeme se zaměřovat na predikci.

Teoreticky by bylo možné použít nesymetrické klouzavé okénko a průměrovat např. pro $h = 2$: $\hat{y}_5 = \frac{1}{5}(y_1 + \cdots + y_5)$, $\hat{y}_6 = \frac{1}{5}(y_2 + \cdots + y_6)$ atd. Tuto variantu si případně propočtěte sami.

Vyrovnání časové řady klouzavým mediánem je jednou z možností, jak potlačit odlehlá pozorování. Pro průběžné potlačování odlehlých pozorování v akciových kursech se doporučuje varianta klouzavého mediánu:

$$\hat{y}_t = \mathrm{med}(y_{t-2}, y_{t-1}, y_t), \qquad t = 3,\ 4,\ \ldots ,$$

kde $\mathrm{med}(y_{t-2}, y_{t-1}, y_t)$ označuje medián z uvedené trojice hodnot. Pro naši časovou řadu je

$$\hat{y}_3 = \text{med}(4000, 3600, 3240) = 3600, \quad \hat{y}_4 = \text{med}(3600, 3240, 1800) = 3240, \quad \ldots .$$

Všimněte si, že v uvedené variantě kouzavého mediánu se k vyrovnání hodnoty y_t použijí jen hodnoty časové řady známé do doby t, viz též poznámka výše u klouzavého průměru.

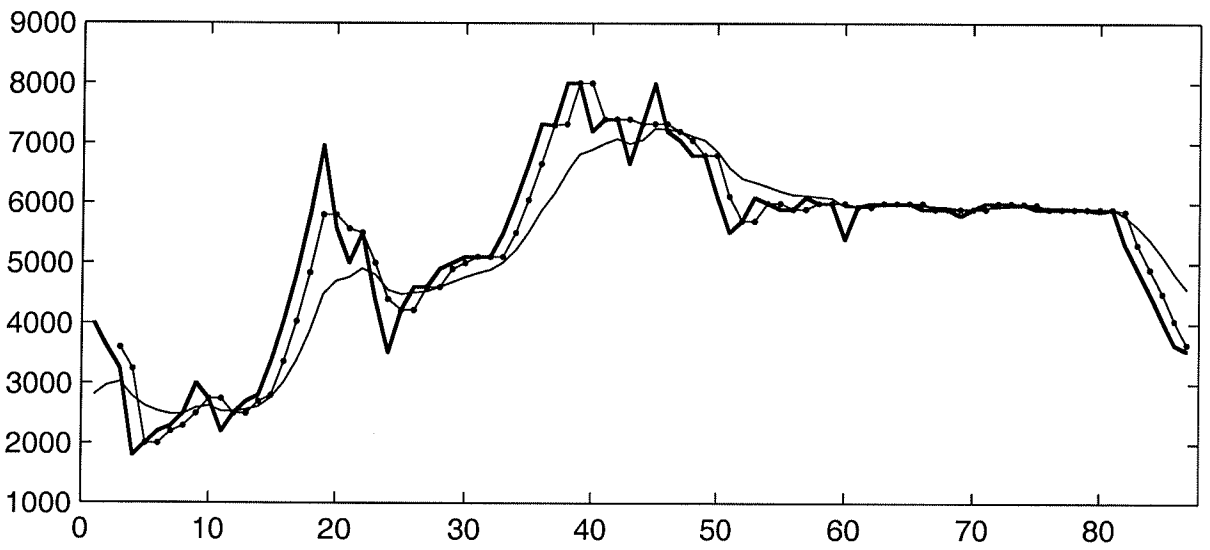
Jednoduché exponenciální vyrovnávání patří k nejoblíbenějším a v praxi nejpoužívanějším metodám. Jeho principem je, že hodnota y_t se vyrovnává váženým průměrem minulých hodnot, přičemž váhy exponenciálně rychle klesají směrem do minulosti. Logika tohoto postupu je, že určité minulé pozorování má na konstrukci vyrovnané hodnoty tím menší vliv, čím je časově vzdálenější. Pro časovou řadu neomezenou do minulosti platí vzorec

$$\hat{y}_t = \alpha y_t + \alpha(1-\alpha)y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} + \cdots \quad \Longrightarrow \quad \hat{y}_t = \alpha y_t + (1-\alpha)\hat{y}_{t-1}.$$

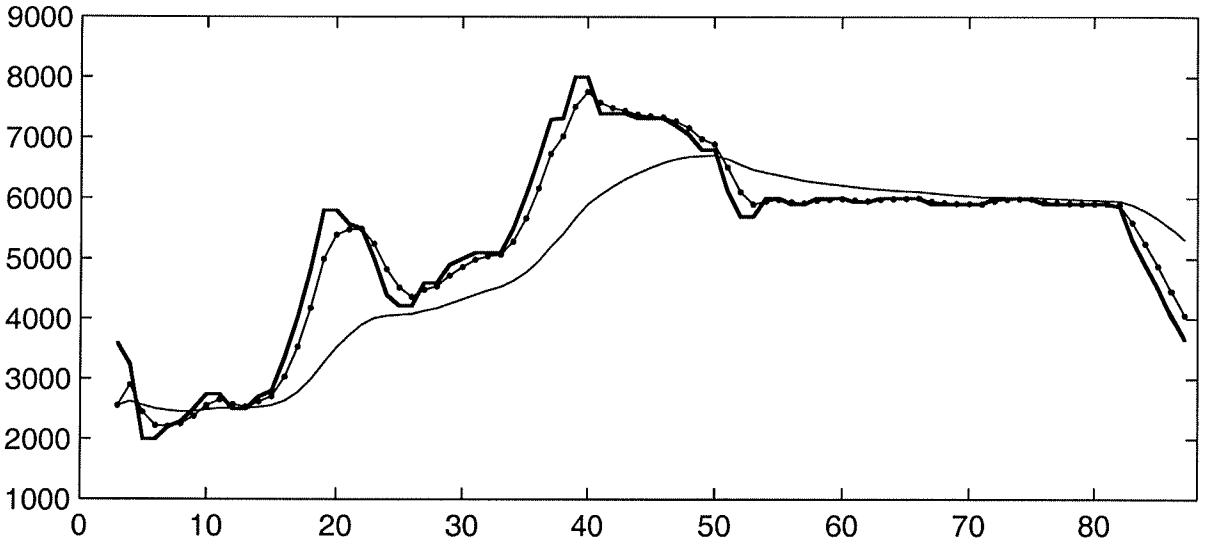
Zde $\alpha \in (0, 1)$ je tzv. vyrovnávací konstanta. Pro vyrovnání řady $y_1, y_2, \ldots, y_n$ je třeba vhodně zvolit první vyrovnanou hodnotu $\hat{y}_1$, konstantu α a dále použít výše uvedený rekurentní vzorec. V našem příkladu zvolme $\hat{y}_1$ jako průměr prvních šesti hodnot a α rovno 0.2 (doporučuje se volit $\alpha \in (0, 0.3)$, je možné tuto volbu pro danou řadu určitým způsobem optimalizovat, ale nebudeme se o to pokoušet). Platí tedy:

$$\begin{aligned}
\hat{y}_1 &= \frac{1}{6}(y_1 + y_2 + \cdots + y_6) = \frac{1}{6}(4000 + 3600 + \cdots + 2200) = 2806.67, \\
\hat{y}_2 &= \alpha y_2 + (1-\alpha)\hat{y}_1 = 0.2 \cdot 3600 + 0.8 \cdot 2806.67 = 2965.33, \\
\hat{y}_3 &= \alpha y_3 + (1-\alpha)\hat{y}_2 = 0.2 \cdot 3240 + 0.8 \cdot 2965.33 = 3020.27, \\
&\ldots
\end{aligned}$$

Na druhém obrázku vidíte opět kurs akcie (tlustá čára), vyrovnání klouzavým mediánem (body spojené obyčejnou čarou) a exponenciální vyrovnání (obyčejná čára).



Hlavním úkolem analýzy akciových kursů je najít vhodné momenty k nákupu či k prodeji. Ukážeme si jednu z možností, jak tyto momenty najít pomocí výše uvedených vyrovnávacích metod. V této souvislosti se mluví o tzv. *akciogramu*. Časovou řadu nejprve „očistíme" pomocí klouzavého mediánu. Tuto očištěnou řadu vyrovnáme pomocí krátkého a dlouhého průměru. Signálem k nákupu akcie je potom okamžik, kdy krátký průměr protne dlouhý průměr zdola (tj. když růst cen překročí systematicky dlouhodobý normál). Signálem k prodeji je naopak okamžik, kdy krátký průměr protne dlouhý průměr shora. Klouzavý medián zkonstruujeme opět z posledních třech hodnot, jako reprezentanta krátkého průměru zvolíme metodu jednoduchého exponenciálního vyrovnávání s $\alpha = 0.5$ a jako reprezentanta dlouhého průměru tutéž metodu s $\alpha = 0.1$. (Uvádí se, že exponenciální vyrovnání s konstantou α přitom zhruba odpovídá klouzavému průměru s konstantou h, kde $\alpha \sim 1/(h+1)$. Námi zvolené $\alpha = 0.1$ tedy odpovídá přibližně klouzavému průměru s $h = 9$.) Připomeňme ještě, že tentokrát provádíme exponenciální vyrovnání očištěné časové řady získané jako klouzavý medián, počáteční hodnotu pro obě varianty α zvolme opět jako průměr prvních šesti členů této očištěné řady. Na posledním obrázku vidíte kurs akcie vyhlazený klouzavým mediánem (tlustá čára), jeho vyrovnání krátkým průměrem (body spojené čarou) a dlouhým průměrem (obyčejná čára). Průsečíky krátkého a dlouhého průměru jsou pak hledané signály k nákupu a prodeji. V rozsáhlé tabulce jsou pak přehledně shrnuty všechny výpočty. Z této tabulky vyčteme, že tyto akcie by bývalo bylo vhodné koupit dne 14. 9. 1993 a opět prodat 15. 3. 1994. O výhodnosti takového počínání svědčí cenový rozdíl 2770 Kč.



Čtenáře, zajímajícího se více o problematiku vyrovnávání a predikce akciových kursů, odkazujeme na publikaci Cipra (1995).

burzovní den	kurs [Kč]	klouzavý medián	krátký průměr	dlouhý průměr	signál k nákupu	signál k prodeji
13. 07. 1993	4000				—	—
20. 07. 1993	3600				—	—
27. 07. 1993	3240	3600	2555.00	2555.00	—	—
03. 08. 1993	1800	3240	2897.50	2623.50	—	—
10. 08. 1993	2000	2000	2448.75	2561.15	—	+
17. 08. 1993	2200	2000	2224.38	2505.04	—	—
24. 08. 1993	2290	2200	2212.19	2474.53	—	—
31. 08. 1993	2500	2290	2251.09	2456.08	—	—
07. 09. 1993	3000	2500	2375.55	2460.47	—	—
14. 09. 1993	2745	2745	2560.27	2488.92	+	—
21. 09. 1993	2200	2745	2652.64	2514.53	—	—
28. 09. 1993	2500	2500	2576.32	2513.08	—	—
05. 10. 1993	2700	2500	2538.16	2511.77	—	—
12. 10. 1993	2800	2700	2619.08	2530.59	—	—
19. 10. 1993	3360	2800	2709.54	2557.53	—	—
26. 10. 1993	4030	3360	3034.77	2637.78	—	—
02. 11. 1993	4835	4030	3532.38	2777.00	—	—
04. 11. 1993	5800	4835	4183.69	2982.80	—	—
09. 11. 1993	6960	5800	4991.85	3264.52	—	—
11. 11. 1993	5570	5800	5395.92	3518.07	—	—
16. 11. 1993	5000	5570	5482.96	3723.26	—	—
18. 11. 1993	5500	5500	5491.48	3900.94	—	—
23. 11. 1993	4400	5000	5245.74	4010.84	—	—
25. 11. 1993	3520	4400	4822.87	4049.76	—	—
30. 11. 1993	4220	4220	4521.44	4066.78	—	—
02. 12. 1993	4600	4220	4370.72	4082.10	—	—
07. 12. 1993	4600	4600	4485.36	4133.89	—	—
09. 12. 1993	4900	4600	4542.68	4180.50	—	—
14. 12. 1993	5000	4900	4721.34	4252.45	—	—
16. 12. 1993	5100	5000	4860.67	4327.21	—	—
06. 01. 1994	5100	5100	4980.33	4404.49	—	—
11. 01. 1994	5100	5100	5040.17	4474.04	—	—
13. 01. 1994	5500	5100	5070.08	4536.64	—	—
18. 01. 1994	6050	5500	5285.04	4632.97	—	—
20. 01. 1994	6655	6050	5667.52	4774.67	—	—
25. 01. 1994	7320	6655	6161.26	4962.71	—	—
27. 01. 1994	7300	7300	6730.63	5196.44	—	—
01. 02. 1994	8000	7320	7025.32	5408.79	—	—
03. 02. 1994	8000	8000	7512.66	5667.91	—	—
08. 02. 1994	7200	8000	7756.33	5901.12	—	—
10. 02. 1994	7400	7400	7578.16	6051.01	—	—
15. 02. 1994	7400	7400	7489.08	6185.91	—	—
17. 02. 1994	6660	7400	7444.54	6307.32	—	—

burzovní den	kurs [Kč]	klouzavý medián	krátký průměr	dlouhý průměr	signál k nákupu	signál k prodeji
22. 02. 1994	7325	7325	7384.77	6409.09	—	—
24. 02. 1994	7995	7325	7354.89	6500.68	—	—
01. 03. 1994	7200	7325	7339.94	6583.11	—	—
03. 03. 1994	7050	7200	7269.97	6644.80	—	—
08. 03. 1994	6800	7050	7159.99	6685.32	—	—
10. 03. 1994	6800	6800	6979.99	6696.79	—	—
14. 03. 1994	6120	6800	6890.00	6707.11	—	—
15. 03. 1994	5510	6120	6505.00	6648.40	—	+
17. 03. 1994	5700	5700	6102.50	6553.56	—	—
21. 03. 1994	6100	5700	5901.25	6468.20	—	—
22. 04. 1994	6000	6000	5950.62	6421.38	—	—
24. 03. 1994	5900	6000	5975.31	6379.24	—	—
28. 03. 1994	5900	5900	5937.66	6331.32	—	—
29. 03. 1994	6100	5900	5918.83	6288.19	—	—
31. 03. 1994	6000	6000	5959.41	6259.37	—	—
05. 04. 1994	6000	6000	5979.71	6233.43	—	—
07. 04. 1994	5400	6000	5989.85	6210.09	—	—
11. 04. 1994	5940	5940	5964.93	6183.08	—	—
12. 04. 1994	6000	5940	5952.46	6158.77	—	—
14. 04. 1994	6000	6000	5976.23	6142.89	—	—
18. 04. 1994	6000	6000	5988.12	6128.61	—	—
19. 04. 1994	6000	6000	5994.06	6115.74	—	—
21. 04. 1994	5900	6000	5997.03	6104.17	—	—
25. 04. 1994	5900	5900	5948.51	6083.75	—	—
26. 04. 1994	5900	5900	5924.26	6065.38	—	—
28. 04. 1994	5790	5900	5912.13	6048.84	—	—
02. 05. 1994	5900	5900	5906.06	6033.96	—	—
03. 05. 1994	6000	5900	5903.03	6020.56	—	—
05. 05. 1994	6000	6000	5951.52	6018.50	—	—
09. 05. 1994	5995	6000	5975.76	6016.65	—	—
10. 05. 1994	5980	5995	5985.38	6014.49	—	—
12. 05. 1994	5900	5980	5982.69	6011.04	—	—
16. 05. 1994	5900	5900	5941.34	5999.94	—	—
17. 05. 1994	5900	5900	5920.67	5989.94	—	—
19. 05. 1994	5900	5900	5910.34	5980.95	—	—
23. 05. 1994	5900	5900	5905.17	5972.85	—	—
24. 05. 1994	5860	5900	5902.58	5965.57	—	—
26. 05. 1994	5900	5900	5901.29	5959.01	—	—
30. 05. 1994	5310	5860	5880.65	5949.11	—	—
31. 05. 1994	4900	5310	5595.32	5885.20	—	—
02. 06. 1994	4500	4900	5247.66	5786.68	—	—
06. 06. 1994	4050	4500	4873.83	5658.01	—	—
07. 06. 1994	3655	4050	4461.92	5497.21	—	—
09. 06. 1994	3550	3655	4058.46	5312.99	—	—

Příklad 12.5.

Na betonovém základě stroje je umístěno čidlo, které snímá vibrace. Předpokládáme, že základ kmitá pouze s jednou vynucenou frekvencí α. Signál byl snímán během časového intervalu $\langle 0, T\rangle$ a digitalizován. Digitalizovaný signál $\{S_t,\ t = 1, \ldots, n\}$ $(n = 256)$ je po sloupcích uveden v následující tabulce:

3.7433657	-6.4257957	-3.0837344	2.9301259	1.7826857	-3.9598588	-1.7287108	-1.7337894
3.8519512	-3.9609909	-3.0446357	1.3573024	4.0220746	-1.6522091	-2.4605195	-0.0400703
4.6582556	-3.5167319	-2.6260253	1.5373048	4.2778695	-2.3445565	-3.1271758	-0.8074759
3.9919812	-2.2583551	-4.5849751	0.3166985	3.7944147	-0.8488091	-5.6091664	-0.3360868
6.6090798	-2.3617430	-3.4705421	0.5002870	6.8643202	-1.2781719	-2.7554944	-0.5360798
5.0377600	-1.0411661	-2.9429033	-1.6162783	5.5326271	2.0528692	-4.9797443	-0.5995217
6.6894572	1.4359917	-4.1144884	-1.6855737	4.8554755	1.1326049	-5.5575819	-2.0889821
4.9316768	1.5700751	-3.6659505	0.0392869	4.1387945	3.5791542	-4.8930044	-2.4997074
5.1953208	1.8726152	-4.4242067	-2.3789989	0.3173294	4.1884822	-3.5465637	-1.1855925
2.4375694	1.5168220	-2.2387721	-1.9027294	4.3103682	3.8658546	-2.8293703	-1.1593468
2.6720131	3.1258395	-3.0903721	-4.0989997	2.0725657	4.7978213	-2.7158324	-3.7234254
4.0633387	3.9910328	-5.4525089	-3.2772139	0.5675735	4.4863960	-2.5270038	-3.5057588
1.6009876	3.7999667	-3.0955508	-3.9973525	-1.1394323	4.5934183	0.3067024	-4.9074472
2.2388011	4.3848757	-4.3078779	-3.8193246	0.4757320	6.1060360	-1.2978220	-4.1308857
0.7378663	5.3433781	-0.7352005	-2.8767573	1.9794758	4.8491130	0.3384853	-3.5574265
0.4247984	6.5342120	-0.5494256	-3.1228166	-0.2040069	3.4287519	0.4023877	-4.9784282
-1.2944887	2.4294681	0.5684087	-3.1107119	-0.1553453	5.7035015	2.4076997	-4.2054677
-1.6878673	3.5363859	1.6967197	-4.2154701	-1.5948476	2.6135879	0.8599128	-4.1297274
0.1255774	5.4013016	3.4535536	-3.8655162	-0.7713127	3.5508003	4.1617318	-4.4871145
-1.3199168	4.4112169	2.2095648	-5.2919644	-3.7096363	1.7154009	2.9886659	-3.9663676
-2.4707826	3.0413119	3.1373588	-4.4297813	-1.5884383	1.6680017	4.4367143	-0.9822004
-2.8204544	3.1674865	4.5235043	-2.3175385	-2.3896497	1.8039371	3.7741477	-3.1854455
-1.5676560	1.1970180	5.3834804	-2.1670137	-3.4493425	1.9029701	3.7305265	-1.9020539
-2.8513333	0.5758177	3.6373281	-1.9836430	-3.3930371	-0.0112274	3.0590925	-2.0616931
-2.6359635	1.3538794	4.3913449	-2.9225139	-2.2933197	0.1753502	3.8087467	0.2337452
-3.0883836	2.0181601	5.0120785	-0.5235816	-4.0923236	-0.8403087	4.7001980	1.6171829
-3.0342044	0.7497049	2.3705844	0.4188553	-3.4067373	-1.0316844	4.5807479	1.5537721
-3.6498705	-0.9760672	5.7587381	0.5172621	-5.6783127	-2.6473217	5.1518081	3.2204558
-5.5885570	-0.0214898	3.2062079	2.0894713	-3.1742096	-1.9767708	3.7105764	2.8506956
-3.9870837	-1.5684503	4.9279936	5.1837792	-2.3968102	-0.7365201	0.3466317	5.4169836
-3.2519137	-2.4416397	2.4651654	2.8939746	-4.5261611	-3.4313151	0.8858662	4.5492665
-3.5476884	-2.5426756	3.3238060	2.3597955	-4.5464580	-4.2567139	1.2013307	5.6352389

Po odečtení průměru $\bar{S}$ budeme signál považovat za časovou řadu $\{Y_t,\ t = 1, \ldots, n\}$, o které budeme předpokládat:

$$Y_t = m(t) + e_t,$$

kde regresní funkce $m(t)$ má tvar:

$$m(t) = C\cos(\theta t + \phi) = A\cos(\theta t) + B\sin(\theta t).$$

$\theta = \alpha 2\pi T/n$ je frekvence, C je amplituda, ϕ je fázové posunutí. Pro konstanty A a B platí $A = C\cos(\phi)$ a $B = -C\sin(\phi)$. Náhodné chyby $\{e_t\}$ reprezentují chyby měření a jiné náhodné vlivy a předpokládá se o nich, že jsou to nezávislé stejně rozdělené náhodné veličiny, pro které $\mathrm{E}\, e_t = 0$ a $\mathrm{Var}\, e_t = \sigma^2$. Odhadněte frekvenci α, amplitudu a fázové posunutí snímaného signálu.

Poznamenejme, že pokud je to možné, je výhodné volit n ve tvaru $n = 2^p$, $p \in N$. Tato volba umožní optimálně využít vlastností rychlé Fourierovy transformace, viz níže.

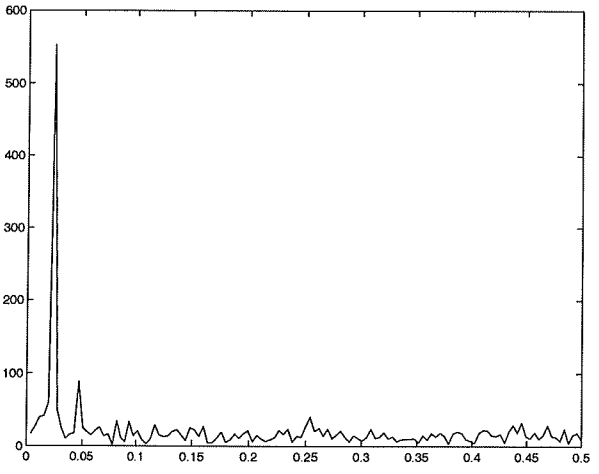
Řešení:

Popsaný problém je problémem nelineární regrese, neboť regresní funkce $m(t)$ je v parametru θ nelineární. Problém se obvykle řeší tak, že nejprve odhadneme pomocí Fourierovy transformace θ a poté metodou nejmenších čtverců odhadneme parametry A a B a z nich spočteme odhady C a ϕ.

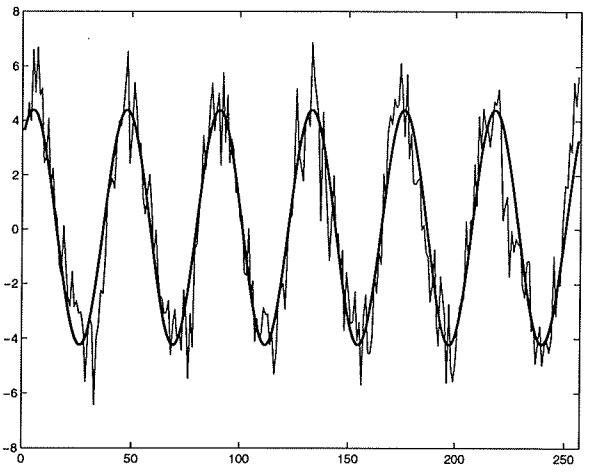
Nejprve odečteme od signálu průměr $\bar{S} = 0.0805$. Frekvenci θ odhadneme pomocí Fourierovy transformace

$$F(\lambda) = \sum_{t=1}^{n} Y_t \mathrm{e}^{-it\lambda}, \quad \lambda \in \langle -\pi, \pi \rangle .$$

Je známo, že modus Fourierovy transformace nabývá v případě, že $m(t) = C\cos(\theta t + \phi)$, velkou hodnotu pro $\lambda = \theta$. Pro výpočet Fourierovy transformace pro $n = 2^p$, $p \in N$ lze použít rychlý počítačový algoritmus, který se nazývá rychlá Fourierova transformace, zkráceně FFT. Rychlá Fourierova transformace počítá hodnoty Fourierovy transformace jen pro Fourierovy frekvence, tj. frekvence ve tvaru $\{\frac{2\pi j}{n}, j = 1, \ldots, n/2\}$. Znamená to, že i náš odhad budeme hledat jen mezi Fourierovými frekvencemi. Speciálně za odhad $\hat{\theta}$ vezmeme tu Fourierovu frekvenci $\lambda_0 = \frac{2\pi j_0}{n}$, pro kterou je hodnota modu Fourierovy tranformace největší, tedy $|F(\frac{2\pi j_0}{n})| = \max_{j=1,\ldots,n/2} |F(\frac{2\pi j}{n})|$.



Modus Fourierovy transformace.



Digitalizovaný signál a jeho odhadnutá hodnota.

Uvedený graf modu Fourierovy transformace zobrazuje funkci

$$\{|F(2\pi j/n)|\} \quad \text{pro} \quad \{j/n,\, j = 1, \ldots, n/2\}.$$

Z něho je patrné, že modus Fourierovy transformace dosáhl maxima pro 6/256, to znamená pro úhlovou frekvenci $2\pi \cdot 6/256$. Za odhad úhlové frekvence θ vezmeme proto $\hat{\theta} = 12\pi/256 \doteq 0.14726$. Poznamenejme, že se někdy místo modu Fourierovy transformace pracuje s kvadrátem modu Fourierovy transformace, který se nazývá periodogram. (Jde až na normující konstantu o funkci $I(\lambda)$ ze skript Jarušková (1996).)

Poté, co jsme odhadli parametr θ pomocí $\hat{\theta}$, je problém odhadu parametrů A a B úlohou lineární regrese s dvěma regresory

$$\{\cos(\hat{\theta}t),\ t = 1, \ldots, n\} \quad \text{a} \quad \{\sin(\hat{\theta}t),\ t = 1, \ldots, n\}$$

bez konstantního členu. Odhady $\hat{A}$ a $\hat{B}$ můžeme tedy najít jako řešení soustavy normálních rovnic:

$$\begin{aligned} A \cdot \sum_{t=1}^{n} \cos^2(\hat{\theta}t) + B \cdot \sum_{t=1}^{n} \cos(\hat{\theta}t)\sin(\hat{\theta}t) &= \sum_{t=1}^{n} Y_t \cos(\hat{\theta}t)\,, \\ A \cdot \sum_{t=1}^{n} \cos(\hat{\theta}t)\sin(\hat{\theta}t) + B \cdot \sum_{t=1}^{n} \sin^2(\hat{\theta}t) &= \sum_{t=1}^{n} Y_t \sin(\hat{\theta}t)\,. \end{aligned}$$

V našem případě, kdy $\hat{\theta}$ patří mezi Fourierovy frekvence, je řešení soustavy normálních rovnic obzvlášť jednoduché:

$$\hat{A} = \frac{2}{n}\sum_{t=1}^{n} Y_t \cos(\hat{\theta}t) = \frac{2}{n}\,\mathrm{Re}\left(F\left(\frac{12\pi}{256}\right)\right) = 3.1835,$$

$$\hat{B} = \frac{2}{n}\sum_{t=1}^{n} Y_t \sin(\hat{\theta}t) = \frac{2}{n}\,\mathrm{Im}\left(F\left(\frac{12\pi}{256}\right)\right) = 2.9165.$$

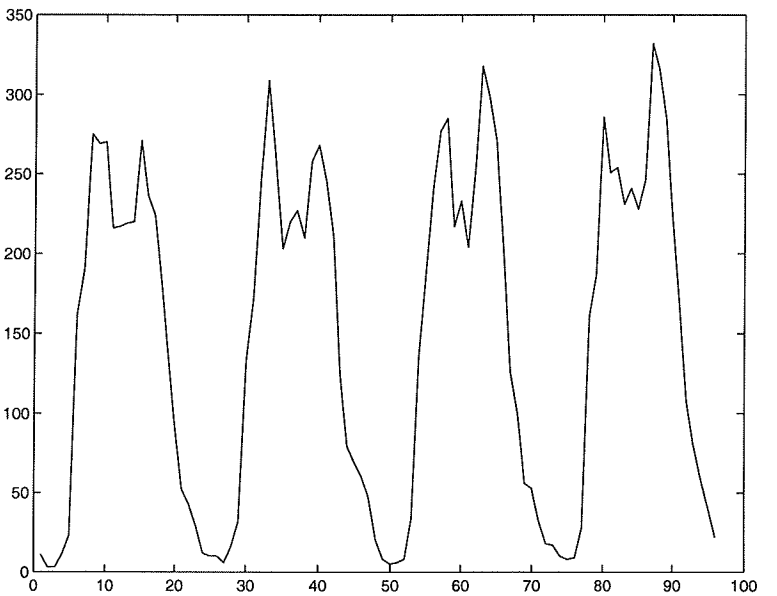
Poznamenejme, že $\mathrm{Re}(z)$ značí reálnou a $\mathrm{Im}(z)$ imaginární část komplexního čísla z. Amplitudu C odhadneme jako $\hat{C} = \sqrt{\hat{A}^2 + \hat{B}^2} = 4.3175$ a fázové posunutí ϕ pomocí $\hat{\phi} = \mathrm{arctg}(-\hat{B}/\hat{A}) = -0.74166$. Digitalizovaný signál spolu s jeho odhadnutou střední hodnotou $\bar{S} + \hat{C}\cos(\hat{\theta}\,t + \hat{\phi})$ je zobrazen na shora uvedeném obrázku.

Čtenáři, který se o problematiku periodických časových řad více zajímá, doporučujeme knihy Anděl (1976) nebo Cipra (1986).

Příklad 12.6.

Ve čtyřech po sobě jdoucích všedních dnech roku 1990 byly na silnici 1. třídy nedaleko Kolína zjištěny hodinové intenzity dopravy (tj. počty vozidel, které projely pozorovacím stanovištěm během určité hodiny). Tyto intenzity jsou uvedeny v tabulce na další straně, řada je též zobrazena na následujícím obrázku. Navrhněte pro tato data vhodný periodický model.

hodina	pondělí 22. 1.	úterý 23. 1.	středa 24. 1.	čtvrtek 25. 1.
00-01	11	10	8	17
01-02	3	10	5	10
02-03	3	6	6	8
03-04	11	16	8	9
04-05	23	32	34	28
05-06	162	133	135	161
06-07	191	172	188	187
07-08	275	248	240	286
08-09	269	309	277	251
09-10	270	259	285	254
10-11	216	203	217	231
11-12	217	220	233	241
12-13	219	227	204	228
13-14	220	210	252	246
14-15	271	258	318	332
15-16	236	268	298	316
16-17	224	246	271	285
17-18	182	211	203	221
18-19	136	125	126	167
19-20	91	79	100	108
20-21	52	69	56	80
21-22	43	60	53	59
22-23	29	47	32	41
23-24	12	20	18	22



Řešení:

Předpokládejme, že hodinové intenzity dopravy tvoří časovou řadu

$$Y_t = m(t) + e_t, \qquad t = 1, 2, \ldots, n,$$

kde $n = 4 \cdot 24 = 96$, $\{e_t\}$ jsou nezávislé náhodné chyby s nulovou střední hodnotou a shodným rozptylem a $m(t)$ je regresní funkce tvaru

$$m(t) = m + \sum_{j=1}^{p} \big(A_j \cos(\theta_j t) + B_j \sin(\theta_j t)\big).$$

Pokusme se nejprve intuitivní úvahou odhadnout, jaké frekvence θ_j mohou být pro naše data významné.

Je zcela zřejmé, že časová řada hodinových intenzit se zhruba opakuje s periodou $T_1 = 24$ hodin. Této periodě odpovídá frekvence $\lambda_1 = 1/24$, tj. $\theta_1 = 2\pi/24$. Prohlédneme-li si pozorně zobrazená data, vidíme, že se během jednoho dne vyskytují dva vrcholy, tj. jejich průběh nemůže být zcela vystižen funkcí s periodou 24. Je rozumné předpokládat (a zkušenosti to potvrzují), že automobilová doprava kolísá též s periodou, odpovídající pracovnímu cyklu, tj. $T_2 = 8$ hodin. Této periodě odpovídá frekvence $\lambda_2 = 1/8$, tj. $\theta_2 = 2\pi/8$.

Zvolíme proto funkci

$$m(t) = m + A_1 \cos(2\pi t/24) + B_1 \sin(2\pi t/24) + A_2 \cos(2\pi t/8) + B_2 \sin(2\pi t/8).$$

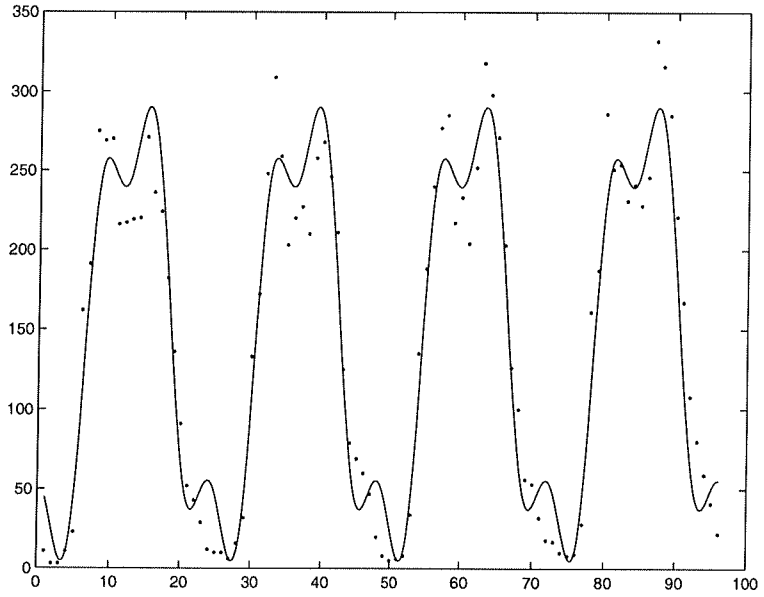
Parametry modelu pak odhadneme následovně:

$$\begin{aligned} m &= \bar{y} = 147.4896, \\ A_1 &= \frac{2}{n} \sum_{t=1}^{n} Y_t \cos(\theta_1 t) = -133.9055, \\ B_1 &= \frac{2}{n} \sum_{t=1}^{n} Y_t \sin(\theta_1 t) = -32.1957, \\ A_2 &= \frac{2}{n} \sum_{t=1}^{n} Y_t \cos(\theta_2 t) = 41.6792, \\ B_2 &= \frac{2}{n} \sum_{t=1}^{n} Y_t \sin(\theta_2 t) = 8.2075. \end{aligned}$$

Data, zobrazená tentokrát jako body, jsou spolu s regresní funkcí

$$\begin{aligned} m(t) = 147.4896 &- 133.9055 \cos(2\pi t/24) - 32.1957 \sin(2\pi t/24) + \\ &+ 41.6792 \cos(2\pi t/8) + 8.2075 \sin(2\pi t/8) \end{aligned}$$

znázorněna na dalším obrázku.



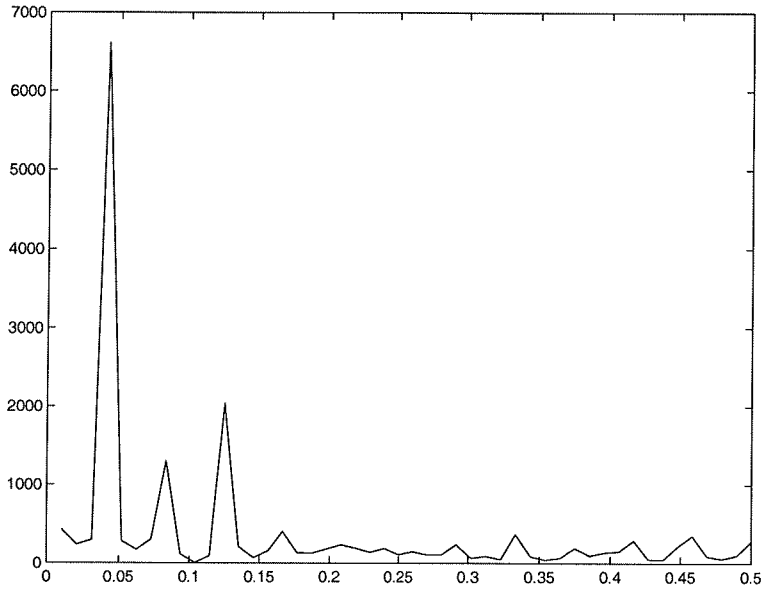
Pokud bychom se nechtěli spokojit s intuitivním odhadem frekvencí, mohli bychom je odvozovat z periodogramu, resp. z modu Fourierovy transformace. Na posledním obrázku je graf funkce

$$|F(2\pi j/n)|, \quad \text{pro } j = 1, 2, \ldots, n/2,$$

kde

$$F(\lambda) = \sum_{t=1}^{n} Y_t \mathrm{e}^{-it\lambda}.$$

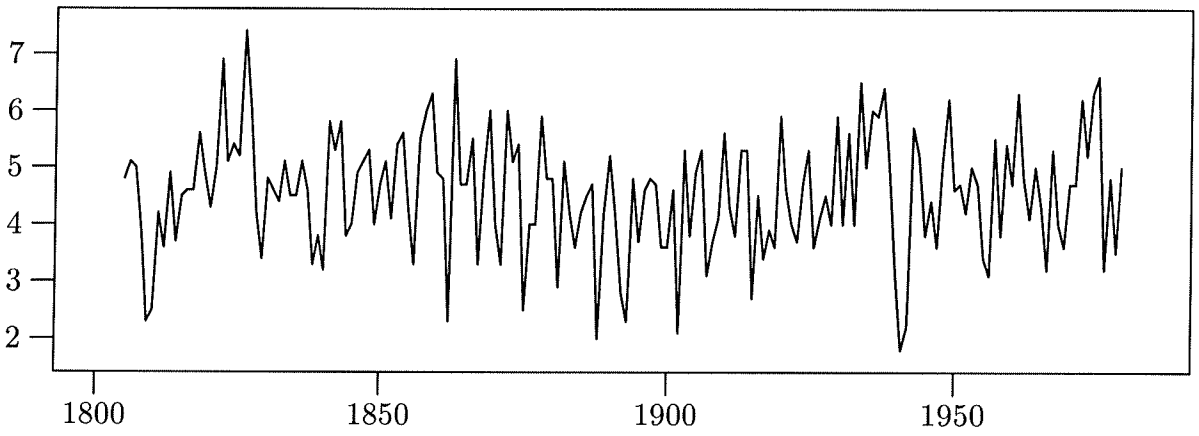
Z něho vidíme, že intuitivně odhadnutým frekvencím λ_1 a λ_2 skutečně odpovídají dva nejvyšší vrcholky modu Fourierovy transformace. Třetí nejvyšší vrcholek by pak odpovídal frekvenci $\lambda_3 = 1/16$, tj. periodě 16 hodin.



Příklad 12.7.

V následující tabulce jsou dány průměrné roční teploty v St. Peterburgu v letech 1805 – 1979, viz obrázek na další straně. Rozhodněte, zda se jedná o řadu nezávislých pozorování nebo zda existuje závislost mezi pozorováními, která jsou si v čase blízká.

rok	teplota	rok	teplota	rok	teplota	rok	teplota	rok	teplota
1805	4.8	1840	3.2	1875	2.5	1910	5.6	1945	3.8
1806	5.1	1841	5.8	1876	4.0	1911	4.3	1946	4.4
1807	5.0	1842	5.3	1877	4.0	1912	3.8	1947	3.6
1808	3.9	1843	5.8	1878	5.9	1913	5.3	1948	5.1
1809	2.3	1844	3.8	1879	4.8	1914	5.3	1949	6.2
1810	2.5	1845	4.0	1880	4.8	1915	2.7	1950	4.6
1811	4.2	1846	4.9	1881	2.9	1916	4.5	1951	4.7
1812	3.6	1847	5.1	1882	5.1	1917	3.4	1952	4.2
1813	4.9	1848	5.3	1883	4.2	1918	3.9	1953	5.0
1814	3.7	1849	4.0	1884	3.6	1919	3.6	1954	4.7
1815	4.5	1850	4.7	1885	4.2	1920	5.9	1955	3.4
1816	4.6	1851	5.1	1886	4.5	1921	4.6	1956	3.1
1817	4.6	1852	4.1	1887	4.7	1922	4.0	1957	5.5
1818	5.6	1853	5.4	1888	2.0	1923	3.7	1958	3.8
1819	4.9	1854	5.6	1889	4.2	1924	4.7	1959	5.4
1820	4.3	1855	4.3	1890	5.2	1925	5.3	1960	4.7
1821	5.0	1856	3.3	1891	4.1	1926	3.6	1961	6.3
1822	6.9	1857	5.5	1892	2.8	1927	4.1	1962	4.8
1823	5.1	1858	6.0	1893	2.3	1928	4.5	1963	4.1
1824	5.4	1859	6.3	1894	4.8	1929	4.0	1964	5.0
1825	5.2	1860	4.9	1895	3.7	1930	5.9	1965	4.3
1826	7.4	1861	4.8	1896	4.6	1931	4.0	1966	3.2
1827	6.1	1862	2.3	1897	4.8	1932	5.6	1967	5.3
1828	4.2	1863	6.9	1898	4.7	1933	4.0	1968	4.0
1829	3.4	1864	4.7	1899	3.6	1934	6.5	1969	3.6
1830	4.8	1865	4.7	1900	3.6	1935	5.0	1970	4.7
1831	4.6	1866	5.5	1901	4.6	1936	6.0	1971	4.7
1832	4.4	1867	3.3	1902	2.1	1937	5.9	1972	6.2
1833	5.1	1868	4.9	1903	5.3	1938	6.4	1973	5.2
1834	4.5	1869	6.0	1904	3.8	1939	5.0	1974	6.3
1835	4.5	1870	4.0	1905	4.9	1940	3.1	1975	6.6
1836	5.1	1871	3.3	1906	5.3	1941	1.8	1976	3.2
1837	4.6	1872	6.0	1907	3.1	1942	2.2	1977	4.8
1838	3.3	1873	5.1	1908	3.7	1943	5.7	1978	3.5
1839	3.8	1874	5.4	1909	4.1	1944	5.2	1979	5.0



Průměrné roční teploty v St. Peterburgu v letech 1805 – 1979.

Řešení:

Základními charakteristikami stacionární časové řady je její střední hodnota $\mathrm{E}\,X$, směrodatná odchylka $\mathrm{sd}\,X$ (resp. rozptyl $\mathrm{Var}\,X$) a autokorelační funkce

$$\rho(k) = \frac{\mathrm{E}(X_i - \mathrm{E}\,X)(X_{i+k} - \mathrm{E}\,X)}{\mathrm{Var}\,X}, \qquad k = 1, 2, \ldots .$$

Hodnota autokorelační funkce $\rho(k)$ vyjadřuje míru lineární závislosti mezi dvěma pozorováními, které jsou v čase vzdáleny o k časových jednotek. Jako odhady shora uvedených charakteristik slouží: výběrový průměr, výběrová směrodatná odchylka a výběrová autokorelační funkce. Pro posunutí (anglicky „lag“) $k = 0, 1, 2, \ldots$, spočteme hodnotu výběrové autokorelační funkce $r(k)$ takto:

$$r(k) = \frac{\frac{1}{n}\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

kde $\{x_i,\ i = 1, \ldots, n\}$ jsou napozorované hodnoty určitého úseku časové řady $\{X_i\}$ a $\bar{x}$ je aritmetický průměr spočtený z $\{x_i,\ i = 1, \ldots, n\}$. Pokud je časová řada $\{X_i\}$ tvořena nezávislými veličinami, pak pro všechna $k \neq 0$ platí $\rho(k) = 0$. Odtud vyplývá, že i odhady $r(k)$, $k = 1, 2, \ldots$ by měly nabývat malých hodnot. Obvykle se považují hodnoty $r(k)$ za malé, jestliže $|r(k)| < 2/\sqrt{n}$. Nerovnost je odvozena z úvahy, že pokud $\{X_i\}$ jsou nezávislé veličiny, pak má $r(k)$ přibližně normální rozdělení $N(0,\ 1/n)$, a tedy $P(|r(k)| > 2/\sqrt{n}) \doteq 0.05$. Pokud však současně zjišťujeme, zda $|r(1)| > 2/\sqrt{n}$, $|r(2)| > 2/\sqrt{n}$, $\ldots$, $|r(l)| > 2/\sqrt{n}$ pro nějaké $l << n$, pak P(alespoň pro jedno $k \leq l$ platí: $|r(k)| > 2/\sqrt{n}$) je vyšší než 0.05. Je tudíž třeba být v zamítání nezávislosti opatrní.

Výběrový průměr spočtený ze shora uvedených dat se rovná $\bar{x} = 4.547$ a výběrová směrodatná odchylka $s = 1.044$. Následující tabulka udává hodnoty výběrové autokorelační funkce pro posunutí $k = 0, \ldots, 23$:

0	1.00000	6	0.01550	12	-0.06870	18	-0.02637
1	0.13434	7	-0.02428	13	0.00129	19	-0.02867
2	0.02991	8	-0.03632	14	0.04337	20	-0.04961
3	-0.08686	9	0.00017	15	0.06298	21	-0.01307
4	0.06723	10	0.06099	16	0.00490	22	-0.02145
5	0.09076	11	0.06327	17	-0.08097	23	0.05064

Žádná z hodnot výběrové autokorelační funkce pro $k = 1, 2, 3, \ldots$ není větší než $2/\sqrt{n} = 0.1512$. Z toho usuzujeme, že průměrné roční teploty je možno považovat za nezávislé náhodné veličiny. (Přesněji řečeno: Neprokázali jsme, že by průměrné roční teploty byly zkorelovány.)

Příklad 12.8.

Tabulka udává průměrné roční průtoky Labe v Děčíně v m^3/s měřené v letech 1852-1988 (údaje jsou uvedeny po sloupcích).

353.917	170.000	282.073	190.024	337.197	450.744
339.250	417.833	288.370	274.726	340.891	332.451
339.750	293.500	445.073	385.590	321.433	265.767
458.417	260.167	295.930	351.172	253.690	317.140
274.250	284.667	270.069	273.444	358.316	284.575
240.500	323.583	222.898	545.645	218.670	195.625
181.333	383.417	254.736	367.119	173.174	167.163
259.167	242.083	278.939	271.741	230.224	231.990
347.000	386.750	310.240	216.187	205.280	398.470
254.583	230.083	325.237	189.496	260.429	253.974
231.250	231.333	243.365	389.716	196.180	348.504
171.750	303.500	282.802	296.004	364.424	338.852
189.333	209.083	380.201	186.337	295.185	377.814
244.583	375.133	358.882	140.565	380.262	489.922
151.250	312.721	283.146	241.032	397.239	429.044
447.667	391.045	287.989	271.047	243.216	417.303
336.417	340.726	313.104	332.096	252.520	283.990
208.167	300.317	414.151	339.767	304.877	215.476
298.667	239.517	427.512	380.101	278.767	241.123
357.250	264.854	342.703	532.498	169.753	325.236
175.250	353.909	170.561	685.822	179.893	472.419
171.583	393.007	259.814	373.824	493.982	393.169
165.417	422.820	460.131	171.741	433.657	

Zjistěte, zda je řadu možno považovat za posloupnost nezávislých veličin. V případě, že nikoliv, rozhodněte, zda je možné data lépe modelovat pomocí AR(1) nebo MA(1) modelu. Využijte zvolený model pro predikci průměrného ročního průtoku v roce 1989.

Řešení:

Nejprve od všech naměřených dat odečtěme celkový průměr $\bar{x} = 305.4776$. Řadu $\{Y_t\}$, která tak vznikne, můžeme považovat za stacionární se střední hodnotou rovnou 0. Na obrázku je zobrazena její výběrová autokorelační funkce pro 25 prvních posunutí.

Hodnota výběrové autokorelační funkce pro posunutí $k = 1$ je rovna $r(1) \doteq 0.335$, a tudíž výrazně překračuje $2/\sqrt{n} \doteq 0.171$. Odtud vyplývá, že roční průtoky nemohou být považovány za nezávislé náhodné veličiny. Optimální model vybíráme z modelů AR(1) a MA(1):

$$Y_t = a\,Y_{t-1} + e_t, \quad |a| < 1,$$
$$Y_t = e_t + b\,e_{t-1}, \quad |b| < 1,$$

kde $\{e_t\}$ jsou náhodné chyby, o kterých se obvykle předpokládá, že to jsou nezávislé stejně rozdělené náhodné veličiny s $\mathrm{E}\,e_i = 0$ a $\mathrm{Var}\,e_i = \sigma^2$. Modely AR(1) a MA(1) jsou speciálním případem modelů ARMA(p, q) splňujících vztah:

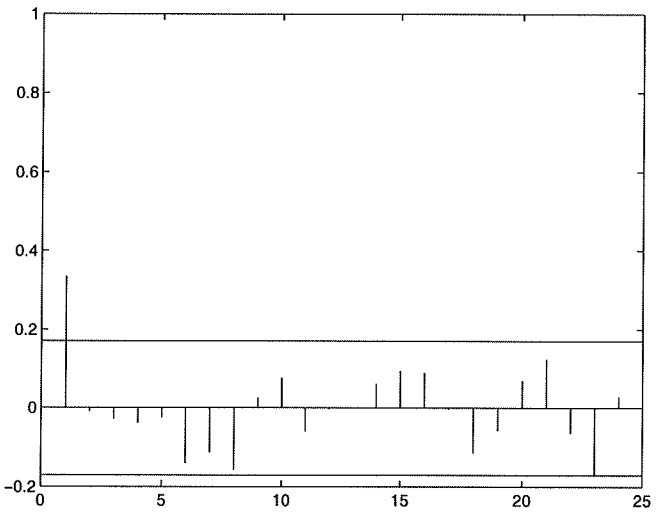
$$Y_t = a_1\,Y_{t-1} + \cdots + a_p\,Y_{t-p} + e_t + b_1\,e_{t-1} + \ldots\, b_q\,e_{t-q}.$$

Pro teoretickou autokorelační funkci $\{\rho(k),\ k = 0, \ldots\}$ stacionární autoregresní posloupnosti AR(1) platí:

$$\rho(k) = a^k, \quad \text{kde} \quad |a| < 1,$$

zatímco pro posloupnost klouzavých součtů MA(1) platí:

$$\rho(1) = b \quad \text{a} \quad \rho(k) = 0 \quad \text{pro} \quad k = 2, 3, \ldots.$$



Výběrová autokorelační funkce ročních průměrných průtoků Labe a vyznačené meze $\pm 2/\sqrt{n}$.

Výběrová autokorelační funkce by měla mít do jisté míry podobné vlastnosti. Prohlédneme-li si dobře výběrovou autokorelační funkci, zdá se, že MA(1) bude lepším modelem. Objektivněji rozhodneme, použijeme-li některé informační kritérium, např.

Akaikeho kritérium. Akaikeho kritérium vybírá z množiny modelů ARMA(p, q) ten model, pro který je $AIC(k)$:

$$AIC(k) = n \ln \widehat{\sigma^2} + 2k$$

minimální. Přirozené číslo $k = p + q$ označuje počet parametrů modelu (pro AR(1) i MA(1) je $k = 1$) a $\widehat{\sigma^2}$ je maximálně věrohodný odhad rozptylu σ^2 náhodných chyb $\{e_t\}$. Najít maximálně věrohodný odhad rozptylu σ^2 je velmi obtížné, a proto se častěji používá metoda nejmenších čtverců, a to buď podmíněná nebo nepodmíněná (více Cipra (1986)). Podmíněná metoda nejmenších čtverců odhaduje autoregresní koeficient a a rozptyl σ^2 v modelu AR(1) následovně:

$$\hat{a} = \arg\min_a \sum_{t=2}^{n} (Y_t - a\,Y_{t-1})^2 \quad \text{a} \quad \widehat{\sigma^2} = \frac{1}{n-1} \sum_{t=2}^{n} (Y_t - \hat{a}\,Y_{t-1})^2,$$

zatímco koeficient b a rozptyl σ^2 v modelu MA(1) takto:

$$\hat{b} = \arg\min_b \sum_{t=2}^{n} (Y_t - b\,\hat{e}_{t-1})^2$$

kde $\hat{e}_1 = Y_1$, $\hat{e}_2 = Y_2 - b\,\hat{e}_1 = Y_2 - b\,Y_1$, $\hat{e}_3 = Y_3 - b\,\hat{e}_2 = Y_3 - b\,(Y_2 - b\,Y_1)$, ...,

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{t=2}^{n} (Y_t - \hat{b}\,\hat{e}_{t-1})^2, \text{ kde } \hat{\hat{e}}_1 = Y_1, \ \hat{\hat{e}}_2 = Y_2 - \hat{b}\,Y_1, \ \hat{\hat{e}}_3 = Y_3 - \hat{b}\,(Y_2 - \hat{b}\,Y_1).$$

V našem případě jsme metodou podmíněných nejmenších čtverců dostali pro model AR(1): $\hat{a} = 0.3376$ a $\widehat{\sigma^2} = 7756.8$ a pro model MA(1): $\hat{b} = 0.39398$ a $\widehat{\sigma^2} = 7599.8$.

Poznamenejme, že pro model MA(1) je minimalizace nejmenších čtverců nelineární úloha, kterou je třeba řešit nějakou iterační numerickou metodou.

Vzhledem k tomu, že pro oba studované modely je k rovno jedné, vybereme z nich ten model, pro který je menší $\widehat{\sigma^2}$, tj. v našem případě MA(1). Přičteme-li ještě na počátku odečtený průměr, modelujeme časovou řadu pomocí:

$$305.4776 + e_t + 0.39398\,e_{t-1}, \qquad \text{kde} \quad \mathrm{E}\,e_t = 0, \ \mathrm{Var}\,e_t = 7599.8.$$

Tento model použijeme také pro predikci průměrného ročního průtoku Labe v roce 1989. Předpokládáme-li, že časová řada tvoří MA(1) posloupnost, pak pro nalezení predikce $\hat{Y}_{t+1}(t)$ pro čas $t + 1$ na základě znalosti hodnot časové řady až do času t použijeme rekurentní vztah:

$$\hat{Y}_{t+1}(t) = \hat{b}(Y_t - \hat{Y}_t(t-1))$$

s počáteční podmínkou $\hat{Y}_2(1) = 0$. Připomeňme, že řada $\{Y_t,\ t = 1, 2, \dots\}$ je naměřenou časovou řadou, od které jsme odečetli celkový průměr. V našem případě $\hat{Y}_{1989}(1988) \doteq$ $\doteq 10.75$. Průměrný průtok pro rok 1989 pak předpovíme hodnotou $\bar{x} + \hat{Y}_{1989}(1988) \doteq$ $\doteq 305.48 + 10.75 = 316.23$ m^3/s.

Odpověď: Průměrné roční průtoky Labe v Děčíně je vhodné modelovat posloupností:

$$305.4776 + e_t + 0.39398\,e_{t-1}, \qquad \text{kde } \mathrm{E}\,e_t = 0, \ \mathrm{Var}\,e_t = 7599.8.$$

Na základě tohoto modelu předpovídáme, že průměrný roční průtok v roce 1989 bude roven 316.23 m^3/s.

Neřešené příklady

Příklad 12.9.

V časových okamžicích $t = 1, 2, \ldots, n$ pozorujeme řadu $Y_1, Y_2, \ldots, Y_n$. Odhadněte metodou nejmenších čtverců parametry c a b v modelu:

$$Y_t = \begin{cases} c + e_t, & \text{pro } t \leq k; \\ c + b\,(t-k) + e_t, & \text{pro } t > k; \end{cases}$$

kde $\{e_t\}$ jsou náhodné chyby. Problém spočívá v proložení dat funkcí, která je na intervalu $\langle 0, k\rangle$ konstantní a na intervalu $(k, n\rangle$ lineární a zároveň je v bodě k spojitá.

Aplikujte na průměrné roční teploty měřené v Klementinu (viz příklad 12.1), kde za k vezměte index odpovídající roku 1900.

Příklad 12.10.

Tabulka udává vývoj průměrného věku mužů a žen v České republice od roku 1947 do roku 1996.

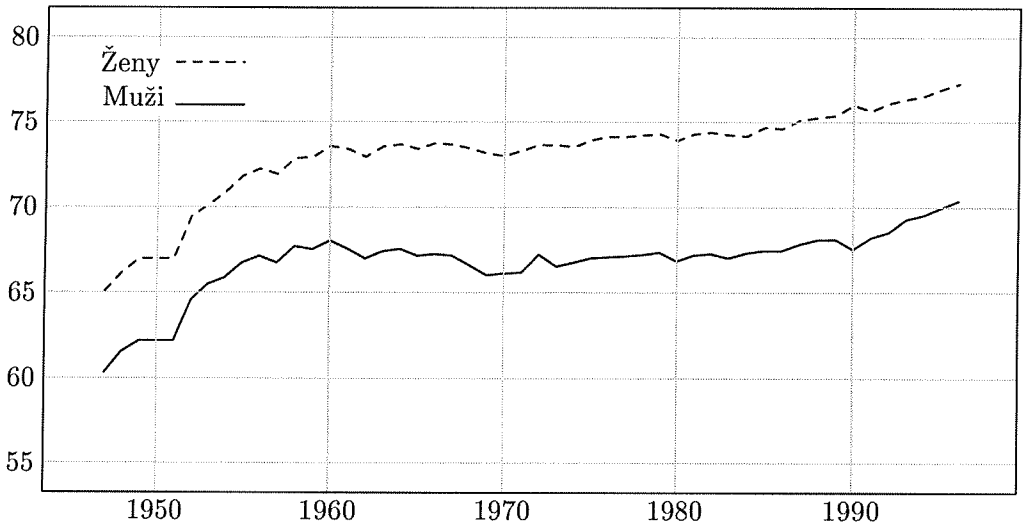
rok	muži	ženy
1947	60.30	65.05
1948	61.54	66.16
1949	62.16	66.97
1950	62.16	66.97
1951	62.16	66.97
1952	64.56	69.41
1953	65.49	70.11
1954	65.86	70.81
1955	66.74	71.79
1956	67.13	72.22
1957	66.74	71.92
1958	67.70	72.85
1959	67.53	72.94
1960	68.03	73.58
1961	67.55	73.41
1962	66.99	72.95
1963	67.41	73.56
1964	67.55	73.68
1965	67.15	73.42
1966	67.25	73.76
1967	67.16	73.67
1968	66.60	73.46
1969	66.02	73.15
1970	66.12	73.01
1971	66.18	73.32

rok	muži	ženy
1972	67.22	73.66
1973	66.52	73.64
1974	66.76	73.54
1975	67.01	73.94
1976	67.07	74.14
1977	67.13	74.14
1978	67.21	74.23
1979	67.35	74.30
1980	66.84	73.92
1981	67.18	74.30
1982	67.27	74.40
1983	67.02	74.25
1984	67.31	74.18
1985	67.46	74.70
1986	67.46	74.62
1987	67.83	75.13
1988	68.09	75.27
1989	68.11	75.40
1990	67.54	76.01
1991	68.21	75.67
1992	68.52	76.11
1993	69.28	76.35
1994	69.53	76.55
1995	69.96	76.94
1996	70.37	77.27

Od roku 1960 se zdá, že průměrná délka života Z_t roste s časem kvadraticky:

$$Z_t = a + bt + ct^2.$$

Budeme-li předpokládat, že průměrná délka života poroste i nadále tímto způsobem, odhadněte průměrnou délku života v roce 2000.



Příklad 12.11.

Uvažujme časovou řadu $\{Y_t\}$, jejíž deterministická složka se rovná $\{\sin \frac{2\pi t}{12}\}$ a náhodná složka $\{e_t\}$ má střední hodnotu nulovou a konstantní rozptyl σ^2:

$$Y_t = m(t) + e_t, \quad \text{kde} \quad m(t) = \sin \frac{2\pi t}{12}.$$

Tuto řadu jsme napozorovali v časech $t = 0, 1, \ldots, 12$. Rozhodli jsme se ji vyhladit klouzavými průměry

$$\hat{m}(t) = \frac{Y_{t-1} + Y_t + Y_{t+1}}{3}.$$

Jakou střední hodnotu a jaký rozptyl budou mít veličiny $\{\hat{m}(t),\ t = 1, \ldots, 11\}$? Ve kterém časovém okamžiku t nedochází při vyhlazení k systematickému zkreslení a ve kterém dochází naopak k největšímu zkreslení?

Příklad 12.12.

Zjistěte, zda průměrné roční teploty měřené v Klementinu v letech 1807-1899 (viz příklad 12.1) mohou být považovány za nezávislé náhodné veličiny.

Příklad 12.13.

Během čtyř dnů za stálého počasí byla v šestihodinových intervalech měřena na určitém místě teplota vzduchu ve stupních Celsia (první hodnota je z šesti hodin ráno prvního dne): 11, 21, 17, 9, 9, 19, 16, 9, 12, 23, 20, 11, 10, 13, 21, 9. Najděte vhodný periodický model pro tuto časovou řadu. Odhadněte teplotu pro další den ve 3 hodiny odpoledne (za předpokladu, že počasí vydrží). Odhadněte též, ve kterou denní dobu daného období dosahovala teplota svého maxima.

Příklad 12.14.

V devíti po sobě jdoucích dnech po havárii byla měřena radioaktivita na určitém místě: 1578.2, 952.8, 568.4, 352.7, 213.1, 128.8, 78.5, 48.2, 28.9. Navrhněte exponenciální regresní model pro tuto časovou řadu, odhadněte velikost radioaktivity pro následující (tj. desátý) den. Odhadněte, který den radioaktivita klesne poprvé pod 5 jednotek.

Příklad 12.15.

V následující tabulce je zachycen vývoj kursu akcie Škody Plzeň na burze cenných papírů v Praze v jednotlivých burzovních dnech období od 29. června 1993 do 9. června 1994 (kursy v Kč jsou chronologicky seřazeny po sloupcích). Proveďte analýzu této časové řady akciogramem podobně jako v příkladu 12.4. Porovnejte četnosti signálů k nákupu a prodeji, když zvolíte pro dlouhý průměr $\alpha = 0.1$, resp. $\alpha = 0.2$.

400	300	398	520	700	815	720	726	700	650	635
360	360	477	624	770	815	750	700	680	650	572
324	330	572	700	810	800	750	700	685	650	520
292	330	686	650	800	800	750	750	650	655	468
234	300	823	677	820	800	750	750	650	665	423
187	315	987	680	845	770	730	730	650	650	425
224	345	790	670	761	847	700	740	650	650	450
250	360	632	685	810	800	660	705	650	645	450

Příklad 12.16.

Ze 400 členů stacionární časové řady byla vypočtena výběrová autokovarianční funkce (uvádíme jen několik prvních členů):

$$\hat{R}(0) = 1.9936, \quad \hat{R}(1) = 0.4017, \quad \hat{R}(2) = -1.2156,$$
$$\hat{R}(3) = -0.6995, \quad \hat{R}(4) = 0.5982, \quad \hat{R}(5) = -0.2361.$$

Rozhodneme-li se data považovat za realizaci úseku autoregresní posloupnosti nejvýše třetího řádu, jaký řád je optimální zvolit?

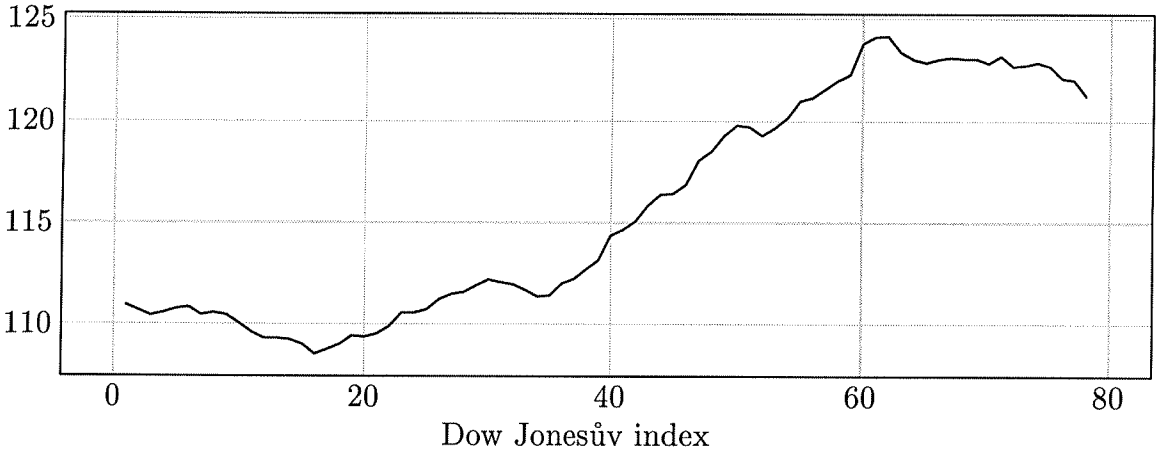
Příklad 12.17.

Následující tabulka udává hodnoty Dow Jonesova burzovního indexu ve dnech 28. srpna až 18. prosince 1972. Údaje jsou řazeny postupně do sloupců. Vyhlaďte časovou řadu pomocí klouzavých průměrů a pomocí normálního jádra.

110.94	109.60	109.53	112.06	114.65	119.70	124.11	123.18
110.69	109.31	109.89	111.96	115.06	119.28	124.14	122.67
110.43	109.31	110.56	111.68	115.86	119.66	123.37	122.73
110.56	109.25	110.56	111.36	116.40	120.14	123.02	122.86
110.75	109.02	110.72	111.42	116.44	120.97	122.86	122.67
110.84	108.54	111.23	112.00	116.88	121.13	123.02	122.09
110.46	108.77	111.48	112.22	118.07	121.55	123.11	122.00
110.56	109.02	111.58	112.70	118.51	121.96	123.05	121.23
110.46	109.44	111.90	113.15	119.28	122.26	123.05	
110.05	109.38	112.19	114.36	119.79	123.79	122.83	

Příklad 12.18.

Z obrázku je patrné, že časová řada popisující chování Dow Jonesova burzovního indexu není stacionární. Řadu lze stacionarizovat tím, že utvoříme řadu postupných diferencí. Pro řadu těchto diferencí najděte pomocí Akaikeho kritéria nejlepší model mezi modely AR(1), AR(2), MA(1), MA(2) a ARMA(1,1). Nalezený model použijte pro predikci hodnoty burzovního indexu pro následující den.



Dow Jonesův index

LITERATURA

[1] Anděl J., *Matematická statistika*, SNTL – Nakladatelství technické literatury, Praha, 1985.

[2] Anděl J., *Statistická analýza časových řad*, SNTL – Nakladatelství technické literatury, Praha, 1976.

[3] Antoch J., Vorlíčková D., *Vybrané metody statistické analýzy dat*, Academia, Praha, 1992.

[4] Beneš V., Dohnal G., *Pravděpodobnost a matematická statistika – doplňkové skriptum*, Vydavatelství ČVUT, Praha, 1993.

[5] Cipra T., *Analýza časových řad s aplikacemi v ekonomii*, SNTL – Nakladatelství technické literatury, Praha, 1986.

[6] Cipra T., *Praktický průvodce finanční a pojistnou matematikou*, Edice HZ, Praha, 1995.

[7] Hátle J., Likeš.J., *Základy počtu pravděpodobnosti a matematické statistiky*, SNTL – Nakladatelství technické literatury, Praha, 1974.

[8] Jarušková D., *Matematická statistika*, Vydavatelství ČVUT, Praha, 1996.

[9] Likeš J., Machek J., *Počet pravděpodobnosti*, SNTL – Nakladatelství technické literatury, Praha, 1981.

[10] Likeš J., Machek J., *Matematická statistika*, SNTL – Nakladatelství technické literatury, Praha, 1988.

[11] Nacházel K., *Stochastické metody ve vodním hospodářství – doplňkové skriptum*, Vydavatelství ČVUT, Praha, 1993.

[12] Rektorys K., *Přehled užité matematiky II*, Nakladatelství Prometheus, Praha, 1995.

[13] Rublík F., *Základy pravdepodobnosti a štatistiky*, Vydavatelství Alfa, Bratislava, 1983.

[14] Stuchlý J., *Matematika IV*, Vydavatelství technické a ekonomické literatury, Bratislava, 1981.

[15] Svešnikov A. A., *Sbírka úloh z teorie pravděpodobnosti, matematické statistiky a teorie náhodných funkcí*, SNTL – Nakladatelství technické literatury, Praha, 1971.

[16] Vorlíček M., Holický M., Špačková M., *Pravděpodobnost a matematická statistika pro inženýry – skriptum*, Ediční středisko ČVUT, Praha, 1982.

[17] Vorlíček M., Holický M., Špačková M., *Numerické tabulky ke skriptu Matematická statistika – doplňkové skriptum*, Vydavatelství ČVUT, Praha, 1994.

[18] Zvára K., *Regresní analýza*, Academia, Praha, 1989.

prof. RNDr. Daniela Jarušková, CSc., RNDr. Martin Hála, CSc.

PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA. Příklady

Vydalo České vysoké učení technické v Praze,
Česká technika – nakladatelství ČVUT, Thákurova 1, 160 41 Praha 6
v roce 2016 jako svou 11 729. publikaci.

Vytiskla Česká technika – nakladatelství ČVUT, výroba, Zikova 4, 166 36 Praha 6
146 stran
1. dotisk 3. vydání. Náklad 50 výtisků. Rozsah 11,76 AA, 12,09 VA

147 = VAKÁT

148 TIRÁŽ