PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA

Prof. RNDr. Daniela Jarušková, CSc.

2015
České vysoké učení technické v Praze

(1) TITULNÍ LIST

ISBN 978-80-01-04829-0

(2) RUB TTML · USM

Obsah

Poděkování

Skripta, která máte před sebou, jsou doplněna obrázky, které pro ně vytvořil Doc. RNDr. Jaromír Antoch, CSc. Chtěla bych mu tímto způsobem poděkovat za spoustu práce, které měl s jejich nakreslením. Zároveň bych chtěla poděkovat svým kolegům RNDr. Martinu Hálovi, CSc. a RNDr. Janě Noskové, Dr. za revizi textu a mnohé cenné připomínky.

Část I. Základní pojmy teorie pravděpodobnosti

1. Náhodné jevy

Ve fyzice jsme studovali mnoho pokusů, kde splnění určitého souboru podmínek má za následek výskyt určitého jevu. Zahřejeme-li vodu při normálním atmosferickém tlaku na 100 ° C , začne vřít. Magnetická střelka kompasu se na území naší republiky vždy ustálí v severojižním směru. Takovým pokusům říkáme *deterministické.*

V přírodě však existují i pokusy, jejichž výsledek i při zachování stejných podmínek může být pokaždé jiný. Takové pokusy se nazývají *náhodné.* Příkladem náhodného pokusu je hod kostkou. Budeme-li házet opakovaně stejnou kostkou, může nám padnout jednou šestka a podruhé třeba jednička.

Výsledku náhodného pokusu říkáme *náhodný jev.* Náhodné jevy budeme značit velkými písmeny $A, B, C, \ldots$

Mezi náhodnými jevy existují dva jevy, které mají zvláštní postavení a budeme je označovat vždy stejně. *Nemožný jev,* to znamená jev, který za daných podmínek nenastane nikdy, označujeme symbolem $\emptyset$. *Jistý jev,* to znamená jev, který za daných podmínek nastává vždy, označujeme symbolem Ω.

S náhodnými jevy můžeme provádět následující operace: *Průnikem* dvou jevů A a B (značíme $A \cap B$) nazýváme jev, který nastává právě tehdy, jestliže nastane jev A a současně jev B. *Sjednocením* dvou jevů A a B (značíme $A \cup B$) nazýváme náhodný jev, který nastane právě tehdy, jestliže nastane alespoň jeden z jevů A a B. Značí-li například jev A „padnutí sudého počtu ok na hrací kostce“ a jev B „padnutí většího počtu ok než čtyři“, znamená jev $A \cap B$ „padnutí šestky“ a jev $A \cup B$ „padnutí jednoho z čísel $\{2,4,5,6\}$“. Průnik a sjednocení můžeme rozšířit na libovolný (dokonce i nekonečný) počet náhodných jevů. Uvažujeme-li například jevy $A_1, A_2, \ldots, A_k$, značí jejich průnik $A_1 \cap A_2 \cap \cdots \cap A_k$ jev, který nastane právě tehdy, nastane-li všech k jevů současně.

O dvou náhodných jevech A a B řekneme, že jsou *disjunktní,* jestliže $A \cap B = \emptyset$, nebo, řečeno slovně, oba jevy nemohou nastat současně. Příkladem disjunktních jevů jsou například „padnutí šestky“ a „padnutí lichého počtu ok“ při jednom hodu kostkou.

Opačný jev k jevu A (značíme $\overline{A}$) nazýváme jev, který nastane právě tehdy, nenastane-li jev A. Jevy A a $\overline{A}$ jsou disjunktní a platí $A \cup \overline{A} = \Omega$. Označíme-li například jev A „padnutí sudého počtu ok na hrací kostce“, označuje jev $\overline{A}$ „padnutí lichého počtu ok“.

Na závěr článku o náhodných jevech zaveďme ještě pojem elementárního jevu. *Elementární jev* A je takový jev, který nelze vyjádřit jako sjednocení dvou jevů různých od A. Elementární jev si lze představit jako „nejjednodušší“ výsledek pokusu. Nechť například náhodný pokus spočívá v hodu kostkou, pak jevy A_1 - „padne jednička“, A_2 - „padne dvojka“, ..., A_6 - „padne šestka“ jsou elementární, zatímco jev C - „padne sudé číslo“ není elementárním jevem, neboť např. $C = E \cup F$, kde E značí jev „padne dvojka“ a F jev „padne čtyřka nebo šestka“.

2. Pravděpodobnost

Představme si, že opakovaně mnohokrát za sebou házíme kostkou, která má naprosto dokonalý tvar krychle a je vyrobena z homogenního materiálu. Budeme-li si zaznamenávat jako úspěšný pokus každý hod, ve kterém nám padne sudý počet ok, zjistíme, že se podíl počtu těchto úspěšných pokusů k celkovému počtu pokusů bude blížit 0.5 (50 %). Číslu 0.5 se obvykle říká pravděpodobnost jevu A „padnutí sudého počtu ok“. Ta se dá v tomto případě, kdy náhodný pokus může mít jen konečný počet výsledků se stejnou možností výskytu, spočítat následovně:

$$P(A) = \frac{m_A}{m},$$

to jest jako *podíl* m_A *- počtu výsledků příznivých jevu* A *ku* m *- počtu všech možných výsledků pokusu.* Protože v našem případě jsou tři příznivé výsledky $\{2, 4, 6\}$ a šest možných výsledků $\{1, 2, 3, 4, 5, 6\}$, je výsledná pravděpodobnost rovna 0.5 (50 %). Tato, takzvaná *klasická definice pravděpodobnosti*, však nepostačuje v případě, jestliže pokus může mít nekonečně mnoho výsledků. Pokusem s nekonečným počtem výsledků je například hod šipkou na terč. Výsledek pokusu, tj. vzdálenost zásahu od středu terče, je číslo z intervalu $\langle 0, a \rangle$, kde a je poloměr terče. Navíc klasická definice pravděpodobnosti předpokládá, že je intuitivně jasný pojem stejné možnosti výskytu různých jevů.

Nevýhody klasické definice pravděpodobnosti odstraňuje *statistická definice.* Představme si posloupnost velkého počtu n realizací nějakého náhodného pokusu. Pro daný jev A označme n_A počet těch realizací, ve kterých nastal jev A. Podíl n_A/n se nazývá relativní četnost jevu A v n pokusech. Zkušenost ukazuje, že relativní četnosti jevu A v dlouhých sériích realizací náhodného pokusu se odchylují málo od určitého čísla, které nazýváme pravděpodobností tohoto jevu.

Statistická definice pravděpodobnosti však není z hlediska matematiky „korektní" definicí. Proto se dnes nejčastěji definuje pravděpodobnost pomocí *axiomatického přístupu.* Axiomatický přístup předpokládá existenci konečné nebo nekonečné neprázdné množiny, tzv. jevového pole, jehož prvky jsou nazývány elementárními jevy. Dále je určen systém podmnožin jevového pole, jehož prvky se nazývají náhodné jevy. Tento systém musí

a) obsahovat jistý jev Ω i nemožný jev $\emptyset$

a zároveň musí splňovat následující pravidla:

b) S každým jevem A musí obsahovat i jev opačný $\overline{A}$,

c) s každým systémem jevů (konečným nebo nekonečným spočetným) obsahuje i sjednocení a průnik těchto jevů.

Každému náhodnému jevu A je přiřazena pravděpodobnost $P(A)$, což je číslo z intervalu $\langle 0, 1\rangle$

$$P: \quad A \longrightarrow P(A),$$

přičemž

1) $P(\Omega) = 1$,

2) pravděpodobnost sjednocení konečně nebo spočetně mnoha disjunktních jevů je rovna součtu jejich pravděpodobností, tj.

$$P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$$

Pro ilustraci axiomatického přístupu k definici pravděpodobnosti uveďme jako příklad konečné jevové pole. Jevové pole je v tomto případě tvořeno konečným počtem k elementárních jevů $\{A_1, \ldots A_k\}$. Systém náhodných jevů se skládá ze všech elementárních jevů $A_1, \ldots A_k$, ze všech sjednocení libovolného počtu jevů vybraných z jevů $A_1, \ldots A_k$ a nemožného jevu $\emptyset$. Jistý jev $\Omega = A_1 \cup A_2 \cup \ldots A_k$.

Příkladem konečného jevového pole je pole, kde náhodné jevy jsou výsledky jednoho hodu kostkou. Jevové pole je zde tvořeno elementárními jevy A_1 - „padne jednička", A_2 - „padne dvojka", ..., A_6 - „padne šestka". Systém náhodných jevů je tvořen nemožným jevem $\emptyset$, elementárními jevy A_i, $i = 1, \ldots, 6$, všemi možnými sjednoceními libovolného počtu jevů vybraných z jevů $A_1, \ldots A_6$. Například jev C - „padne sudý počet ok" lze vyjádřit jako $C = A_2 \cup A_4 \cup A_6$. Definujme pravděpodobnosti jevů z konečného jevového pole následovně. Pro každý jev A_i, $i = 1, \ldots, k$, platí $P(A_i) = 1/k$. V případě hodu symetrickou kostkou $k = 6$ a $P(A_i) = 1/6$, $i = 1, \ldots, 6$. Pravděpodobnost každého

jiného jevu je součet pravděpodobností těch jevů A_i, jejichž je sjednocením. Například pro jev C - „padne sudý počet ok“ platí

$$P(C) = P(A_2) + P(A_4) + P(A_6) = 3/6 = 1/2.$$

Všimněme si, že konečné jevové pole a pravděpodobnost definovaná shora popsaným způsobem splňují požadavky axiomatické definice. Navíc se takto zavedená pravděpodobnost shoduje s klasickou definicí pravděpodobnosti. Obdobná shoda nastane vždy, pokud má pokus konečný počet vzájemně se vylučujících výsledků, přičemž tyto výsledky jsou stejně pravděpodobné, nebo jinak řečeno, jedná se o konečné jevové pole, přičemž výskyt každého elementárního jevu má stejnou pravděpodobnost.

Poznámka

Všimněme si ještě skutečnosti, že *pravidlo o sčítání pravděpodobností platí pouze pro disjunktní jevy.*

Jestliže jev A označuje „padnutí šestky“ a jev B „padnutí sudého počtu ok“, pak pravděpodobnost jevu $A \cup B$ - „buď padne šestka nebo sudé číslo“ se rovná $P(A \cup B) = 1/2$, zatímco $P(A) + P(B) = 1/6 + 1/2 = 2/3$.

V následujících příkladech předpokládáme, že uvažované hrací kostky jsou symetrické, tj. všechny jejich strany padají se stejnou pravděpodobností.

Příklad 1.

Házíme dvěma kostkami. S jakou pravděpodobností padne alespoň na jedné kostce šestka?

Řešení:

Stejně pravděpodobné výsledky pokusu lze zapsat do následující tabulky:

11	12	13	14	15	16
21	22	23	24	25	26
31	32	33	34	35	36
41	42	43	44	45	46
51	52	53	54	55	56
61	62	63	64	65	66

Tabulka 1.

Označíme-li si jednu kostku jako první a zbývající jako druhou (například si můžeme představit, že první kostka je modrá a druhá červená), znamená prvé číslo ve dvojici

v tabulce 1 výsledek na první kostce a druhé výsledek na druhé kostce. Všech možných výsledků je 36. Výsledků příznivých jevu, že alespoň na jedné kostce, tedy buď na první nebo na druhé, padne šestka, je 11. Hledaná pravděpodobnost je 11/36. □

Příklad 2.

Jaká je pravděpodobnost, že při jednom hodu kostkou nepadne šestka?

Řešení:

Pravděpodobnost jevu $\overline{A}$ opačného k jevu A splňuje vztah $P(\overline{A}) = 1 - P(A)$, neboť jevy A a $\overline{A}$ jsou disjunktní a jejich sjednocení tvoří jistý jev. Odtud $P(\overline{A}) + P(A) = P(A \cup \overline{A}) = P(\Omega) = 1$.

Jev $\overline{B}$ - „na kostce nepadne šestka" je jev opačný k jevu B - „na kostce padne šestka" a tedy $P(\overline{B}) = 1 - P(B) = 1 - 1/6 = 5/6$. Samozřejmě bylo možné spočítat tuto pravděpodobnost také přímo.

□

3. Podmíněná pravděpodobnost

Z osudí, ve kterém jsou dvě bílé a dvě černé koule, taháme koule, aniž je vracíme zpět. S jakou pravděpodobností bude druhá vytažená koule bílá? Pokus má 6 stejně pravděpodobných výsledků {BBČČ, BČBČ, BČČB, ČBČB, ČČBB, ČBBČ}, přičemž příznivé jsou 3 - {BBČČ, ČBČB, ČBBČ}. Hledaná pravděpodobnost je tedy 1/2. Jestliže však jsme v situaci, kdy po prvním tahu víme, že první vytažená koule byla bílá, pak pravděpodobnost, že druhá vytažená koule bude bílá, je 1/3. Informace, že první vytažená koule je bílá, snížila naději na to, že druhá vytažená koule bude bílá.

Pravděpodobnost, že nastane jev A za podmínky, že nastal jev B, se nazývá *podmíněná pravděpodobnost* a značí se $P(A|B)$. Podmíněnou pravděpodobnost lze spočítat ze vztahu

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Příklad 3.

Na výrobě určité součástky se podílejí čtyři výrobci. Označme je pro jednoduchost A, B, C, D. Jakost výrobku je označena jako $I.$, $II.$, $III.$, přičemž $I.$ jakost je nejlepší. Během dlouhé doby bylo zjištěno, jaký podíl z celkového množství součástek na trhu je dané kvality a zároveň je vyroben daným výrobcem, viz tabulka 2. Přáli bychom si zakoupit výrobek $I.$ jakosti. Od jakého výrobce je nejlépe výrobek koupit?

	A	B	C	D
I.	0.05	0.10	0.05	0.02
II.	0.35	0.10	0.05	0.01
III.	0.10	0.10	0.05	0.02

Tabulka 2.

Řešení:

Chceme zjistit, která z podmíněných pravděpodobností $P(I|A)$, $P(I|B)$, $P(I|C)$ resp. $P(I|D)$ je největší, neboť $P(I|A)$ vyjadřuje, jaký podíl tvoří výrobky $I.$ jakosti ze všech výrobků výrobce A, obdobně $P(I|B)$ vyjadřuje, jaký je podíl výrobků $I.$ jakosti mezi výrobky výrobce B atd. Tyto pravděpodobnosti jsou postupně rovny $1/10$, $1/3$, $1/3$, $2/5$. Pokud si přejeme koupit výrobek $I.$ jakosti, pak je nejvýhodnější nakupovat od posledního výrobce D. Mezi jeho výrobky jsou $2/5$, to je 40 %, výrobků $I.$ jakosti. □

4. Věta o úplné pravděpodobnosti a Bayesova věta

Občas se můžeme setkat se situací, kdy nás zajímá pravděpodobnost výskytu nějakého jevu A, jestliže známe pravděpodobnosti výskytu tohoto jevu za různých podmínek, tj. $P(A|B_1), \ldots, P(A|B_k)$, a také pravděpodobnosti, s jakými tyto podmínky nastanou, tj. $P(B_1), \ldots, P(B_k)$. Odpověď na to, jak v takovéto situaci spočítat pravděpodobnost jevu A, dává věta o úplné pravděpodobnosti.

Věta o úplné pravděpodobnosti

Uvažujme systém disjunktních jevů $B_1, \ldots, B_k$ takových, že $\cup_{i=1}^{k} B_i = \Omega$. Pak

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k).$$

Důkaz věty je jednoduchý. Zřejmě lze jev A rozložit následovně

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_k).$$

Jevy $B_1, \ldots, B_k$ jsou vzájemně disjunktní, a proto i jevy $A \cap B_1, \ldots, A \cap B_k$ jsou vzájemně disjunktní. Odtud plyne

$$P(A) = P(A \cap B_1) + \cdots + P(A \cap B_k).$$

Z definice podmíněné pravděpodobnosti pro každé $i = 1, \ldots, k$, plyne $P(A \cap B_i) = P(A|B_i)P(B_i)$ a odtud pak

$$P(A) = P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k).$$

Příklad 4.

Na výrobě určitého výrobku se podílejí tři výrobci, přičemž první výrobce vyrobí 50 %, druhý 30 % a třetí 20 % veškeré produkce. Je známo, že se u některých výrobků vyskytne výrobní vada, kterou je třeba opravit ještě během záruční lhůty. U prvního výrobce tvoří podíl takových výrobků 25 %, u druhého 20 % a u třetího jen 10 %. Jaká část z celkového množství výrobků bude vyžadovat opravu během záruční doby?

Řešení:

Označme

A ... náhodně vybraný výrobek vyžaduje opravu během záruční lhůty,

B_1 ... náhodně vybraný výrobek byl vyroben prvním výrobcem,

B_2 ... náhodně vybraný výrobek byl vyroben druhým výrobcem,

B_3 ... náhodně vybraný výrobek byl vyroben třetím výrobcem.

Zřejmě $P(B_1) = 0.5$, $P(B_2) = 0.3$ a $P(B_3) = 0.2$. Dále známe podmíněné pravděpodobnosti $P(A|B_1) = 0.25$, $P(A|B_2) = 0.2$ a $P(A|B_3) = 0.1$. Všechny výrobky, které budou potřebovat opravu, lze rozdělit na výrobky, které jsou vyrobeny prvním výrobcem $A \cap B_1$, na výrobky, které jsou vyrobeny druhým výrobcem $A \cap B_2$, a na výrobky, které jsou vyrobeny třetím výrobcem $A \cap B_3$. Tyto výrobky budou postupně tvořit 12.5 %, 6 % a 2 % celkové produkce, neboť

$$\begin{aligned} P(A \cap B_1) &= P(A|B_1)\, P(B_1) = 0.25 \cdot 0.5 = 0.125, \\ P(A \cap B_2) &= P(A|B_2)\, P(B_2) = 0.20 \cdot 0.3 = 0.060, \\ P(A \cap B_3) &= P(A|B_3)\, P(B_3) = 0.10 \cdot 0.2 = 0.020, \end{aligned}$$

což dohromady činí 20.5 % celkové produkce, neboť

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) = 0.125 + 0.06 + 0.02 = 0.205.$$

Odpověď: Z celkového množství všech vyrobených výrobků bude 20.5 % potřebovat opravu ještě v záruční době. □

Příklad 5.

Uvažujme situaci stejnou jako v předchozím příkladě. Jaký podíl výrobků, které potřebovaly opravu v záruční době, je tvořen výrobky vyrobenými prvním výrobcem?

Řešení:

Výrobky, které potřebují opravu a současně jsou vyrobeny prvním výrobcem, tvoří 12.5 % všech výrobků. Z celkového množství všech výrobků činí 20.5 % výrobky, které potřebují opravu. Výrobky, které jsou vyrobeny prvním výrobcem, tedy tvoří 60.98 % (= $0.125/0.205 \cdot 100$ %) výrobků, které potřebují opravu v záruční době.

Pravděpodobnost, kterou jsme zde vypočetli, je podmíněná pravděpodobnost jevu B_1 za podmínky, že nastal jev A:

$$P(B_1|A) = \frac{P(A \cap B_1)}{P(A)} = \frac{P(A|B_1)P(B_1)}{P(A)} = \frac{0.125}{0.205} = 0.6098.$$

□

Výrobky vyrobené prvním výrobcem tvoří 50 % celkové výroby, ale mezi výrobky vyžadujícími opravu v záruční době tvoří 60.98 %. V této souvislosti mluvíme o apriorní pravděpodobnosti jevu B_1, tj. $P(B_1) = 0.5$, a o aposteriorní pravděpodobnosti jevu B_1, víme-li, že nastal jev A, tj. $P(B_1|A) = 0.6098$. Bayesova věta podává způsob výpočtu aposteriorní pravděpodobnosti.

Bayesova věta

Uvažujme systém disjunktních jevů $B_1, \ldots, B_k$ takových, že $\cup_{i=1}^{k} B_i = \Omega$. Pak

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + \cdots + P(A|B_k)P(B_k)}, \quad j = 1, \ldots, k.$$

Důkaz:

Podle definice je podmíněná pravděpodobnost jevu B_j za podmínky, že nastal jev A, rovna

$$P(B_j|A) = \frac{P(B_j \cap A)}{P(A)}.$$

Pravděpodobnost průniku $B_j \cap A$ je rovna:

$$P(B_j \cap A) = P(A|B_j)P(B_j)$$

a pravděpodobnost jevu A lze spočítat podle věty o úplné pravděpodobnosti

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_k)P(B_k).$$

5. Nezávislost náhodných jevů

Nezávislost dvou jevů A a B se v praxi projevuje tak, že výskyt jevu A neovlivňuje počet výskytů jevu B a naopak. Tak například, jestliže házíme dvěma kostkami - červenou a modrou, pak jev „padnutí šestky na červené kostce“ a jev „padnutí šestky na modré kostce“ jsou považovány za nezávislé jevy, neboť výskyt šestek na jedné kostce neovlivní výskyt šestek na druhé kostce. Jestliže ale uvažujeme dva náhodné jevy, z nichž první označuje „vznik náledí v určité oblasti“ a druhý „výskyt automobilové nehody v této oblasti“, pak tyto jevy považujeme za jevy závislé, neboť výskyt náledí ovlivňuje počet automobilových nehod.

Matematicky je *nezávislost dvou jevů* A a B definována vztahem

$$P(A \cap B) = P(A) \cdot P(B).$$

Jestliže jev A označuje „padnutí šestky na červené kostce“, jev B označuje „padnutí šestky na modré kostce“ a jev $A \cap B$ označuje jev „na obou kostkách padnou šestky“, pak $P(A) = P(B) = 1/6$ a $P(A \cap B) = 1/36$. Tedy skutečně $P(A \cap B) = P(A) \cdot P(B)$ a jevy A a B jsou nezávislé.

Jestliže jsou jevy A a B nezávislé, pak

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A),$$

což znamená, že pro nezávislé jevy je podmíněná pravděpodobnost rovna nepodmíněné. To odpovídá naší intuitivní představě, že nezávislost jevů A a B spočívá v tom, že skutečnost, že nastal jeden z jevů, neovlivní pravděpodobnost, že nastane jev druhý.

V praxi se často setkáváme s nezávislým opakováním pokusu. Jestliže hodíme dvakrát po sobě za stejných podmínek kostkou, pak „padnutí šestky v prvém hodu“ a „padnutí šestky v druhém hodu“ jsou obecně považovány za nezávislé jevy, neboť výsledek prvního hodu neovlivňuje výsledek v druhém hodu. Obdobně „vytažení určitého čísla ve Sportce v po sobě následujících týdnech“ modelujeme jako nezávislé jevy, neboť je rozumné předpokládat, že „šance“ určitého čísla být vytaženo je v každém tahu stejná, neovlivněná výsledky předchozích tahů. Za takovéhoto předpokladu je úvaha, že větší pravděpodobnost vytažení má číslo, které již dlouho nebylo taženo, mylná.

Příklad 6.

Házíme opakovaně kostkou. Předpokládejme, že výsledky při jednotlivých opakováních hodů jsou nezávislé. S jakou pravděpodobností nám alespoň jednou v pěti po sobě jdoucích hodech padne šestka?

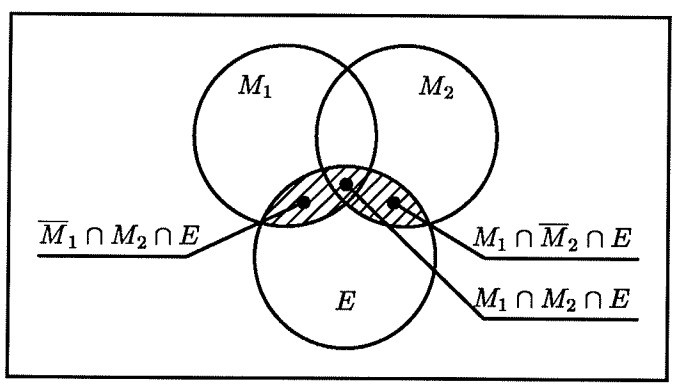
Řešení:

Označme A jev „v pěti po sobě jdoucích hodech padne alespoň jedna šestka“, pak opačný jev $\overline{A}$ značí „v pěti po sobě jdoucích hodech nepadne ani jedna šestka.“ Dále označme A_1 jev „v 1. hodu nepadne šestka“, ..., A_5 jev „v 5. hodu nepadne šestka“. Jevy $A_1, \ldots, A_5$ jsou nezávislé a každý má pravděpodobnost výskytu 5/6. Odtud $P(\overline{A}) = A_1 \cap A_2 \cap \cdots \cap A_5 = (5/6)^5 = 0.402$ a $P(A) = 1 - P(\overline{A}) = 1 - (5/6)^5 = 0.598$. □

Příklad 7.

Na betonárce pracují dvě kontinuální míchačky, každá s pravděpodobností poruchy 0.01. Pravděpodobnost poruchy v dodávce elektrické energie je 0.05. Určete pravděpodobnost, že betonárka bude pracovat alespoň na poloviční výkon.

Řešení:



Obrázek 1.

M_1	... 1. míchačka nemá poruchu	$P(M_1)=0.99$
$\overline{M}_1$	... 1. míchačka má poruchu	$P(\overline{M}_1)=0.01$
M_2	... 2. míchačka nemá poruchu	$P(M_2)=0.99$
$\overline{M}_2$	... 2. míchačka má poruchu	$P(\overline{M}_2)=0.01$
E	... dodávka elektrické energie nemá poruchu	$P(E) =0.95$
$\overline{E}$	... dodávka elektrické energie má poruchu	$P(\overline{E}) =0.05$
A	... betonárka pracuje alespoň na poloviční výkon	$P(A) =?$

Na obrázku 1 vyšrafovaná množina odpovídá jevu A. Z obrázku je patrné, že jev A se dá napsat jako sjednocení tří disjunktních jevů

$$A = (M_1 \cap M_2 \cap E) \cup (M_1 \cap \overline{M}_2 \cap E) \cup (\overline{M}_1 \cap M_2 \cap E),$$

a tedy

$$P(A) = P(M_1 \cap M_2 \cap E) + P(M_1 \cap \overline{M}_2 \cap E) + P(\overline{M}_1 \cap M_2 \cap E).$$

Předpokládáme-li, že poruchy na jednotlivých míchačkách a poruchy v dodávce elektrické energie vznikají nezávisle, platí

$$P(A) = P(M_1)P(M_2)P(E) + P(M_1)P(\overline{M}_2)P(E) + P(\overline{M}_1)P(M_2)P(E) = 0.9499.$$

Betonárka bude pracovat alespoň na poloviční výkon s pravděpodobností 0.9499. □

Část II. Náhodné veličiny a jejich rozdělení

6. Náhodná veličina

Většina náhodných pokusů, se kterými se setkáváme v technických aplikacích, má výsledek vyjádřitelný číslem. Někdy je náhodné kolísání výsledků dáno existencí náhodných chyb měření. Vyměřuje-li například geodet vzdálenost dvou bodů, vždy je měření zatíženo náhodnými chybami způsobenými nepřesností měření i geodeta samého. Náhodné chyby budou samozřejmě tím menší, čím přesnější bude geodet pracovat a čím kvalitnější bude mít přístroje. V jiných případech je náhodnost přímo obsažena v daných pokusech. Zkoumáme-li krychelnou pevnost betonu, budou se výsledky vzájemně lišit podle toho, jak náhodně kolísá kvalita surovin a složení směsi. Naměřená vzdálenost nebo krychelná pevnost betonu jsou příkladem náhodných veličin.

Náhodná veličina je taková veličina, která mění své hodnoty v závislosti na náhodě. Z praktických hledisek se budeme zabývat dvěma typy náhodných veličin - *diskrétními a spojitými.*

Diskrétní náhodná veličina může nabývat konečně nebo spočetně hodnot (to je hodnot, které lze očíslovat 0,1,2,...). V aplikacích se dá často vyjádřit jako počet, např. počet ok při vrhu kostkou, počet let za století, kdy průměrný průtok v řece překročil danou mez, nebo počet autonehod během víkendu na území města Prahy.

Spojitá náhodná veličina může nabývat všech hodnot z určitého intervalu. Příklady takovýchto veličin jsou doba bezporuchového chodu zařízení, okamžitý průtok v řece, chyba v měření vzdálenosti dvou bodů apod.

Náhodné veličiny budeme značit velkými písmeny X, Y,...

7. Diskrétní náhodná veličina

Pravděpodobnostní chování náhodné veličiny X je dáno *rozdělením pravděpodobnosti náhodné veličiny* X. V případě diskrétní náhodné veličiny je rozdělení dáno

a) výčtem hodnot I, kterých veličina může nabývat,

b) pravděpodobnostmi, s jakými jednotlivé hodnoty $x \in I$ nabývá, tj. $P(X = x)$, $x \in I$.

Uveďme několik příkladů rozdělení diskrétních náhodných veličin.

Příklad 8.

Uvažujme náhodnou veličinu X, která označuje počet líců ve třech po sobě jdoucích nezávislých hodech mincí. Jaké je její rozdělení?

Řešení:

Náhodná veličina X může nabývat hodnot $\{0, 1, 2, 3\}$. Hodnotu 0 nabude, jestliže ve všech třech případech padne rub, tedy RRR. V prvním hodu padne rub s pravděpodobností 1/2, v druhém 1/2 a v třetím opět 1/2. Hody jsou nezávislé, a proto jev RRR má pravděpodobnost 1/8, a tedy $P(X = 0) = 1/8$. Náhodná veličina X nabude hodnotu 1, jestliže nastane buď jev LRR nebo jev RLR nebo RRL. Každý z těchto jevů má pravděpodobnost 1/8, a proto $P(X = 1) = 1/8 + 1/8 + 1/8 = 3/8$. Obdobně zjistíme, že $P(X = 2) = 3/8$ a $P(X = 3) = 1/8$. Tabulka 3 udává přehledně rozdělení veličiny X.

x	0	1	2	3
P(X=x)	1/8	3/8	3/8	1/8

Tabulka 3.

□

Poznámka:

Všimněme si, že součet pravděpodobností v druhém řádku tabulky 3 je roven jedné. Tento fakt musí být splněn vždy, neboť pro různé hodnoty i a $j \in I$ jsou jevy $(X = i)$ a $(X = j)$ disjunktní a sjednocení všech jevů $\cup_{i \in I}(X = i)$ tvoří jistý jev Ω. Platí vždy $\sum_{x \in I} P(X = i) = 1$.

Příklad 9.

Uvažujme terč, který je tvořen středem a mezikružím. Zásah do středu je hodnocen 10 body, zásah do mezikruží 5 body. Určitý střelec střílí tak, že zasáhne střed s pravděpodobností 0.7, mezikruží s pravděpodobností 0.2 a netrefí se vůbec do terče s pravděpodobností 0.1. Náhodná veličina X je součet bodů získaných tímto střelcem po dvou výstřelech. Určete její rozdělení.

Řešení:

Označme

S - zásah do středu,

M - zásah do mezikruží,

V - zásah mimo terč.

Vypišme všechny možné výsledky střelby, tj. kolik dávají bodů a s jakou pravděpodobností se vyskytují:

výsledek	VV	VM	MV	VS	SV	MM	MS	SM	SS
počet bodů	0	5	5	10	10	10	15	15	20
pravděpodobnost	0.01	0.02	0.02	0.07	0.07	0.04	0.14	0.14	0.49

Tabulka 4.

Rozdělení náhodné veličiny X je tedy dáno následující tabulkou:

x	0	5	10	15	20
$P(X = x)$	0.01	0.04	0.18	0.28	0.49

Tabulka 5.

□

8. Charakteristiky diskrétní náhodné veličiny

Polohu hodnot náhodné veličiny charakterizuje nejlépe *střední hodnota* (očekávaná střední hodnota) náhodné veličiny X, která se obvykle značí $\mathrm{E}\,X$. V případě diskrétní náhodné veličiny X je definována vztahem:

$$\mathrm{E}\,X = \sum_{x \in I} x\,P(X = x),$$

kde I je množina hodnot, jichž náhodná veličina X nabývá.

Příklad 10.

Spočtěte střední hodnotu náhodné veličiny X, která označuje počet líců ve třech po sobě jdoucích nezávislých hodech mincí (viz příklad 8).

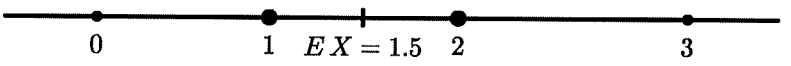
Řešení:

$$\mathrm{E}\,X = 0 \cdot (1/8) + 1 \cdot (3/8) + 2 \cdot (3/8) + 3 \cdot (1/8) = 12/8 = 1.5.$$

□

Názornou představu o tom, kde se střední hodnota na číselné ose nachází, dává následující postup. Představme si hodnoty náhodné veličiny X jako hmotné body umístěné na číselné ose. Jejich umístění vzhledem k počátku odpovídá daným hodnotám náhodné veličiny a jejich hmotnost příslušným pravděpodobnostem. Střední hodnota je pak shodná

s těžištěm takovéto hmotné soustavy. Pro příklad 10 je ze symetrie ihned patrné, že $\mathrm{E}\,X = 1.5$, viz obrázek 2.



Obrázek 2.

Uvažujme náhodný pokus, jehož výsledkem je realizace náhodné veličiny X. Jestliže pokus za stejných podmínek n krát nezávisle zopakujeme, získáme vektor $(X_1, \ldots, X_n)$. Pokus může například spočívat ve třech hodech mincí, přičemž náhodná veličina označuje počet výskytů líců. Zopakujeme-li n krát pokus, pak X_1 značí počet výskytů líců v první trojici hodů, X_2 počet líců v druhé trojici a tak dále. Za velmi obecných podmínek platí, že se aritmetický průměr $\overline{X} = (\sum X_i)/n$ bude pro velká n blížit střední hodnotě $\mathrm{E}\,X$. Můžeme si tedy intuitivně představovat střední hodnotu jako jakýsi dlouhodobý průměr.

Shora uvedená úvaha se často používá při hazardních hrách, kde uvažovanou náhodnou veličinou je zisk při jedné hře. Jestliže pro velká n platí, že $(\sum X_i)/n \approx \mathrm{E}\,X$, pak celkový zisk v n hrách se chová přibližně jako $n \cdot \mathrm{E}\,X$. Záleží tedy na střední hodnotě, zda je hra spravedlivá ($\mathrm{E}\,X = 0$) nebo pro hráče výhodná ($\mathrm{E}\,X > 0$) nebo nevýhodná ($\mathrm{E}\,X < 0$).

Příklad 11.

Hráč hází kostkou. Padne-li šestka, získává 6 Kč, padne-li jiné číslo, ztrácí 1 Kč. Je hra pro hráče výhodná nebo nevýhodná?

Řešení:

Náhodná veličina X označující zisk hráče během jedné hry má rozdělení dané tabulkou:

x	-1	6
$P(X = x)$	5/6	1/6

Tabulka 6.

Spočtěme střední hodnotu: $\mathrm{E}\,X = (-1) \cdot (5/6) + 6 \cdot (1/6) = 1/6 > 0$. Hra je tedy pro hráče výhodná. Je možné trochu nepřesně říci, že v průměru získá hráč na každé hře 1/6 Kč. □

Rozptýlenost (kolísavost, variabilitu) hodnot náhodné veličiny vyjadřuje nejlépe *rozptyl*, který se značí $\mathrm{Var}\,X$. V případě diskrétní náhodné veličiny je definován vztahem:

$$\mathrm{Var}\,X = \sum_{x \in I} (x - \mathrm{E}\,X)^2 \cdot P(X = x) = \left(\sum_{x \in I} x^2 \cdot P(X = x) \right) - (\mathrm{E}\,X)^2.$$

Poznamenejme, že pro výpočet je vhodnější druhý výraz. Často jako charakteristiku rozptýlenosti uvažujeme druhou odmocninu z rozptylu, které se říká *směrodatná odchylka* a budeme ji značit $\operatorname{sd} X$.

Příklad 12.

Spočtěte rozptyl náhodné veličiny X označující počet líců ve třech po sobě jdoucích hodech symetrickou mincí (viz příklad 8).

Řešení:

$$\operatorname{Var} X = 0^2 \cdot (1/8) + 1^2 \cdot (3/8) + 2^2 \cdot (3/8) + 3^2 \cdot (1/8) - 1.5^2 = 3/4 = 0.75.$$

□

Uvažujeme dvě hry. V první hře házíme mincí a získáváme 1 Kč, jestliže padne líc, a ztrácíme 1 Kč, jestliže padne rub. V druhé hře rovněž házíme mincí a získáváme 10 Kč, jestliže padne líc, a ztrácíme 10 Kč, jestliže padne rub. Označme zisk v první hře X_1 a v druhé X_2. Rozdělení náhodných veličin X_1 a X_2 je následující:

x	-1	1
$P(X_1 = x)$	1/2	1/2

Tabulka 7

x	-10	10
$P(X_2 = x)$	1/2	1/2

Tabulka 8

Obě hry jsou zřejmě spravedlivé, neboť $\operatorname{E} X_1 = 0$ a $\operatorname{E} X_2 = 0$. Sledujme však celkový zisk během n her, tj. $\sum X_{1i}$, resp. $\sum X_{2i}$, kde X_{1i} označuje zisk v i-tém opakování první hry a X_{2i} označuje zisk v i-tém opakování druhé hry. V případě druhé hry bude celkový zisk kolísat kolem nuly daleko s většími rozkmity než u hry první. To je způsobeno tím, že rozptyl celkového zisku v n hrách se v tomto případě rovná součtu rozptylů zisku v jednotlivých hrách a ten je daleko větší v druhé hře než v první, neboť

$$\operatorname{Var} X_1 = (-1)^2 \cdot (1/2) + 1^2 \cdot (1/2) = 1,$$

$$\operatorname{Var} X_2 = (-10)^2 \cdot (1/2) + (10)^2 \cdot (1/2) = 100.$$

Příklad 13.

Při zkoušení přístrojů dvou typů A a B byla určena pravděpodobnost výskytu poruch za určité období (viz tabulka 9), jejichž oprava stojí postupně 100, 200 a 300 Kč. Nákupčí má pro závod nakoupit velké množství těchto přístrojů. Pro který typ by se měl rozhodnout?

Cena za opravu		100 Kč	200 Kč	300 Kč
Pravděpodobnost	Typ A	0.20	0.06	0.04
výskytu potuch	Typ B	0.06	0.04	0.10

Tabulka 9.

Řešení:

Za lepší přístroj bychom měli považovat ten typ, jehož očekávaná střední hodnota výdajů za opravy je menší, neboť při velkém počtu přístrojů se průměrný výdaj na opravu jednotlivého přístroje bude pohybovat kolem této hodnoty. Spočtěme střední hodnoty výdajů za opravy u obou přístrojů:

$$\mathrm{E}\,X_A = 100 \cdot 0.20 + 200 \cdot 0.06 + 300 \cdot 0.04 = 44 \quad (\text{Kč}),$$

$$\mathrm{E}\,X_B = 100 \cdot 0.06 + 200 \cdot 0.04 + 300 \cdot 0.10 = 44 \quad (\text{Kč}).$$

Střední hodnota je u obou typů stejná. Podle čeho se máme tedy rozhodnout dále? Dalším kritériem by mohl být rozptyl, neboť výdaje za opravy přístroje s menším rozptylem budou kolísat kolem hodnoty 44 Kč daleko méně než u přístroje s větším rozptylem. Nemáme-li tedy hazardní povahu a snažíme-li se například mít v pokladně vždy dost peněz na případné výdaje za opravy, je lépe se rozhodnout pro přístroj s menším rozptylem. Spočtěme rozptyly a směrodatné odchylky výdajů za opravy u obou přístrojů:

$$\mathrm{Var}\,X_A = 100^2 \cdot 0.20 + 200^2 \cdot 0.06 + 300^2 \cdot 0.04 - 44^2 = 6064,$$

$$\mathrm{Var}\,X_B = 100^2 \cdot 0.06 + 200^2 \cdot 0.04 + 300^2 \cdot 0.10 - 44^2 = 9264,$$

$$\mathrm{sd}\,X_A = 77.9 \ (\text{Kč}),$$

$$\mathrm{sd}\,X_B = 96.2 \ (\text{Kč}).$$

Přijmeme-li tedy pro výhodnost nákupu kritérium menšího rozptylu, je přístroj typu A lepší než přístroj typu B. □

Nakonec ještě uveďme pravidla pro výpočet střední hodnoty a rozptylu transformované veličiny $Y = \alpha X + \beta$:

$$\mathrm{E}\,(\alpha X + \beta) = \alpha\,\mathrm{E}\,X + \beta,$$

$$\mathrm{Var}\,(\alpha X + \beta) = \alpha^2\,\mathrm{Var}\,X,$$

$$\mathrm{sd}\,(\alpha X + \beta) = |\alpha|\,\mathrm{sd}\,X.$$

9. Některé typy diskrétně rozdělených náhodných veličin

V praxi se opakovaně setkáváme s některými rozděleními diskrétních veličin. Tato rozdělení jsou určitého typu, přičemž dvě rozdělení stejného typu se liší jen hodnotou parametrů, které určují konkrétní rozdělení náhodné veličiny.

Binomické rozdělení Bi (n,p)

Náhodná veličina X má *binomické rozdělení* s parametry n a p, jestliže nabývá hodnot $x = 0, 1, \ldots, n$ s pravděpodobnostmi

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Parametr n je přirozené číslo a $p \in (0, 1)$.

Binomickým rozdělením $Bi(n, p)$ se řídí náhodná veličina, která označuje počet výskytů jevu A v n nezávislých pokusech, jestliže pravděpodobnost výskytu tohoto jevu A je v každém jednotlivém pokusu rovna témuž číslu p. Někdy se situace také popisuje tak, že určitý pokus má pouze dva možné výsledky - úspěch a neúspěch, přičemž pravděpodobnost úspěchu je rovna p a pravděpodobnost neúspěchu je $1 - p$. Počet úspěchů v sérii n nezávislých pokusů se řídí binomickým rozdělením $Bi(n, p)$.

Poznamenejme, že

$$\mathrm{E}\, X = n \cdot p \qquad \text{a} \qquad \mathrm{Var}\, X = n \cdot p \cdot (1-p).$$

Binomickému rozdělení s parametrem $n = 1$ říkáme *alternativní* $A(p)$. Rozdělení náhodné veličiny X řídící se alternativním rozdělením $A(p)$ je dáno tabulkou 10:

x	0	1
P(X=x)	1-p	p

Tabulka 10.

Zřejmě

$$\mathrm{E}\, X = p \qquad \text{a} \qquad \mathrm{Var}\, X = p \cdot (1-p).$$

Příklad 14.

Házíme třikrát po sobě symetrickou mincí. Určete rozdělení náhodné veličiny X označující počet líců v těchto hodech (viz příklad 8).

Řešení:

Pokusem je zde hod mincí. Za úspěch se považuje padnutí líce. Pravděpodobnost padnutí líce na minci, která je z hlediska pravděpodobnosti symetrická, je $p = 1/2$. Tato

pravděpodobnost je v každém hodu stejná. Pokus opakujeme třikrát a tedy $n = 3$. Náhodná veličina X označující počet líců ve třech hodech se řídí binomickým rozdělením $Bi(n = 3, p = 1/2)$. Dosadíme-li do vzorce pro výpočet pravděpodobností binomického rozdělení, dostaneme:

$$P(X = 0) = \binom{3}{0}(1/2)^0(1/2)^3 = 1 \cdot 1 \cdot (1/8) = 1/8,$$

$$P(X = 1) = \binom{3}{1}(1/2)^1(1/2)^2 = 3 \cdot (1/2) \cdot (1/4) = 3/8,$$

$$P(X = 2) = \binom{3}{2}(1/2)^2(1/2)^1 = 3 \cdot (1/4) \cdot (1/2) = 3/8,$$

$$P(X = 3) = \binom{3}{3}(1/2)^3(1/2)^0 = 1 \cdot (1/8) \cdot 1 = 1/8.$$

Všimněme si, že výsledky se shodují s příkladem 8. □

Příklad 15.

Student složí zkoušku, jestliže v testu odpoví správně alespoň na čtyři z pěti otázek. U každé otázky jsou čtyři možné odpovědi, z nichž jediná je správná. S jakou pravděpodobností student složí zkoušku, jestliže se vůbec nepřipravoval a odpovědi volí náhodně?

Řešení:

Pokusem je zde odpověď na otázku. Pokus končí úspěchem, jestliže byla otázka zodpovězena správně. Student vybírá náhodně ze čtyř možných odpovědí, a proto $p = 1/4$. Počet pokusů je $n = 5$. Náhodná veličina X označující počet správných odpovědí má tedy binomické rozdělení $Bi(n = 5, p = 1/4)$. Student složí zkoušku, jestliže správně zodpoví alespoň 4 otázky, tj. 4 nebo 5. Odtud

$$P(X \geq 4) = P(X = 4) + P(X = 5) = \binom{5}{4}\left(\frac{1}{4}\right)^4\frac{3}{4} + \binom{5}{5}\left(\frac{1}{4}\right)^5$$

$$= 5\frac{3}{1024} + \frac{1}{1024} = \frac{16}{1024} = 0.0156.$$

Student složí zkoušku bez přípravy s pravděpodobností 0.0156.

□

Hypergeometrické rozdělení

Náhodná veličina X má hypergeometrické rozdělení s parametry N, A, n, jestliže

$$P(X=x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}} \qquad \text{pro} \quad x = \max(0, A-N+n), \ldots, \min(A,n).$$
$$= \quad 0 \qquad\qquad\qquad\qquad\qquad\qquad \text{jinak.}$$

Parametry N, A, n jsou přirozená čísla splňující $1 \leq n \leq N$, $1 \leq A \leq N$.

Nejčastější interpretaci parametrů uvidíme z následujícího modelu. Mějme soubor N jednotek, z nichž A jednotek má sledovanou vlastnost. Z tohoto souboru vybereme náhodně najednou nebo postupně bez vracení n jednotek. Náhodná veličina X označující počet vybraných jednotek vykazujících sledovanou vlastnost se řídí hypergeometrickým rozdělením. Nechť je například v osudí N koulí, přičemž A koulí je bílých a $N-A$ černých. Z promíchaného osudí vytáhneme bez vracení n koulí. Počet vytažených bílých koulí se bude řídit hypergeometrickým rozdělením s parametry N, A, n.

Střední hodnota náhodné veličiny řídící se hypergeometrickým rozdělením

$$\mathrm{E}\, X = \frac{n \cdot A}{N}$$

a rozptyl

$$\mathrm{Var}\, X = n \frac{A}{N}\left(1 - \frac{A}{N}\right)\left(\frac{N-n}{N-1}\right).$$

Příklad 16.

Je známo, že mezi 10 součástkami jsou 3 vadné. Sestavíme-li přístroj ze 3 náhodně vybraných součástek, jaká je pravděpodobnost, že bude fungovat?

Řešení:

Náhodná veličina X označující počet vadných součástek v přístroji se řídí hypergeometrickým rozdělením, přičemž $N = 10$ (počet všech součástek), $A = 3$ (počet vadných součástek) a $n = 3$ (počet součástek použitých při výrobě přístroje). Přístroj bude fungovat, nebude-li obsahovat žádnou vadnou součástku. Spočtěme pravděpodobnost, s jakou se náhodná veličina X označující počet vadných součástek v přístroji rovná 0, tj.

$$P(X=0) = \frac{\binom{3}{0}\binom{7}{3}}{\binom{10}{3}} = 0.292.$$

Přístroj bude fungovat s pravděpodobností 0.292.

□

Příklad 17.

V určitém týdnu jsme vsadili ve Sportce 1 sloupec. S jakou pravděpodobností vyhrajeme v jednom tahu 5. pořadí, tj. uhádneme 3 čísla ze 6 tažených?

Řešení:

Náhodná veličina X označující počet uhádnutých čísel se řídí hypergeometrickým rozdělením, přičemž $N = 49$ (počet všech čísel ve sloupci), $A = 6$ (počet čísel, které jsme zaškrtli ve sloupci) a $n = 6$ (počet tažených čísel), a tedy

$$P(X = 3) = \frac{\binom{6}{3}\binom{43}{3}}{\binom{49}{6}} = 0.01765.$$

Jestliže vsadíme pouze jeden sloupec, vyhrajeme v jednom tahu 5. pořadí s pravděpodobností 0.01765.

□

Příklad 18.

Mějme osudí, v kterém je A bílých a $(N - A)$ černých koulí. Pokus spočívá ve vytažení n koulí. Náhodná veličina X označuje počet bílých vytažených koulí, jestliže po každém vytažení kouli vracíme zpět. Náhodná veličina Y označuje počet bílých vytažených koulí, jestliže vytažené koule již zpět nevracíme. Určete rozdělení X a Y.

Řešení:

Náhodná veličina X má binomické rozdělení s parametry n a $p = A/N$, neboť pravděpodobnost vytažení bílé koule je při každém tahu stejná, nezávisí na výsledcích předchozích tahů a rovná se A/N. Naopak náhodná veličina Y má hypergeometrické rozdělení s parametry N, A, n, neboť pravděpodobnost vytažení bílé koule se s přibývajícími tahy mění a závisí na tom, jaké koule jsme vytáhli v předchozích tazích.

□

Poissonovo rozdělení Po (λ)

Náhodná veličina X má *Poissonovo rozdělení* s parametrem $\lambda > 0$, jestliže

$$\begin{aligned} P(X = x) &= \frac{e^{-\lambda}\lambda^x}{x!} && \text{pro} \quad x = 0, 1, \ldots \\ &= 0 && \text{jinak.} \end{aligned}$$

Parametru λ se říká intenzita, neboť platí $\mathrm{E}\, X = \lambda$. Navíc platí rovněž $\mathrm{Var}\, X = \lambda$.

Poissonovým rozdělením se řídí náhodné veličiny, označující počet částeček v určitém objemu dobře zamíchané směsi či počet událostí v nějakém časovém intervalu. Typickým příkladem je počet radioaktivních pulsů zaznamenaných Geigrovým přístrojem během intervalu $(0, t)$. Předpokládejme, že pravděpodobnost výskytu pulsů v čase se řídí následujícími pravidly:

1. Existuje parametr $\alpha > 0$ takový, že v krátkém časovém intervalu Δt je pravděpodobnost jediného zaznamenaného pulsu rovna $\alpha \Delta t + o(\Delta t)$.

2. Pravděpodobnost, že v časovém intervalu o délce Δt zaznamenáme více než jeden impuls je $o(\Delta t)$, a tedy pravděpodobnost, že se nezaznamená vůbec žádný puls je rovna $1 - \alpha \Delta t + o(\Delta t)$.

3. Počet pulsů zaznamenaných během časového intervalu o délce Δt je nezávislý na počtu pulsů, který byl zaznamenán před tímto intervalem.

Jsou-li splněny shora uvedené předpoklady, pak počet pulsů X zaznamenaných během časového intervalu $(0, t)$ má Poissonovo rozdělení s parametrem $\lambda = \alpha t$, tj.

$$\begin{aligned} P(X = x) &= \frac{e^{-\alpha t}(\alpha t)^x}{x!} && \text{pro} \quad x = 0, 1, \ldots \\ &= \quad 0 && \text{jinak.} \end{aligned}$$

Očekávaný počet pulsů $\mathrm{E}\, X$ během časového intervalu $(0, t)$ je roven αt, takže očekávaný počet pulsů v intervalu $(0, 1)$ je roven α.

Jestliže v předpokladech 1.–3. nahradíme slovo puls slovem událost, dostaneme obecnou definici *Poissonova procesu.* Poissonovým procesem se modeluje např. počet přicházejících či obsluhujících zákazníků, počet vozidel, které projedou za určitý čas křižovatkou atd.

Příklad 19.

K holiči chodí „v průměru" čtyři zákazníci za hodinu. S jakou pravděpodobností přijde během půl hodiny alespoň jeden zákazník?

Řešení:

Předpokládáme-li, že zákazníci chodí k holiči zcela náhodně, pak náhodná veličina X označující počet zákazníků, kteří tam přijdou během půl hodiny, se bude řídit Poissonovým rozdělením. „Průměrně" přijdou k holiči dva zákazníci za půl hodiny a tedy $\lambda = 2$. K holiči přijde alespoň jeden zákazník, přijde-li jich tam jeden nebo více, a tedy

$$P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-2} \doteq 1 - 0.135 = 0.865.$$

Během půl hodiny přijde k holiči alespoň jeden zákazník s pravděpodobností 0.865.

□

Poznámka

Poissonova rozdělení se často používá pro aproximaci binomického rozdělení, jestliže pro parametry binomického rozdělení platí, že n je velké a p je malé. V tomto případě se hodnota parametru λ aproximujícího Poissonova rozdělení vypočte ze vztahu $\lambda = np$.

Příklad 20.

Na telefonní ústřednu je napojeno 300 účastníků. Každý z nich bude volat ústřednu během hodiny s pravděpodobností 0.01. Jaká je pravděpodobnost toho, že během hodiny zavolají ústřednu 4 účastníci?

Řešení:

Náhodná veličina označující počet účastníků volající ústřednu během hodiny má binomické rozdělení s parametry $n = 300$ a $p = 0.01$. Binomické rozdělení je možno v našem případě aproximovat Poissonovým rozdělením s parametrem $\lambda = 3$. Znamená to, že pravděpodobnost $P(X = 4) = \binom{300}{4} 0.01^4 0.99^{296}$ je možno aproximovat pravděpodobností $P(X = 4) = e^{-3}\, 3^4/4! \doteq 0.168$. Pravděpodobnost toho, že během hodiny zavolají ústřednu 4 účastníci, je přibližně rovna 0.168. □

10. Spojitá náhodná veličina

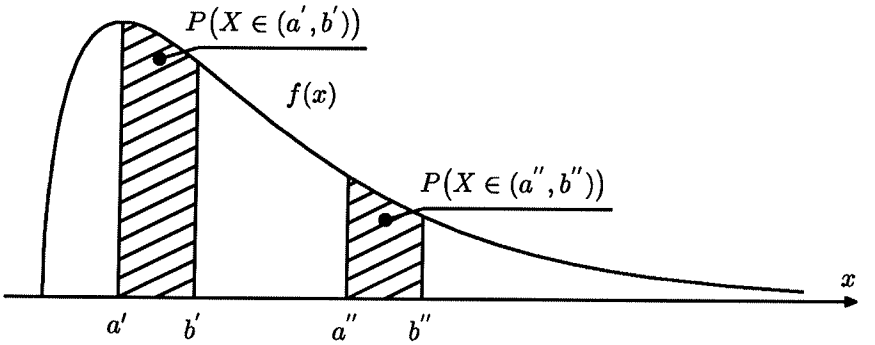
V případě spojité náhodné veličiny nás obvykle zajímá, s jakou pravděpodobností se náhodná veličina realizuje uvnitř nějakého konečného nebo nekonečného intervalu (a, b). Zajímat nás může například pravděpodobnost toho, že průtok v řece Váh překročí 200 m^3/s,že se výška náhodně vybraného člověka se pohybuje mezi 170 až 190 cm, že tlak, při kterém praská betonová kostka, bude menší než 20 MPa. Tyto pravděpodobnosti $P(X \in (a, b))$ se spočítají jako integrály z nezáporné funkce $f(x)$, tj. $\int_a^b f(x)\, dx$. Jinak řečeno náhodná veličina X má rozdělení spojitého typu, existuje-li nezáporná reálná funkce $f(x)$ taková, že pro libovolný interval (a, b) platí

$$P\big(X \in (a, b)\big) = \int_a^b f(x)\, dx.$$

Funkci $f(x)$ se říká *hustota pravděpodobnosti* a musí pro ni platit

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Podívejme se na obrázek 3. Předpokládejme, že známe tvar funkce $f(x)$. Pravděpodobnosti $P\big(X \in (a', b')\big)$ a $P\big(X \in (a'', b'')\big)$ odpovídají velikosti vyšrafovaných ploch.



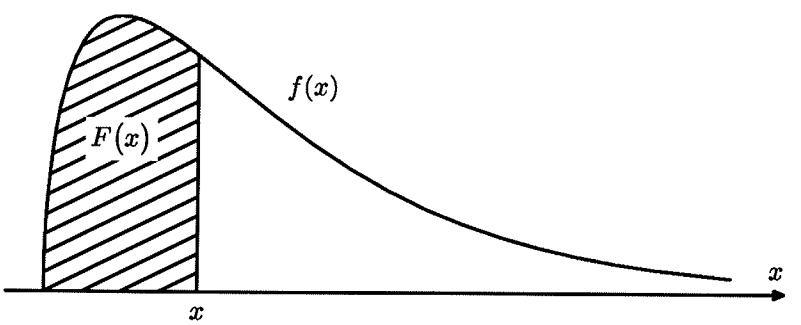
Obrázek 3.

Všimněme si, že intervaly (a', b') a (a'', b'') jsou stejně dlouhé, avšak $P\big(X \in (a', b')\big)$ je mnohem větší než $P\big(X \in (a'', b'')\big)$. Lze tedy předpokládat, že náhodná veličina X se bude mnohem častěji realizovat tam, kde je hustota $f(x)$ velká, než tam, kde je malá.

Kromě hustoty $f(x)$ se v teorii pravděpodobnosti velmi často pracuje s funkcí

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = P\big(X \in (-\infty, x)\big),$$

které se říká *distribuční funkce*. Názorně si můžeme představit distribuční funkci v bodě x jako plochu pod hustotou do bodu x, viz obrázek 4.



Obrázek 4.

Platí $\lim_{x\to-\infty} F(x) = 0$ a $\lim_{x\to\infty} F(x) = 1$. Hustota je zřejmě derivace distribuční funkce.

Příklad 21.

Autobusy přijíždějí na zastávku přesně v pětiminutových intervalech. Jistý pán přijde náhodně na zastávku a zjistí nepříjemnou skutečnost. Nepřijede-li mu autobus do dvou

minut, přijde pozdě do zaměstnání. S jakou pravděpodobností přijde pozdě do zaměstnání?

Řešení:

Náhodná veličina X zde označuje dobu čekání na autobus. Tato doba nemůže být záporná ani větší než 5, odtud $f(x) = 0$ pro $x < 0$ a $x > 5$. Pán přišel na zastávku náhodně, a proto například pravděpodobnost, že doba čekání na autobus se bude pohybovat mezi 2–3 minutami je stejná jako pravděpodobnost, že se doba čekání bude pohybovat mezi 3–4 minutami. Obdobně tato pravděpodobnost bude stejná pro libovolné stejně dlouhé časové intervaly v rozmezí 0 až 5 minut. To ovšem znamená, že hustota bude na intervalu $\langle 0, 5\rangle$ konstantní. Navíc musí platit $\int_{-\infty}^{\infty} f(x)\,dx = \int_0^5 f(x)\,dx = 1$, a proto $f(x) = 1/5$ pro $x \in \langle 0, 5\rangle$. Doba čekání na autobus má tedy hustotu

$$\begin{aligned} f(x) &= 1/5 && \text{pro} \quad x \in \langle 0, 5\rangle, \\ &= 0 && \text{pro} \quad x \notin \langle 0, 5\rangle. \end{aligned}$$

Pán přijde pozdě, jestliže $X > 2$, a tedy $P(X > 2) = \int_2^5 (1/5)\,dx = 3/5$. Pán přijde pozdě do zaměstnání s pravděpodobností $3/5 = 0.6$.

□

Rozdělení spojité náhodné veličiny udává hustota. Praktická znalost toho, jakým rozdělením se určitá náhodná veličina řídí, je obvykle dána dlouhodobou zkušeností. Někdy se typ rozdělení dá odvodit z teoretických úvah.

11. Charakteristiky spojité náhodné veličiny

Nejvýznamnější charakteristikou je opět *střední hodnota*, která udává polohu hodnot náhodné veličiny. V případě spojité náhodné veličiny X je definována vztahem:

$$\mathrm{E}\,X = \int_{-\infty}^{\infty} x\, f(x)\,dx\,.$$

Příklad 22.

Doba X do vybití baterie určovaná v rocích se řídí rozdělením s hustotou

$$\begin{aligned} f(x) &= e^{-x} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0. \end{aligned}$$

Spočtěte její střední hodnotu.

Řešení:

$$\mathrm{E}\,X = \int_{-\infty}^{\infty} x\,f(x)\,dx = \int_{0}^{\infty} x\,e^{-x}\,dx = [-x\,e^{-x}]_0^{\infty} + \int_{0}^{\infty} e^{-x}\,dx = 1.$$

Střední doba do vybití baterie je 1 rok.

□

Názornou představu o tom, kde se střední hodnota na číselné ose nachází, dává postup analogický jako v diskrétním případě. Představme si, že je hmota rozmístěna na přímce podle hustoty $f(x)$, pak poloha střední hodnoty na přímce se bude shodovat s těžištěm takovéto přímky. Z této představy ihned vyplývá, že u symetricky rozdělených náhodných veličin se střední hodnota shoduje se středem symetrie. Tvrzení o konvergenci aritmetického průměru ke střední hodnotě platí i v případě spojité náhodné veličiny, a tudíž zůstává zachována i intuitivní představa střední hodnoty jako dlouhodobého průměru.

Rozptýlení (kolísavost, variabilitu) náhodné veličiny opět vyjadřuje *rozptyl* $\mathrm{Var}\,X$, který je v případě spojité náhodné veličiny definován následovně:

$$\mathrm{Var}\,X = \int_{-\infty}^{\infty} (x - \mathrm{E}\,X)^2\,f(x)\,dx = \Big(\int_{-\infty}^{\infty} x^2\,f(x)\,dx\Big) - \big(\mathrm{E}\,X\big)^2.$$

Odmocnině z rozptylu se říká *směrodatná odchylka.*

Příklad 23.

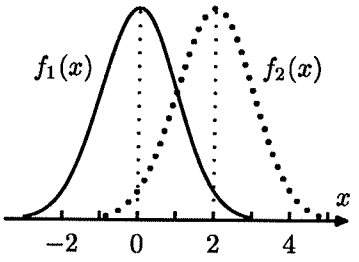
Spočtěte rozptyl náhodné veličiny udávající dobu do vybití baterie, přičemž hustota této veličiny je stejná jako v příkladu 22.

Řešení:

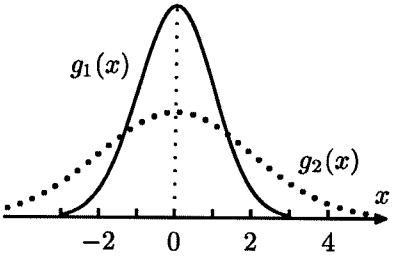
$$\mathrm{Var}\,X = \Big(\int_{-\infty}^{\infty} x^2\,f(x)\,dx\Big) - \big(\mathrm{E}\,X\big)^2 = \int_{0}^{\infty} x^2\,e^{-x}\,dx - 1 = 1.$$

□

Pro lepší představu souvislosti mezi hustotou a jejími charakteristikami střední hodnotou a rozptylem si prohlédněte následující obrázky. Na obrázku 5 vidíme hustotu $f_1(x)$ náhodné veličiny X_1 a hustotu $f_2(x)$ náhodné veličiny X_2, přičemž $\mathrm{E}\,X_1 < \mathrm{E}\,X_2$ a $\mathrm{Var}\,X_1 = \mathrm{Var}\,X_2$. Na obrázku 6 vidíme hustotu $g_1(x)$ náhodné veličiny Y_1 a hustotu $g_2(x)$ náhodné veličiny Y_2, přičemž $\mathrm{E}\,Y_1 = \mathrm{E}\,Y_2$ a $\mathrm{Var}\,Y_1 < \mathrm{Var}\,Y_2$.



Obrázek 5.



Obrázek 6.

Příklad 24.

Na trhu jsou dva typy součástek za stejnou cenu. Doba životnosti X_1 prvního typu součástek má hustotu

$$\begin{aligned} f_1(x) &= e^{-x} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0, \end{aligned}$$

doba životnosti X_2 druhého typu součástek má hustotu

$$\begin{aligned} f_2(x) &= 2\,e^{-2x} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0. \end{aligned}$$

Doby životnosti jsou udány v rocích. Pro nákup jakého typu součástek se rozhodneme?

Řešení:

Za rozhodující kritérium můžeme zvolit například střední hodnotu, neboť u součástek s delší dobou životnosti můžeme očekávat, že nám „v průměru" déle vydrží. Vzhledem k tomu, že

$$\begin{aligned} \mathrm{E}\,X_1 &= \int_0^\infty x\,e^{-x}\,dx = 1, \\ \mathrm{E}\,X_2 &= \int_0^\infty 2x\,e^{-2x}\,dx = 1/2, \end{aligned}$$

rozhodneme se pro nákup součástky prvního typu.

□

Příklad 25.

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

$$\begin{aligned} f_1(x) &= (1/2)\,e^{x} && \text{pro} \quad x \leq 0, \\ &= (1/2)\,e^{-x} && \text{pro} \quad x > 0, \end{aligned}$$

a odchylka X_2 v dávkách 2. dávkovače má rozdělení s hustotou

$$\begin{aligned} f_1(x) &= e^{2x} && \text{pro} \quad x \leq 0, \\ &= e^{-2x} && \text{pro} \quad x > 0. \end{aligned}$$

Který dávkovač je lepší?

Řešení:

Hustoty obou veličin X_1 a X_2 jsou symetrické kolem 0, a proto $\mathrm{E}\, X_1 = \mathrm{E}\, X_2 = 0$. To znamená, že ani jeden z dávkovačů nemá systematickou chybu nebo jinak řečeno, velikosti dávek se skutečně pohybují kolem žádané (nastavené) hodnoty. Kritérium pro rozhodování může být rozptyl, neboť dávkovač, jehož odchylky v dávkách budou mít menší rozptyl, bude přesnější. Spočteme rozptyly pro oba dávkovače:

$$\begin{aligned} \mathrm{Var}\, X_1 &= \int_{-\infty}^{0} x^2 \tfrac{1}{2} e^{x}\, dx + \int_{0}^{\infty} x^2 \tfrac{1}{2} e^{-x}\, dx = 2 \int_{0}^{\infty} x^2 \tfrac{1}{2} e^{-x}\, dx = 2, \\ \mathrm{Var}\, X_2 &= 2 \int_{0}^{\infty} x^2 e^{-2x}\, dx = \frac{1}{2}. \end{aligned}$$

Z porovnání obou rozptylů vyplývá, že druhý dávkovač je přesnější.

□

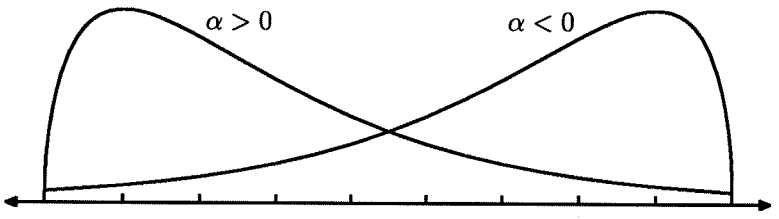
Podobně jako pro diskrétní náhodné veličiny platí

$$\begin{aligned} \mathrm{E}\,(\alpha X + \beta) &= \alpha\, \mathrm{E}\, X + \beta, \\ \mathrm{Var}\,(\alpha X + \beta) &= \alpha^2\, \mathrm{Var}\, X, \\ \mathrm{sd}\,(\alpha X + \beta) &= |\alpha|\, \mathrm{sd}\, X. \end{aligned}$$

Užitečnou charakteristikou pro zjišťování asymetrie je *šikmost*

$$\alpha_3 = \frac{\int_{-\infty}^{\infty} (x - \mathrm{E}\, X)^3 f(x)\, dx}{(\mathrm{Var}\, X)^{(3/2)}}.$$

Pro symetrická rozdělení platí $\alpha_3 = 0$. Pro rozdělení „sešikmená doleva“ je $\alpha_3 > 0$, pro rozdělení „sešikmená doprava“ je $\alpha_3 < 0$ (viz obrázek 7).



Obrázek 7.

Kromě šikmosti se též zavádí *špičatost*:

$$\alpha_4 = \frac{\int_{-\infty}^{\infty}(x - \mathrm{E}\,X)^4 f(x)\,dx}{(\mathrm{Var}\,X)^2} - 3.$$

Dalšími důležitými charakteristikami náhodné veličiny X jsou *horní kvantily. Horním* $100 \cdot p$ *% kvantilem* náhodné veličiny X nazveme číslo u_p takové, že pro dané $p\,(0 < p < 1)$ platí

$$P(X > u_p) = p.$$

Například jednoprocentním horním kvantilem náhodné veličiny udávající průtok v řece je taková horní mez, kterou překročí pouze tzv. „stoletá voda", tedy mez, která je překročena pouze v 1% případů.

Dolním $100 \cdot p$ *% kvantilem* náhodné veličiny X nazveme číslo v_p takové, že pro dané $p\,(0 < p < 1)$ platí

$$P(X < v_p) = p.$$

Zřejmě $v_p = u_{1-p}$.

Horní resp. dolní 50 % kvantil se nazývá *medián.*

Příklad 26.

Najděte kritickou dobu, kterou vydrží jen 5 % baterií, řídí-li se rozdělení doby životnosti hustotou z příkladu 22.

Řešení:

Kritická doba u, která nás zajímá, je vlastně 5 % horní kvantil, a tedy $P(X > u) = \int_u^{\infty} e^{-x}\,dx = 0.05$. Odtud $e^{-u} = 0.05$ a $u = -\ln(0.05) \doteq 3$. Kritickou dobou, kterou vydrží jen 5 % baterií, jsou přibližně tři roky.

□

12. Některé typy spojitě rozdělených náhodných veličin

V předchozím výkladu jsme se setkali s některými spojitě rozdělenými náhodnými veličinami, které patří ke známým a v praxi často používaným typům. Nyní si je probereme podrobněji.

Jednotlivá rozdělení téhož typu se liší jen jinou volbou hodnot parametrů. Abychom zdůraznili závislost hustoty na hodnotách parametrů, budeme je uvádět v argumentu hustoty za středník. Například hustotu $f(x)$ s parametrem θ budeme značit $f(x;\theta)$.

Rovnoměrné rozdělení R (α, β)

Náhodná veličina X má rovnoměrné rozdělení na intervalu $\langle \alpha, \beta \rangle$, jestliže pro hustotu $f(x; \alpha, \beta)$ platí

$$\begin{aligned} f(x; \alpha, \beta) &= \frac{1}{\beta - \alpha} && \text{pro} \quad x \in \langle \alpha, \beta \rangle, \\ &= 0 && \text{pro} \quad x \notin \langle \alpha, \beta \rangle. \end{aligned}$$

Platí

$$\mathrm{E}\, X = \frac{(\alpha + \beta)}{2} \quad \text{a} \quad \mathrm{Var}\, X = \frac{(\beta - \alpha)^2}{12}.$$

Příklad 27.

Určete rozdělení doby čekání na autobus z příkladu 21.

Řešení:

Doba čekání na autobus z příkladu 21 má rovnoměrné rozdělení $R(0, 5)$.

□

Příklad 28.

Součástí programového vybavení počítačů bývají často generátory náhodných čísel z rovnoměrného rozdělení, viz článek 41. Předpokládejme, že získané náhodné číslo je skutečně realizací náhodné veličiny X řídící se $R(0, 1)$. S jakou pravděpodobností bude mít takto vygenerované náhodné číslo na prvním místě jedničku?

Řešení

Náhodné číslo bude mít na prvním místě jedničku s pravděpodobností

$$p = P\big(X \in \langle 0.1, 0.2 \rangle\big) = \int_{0.1}^{0.2} 1\, dx = 0.1.$$

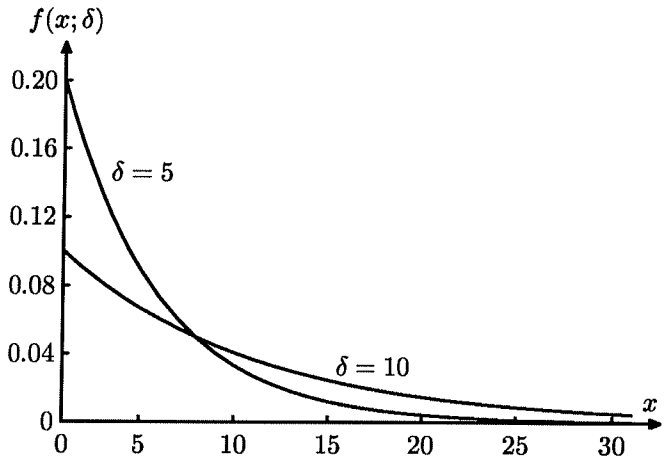
□

Exponenciální rozdělení E (δ)

Náhodná veličina X má exponenciální rozdělení s parametrem $\delta > 0$, jestliže hustota $f(x; \delta)$ má tvar

$$\begin{aligned} f(x; \delta) &= \frac{1}{\delta} \cdot e^{-(1/\delta)\, x} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0. \end{aligned}$$

Na obrázku 8 si prohlédněte hustoty exponenciálního rozdělení pro $\delta = 5$ a $\delta = 10$.



Obrázek 8.

Platí $\mathrm{E}\,X = \delta$ a $\mathrm{Var}\,X = \delta^2$. Exponenciálním rozdělením se řídí doby životnosti v teorii spolehlivosti nebo doby čekání na zákazníka v teorii obsluhy apod.

Příklad 29.

Určete rozdělení náhodných veličin z příkladu 24.

Řešení:

Náhodná veličina X_1 se řídí exponenciálním rozdělením s parametrem $\delta = 1$. Náhodná veličina X_2 se řídí rovněž exponenciálním rozdělením, avšak s parametrem $\delta = 1/2$.

□

Normální rozdělení N (μ, σ^2)

Náhodná veličina X má normální rozdělení s parametry $\mu \in R^1$, $\sigma^2 > 0$, jestliže pro hustotu $f(x;\mu,\sigma^2)$ platí

$$f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{-\frac{(x-\mu)^2}{2\,\sigma^2}\Big\} \qquad \text{pro} \quad x \in R^1.$$

Hustota $f(x;\mu,\sigma^2)$ je zvonovitého tvaru, symetrická kolem parametru μ. V případě normálního rozdělení lze snadno nahlédnout, jaký mají oba parametry význam, neboť

$$\mathrm{E}\,X = \mu \quad \text{a} \quad \mathrm{Var}\,X = \sigma^2.$$

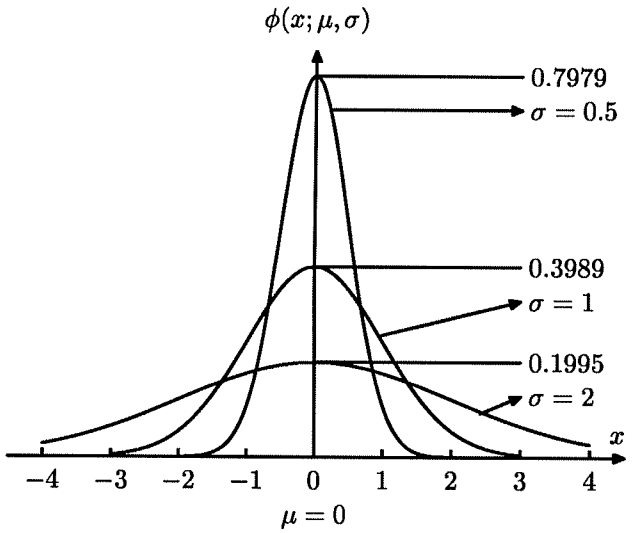
Znamená to tedy, že realizace náhodné veličiny X s normálním rozdělením $N(\mu,\sigma^2)$ se pohybují kolem hodnoty μ, přičemž jejich rozptýlení kolem parametru μ je dáno parametrem σ^2, resp. σ, kde $\sigma > 0$. Představu o velikosti rozptýlení veličiny X dávají následující

pravděpodobnosti:

$$P\big(X \in (\mu-\sigma, \mu+\sigma)\big) = 0.689,$$
$$P\big(X \in (\mu-2\sigma, \mu+2\sigma)\big) = 0.954,$$
$$P\big(X \in (\mu-3\sigma, \mu+3\sigma)\big) = 0.997.$$

Všimněme si, že náhodná veličina X s normálním rozdělením $N(\mu, \sigma^2)$ může sice teoreticky nabývat jakékoliv reálné hodnoty, ale fakticky se realizuje s pravděpodobností 0.997, tedy téměř rovnou 1, v mezích $(\mu - 3\sigma, \mu + 3\sigma)$.

Na obrázku 9 vidíme hustoty normálního rozdělení pro $\mu = 0$ a různé hodnoty σ.



Obrázek 9.

Významnou roli mezi všemi normálními rozděleními hraje rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$, kterému se říká standardní (normované) rozdělení. Jeho hustota $\phi(x)$ má tvar

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad \text{pro} \quad x \in R^1$$

a pro jeho distribuční funkci platí

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \qquad \text{pro} \quad x \in R^1.$$

Distribuční funkci $\Phi(x)$ nelze vyjádřit pomocí elementárních funkcí a výpočet je možný pouze numericky. Hodnoty distribuční funkce standardního normálního rozdělení lze najít v každých statistických tabulkách. Ze symetrie funkce $\phi(x)$ vyplývá pro distribuční funkci: $\Phi(-x) = 1 - \Phi(x)$ pro $x \in R^1$.

Jestliže náhodná veličina X má normální rozdělení $N(\mu, \sigma^2)$, pak náhodná veličina $Y = (X-\mu)/\sigma$ má standardní normální rozdělení. Odtud vyplývá vztah mezi distribuční funkcí

F normálního rozdělení s obecnými parametry μ, σ^2 a distribuční funkcí Φ standardního normálního rozdělení.

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Pravděpodobnost, že se náhodná veličina $X \sim N(\mu, \sigma^2)$ realizuje uvnitř intervalu (a, b), lze vyjádřit:

$$P(X \in (a,b)) = F(b) - F(a) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

Normální rozdělení je nejpoužívanějším rozdělením v teorii pravděpodobnosti a matematické statistice. Používá se tam, kde náhodná veličina vzniká jako součet velikého množství nezávislých vlivů. Tak vznikají například chyby měření apod.

Příklad 30.

Měření vzdálenosti od objektu je spojeno se systematickými i náhodnými chybami. Systematická chyba zkracuje vzdálenost a je rovna 0.05 m. Náhodné chyby mají normální rozdělení se směrodatnou odchylkou $\sigma = 0.1\,\text{m}$. Určete pravděpodobnost, že změřená vzdálenost nepřekročí vzdálenost skutečnou.

Řešení:

Chyba měření X, to je rozdíl mezi naměřenou hodnotou a skutečností, má normální rozdělení. Vzhledem k tomu, že přístroj má systematickou chybu, budou se chyby měření pohybovat kolem -0.05 m, a tedy $X \sim N(\mu = -0.05, \sigma^2 = 0.1^2)$. Změřená vzdálenost nepřekročí vzdálenost skutečnou, jestliže rozdíl mezi naměřenou hodnotou a skutečností bude záporný, tj. $X < 0$.

$$P(X < 0) = F(0) = \Phi\left(\frac{0-(-0.05)}{0.1}\right) = \Phi(0.5) = 0.6915.$$

Změřená vzdálenost nepřekročí skutečnou vzdálenost s pravděpodobností 0.6915.

□

Příklad 31.

Soustruh na vysoustružení hřídelek nemá systematickou chybu. Náhodné chyby mají normální rozdělení se směrodatnou odchylkou 0.5 mm. Vadné hřídelky jsou takové, jejichž poloměr se liší v absolutní hodnotě od správné hodnoty o víc než dovoluje toleranční mez. Jak stanovit toleranční mez, abychom vyřadili z výroby jako vadné 5 % nejhorších výrobků?

Řešení:

Náhodná veličiny X označující chybu soustružení má normální rozdělení. Soustruh nemá systematickou chybu, a tudíž hodnoty poloměru vysoustružených hřídelek se pohybují kolem správné hodnoty z výkresu. Odtud plyne, že střední chyba soustružení μ se rovná 0. Směrodatná odchylka $\sigma = 0.5$. Toleranční mez t chceme stanovit tak, aby se chyba soustružení 95 % výrobků pohybovala v mezích $(-t, t)$, tj. $P\big(X \in (-t,t)\big) = 0.95$. Odtud $\Phi\big(t/\sigma\big) - \Phi\big(-t/\sigma\big) = \Phi\big(t/\sigma\big) - \big(1 - \Phi\big(t/\sigma\big)\big) = 0.95$, a tedy $\Phi\big(t/\sigma\big) = 0.975$. Předchozí rovnost je splněna pro argument $t/\sigma = 1.96$, to znamená pro $t = 0.98$. Zvolíme-li toleranční mez pro absolutní odchylky od správné hodnoty 0.98 mm, vyloučíme 5 % nejhorších výrobků. □

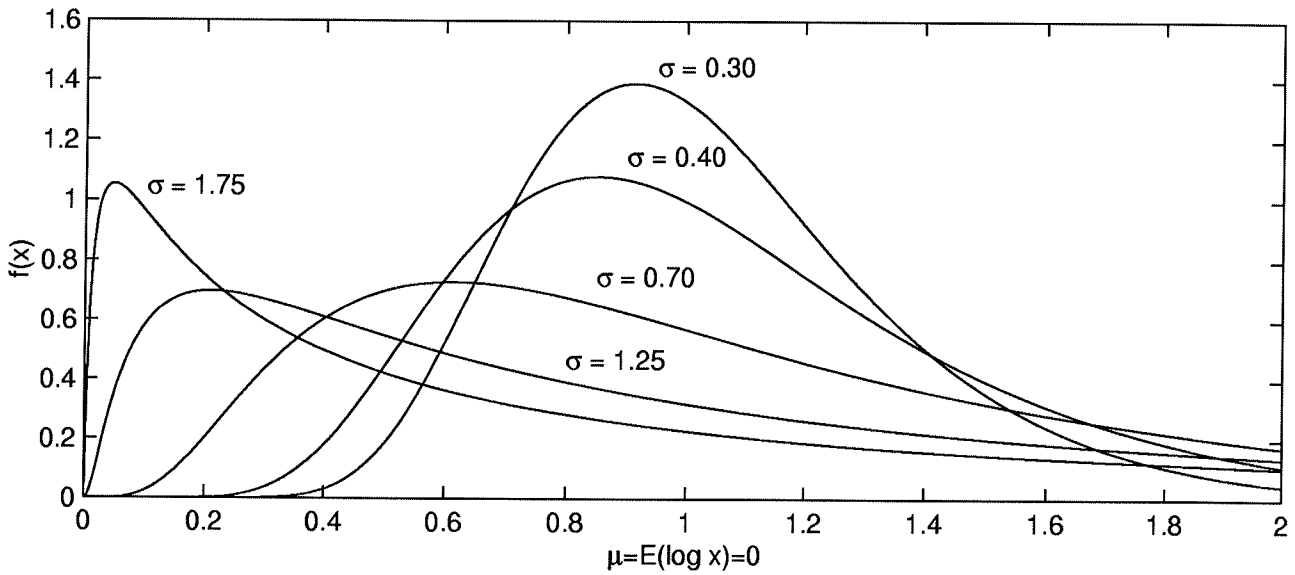
Logaritmicko–normální rozdělení LN (μ, σ^2, x_0)

Náhodná veličina X má logaritmicko–normální (též lognormální) rozdělení s parametry $\mu \in R^1$, $\sigma^2 > 0$ a $x_0 \in R^1$, jestliže její hustota pravděpodobnosti $f(x; \mu, \sigma^2, x_0)$ splňuje vztah:

$$f(x;\mu,\sigma^2,x_0) = \frac{1}{\sigma(x-x_0)\sqrt{2\pi}} \exp\left\{ -\frac{(\ln(x-x_0)-\mu)^2}{2\sigma^2} \right\} \qquad \text{pro} \quad x > x_0,$$

$$= 0 \qquad \text{pro} \quad x \le x_0.$$

Na obrázku 10 vidíme hustoty logaritmicko–normální rozdělení pro $\mu = 0, x_0 = 0$ a různé hodnoty σ.



Obrázek 10.

Náhodná veličina $Y = \ln(X - x_0)$, kde X má rozdělení $LN(\mu, \sigma^2, x_0)$ má normální rozdělení $N(\mu, \sigma^2)$. Odtud je možno odvodit vztah mezi distribuční funkcí $F(x)$ logaritmicko-normálního rozdělení s obecnými parametry μ, σ^2, x_0 a distribuční funkcí standardního normálního rozdělení

$$\begin{aligned} F(x) &= \Phi\left(\frac{\ln(x - x_0) - \mu}{\sigma}\right) && \text{pro} \quad x > x_0, \\ &= 0 && \text{pro} \quad x \leq x_0, \end{aligned}$$

který lze využít pro výpočet pravděpodobnosti realizace náhodné veličiny s logaritmicko-normálním rozdělením uvnitř intervalu (a, b) (předpokládáme $a,\ b > x_0$):

$$P\big(X \in (a,b)\big) = F(b) - F(a) = \Phi\left(\frac{\ln(b - x_0) - \mu}{\sigma}\right) - \Phi\left(\frac{\ln(a - x_0) - \mu}{\sigma}\right).$$

Pro střední hodnotu a rozptyl platí

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

Příklad 32.

Bylo zjištěno, že doba bezporuchové činnosti určitého výrobku udána v letech má $LN(\mu = 1.2, \sigma^2 = 0.25, x_0 = 0)$. Jakou dobu vydrží bez poruchy 5 % výrobků?

Řešení:

Hledáme 5 % horní kvantil $LN(\mu = 1.2, \sigma^2 = 0.25, x_0 = 0)$, a tedy číslo u, pro které platí $P(X > u) = 1 - F(u) = 0.05$. Odtud $\Phi\big(\frac{\ln u - 1.2}{0.5}\big) = 0.95$, a tedy $\frac{\ln u - 1.2}{0.5} = 1.645$ a $u = 7.557$. Kritickou dobou, kterou vydrží bez poruchy jen 5 % výrobků, je přibližně 7.6 let.

□

Pearsonovo rozdělení typu III

Pro modelování průtoků řek se používá Pearsonovo rozdělení typu III s hustotou

$$\begin{aligned} f(x; \alpha, \kappa, \nu) &= \alpha^\kappa \frac{(x - \nu)^{\kappa - 1}}{\Gamma(\kappa)} \exp\{-\alpha(x - \nu)\} && \text{pro} \quad x > \nu, \\ &= 0 && \text{pro} \quad x \leq \nu, \end{aligned}$$

kde $\alpha, \kappa > 0$ a $\nu \in R^1$.

Funkce $\Gamma(\beta)$ se nazývá gamma funkce a je definována vztahem $\Gamma(\beta) = \int_0^\infty x^{\beta-1} e^{-x}\, dx$. Pro přirozená n platí $\Gamma(n) = (n-1)!$.

Základní charakteristiky tohoto rozdělení splňují:

$$\begin{aligned} \mathrm{E}\, X &= \frac{\kappa}{\alpha} + \nu, \\ \mathrm{Var}\, X &= \frac{\kappa}{\alpha^2}, \\ \text{šikmost } \alpha_3 &= \frac{2}{\sqrt{\kappa}}. \end{aligned}$$

Poznámka:

Pearsonovo rozdělení typu III se někdy nazývá také gamma rozdělení.

Příklad 33.

Předpokládejme, že náhodná veličina X se řídí Pearsonovým rozdělením typu III se střední hodnotou 142, směrodatnou odchylkou 35 a šikmostí 0.55. S jakou pravděpodobností nabude hodnoty menší než 110?

Řešení:

Ze vztahů pro charakteristiky Pearsonova rozdělení můžeme postupně vypočítat:

$$\begin{aligned} \kappa &= \frac{4}{\alpha_3^2} = \frac{4}{0.55^2} = 13.22314, \\ \alpha &= \sqrt{\frac{\kappa}{\mathrm{Var}\, X}} = \sqrt{\frac{13.22314}{35^2}} = 0.103896, \\ \nu &= \mathrm{E}\, X - \frac{\kappa}{\alpha} = 14.72715. \end{aligned}$$

Hledaná pravděpodobnost by pak byla rovna $\int_\nu^{110} f(x; \alpha, \kappa, \nu)\, dx$. Tento integrál však lze vypočítat jen numericky. Nemáme-li k dispozici počítač vybavený vhodným programem, je schůdnější cestou využití statistických tabulek. Veličina $Y = \frac{X - \mathrm{E}\, X}{\mathrm{sd}\, X}$ je tzv. normovaná veličina vzniklá z X. Přitom je obecně vždy $\mathrm{E}\, Y = 0$, $\mathrm{Var}\, Y = 1$, $\alpha_3(Y) = \alpha_3(X)$. Má-li X Pearsonovo rozdělení typu III, bude mít Y též Pearsonovo rozdělení typu III, které bude normované, s nezměněnou šikmostí. Proto

$$P(X < 110) = P\Big(Y < \frac{110 - 142}{35}\Big) \doteq F_Y(-0.914),$$

kde F_Y je distribuční funkce normovaného Pearsonova rozdělení typu III se šikmostí 0.55. V příslušné tabulce při dvojnásobné lineární interpolaci (je třeba interpolovat vzhledem k nezávislé proměnné i vzhledem k šikmosti) najdeme $F_Y(-0.914) \doteq 0.1814$. □

Rozdělení χ^2, t, F

Ve statistických aplikacích se vyskytují ještě další typy rozdělení, u kterých nebudeme uvádět přesný tvar hustot. Čtenář si je může vyhledat např. v Andělovi (1976). Ve výkladu matematické statistiky však budeme pracovat s kvantily těchto rozdělení, a proto zde zavedeme příslušná označení.

Rozdělení χ^2

Hustota rozdělení χ^2 (chí kvadrát) je kladná pouze pro kladné hodnoty argumentu. Má jediný parametr ν, jehož hodnotou může být pouze přirozené číslo. Říká se mu počet stupňů volnosti.

Uvažujeme posloupnost nezávislých stejně rozdělených náhodných veličin $X_1, \ldots, X_\nu$ (viz článek 15) řídících se normálním rozdělením $N(0,1)$, pak náhodná veličina

$$X = X_1^2 + X_2^2 + \cdots + X_\nu^2$$

se řídí χ^2 rozdělením o ν stupních volnosti.

Dolní i horní kvantily jsou uvedeny ve statistických tabulkách. Horní $100\,\mathrm{p}$ % kvantil χ^2 rozdělení o ν stupních volnosti budeme značit $\chi_p^2[\nu]$.

Rozdělení t

Hustota rozdělení t (Studentova) je symetrická kolem nuly. Její jediný parametr ν nabývá hodnot z množiny přirozených čísel a říká se mu rovněž počet stupňů volnosti.

Uvažujeme dvě nezávislé náhodné veličiny U a V, přičemž veličina U se řídí normálním rozdělením $N(0,1)$ a veličina V se řídí χ^2 rozdělením o ν stupních volnosti, pak náhodná veličina

$$X = \frac{U}{\sqrt{V}}\sqrt{\nu}$$

se řídí t rozdělením o ν stupních volnosti.

Kvantily t rozdělení jsou opět ve statistických tabulkách. Horní $100\,\mathrm{p}$ % kvantil t rozdělení o ν stupních volnosti budeme značit $t_p[\nu]$. Ze symetrie t rozdělení vyplývá, že $100\,\mathrm{p}$ % horní a dolní kvantily se liší pouze znaménkem.

Rozdělení F

Hustota rozdělení F (Fisher – Snedeckorova) je kladná jen pro kladné hodnoty argumentu. Má dva parametry ν_1, ν_2 nazývané počty stupňů volnosti, jejichž hodnoty mohou být pouze přirozená čísla.

Uvažujme dvě nezávislé náhodné veličiny W a Z, přičemž veličina W se řídí χ^2 rozdělením o ν_1 stupních volnosti a veličina Z se řídí χ^2 rozdělením o ν_2 stupních volnosti, pak náhodná veličina

$$X = \frac{W/\nu_1}{Z/\nu_2}$$

se řídí F rozdělením o ν_1 a ν_2 stupních volnosti.

Horní 100 p % kvantily F rozdělení o ν_1 a ν_2 stupních volnosti $F_p[\nu_1, \nu_2]$ jsou tabelované ve statistických tabulkách. Tvar hustoty, a tudíž i hodnoty kvantilů závisí na pořadí stupňů volnosti, a proto $F_p[\nu_1, \nu_2] \neq F_p[\nu_2, \nu_1]$, jestliže $\nu_1 \neq \nu_2$.

Část III. Náhodné vektory a jejich rozdělení

13. Náhodný vektor

Výsledkem náhodného pokusu často není jediné číslo, ale $n-$tice reálných čísel. Sledujeme-li například pohyb hmotného bodu v prostoru, zaznamenáváme všechny tři jeho souřadnice. Jindy zjišťujeme tlak a teplotu zároveň apod. Ve všech takových případech pozorujeme $n-$tici náhodných veličin $\boldsymbol{X} = (X_1, \ldots, X_n)'$ (značka $'$ znamená transpozici), kterou nazýváme *náhodný vektor.*

V praxi se často setkáváme s tím, že jsou všechny souřadnice náhodného vektoru $\boldsymbol{X} = (X_1, \ldots, X_n)'$ buď pouze diskrétní nebo pouze spojité. Pokud jsou všechny složky diskrétní, říkáme, že vektor $\boldsymbol{X} = (X_1, \ldots, X_n)'$ má vícerozměrné diskrétní rozdělení. Vícerozměrné (sdružené) diskrétní rozdělení je dáno výčtem všech vektorů $(x_1, \ldots, x_n)'$, které může náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)'$ nabývat a příslušnými pravděpodobnostmi $P(X_1 = x_1, \ldots, X_n = x_n)$. Pokud má náhodný vektor jen dvě složky, tj. pracujeme-li s vektorem $\boldsymbol{X} = (X_1, X_2)'$, a pokud počet hodnot, které složky nabývají, je malý, je možno zadat rozdělení tabulkou.

Příklad 34.

Následující tabulka udává pravděpodobnosti, s jakými žáci posledního ročníku základní školy získají známky $1, 2, 3, 4, 5$ z matematiky a fyziky.

F \ M	1	2	3	4	5
1	0.11	0.12	0.05	0.01	0.00
2	0.08	0.09	0.07	0.02	0.00
3	0.05	0.07	0.09	0.07	0.01
4	0.02	0.01	0.01	0.06	0.03
5	0.00	0.00	0.00	0.02	0.01

Tabulka 11.

a) S jakou pravděpodobností má náhodně vybraný žák jedničku z obou předmětů?

b) S jakou pravděpodobností má náhodně vybraný žák z obou předmětů lepší známku než trojku?

c) S jakou pravděpodobností má náhodně vybraný žák jedničku z matematiky?

d) S jakou pravděpodobností má náhodně vybraný žák alespoň z jednoho z těchto předmětů horší známku než dvojku?

Řešení:

Uvažujme náhodný vektor $(M, F)'$, kde náhodná veličina M označuje známku z matematiky a náhodná veličina F známku z fyziky. Všechny možné hodnoty vektoru $(M, F)'$ jsou uspořádané dvojice $(i, j)'$, kde $i = 1, 2, 3, 4, 5$ a $j = 1, 2, 3, 4, 5$.

a) $P(M = 1, F = 1) = 0.11$,

b) $P(M \leq 2, F \leq 2) =$
$= P(M = 1, F = 1) + P(M = 2, F = 2) + P(M = 1, F = 2) + P(M = 2, F = 1) = 0.40$,

c) $P(M = 1) = \sum_{j=1}^{5} P(M = 1, F = j) = 0.26$

d) $P(M > 2 \cup F > 2) = \sum_{i=1}^{2} \sum_{j=3}^{5} P(M = i, F = j) + \sum_{i=3}^{5} \sum_{j=1}^{5} P(M = i, F = j) =$
$= 0.60$.

K témuž výsledku můžeme dojít také následovně $P(M > 2 \cup F > 2) = 1 - P(M \leq 2 \cap F \leq 2) = 0.60$.

□

Zajímáme-li se o rozdělení menšího počtu souřadnic (velmi často například o rozdělení pouze jedné souřadnice), pak se takovému rozdělení říká *marginální*. Marginální rozdělení veličiny X_i je dáno pravděpodobnostmi $\{P(X_i = x_i)\}$ pro všechny hodnoty $\{x_i\}$, které může veličina X_i nabývat. Pravděpodobnost $P(X_i = x_i)$ se získá vysčítáním pravděpodobností $P(X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x_i, X_{i+1} = x_{i+1}, \ldots, X_n = x_n)$ přes všechny možné hodnoty $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, tj.

$$P(X_i = x_i) =$$
$$\sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_n} P(X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_i = x_i, X_{i+1} = x_{i+1}, \ldots, X_n = x_n)$$

Příklad 35.

S pomocí tabulky 11 určete rozdělení náhodné veličiny M, která označuje známku z matematiky v posledním ročníku základní školy.

Řešení:

x	1	2	3	4	5
$P(M = x)$	0.26	0.29	0.22	0.18	0.05

Tabulka 12

□

Ve shodě s definicí podmíněné pravděpodobnosti definujeme *podmíněnou pravděpodobnost toho, že náhodná veličina X_1 nabude hodnoty x_1 za podmínky, že veličina X_2 nabyla hodnoty x_2* pomocí

$$P(X_1 = x_1 | X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)},$$

jestliže $P(X_2 = x_2) > 0$. Pravděpodobnosti $P(X_1 = x_1 | X_2 = x_2)$ pro všechny možné hodnoty x_1 určují *podmíněné rozdělení veličiny X_1 za podmínky $X_2 = x_2$*.

Příklad 36.

S pomocí tabulky 12 určete rozdělení náhodné veličiny M za podmínky, že $F = 1$.

Řešení:

x	1	2	3	4	5
$P(M = x \mid F = 1)$	0.379	0.414	0.172	0.034	0.000

Tabulka 13.

Interpretujeme-li pravděpodobnosti pomocí četností, pak si všimněme, že podíl žáků, kteří mají jedničku z matematiky tvoří 26 % z celkového množství všech žáků, zatímco jejich podíl mezi těmi žáky, kteří mají jedničku z fyziky, tvoří 37.9 %.

□

Náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)'$ má rozdělení spojitého typu, existuje-li nezáporná reálná funkce $f(x_1, \ldots, x_n)$ taková, že

$$P\big(X_1 \in (a_1, b_1), \ldots, X_n \in (a_n, b_n)\big) = \int_{a_1}^{b_1} \cdots \int_{a_n}^{b_n} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n.$$

Funkci $f(x_1, \ldots, x_n)$ říkáme *sdružená hustota*. Zřejmě musí platit

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = 1.$$

Hustota $f_i(x_i)$ pouze jediné složky X_i, $i = 1, \ldots, n$ se získá ze sdružené hustoty následovně:

$$f_i(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n.$$

Této hustotě se říká *marginální hustota* veličiny X_i.

Podmíněná hustota pravděpodobnosti $f(x_1 | x_2)$ veličiny X_1 za podmínky $X_2 = x_2$ je dána výrazem

$$f(x_1 | x_2) = \frac{f(x_1, x_2)}{f_2(x_2)},$$

jestliže $f_2(x_2) > 0$. Podmíněná pravděpodobnost toho, že veličina X_1 nabude hodnoty z množiny A za podmínky, že $X_2 = x_2$ se spočte pomocí podmíněné hustoty:

$$P(X_1 \in A | X_2 = x_2) = \int_A f(x_1 | x_2)\, dx_1.$$

14. Charakteristiky rozdělení náhodného vektoru

Nejvýznamnější charakteristikou náhodného vektoru $\boldsymbol{X} = (X_1, \ldots, X_n)'$ je jeho *střední hodnota* - vektor $(\mathrm{E}\, X_1, \ldots, \mathrm{E}\, X_2)'$. Jeho $i-$tá složka $\mathrm{E}\, X_i$, $i = 1, \ldots, n$, je střední hodnotou veličiny X_i. Pro diskrétní rozdělení

$$\mathrm{E}\, X_i = \sum_{x_1} \cdots \sum_{x_i} \cdots \sum_{x_n} x_i\, P(X_1 = x_1, \ldots, X_n = x_n) = \sum_{x_i} x_i\, P(X_i = x_i),$$

zatímco pro spojité rozdělení

$$\mathrm{E}\, X_i = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_i\, f(x_1, \ldots, x_n)\, dx_1 \ldots dx_n = \int_{-\infty}^{\infty} x_i\, f_i(x_i)\, dx_i.$$

Druhou významnou charakteristikou je *kovarianční matice*

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \vdots & \vdots & \ldots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix},$$

jejíž prvky σ_{ij} se nazývají *kovariance* náhodných veličin X_i, X_j a značí se $\mathrm{cov}(X_i, X_j)$. Pro diskrétní rozdělení

$$\begin{aligned} \sigma_{ij} &= \sum_{x_1} \cdots \sum_{x_n} (x_i - \mathrm{E}\, X_i)(x_j - \mathrm{E}\, X_j)\, P(X_1 = x_1, \ldots, X_n = x_n) \\ &= \sum_{x_i} \sum_{x_j} (x_i - \mathrm{E}\, X_i)(x_j - \mathrm{E}\, X_j)\, P(X_i = x_i, X_j = x_j) \\ &= \sum_{x_i} \sum_{x_j} x_i x_j\, P(X_i = x_i, X_j = x_j) - (\mathrm{E}\, X_i)(\mathrm{E}\, X_j) \\ &= \mathrm{E}\,(X_i X_j) - (\mathrm{E}\, X_i)(\mathrm{E}\, X_j), \end{aligned}$$

zatímco pro spojité rozdělení

$$\begin{aligned}\sigma_{ij} &= \int_{\infty}^{\infty} \dots \int_{\infty}^{\infty} (x_i - \mathrm{E}\, X_i)(x_j - \mathrm{E}\, X_j) f(x_1, \dots, x_n)\, dx_1 \dots x_n \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_i x_j\, f_{ij}(x_i, x_j)\, dx_i dx_j \;-\; (\mathrm{E}\, X_i)(\mathrm{E}\, X_j) \\ &= \mathrm{E}\,(X_i X_j) \;-\; (\mathrm{E}\, X_i)(\mathrm{E}\, X_j).\end{aligned}$$

Všimněme si, že $\sigma_{ii} = \mathrm{cov}(X_i, X_i) = \mathrm{Var}\, X_i$ a $\sigma_{ij} = \sigma_{ji}$.

Pokud s velkými pravděpodobnostmi nabývají veličiny X_i a X_j současně obě dvě velkých, tj. nadprůměrných hodnot, nebo obě dvě malých, tj. podprůměrných hodnot, pak je kovariance $\mathrm{cov}(X_i, X_j)$ kladná. Pokud s velkými pravděpodobnostmi dochází k tomu, že současně jedna z veličin nabývá velkých, tj. nadprůměrných hodnot, a druhá malých, tj. podprůměrných hodnot, a naopak, pak je $\mathrm{cov}(X_i, X_j)$ záporná.

Příklad 37.

Spočtěte $\mathrm{cov}(M, F)$, kde M a F jsou náhodné veličiny z příkladu 34.

Řešení:

$$\begin{aligned}\mathrm{cov}(M,F) = \;& 1\cdot 1\cdot 0.11 + 1\cdot 2\cdot 0.12 + 1\cdot 3\cdot 0.05 + 1\cdot 4\cdot 0.01 + 1\cdot 5\cdot 0.00 + \\ &+ 2\cdot 1\cdot 0.08 + 2\cdot 2\cdot 0.09 + 2\cdot 3\cdot 0.07 + 2\cdot 4\cdot 0.02 + 2\cdot 5\cdot 0.00 + \\ &+ 3\cdot 1\cdot 0.05 + 3\cdot 2\cdot 0.07 + 3\cdot 3\cdot 0.09 + 3\cdot 4\cdot 0.07 + 3\cdot 5\cdot 0.01 + \\ &+ 4\cdot 1\cdot 0.02 + 4\cdot 2\cdot 0.01 + 4\cdot 3\cdot 0.01 + 4\cdot 4\cdot 0.06 + 4\cdot 5\cdot 0.03 + \\ &+ 5\cdot 1\cdot 0.00 + 5\cdot 2\cdot 0.00 + 5\cdot 3\cdot 0.00 + 5\cdot 4\cdot 0.02 + 5\cdot 5\cdot 0.01 - \\ &- \big(1\cdot 0.26 + 2\cdot 0.29 + 3\cdot 0.22 + 4\cdot 0.18 + 5\cdot 0.05\big)\cdot \\ &\cdot \big(1\cdot 0.29 + 2\cdot 0.26 + 3\cdot 0.29 + 4\cdot 0.13 + 5\cdot 0.03\big) = \\ =\;& 6.50 - 2.47\cdot 2.35 = 0.6955.\end{aligned}$$

Kovariance je kladná, což odpovídá tomu, že se dosti často vyskytují žáci, kteří mají z obou předmětů dobrou známku a také se často vyskytují žáci, kteří mají z obou těchto

předmětů špatnou známku. Naopak poměrně zřídka se setkáváme s žáky, kteří mají dobrou známku z jednoho ze sledovaných předmětů a špatnou známku z druhého předmětu.

□

Normujeme-li kovarianci $\mathrm{cov}(X_i, X_j)$ směrodatnými odchylkami náhodných veličin X_i a X_j, tj. $\mathrm{sd}\, X_i = \sqrt{\mathrm{Var}\, X_i}$ a $\mathrm{sd}\, X_j = \sqrt{\mathrm{Var}\, X_j}$, získáme charakteristiku, která se nazývá *korelace* náhodných veličin X_i a X_j, jinak též *korelační koeficient*:

$$\mathrm{corr}(X_i, X_j) = \frac{\mathrm{cov}(X_i, X_j)}{(\mathrm{sd}\, X_i) \cdot (\mathrm{sd}\, X_j)}.$$

Hodnota kovariance se mění se změnou měřítka, ve kterém jsou veličiny udávány. Výhodou korelačního koeficientu je, že se jeho hodnota nemění, změníme-li měřítko jedné nebo obou veličin lineárně, tj. jestliže $Y_1 = aX_1 + b_1$, $Y_2 = cX_2 + d$, kde $a > 0, c > 0$, pak $\mathrm{corr}(Y_1, Y_2) = \mathrm{corr}(X_1, X_2)$. Označuje-li např. X_1 teplotu vzduchu měřenou ve stupních Fahrenheita a X_2 množství srážek měřených v palcích a Y_1 je tatáž teplota měřená ve stupních Celsia a Y_2 množství srážek měřené v mm, pak bude v obou případech korelační koeficient mezi teplotou a srážkami stejný. Korelační koeficient může nabývat pouze hodnot z intervalu $\langle -1, 1 \rangle$.

Následující příklad ukazuje výpočet charakteristik náhodného vektoru, který má spojité rozdělení.

Příklad 38.

Nechť je bod B náhodně vybrán z trojúhelníku T s vrcholy (0,0), (0,1), (1,1). Najděte sdruženou hustotu vektoru $(X, Y)'$, kde X označuje $x-$ovou a Y označuje $y-$ovou souřadnici bodu B. Najděte marginální hustoty X a Y, vektor středních hodnot a kovarianční matici vektoru $(X, Y)'$. Najděte $\mathrm{corr}(X, Y)$.

Řešení:

Bod B je vybrán z trojúhelníku $T = \{0 \leq x \leq 1,\ 0 \leq y \leq 1 - x\}$ náhodně, a proto musí být hustota $f(x, y)$ uvnitř trojúhelníka konstantní. Protože $\iint_T f(x, y)\, dxdy = 1$ a obsah trojúhelníka T je $1/2$, platí

$$\begin{aligned} f_(x, y) &= 2 && \text{pro} \quad (x, y) \in T, \\ &= 0 && \text{pro} \quad (x, y) \notin T. \end{aligned}$$

Ze symetrie X a Y plyne, že marginální hustota, střední hodnota a rozptyl náhodné veličiny X jsou stejné jako náhodné veličiny Y. Marginální hustota

$$\begin{aligned} f_1(x) &= \int_0^{1-x} 2\,dy = 2(1-x) && \text{pro} \quad x \in \langle 0,1\rangle, \\ &= 0 && \text{pro} \quad x \notin \langle 0,1\rangle. \end{aligned}$$

Dále

$$\begin{aligned} \mathrm{E}\,X &= \int_0^1 \int_0^{1-x} 2\,x\,dy\,dx = \int_0^1 2\,x\,(1-x)\,dx = \frac{1}{3}, \\ \mathrm{Var}\,X &= \int_0^1 \int_0^{1-x} 2\,x^2\,dy\,dx - \left(\frac{1}{3}\right)^2 = \int_0^1 2\,x^2(1-x)\,dx - \left(\frac{1}{3}\right)^2 = \frac{1}{18}, \\ \mathrm{cov}(X,Y) &= \int_0^1 \int_0^{1-x} 2\,xy\,dy\,dx - \frac{1}{3}\cdot\frac{1}{3} = -\frac{1}{36}, \\ \mathrm{corr}(X,Y) &= \frac{-1/36}{\sqrt{1/18}\sqrt{1/18}} = -\frac{1}{2}. \end{aligned}$$

Střední hodnota vektoru $(X,Y)'$ je tedy $(1/3, 1/3)'$ a kovarianční matice

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1/18 & -1/36 \\ -1/36 & 1/18 \end{pmatrix}.$$

□

15. Nezávislost náhodných veličin

O dvou náhodných veličinách si intuitivně představujeme, že jsou nezávislé, jestliže výsledek realizace jedné z nich neovlivní výsledek realizace druhé a naopak. Matematicky lze tuto skutečnost vyjádřit následovně:

Náhodné veličiny X a Y jsou *nezávislé*, jestliže pro libovolné intervaly A a B platí:

$$P((x \in A) \cap (Y \in B)) = P(X \in A) \cdot P(Y \in B).$$

Pro diskrétně rozdělené náhodné veličiny to znamená, že pro libovolné x a y:

$$P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y).$$

Pro spojitě rozdělené náhodné veličiny to znamená, že sdružená hustota je součinem marginálních hustot

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

Odtud plyne, že pro nezávislé náhodné veličiny platí

$$\mathrm{E}(XY) = \sum_x \sum_y xy\, P(X=x, Y=y) = \Big(\sum_x x\, P(X=x)\Big)\Big(\sum_y y\, P(Y=y)\Big) =$$

$$= (\mathrm{E}\,X)(\mathrm{E}\,Y),$$

$$\mathrm{E}(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dxdy = \Big(\int_{-\infty}^{\infty} x\, f_1(x)\, dx\Big)\Big(\int_{-\infty}^{\infty} y\, f_2(y)\, dy\Big) = (\mathrm{E}\,X)(\mathrm{E}\,Y),$$

a tedy $\mathrm{cov}(X,Y) = 0$. Zdůrazněme ještě jednou slovně důležitou skutečnost, že kovariance resp. korelace, dvou nezávislých náhodných veličin je nulová.

Příklad 39.

Jsou veličiny M a F z příkladu 34 nezávislé?

Řešení:

$$P(M=1, F=1) = 0.11 \neq P(M=1)\cdot P(F=1) = 0.26\cdot 0.29 = 0.0754.$$

Veličiny M a F nejsou nezávislé. Závislost veličin M a F plyne také z faktu, že $\mathrm{corr}(M,F) \neq 0$. □

16. Charakteristiky lineární kombinace náhodných veličin

Pro libovolné náhodné veličiny X a Y platí

$$\mathrm{E}(aX + bY) = a\,\mathrm{E}\,X + b\,\mathrm{E}\,Y,$$

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}\,X + 2ab\,\mathrm{cov}(X,Y) + b^2\mathrm{Var}\,Y.$$

Speciálně pro nezávislé náhodné veličiny platí

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}\,X + b^2\mathrm{Var}\,Y.$$

Odtud například plyne, že pro součet dvou nezávislých náhodných veličin platí

$$\mathrm{E}(X+Y) = \mathrm{E}\,X + \mathrm{E}\,Y \quad \text{a} \quad \mathrm{Var}(X+Y) = \mathrm{Var}\,X + \mathrm{Var}\,Y$$

a pro rozdíl dvou nezávislých náhodných veličin platí

$$\mathrm{E}(X-Y) = \mathrm{E}\,X - \mathrm{E}\,Y \quad \text{a} \quad \mathrm{Var}(X-Y) = \mathrm{Var}\,X + \mathrm{Var}\,Y.$$

Příklad 40.

Uvažujme dvě nezávislé stejně rozdělené náhodné veličiny X a Y nabývající pouze dvou hodnot 1 a -1, přičemž

$$P(X=1) = P(X=-1) = 1/2,$$
$$P(Y=1) = P(Y=-1) = 1/2.$$

Spočtěte střední hodnotu a rozptyl náhodné veličiny $S = X+Y$ a veličiny $R = X-Y$.

Řešení:

Sdružené rozdělení veličin X a Y je následující:

$$\begin{aligned}
P(X=1, Y=1) &= P(X=1)\,P(Y=1) &&= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},\\
P(X=1, Y=-1) &= P(X=1)\,P(Y=-1) &&= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},\\
P(X=-1, Y=1) &= P(X=-1)\,P(Y=1) &&= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},\\
P(X=-1, Y=-1) &= P(X=-1)\,P(Y=-1) &&= \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4},
\end{aligned}$$

Zřejmě $\mathrm{E}\,X = \mathrm{E}\,Y = 0$ a $\mathrm{Var}\,X = \mathrm{Var}\,Y = (-1)^2\cdot(1/2) + 1^2\cdot(1/2) = 1$.

Náhodná veličina S nabývá hodnot -2, 0, 2, přičemž

$$\begin{aligned}
P(S=-2) &= P(X=-1, Y=-1) = \frac{1}{4},\\
P(S=0) &= P\big((X=-1, Y=1)\cup(X=1, Y=-1)\big) =\\
&= P(X=-1, Y=1) + P(X=1, Y=-1) = \frac{1}{4}+\frac{1}{4} = \frac{1}{2},\\
P(S=2) &= P(X=1, Y=1) = \frac{1}{4}.
\end{aligned}$$

Odtud $\mathrm{E}\,S = 0$ a $\mathrm{Var}\,S = (-2)^2\cdot(1/4) + 2^2\cdot(1/4) = 2$. Tentýž výsledek jsme mohli též získat z pravidla o počítání střední hodnoty a rozptylu pro součet dvou nezávislých náhodných veličin:

$$\mathrm{E}\,S = \mathrm{E}\,X + \mathrm{E}\,Y = 0,$$
$$\mathrm{Var}\,S = \mathrm{Var}\,X + \mathrm{Var}\,Y = 1+1 = 2.$$

Je zajímavé, že náhodná veličina R má stejné rozdělení jako veličina S. Nabývá totiž rovněž hodnot -2, 0, 2 s pravděpodobnostmi

$$\begin{aligned} P(R=-2) &= P(X=-1,Y=1) = \frac{1}{4}, \\ P(R=0) &= P\big((X=1,Y=1)\cup(X=-1,Y=-1)\big) = \frac{1}{4}+\frac{1}{4} = \frac{1}{2}, \\ P(R=2) &= P(X=1,Y=-1) = \frac{1}{4}. \end{aligned}$$

Odtud $\mathrm{E}\,R = 0$ a $\mathrm{Var}\,R = (-2)^2 \cdot (1/4) + 2^2 \cdot (1/4) = 2$. Shodný výsledek jsme mohli získat též z pravidla o počítání střední hodnoty a rozptylu rozdílu dvou nezávislých náhodných veličin:

$$\begin{aligned} \mathrm{E}\,R &= \mathrm{E}\,X - \mathrm{E}\,Y = 0, \\ \mathrm{Var}\,R &= \mathrm{Var}\,X + \mathrm{Var}\,Y = 1+1 = 2. \end{aligned}$$

□

Výpočet charakteristik lineární kombinace dvou náhodných veličin je speciálním případem výpočtu střední hodnoty a rozptylu pro lineární kombinaci veličin $X_1, \ldots, X_n$, tj. veličinu $c_1X_1 + \cdots + c_nX_n$, kde $c_1 \in R^1, \ldots, c_n \in R^1$:

$$\mathrm{E}(c_1X_1 + \cdots + c_nX_n) = \sum_{i=1}^{n} c_i\,\mathrm{E}\,X_i, \tag{16.1}$$

$$\mathrm{Var}(c_1X_1 + \cdots + c_nX_n) = \sum_{i=1}^{n} c_i^2 \mathrm{Var}\,X_i + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} c_ic_j\,\mathrm{cov}(X_i,X_j). \tag{16.2}$$

Pro nezávislé náhodné veličiny $X_1, \ldots, X_n$ platí

$$\mathrm{Var}\left(\sum_{i=1}^{n} c_i\,X_i\right) = \sum_{i=1}^{n} c_i^2\,\mathrm{Var}\,X_i.$$

Příklad 41.

Uvažujme $n-$tici nezávislých stejně rozdělených náhodných veličin $X_1, \ldots, X_n$ se střední hodnotou $\mathrm{E}\,X_i = \mu$ a $\mathrm{Var}\,X_i = \sigma^2$. Spočtěte střední hodnotu a rozptyl aritmetického průměru $\overline{X} = \sum_{i=1}^{n} X_i/n$.

Řešení:

Aritmetický průměr $\overline{X}$ je lineární kombinací veličin $X_1, \ldots, X_n$, kde $c_1 = c_2 = \cdots = c_n = 1/n$. Odtud

$$\mathrm{E}\,\overline{X} = \frac{1}{n}\mathrm{E}\,X_1 + \cdots + \frac{1}{n}\mathrm{E}\,X_n = \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu,$$
$$\mathrm{Var}\,\overline{X} = \left(\frac{1}{n}\right)^2 \mathrm{Var}\,X_1 + \cdots + \left(\frac{1}{n}\right)^2 \mathrm{Var}\,X_n = \left(\frac{1}{n}\right)^2 \sigma^2 + \cdots + \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

Znamená to, že aritmetický průměr kolísá kolem téže střední hodnoty μ, avšak s menšími odchylkami než původní veličiny.

□

17. Vícerozměrné normální rozdělení

Náhodný vektor $\boldsymbol{X} = (X_1, \ldots, X_n)'$ má *vícerozměrné normální rozdělení* s parametry $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$, jestliže sdružená hustota má tvar

$$f(x_1, \ldots, x_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{(n/2)}(\det \boldsymbol{\Sigma})^{1/2}} \exp\Big\{ - \frac{(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}{2} \Big\},$$

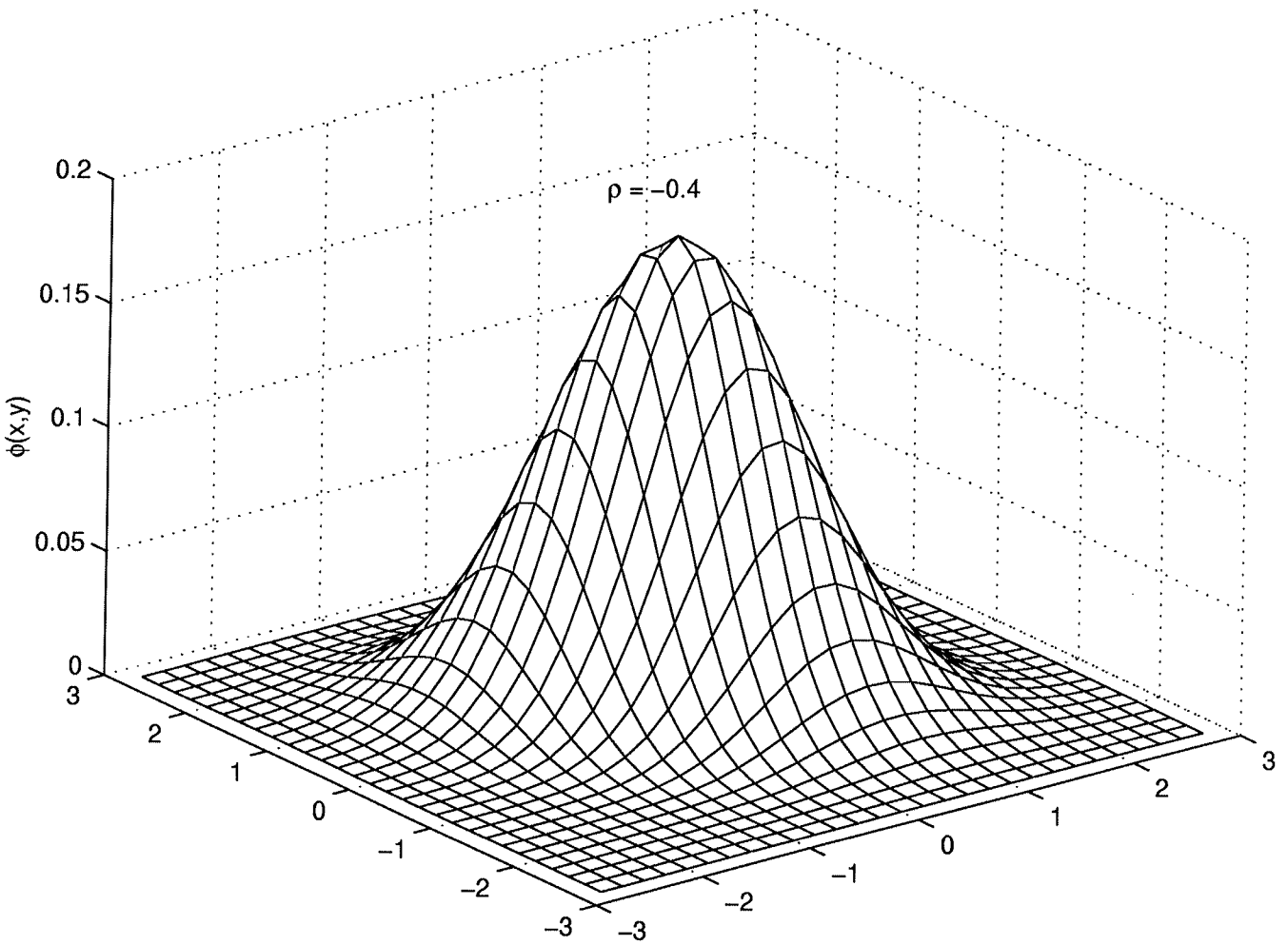
kde $\boldsymbol{\mu}$ je reálný vektor, $\boldsymbol{\Sigma}$ symetrická, pozitivně definitní matice. Význam parametrů $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ je dán faktem, že střední hodnota vektoru $\boldsymbol{X}$ je rovna $\boldsymbol{\mu}$ a kovarianční matice je rovna $\boldsymbol{\Sigma}$.

Speciálním případem vícerozměrného normálního rozdělení je *dvojrozměrné normální rozdělení*. Zavedeme-li $\sigma_1 > 0$, $\sigma_2 > 0, \rho \in R^1$ vztahy $\sigma_{11} = \sigma_1^2$, $\sigma_{22} = \sigma_2^2$ a $\sigma_{12} = \sigma_{21} = \rho\sigma_1\sigma_2$, kde $\rho = \mathrm{corr}(X_1, X_2)$ je korelace náhodných veličin X_1, X_2, pak hustota náhodného vektoru $(X_1, X_2)'$ řídícího se dvojrozměrným normálním rozdělením je dána vztahem

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) =$$

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \right) \right\}$$

pro $(x_1, x_2) \in R^2$.



Obrázek 11.

Na obrázku 11 je znázorněna hustota dvojrozměrného normálního rozdělení s parametry $\mu_1 = 0$, $\mu_2 = 0$, $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho = -0.4$.

Předpokládáme-li, že se vektor $(X_1, X_2)'$ řídí rozdělením $N(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, pak náhodná veličina X_1 má normální rozdělení $N(\mu_1, \sigma_1^2)$ s hustotou

$$f_1(x_1; \mu_1, \sigma_1^2) = \frac{1}{\sqrt{2\pi}\,\sigma_1} \exp\left\{-\frac{(x_1 - \mu_1)^2}{2\sigma_1^2}\right\}$$

a veličina X_2 má normální rozdělení $N(\mu_2, \sigma_2^2)$ s hustotou

$$f_1(x_2; \mu_2, \sigma_2^2) = \frac{1}{\sqrt{2\pi}\,\sigma_2} \exp\left\{-\frac{(x_2 - \mu_2)^2}{2\sigma_2^2}\right\}.$$

Podmíněné rozdělení veličiny X_1 při daném $X_2 = x_2$ je $N\big(\mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1-\rho^2)\big)$ a podmíněné rozdělení veličiny X_2 při daném $X_1 = x_1$ je $N\big(\mu_2 + \rho\,\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1-\rho^2)\big)$.

Příklad 42.

Předpokládejme, že výška X_1 (udávaná v cm) a hmotnost X_2 (udávaná v kg) dvacetiletého muže se řídí dvojrozměrným normálním rozdělením s parametry $\mu_1 = 178$, $\mu_2 = 80, \sigma_1^2 = 5^2$, $\sigma_2^2 = 5^2, \rho = 0.35$. Jestliže víme, že muž, o kterého se zajímáme, měří 190 cm, s jakou pravděpodobností bude vážit více než 90 kg?

Řešení:

Podmíněné rozdělení váhy X_2 za podmínky $X_1 = 190$ je normální s parametry

$$\mu = 80 + 0.35\frac{5}{5}(190 - 178) = 84.2,$$
$$\sigma^2 = 5^2(1 - 0.35^2) = 21.9375.$$

Odtud

$$P(X_2 > 90|X_1 = 190) = 1 - \Phi\left(\frac{90 - 84.2}{\sqrt{21.9375}}\right) = 1 - \Phi(1.238) = 0.108.$$

Jestliže víme, že muž, o kterého se zajímáme, měří 190 cm, pak bude vážit více než 90 kg s pravděpodobností 0.108 (10.8 %). Všimněte si, že podíl mužů mezi dvacetiletými s váhou vyšší než 90 kg tvoří jen 2.2 %, neboť $1 - \Phi(\frac{90-80}{5}) = 1 - \Phi(2) = 0.0228$. □

Poznámka

Všimněme si, že je-li korelační koeficient $\rho = 0$, pak

$$f(x_1, x_2; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho) = f_1(x_1; \mu_1, \sigma_1^2) \cdot f_2(x_2; \mu_2, \sigma_2^2).$$

Odtud vyplývá, že je-li $\rho = 0$, pak jsou náhodné veličiny nezávislé. *Pro normálně rozdělené náhodné veličiny platí, že jsou nezávislé tehdy a jen tehdy, jestliže* $\rho = 0$.

Řídí-li se vektor $\boldsymbol{X} = (X_1, \ldots, X_n)'$ normálním rozdělením $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} \quad \text{a} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1n} \\ \vdots & \ldots & \vdots \\ \sigma_{n1} & \ldots & \sigma_{nn} \end{pmatrix},$$

pak lineární kombinace $Y = c_1X_1 + \cdots + c_nX_n$ má normální rozdělení se střední hodnotou $c_1\mu_1 + \cdots + c_n\mu_n$ a rozptylem $c_1^2\sigma_{11} + c_2^2\sigma_{22} + \cdots + c_n^2\sigma_{nn} + 2c_1c_2\sigma_{12} + 2c_1c_3\sigma_{13} + \cdots + 2c_1c_n\sigma_{1n} + 2c_2c_3\sigma_{23} + \cdots + 2c_2c_n\sigma_{2n} + \cdots + 2c_{n-1}c_n\sigma_{n-1,n}$, viz (16.1) a (16.2). Odtud například vyplývá, že průměr $\overline{X}$ spočtený z n nezávislých náhodných veličin $X_1, \ldots, X_n$ řídících se $N(\mu, \sigma^2)$ se řídí normálním rozdělením $N(\mu, \sigma^2/n)$.

Příklad 43.

Tomáš a Jan soutěží, kdo vypije rychleji jedno velké pivo. Doba (měřená v sekundách), za kterou Tomáš vypije pivo, je normálně rozdělená náhodná veličina s rozdělením $N(\mu = 12, \sigma^2 = 9)$. Janova doba pití je rovněž normálně rozdělená náhodná veličina s rozdělením $N(\mu = 10, \sigma^2 = 4)$. Najděte pravděpodobnost, s jakou Tomáš porazí Jana.

Řešení:

Předpokládejme, že doba T, po kterou pije pivo Tomáš, a doba J, po kterou ho pije Jan, jsou nezávislé náhodné veličiny. Rozdíl v těchto dobách $T - J$ má normální rozdělení $N(\mu = 12 - 10, \sigma^2 = 9 + 4)$. Odtud $P(T < J) = P(T - J < 0) = F(0) = \Phi(-2/\sqrt{13}) = 0.29$.

Tomáš porazí Jana s pravděpodobností 0.29.

□

Příklad 44.

Astronom si přeje zjistit vzdálenost (ve světelných letech) mezi svou observatoří a vzdálenou hvězdou. Aby omezil vliv náhodných chyb měření, rozhodne se měření vícekrát zopakovat a k odhadu použít aritmetický průměr. Jeho přístroj nemá systematickou chybu a ze zkušenosti ví, že náhodné chyby mají normální rozdělení se směrodatnou odchylkou $\sigma = 2$. Kolik měření má provést, aby si byl rozumně jist, že spočtený průměr se v absolutní hodnotě liší od skutečné vzdálenosti o méně než $1/2$ světelného roku?

Řešení:

Aritmetický průměr n měření má normální rozdělení $N(d, 4/n)$, kde d je skutečná vzdálenost. Je zřejmé, že i kdyby astronom provedl jakékoli množství měření, přesto může nastat situace, kdy se bude jeho průměr lišit od skutečné vzdálenosti o více než $1/2$ světelného roku. Jde o to, aby se tato situace vyskytla jen velmi zřídka, dejme tomu v 5% případů, tj. aby $P(|\overline{X} - d| > 1/2) = 0.05$. Odtud $P(\overline{X} \in (d - 1/2, d + 1/2)) = \Phi(\sqrt{n}/4) - \Phi(-\sqrt{n}/4) = 0.95$, a tedy $\sqrt{n}/4 = 1.96$ čili $n = 61.5 \doteq 62$.

Jestliže se astronom spokojí s 95% spolehlivostí, pak by měl provést 62 měření.

□

18. Centrální limitní věta

V teorii pravděpodobnosti existuje mnoho různých verzí centrální limitní věty. My zde uvedeme jen její nejjednodušší verzi:

Nechť $X_1, \ldots, X_n$ jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou $\mathrm{E}\, X_i = \mu$ a rozptylem $\mathrm{Var}\, X_i = \sigma^2$, $i = 1, \ldots, n$, pro které platí $\mathrm{E}\,|X_i|^3 < \infty$. Pak pro velká n součet těchto veličin $S_n = \sum_{i=1}^n X_i$ má přibližně normální rozdělení $N(n \cdot \mu, n \cdot \sigma^2)$ a aritmetický průměr $\overline{X} = \sum_{i=1}^n X_i/n$ má přibližně normální rozdělení $N(\mu, \sigma^2/n)$.

Z předchozího tvrzení například vyplývá, že opakujeme-li mnohokrát měření určité charakteristiky, pak součet, resp. průměr naměřených hodnot má (za velmi obecných předpokladů) přibližně normální rozdělení bez ohledu na to, jaké rozdělení měla původní měření.

Příklad 45.

Průměrná váha zavazadla cestujícího v turistické 2. třídě na trase Praha – Paříž je 20 kg a standardní odchylka je 7 kg. Průměrná váha zavazadla cestujícího v 1. třídě (business class) je 12.5 kg a směrodatná odchylka je 4 kg. Jestliže je v letadle 12 cestujících v 1. třídě a 50 v 2. třídě, jaká je pravděpodobnost, že celková váha všech zavazadel překročí 1200 kg?

Řešení:

Celková váha zavazadel cestujících v 1. třídě $B = X_1 + \cdots + X_{12}$, kde X_i je váha zavazadla $i-$tého cestujícího v 1. třídě, má přibližně normální rozdělení se střední hodnotou $\mu_B = 12 \cdot 12.5 = 150$ a rozptylem $\sigma_B^2 = 12 \cdot 16 = 192$. Celková váha zavazadel cestujících v 2. třídě $T = Y_1 + \cdots + Y_{50}$, kde Y_j je váha zavazadla $j-$tého cestujícího v 2. třídě, má přibližně normální rozdělení se střední hodnotou $\mu_T = 50 \cdot 20 = 1000$ a rozptylem $\sigma_T^2 = 50 \cdot 49 = 2450$. Celková váha všech zavazadel $B + T$ má přibližně normální rozdělení se střední hodnotou $\mu_B + \mu_T = 1150$ a rozptylem $\sigma_B^2 + \sigma_T^2 = 192 + 2450 = 2642$. Pravděpodobnost

$$P(B + T > 1200) = 1 - \Phi\left(\frac{1200 - 1150}{\sqrt{2642}}\right) = 1 - \Phi(0.9728) = 1 - 0.835 = 0.165.$$

Pravděpodobnost, že celková váha všech zavazadel překročí 1200 kg, je přibližně rovna 0.165.

□

Aproximace binomického rozdělení normálním rozdělením.

Uvažujeme náhodnou veličinu X, která má binomické rozdělení $Bi(n,p)$. Pravděpodobnost, že náhodná veličina X nabude některé hodnoty z intervalu $\langle a,b\rangle$, kde a a b jsou přirozená čísla, lze spočítat přesně pomocí binomického rozdělení následovně:

$$P(X \in \langle a,b\rangle) = \sum_{x=a}^{b} \binom{n}{x} p^x (1-p)^{n-x}.$$

Pokud n je velké, může být předchozí výpočet dosti náročný. Vzhledem k tomu, že náhodná veličina X označuje počet výskytů určitého jevu A v n nezávislých pokusech, lze ji vyjádřit pomocí veličin $Y_1,\ldots,Y_n$:

$$X = Y_1 + Y_2 + \cdots + Y_n,$$

kde Y_i, $i = 1,\ldots,n$ nabývá hodnoty 1, jestliže v i–tém pokusu jev A nastal, a hodnoty 0, jestliže nenastal. Náhodné veličiny $Y_1,\ldots,Y_n$ jsou nezávislé, přičemž mají stejné alternativní rozdělení $A(p)$ se střední hodnotou $\mathrm{E}\,Y_i = p$ a rozptylem $\mathrm{Var}\,Y_i = p(1-p)$. Použijeme-li však centrální limitní větu, pak pro velké n má X přibližně normální rozdělení $N\big(np, np(1-p)\big)$, a tedy

$$P(X \in \langle a,b\rangle) \doteq \Phi\left(\frac{b-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{np(1-p)}}\right).$$

Ještě lepší přiblížení dostaneme, použijeme-li tzv. opravu na spojitost, která zohledňuje fakt, že X má diskrétní a ne spojité rozdělení:

$$P(X \in \langle a,b\rangle) \doteq \Phi\left(\frac{b+0.5-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-0.5-np}{\sqrt{np(1-p)}}\right).$$

Část IV. Náhodný výběr

19. Úvod do matematické statistiky

Hlavním úkolem matematické statistiky je zpracování dat, která vykazují náhodné kolísání. Pro ilustraci uveďme hned na počátku příklad.

Příklad 46.

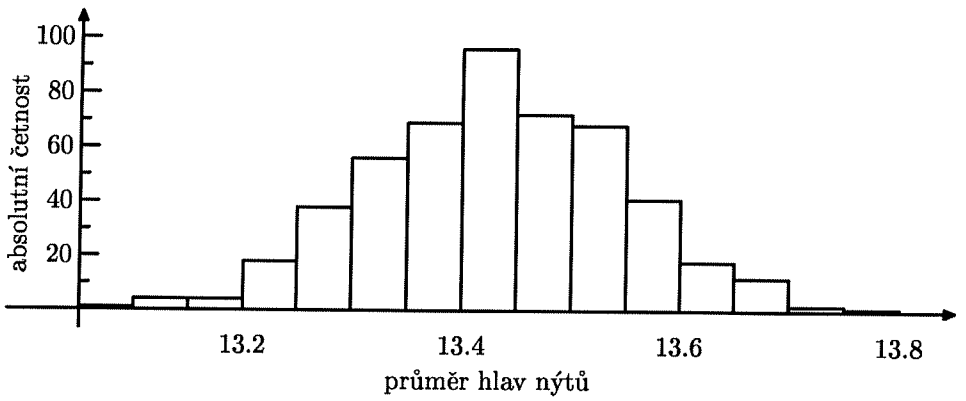
Při kontrole jakosti bylo vybráno 500 nýtů a u každého byl změřen průměr jeho hlavičky. Pro lepší přehlednost byly naměřené výsledky rozděleny do intervalů o délce 0.05 mm, a tak byla získána následující tabulka:

velikost hlavičky nýtu (mm)	počet nýtů - - absolutní četnost	relativní četnost
13.045 – 13.095	1	0.002
13.095 – 13.145	4	0.008
13.145 – 13.195	4	0.008
13.195 – 13.245	18	0.036
13.245 – 13.295	38	0.076
13.295 – 13.345	56	0.112
13.345 – 13.395	69	0.138
13.395 – 13.445	96	0.192
13.445 – 13.495	72	0.144
13.495 – 13.545	68	0.136
13.545 – 13.595	41	0.082
13.595 – 13.645	18	0.036
13.645 – 13.695	12	0.024
13.695 – 13.745	2	0.004
13.745 – 13.795	1	0.002

Tabulka 14.

Naměřeným hodnotám říkáme *pozorování* nebo *vstupní data.* Počtu pozorování, která nabudou dané hodnoty nebo padnou do daných mezí, se říká *absolutní četnost.* Počtu všech pozorování dohromady říkáme *rozsah souboru.* V našem případě rozsah souboru $n = 500$. Součet všech absolutních četností je roven rozsahu souboru. *Relativní četnost* je podíl

absolutní četnosti a rozsahu souboru. Z předchozího je zřejmé, že součet všech relativních četností musí být roven jedné. Relativní četnost se často udává také v procentech, neboť vyjadřuje podíl počtu takových výsledků, které nabudou dané hodnoty (resp. padnou do daných mezí), k celkovému počtu výsledků. Pro lepší názornost je možno zakreslit absolutní (resp. relativní) četnosti do grafu, kterému se říká *histogram*. Uveďme histogram relativních četností pro data z příkladu 46.



Obrázek 12.

Z našeho kontrolního vzorku můžeme spočítat průměrnou velikost hlavičky, najít největší a nejmenší hodnotu hlavičky a různým způsobem vzorek popsat. Tímto způsobem postupuje popisná statistika. Úkolem pracovníků kontrolního oddělení však není popsat zkušební vzorek, nýbrž udělat závěry týkající se celé výroby. Pracovníky může například zajímat, jestli se razicí hlavice neopotřebovala a hlavičky nýtů nejsou systematicky větší než hodnota, na jakou byl stroj nastaven. Dejme tomu, že byl stroj nastaven na hodnotu 13.35 mm, přičemž aritmetický průměr zkušebního vzorku byl 13.426 mm. Tento průměr se samozřejmě liší od hodnoty 13.35 mm. Nyní může vyvstat otázka, jestli je tento rozdíl způsoben systematickou chybou, vzniklou opotřebením razicí hlavice, nebo náhodným kolísáním způsobeným náhodnými chybami výroby, neboť průměr počítaný z jiného vzorku by samozřejmě mohl být jiný. Jindy kontrolního pracovníka může zajímat podíl výrobků z celé výroby vyjádřený v procentech, které jsou mimo toleranční meze dané normou apod. K tomu, abychom mohli dělat na základě vzorku určité závěry týkající se celku, musíme použít teorii pravděpodobnosti. Použití teorie pravděpodobnosti je opodstatněné tím, že výsledky určitých druhů náhodných pokusů se chovají jako náhodné veličiny s určitým typem rozdělení v tom smyslu, že relativní četnosti výsledků padnoucích do množiny A velice dobře souhlasí s pravděpodobnostmi, s jakými se náhodná veličina s určitým rozdělením realizuje uvnitř množiny A. Ukažme si, jak relativní četnosti z příkladu 46 souhlasí

s pravděpodobnostmi, s jakými se náhodná veličina $X \sim N(\mu = 13.426, \sigma^2 = 0.013225)$ realizuje uvnitř daných intervalů.

velikost hlavičky nýtu (mm)	relativní četnost	$P(X \in (a,b))$
13.045 – 13.095	0.002	0.002
13.095 – 13.145	0.008	0.005
13.145 – 13.195	0.008	0.015
13.195 – 13.245	0.036	0.036
13.245 – 13.295	0.076	0.069
13.295 – 13.345	0.112	0.113
13.345 – 13.395	0.138	0.153
13.395 – 13.445	0.192	0.172
13.445 – 13.495	0.144	0.160
13.495 – 13.545	0.136	0.123
13.545 – 13.595	0.082	0.080
13.595 – 13.645	0.036	0.042
13.645 – 13.695	0.024	0.019
13.695 – 13.745	0.004	0.007
13.745 – 13.795	0.002	0.002

Tabulka 15.

Mezi relativními četnostmi a příslušnými pravděpodobnostmi vidíme velmi dobrou shodu. Znamená to, že normální rozdělení je dobrým modelem pro průměr hlaviček nýtů. O tom, jakého rozdělení pro modelování výsledku pokusu použít, se často rozhodujeme na základě analýzy podobných pokusů z dřívějška. Někdy však pro nalezení vhodného rozdělení musíme provést podrobnější rozbor dat. Jedna z možností, jak nalézt vyhovující model, je popsána v článku 32.

20. Náhodný výběr a jeho statistiky

Opakujeme-li n–krát nezávisle po sobě pokus, získáme náhodný výběr $(X_1, \ldots, X_n)$, jehož složky jsou nezávislé. Zachováme-li během pokusu stejné podmínky, pak rozdělení všech složek bude stejné. Takovému vektoru říkáme *náhodný výběr o rozsahu n z určitého rozdělení.* Vybereme-li zkušební vzorek o 500 nýtech (viz příklad 46), pak o vektoru

$(X_1, \ldots, X_{500})$, kde X_i označuje průměr hlavičky $i-$tého nýtu, můžeme například říci, že je to náhodný výběr z normálního rozdělení.

Uvažujme náhodný výběr $(X_1, \ldots, X_n)$. Náhodné veličině $T_n = T(X_1, \ldots, X_n)$, kde T je nějaká funkce n proměnných, říkáme *statistika*. Statistiky jsou určitými charakteristikami výběru. S jejich pomocí se snažíme získat z náhodného výběru nějaké další, pro nás zajímavé informace.

Nejpoužívanější statistikou je *výběrový průměr* definovaný následovně:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Výběrový průměr se zřejmě shoduje s všeobecně známým aritmetickým průměrem. Spočte se tak, že sečteme hodnoty veličin $X_1, \ldots, X_n$ a součet vydělíme rozsahem výběru, nebo jinak řečeno, sečteme příslušná data a součet vydělíme jejich počtem.

Výběrovému průměru se také říká *první výběrový obecný moment.*

Příklad 47.

Student Milan obdržel z matematiky v posledním ročníku střední školy známky:

$$1, \quad 3, \quad 1, \quad 1, \quad 2, \quad 3, \quad 1, \quad 2, \quad 1, \quad 1.$$

Jaká průměrná známka mu vychází na vysvědčení?

Řešení:

Milanovy známky si pro přehled můžeme zapsat do tabulky:

známka	1	2	3	4	5
počet obdržených známek	6	2	2	0	0

Tabulka 16.

Průměr spočteme tak, že příslušné známky vynásobíme počtem, kolikrát se v seznamu vyskytují, tedy absolutní četností, sečteme a vydělíme celkovým počtem získaných známek:

$$\overline{x} = (1 \cdot 6 + 2 \cdot 2 + 3 \cdot 2 + 4 \cdot 0 + 5 \cdot 0)/10 = 1.6\,.$$

Průměr $\overline{x}$ se tedy rovná 1.6. Jestliže by učitel nepřihlédl k Milanově aktivitě během vyučování, dostal by na vysvědčení nejspíš dvojku.

□

Význam výběrového průměru spočívá v tom, že nám charakterizuje, kde jsou data na reálné ose umístěna, nebo, jinak řečeno, „jaké hodnoty nabývají data jako celek". Přesněji říkáme, že výběrový průměr lze použít jako *míru polohy dat.*

Příklad 48.

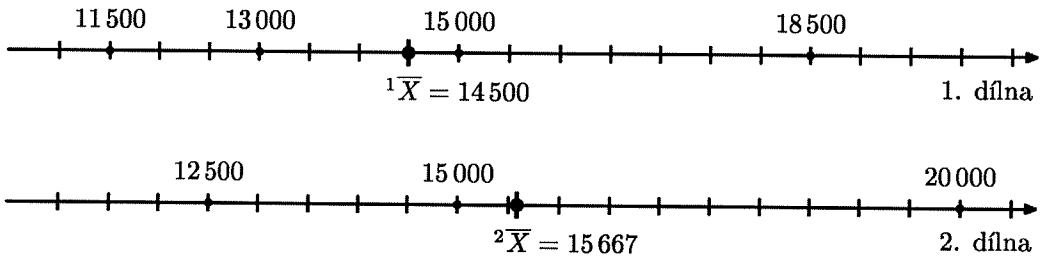
Porovnejte velikosti platů ve dvou sousedních dílnách téhož podniku. Platy pracovníků v první dílně jsou: 13 000 Kč, 11 500 Kč, 18 500 Kč, 15 000 Kč. Platy pracovníků v druhé dílně jsou: 15 000 Kč , 12 500 Kč , 20 000 Kč.

Řešení:

Průměrný plat v první dílně $\overline{x}_1 = (13\,000 + 11\,500 + 18\,500 + 15\,000)/4 = 14\,500$ (Kč).

Průměrný plat v druhé dílně $\overline{x}_2 = (15\,000 + 12\,500 + 20\,000)/3 \doteq 15\,833$ (Kč).

Platy pracovníků v druhé dílně jsou v průměru vyšší, což přirozeně neznamená, že každý pracovník z druhé dílny má vyšší plat než libovolný pracovník z první dílny, ale že „platy v druhé dílně jsou jako celek vyšší než v první dílně", viz obrázek 13. Kdyby se peníze rozdělily stejnoměrně mezi pracovníky, připadlo by na jednoho pracovníka druhé dílny víc.



Obrázek 13.

□

Mezi velmi významné statistiky patří *výběrový rozptyl*, definovaný následovně:

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = (\frac{1}{n}\sum_{i=1}^{n}X_i^2) - \bar{X}^2.$$

Výběrovému rozptylu se také říká *druhý výběrový centrální moment.* Význam výběrového rozptylu spočívá v tom, že ho lze použít jako *míru rozptýlenosti (variability).* Čím větší je výběrový rozptyl, tím více výsledky pokusu kolísají.

Příklad 49.

Porovnejme výsledky písemných prací z matematiky studentů dvou studijních skupin. Výsledky jsou uvedeny v následujících tabulkách:

První skupina

známka	1	2	3	4
počet získaných známek	5	8	2	10

Tabulka 17.

Druhá skupina

známka	1	2	3	4
počet získaných známek	0	10	8	2

Tabulka 18.

Řešení:

Nejprve spočtěme průměry v obou skupinách:

$$\bar{x}_1 = (1 \cdot 5 + 2 \cdot 8 + 3 \cdot 2 + 4 \cdot 10)/25 = 2.68,$$

$$\bar{x}_2 = (1 \cdot 0 + 2 \cdot 10 + 3 \cdot 8 + 4 \cdot 2)/20 = 2.6.$$

Průměry v obou skupinách se téměř shodují, přesto se výsledky zjevně liší. V první skupině je poměrně dost výborných výsledků, ale i poměrně dost nedostatečných výsledků, zatímco ve druhé skupině se výsledky nejčastěji pohybují mezi dvojkou a trojkou. Znamená to, že první skupina je nevyrovnaná, je tam velký podíl velmi dobrých studentů i velký podíl velmi slabých studentů. Druhá skupina je poměrně vyrovnaná. Spočtěme nyní výběrové rozptyly u obou skupin:

$$\sigma_{n,1}^2 = (1^2 \cdot 5 + 2^2 \cdot 8 + 3^2 \cdot 2 + 4^2 \cdot 10)/25 - 2.68^2 = 1.4176,$$

$$\sigma_{n,2}^2 = (1^2 \cdot 0 + 2^2 \cdot 10 + 3^2 \cdot 8 + 4^2 \cdot 2)/20 - 2.6^2 = 0.44.$$

Výběrový rozptyl u druhé skupiny je daleko menší než u první skupiny. Tento výsledek jiným způsobem potvrdil, že výsledky druhé skupiny jsou méně rozptýlené než výsledky první skupiny.

□

Místo výběrového rozptylu se často jako míra rozptýlenosti používá jeho odmocnina, tj. $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, která se nazývá *výběrová směrodatná odchylka*. Na kalkulačkách je často značena jako σ_n (výjimečně σ), proto i my budeme zachovávat toto značení.

Poznámka

V některých knihách a statistických programových systémech se názvu výběrový rozptyl používá pro statistiku $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ a názvu výběrová směrodatná odchylka pro statistiku $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, která je také uvedena na kalkulačkách, většinou s označením σ_{n-1} (někdy též s označením s, které budeme používat i my). O jejím přesném významu si povíme v článku 25. Je zřejmé, že statistika s je také mírou rozptýlenosti a liší se od σ_n pouze v normování. Mezi σ_n^2 a s^2, resp. mezi σ_n a s, jsou vztahy:

$$s^2 = \frac{n}{n-1}\sigma_n^2, \qquad s = \sqrt{\frac{n}{n-1}}\sigma_n.$$

Příklad 50.

Pro údaje z příkladu 48 spočtěte výběrové směrodatné odchylky $\sigma_{n,1}$, $\sigma_{n,2}$.

Řešení:

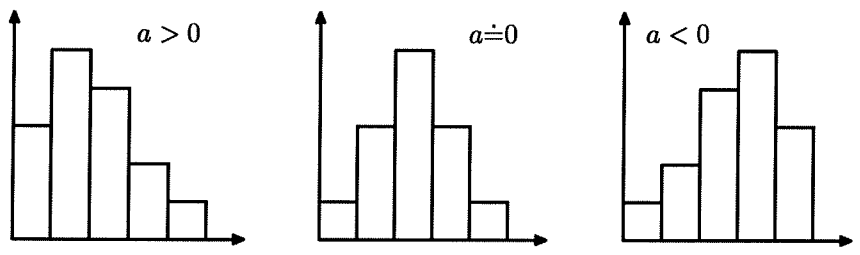
$\sigma_{n,1} = 2622$, a $\sigma_{n,2} = 3118$.

□

Jako míra nesymetrie v rozložení dat může sloužit *výběrový koeficient šikmosti*, zkráceně *šikmost*. Šikmost je dána vztahem:

$$A_3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^{3/2}}.$$

Vztah mezi histogramem vytvořeným z dat a šikmostí je nejlépe možno pochopit z následujícího obrázku.



Obrázek 14.

Výběrový koeficient špičatosti, zkráceně *špičatost*, je definován takto:

$$A_4 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2} - 3.$$

21. Uspořádaný výběr a jeho statistiky

Mějme náhodný výběr $(X_1, \ldots, X_n)$. Nechť $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ je týž náhodný výběr uspořádaný podle velikosti. Pak $X_{(i)}$, $i = 1, \ldots, n$, se nazývá i–tá *pořádková statistika* a vektor $(X_{(1)}, \ldots, X_{(n)})$ *uspořádaný výběr.*

Hodnotě, pro kterou platí, že 50 % pozorování je menších nebo rovných této hodnotě a 50 % pozorování je větších nebo rovných této hodnotě, říkáme *výběrový medián.* Výběrový medián se obvykle vyjadřuje pomocí pořádkových statistik následovně:

$$\widetilde{X} = X_{([n/2]+1)} \qquad \text{pro} \quad n \quad \text{liché},$$

$$\widetilde{X} = \frac{X_{(n/2)} + X_{(n/2+1)}}{2} \qquad \text{pro} \quad n \quad \text{sudé},$$

$[x]$ označuje celou část čísla x.

Hodnotě, pro kterou platí, že 100 p % pozorování je menších nebo rovných této hodnotě a 100 (1-p) % pozorování je větších nebo rovných této hodnotě, říkáme *100 p % dolní výběrový kvantil.* Hodnotě, pro kterou platí, že 100 (1-q) % pozorování je menších nebo rovných této hodnotě a 100 q % pozorování je větších nebo rovných této hodnotě, říkáme *100 q % horní výběrový kvantil.* Takto definované kvantily by však nemusely být ve všech případech jednoznačné. Častěji se proto používá definice pomocí pořádkových statistik. Výběrový 100 p % dolní kvantil se definuje:

$$\widetilde{X}_{100\,p} = X_{([np]+1)} \qquad \text{pokud} \quad np \quad \text{není celé číslo},$$

$$\widetilde{X}_{100\,p} = \frac{X_{(np)} + X_{(np+1)}}{2} \qquad \text{pokud} \quad np \quad \text{je celé číslo}.$$

Obdobně by bylo možno pomocí pořádkových statistik definovat i horní výběrové kvantily.

Z předchozího je zřejmé, že výběrový medián je 50 % výběrovým horním (resp. dolním) kvantilem. Kromě výběrového mediánu se často používá dolní 25 % výběrový kvantil, kterému říkáme *dolní výběrový kvartil*, a horní 25 % výběrový kvantil, kterému říkáme *horní výběrový kvartil.*

Výběrové kvantily jsou rovněž mírou polohy dat. Rozdíly mezi 100 p % horním kvantilem a 100 p % dolním kvantilem slouží jako míra rozptýlenosti dat. Velmi častou mírou rozptýlenosti je *mezikvartilové rozpětí*, tj. rozdíl mezi horním a dolním kvartilem.

Příklad 51.

Pokus spočíval ve zjištění výšek tří náhodně vybraných studentů stavební fakulty. Zjištěné výšky byly: 167, 172 a 170 cm. Určete výběrový medián a výběrový průměr. Dále předpokládejme, že do výběru byl ještě přibrán student Vonásek s výškou 198 cm. Určete znovu medián a průměr.

Řešení:

V prvním výběru první pořádková statistika $X_{(1)}$ nabyla hodnoty 167 cm, druhá $X_{(2)} = 170$ cm a třetí $X_{(3)} = 172$ cm. Výběrový medián je roven 170 cm a výběrový průměr 169.67 cm. V druhém výběru, kde byl přibrán ještě student Vonásek, $X_{(1)} = 167$ cm, $X_{(2)} = 170$ cm, $X_{(3)} = 172$ cm, $X_{(4)} = 198$ cm, výběrový medián je roven 171 cm a výběrový průměr 176.75 cm.

□

Na příkladě 51 je vidět, že odlehlé pozorování 198 cm zvýšilo průměr o 7 cm, zatímco medián se zvýšil o 1 cm. Výběrový medián je tedy mírou polohy, která na rozdíl od průměru potlačuje vliv odlehlých pozorování.

Výběrový medián je speciálním případem takzvaných *useknutých průměrů*, které se spočtou následovně:

$$\overline{X}_a = \frac{1}{n-2a} \sum_{i=a+1}^{n-a} X_{(i)}.$$

Useknuté průměry se používají jako míra polohy v případě, kdy chceme potlačit odlehlá pozorování. Příkladem useknutého průměru je průměr spočtený z náhodného výběru, ze kterého jsme vynechali nejmenší a největší pozorování. Useknutých průměrů se někdy používá při bodování sportovních výkonů.

22. Přehled běžně užívaných popisných statistik

výběrový průměr	$\overline{X} = \frac{1}{n}\sum X_i$	
výběrový medián	$X_{([n/2]+1)}$	pro n liché
	$\frac{X_{(n/2)} + X_{n/2+1)}}{2}$	pro n sudé
modus	hodnota ve výběru s největší četností	
geometrický průměr	$(X_1 \cdot X_2 \cdot \ldots \cdot X_n)^{1/n}$	
výběrový rozptyl	$\sigma_n^2 = \frac{\sum(X_i - \overline{X})^2}{n}$	někdy $s^2 = \frac{\sum(X_i - \overline{X})^2}{n-1}$
výběrová směrodatná odchylka	$\sigma_n = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n}}$	někdy $s = \sqrt{\frac{\sum(X_i - \overline{X})^2}{n-1}}$
standardní chyba	$\sqrt{\frac{\sum(X_i - \overline{X})^2}{n(n-1)}}$	
minimální hodnota ve výběru	$X_{(1)} = \min_{1 \le i \le n} X_i$	
maximální hodnota ve výběru	$X_{(n)} = \max_{1 \le i \le n} X_i$	
dolní výběrový kvartil	$X_{([n/4]+1)}$	pokud n není dělitelno 4
	$\frac{X_{(n/4)} + X_{(n/4+1)}}{2}$	pokud n je dělitelno 4
horní výběrový kvartil	$X_{([3n/4]+1)}$	pokud n není dělitelno 4
	$\frac{X_{(3n/4)} + X_{(3n/4+1)}}{2}$	pokud n je dělitelno 4
rozpětí	$X_{(n)} - X_{(1)}$	
mezikvartilové rozpětí	horní výběrový kvartil − dolní výběrový kvartil	
výběrová šikmost	$\frac{\frac{1}{n}\sum(X_i - \overline{X})^3}{\sigma_n^3}$	někdy $\frac{\frac{1}{n-1}\sum(X_i - \overline{X})^3}{s^3}$
výběrová špičatost	$\frac{\frac{1}{n}\sum(X_i - \overline{X})^4}{\sigma_n^4} - 3$	někdy $\frac{\frac{1}{n-1}\sum(X_i - \overline{X})^4}{s^4} - 3$

Příklad 52.

Pro zjištění vhodnosti nově zaváděné vyučovací metody byl proveden následující pokus. Byla vybrána skupina studentů, takzvaná pokusná, která byla vyučována novou metodou. Na závěr studenti vypracovali test, ve kterém byly jednotlivé příklady obodovány a spočten součet získaných bodů. Čím více bodů student získal, tím byl jeho výsledek lepší. Výsledky testu byly porovnány s výsledky téhož testu, dosaženými jinou kontrolní skupinou, jejíž vyučování probíhalo podle starých metod. Počty získaných bodů jsou uvedeny v následujících tabulkách. Spočtěte základní popisné statistiky u obou skupin.

Pokusná skupina

počet bodů	4	6	7	9	10	11	11.5	12	13	14	15	16	17	18	19	20	22
počet studentů	3	4	2	4	10	5	2	13	4	2	2	4	6	2	2	6	2

Tabulka 19.

Kontrolní skupina

počet bodů	0	3	4	5	6	7	8	9	10	11	12	13	15	18	19	20
počet studentů	2	2	2	7	2	2	6	8	2	6	11	3	4	4	2	2

Tabulka 20.

Řešení:

	Pokusná skupina	Kontrolní skupina
velikost vzorku	73	65
výběrový průměr	12.79	10.18
výběrový medián	12.00	10.00
výběrový modus	12.00	12.00
geometrický průměr	11.95	0.00
výběrový rozptyl	19.44	21.84
výběrová směrodatná odchylka	4.41	4.67
standardní chyba	0.52	0.58
minimální hodnota ve výběru	4.00	0.00
maximální hodnota ve výběru	22.00	20.00

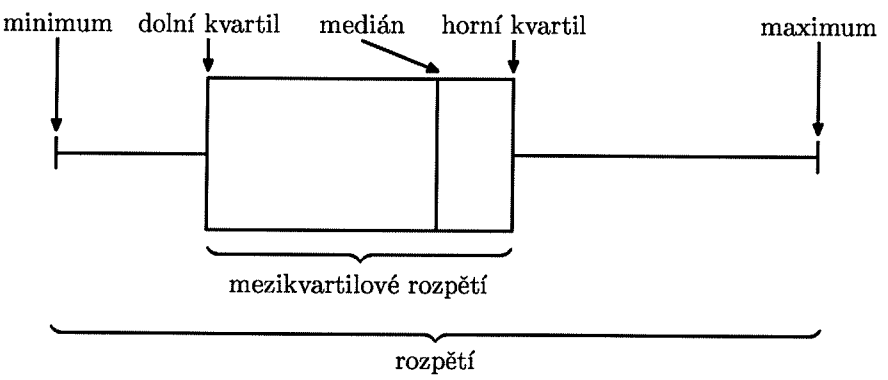
	Pokusná skupina	Kontrolní skupina
dolní výběrový kvartil	10.00	7.00
horní výběrový kvartil	16.00	12.00
rozpětí	18.00	20.00
mezikvartilové rozpětí	6.00	5.00
výběrová šikmost	0.16	0.18
výběrová špičatost	-0.51	-0.22

□

Poznámka

Všimněme si, že téměř všechny míry polohy, tj. průměr, medián, geometrický průměr, minimum, maximum, dolní a horní kvartily jsou pro pokusnou skupinu větší než pro kontrolní skupinu. Pokusná skupina tedy jako celek získala lepší výsledky. Přesnější statistické ohodnocení pokusu bychom získali pomocí testování hypotéz.

Popisné statistiky, o kterých jsme se zmínili, vyjadřují vždy pouze jedinou vlastnost datového souboru. Souhrnnější představu získáme pomocí *krabicového grafu.* Anglický název „box-whisker plot" znamenající „krabička s vousy" dobře vyjadřuje tvar tohoto grafu.

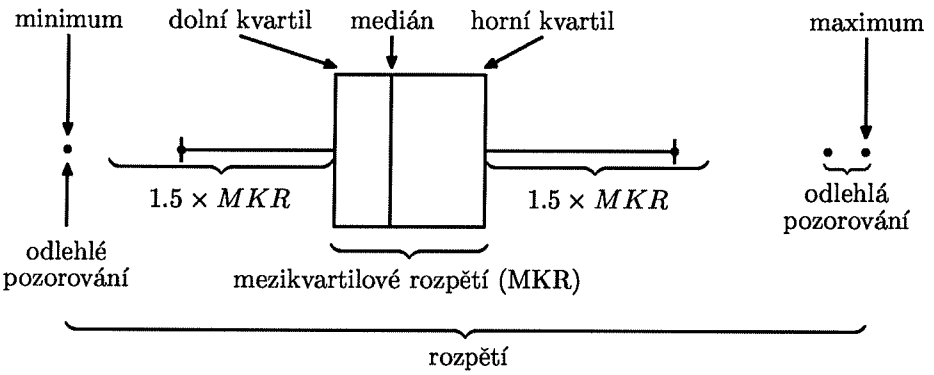


Obrázek 15.

Protože krabicový graf používáme často v případech, kdy zpracováváme menší počet dat, je jeho základní část, tzv. „krabička" vytvořena pomocí statistik, které nejsou citlivé k odlehlým pozorováním, to jest dolním a horním kvartilem. Prostřední čára v krabičce pak odpovídá mediánu. V případě, že v datech nejsou odlehlá pozorování, sahají „vousy", to

znamená postranní čáry, do vzdálenosti minima, respektive maxima. Za odlehlé pozorování se považuje hodnota ležící od okraje krabičky dále než 1.5 násobek mezikvartilového rozpětí. Jestliže je mezi daty například příliš velké, tj. odlehlé pozorování, sahá horní „vous" jen do vzdálenosti největšího pozorování méně vzdáleného od horního okraje krabičky než 1.5 násobek mezikvartilového rozpětí a odlehlé pozorování je znázorněno hvězdičkou. Analogicky se postupuje také pro malá odlehlá pozorování.

Obrázek 15 ukazuje tvar krabicového grafu v případě, že v datech nejsou odlehlá pozorování, zatímco obrázek 16 odpovídá situaci, když se mezi daty vyskytují odlehlé hodnoty.



Obrázek 16.

Krabicový graf poskytuje informaci o poloze, variabilitě a symetrii dat. Lze z něj též ihned zjistit, zda jsou v datech odlehlá pozorování.

Na závěr tohoto článku zdůrazněme, že jakákoliv statistika je rovněž náhodná veličina. Pokud by výsledek pokusů byl vlivem náhody jiný a náhodný výběr by nabyl jiných hodnot, pak by statistika, která je funkcí těchto hodnot, měla jiný výsledek. V některých jednodušších případech je možno rozdělení jednotlivých statistik spočítat. Již v článku 17 jsme hovořili o tom, že v případě výběru $(X_1, \ldots, X_n)$ z normálního rozdělení $N(\mu, \sigma^2)$ má průměr $\overline{X}$ normální rozdělení $N(\mu, \sigma^2/n)$, respektive veličina $(\overline{X} - \mu)\sqrt{n}/\sigma$ má standardní normální rozdělení $N(0,1)$. Neznáme-li rozptyl σ^2, lze ho nahradit odhadem s^2, přičemž statistika $(n-1)\,s^2/\sigma^2$ se řídí χ^2 rozdělením s $n-1$ stupni volnosti. Díky nezávislosti $\overline{X}$ a s se statistika $(\overline{X} - \mu)\sqrt{n}/s$ řídí t rozdělením o $n-1$ stupních volnosti, viz článek 10.

Část V. Teorie odhadu

23. Bodové odhady

Pro lepší pochopení myšlenek teorie odhadu začněme příkladem.

Příklad 53.

Představme si, že v ruce držíme konkrétní minci. Jaká je pravděpodobnost, že na této minci padne líc?

Řešení:

Začněme úvahou. Mince může být vyrobena z nehomogenního materiálu, její tvar může být změněn dlouhodobým používáním a podobně. Těžko tedy můžeme tvrdit, že tato mince je z hlediska pravděpodobnosti symetrická, to je taková, aby obě její strany padaly se stejnou pravděpodobností. Pravděpodobnost p toho, že padne líc na konkrétní minci, je jakási skrytá vlastnost dané mince, kterou nemůžeme explicitně zjistit. Přesto však nejsme úplně bezradní a pomůžeme si tak, že budeme mincí házet a zapisovat si výsledky. Výsledek pokusu pro $n = 10$ hodů mincí můžeme zapsat například R, R, L, L, L, R, L, L, R, L. Tentýž výsledek můžeme zapsat jinak 0, 0, 1, 1, 1, 0, 1, 1, 0, 1, kde v zápise píšeme 0, jestliže padne rub, a 1, jestliže padne líc. Druhý zápis je vlastně realizací náhodného výběru $(X_1, \ldots, X_n)$ o rozsahu $n = 10$ z alternativního rozdělení $A(p)$, viz tabulka 10. Připomeňme, že platí $\mathrm{E}\, X = p$, $\mathrm{Var}\, X = p\,(1-p)$.

Ihned nás napadne, že bude-li rozsah výběru dost velký, bude relativní četnost jevu, že padne líc, blízká pravděpodobnosti p. Relativní četnost se dá v tomto případě vyjádřit jako $\frac{1}{n}\sum X_i$, což je vlastně průměr $\overline{X}$. O tom, že pro velká n je za dosti obecných podmínek průměr $\overline{X}$ blízký střední hodnotě $\mathrm{E}\, X$ (v našem případě $\mathrm{E}\, X = p$), jsme již mluvili. Výběrový průměr $\overline{X}$ má v tomto případě vedle své blízkosti k odhadované pravděpodobnosti p i jiné dobré vlastnosti, o kterých budeme mluvit později. Proto může sloužit jako odhad pravděpodobnosti p.

□

Na příkladu 53 jsme ilustrovali základní myšlenky teorie odhadu. V teorii odhadu vycházíme z toho, že známe typ rozdělení náhodné veličiny X a na základě realizace náhodného výběru $(X_1, \ldots, X_n)$ z tohoto rozdělení chceme odhadnout jeho parametry. Odhadem parametru je funkce $T_n = T(X_1, \ldots, X_n)$ – tedy statistika. Při dané realizaci $X_1 = x_1, \ldots, X_n = x_n$ nabude statistika T_n nějaké konkrétní hodnoty, kterou prohlásíme

za *bodový odhad parametru.* V příkladu 53 jsme věděli, že náhodná veličina X má alternativní rozdělení, a naším úkolem bylo odhadnout jeho parametr p. Za bodový odhad parametru p jsme vzali průměr pozorování $X_1, \ldots, X_n$.

Bodové odhady musí mít určité vlastnosti, podle kterých posuzujeme jejich vhodnost k odhadování neznámého parametru θ. Jednou z podstatných vlastností je jejich nestrannost. Odhad S parametru θ je nestranný tehdy, jestliže $\mathrm{E}\, S = \theta$, což znamená, že se tento odhad pohybuje kolem skutečné hodnoty parametru θ. Ze všech nestranných odhadů je nejlepší ten, který má nejmenší rozptyl, neboť to znamená, že tento odhad nejméně kolísá kolem skutečné hodnoty parametru θ. Najít *nejlepší nestranný odhad* je v některých případech složité, a proto se používají i některé jiné odhady, které mají také dobré vlastnosti. Mezi nejznámější takové odhady patří *maximálně věrohodné odhady.* Myšlenka maximálně věrohodného odhadu spočívá v tom, že za odhad parametru θ vybereme takovou hodnotu tohoto parametru, která je při dané realizaci výběru $X_1 = x_1, \ldots, X_n = x_n$ nejvěrohodnější. Konkrétně to znamená, že se maximalizuje věrohodnostní funkce $L(\theta; x_1, \ldots, x_n)$. Pro diskrétní náhodnou veličinu s rozdělením $p(x;\theta) = P_\theta(X = x)$, $x \in I$, se věrohodnostní funkce spočte ze vztahu:

$$L(\theta; x_1, \ldots, x_n) = p(x_1;\theta) \cdot p(x_2;\theta) \cdot \ldots \cdot p(x_n;\theta).$$

Pro spojitou náhodnou veličinu s hustotou $f(x;\theta)$ se věrohodnostní funkce vypočte následovně:

$$L(\theta; x_1, \ldots, x_n) = f(x_1;\theta) \cdot f(x_2;\theta) \cdot \ldots \cdot f(x_n;\theta).$$

Maximálně věrohodný odhad θ^* pak maximalizuje věrohodnostní funkci; jedná se tedy o takové θ^*, pro něž

$$L(\theta^*; x_1, \ldots, x_n) = \max_\theta L(\theta; x_1, \ldots, x_n).$$

24. Bodové odhady parametrů pro vybrané typy rozdělení

Pro některá rozdělení, se kterými jsme se předchozích kapitolách setkali, uveďme nejlepší nestranné a maximálně věrohodné odhady. Všimněme si, že se v těchto jednoduchých případech odhady vůbec, nebo příliš, neliší.

Rozdělení	nejlepší nestranné odhady	maximálně věrohodné odhady
Alternativní rozdělení $p(x;p) = p^x(1-p)^{1-x},\ x = 0,1$	$\widehat{p} = \overline{X} = \frac{\sum X_i}{n}$	$p^* = \overline{X} = \frac{\sum X_i}{n}$
Poissonovo rozdělení $p(x;\lambda) = \frac{e^{-x}\lambda^x}{x!},\ x = 0,1,2,\ldots$	$\widehat{\lambda} = \overline{X} = \frac{\sum X_i}{n}$	$\lambda^* = \overline{X} = \frac{\sum X_i}{n}$
Normální rozdělení $f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in (-\infty,\infty)$	$\widehat{\mu} = \overline{X} = \frac{\sum X_i}{n}$ $\widehat{\sigma^2} = s^2 = \frac{\sum(X_i-\overline{X})^2}{n-1}$	$\mu^* = \overline{X} = \frac{\sum X_i}{n}$ $\sigma^{2*} = \sigma_n^2 = \frac{\sum(X_i-\overline{X})^2}{n}$
Logaritmicko-normální rozdělení $f(x;\mu,\sigma^2,0) = \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{x} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}},\ x \in (0,\infty)$	$\widehat{\mu} = \frac{\sum \ln X_i}{n}$ $\widehat{\sigma^2} = \frac{\sum(\ln X_i-\widehat{\mu})^2}{n-1}$	$\mu^* = \frac{\sum \ln X_i}{n}$ $\sigma^{2*} = \frac{\sum(\ln X_i-\mu^*)^2}{n}$
Exponenciální rozdělení $f(x;\delta) = \frac{1}{\delta}e^{-\frac{x}{\delta}},\ x \in (0,\infty)$	$\widehat{\delta} = \overline{X} = \frac{\sum X_i}{n}$	$\delta^* = \overline{X} = \frac{\sum X_i}{n}$

Tabulka 21.

Příklad 54.

Stejným přístrojem bylo provedeno 5 nezávislých měření úhlu. Výsledky jsou dány v gradech: 60.054, 60.057, 60.057, 60.056, 60.056. Předpokládejme, že přístroj nemá systematickou chybu. Odhadněte skutečnou velikost úhlu.

Řešení:

Skutečná velikost úhlu je určitá hodnota, kterou však měříme vždy s nějakou chybou. Dlouhodobým pozorováním bylo zjištěno, že chyby měření mívají normální rozdělení. Jestliže naše měření není zatíženo systematickou chybou, budou se změřené hodnoty pohybovat kolem skutečné velikosti úhlu. Skutečná velikost úhlu je tedy střední hodnota, kterou máme odhadnout. Nejlepším nestranným odhadem střední hodnoty μ normálního rozdělení je průměr $\overline{X} = 60.056$. Spočtěme ještě doplňkové hodnoty $s = 0.0012$ a $\sigma_n = 0.0011$. □

Příklad 55.

Při zkouškách 10 přístrojů téhož typu byl registrován okamžik, kdy každý z nich přestal pracovat. Výsledky pokusu jsou uvedeny v tabulce 22.

číslo přístroje	1	2	3	4	5	6	7	8	9	10
doba (hod)	200	350	600	450	400	400	500	350	450	550

Tabulka 22.

Předpokládáme, že doba životnosti přístroje má exponenciální rozdělení. Odhadněte parametr δ.

Řešení:

Nejlepším nestranným i maximálně věrohodným odhadem parametru δ je průměr:

$$\widehat{\delta} = \delta^* = \overline{X} = 425 \text{ (hod)}.$$

□

Kromě metody maximální věrohodnosti se někdy používá *momentová metoda*. Výhodou momentové metody je její jednoduchost. Spočívá v porovnání teoretických a výběrových momentů. S úspěchem se například používá pro odhady parametrů Pearsonova rozdělení typu III. Podrobnější informace o odhadování parametrů rozdělení používaných ve vodohospodářství najde čtenář v Nacházelovi (1986).

Na závěr udejme ještě odhady parametrů dvojrozměrného normálního rozdělení $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

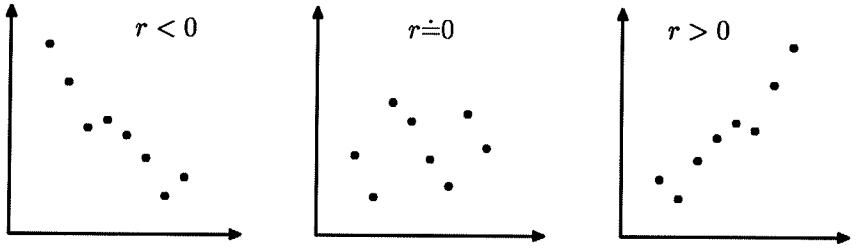
Uvažujme náhodný výběr $((X_1, Y_1)', \ldots, (X_n, Y_n)')$ z tohoto rozdělení. Parametry dvojrozměrného normálního rozdělení se obvykle odhadují metodou maximální věrohodnosti:

$$\mu_1^* = \overline{X}, \quad \mu_2^* = \overline{Y}, \quad \sigma_1^{2*} = \sigma_n^2(x), \quad \sigma_2^{2*} = \sigma_n^2(y) \quad \text{a} \quad \rho^* = r,$$

kde statistika r je definována:

$$r = \frac{\frac{1}{n}\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\left(\frac{1}{n}\sum(X_i - \overline{X})^2\right)\left(\frac{1}{n}\sum(Y_i - \overline{Y})^2\right)}} = \frac{\frac{1}{n}\sum X_i Y_i - \overline{X}\,\overline{Y}}{\sigma_n(x)\sigma_n(y)}$$

a nazývá se *výběrový korelační koeficient*. Jak již víme, je korelační koeficient mírou závislosti mezi veličinami X a Y, a proto je jeho odhad r velmi důležitý. Platí, že $r \in \langle -1, 1 \rangle$. Jestliže je výběrový korelační koeficient r v absolutní hodnotě blízký jedné, pak usuzujeme na velkou závislost mezi X a Y. Naopak, je-li v absolutní hodnotě velmi blízký nule, pak usuzujeme na nezávislost. Znaménko r udává způsob závislosti, viz obrázek 17.



Obrázek 17.

Příklad 56.

Na dvou letištích v Miláně byly zaznamenány maximální roční rychlosti větru v m/s. Uvedené hodnoty vznikly zprůměrováním okamžité rychlosti během 10 minut.

Rok	Malpensa	Linate	Rok	Malpensa	Linate
	x_i	y_i		x_i	y_i
1951	18.468	16.416	1963	11.286	10.260
1952	16.416	10.260	1964	12.825	13.851
1953	14.364	29.241	1965	12.825	11.286
1954	12.825	11.799	1966	15.390	13.851
1955	11.799	27.189	1967	19.494	13.380
1956	14.364	10.260	1968	15.390	13.851
1957	18.468	9.747	1969	14.364	12.312
1958	15.390	11.286	1970	15.390	12.825
1959	13.851	27.702	1971	18.468	15.390
1960	13.338	14.364	1972	15.390	13.338
1961	21.546	19.494	1973	17.955	23.598
1962	16.416	14.877			

Tabulka 23.

Předpokládáme-li, že dvojice maximální roční rychlost větru na letišti v Malpensa a maximální roční rychlost větru na letišti v Linate jsou výběrem z dvojrozměrného normálního rozdělení, odhadněte parametry tohoto rozdělení.

Řešení:

Spočteme-li maximálně věrohodné odhady, obdržíme:

$$\mu_M^* = \overline{X} = 15.4792, \qquad \mu_L^* = \overline{Y} = 15.5023,$$

$$\sigma_M^* = \sigma_n(x) = 2.5612, \qquad \sigma_L^* = \sigma_n(y) = 5.7397,$$

$$\rho^* = r = -0.0072.$$

Výběrový korelační koeficient r je velmi malý. Z toho je možno usuzovat na nezávislost mezi maximální ročními rychlostmi větru na obou letištích.

Spolehlivější závěr týkající se nezávislosti by bylo možno získat použitím testování hypotéz.

□

25. Intervaly spolehlivosti

Bodový odhad parametru θ, získaný z výběru $(X_1, \ldots, X_n)$, nám neříká nic o tom, s jakou přesností byl odhad pořízen. Častěji než bodový odhad nás může zajímat interval, ve kterém s velkou pravděpodobností neznámý parametr θ leží. Zvolíme-li si předem pravděpodobnost

blízkou jedné, to je například 0.90, 0.95, 0.99, obecně $1-\alpha$, pak interval (θ_1, θ_2) takový, že

$$P\big(\theta \in (\theta_1, \theta_2)\big) = 1-\alpha,$$

nazýváme *100 (1 - α)% intervalem spolehlivosti* pro parametr θ.

Pro ilustraci uveďme tvar $100\,(1-\alpha)\%$ intervalu spolehlivosti pro parametr p alternativního rozdělení, počítaný na základě realizace náhodného výběru $(X_1, \ldots, X_n)$ z tohoto rozdělení, za předpokladu, že n je velké:

$$\big(\widehat{p} - u_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p} + u_{\alpha/2} \cdot \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\big).$$

Interval je symetrický kolem bodového odhadu $\widehat{p} = \frac{\sum X_i}{n} = \frac{n_1}{n}$, kterým je relativní četnosti jedniček v realizaci vektoru $(X_1, \ldots, X_n)$. Délka intervalu se zvětšuje s požadovanou spolehlivostí $\big(u_{\alpha/2}$ je $(100 \cdot \alpha/2)\%$ horní kvantil $N(0,1)\big)$ a zmenšuje s počtem dat n.

Nejčastěji používaným $100\,(1-\alpha)\%$ intervalem spolehlivosti je interval pro parametr μ normálního rozdělení $N(\mu, \sigma^2)$ počítaný na základě realizace náhodného výběru $(X_1, \ldots, X_n)$ z tohoto rozdělení:

$$\big(\overline{X} - t_{\alpha/2}[n-1]\,\frac{s}{\sqrt{n}},\ \overline{X} + t_{\alpha/2}[n-1]\,\frac{s}{\sqrt{n}}\big).$$

Všimněme si, že interval spolehlivosti je symetrický kolem výběrového průměru, který je bodovým odhadem parametru μ. Interval je tím větší, čím je větší s, to znamená, když data více kolísají, a je tím menší, čím je větší rozsah výběru n, to jest počet dat. Interval spolehlivosti také závisí na pravděpodobnosti, s jakou žádáme, aby pokrýval neznámý parametr μ. S rostoucím $(1-\alpha)$ roste také horní $100\,\alpha/2\,\%$ kvantil $t-$ rozdělení o $n-1$ stupních volnosti $t_{\alpha/2}[n-1]$ a interval se zvětšuje.

Příklad 57.

Bylo provedeno $n = 5$ měření rychlosti toku v potrubí. Měření dalo následující výsledky: 1.146, 1.210, 1.200, 1.151, 1.259 (m/s). Sestrojte 90%, 95% a 99% interval spolehlivosti pro skutečnou hodnotu průtoku.

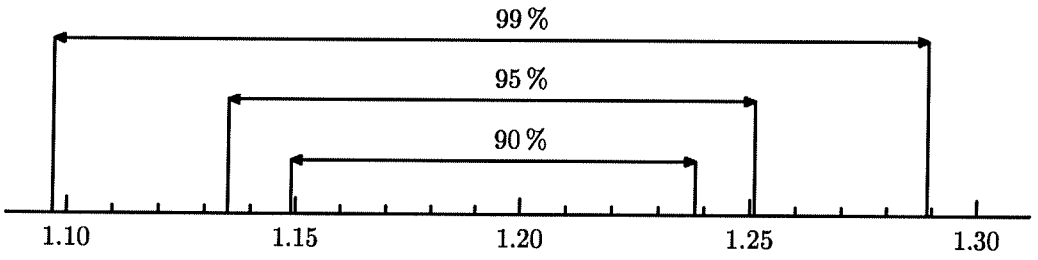
Řešení:

Výsledky měření mívají obvykle normální rozdělení. Skutečná hodnota má zde zřejmě význam střední hodnoty, kolem které se jednotlivá měření pohybují. Spočetli jsme $\overline{x} = 1.1932$ a $s = 0.0465478$. Potřebné kvantily je třeba vyhledat v tabulce t rozdělení: $t_{0.05}[4] = 2.1318$, $t_{0.025}[4] = 2.7764$, $t_{0.005}[4] = 4.6041$. Výsledné intervaly spolehlivosti jsou následující:

90% interval spolehlivosti: (1.149, 1.238),

95% interval spolehlivosti: (1.135, 1.251),

99% interval spolehlivosti: (1.097, 1.289).



Obrázek 18.

□

Někdy nechceme konstruovat $100\,(1-\alpha)\%$ interval spolehlivosti pro střední hodnotu, ale chceme s velkou spolehlivostí předpovídat, v jakém rozmezí bude ležet nějaká budoucí hodnota náhodné veličiny X. V takovém případě mluvíme o *$100\,(1-\alpha)$ % predikčním intervalu.* Pro normální rozdělení $N(\mu, \sigma^2)$ se známými parametry μ a σ^2 padne budoucí hodnota veličiny X s pravděpodobností $1-\alpha$ do intervalu $(\mu - u_{\alpha/2}\,\sigma, \mu + u_{\alpha/2}\,\sigma)$. Jestliže však μ a σ neznáme, musíme je nahradit odhady $\widehat{\mu} = \overline{X}$, $\widehat{\sigma} = s$ a $100\,(1-\alpha)\%$ predikční interval má tvar:

$$\left(\overline{X} - t_{\alpha/2}[n-1]s\sqrt{1+\frac{1}{n}},\, \overline{X} + t_{\alpha/2}[n-1]s\sqrt{1+\frac{1}{n}}\right).$$

Část VI. Testování hypotéz

26. Úvod do testování hypotéz

Statistické hypotézy v našem výkladu jsou určitá tvrzení týkající se rozdělení náhodných veličin. Dané hypotézy mohou tvrdit, že parametr známého rozdělení leží v určité množině nebo že náhodná veličina má rozdělení určitého typu apod. Pro lepší pochopení uveďme příklad.

Příklad 58.

Představme si, že hrajeme hru, o které jsme mluvili v článku 8. Házíme mincí. Jestliže nám padne líc, získáváme 1 Kč, a jestliže padne rub, dáváme 1 Kč spoluhráči. Před začátkem hry náš spoluhráč tvrdí, že hra bude spravedlivá, neboť na minci, s kterou házíme, padají obě dvě strany se stejnou pravděpodobností. My však pojmeme podezření, že jeho tvrzení je lživé a že mince je uzpůsobena tak, aby na ní častěji padal rub. Jak se můžeme o pravdivosti svého podezření přesvědčit?

Řešení:

Své tvrzení si můžeme ověřit tak, že opakovaně házíme mincí. Jestliže podíl případů, ve kterých padne líc, ku celkovému počtu hodů bude mnohem menší než 50 %, naše podezření se silně zvětší.

Jak však rozhodnout, kdy podíl líců ku celkovému počtu hodů se odchyluje od 50 % díky náhodě a kdy je to způsobeno porušením symetrie mince? Jak stanovit mez, podle které budeme rozhodovat, zda je spočtený podíl již tak malý, abychom byli oprávněni tvrzení o symetrii mince napadnout? Matematická statistika se snaží na tyto a podobné otázky kladené v analogických případech odpovědět. Na řešení našeho příkladu si ilustrujme, jak při tom postupuje.

Úlohu z našeho příkladu lze popsat následovně: Náhodná veličina X nabývající hodnoty 1, jestliže padne líc, a hodnoty 0, jestliže padne rub, se řídí alternativním rozdělením s neznámým parametrem p, který označuje pravděpodobnost padnutí líce v jednom hodu, viz příklad 53. Náš spoluhráč tvrdil, že mince je symetrická, tedy $p = 1/2$. Tomuto tvrzení říkáme *nulová hypotéza* a značíme $H_0 : p = 1/2$. Naše podezření však bylo, že $p < 1/2$, to jest, že na minci padá líc méně často než rub. Tomuto tvrzení říkáme *alternativní hypotéza* a značíme $A : p < 1/2$. Abychom prokázali alternativu A nebo jinými slovy zamítli hypotézu H_0, házeli jsme opakovaně mincí. Výsledky házení jsme zapsali do vektoru, přičemž stejně jako v příkladě 53 jsme líc nahradili jedničkou a rub nulou. Získali jsme tak vektor, který může například vypadat následovně: $(0, 0, 1, 1, \ldots, 1, 0)$. Tento vektor můžeme považovat za realizaci náhodného výběru $(X_1, \ldots, X_n)$ z alternativního rozdělení $A(p)$. Hypotézu H_0 zamítáme, jestliže $\sum X_i/n < C$, to znamená, že podíl líců ku celkovému počtu hodů je menší než konstanta C. Konstantě C říkáme *kritická*

hodnota. Nyní vyvstává otázka, jak konstantu C stanovit. Máme zvolit $C = 0.3\,(30\,\%)$ nebo $C = 0.4\,(40\,\%)$ nebo nějaké jiné číslo? I v případě, že zvolíme $C = 0.1\,(10\,\%)$ a mince je symetrická, může se stát, i když velmi zřídka, že podíl líců ku celkovému počtu hodů bude menší než $10\,\%$, to znamená, že platí $\sum X_i/n < C$. Naší snahou je zvolit konstantu C tak, aby v případě platnosti nulové hypotézy pravděpodobnost toho, že nastane situace, kdy $\sum X_i/n < C$, byla malá. Nechceme totiž neoprávněně osočovat spoluhráče. Chybě, které bychom se tímto způsobem dopustili, tj. zamítli nulovou hypotézu H_0, ačkoli je správná, říkáme *chyba prvního druhu.* Dopustit se však můžeme i chyby, že nezamítneme nulovou hypotézu H_0, ačkoli správná není. V tomto případě uděláme *chybu druhého druhu.* Je přirozené požadovat, aby pravděpodobnost obou těchto chyb byla co možná nejmenší. Rozhodujeme-li však na základě již provedené realizace náhodného výběru $(X_1, \ldots, X_n)$, nelze pravděpodobnost obou možných chyb udělat současně tak malé, jak bychom si přáli. Obvykle se trvá jen na požadavku, aby pravděpodobnost chyby prvního druhu byla rovna α, kde $\alpha \in (0, 1)$. Číslu α se říká *hladina významnosti testu* a volí se nejčastěji $\alpha = 0.05$, někdy $\alpha = 0.01$. V našem příkladě bychom se tedy snažili zvolit konstantu C tak, aby za platnosti nulové hypotézy $H_0 : p = 1/2$ platilo $P(\sum X_i/n < C) = \alpha$. Podařilo-li by se nám takovou konstantu C najít, získali bychom zamítací pravidlo, na základě kterého bychom mohli rozhodnout, zda můžeme nulovou hypotézu zamítnout nebo ne.

Pro jednoduché případy byla ve statistice zkonstruována pravidla, podle kterých nulovou hypotézu zamítáme. Použijeme-li těchto pravidel, máme zajištěno, že se dopustíme chyby zamítnutí nulové hypotézy v případě její platnosti pouze s pravděpodobností nejvýše se rovnající hladině významnosti α. Zamítací pravidla jsou konstruována navíc tak, aby při zachování požadavku na pravděpodobnost chyby 1. druhu byla pravděpodobnost chyby 2. druhu co nejmenší. Zamítací pravidlo má obvykle tvar nerovnosti mezi testovou statistikou a kritickou hodnotou. *Nulovou hypotézu H_0 zamítáme, platí-li nerovnost daná zamítacím pravidlem.*

Zopakujeme si tedy celý postup, který se nazývá *test nulové hypotézy H_0 proti alternativě A.* V úloze, kterou chceme řešit, je dána nulová hypotéza H_0, o které pochybujeme, že je pravdivá. Přejeme si naopak prokázat, že pravdivá je jiná hypotéza A – alternativa. Zvolíme hladinu významnosti α, obvykle $\alpha = 0.05$ nebo $\alpha = 0.01$. Na základě realizace náhodného výběru $(X_1, \ldots, X_n)$ pomocí statistického zamítacího pravidla určíme, zda je možno hypotézu H_0 zamítnout nebo ne. V našem příkladě má zamítací pravidlo pro velký rozsah výběru tvar:

$$\frac{1}{n}\sum X_i < \frac{-u_\alpha}{2\sqrt{n}} + 0.5,$$

kde u_α je $100\,\alpha\,\%$ horní kvantil $N(0, 1)$. Pro rozsah výběru $n = 100$ a hladinu významnosti $\alpha = 0.05$ zamítáme hypotézu o symetrii mince, jestliže podíl líců ku celkovému počtu hodů bude menší než 0.418 $(41.8\,\%)$, neboť $u_{0.05} = 1.645$. □

V příkladě 58 jsme testovali hypotézu $H_0 : p = 1/2$ proti alternativě $A : p < 1/2$. Této alternativě se říká *jednostranná alternativa* a testu *jednostranný test.* Budeme-li nestranným soudcem ve shora popsané hře, bude nás spíše zajímat alternativa $A : p \neq 1/2$, to znamená, zda-li ani jeden z hráčů není ve výhodě. Takové alternativě se říká *oboustranná alternativa* a testu *oboustranný test.* Pro oboustrannou alternativu $A : p \neq 1/2$, hladinu významnosti α a velký rozsah výběru n má zamítací pravidlo tvar:

$$|\frac{1}{n}\sum X_i - 0.5| > \frac{u_{\alpha/2}}{2\sqrt{n}},$$

kde $u_{\alpha/2}$ je $100\,(\alpha/2)\%$ horní kvantil $N(0,1)$. Všimněme si, že zamítací pravidlo pro oboustrannou alternativu je jiné než pro jednostrannou alternativu.

Shora uvedené testy jsou speciálním případem testů pro parametr alternativního rozdělení $A(p)$:

1. Nulová hypotéza $H_0 : \quad p = p_0$

   Alternativa $A : \quad p \neq p_0$

   Zamítací pravidlo : $|\frac{1}{n}\sum X_i - p_0| > u_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$

2. Nulová hypotéza $H_0 : \quad p = p_0$

   Alternativa $A : \quad p > p_0$

   Zamítací pravidlo : $\frac{1}{n}\sum X_i > p_0 + u_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$

3. Nulová hypotéza $H_0 : \quad p = p_0$

   Alternativa $A : \quad p < p_0$

   Zamítací pravidlo : $\frac{1}{n}\sum X_i < p_0 - u_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$

Statistika $\frac{1}{n}\sum X_i = \frac{n_1}{n}$ zde odpovídá relativní četnosti výskytu jedniček při realizaci výběru $(X_1, \ldots, X_n)$ z alternativního rozdělení.

Poznámka

V předchozím výkladu jsme uvedli, že obvyklý postup při testování hypotéz spočívá v tom, že nejprve zvolíme hladinu významnosti testu (nejčastěji $\alpha = 0.05$ nebo $\alpha = 0.01$) a poté zjišťujeme, zda hodnota testové statistiky překračuje kritickou hodnotu odpovídající zvolené hladině významnosti. Jestliže k překročení dojde, pak podle zamítacího pravidla je možno nulovou hypotézu zamítnout. Můžeme však použít i jiný myšlenkový postup. Z daného výběru spočítat hodnotu testové statistiky a zjistit, jakou hladinu významnosti bychom museli volit, aby při této hodnotě testové statistiky bylo možno nulovou hypotézu ještě zamítnout. Přesněji řečeno, ke každé spočítané hodnotě testové statistiky existuje jakási hraniční hodnota - jakmile volíme hladinu významnosti větší než tuto hraniční hodnotu, lze ještě nulovou hypotézu zamítnout, jakmile však ji volíme menší, nulovou hypotézu již zamítnout nemůžeme. Hraniční hodnotě s touto vlastností říkáme p - *hodnota.* Jestliže nulovou hypotézu zamítneme právě tehdy, když $p-$ hodnota je menší než 0.05 (resp. 0.01), shoduje se tento postup se zamítacím postupem popsaným v předchozím výkladu, kde $\alpha = 0.05$ (resp. 0.01). Spočtená $p-$ hodnota nám však dává podrobnější informaci o výsledku našeho testu. Zvolíme-li $\alpha = 0.05$ a $p-$ hodnota odpovídající našim datům se rovná 0.049, pohybujeme se na hranici zamítnutí. Je-li $p-$ hodnota spočtená z našich dat velmi malá, například 0.001, je zamítnutí nulové hypotézy velice oprávněné. Čím větší je spočtená $p-$ hodnota, tím je zamítnutí nulové hypotézy méně oprávněné.

27. Jednovýběrová analýza pro normální rozdělení

Uvažujme náhodný výběr $(X_1, \ldots, X_n)$ o rozsahu n z rozdělení $N(\mu, \sigma^2)$. Uveďme některé testy týkající se parametrů μ a σ^2. Nejprve ještě připomeňme, jaké statistiky jsou nejlepšími odhady parametrů μ a σ^2: $\widehat{\mu} = \overline{X}$ a $\widehat{\sigma^2} = s^2 = \frac{1}{n-1}\sum(X_i - \overline{X})^2$. Ve všech uvedených případech α označuje hladinu významnosti testu. Nejprve uveďme testy týkající se parametru μ:

4. Nulová hypotéza $H_0: \quad \mu = \mu_0$

   Alternativa $A: \quad \mu \neq \mu_0$

   Zamítací pravidlo: $t = \dfrac{|\overline{X} - \mu_0|}{s}\sqrt{n} > t_{\alpha/2}[n-1]$

5. Nulová hypotéza $H_0: \quad \mu = \mu_0$

   Alternativa $A: \quad \mu > \mu_0$

   Zamítací pravidlo : $t = \dfrac{\overline{X} - \mu_0}{s}\sqrt{n} > t_\alpha[n-1]$

6. Nulová hypotéza $H_0: \quad \mu = \mu_0$

   Alternativa $A: \quad \mu < \mu_0$

   Zamítací pravidlo : $t = \dfrac{\overline{X} - \mu_0}{s}\sqrt{n} < -t_\alpha[n-1]$

Připomeňme, že $t_p[\nu]$ je $100p\%$ horní kvantil t rozdělení o ν stupních volnosti.

Příklad 59.

Skupina studentů měřila vzdálenost dvou orientačních bodů, které byly vybrány tak, aby byly přesně ve vzdálenosti 100.00 m. Studenti naměřili 26 následujících údajů:

1	2	3	4	5	6	7	8	9
100.02	100.01	99.98	100.09	100.01	100.01	100.01	99.99	100.05
10	11	12	13	14	15	16	17	18
100.03	100.01	99.96	100.02	100.04	100.00	99.98	99.99	100.03
19	20	21	22	23	24	25	26	
100.07	100.01	100.00	99.96	100.01	100.02	100.02	100.00	

Tabulka 24.

Otestujte na hladině významnosti $\alpha = 0.05$, zda přístroj, kterým bylo měření provedeno, nemá systematickou chybu.

Řešení:

Pokud přístroj nemá systematickou chybu, budou naměřené hodnoty kolísat kolem 100.00 m. Z hlediska nevhodnosti přístroje nás zajímá jak systematická chyba, která při měření zkracuje vzdálenost, tak i taková, která ji prodlužuje. To znamená, že volíme oboustrannou alternativu. Za předpokladu, že chyby měření mají normální rozdělení, budeme postupovat podle případu 4.

Hladina významnosti $\alpha = 0.05$.

$$H_0: \quad \mu = \mu_0 = 100\,\mathrm{m}$$
$$A: \quad \mu \neq \mu_0 = 100\,\mathrm{m}$$

Hypotézu H_0 zamítneme ve prospěch alternativy A, jestliže

$$t = \frac{|\overline{X} - \mu_0|}{s}\sqrt{n} > t_{0.025}[n-1].$$

V našem případě $\overline{X} = 100.0123$, $s = 0.029305$ a $n = 26$. Po dosazení do vzorce je hodnota testové statistiky $t = 2.142$ větší než $t_{0.025}[25] = 2.0595$, a proto hypotézu H_0 zamítáme. Závěr tedy zní, že přístroj má systematickou chybu.

Podívejme se však na výsledky testu blíže. Hodnotě testové statistiky $t = 2.142$ odpovídá $p-$ hodnota 0.042. To znamená, že se pohybujeme na hranici zamítnutí nulové hypotézy, viz poznámka v článku 26. Kdybychom na počátku volili hladinu významnosti $\alpha = 0.01$, nulovou hypotézu bychom zamítnout nemohli. Za této situace bychom nejspíš doporučili, dříve než bude přístroj zaslán do opravy, udělat několik dalších kontrolních měření, aby se existence systematické chyby přístroje potvrdila.

□

Příklad 60.

Ochránci přírody pojali podezření, že podnik vypouští do řeky víc škodlivého odpadu, než uvádí ve své zprávě. Ve zprávě je uvedeno, že průměrné množství škodlivého odpadu vypouštěného do řeky je $2.0\,\mu$g/l. Výsledky měření množství škodlivé látky v odebraných vzorcích jsou uvedeny v následující tabulce:

číslo vzorku	1	2	3	4	5	6	7	8	9	10
množství škodlivin (μg/l)	2.2	2.2	1.8	2.1	1.7	1.9	1.7	2.2	2.3	2.0

Tabulka 25.

Podařilo se podniku vyšší znečištění vody dokázat? (Předpokládejme, že množství vypouštěných škodlivin se řídí normálním rozdělením.)

Řešení:

V tomto příkladě nás zajímá, zda obsah škodlivin systematicky nepřekračuje hodnotu $\mu_0 = 2.0\,\mu$g/l. Testujeme tedy nulovou hypotézu $H_0 : \mu = \mu_0$ proti jednostranné alternativě $A : \mu > \mu_0$. Předpokládáme-li, že naměřené obsahy škodlivin jsou realizací náhodného výběru

$(X_1, \ldots, X_n)$ z normálního rozdělení, pak nulovou hypotézu H_0 zamítneme ve prospěch alternativy A, jestliže:

$$t = \frac{\overline{X} - \mu_0}{s} \sqrt{n} > t_\alpha[n-1].$$

V našem případě $\overline{X} = 2.01$, $s = 0.22336$ a $n = 10$. Dosazením do vzorce spočteme hodnotu testové statistiky $t = 0.14158$. Volíme-li hladinu významnosti $\alpha = 0.05$, srovnáváme statistiku t s 5% horním kvantilem t rozdělení o 9 stupních volnosti $t_{0.05}[9] = 1.8331$. Protože platí $t = 0.14158 < t_{0.05}[9] = 1.8331$, nemůžeme nulovou hypotézu zamítnout. Ochráncům přírody se nepodařilo podniku vyšší znečištění vody prokázat.

□

Poznámka

Vraťme se ještě k příkladu 60. Náš závěr neznamená, že podnik neznečišťuje vodu ve větší míře, než uvádí ve své zprávě. Znamená pouze, že se podniku na základě měření nepodařilo vyšší znečištění dokázat. To může být způsobeno například malým rozsahem výběru.

Dále uveďme jeden test týkající se parametru σ^2, resp. σ.

7. Nulová hypotéza $H_0: \quad \sigma^2 = \sigma_0^2 \quad (\text{resp. } \sigma = \sigma_0)$

   Alternativa $A: \quad \sigma^2 > \sigma_0^2 \quad (\text{resp. } \sigma > \sigma_0)$

   Zamítací pravidlo: $\chi^2 = \dfrac{(n-1)\,s^2}{\sigma_0^2} > \chi_\alpha^2[n-1]$

Připomeňme, že $\chi_p^2[\nu]$ je $100p\%$ horní kvantil χ^2 rozdělení o ν stupních volnosti.

Příklad 61.

Při uvedení do provozu byl dávkovač seřízen tak, aby směrodatná odchylka $\sigma = 0.4$. Po čase byla provedena kontrola, zda se provozní parametry dávkovače nezhoršily. Vybraný vzorek o rozsahu $n = 11$ dal následující hodnoty: 50.12, 49.65, 48.85, 50.56, 50.23, 49.13, 49.10, 50.77, 49.34, 49.86, 50.45. Otestujte, zda se přesnost dávkovače nezhoršila. (Předpokládejme, že velikost dávek se řídí normálním rozdělením.)

Řešení:

Zhoršení přesnosti dávkovače znamená zvětšení směrodatné odchylky. Odtud volíme nulovou hypotézu $H_0: \sigma = \sigma_0 = 0.4$ a alternativu $A: \sigma > \sigma_0 = 0.4$. Nejprve spočtěme základní statistiky $\overline{X} = 49.8236$, $s = 0.65584$ a $s^2 = 0.430125$ a poté testovou statistiku $\chi^2 = \frac{0.430125}{0.16} \cdot 10 = 26.88$. Zvolíme-li hladinu významnosti $\alpha = 0.05$, budeme porovnávat hodnotu testové statistiky χ^2

s $\chi^2_{0.05}[10]$. Vzhledem k tomu, že $\chi^2 = 26.88 > \chi^2_{0.05} = 18.307$, musíme nulovou hypotézu zamítnout. Přesnost dávkovače se tedy zhoršila.

□

28. Dvouvýběrová analýza pro normální rozdělení

Uvažujme náhodný výběr $(X_1, \ldots, X_n)$ z rozdělení $N(\mu_1, \sigma_1^2)$ a výběr $(Y_1, \ldots, Y_m)$ z rozdělení $N(\mu_2, \sigma_2^2)$. Nechť náhodné výběry $(X_1, \ldots, X_n)$ a $(Y_1, \ldots, Y_m)$ jsou nezávislé. Označme:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad s^2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$
$$\overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i, \quad s^2(y) = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \overline{Y})^2.$$

Uveďme některé testy týkající se parametrů μ_1, μ_2, σ_1^2, σ_2^2:

8. Nulová hypotéza H_0 : $\sigma_1^2 = \sigma_2^2$

   Alternativa A : $\sigma_1^2 \neq \sigma_2^2$

   Zamítací pravidlo: $F = \dfrac{\max\big(s^2(x), s^2(y)\big)}{\min\big(s^2(x), s^2(y)\big)} > F_{\alpha/2}[\nu_1, \nu_2]$,

   kde $\nu_1 = n-1$ a $\nu_2 = m-1$, jestliže $s^2(x) > s^2(y)$,

   $\nu_1 = m-1$ a $\nu_2 = n-1$, jestliže $s^2(x) < s^2(y)$.

Připomeňme, že $F_{\alpha/2}[\nu_1, \nu_2]$ je $100\,(\alpha/2)\%$ horní kvantil F rozdělení o ν_1, ν_2 stupních volnosti.

9. Nulová hypotéza H_0 : $\sigma_1^2 = \sigma_2^2$

   Alternativa A : $\sigma_1^2 > \sigma_2^2$

   Zamítací pravidlo: $F = \dfrac{s^2(x)}{s^2(y)} > F_\alpha[n-1, m-1]$

10. Nulová hypotéza H_0 : $\mu_1 = \mu_2$

Alternativa A : $\mu_1 \neq \mu_2$

a) Zamítací pravidlo : (za předpokladu $\sigma_1^2 = \sigma_2^2$)

$$t = \frac{|\overline{X} - \overline{Y}|}{\sqrt{(n-1)s^2(x) + (m-1)s^2(y)}} \sqrt{n+m-2} \sqrt{\frac{nm}{n+m}} >$$

$$> t_{\alpha/2}[n+m-2]$$

b) Zamítací pravidlo : (za předpokladu $\sigma_1^2 \neq \sigma_2^2$)

$$t = \frac{|\overline{X} - \overline{Y}|}{\sqrt{\frac{s^2(x)}{n} + \frac{s^2(y)}{m}}} > t_{\alpha/2}[\nu],$$

kde $\frac{1}{\nu} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$ a $c = \frac{s^2(x)/n}{(s^2(x)/n) + (s^2(y)/m)}$.

11. Nulová hypotéza H_0 : $\mu_1 = \mu_2$

Alternativa A : $\mu_1 > \mu_2$

a) Zamítací pravidlo : (za předpokladu $\sigma_1^2 = \sigma_2^2$)

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{(n-1)s^2(x) + (m-1)s^2(y)}} \sqrt{n+m-2} \sqrt{\frac{nm}{n+m}} >$$

$$> t_{\alpha}[n+m-2]$$

b) Zamítací pravidlo : (za předpokladu $\sigma_1^2 \neq \sigma_2^2$)

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s^2(x)}{n} + \frac{s^2(y)}{m}}} > t_{\alpha}[\nu],$$

kde $\frac{1}{\nu} = \frac{c^2}{n-1} + \frac{(1-c)^2}{m-1}$ a $c = \frac{s^2(x)/n}{(s^2(x)/n) + (s^2(y)/m)}$.

Příklad 62.

Ve výrobně betonu bylo vyrobeno a vyzkoušeno 11 krychlí betonu B 15 ze dvou různých záměsí. Otestujte, zda oba dva druhy betonu jsou stejně kvalitní. (Předpokládejme, že pevnost je normálně rozdělená náhodná veličina.)

číslo zkoušky	pevnost (MPa) 1. záměs	pevnost (MPa) 2.záměs
1	18	19
2	19	20
3	19	21
4	21	20
5	22	18
6	20	22
7	19	21
8	21	19
9	22	21
10	18	20
11	19	21

Tabulka 26.

Řešení:

Nechť X označuje krychelnou pevnost betonu 1. záměsi a Y krychelnou pevnost betonu 2. záměsi. Podle předpokladu $X \sim N(\mu_1, \sigma_1^2)$ a $Y \sim N(\mu_2, \sigma_2^2)$. Oba dva druhy betonů jsou stejně kvalitní, jestliže $\mu_1 = \mu_2$. Odtud nulová hypotéza $H_0 : \mu_1 = \mu_2$ a alternativa $A : \mu_1 \neq \mu_2$. Z nezávislých výběrů $(X_1, \ldots, X_{11})$ a $(Y_1, \ldots, Y_{11})$ spočteme: $\overline{X} = 19.8182$, $s(x) = 1.47093$ a $\overline{Y} = 20.1818$, $s(y) = 1.16775$. V případě shodných rozsahů je testová statistika t shodná pro 10 a) i 10 b) a rovná se $t = 0.642161$. Zvolme hladinu významnosti $\alpha = 0.05$. Za předpokladu $\sigma_1^2 = \sigma_2^2$ se pravá strana zamítacího pravidla rovná hornímu kvantilu $t_{0.025}[20] = 2.086$, viz 10 a). Předpokládáme-li $\sigma_1^2 \neq \sigma_2^2$, lze pravou stranu zamítacího pravidla aproximovat horním kvantilem $t_{0.025}[19] = 2.093$, viz 10 b). Zamítací pravidlo není splněno ani v jednom z těchto dvou případů. Nulovou hypotézu tedy nemůžeme zamítnout. Nepodařilo se nám prokázat rozdíl v kvalitě betonu 1. a 2. záměsi.

□

Příklad 63.

Ve dvou cementárnách byla provedena kontrola dávkovačů. Zvážením 13 náhodně vybraných pytlů cementu z první cementárny byly získány hodnoty: 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 (kg). Obdobně zvážením 9 náhodně vybraných pytlů z druhé cementárny byly získány hodnoty: 50.3, 50.7, 48.2, 50.1, 49.9, 51.1, 49.8, 48.9, 50.3 (kg). Otestujte zda oba dávkovače pracují stejně.

Řešení:

Výběry získané z první a druhé cementárny jsou nezávislé. Předpokládáme-li, že hmotnosti pytlů jsou normálně rozdělené, můžeme použít postupu 8. a 10.

Nejprve otestujme, zda jsou jejich rozptyly stejné. Testujeme nulovou hypotézu: $H_0 : \sigma_1^2 = \sigma_2^2$ proti alternativě $A : \sigma_1^2 \neq \sigma_2^2$. Spočtěme testovou statistiku $F = \frac{\max(s^2(x),s^2(y))}{\min(s^2(x),s^2(y))} = \frac{s^2(x)}{s^2(y)} = \frac{4.44231}{0.79194} = 5.609$. Zvolíme-li hladinu významnosti $\alpha = 0.05$, pak $F = 5.609 > F_{0.025}[12, 8] = 4.200$. Nulovou hypotézu musíme zamítnout. Prokázali jsme, že oba dávkovače nejsou stejně přesné.

Dále otestujme, zda jeden z dávkovačů nedává systematicky větší dávky než druhý. Testujeme nulovou hypotézu $H_0 : \mu_1 = \mu_2$ proti alternativě $A : \mu_1 \neq \mu_2$. Hladinu významnosti opět uvažujeme $\alpha = 0.05$. Vzhledem k tomu, že jsme prokázali, že $\sigma_1^2 \neq \sigma_2^2$, musíme použít zamítacího pravidla z 10 b). Statistika $t = \frac{|49.131-49.922|}{\sqrt{(4.442/13)+(0.792/9)}} = 1.206$. Spočtěme $\nu = 17.3$, kritická hodnota $t_{0.025}[17] = 2.1098$ a $t_{0.025}[18] = 2.1009$. Vzhledem k tomu, že testová statistika $t = 1.206$ je menší než obě kritické hodnoty, nemůžeme nulovou hypotézu zamítnout. Nepodařilo se nám prokázat, že jeden dávkovač dává systematicky větší dávky než druhý. Podařilo se nám však prokázat, že dávkovače nejsou stejně přesné. Dávkovač v druhé cementárně je zřejmě přesnější než dávkovač v první cementárně. □

29. Párový test

V případě, že náhodné výběry $(X_1, \ldots, X_n)$ a $(Y_1, \ldots, Y_n)$ nejsou nezávislé, nemůžeme použít předchozí dvouvýběrovou analýzu. Někdy však můžeme s výhodou použít párového testu. Myšlenka párového testu spočívá v tom, že je-li $\big((X_1, Y_1)', \ldots, (X_n, Y_n)'\big)$ výběr z dvourozměrného normálního rozdělení $N\big((\mu_1, \mu_2)', \boldsymbol{\Sigma}\big)$, pak výběr $(Z_1, \ldots, Z_n)$, kde $Z_i = Y_i - X_i$, $i = 1, \ldots, n$, je výběr z normálního rozdělení $N\big(\mu(z) = (\mu_2 - \mu_1), \sigma^2(z)\big)$. Hypotézy týkající se porovnání středních hodnot μ_1 a μ_2 je možno převést na hypotézy týkající se parametru $\mu(z)$ a použít zamítacích pravidel z článku 27.

12.	Nulová hypotéza	$H_0 : \mu_1 = \mu_2$	$\Longleftrightarrow$	$H_0 : \mu(z) = 0$
	Alternativa	$A : \mu_1 \neq \mu_2$	$\Longleftrightarrow$	$A : \mu(z) \neq 0$
13.	Nulová hypotéza	$H_0 : \mu_1 = \mu_2$	$\Longleftrightarrow$	$H_0 : \mu(z) = 0$
	Alternativa	$A : \mu_1 > \mu_2$	$\Longleftrightarrow$	$A : \mu(z) < 0$
14.	Nulová hypotéza	$H_0 : \mu_1 = \mu_2$	$\Longleftrightarrow$	$H_0 : \mu(z) = 0$
	Alternativa	$A : \mu_1 < \mu_2$	$\Longleftrightarrow$	$A : \mu(z) > 0$

Příklad 64.

Nový trenér navrhl novou trénovací metodu, o které však ostatní trenéři tvrdili, že nepovede ke zlepšení výkonu. Aby dokázal, že nová metoda vede ke zlepšení výkonu, převzal trenér družstvo skokanů, u kterých byly zaznamenány jejich poslední výkony, viz tabulka 27. Po třech měsících tréninku podle nové metody byly opět změřeny jejich výkony. Podařilo se trenérovi prokázat vhodnost jeho metody? (Předpokládejme, že výkony mají normální rozdělení.)

Řešení:

Výkon $i-$tého sportovce na počátku trénovacího období označme X_i a výkon téhož sportovce po třech měsících označme Y_i. Předpokládejme, že $(X_i, Y_i)' \sim N\big(\boldsymbol{\mu} = (\mu_1, \mu_2)', \boldsymbol{\Sigma}\big), i = 1, \ldots, 16$.

Trenér chce prokázat, že jeho trénovací metoda vede obecně ke zlepšení výkonu. Samozřejmě to neznamená, že výkon každého sportovce se zlepší, ale že se zvedne hodnota, kolem které se délky skoků pohybují, tedy, že platí $A : \mu_1 < \mu_2$. Náhodné veličiny X_i a Y_i jsou závislé, neboť například obě závisí na fyzické zdatnosti $i-$tého sportovce. Budeme-li měřit výkony sportovce s větší fyzickou zdatností, budou obě hodnoty X_i a Y_i větší apod. Použití párového testu je zde nutné.

číslo skokana	délka skoku na počátku trénovacího období (cm)	délka skoku po ukončení trénovacího období (cm)	rozdíl ve výkonu (cm)
1	737	740	3
2	725	724	−1
3	741	750	9
4	742	741	−1
5	727	735	8
6	731	729	−2
7	742	741	−1
8	725	729	4
9	725	729	4
10	731	728	−3
11	731	733	2
12	727	733	6
13	727	728	1
14	727	730	3
15	727	726	−1
16	739	742	3

Tabulka 27.

Zaveďme nové náhodné veličiny $Z_i = Y_i - X_i$, $i = 1, \ldots, 16$ a testujme nulovou hypotézu $H_0 : \mu(z) = 0$ proti alternativě $A : \mu(z) > 0$. Zvolíme-li $\alpha = 0.05$, pak zamítací pravidlo má tvar:

$$t = \frac{\overline{Z}}{s(z)}\sqrt{n} > t_{0.05}[n-1],$$

kde $t_{0.05}[n-1]$ je 5% horní kvantil t rozdělení. Spočtěme $\overline{Z} = 2.125$, $s(z) = 3.55668$, $n = 16$ a odtud dosazením do vzorce spočtěme hodnotu testové statistiky $t = 2.3899$. Hodnota testové statistiky je větší než $t_{0.05}[15] = 1.7531$, a proto se nulová hypotéza zamítá. Trenérovi se podařilo prokázat vhodnost jeho metody.

□

Poznámka

Zdůrazněme znovu rozdíl mezi párovým testem, viz článek 29, a dvouvýběrovým testem, viz článek 28. Dvouvýběrový test můžeme použít jen v případě, že skutečně máme zajištěnu nezávislost všech veličin $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Užijeme-li dvouvýběrový test v situaci, pro kterou je nezbytný test párový, vede to zpravidla k nesmyslným výsledkům. Naproti tomu v případě $m = n$ není hrubou chybou použít párový test i v situaci, pro kterou je vhodnější dvouvýběrový test. Dojde jen k méně efektivnímu zpracování informace obsažené v datech.

30. Analýza rozptylu – jednoduché třídění

Uvažujme následující zobecnění dvouvýběrové analýzy pro normální rozdělení. Máme m nezávislých výběrů $(X_{11}, \ldots, X_{1n_1}), (X_{21}, \ldots, X_{2n_2}), \ldots, (X_{m1}, \ldots, X_{mn_m})$ z normálních rozdělení $N(\mu_1, \sigma^2)$, $N(\mu_2, \sigma^2), \ldots, N(\mu_m, \sigma^2)$. Důležitou otázkou v aplikacích je, zda jsou střední hodnoty $\mu_1, \ldots, \mu_m$ shodné, případně které se liší. Rozptyly se předpokládají stejné – rovné neznámému parametru σ^2.

15. Nulová hypotéza H_0 : $\mu_1 = \mu_2 = \ldots = \mu_m$

Alternativa A : existuje $i \neq j$ tak, že $\mu_i \neq \mu_j$

Zamítací pravidlo: $F = \dfrac{(N-m)S_A}{(m-1)S_e} > F_\alpha[m-1, N-m],$

kde $N = \sum_{k=1}^{m} n_k$ je součet rozsahů všech souborů. Statistika S_A se nazývá *meziúrovňový součet čtverců* a měří, jak daleko jsou průměry jednotlivých výběrů od celkového průměru všech pozorování. Veličina S_e se nazývá *residuální součet čtverců*, přičemž $S_e/(N-m)$ je odhadem neznámého rozptylu σ^2. Nulová hypotéza je zamítnuta tehdy, jestliže se průměry jednotlivých výběrů liší výrazně od celkového průměru, přičemž je třeba vzít v úvahu, zda případné odchylky nejsou způsobeny jen vlastním kolísáním dat, které vyjadřuje jejich rozptyl σ^2.

Statistiky S_A a S_e můžeme vypočítat na základě dále uvedeného postupu. Z každého souboru $(X_{i1}, \ldots, X_{in_i})$ vypočteme statistiky

$$\begin{aligned} X_{i\cdot} &= \sum_{j=1}^{n_i} X_{ij} && \text{– součet hodnot,} \\ \overline{X}_{i\cdot} &= \frac{1}{n_i} X_{i\cdot} && \text{– výběrový průměr,} \\ \sum_{j=1}^{n_i} X_{ij}^2 & && \text{– součet druhých mocnin,} \end{aligned}$$

a pro všechna data

$$\begin{aligned} X_{\cdot\cdot} &= \sum_{i=1}^{m} X_{i\cdot} && \text{– celkový součet hodnot všech souborů,} \\ \overline{X}_{\cdot\cdot} &= \frac{1}{N} X_{\cdot\cdot} && \text{– výběrový průměr hodnot všech souborů.} \end{aligned}$$

Dále definujme

$$\begin{aligned} S_A &= \sum_{i=1}^{m} n_i \left(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot}\right)^2 = \left(\sum_{i=1}^{m} \frac{1}{n_i} X_{i\cdot}^2\right) - \frac{1}{N} X_{\cdot\cdot}^2, \\ S_T &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{\cdot\cdot}\right)^2 = \left(\sum_{i=1}^{m} \sum_{j=1}^{n_i} X_{ij}^2\right) - \frac{1}{N} X_{\cdot\cdot}^2, \\ S_e &= S_T - S_A. \end{aligned}$$

Poznámka

Uveďme především, že není správné místo výše uvedeného zamítacího pravidla použít dvouvýběrovou analýzu pro otestování nulových hypotéz $H_0^{ij} : \mu_i = \mu_j$ pro všechny možné kombinace $i \neq j$. Kumulací více testů se totiž změní hladina významnosti.

Příklad 65.

V pěti výrobnách oceli téže třídy byly na celkem 29 vzorcích změřeny pevnosti v kluzu. Otestujte, zda jednotlivé výrobny produkují stejně kvalitní ocel.

Řešení:

Předpokládejme, že veličina X_i označující pevnost v kluzu oceli z i–té výrobny $(i = 1, \ldots, 5)$ má normální rozdělení s parametry μ_i a σ^2. Stejná kvalita oceli zřejmě znamená, že platí $\mu_1 = \cdots = \mu_5$.

označení výrobny	zjištěné pevnosti (MPa)	$\overline{X}_{i\cdot}$	$X_{i\cdot}$	$\sum_{i=1}^{n_i} X_{ij}^2$
I	326, 334, 340, 344, 326, 338	334.6667	2008	672 288
II	350, 348, 356, 343, 351, 348	349.3333	2096	732 294
III	328, 319, 309, 333, 330, 342	326.8333	1961	641 579
IV	339, 346, 320, 344, 358, 352	343.1667	2059	707 441
V	309, 302, 321, 336, 312,	316	1580	499 966

Tabulka 28.

Stejná kvalita oceli zřejmě znamená, že platí $\mu_1 = \cdots = \mu_5$. Z nezávislých výběrů $(X_{i1}, \ldots, X_{in_i})$ spočteme potřebné statistiky, viz pravá část tabulky 28. Celkový součet $X_{\cdot\cdot} = 9\,704$ a průměr všech dat $\overline{X}_{\cdot\cdot} = 334.6$. Odtud $S_A = 3\,834.49$, $S_T = 6\,408.83$, $S_e = 2\,574.34$. Testová statistika $F = 8.937$ je větší než $F_{0.05}[4, 24] = 2.776$ i než $F_{0.01}[4, 24] = 4.218$. Nulovou hypotézu tedy zamítáme na hladině $\alpha = 0.05$ i na hladině $\alpha = 0.01$. Podařilo se prokázat významný rozdíl v kvalitě oceli z různých výroben. □

Zamítneme-li nulovou hypotézu $H_0 : \mu_1 = \cdots = \mu_m$, je přirozené se ptát, mezi kterými hodnotami je významný rozdíl. Podle takzvaného Scheffeho kritéria se významně liší ty střední hodnoty μ_i a μ_j, pro které

$$F^{ij} = |\overline{X}_{i\cdot} - \overline{X}_{j\cdot}| > \sqrt{\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \frac{(m-1)S_e}{(N-m)} F_\alpha[m-1, N-m]}.$$

Příklad 66.

Označte dvojice výroben z příkladu 65, které produkují ocel různé kvality.

Řešení:

Veličiny F^{ij} můžeme zapsat do tabulky:

F^{ij}	1	2	3	4	5
1	-	14.7	7.8	8.5	18.7
2	-	-	22.5	6.2	33.3
3	-	-	-	16.3	10.8
4	-	-	-	-	27.2

Tabulka 29.

Pravá strana zamítacího pravidla je pro $\alpha = 0.05$ rovna 19.94 ($i < 5, j < 5$) a 20.91 ($i = 5$ nebo $j = 5$). Významně se liší dvojice výroben (II, III), (II, V), (IV, V). □

31. Testy nulovosti korelačního koeficientu

V aplikacích nás často zajímá, zda jsou veličiny X a Y závislé nebo nezávislé. Řídí-li se náhodný vektor $(X,Y)'$ dvourozměrným normálním rozdělením $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, kde

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

pak jsou veličiny X a Y nezávislé pravě tehdy, je-li korelační koeficient $\rho = 0$, viz poznámka v článku 17. Testování nulovosti korelačního koeficientu je tedy velmi důležité.

Uvažujme náhodný výběr $(X_1,Y_1)', \ldots, (X_n,Y_n)'$ z dvourozměrného normálního rozdělení $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Nechť r označuje výběrový korelační koeficient, definovaný v článku 24. Uveďme testy nulovosti korelačního koeficientu ρ proti různým alternativám:

16. Nulová hypotéza $H_0: \quad \rho = 0$

Alternativa $A: \quad \rho \neq 0$

Zamítací pravidlo: $t = \dfrac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} > t_{\alpha/2}[n-2]$

17. Nulová hypotéza $H_0: \quad \rho = 0$

Alternativa $A: \quad \rho > 0$

Zamítací pravidlo: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} > t_{\alpha}[n-2]$

18. Nulová hypotéza $H_0: \quad \rho = 0$

Alternativa $A: \quad \rho < 0$

Zamítací pravidlo: $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} < -t_{\alpha}[n-2]$

Příklad 67.

V tabulce 30 jsou uvedeny hodnoty X_i - průměrné roční průtoky Labe v Ústí nad Labem a hodnoty Y_i - roční srážkové úhrny v povodí Labe měřené v letech 1941 – 1956. Předpokládejme, že $(X_i,Y_i)'$, $i = 1,\ldots,16$, tvoří náhodný výběr z dvourozměrného normálního rozdělení. Otestujte na hladině významnosti $\alpha = 0.05$ nulovou hypotézu $H_0: \rho = 0$ proti alternativě $A: \rho > 0$.

rok	průtok (m^3/s)	srážkový úhrn (mm)
1941	699.0	844
1942	314.8	551
1943	155.9	479
1944	387.6	767
1945	314.6	742
1946	316.8	729
1947	254.7	568
1948	338.9	666
1949	226.1	702
1950	187.4	667
1951	209.6	562
1952	240.9	671
1953	217.9	503
1954	221.2	729
1955	363.6	708
1956	342.5	706

Tabulka 30.

Řešení:

Výběrový korelační koeficient $r = 0.677$. Spočtěme hodnotu testové statistiky $t = 3.442$ a porovnejme ji s 5% horním kvantilem t rozdělení o 14 stupních volnosti $t_{0.05}[14] = 1.7613$, viz 17. Vzhledem k tomu, že $t = 3.442 > t_{0.05}[14] = 1.7613$, zamítáme nulovou hypotézu ve prospěch alternativy. Tento výsledek potvrzuje skutečnost, kterou jsme tušili, že „s růstem srážkové činnosti se zvětšují i průměrné roční průtoky".

□

32. Test χ^2 dobré shody

Vhodnost nebo nevhodnost použití určitého rozdělení jako modelu pro napozorovaná data je možno posoudit tak, že pro libovolnou množinu A porovnáváme relativní četnost, se kterou padnou data do této množiny, a pravděpodobnost, s jakou se náhodná veličina s rozdělením, které chceme pro modelování použít, realizuje uvnitř množiny A, viz článek 19. Je zřejmé, že nemůžeme toto porovnání provádět pro každou množinu A. Většinou zvolíme nějaký systém disjunktních množin $A_1, \ldots, A_l$ (říká se jim někdy třídy), které pokrývají množinu možných hodnot teoretického rozdělení, o kterém chceme rozhodnout, zda-li je to dobrý model nebo ne. Nechť pro $i = 1, \ldots, l$, značí n_i absolutní četnost množiny A_i, to znamená počet dat, které

padnou do množiny A_i, a podíl n_i/n, kde n je rozsah výběru, značí relativní četnost množiny A_i. Čím je shoda mezi relativními četnostmi n_i/n, $i = 1, \ldots, l$, a pravděpodobnostmi $p_i = P(X \in A_i)$, $i = 1, \ldots, l$, větší, tím je vybraný model vhodnější. Dobrá shoda mezi relativními četnostmi n_i/n a pravděpodobnostmi p_i, $i = 1, \ldots, l$, nastává právě tehdy, jestliže je dobrá shoda mezi *skutečnými (empirickými) absolutními četnostmi* n_i, $i = 1, \ldots, l$, a takzvanými *teoretickými četnostmi* np_i, $i = 1, \ldots, l$. Jednou z možností, jak tuto shodu měřit, je použít statistiku χ^2:

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - np_i)^2}{np_i}. \tag{32.1}$$

Je zřejmé, že čím se skutečné četnosti n_i a teoretické četnosti np_i více shodují, tím je statistika χ^2 menší, a naopak, čím se více liší, tím je statistika χ^2 větší. Na této myšlence je založen hojně užívaný test χ^2 dobré shody. Platí-li totiž, že pozorování jsou skutečně realizací náhodného výběru z rozdělení, pro které platí $p_i = P(X \in A_i)$, $i = 1, \ldots, l$, pak je pro velká n statistika (32.1) rozdělena přibližně podle χ^2 rozdělení s $l - 1$ stupni volnosti. Hypotézu H_0, že data pocházejí z daného rozdělení zamítneme, jestliže $\chi^2 > \chi^2_\alpha[l-1]$, kde $\chi^2_\alpha[l-1]$ je $100\alpha\,\%$ horní kvantil χ^2 rozdělení o $l - 1$ stupních volnosti. Pro praktické použití testu se doporučuje, aby teoretické četnosti všech tříd nebyly menší než 5.

Příklad 68.

V tenké vrstvě roztoku zlata se registroval počet částic zlata, které se dostaly do zorného pole mikroskopu. Pozorování se prováděla pravidelně vždy po uplynutí stejně dlouhého časového intervalu. Výsledky jsou uvedeny v tabulce 31.

počet částic	0	1	2	3	4	5	6	7
absolutní četnost	112	168	130	68	32	5	1	1

Tabulka 31.

Ověřte pomocí testu χ^2 dobré shody, zda data jsou realizací náhodného výběru z Poissonova rozdělení s parametrem $\lambda = 1.5$.

Řešení:

Náhodná veličina řídící se Poissonovým rozdělením nabývá hodnot $0, 1, 2, \ldots$. Obor hodnot rozdělme do šesti tříd - $\{0\}$, $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\,\text{a více}\}$. Rozsah výběru $n = 517$. Pro jednotlivé třídy porovnejme skutečné četnosti n_i a teoretické četnosti np_i, kde $p_i = P(X = i) = \frac{e^{-1.5}(1.5)^i}{i!}$, $i = 0, \ldots, 4$, a $p_5 = P(X \geq 5) = 1 - P(X \leq 4) = 1 - \sum_{i=0}^{4} p_i$.

počet částic	skutečná četnost n_i	teoretická četnost np_i	$\frac{(n_i-np_i)^2}{np_i}$
0	112	115.358	0.097766
1	168	173.037	0.146649
2	130	129.778	0.000379
3	68	64.889	0.149148
4	32	24.333	2.415484
5 a více	7	9.607	0.705928

Tabulka 32.

Nulová hypotéza H_0 tvrdí, že množství zlata v zorném poli mikroskopu je výběr z Poissonova rozdělení $Po(\lambda = 1.5)$. Abychom mohli o platnosti hypotézy H_0 rozhodnout, je třeba spočítat hodnotu statistiky χ^2, která se rovná součtu posledního sloupce v tabulce 32, tj. $\chi^2 = 3.51535$. Srovnáme-li hodnotu $\chi^2 = 3.51535$ s 5 % horním kvantilem χ^2 rozdělení o 5 stupních volnosti $\chi^2_{0.05}[5] = 11.07$, vidíme ihned, že hodnota testové statistiky χ^2 je daleko menší než 5 % horní kvantil χ^2 rozdělení. Navíc odpovídající $p-$hodnota je rovna 0.621. Odtud vidíme, že tvrzení o vhodnosti Poissonova rozdělení s parametrem $\lambda = 1.5$ pro modelování dat nemůžeme zamítnout. Naopak, dokonce se zdá, že dané Poissonovo rozdělení je velmi vhodným modelem pro naše data. □

Ve většině praktických případů chceme ověřit, zda se dají data modelovat pomocí určitého rozdělení, aniž známe předem přesné parametry tohoto rozdělení. Například chceme ověřit, zda pro modelování dat lze použít normálního rozdělení apod. V těchto případech je třeba v testu χ^2 dobré shody použít místo konkrétních hodnot parametrů jejich odhady. Odhad parametrů se provádí modifikovanou metodou minimálního χ^2, která je popsána např. v Likešovi a Machkovi (1983). Při testování, zda data pocházejí z daného rozdělení, srovnáváme pak statistiku χ^2 s kvantilem χ^2 rozdělení o $l-k-1$ stupních volnosti, kde l je počet tříd a k je počet odhadovaných parametrů. Na příkladě si ukážeme použití této metody pro normální a logaritmicko-normální rozdělení.

Příklad 69.

Na dálnici v době od 16:05 do 16:15 byly měřeny časové odstupy v sekundách mezi jednotlivými vozidly. Naměřené hodnoty jsou zaznamenány v tabulce 33.

časové odstupy	$\langle 0,1)$	$\langle 1,2)$	$\langle 2,3)$	$\langle 3,4)$	$\langle 4,5)$	$\langle 5,6)$	$\langle 6,7)$	$\langle 7,8)$
četnost	3	28	20	16	21	14	10	4
časové odstupy	$\langle 8,9)$	$\langle 9,10)$	$\langle 10,12)$	$\langle 12,14)$	$\langle 14,16)$	$\langle 16,18)$	$\langle 18,20)$	$\langle 20,25)$
četnost	6	2	5	1	1	0	0	1

Tabulka 33.

Najděte vhodný model.

Řešení:

V tomto případě neznáme konkrétní hodnoty náhodného výběru, známe jen počty dat, které leží v daných intervalech - takzvané *skupinové rozdělení četností*. Nejprve otestujme, zda výběr pochází z normálního rozdělení. Modifikovaná metoda minimálního χ^2 vede pro parametry μ a σ^2 normálního rozdělení na odhady:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{l} n_i\,\xi_i, \quad \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{l} n_i\,(\xi_i - \widehat{\mu})^2,$$

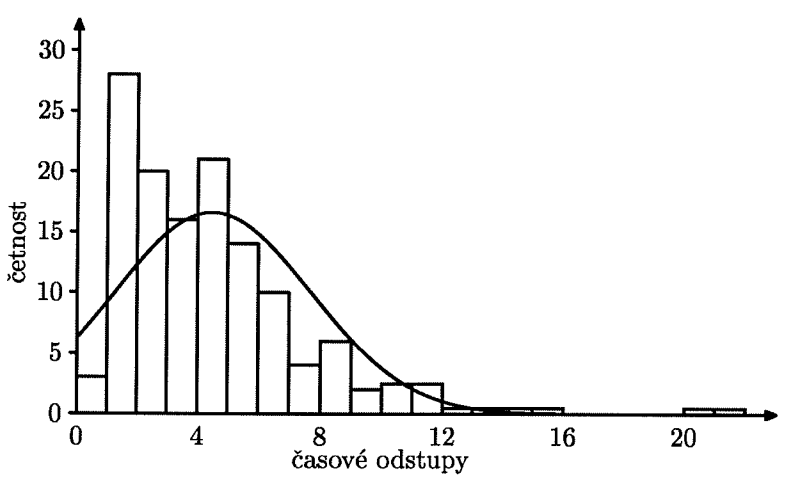
kde ξ_i je střed $i-$té třídy.

ξ_i	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	11	13	15	22.5
n_i	3	28	20	16	21	14	10	4	6	2	5	1	1	1

Tabulka 34.

Použijeme-li předchozí vztahy, dostáváme $\widehat{\mu} = 4.48106$ a $\widehat{\sigma} = 3.197659$.

Obor hodnot normálního rozdělení R^1 rozdělme do tříd $(-\infty, 0)$, $\langle 0,1)$, $\langle 1,2)$, ..., $\langle 8,9)$, $\langle 9,\infty)$. Pro každou třídu $\langle r_i, r_{i+1})$ spočteme teoretickou četnost $n\,p_i$, kde $p_i = P\big(X \in \langle r_i, r_{i+1})\big)$ a $X \sim N(\widehat{\mu}, \widehat{\sigma^2})$. Odtud $p_i = F(r_{i+1}) - F(r_i) = \Phi\left(\frac{r_{i+1}-\widehat{\mu}}{\widehat{\sigma}}\right) - \Phi\left(\frac{r_i-\widehat{\mu}}{\widehat{\sigma}}\right)$. Porovnejme skutečné i teoretické četnosti.



Obrázek 19.

časový odstup (s)	skutečná četnost n_i	teoretická četnost np_i	$\frac{(n_i-np_i)^2}{np_i}$
$(-\infty, 0)$	0	10.633	10.6331
$\langle 0, 1)$	3	7.604	2.7875
$\langle 1, 2)$	28	10.658	28.2155
$\langle 2, 3)$	20	13.559	3.0601
$\langle 3, 4)$	16	15.603	0.0077
$\langle 4, 5)$	21	16.401	1.2894
$\langle 5, 6)$	14	15.596	0.1633
$\langle 6, 7)$	10	13.459	0.8891
$\langle 7, 8)$	4	10.542	4.0593
$\langle 8, 9)$	6	7.493	0.2975
$\langle 9, \infty)$	10	10.401	0.0155

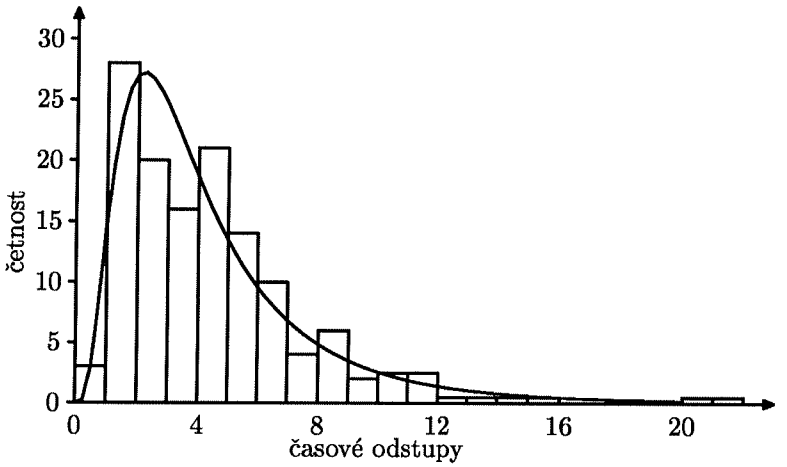
Tabulka 35.

Hodnotu $\chi^2 = \sum_{i=1}^{l} \frac{(n_i-np_i)^2}{np_i} = 51.5181$ srovnáme s 5% horním kvantilem χ^2 rozdělení o $l-k-1 = 11-2-1 = 8$ stupních volnosti, tj. $\chi^2_{0.05}[8] = 15.507$. Ze srovnání hodnoty testové statistiky χ^2 a příslušné kritické hodnoty vyplývá, že nulovou hypotézu zamítáme, jinak řečeno, normální rozdělení není v tomto případě vhodným modelem. Spočteme-li k hodnotě testové statistiky $\chi^2 = 51.4181$ odpovídající $p-$hodnotu, vyjde $2.18 \cdot 10^{-8}$, což je číslo daleko, daleko menší než 0.05. Tento výsledek ještě mnohem výrazněji zdůrazňuje nevhodnost normálního rozdělení pro modelování napozorovaných dat.

Rozdělení, které v důsledku výrazné šikmosti histogramu připadá v úvahu, je rozdělení logaritmicko-normální. Odhadněme proto parametry logaritmicko-normálního rozdělení $LN(\mu, \sigma^2, x_0 = 0)$:

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum n_i \ln \xi_i = 1.269639, \\ \widehat{\sigma} &= \sqrt{\frac{1}{n}\sum n_i (\ln \xi_i - \hat{\mu})^2} = 0.699603. \end{aligned}$$

Obor hodnot logaritmicko-normálního rozdělení rozdělme do tříd $(0, 2), \langle 2, 3), \ldots, \langle 7, 8), \langle 8, 10), \langle 10, \infty)$. Pro každou třídu $\langle r_i, r_{i+1})$ spočtěme teoretickou četnost np_i, kde $p_i = \Phi\left(\frac{\ln(r_{i+1})-\widehat{\mu}}{\widehat{\sigma}}\right) - \Phi\left(\frac{\ln(r_i)-\widehat{\mu}}{\widehat{\sigma}}\right)$ a porovnejme skutečné a teoretické četnosti.



Obrázek 20.

časový odstup (s)	skutečná četnost n_i	teoretická četnost np_i	$\frac{(n_i-np_i)^2}{np_i}$
$(0,2)$	31	27.055	0.5752
$\langle 2,3)$	20	26.199	1.4666
$\langle 3,4)$	16	21.487	1.4011
$\langle 4,5)$	21	15.866	1.6614
$\langle 5,6)$	14	11.332	0.6281
$\langle 6,7)$	10	8.036	0.4799
$\langle 7,8)$	4	5.719	0.5169
$\langle 8,10)$	8	7.078	0.1201
$\langle 10,\infty)$	8	9.228	0.1634

Tabulka 36.

Hodnotu statistiky $\chi^2 = \sum_{i=1}^{l} \frac{(n_i-np_i)^2}{np_i} = 7.0126$ porovnáme s 5 % horním kvantilem χ^2 rozdělení o $l-k-1 = 9-2-1 = 6$ stupních volnosti $\chi^2_{0.05}[6] = 12.592$. Nulovou hypotézu o vhodnosti logaritmicko-normálního rozdělení nezamítáme. Příslušná $p-$hodnota je rovna 0.3297, což je číslo značně větší než 0.05. Tento fakt nám znovu potvrzuje, že logaritmicko-normální rozdělení je vhodným modelem pro naše data.

□

Část VII. Regrese

33. Lineární regrese s jednou vysvětlující proměnnou

Začněme příkladem. Z fyziky si ještě dobře pamatujeme Hookův zákon. Je-li tyč namáhána vnější silou F, pak přírůstek délky Y je přímo úměrný síle F. Budeme-li provádět tento pokus prakticky, to jest měřit sílu i prodloužení tyče, nebudou hodnoty prodloužení ležet zcela přesně v přímce, ale budou kolem této přímky mírně kolísat. Malé odchylky od přímky budou způsobeny nejspíš chybami měření.

Uveďme ještě další příklad. Počet získaných bodů u písemné zkoušky z matematiky je úměrný počtu hodin, které student věnoval přípravě. Přesto, i kdybychom znali přesnou dobu, kterou se student na zkoušku připravoval, nemůžeme předem přesně říci, kolik bodů u zkoušky získá. Závislá proměnná Y - počet získaných bodů závisí na nezávislé proměnné X - počtu hodin přípravy, ale její hodnotu ovlivňují ještě další faktory, například momentální indispozice studenta, studentovy vlohy apod.

V obou těchto případech bylo možno Y vyjádřit jako součet hodnoty nějaké funkce f v bodě x, kde x označuje hodnotu, kterou nabyla proměnná X, a náhodné chyby e:

$$Y \quad = \quad f(x) \quad + \quad e.$$

Náhodná chyba e vzniká například jako chyba měření nebo působením jiných náhodných vlivů. Veličina Y je zřejmě náhodná, neboť vzniká jako součet nenáhodné funkce $f(x)$ a náhodné chyby e.

Prvním úkolem statistického zkoumání je najít regresní funkci f, známe-li n dvojic $(x_1, y_1), \ldots, (x_n, y_n)$, kde x_i je hodnota *nezávislé, vysvětlující proměnné* X_i a y_i hodnota odpovídající *závislé, vysvětlované proměnné* Y_i, přičemž předpokládáme, že

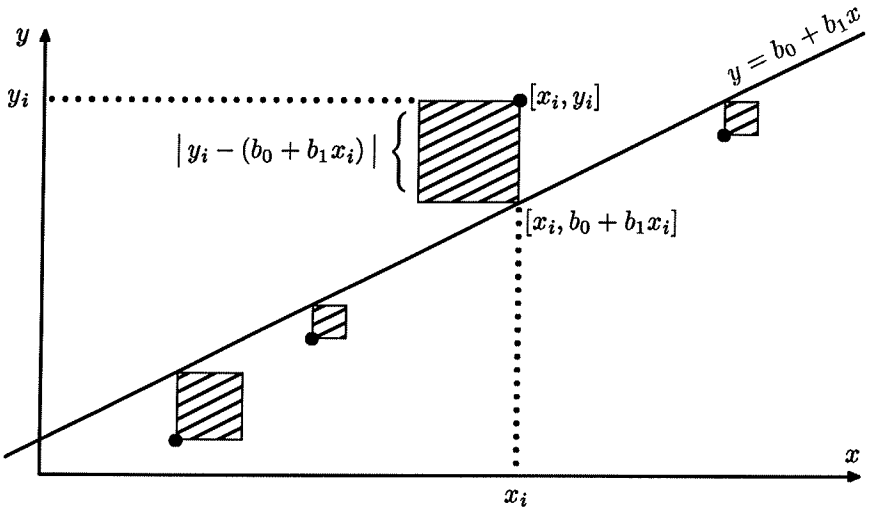
$$Y_i = f(x_i) + e_i, \quad i = 1, \ldots, n.$$

Víme-li předem, že regresní funkce $f(x)$ má tvar

$$f(x) = b_0 + b_1\, x,$$

pak mluvíme o *lineární regresi s jednou vysvětlující proměnnou*, případně o *jednoduché lineární regresi.*

Prvním krokem v případě lineární regrese je odhadnutí koeficientů b_0 a b_1. Nejčastěji používanou metodou pro jejich odhad je *metoda nejmenších čtverců.* Tato metoda hledání b_0 a b_1 spočívá v minimalizaci součtu kvadrátů chyb $g(b_0, b_1) = \sum_{i=1}^{n} \big(Y_i - (b_0 + b_1\, x_i)\big)^2$. Na obrázku 21 odpovídá funkce $g(b_0, b_1)$ vyšrafované ploše.



Obrázek 21.

Minimum funkce $g(b_0, b_1)$ hledáme tak, že položíme parciální derivace

$$\frac{\partial g(b_0, b_1)}{\partial b_0} \quad \text{a} \quad \frac{\partial g(b_0, b_1)}{\partial b_1}$$

rovny nule. Vznikne tak lineární soustava dvou rovnic pro dvě neznámé b_0 a b_1:

$$\begin{aligned} b_0\, n \quad &+ \quad b_1 \sum x_i \quad = \quad \sum Y_i, \\ b_0 \sum x_i \quad &+ \quad b_1 \sum x_i^2 \quad = \quad \sum x_i Y_i. \end{aligned}$$

Řešením získáme odhady:

$$\begin{aligned} \widehat{b}_0 &= \overline{Y} - \widehat{b}_1\, \overline{x}, \\ \widehat{b}_1 &= \frac{\sum (x_i - \overline{x})\, Y_i}{\sum (x_i - \overline{x})^2} = \frac{\sum x_i\, Y_i - n\, \overline{x}\, \overline{Y}}{n\, \sigma_n^2(x)}. \end{aligned}$$

Pokud náhodné chyby $e_1, \ldots, e_n$ jsou nezávislé, stejně rozdělené s normálním rozdělením $N(0, \sigma^2)$, pak jsou odhady $\widehat{b}_0$, $\widehat{b}_1$ nejlepší nestranné odhady b_0, b_1.

Optimální přímka získaná metodou nejmenších čtverců je $y = \widehat{b}_0 + \widehat{b}_1\, x$. Nepřesnost, které jsme se při prokládání nejlepší přímky ve smyslu metody nejmenších čtverců dopustili, se rovná

$$S_e = \sum \left(Y_i - (\widehat{b}_0 + \widehat{b}_1\, x_i)\right)^2 = \sum Y_i^2 - \widehat{b}_0 \sum Y_i - \widehat{b}_1 \sum x_i Y_i.$$

Veličina S_e se nazývá *residuální součet čtverců.* Celkový rozptyl vysvětlované proměnné Y se dá vyjádřit jako součet *nevysvětlitelné části rozptylu* , která se rovná residuálnímu součtu čtverců

S_e, a *vysvětlitelné části rozptylu* $\sum\big((\widehat{b}_0+\widehat{b}_1\,x_i)-\overline{Y}\big)^2=\widehat{b}_1^2\sum(x_i-\overline{x})^2$:

$$\sum(Y_i-\overline{Y})^2=\sum\big(Y_i-(\widehat{b}_0+\widehat{b}_1\,x_i)\big)^2+\widehat{b}_1^2\sum(x_i-\overline{x})^2.$$

Podíl vysvětlitelné části rozptylu k celkovému rozptylu

$$d=\frac{\widehat{b}_1^2\sum(x_i-\overline{x})^2}{\sum(Y_i-\overline{Y})^2}$$

se nazývá *koeficient determinace*. Platí $d=r^2$, kde r je výběrový korelační koeficient. Čím lépe funkce $y=\widehat{b}_0+\widehat{b}_1\,x$ vyjadřuje závislost Y na X, to znamená, čím „blíže jsou body o souřadnicích (x_i,y_i), $i=1,\dots,n$, přímce $y=\widehat{b}_0+\widehat{b}_1\,x$“, tím je nevysvětlitelná část rozptylu menší a koeficient determinace se blíží 1 (100 %).

Pro sledování chování odhadů b_0 a b_1 z hlediska pravděpodobnosti budeme nadále předpokládat, že náhodné veličiny e_i, $i=1,\dots,n$, jsou nezávislé, stejně rozdělené s normálním rozdělením $N(0,\sigma^2)$. Odhadem parametru σ^2 je *residuální rozptyl*:

$$s_r^2=\frac{S_e}{n-2}=\frac{1}{n-2}\sum\big(Y_i-(\widehat{b}_0+\widehat{b}_1\,x_i)\big)^2=\frac{1}{n-2}\Big(\sum Y_i^2-\widehat{b}_0\sum Y_i-\widehat{b}_1\sum x_iY_i\Big).$$

Odhady $\widehat{b}_0$ a $\widehat{b}_1$ jsou stejně jako s_r^2 opět náhodné veličiny. Z předpokladu o normálním rozdělení chyb $e_1,\dots,e_n$ vyplývá, že rovněž veličiny $Y_1,\dots,Y_n$ mají normální rozdělení. Dále je zřejmé, že veličiny $\widehat{b}_0$ a $\widehat{b}_1$ jsou lineárními kombinacemi veličin $Y_1,\dots,Y_n$, a tudíž jsou rovněž normálně rozdělené, přičemž

$$\mathrm{E}\,\widehat{b}_0=b_0,\quad \mathrm{Var}\,\widehat{b}_0=\sigma_{\widehat{b}_0}^2=\sigma^2\,\frac{\sum x_i^2}{n\sum(x_i-\overline{x})^2},$$
$$\mathrm{E}\,\widehat{b}_1=b_1,\quad \mathrm{Var}\,\widehat{b}_1=\sigma_{\widehat{b}_1}^2=\sigma^2\,\frac{1}{\sum(x_i-\overline{x})^2}.$$

Znamená to například, že s pravděpodobností 95 % platí

$$|\widehat{b}_1-b_1|<1.96\,\sigma\,\frac{1}{\sqrt{\sum(x_i-\overline{x})^2}}.$$

Ve většině případů však parametr σ, který udává velikost chyb, neznáme, a musíme ho nahradit odhadem s_r.

$100\,(1-\alpha)\%$ intervaly spolehlivosti pro koeficienty b_0 a b_1 mají tvar:

$$(\widehat{b}_0-t_{\alpha/2}[n-2]\,s_{\widehat{b}_0},\ \ \widehat{b}_0+t_{\alpha/2}[n-2]\,s_{\widehat{b}_0}),$$
$$(\widehat{b}_1-t_{\alpha/2}[n-2]\,s_{\widehat{b}_1},\ \ \widehat{b}_1+t_{\alpha/2}[n-2]\,s_{\widehat{b}_1}),$$

kde $s_{\widehat{b}_0}$ je odhad směrodatné odchylky odhadu parametru b_0:

$$s_{\widehat{b}_0} = s_r \frac{\sqrt{\sum x_i^2}}{\sqrt{n \sum (x_i - \overline{x})^2}}$$

a $s_{\widehat{b}_1}$ je odhad směrodatné odchylky odhadu parametru b_1:

$$s_{\widehat{b}_1} = s_r \frac{1}{\sqrt{\sum (x_i - \overline{x})^2}}.$$

Chceme-li ověřit, zda naše data vykazují trend, to je, zda jimi lze rozumně proložit přímka, která není rovnoběžná s osou x, testujeme hypotézu $H_0 : b_1 = 0$ proti alternativě $A : b_1 \neq 0$. Zamítací pravidlo pro hladinu významnosti α je dáno následovně:

$$t = \frac{|\widehat{b}_1|}{s_{\widehat{b}_1}} > t_{\alpha/2}[n-2].$$

Pokud chceme prokázat, že trend proložené přímky je kladný, to znamená, že s růstem nezávislé veličiny má závislá veličina rovněž tendenci růst, testujeme nulovou hypotézu $H_0 : b_1 = 0$ proti alternativě $A : b_1 > 0$. Zamítací pravidlo pro hladinu významnosti α je dáno:

$$t = \frac{\widehat{b}_1}{s_{\widehat{b}_1}} > t_{\alpha}[n-2].$$

Analogicky odvodíme test pro jednostrannou alternativu $A : b_1 < 0$.

Testujeme-li nulovou hypotézu $H_0 : b_0 = 0$ proti alternativě $A : b_0 \neq 0$, pak zamítací pravidlo je dáno:

$$t = \frac{|\widehat{b}_0|}{s_{\widehat{b}_0}} > t_{\alpha/2}[n-2].$$

Analogicky odvodíme i zamítací pravidla pro jednostranné alternativy.

Často si přejeme odhadnout střední hodnotu závisle proměnné, nabude-li nezávisle proměnná nějaké určité hodnoty x. Bodový odhad regresní funkce $y(x)$ v bodě x se spočte ze vztahu:

$$\widehat{y}(x) = \widehat{b}_0 + \widehat{b}_1 x$$

a $100\,(1-\alpha)\,\%$ *interval spolehlivosti pro* $y(x)$ má tvar:

$$(\widehat{y}(x) - t_{\alpha/2}[n-2]\, s_{\widehat{y}(x)},\ \widehat{y}(x) + t_{\alpha/2}[n-2]\, s_{\widehat{y}(x)}),$$

kde

$$s_{\widehat{y}(x)} = s_r \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum (x_i - \overline{x})^2}}.$$

Někdy nás místo střední hodnoty neznámé veličiny Y při dané hodnotě x zajímá, jak se bude chovat jedna konkrétní hodnota Y při daném x. Pokud bychom znali koeficienty b_0, b_1 a rozptyl σ^2, pak by tato hodnota ležela s pravděpodobností $(1-\alpha)$ v intervalu $(b_0 + b_1 x - u_{\alpha/2}\sigma,\ b_0 + b_1 x + u_{\alpha/2}\sigma)$, kde $u_{\alpha/2}$ je $100\,\alpha/2\,\%$ horní kvantil standardního normálního rozdělení. Jestliže neznáme b_0, b_1 ani σ^2, pak je musíme nahradit odhady $\widehat{b}_0$, $\widehat{b}_1$ a s_r^2, čímž se ovšem dopouštíme

jisté chyby, která způsobí, že interval, v kterém předpokládáme, že bude ležet hodnota Y při daném x, bude větší. Tento $100\,(1-\alpha)\,\%$ *predikční interval* má tvar:

$$(\widehat{y}(x) - t_{\alpha/2}[n-2]\, s_{Y(x)},\ \widehat{y}(x) + t_{\alpha/2}[n-2]\, s_{Y(x)}),$$

kde odhad směrodatné odchylky předpovědi $s_{Y(x)}$ spočteme následovně:

$$s_{Y(x)} = s_r \sqrt{1 + \frac{1}{n} + \frac{(x-\overline{x})^2}{\sum (x_i - \overline{x})^2}}.$$

Všimněme si, že odhad směrodatné odchylky střední hodnoty $s_{\widehat{y}(x)}$ i odhad směrodatné odchylky predikce $s_{Y(x)}$ jsou tím menší, čím je hodnota x blíže aritmetickému průměru $\overline{x}$. Odtud plyne, že rovněž interval spolehlivosti, resp. predikční interval, má pro různé hodnoty x různou šířku, přičemž je tím užší, čím je x bližší aritmetickému průměru. To znamená například, že pásy spolehlivosti konstruované kolem přímky $y(x) = \widehat{b}_0 + \widehat{b}_1 x$ nemají hranice rovnoběžné s touto přímkou. Dále si všimněme, že oba odhady $s_{\widehat{y}(x)}$ i $s_{Y(x)}$ jsou tím menší, čím více pozorování máme a čím jsou hodnoty nezávisle proměnné více rozptýleny, neboť pak je hodnota $\sum (x_i - \overline{x})^2$ větší. Přesnost odhadu střední hodnoty $y(x)$ i přesnost předpovědi hodnoty Y při daném x můžeme zlepšit, provedeme-li více měření, přičemž se snažíme o to, aby hodnoty nezávisle proměnné měly co největší rozptyl.

Poznámka

U uvedeném výkladu jsme předpokládali, že vysvětlující proměnná je nenáhodná veličina, jejíž hodnoty x jsme mohli sami stanovit. Všechny předchozí vztahy zůstávají v platnosti i v případě, že vysvětlující proměnná X je náhodná veličina, přičemž rozdělení Y_i při dané hodnotě $X_i = x_i$, $i = 1, \ldots, n$, je normální se střední hodnotou $f(x_i)$ a rozptylem σ^2 a je zachován i předpoklad o nezávislosti. Podrobnější vysvětlení o tomto případě najde čtenář v Likešovi a Machkovi (1983).

Příklad 70.

Ve zkušebně výrobků chtěli zjistit, jaká je závislost mezi výkonností výrobku v testu a výkonností téhož výrobku za skutečných podmínek provozu. Za tímto účelem bylo vybráno 10 výrobků. U každého z nich byla změřena jak výkonnost v testu X, tak výkonnost za skutečných podmínek provozu Y. Naměřené údaje jsou uvedeny v tabulce 37.

x	68	54	90	64	61	51	79	51	83	48
y	55	38	95	63	58	40	74	32	84	45

Tabulka 37.

Předpokládejme, že vztah mezi výkonností v testu a výkonností v provozu je lineární. Použijte lineární regresi (metodu nejmenších čtverců) pro odhad koeficientů b_0 a b_1 a rozhodněte, zda lineární model je pro popis závislosti Y na X vhodný. Odhadněme výkonnost v provozu výrobku, jestliže jeho výkonnost v testu byla $x = 58$.

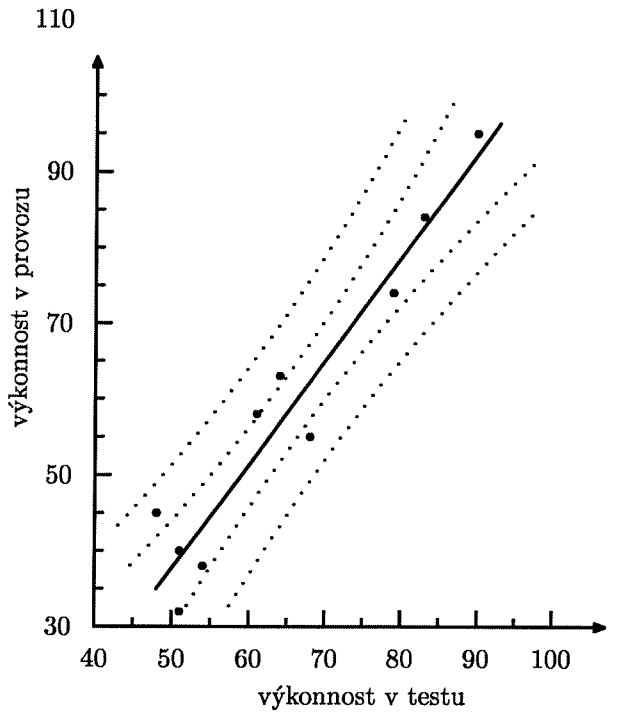
Řešení:

Výběrový korelační koeficient $r = 0.9588$ a koeficient determinace $d = 91.93\,\%$. Proto lze usuzovat na silnou lineární závislost mezi veličinami X a Y. Spočtěme odhady $\widehat{b_0}$ a $\widehat{b_1}$ koeficientů b_0 a b_1 a odhadněme směrodatné odchylky těchto odhadů. Dále spočtěme příslušné statistiky t vhodné pro testování nulovosti koeficientů b_0 a b_1 a odpovídající $p-$hodnoty.

	odhad parametru	odhad směrodatné odchylky odhadu	statistika t	$p-$hodnota
b_0	-29.1158	9.3798	-3.1041	0.01457
b_1	1.3485	0.1413	9.5463	0.00001

Tabulka 38.

Údaje z tabulky 37 i proložená optimální přímka jsou na obrázku 22.



Obrázek 22.

Použijeme-li údaje z tabulky 38 a toho, že 2.5 % horní kvantil t rozdělení o 8 stupních volnosti $t_{0.025}[8] = 2.306$, můžeme najít 95 % interval spolehlivosti pro parametr b_0: $(-50.75,\ -7.49)$ a pro parametr b_1: (1.022, 1.674).

Hypotézu $H_0 : b_1 = 0$ proti alternativě $A : b_1 \neq 0$ na hladině $\alpha = 0.05$ zamítáme, neboť $|t| = 9.546 > t_{0.025}[8] = 2.306$. Zamítnutí nulové hypotézy vyplývá rovněž z toho, že $p-$hodnota 0.00001 je menší než $\alpha = 0.05$.

Víme-li, že výkonnost jistého výrobku v testu byla $x = 58$, pak očekávaná výkonnost v provozu bude

$$\widehat{y}(58) = -29.1158 + 1.3485 \cdot 58 = 49.1.$$

Odhad směrodatné odchylky predikce je roven $s_{Y(x)} = 6.652$, a tudíž výkonnost ve skutečných podmínkách provozu se s 95 % spolehlivostí bude pohybovat v mezích (33.8, 64.4).

□

Lineární regresi je možno použít i v případě, že mezi proměnnými X a Y předpokládáme jiné vztahy než lineární. Jedná se o případy, kdy můžeme získat lineární vztah vhodnou transformací původních proměnných X a Y, například:

1) $Y = a\,X^b$,

2) $Y = e^{a+bX}$,

3) $Y = 1/(a + bX)$.

V případě 1) zlogaritmováním dostaneme:

$$\ln Y = \ln a \ + b \ln X$$

a po substituci $Y^* = \ln Y$, $b_0 = \ln a$, $b_1 = b$, $X^* = \ln X$ vztah:

$$Y^* = b_0 + b_1 X^*.$$

V případě 2) opět zlogaritmováním získáváme vztah:

$$\ln Y = a \ + b\,X$$

a substituce $Y^* = \ln Y$, $b_0 = a$, $b_1 = b$, $X^* = X$ opět vede na vztah:

$$Y^* = b_0 + b_1 X^*.$$

V případě 3) substitucí $Y^* = 1/Y$, $b_0 = a$, $b_1 = b$, $X^* = X$ získáváme opět:

$$Y^* = b_0 + b_1 X^*.$$

Je třeba si však uvědomit, že v těchto případech při posuzování pravděpodobnostního chování odhadnutých parametrů musíme být opatrní, neboť při transformaci se mění i rozdělení veličin X a Y.

34. Lineární regrese s více vysvětlujícími proměnnými

V předchozím modelu jsme uvažovali jednu vysvětlovanou proměnnou, kterou jsme se snažili vyjádřit jako lineární funkci jediné vysvětlující proměnné. V některých případech však může být vysvětlujících proměnných víc. Okamžitý průtok v řece je závislý nejen na množství srážek v posledních dnech, ale i na podílu zalesnění povodí řeky a pravděpodobně i na dalších veličinách.

Předpokládejme, že máme k vysvětlujících proměnných a jednu vysvětlovanou proměnnou. Předpokládejme dále, že vysvětlující proměnné nabyly hodnoty $(x_{i1},\dots,x_{ik})$, $i=1,\dots,n$. Mezi vysvětlovanou proměnnou Y_i a hodnotami $x_{i1},\dots,x_{ik}$ pro $i=1,\dots,n$ platí vztah:

$$Y_i=\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik}+e_i,\quad i=1,\dots,n, \tag{34.1}$$

kde e_i je náhodná chyba. Předchozí vztah lze zapsat maticově:

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e},$$

kde $\boldsymbol{Y}$ je n–rozměrný sloupcový vektor hodnot vysvětlované proměnné, $\boldsymbol{X}$ je matice konstant typu $n\times(k+1)$, $\boldsymbol{\beta}$ je $(k+1)$-rozměrný sloupcový vektor neznámých parametrů a $\boldsymbol{e}$ je n-rozměrný sloupcový vektor chyb. Úkolem je opět co nejlépe odhadnout vektor parametrů $\boldsymbol{\beta}=(\beta_0,\dots,\beta_k)'$. Použijeme-li již známé metody nejmenších čtverců, najdeme odhady $\widehat{\beta_0},\dots,\widehat{\beta_k}$ minimalizací funkce

$$g(\beta_0,\dots,\beta_k)=\sum\big(Y_i-(\beta_0+\beta_1x_{i1}+\cdots+\beta_kx_{ik})\big)^2.$$

Zderivováním podle jednotlivých parametrů a položením parciálních derivací rovným nule, dostaneme soustavu lineárních rovnic:

$$\begin{aligned}
\beta_0 n &+ \beta_1\sum_{i=1}^n x_{i1} &&+\beta_2\sum_{i=1}^n x_{i2} &&+\ldots+\beta_k\sum_{i=1}^n x_{ik} &&=\sum_{i=1}^n Y_i,\\
\beta_0\sum_{i=1}^n x_{i1}&+\beta_1\sum_{i=1}^n x_{i1}^2 &&+\beta_2\sum_{i=1}^n x_{i1}x_{i2} &&+\ldots+\beta_k\sum_{i=1}^n x_{i1}x_{ik} &&=\sum_{i=1}^n x_{i1}Y_i,\\
&&&\vdots\\
\beta_0\sum_{i=1}^n x_{ik}&+\beta_1\sum_{i=1}^n x_{ik}x_{i1}&&+\beta_2\sum_{i=1}^n x_{ik}x_{i2} &&+\ldots+\beta_k\sum_{i=1}^n x_{ik}^2 &&=\sum_{i=1}^n x_{ik}Y_i.
\end{aligned}$$

Shora popsaná soustava se nazývá *soustavu normálních rovnic.* Vyřešením soustavy normálních rovnic získáme odhad $\widehat{\boldsymbol{\beta}}=(\widehat{\beta_0},\dots,\widehat{\beta_k})'$ vektoru parametrů $\boldsymbol{\beta}=(\beta_0,\dots,\beta_k)'$. Maticově lze

soustavu normálních rovnic napsat ve tvaru $(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y}$, kde

$$(\boldsymbol{X}'\boldsymbol{X}) = \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \dots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \dots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \dots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}, \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_{i1}Y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}Y_i \end{pmatrix}$$

Je-li matice $(\boldsymbol{X}'\boldsymbol{X})$ regulární, pak platí $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ je stejně jako matice $(\boldsymbol{X}'\boldsymbol{X})$ symetrická, typu $(k+1)\times(k+1)$:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} v_{00} & v_{01} & \dots & v_{0k} \\ v_{01} & v_{11} & \dots & v_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ v_{0k} & v_{1k} & \dots & v_{kk} \end{pmatrix}. \tag{34.2}$$

Jestliže proložíme body $(x_{i1},\dots,x_{ik},Y_i)$, $i=1,\dots,n$ nadrovinou $y(x_1,\dots,x_n) = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \dots + \widehat{\beta_k}x_k$, pak se ve smyslu metody nejmenších čtverců dopustíme „nepřesnosti" $S_e = \sum\big(Y_i - (\widehat{\beta_0}+\widehat{\beta_1}x_{i1}+\dots+\widehat{\beta_k}x_{ik})\big)^2 = \sum Y_i^2 - \widehat{\beta_0}\sum Y_i - \widehat{\beta_1}\sum x_{i1}Y_i - \dots - \widehat{\beta_k}\sum x_{ik}Y_i$, které říkáme *residuální součet čtverců*. Výrazu $s_r^2 = \frac{1}{n-k-1}S_e$ říkáme *residuální rozptyl*. Residuální rozptyl je tím menší, čím lépe nezávisle proměnné vysvětlují závislou proměnnou pomocí lineárního modelu. Vhodnost použití daného lineárního modelu je možno posuzovat podle *koeficientu determinace* $d = 1 - \frac{S_e}{\sum(Y_i-\overline{Y})^2}$. Koeficient determinace vyjádřený v procentech udává, jakou část z celkového rozptylu $\sum(Y_i-\overline{Y})^2$ lze vysvětlit daným lineárním modelem.

Předpokládáme-li, že náhodné chyby e_i, $i=1,\dots,n$ jsou nezávislé, stejně rozdělené, řídící se normálním rozdělením $N(0,\sigma^2)$, pak odhady $\widehat{\beta_i}$, $i=0,\dots,n$, mají rovněž normální rozdělení se střední hodnotou β_i a rozptylem $\sigma^2_{\widehat{\beta_i}} = \sigma^2 v_{ii}$, kde v_{ii} značí prvek v $i+1$ řádku a $i+1$ sloupci matice $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, viz (34.2). Rozptyl $\sigma^2_{\widehat{\beta_i}}$ lze odhadnout pomocí $s^2_{\widehat{\beta_i}} = s_r^2 v_{ii}$ a směrodatnou odchylku $\sigma_{\widehat{\beta_i}}$ pomocí $s_{\widehat{\beta_i}} = s_r\sqrt{v_{ii}}$. Odtud lze najít $100\,(1-\alpha)\,\%$ interval spolehlivosti pro parametr β_i:

$$\big(\widehat{\beta_i} - t_{\alpha/2}[n-k-1]\,s_{\widehat{\beta_i}},\ \widehat{\beta_i} + t_{\alpha/2}[n-k-1]\,s_{\widehat{\beta_i}}\big)$$

a otestovat, zda se parametr β_i nerovná nule. Jestliže se totiž parametr u určité vysvětlující proměnné rovná nule, znamená to, že vysvětlovaná proměnná na této vysvětlující proměnné nezávisí a že je možno ji z modelu vypustit. Nulová hypotéza $H_0 : \beta_i = 0$ proti alternativě $A : \beta_i \neq 0$ se zamítá, jestliže

$$t_i = \frac{|\widehat{\beta_i}|}{s_{\widehat{\beta_i}}} > t_{\alpha/2}[n-k-1].$$

Testujeme-li obecněji $H_0 : \beta_i = \beta_i^0$ proti alternativě $A : \beta_i \neq \beta_i^0$, pak nulovou hypotézu zamítáme, jestliže

$$t_i = \frac{|\widehat{\beta}_i - \beta_i^0|}{s_{\widehat{\beta}_i}} > t_{\alpha/2}[n-k-1].$$

Poznámka

Shora popsané výsledky je možno rozšířit i na případ, že $(X_{i1}, \ldots, X_{ik})$, $i = 1, \ldots, n$, jsou vektory náhodných veličin, jestliže veličiny Y_i mají při $X_{i1} = x_{i1}, \ldots, X_{ik} = x_{ik}$ podmíněné normální rozdělení se střední hodnotou $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$ a rozptylem σ^2 a jsou nezávislé.

Příklad 71.

V továrně na výrobu kyseliny dusičné ze čpavku byla sledována výkonnost výroby během 21 dní. Nezávisle proměnné byly: proud vzduchu v absorpční věži, teplota chladicí vody a koncentrace cirkulující kyseliny. Třetí proměnná byla vyjádřena v %, odečteno 50 a výsledek vynásoben 10. Závisle proměnná byl podíl čpavku (vyjádřený v % a výsledek vynásobený 10), který uniká neabsorbovaný z absorpční věže, a je tudíž mírou výkonnosti výroby - čím menší množství uniká, tím je provoz výkonnější. Naměřené hodnoty jsou v tabulce 39, ve které x_1 označuje proud vzduchu, x_2 teplotu chladicí vody, x_3 koncentraci kyseliny a y únik čpavku.

den	x_1	x_2	x_3	y	den	x_1	x_2	x_3	y	den	x_1	x_2	x_3	y
1	80	27	89	42	8	62	24	93	20	15	50	18	89	8
2	80	27	88	37	9	58	23	87	15	16	50	18	86	7
3	75	25	90	37	10	58	18	80	14	17	50	19	72	8
4	62	24	87	28	11	58	18	89	14	18	50	19	79	8
5	62	22	87	18	12	58	17	88	13	19	50	20	80	9
6	62	23	87	18	13	58	18	82	11	20	56	20	91	15
7	62	24	93	19	14	58	19	93	12	21	70	20	91	15

Tabulka 39.

Najděte nejlepší lineární model, odhadněte jeho parametry a najděte pro ně 95 % intervaly spolehlivosti.

Řešení:

Hledáme lineární model ve tvaru

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e.$$

Tabulka 40 udává odhady parametrů, odhady jejich směrodatných odchylek, dále statistiky t_j pro testování hypotéz, že se daný parametr rovná nule, a příslušné $p-$hodnoty

nezávisle proměnná	odhad parametru	odhad směrodatné odchylky odhadu	statistika t_j	$p-$hodnota
konstanta	-39.9197	11.8960	-3.3557	0.0038
proud vzduchu X_1	0.7156	0.1349	5.3066	0.0001
teplota chladicí vody X_2	1.2953	0.3680	3.5196	0.0026
koncentrace kyseliny X_3	-0.1521	0.1563	-0.9733	0.3440

Tabulka 40.

Najdeme-li v tabulkách kvantil $t_{0.025}[17] = 2.1098$, pak hypotézy o nulovosti koeficientů u nezávisle proměnných: konstanta, proud vzduch X_1 a teplota vody X_2 musíme zamítnout, zatímco hypotézu o nulovosti koeficientu u proměnné X_3 zamítnout nemůžeme. Navíc nejmenší hladina významnosti, na které by bylo možno tuto hypotézu zamítnout, to znamená příslušná $p-$hodnota, je velmi vysoká, neboť se rovná 0.344. Tento fakt potvrzuje, že proměnná Y na proměnné X_3 nejspíš nezávisí a můžeme ji z modelu vyřadit. Proveďme ještě jednou odhady pro upravený model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$.

nezávisle proměnná	odhad parametru	odhad směrodatné odchylky odhadu	statistika t_j	$p-$hodnota
konstanta	-50.3588	5.1383	-9.8006	0.0000
proud vzduchu X_1	0.6712	0.1267	5.2976	0.0000
teplota chladicí vody X_2	1.2954	0.3675	3.5249	0.0024

Tabulka 41.

Všechny statistiky t_j jsou větší než kvantil $t_{0.025}[18] = 2.1009$ a také všechny $p-$hodnoty jsou daleko menší než 0.05. Odtud vyplývá, že z modelu nelze vyloučit již žádnou další nezávisle proměnnou.

Optimální lineární model má tvar:

$$Y = -50.3588 + 0.6712\, X_1 + 1.2954\, X_2.$$

Odhad $\widehat{\sigma} = s_r = 3.2386$. Odpovídající koeficient determinace $d = 0.904$, což znamená, že lineární model je vhodný pro vyjádření závislosti Y na X_1 a X_2.

Vypočteme ještě 95 % intervaly spolehlivosti pro jednotlivé parametry:

parametr	dolní mez	horní mez
β_0	-61.1567	-39.5610
β_1	0.4049	0.9374
β_2	0.5231	2.0676

Tabulka 42.

□

V modelu (34.1) jsme předpokládali existenci absolutního členu β_0. Někdy však z povahy problému vyplývá, že v případě, že se nezávisle proměnné rovnají nule, pak i závisle proměnná by měla být rovna nule. Zde bychom měli volit model bez absolutního členu:

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + e_i,\ i = 1, \ldots, n.$$

Předpokládáme-li, že veličiny $\{e_i\}$ jsou nezávislé, stejně rozdělené s $N(0, \sigma^2)$, pak můžeme vektor $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ odhadnout pomocí metody nejmenších čtverců

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{Y},$$

kde

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} \sum x_{i1}^2 & \ldots & \sum x_{i1}x_{ik} \\ \vdots & \ldots & \vdots \\ \sum x_{i1}x_{ik} & \ldots & \sum x_{ik}^2 \end{pmatrix}, \quad \boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} \sum x_{i1}Y_i \\ \vdots \\ \sum x_{ik}Y_i \end{pmatrix}.$$

Rozptyl σ^2 se pak odhaduje pomocí $S_e^2/(n-k)$, kde

$$S_e^2 = \sum (Y_i - (\widehat{\beta_1} x_{i1} + \cdots + \widehat{\beta_k} x_{ik}))^2 = \sum Y_i^2 - \widehat{\beta_1} \sum x_{i1}Y_i - \cdots - \widehat{\beta_k} \sum x_{ik}Y_i.$$

Speciální případ nastává pokud model obsahuje pouze jednu nezávislou proměnnou:

$$Y_i = \beta\, x_i + e_i,\ i = 1, \ldots, n.$$

Pak

$$\widehat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2} \quad \text{a} \quad \widehat{\sigma^2} = s_r^2 = \frac{\sum Y_i^2 - \widehat{\beta} \sum x_i Y_i}{n-1}.$$

$100\,(1-\alpha)\,\%$ interval spolehlivosti pro β;

$$(\widehat{\beta} - t_{\alpha/2}[n-1]\, s_{\widehat{\beta}},\ \widehat{\beta} + t_{\alpha/2}[n-1]\, s_{\widehat{\beta}}),$$

kde $s_{\widehat{\beta}} = s_r / \sqrt{\sum x_i^2}$. Pro testování $H_0 : \beta = \beta_0$ proti $A : \beta \neq \beta_0$ má zamítací pravidlo tvar:

$$|\widehat{\beta} - \beta_0| > t_{\alpha/2}[n-1]\, s_{\widehat{\beta}}.$$

35. Polynomická regrese

Uvažujeme jednu vysvětlující proměnnou X a jednu vysvětlovanou proměnnou Y. Předpokládejme, že vysvětlující proměnná X nabyla hodnoty x. Dále předpokládejme, že mezi vysvětlovanou a vysvětlující proměnnou je vztah:

$$Y = b_0 + b_1\, x + \cdots + b_p\, x^p + e,$$

kde e je náhodná chyba. Je-li stupeň polynomu $b_0 + b_1 x + \cdots + b_p x^p$ roven p, pak mluvíme o *polynomické regresi stupně p.*

Uvažujeme-li n dvojic (x_i, Y_i), kde x_i, $i = 1, \ldots, n$, jsou hodnoty, kterých nabyla vysvětlující proměnná, pak odhady parametrů $b_0, \ldots, b_p$ spočtené metodou nejmenších čtverců jsou řešením *soustavy normálních rovnic*:

$$\begin{aligned} b_0 n \quad &+ b_1 \sum_{i=1}^{n} x_i \quad + b_2 \sum_{i=1}^{n} x_i^2 \quad + \ldots + b_p \sum_{i=1}^{n} x_i^p = \sum_{i=1}^{n} Y_i \\ &\vdots \\ b_0 \sum_{i=1}^{n} x_i^p + b_1 \sum_{i=1}^{n} x_i^{(p+1)} &+ b_2 \sum_{i=1}^{n} x_i^{(p+2)} + \ldots + b_p \sum_{i=1}^{n} x_i^{2p} = \sum_{i=1}^{n} x_i^p Y_i. \end{aligned}$$

Všechny závěry týkající se intervalů spolehlivosti pro $b_0, \ldots, b_p$ a testování nulovosti těchto parametrů, které byly uvedeny v článku o regresi s více vysvětlujícími proměnnými, zůstávají v platnosti i pro polynomickou regresi. Tento fakt vyplývá z toho, že polynomická regrese stupně p je speciálním případem regrese s p vysvětlujícími proměnnými, kde vysvětlující proměnné jsou $X, X^2, \ldots, X^p$.

Poznámka

Stejně jako v předchozích dvou článcích můžeme předpokládat, že vysvětlující proměnná X je náhodná. Předpokládá se pak, že veličiny Y_i mají při $X_i = x_i$ podmíněné normální rozdělení se střední hodnotou $b_0 + b_1 x_i + \cdots + b_p x_i^p$ a rozptylem σ^2 a jsou nezávislé.

Příklad 72.

U automobilu Trabant se měřila spotřeba paliva Y_i (v litrech na 100 km) v závislosti na jeho rychlosti x_i. Rychlost x_i budeme pro zjednodušení výpočtů uvádět v (km/h)·10^{-1}, například $x_i = 5$ tedy znamená rychlost 50 km/h. Zkoušky probíhaly na rovině za stejných podmínek. Vůz jel stále se zařazeným 4. rychlostním stupněm. Byly naměřeny tyto výsledky:

x_i	4	5	6	7	8	9	10
Y_i	6.1	5.8	6.0	6.5	6.8	8.1	10.0

Tabulka 43.

Na obrázku 23 jsou tato data znázorněna graficky. Použijte kvadratickou regresi pro vyjádření závislosti Y_i na x_i.

Řešení:

Pro kvadratickou regresi platí:

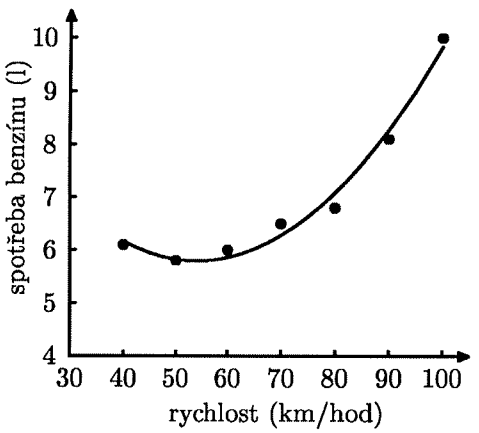
$$Y_i = b_0 + b_1 x_i + b_2 x_i^2 + e_i, \quad i = 1, \ldots, 7.$$

Předpokládáme-li, že náhodné chyby e_i, $i = 1, \ldots, 7$ jsou nezávislé, stejně rozdělené náhodné veličiny řídící se $N(0, \sigma^2)$, pak nejlepší nestranné odhady parametrů b_0, b_1, b_2 získáme metodou nejmenších čtverců. Najděme tyto odhady, odhady jejich směrodatných odchylek, hodnoty statistik pro testování nulovosti parametrů a odpovídající $p-$ hodnoty.

	odhad parametru	odhad směrodatné odchylky odhadu	statistika t_j	$p-$hodnota
b_0	11.392857	1.163022	9.7959	0.0006
b_1	-2.072619	0.351065	-5.9038	0.0041
b_2	0.191667	0.024886	7.7017	0.0015

Tabulka 44.

Naměřená data i proložená parabola jsou na obrázku 23.



Obrázek 23.

Residuální součet čtverců $s_r^2 = 0.052$. Otestujme na hladině významnosti $\alpha = 0.05$ hypotézu $H_0 : b_2 = 0$ proto alternativě $A : b_2 \neq 0$. Testová statistika $t_2 = 7.7017$ je větší než 2.5 % horní kvantil t rozdělení $t_{0.025}[4] = 2.7764$, a tudíž nulovou hypotézu zamítáme. Totéž je potvrzeno tím, že příslušná $p-$ hodnota 0.0015 je menší než hladina významnosti $\alpha = 0.05$. Odtud vyplývá, že závislost spotřeby na rychlosti není lineární.

□

Část VIII. Časové řady

36. Úvod do teorie časových řad

Doposud jsme se zabývali pouze jednou náhodnou proměnnou nebo nejvýše vektorem náhodných veličin. Nyní budeme uvažovat posloupnost náhodných veličin, to znamená $\{X_i\}_{i=a}^{\infty} = X_a, X_{a+1}, \ldots$ nebo $\{X_i\}_{i=-\infty}^{\infty} = \ldots X_{-2}, X_{-1}, X_0, X_1, X_2, X_3, \ldots$ Jestliže indexem posloupnosti je čas, což je nejčastější případ, pak této posloupnosti říkáme *časová řada.* Data, která vytvářejí časovou řadu, vznikají jako pozorování chronologicky uspořádaná v čase. Příkladem časových řad je velikost národního důchodu v České republice v jednotlivých letech nebo průměrné roční průtoky v řece Váh. V obou těchto případech znamená časový index určitý letopočet. Časový index však může označovat i dny, týdny nebo například okamžiky, ve kterých měříme napětí na vibrující konstrukci apod. Úkolem statistické analýzy je obvykle najít „vnitřní mechanismus", podle kterého se daná časová řada řídí, známe-li realizaci určitého úseku časové řady, např. $X_1, \ldots, X_n$. Na základě znalosti tohoto „mechanismu" je pak možné například najít předpovědi časové řady v následujících období.

37. Časové řady s deterministickým trendem a nezávislými chybami

Nejjednodušším příkladem časové řady je posloupnost nezávislých, stejně rozdělených náhodných veličin $\{e_t\}$ se střední hodnotou nulovou a konečným rozptylem, které se říká *bílý šum.* V našem výkladu budeme navíc předpokládat, že každá veličina e_t má normální rozdělení $N(0, \sigma^2)$. Takováto posloupnost je speciálním případem časové řady, která se dá vyjádřit jako součet deterministické (nenáhodné) složky $\mu(t)$ a řady $\{e_t\}$:

$$X_t = \mu(t) + e_t.$$

Předpokládejme, že známe realizaci úseku časové řady $X_1, \ldots, X_n$. Použijeme-li pro popis řady shora zavedeného modelu, můžeme deterministickou složku odhadnout pomocí regrese, kde vysvětlující proměnnou je čas.

Příklad 73.

Spotřeba pitné vody v Praze neustále roste. Na základě zjištěné spotřeby v letech 1969 - 1983 odhadněte spotřebu v roce 1988.

rok	spotřeba (l)	rok	spotřeba
1969	$1.18946 \cdot 10^8$	1977	$1.65366 \cdot 10^8$
1970	$1.25754 \cdot 10^8$	1978	$1.76240 \cdot 10^8$
1971	$1.30503 \cdot 10^8$	1979	$1.81118 \cdot 10^8$
1972	$1.39776 \cdot 10^8$	1980	$1.90983 \cdot 10^8$
1973	$1.43598 \cdot 10^8$	1981	$2.00118 \cdot 10^8$
1974	$1.45071 \cdot 10^8$	1982	$2.16796 \cdot 10^8$
1975	$1.51656 \cdot 10^8$	1983	$2.23448 \cdot 10^8$
1976	$1.60480 \cdot 10^8$		

Tabulka 45.

Řešení:

Použijeme-li lineární regresi

$$X_t = b_0 + b_1 (t - 1968) + e_t,$$

pak pro odhady získané metodou nejmenších čtverců platí:

$$\widehat{b_0} = 1.06968 \cdot 10^8, \quad \widehat{b_1} = 7.21125 \cdot 10^6.$$

Data z tabulky 45 a proložená přímka

$$y = 1.06968 \cdot 10^8 + 7.21125 \cdot 10^6 \cdot (t - 1968)$$

jsou graficky znázorněna na obrázku 24.

Použijeme-li kvadratickou regresi

$$X_t = b_0 + b_1 (t - 1968) + b_2 (t - 1968)^2 + e_t,$$

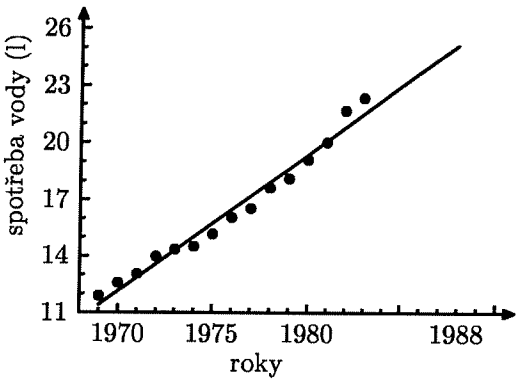
pak pro odhady získané metodou nejmenších čtverců platí:

$$\widehat{b_0} = 1.18925 \cdot 10^8, \quad \widehat{b_1} = 2.99105 \cdot 10^6 \quad \text{a} \quad \widehat{b_2} = 2.63763 \cdot 10^5.$$

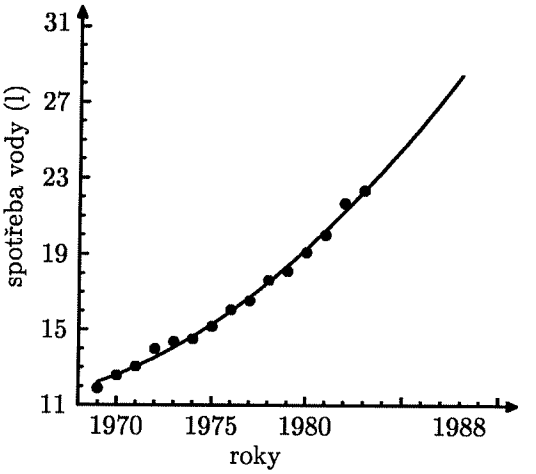
Data z tabulky 45 a proložená parabola

$$y = 1.18925 \cdot 10^8 + 2.99105 \cdot 10^6 (t - 1968) + 2.63763 \cdot 10^5 (t - 1968)^2$$

jsou graficky znázorněna na obrázku 25.



Obrázek 24.



Obrázek 25.

Použijeme-li lineárního vztahu, pak předpověď spotřeby na rok 1988 činí $2.51193 \cdot 10^8$ litrů, použijeme-li kvadratického vztahu, pak je předpověď spotřeby $2.84251 \cdot 10^8$ litrů. Pro zajímavost uveďme, že skutečná spotřeba vody v Praze v roce 1988 byla $2.377153 \cdot 10^8$ litrů.

□

Poznámka

Poznamenejme, že předpovědi spotřeby vody se pro praxi provádějí jemnějšími metodami, kde se bere v úvahu podrobnější rozbor denních spotřeb.

Příklad 74.

V celém světě roste počet obyvatel velkých měst a městských aglomerací. Z údajů v tabulce 46 se pokuste odhadnout počet obyvatel Las Vegas a okolí v roce 1990.

rok	1920	1930	1940	1950	1960	1970	1980
počet obyvatel	4 895	8 532	16 414	48 289	127 288	273 288	461 816

Tabulka 46.

Poznamenejme, že pro růst populace se obvykle používá exponenciální model:

$$X_t = e^{(a+bt)} \cdot u_t,$$

kde $\{u_t\}$ jsou nezávislé náhodné veličiny řídící se logaritmicko-normálním rozdělením $LN(\mu = 0, \sigma^2, x_0 = 0)$.

Řešení:

Vzhledem k tomu, že známe údaje o počtu obyvatel pouze v letech $s = 1920, 1930, \ldots, 1980$, proveďme nejprve časovou transformaci $t = \frac{s-1910}{10}$. Získáme tak časovou řadu $\{X_t\}$, kde X_1 značí počet obyvatel v roce 1920, X_2 počet obyvatel v roce 1930, apod. Naším úkolem je ze znalosti úseku časové řady $X_1, \ldots, X_7$ odhadnout X_8, tj. počet obyvatel v roce 1990. Pro odhad

hodnoty X_8 použijeme model:

$$X_t = e^{(a+bt)} \cdot u_t,$$

kde $\{u_t\}$ jsou nezávislé veličiny mající $LN(\mu = 0, \sigma^2, x_0 = 0)$ rozdělení. Odtud po zlogaritmování získáme:

$$\ln X_t = a \ + b\,t + e_t,$$

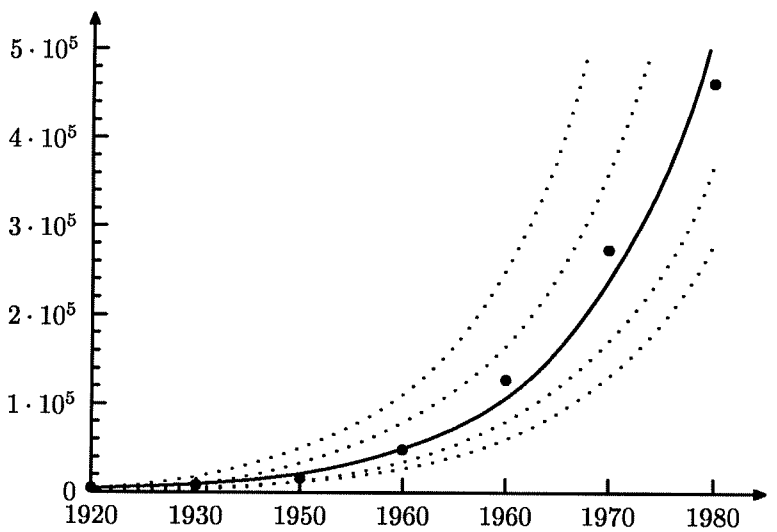
kde $\{e_t\}$ jsou nezávislé veličiny řídící se rozdělením $N(\mu = 0, \sigma^2)$. Zlogaritmováním jsme úlohu převedli na problém lineární regrese. Spočtěme odhady koeficientů a i b:

$$\widehat{a} = 7.53303, \quad \widehat{b} = 0.80795.$$

Po dosazení získáme vztah pro růst počtu obyvatel

$$X_t = e^{7.53303+0.808\,t},$$

který je spolu s hodnotami $X_1, \ldots, X_7$ graficky znázorněn na obrázku 26.



Obrázek 26.

Použijeme-li exponenciálního modelu, pak odhad počtu obyvatel v Las Vegas a okolí v roce 1990 činí: $\widehat{X_8} = 1.1986 \cdot 10^6$. □

Pro časovou řadu, u které se projevují výrazné periodické vlastnosti, se často používá model:

$$X_t = \mu(t) + e_t,$$

kde

$$\mu(t) = \mu + \sum_{j=1}^{p} C_j \cos(2\pi\lambda_j t + \phi_j)$$

nebo ekvivalentně

$$\mu(t) = \mu + \sum_{j=1}^{p} \big(A_j \cos(2\pi\lambda_j t) + B_j \sin(2\pi\lambda_j t)\big),$$

kde $A_j = C_j \cos\phi_j$ a $B_j = -C_j \sin\phi_j$.

Cílem statistického šetření je obvykle odhadnout konstanty μ, $A_1, \ldots, A_p$, $B_1, \ldots, B_p$ a odtud příslušné amplitudy $C_1, \ldots, C_p$ a fázová posunutí $\phi_1, \ldots, \phi_p$, známe-li předem frekvence $\omega_1 = 2\pi\lambda_1, \ldots, \omega_p = 2\pi\lambda_p$, se kterými se daná časová řada periodicky mění. Předpokládejme, že jsme napozorovali hodnoty časové řady $X_1, \ldots, X_n$, pak odhad parametrů μ, $A_1, \ldots, A_p$, $B_1, \ldots, B_p$ je dán následovně:

$$\widehat{\mu} = \frac{1}{n}\sum_{t=1}^{n} X_t,$$

$$\widehat{A_j} = \frac{2}{n}\sum_{t=1}^{n}(X_t - \overline{X})\cos(\omega_j t), \quad j = 1, \ldots, p,$$

$$\widehat{B_j} = \frac{2}{n}\sum_{t=1}^{n}(X_t - \overline{X})\sin(\omega_j t), \quad j = 1, \ldots, p.$$

Příklad 75.

V tabulce 47 jsou udány průměrné roční průtoky řeky Niger v Coulicouro (Mali) měřené v letech 1907 - 1957.

pořadí t	rok	průtok X_t	pořadí t	rok	průtok X_t	pořadí t	rok	průtok X_t
1	1907	39.793	18	1924	77.575	35	1941	43.707
2	1908	42.892	19	1925	83.119	36	1942	35.063
3	1909	69.040	20	1926	59.309	37	1943	41.967
4	1910	43.653	21	1927	69.366	38	1944	35.118
5	1911	56.700	22	1928	76.161	39	1945	43.598
6	1912	45.990	23	1929	72.899	40	1946	54.195
7	1913	28.757	24	1930	71.269	41	1947	44.414
8	1914	32.889	25	1931	61.538	42	1948	58.983
9	1915	48.926	26	1932	62.571	43	1949	48.871
10	1916	48.491	27	1933	57.461	44	1950	53.275
11	1917	52.024	28	1934	51.590	45	1951	75.291
12	1918	56.265	29	1935	50.883	46	1952	58.624
13	1919	43.035	30	1936	60.885	47	1953	69.583
14	1920	43.653	31	1937	45.229	48	1954	73.497
15	1921	36.803	32	1938	51.970	49	1955	72.465
16	1922	52.242	33	1939	48.002	50	1956	48.001
17	1923	54.797	34	1940	41.369	51	1957	73.660

Tabulka 47.

Průtoky v tabulce 47 jsou uvedeny v cfs 10^{-3}. Odhadněte parametry μ, A, B, jestliže se rozhodneme modelovat průměrné roční průtoky řadou:

$$X_t = \mu + A\ \cos(\omega t) + B \sin(\omega t) + e_t, \quad t = 1, 2, \ldots,$$

kde $\omega = 0.2439$.

Řešení:

Pro odhady parametrů platí:

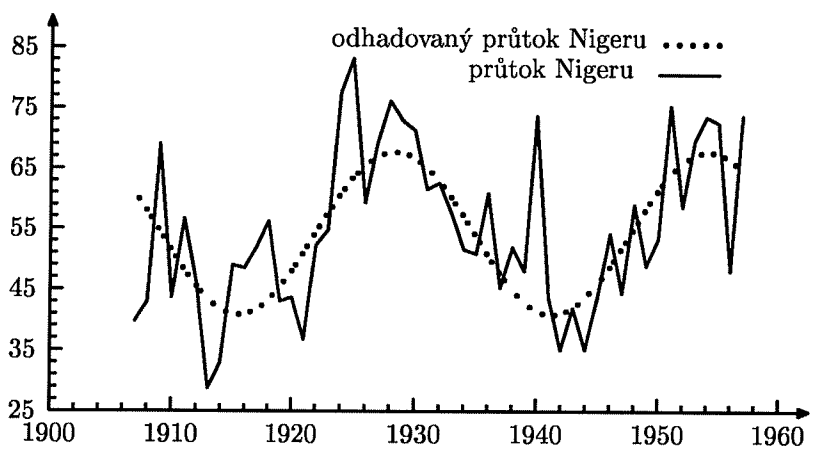
$$\widehat{\mu} = \overline{x} = 54.2639,$$

$$\widehat{A} = \frac{2}{51} \sum_{t=1}^{51} (X_t - 54.2639) \cos(0.2439\, t) = 9.2199,$$

$$\widehat{B} = \frac{2}{51} \sum_{t=1}^{51} (X_t - 54.2639) \sin(0.2439\, t) = -9.8651.$$

Na obrázku 27 jsou zakresleny skutečné průměrné roční průtoky Nigeru a jejich odhady spočtené z modelu:

$$X_t = 54.2639 + 9.2199\ \cos(0.2439\, t) - 9.8651\ \sin(0.2439\, t).$$



Obrázek 27.

38. Periodogram

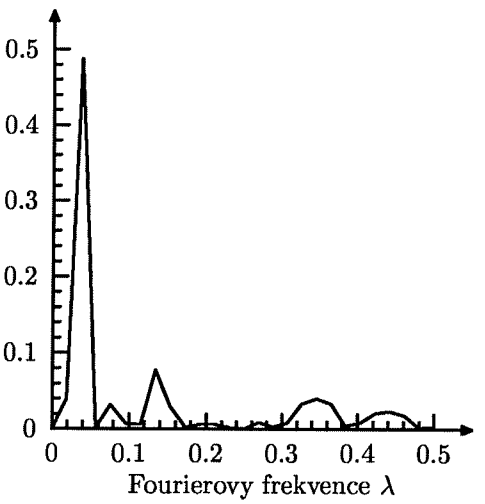
V některých případech jsou frekvence, se kterými se časová řada periodicky mění, předem známy. Frekvence vibrace stavební konstrukce může být určena tvarem konstrukce nebo je dána frekvencí vybuzující vnější síly. U ekonomických časových řad je frekvence zase dána sezónními vlivy apod. Často však frekvence určující chování dané časové řady předem neznáme. V takovém případě je možno pokusit se vyhledat je pomocí periodogramu.

Periodogram $I(\lambda)$ je kvadrát absolutní velikosti *Fourierovy transformace* daného úseku časové řady $X_1, \ldots, X_n$:

$$I(\lambda) = \left| \frac{1}{\sqrt{2\pi n}} \sum_{t=1}^{n} X_t \, e^{-it2\pi\lambda} \right|^2 .$$

Argument λ, označující frekvence, nabývá hodnot z intervalu $\langle -1/2, 1/2 \rangle$, resp. $\omega = 2\pi\lambda$ z intervalu $\langle -\pi, \pi \rangle$. Periodogram $I(\lambda)$ je sudá funkce. V důsledku toho se často zkoumá pouze pro $\lambda \in \langle 0, 1/2 \rangle$, resp. $\omega \in \langle 0, \pi \rangle$. Hodnoty periodogramu se obvykle počítají jen pro takzvané *Fourierovy frekvence* $\lambda_i = \frac{i}{n}$, resp. $\omega_i = \frac{2\pi i}{n}$, $i = 1, \ldots, [n/2]$. Periodogram je dobrým ukazatelem periodicit, neboť jestliže deterministická složka původní řady je periodická s frekvencí λ, resp. $\omega = 2\pi\lambda$, nabývá peridogram pro tuto frekvenci velkou hodnotu.

Prohlédněme si obrázek 28, kde je znázorněn periodogram počítaný z dat příkladu 75, od kterých byl odečten aritmetický průměr. Největší hodnotu nabývá pro $\lambda = 0.038$, což odpovídá $\omega = 0.244$.

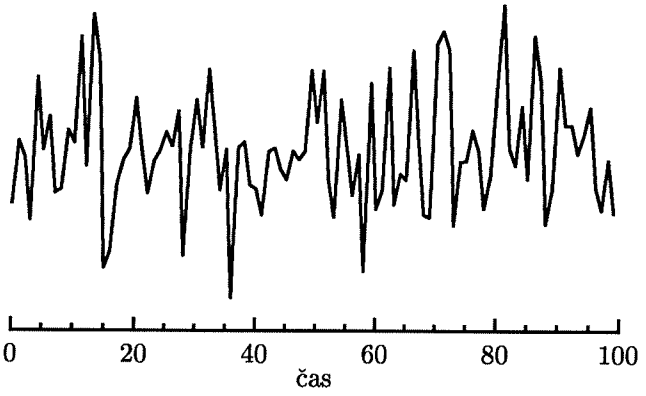


Obrázek 28.

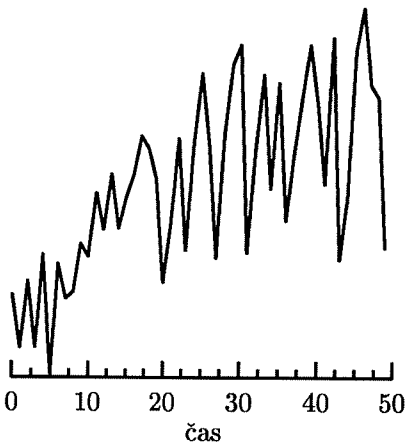
Pro výpočet hodnot periodogramu na počítačích se obvykle používá numerický postup, který se nazývá *rychlá Fourierova transformace - FFT.*

39. Stacionární časové řady

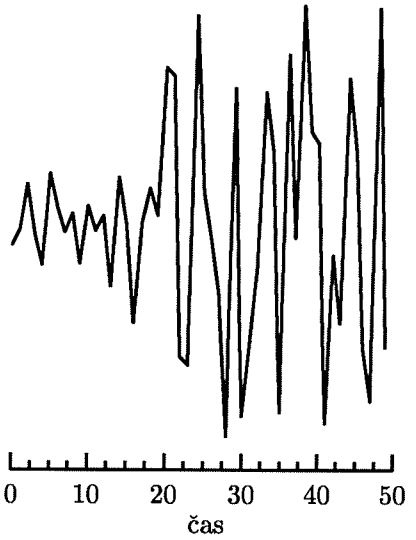
Velmi často nás zajímají řady, jejichž vlastnosti se z hlediska pravděpodobnosti v průběhu času nemění. Takovým řadám říkáme stacionární. Typická realizace stacionární časové řady je na obrázku 29. Na obrázcích 30 a 31 jsou naopak realizace nestacionárních řad.



Obrázek 29.



Obrázek 30.



Obrázek 31.

Při studiu časových řad se často spokojíme jen s požadavkem, aby se s časem neměnily alespoň základní charakteristiky řady, to jest:

1) všechny členy časové řady kolísají kolem stejné hodnoty μ se stejným rozptylem:

$$\mathrm{E}\, X_t = \mu, \quad \mathrm{Var}\, X_t = \sigma^2,$$

2) závislost mezi dvěma členy časové řady závisí pouze na časové vzdálenosti mezi těmito členy, nikoliv na jejich postavení uvnitř časové řady:

$$R(t, t+\tau) = \mathrm{cov}(X_t, X_{t+\tau}) = R(\tau),$$

$$\rho(t, t+\tau) = \mathrm{corr}(X_t, X_{t+\tau}) = \rho(\tau).$$

Připomeňme, že $\mathrm{cov}(X_t, X_{t+\tau})$ značí kovarianci veličin X_t, $X_{t+\tau}$ a $\mathrm{corr}(X_t, X_{t+\tau})$ značí korelaci X_t, $X_{t+\tau}$, viz článek 14. Funkci $R(\tau)$ se říká *autokovarianční funkce* řady $\{X_t\}$ a funkci $\rho(\tau)$ se říká *autokorelační funkce* řady $\{X_t\}$. Argumentu τ se říká *zpoždění*. Mezi autokorelační a autokovarianční funkcí je zřejmě vztah $\rho(\tau) = \frac{R(\tau)}{\sigma^2}$. Časové řady, které splňují podmínky 1) a 2) se nazývají *slabě stacionární*.

Poznamenejme, že o časových řadách na obrázcích 30 a 31 by bylo možno prokázat, že nejsou slabě stacionární. Střední hodnota časové řady z obrázku 30 $\mathrm{E}\,X_t = \mu(t)$ zřejmě roste s přibývajícím časem. Střední hodnota časové řady z obrázku 31 $\mathrm{E}\,X_t = \mu(t)$ se sice nejspíš nemění, ale s přibývajícím časem se zvětšuje rozptyl řady $\mathrm{Var}\,X_t = \sigma^2(t)$.

Za velmi obecných předpokladů lze parametry slabě stacionární řady μ, σ^2, autokovarianční funkci $R(\tau)$ a autokorelační funkci $\rho(\tau)$ odhadnout z realizace úseku časové řady $X_1, \ldots, X_n$ následovně:

$$\widehat{\mu} = \overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t, \quad \widehat{\sigma^2} = \sigma_n^2 = \frac{1}{n}\sum_{t=1}^{n}(X_t - \overline{X})^2,$$

$$\widehat{R}(\tau) = \frac{1}{n}\sum_{t=1}^{n-|\tau|}(X_{t+|\tau|} - \overline{X})(X_t - \overline{X}), \quad \widehat{\rho}(\tau) = \frac{\widehat{R}(\tau)}{\sigma_n^2}.$$

Funkci $\widehat{R}(\tau)$ se říká *výběrová autokovarianční funkce* a funkci $\widehat{\rho}(\tau)$ se říká *výběrová autokorelační funkce.*

Všimněme si, že autokovarianční funkce, autokorelační funkce, výběrová autokovarianční funkce a výběrová autokorelační funkce jsou sudé funkce. Často proto uvažujeme jejich hodnoty pouze pro kladné argumenty.

Poznámka

Poznamenejme, že vlastnosti odhadu autokovarianční, resp. autokorelační funkce se zhoršují s růstem absolutní hodnoty argumentu $|\tau|$. Není tedy vhodné odhadovat autokovarianční, resp. autokorelační funkci, pro všechny teoreticky možné argumenty $|\tau| \leq n-1$, nýbrž jen pro jejich určitou část.

40. Autoregresní posloupnosti

Uvažujme časovou řadu, jejíž členy splňují vztah:

$$X_t = a\,X_{t-1} + e_t, \quad t = \ldots, -2, -1, 0, 1, 2, 3, \ldots,$$

kde náhodné veličiny $\{e_t\}$ jsou nezávislé, stejně rozdělené, řídící se rozdělením $N(0, \sigma^2)$. Této časové řadě se říká *autoregresní posloupnost 1. řádu* a značí se $AR(1)$. Platí-li $|a| < 1$, pak shora popsaná časová řada je slabě stacionární.

Pro její autokovarianční funkci platí:

$$R(\tau) = \frac{\sigma^2}{1-a^2} \cdot a^{|\tau|}$$

a pro autokorelační funkci:

$$\rho(\tau) = a^{|\tau|}.$$

Speciálně $a = \rho(1)$. Odtud plyne, že nejjednodušším odhadem parametru a je $\widehat{a} = \frac{\widehat{R}(1)}{\widehat{R}(0)} = \widehat{\rho}(1)$.

Příklad 76.

	leden	únor	březen	duben	květen	červen
1.	69.2	53.6	424.0	92.1	238.0	111.5
2.	65.4	52.2	591.5	96.0	211.0	105.5
3.	62.5	47.1	494.5	96.0	192.0	100.0
4.	58.6	42.0	352.8	96.0	173.1	86.8
5.	58.6	44.0	260.6	100.0	161.8	83.4
6.	64.0	37.0	253.3	122.1	152.8	78.1
7.	75.4	34.0	218.0	148.1	146.0	74.8
8.	86.0	37.2	194.3	146.0	152.8	76.5
9.	84.0	56.5	173.1	132.9	171.0	76.5
10.	79.0	53.6	159.5	128.4	194.3	80.0
11.	73.5	63.9	152.8	130.6	187.4	80.0
12.	72.6	67.6	141.6	122.1	183.0	86.8
13.	73.5	66.3	126.3	113.5	220.5	85.0
14.	79.0	63.9	100.0	113.5	243.0	80.0
15.	90.0	50.6	70.0	109.5	245.5	71.5
16.	107.0	44.8	46.1	107.3	253.1	65.3
17.	116.0	45.2	53.6	101.8	243.1	63.9
18.	116.0	44.0	88.5	94.0	211.0	60.9
19.	107.0	45.3	81.7	90.4	185.0	66.3
20.	97.0	45.0	71.5	94.0	166.4	66.9
21.	87.0	46.5	80.0	149.0	148.1	68.4
22.	80.0	45.0	78.1	251.0	141.6	66.9
23.	76.0	60.2	88.5	496.0	137.3	62.2
24.	79.0	76.5	86.8	342.0	132.9	56.5
25.	78.0	60.9	81.7	294.4	126.3	52.2
26.	72.9	53.6	76.5	270.8	120.0	59.4
27.	69.0	78.1	73.3	303.0	113.5	53.6
28.	62.4	367.0	74.8	293.3	107.3	51.0
29.	49.0		83.0	303.0	111.5	46.1
30.	46.5		100.0	255.5	126.3	45.0
31.	49.6		92.1		115.8	

	červenec	srpen	září	říjen	listopad	prosinec
1.	40.3	78.1	355.7	336.0	51.0	92.1
2.	39.4	171.0	364.1	324.7	51.0	166.4
3.	38.4	128.4	660.0	268.0	51.0	141.6
4.	37.4	109.5	770.0	223.0	53.6	126.3
5.	40.4	122.1	778.0	192.0	63.9	103.4
6.	41.3	122.1	702.0	173.1	70.0	90.4
7.	43.2	132.9	563.0	189.7	62.2	83.4
8.	48.9	105.5	563.0	220.5	55.0	80.0
9.	61.4	105.5	698.0	361.3	51.0	76.5
10.	51.4	132.9	670.6	286.8	59.4	74.8
11.	46.4	130.6	552.3	240.5	58.0	71.5
12.	42.2	124.2	452.0	201.1	53.6	65.3
13.	39.4	109.5	361.3	194.3	53.6	65.3
14.	38.4	109.5	319.0	180.1	70.0	66.9
15.	34.6	161.0	408.8	161.8	70.0	71.5
16.	33.2	425.0	501.0	141.6	65.3	65.3
17.	34.6	589.0	350.2	122.1	60.9	65.3
18.	33.2	411.0	286.8	113.5	59.4	74.8
19.	30.2	463.0	240.5	109.5	56.5	96.0
20.	36.5	739.0	238.0	103.4	53.6	90.4
21.	52.8	573.9	250.7	101.8	53.6	80.0
22.	51.4	484.6	228.2	90.4	55.0	73.3
23.	101.8	428.5	213.3	92.1	55.0	71.5
24.	149.8	355.7	178.0	90.4	53.6	73.3
25.	111.2	484.6	155.0	85.0	56.5	98.0
26.	99.2	549.0	146.0	76.5	56.5	111.5
27.	73.4	552.3	220.0	66.9	63.9	107.3
28.	61.3	442.3	211.0	66.9	65.3	111.5
29.	54.2	370.0	206.1	58.0	68.4	101.8
30.	54.2	305.8	324.7	56.5	68.4	96.0
31.	57.0	308.2		51.0		111.5

Tabulka 48.

V tabulce 48 jsou uvedeny denní průtoky Vltavy v roce 1926 v místě dnešní přehrady Orlík. Odhadněte z těchto dat za předpokladu stacionarity autokorelační funkci časové řady denních průtoků pro argument $|\tau| \leq 24$. Odhadněte parametry μ a a, použijeme-li pro časovou řadu denních průtoků zobecněný autoregresní model 1. řádu:

$$X_t - \mu = a\,(X_{t-1} - \mu) + e_t.$$

Řešení:

Nejprve odhadněme autokorelační funkci $\rho(\tau)$ pomocí výběrové autokorelační funkce

$$\widehat{\rho}(\tau) = \frac{\sum_{i=1}^{n-|\tau|}(X_{i+|\tau|} - \overline{X})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

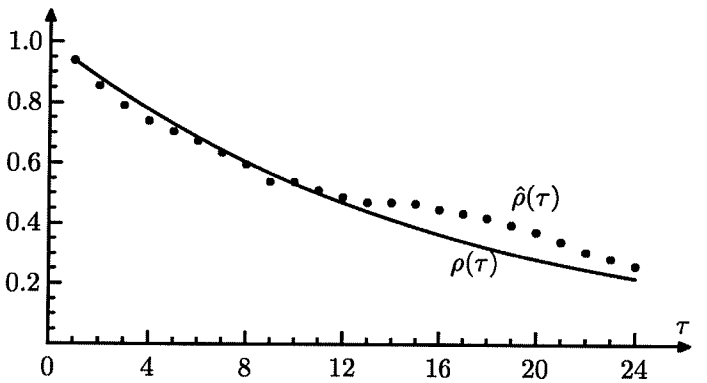
Vzhledem k symetrii (sudosti) funkce $\widehat{\rho}(\tau)$ je možno provést odhad jen pro kladné hodnoty argumentu.

zpoždění τ	odhad autokorelační funkce $\rho(\tau)$ $\widehat{\rho}(\tau)$	zpoždění τ	odhad autokorelační funkce $\rho(\tau)$ $\widehat{\rho}(\tau)$	zpoždění τ	odhad autokorelační funkce $\rho(\tau)$ $\widehat{\rho}(\tau)$
1	0.93892	9	0.56968	17	0.43478
2	0.85560	10	0.53904	18	0.42044
3	0.78972	11	0.51080	19	0.39684
4	0.73936	12	0.48810	20	0.37365
5	0.70447	13	0.47059	21	0.34149
6	0.67407	14	0.47049	22	0.30703
7	0.63485	15	0.46657	23	0.28557
8	0.59676	16	0.44794	24	0.26184

Tabulka 49.

Dále odhadněme parametry μ a a zobecněného autoregresního modelu 1. řádu:

$$\widehat{\mu} = \overline{x} = 148.757 \quad \text{a} \quad \widehat{a} = \widehat{\rho}(1) = 0.93892.$$



Obrázek 32.

Na obrázku 32 je pro srovnání zobrazena výběrová autokorelační funkce časové řady denních průtoků Vltavy v roce 1926 a teoretická autokorelační funkce autoregresní posloupnosti $AR(1)$

$$Y_t = 0.93892\, Y_{t-1} + e_t,$$

tj. funkce $\rho(\tau) = 0.93892^{\tau}$ pro $\tau = 1, 2, \ldots, 24$.

□

Autoregresní posloupnost 1. řádu je speciálním případem *autoregresní posloupnosti k-tého řádu* $AR(k)$:

$$X_t = a_1 X_{t-1} + \cdots + a_k X_{t-k} + e_t, \quad t = \ldots, -2, -1, 0, 1, 2, \ldots,$$

kde $\{e_t\}$ jsou nezávislé, stejně rozdělené náhodné veličiny, řídící se $N(0, \sigma^2)$. Autoregresních posloupností se často používá k modelování průtoků řek. Jestliže všechny kořeny polynomu $A(z) = z^k - a_1 z^{k-1} - \cdots - a_k$ leží uvnitř jednotkového kruhu $\{z \in C; |z| < 1\}$, pak je posloupnost $AR(k)$ slabě stacionární.

Pro odhad autoregresních koeficientů $a_1, \ldots, a_k$ je možno použít metody maximální věrohodnosti nebo metody nejmenších čtverců. Odhady metodou nejmenších čtverců získáme minimalizací výrazu

$$\min_{a_1,\ldots,a_k} \sum_{t=k+1}^{n} \left(X_t - (a_1 X_{t-1} + \cdots + X_{t-k})\right)^2.$$

Jednodušší metodou pro získání odhadů autoregresních koeficientů je hledat tyto odhady jako řešení lineární soustavy Yule-Walkerových rovnic, která lze maticově zapsat:

$$\begin{pmatrix} \widehat{R}(0) & \widehat{R}(1) & \ldots & \widehat{R}(k-1) \\ \widehat{R}(1) & \widehat{R}(0) & \ldots & \widehat{R}(k-2) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{R}(k-1) & \widehat{R}(k-2) & \ldots & \widehat{R}(0) \end{pmatrix} \cdot \begin{pmatrix} \widehat{a_1} \\ \widehat{a_2} \\ \vdots \\ \widehat{a_k} \end{pmatrix} = \begin{pmatrix} \widehat{R}(1) \\ \widehat{R}(2) \\ \vdots \\ \widehat{R}(k) \end{pmatrix}$$

Poznámka

Box a Jenkins, známí odborníci v teorii časových řad, ukázali, že pro případ, kdy n je velké a kořeny polynomu $A(z)$ neleží blízko jednotkové kružnice $\{z \in C, |z| = 1\}$, jsou rozdíly mezi odhady autoregresních koeficientů získaných různými metodami velmi malé. V případě, že kořeny polynomu $A(z)$ leží blízko jednotkové kružnice, vede řešení Yule-Walkerových rovnic ke špatným odhadům a je třeba použít maximálně věrohodné odhady.

Příklad 77.

Z 500 členů časové řady byla spočtena výběrová autokovarianční funkce $\widehat{R}(\tau)$ pro hodnoty argumentu $\tau = 0, 1, \ldots, 14$:

zpoždění τ	výběrová autokovarianční funkce $\widehat{R}(\tau)$	zpoždění τ	výběrová autokovarianční funkce $\widehat{R}(\tau)$	zpoždění τ	výběrová autokovarianční funkce $\widehat{R}(\tau)$
0	1.9838	5	0.6481	10	0.1715
1	0.4131	6	-0.2729	11	-0.0076
2	-1.2341	7	-0.5342	12	-0.1238
3	-0.7131	8	-0.0307	13	-0.1726
4	0.6067	9	0.3176	14	-0.0036

Tabulka 50.

Předpokládejme, že napozorovaná data jsou realizací úseku autoregresní posloupnosti $AR(2)$:

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + e_t, \quad t = \ldots, -2, -1, 0, 1, 2, 3, \ldots$$

Pomocí Yule-Walkerových rovnic odhadněte autoregresní koeficienty a_1, a_2.

Řešení:

V případě autoregresní posloupnosti $AR(2)$ mají Yule-Walkerovy rovnice pro odhad autoregresních koeficientů tvar:

$$\begin{aligned} \widehat{R}(0)\,\widehat{a_1} &+ \widehat{R}(1)\cdot\widehat{a_2} &&= \widehat{R}(1), \\ \widehat{R}(1)\,\widehat{a_1} &+ \widehat{R}(0)\cdot\widehat{a_2} &&= \widehat{R}(2). \end{aligned}$$

Odtud

$$\widehat{a_1} = \frac{\widehat{R}(1)\big(\widehat{R}(0) - \widehat{R}(2)\big)}{\widehat{R}(0)^2 - \widehat{R}(1)^2} = 0.3531, \quad \widehat{a_2} = \frac{\widehat{R}(0)\widehat{R}(2) - \widehat{R}(1)^2}{\widehat{R}(0)^2 - \widehat{R}(1)^2} = -0.6956.$$

□

V příkladech 76 a 77 jsme při odhadování autoregresních koeficientů předpokládali, že známe řád autoregresních posloupnosti. Ve většině praktických případů však řád autoregresní posloupnosti předem neznáme a je třeba ho odhadnout. V poslední době byla vytvořena pro odhadování řádu autoregresní posloupností celá řada informačních kritérií, z nichž nejjednodušší je Akaikeho kritérium, které spočívá v minimalizaci výrazu:

$$AIC(k) = n \ln\big(\widehat{R}(0) - \widehat{a_1}\,\widehat{R}(1) - \cdots - \widehat{a_k}\,\widehat{R}(k)\big) + 2k,$$

kde $\widehat{a_1}, \ldots, \widehat{a_k}$ jsou maximálně věrohodné odhady autoregresních koeficientů a $\widehat{R}(0), \ldots, \widehat{R}(k)$ jsou hodnoty výběrové autokovarianční funkce. Ve většině případů můžeme v Akaikeho kritériu použít místo maximálně věrohodných odhadů odhady získané z Yule-Walkerových rovnic, viz shora uvedená poznámka.

Příklad 78.

Pomocí Akaikeho kritéria odhadněte řád autoregresní posloupnosti z příkladu 77.

Řešení:

Uvažujme hodnotu Akaikeho kritéria $AIC(k)$ pro $k = 1, 2, 3, 4$:

$k=1$	$\widehat{a_1}$	=	0.2082	$AIC(1)$	= 323.3
$k=2$	$\widehat{a_1}$	=	0.3531	$AIC(2)$	= -6.4
	$\widehat{a_2}$	=	-0.6956		
$k=3$	$\widehat{a_1}$	=	0.3602	$AIC(3)$	= -4.4
	$\widehat{a_2}$	=	-0.6992		
	$\widehat{a_3}$	=	0.0102		
$k=4$	$\widehat{a_1}$	=	0.3602	$AIC(4)$	= -2.4
	$\widehat{a_2}$	=	-0.7018		
	$\widehat{a_3}$	=	0.0115		
	$\widehat{a_4}$	=	-0.0037		

Tabulka 51.

Uvažujeme-li pouze řád $k = 1, 2, 3, 4$, pak Akaikeho kritérium $AIC(k)$ nabývá nejmenší hodnotu pro $k = 2$. Použijeme-li Akaikeho kritérium, pak bychom mezi autoregresními posloupnostmi řádu $k = 1, 2, 3, 4$ vybrali jako nejlepší model pro naše data $AR(2)$.

□

Část IX. Simulační metody

41. Statistické modelování a metody Monte Carlo

V mnohých úlohách, týkajících se vyšetřování vlastností stavebních konstrukcí, ve kterých sledované parametry jsou náhodné veličiny nebo náhodné procesy, je velmi obtížné najít analytické řešení. Takové úlohy vznikají například při vyšetřování únavové pevnosti namáhaných částí konstrukce při náhodném dynamickém zatížení, jako jsou přejezdy vozidel po mostních konstrukcích, nápory větru na vysokopodlažní budovy apod. Příkladem úlohy z hydrologie, kde je rovněž obtížné najít analytické řešení, je zkoumání zabezpečenosti pitnou vodou při různých způsobech odběru, kde vstupními parametry jsou náhodné přítoky.

V takových případech je jednou z možností, jak získat informace o náhodné reakci konstrukce a její spolehlivosti, viz Augusti et al. (1993), či informace o chování vodohospodářské soustavy, viz Kos and Zeman (1976), apod. danou úlohu simulovat.

Typické schéma simulační úlohy je následující:

1) Sestavit matematický model adekvátní reálné situaci.

2) Připravit plán experimentu a generovat náhodnou složku, která vstupuje do modelu.

3) Dosadit do matematického modelu a zjistit výstupní hodnoty.

4) Zpracovat a interpretovat výsledky.

Simulační metody jsou metody, které modelují reálné jevy náhodné povahy. V souvislosti s pojmem simulační metody se též vyskytuje pojem metody Monte Carlo. *Metody Monte Carlo* jsou metody řešení matematických úloh pomocí náhodných pokusů, realizovaných zpravidla na počítačích.

Typická úloha, kterou lze řešit metodami Monte Carlo, je následující. Uvažujme vstupní náhodné veličiny $Y_1, \ldots, Y_r$ se známou hustotou $f(y_1, \ldots, y_r)$. Zajímá nás rozdělení výstupní náhodné veličiny $Z = \phi(Y_1, \ldots, Y_r)$, kde $\phi(y_1, \ldots, y_r)$ je známá funkce r proměnných. Jestliže nedovedeme řešit tuto úlohu analyticky, můžeme ji řešit pomocí metod Monte Carlo tak, že opakovaně pomocí náhodných čísel simulujeme hodnoty vstupního náhodného vektoru $(Y_1, \ldots, Y_r)$, dosadíme vždy do ϕ a registrujeme „realizace" Z. Tento postup mnohokrát opakujeme a dostaneme empirické rozdělení Z obvykle ve formě histogramu četností.

Podstata simulačních metod tkví v mnohonásobném opakování náhodného pokusu, přičemž toto opakování se provádí uměle. Podmínkou úspěchu je, mimo jiné, možnost mnohokrát kvalitně pokus opakovat. K tomu je zapotřebí dostatečné množství náhodných čísel. Obvykle vycházíme z náhodných čísel z rovnoměrného rozdělení. Pod pojmem *náhodná čísla z rovnoměrného rozdělení* $R(0,1)$ rozumíme konečnou posloupnost čísel z intervalu $\langle 0, 1\rangle$, kterou lze považovat za realizaci náhodného výběru z rovnoměrného rozdělení $R(0,1)$.

Dříve se pro vyhledávání náhodných čísel z $R(0,1)$ používaly tzv. *tabulky náhodných čísel.* V historii jich byla publikována celá řada. Některé byly vytvořeny na základě sestav ze sčítání lidu, jiné byly pořízeny pomocí „elektronické rulety" apod.

S rozvojem počítačů se v dnešní době většinou užívá *generátorů náhodných čísel*, založených na aritmetických procedurách, které vytvářejí náhodná čísla pomocí speciálních rekurentních vzorců. Těmto číslům říkáme někdy také *pseudonáhodná čísla*, neboť vznikají sice nenáhodným způsobem, ale mají přitom vlastnosti náhodných čísel.

Nejčastěji užívanými generátory jsou *kongruenční generátory.* V tomto případě je posloupnost pseudonáhodných čísel z intervalu $\langle 0, M)$ nejčastěji vytvářena pomocí rekurentního vztahu

$$Y_{n+1} = a\ Y_n + b \pmod{M},$$

kde a, b, M jsou vhodné konstanty. Výraz „mod M" (čti modulo M) znamená, že číslo Y_{n+1} je zbytek po dělení čísla $a\,Y_n + b$ číslem M. Pseudonáhodná čísla X_n, $n = 1, 2, \ldots$ z intervalu $\langle 0, 1)$ získáme transformací $X_n = Y_n/M$, $n = 1, 2, \ldots$.

Je patrné, že takto vytvořená posloupnost pseudonáhodných čísel nemusí být skutečnou realizací náhodného výběru z $R(0,1)$. Její hlavní nevýhodou je periodičnost, to znamená, že se generovaná čísla po určité době začnou opakovat. Generátory musí být tedy vytvořeny tak, aby perioda posloupnosti poskytovaných čísel byla co největší. Na generátor máme celou řadu požadavků týkajících se statistických vlastností pseudonáhodných čísel. Požadujeme samozřejmě, aby se pro velký počet generovaných pseudonáhodných čísel histogram dobře shodoval s hustotou rozdělení $R(0,1)$. Jiný velmi důležitý požadavek je, aby sousední nebo vzdálenější členy posloupnosti nebyly korelované apod. Vlastnosti kongruenčních generátorů závisí na volbě konstant a, b, M a počáteční hodnotě X_0, a tato volba zase úzce souvisí s typem počítače, na kterém je generátor používán. Pro testování vhodnosti určitého generátoru existuje celá řada statistických testů. Vytvoření konkrétních generátorů je záležitostí numerické matematiky a nebudeme se jím podrobněji zabývat.

Nadále budeme předpokládat, že máme k dispozici dokonalý generátor náhodných čísel z rozdělení $R(0,1)$.

42. Obecné metody pro generování náhodného výběru z daného rozdělení

V článku 41 jsme hovořili o generování náhodných čísel z rozdělení $R(0,1)$. V praxi však často potřebujeme generovat náhodná čísla i z jiných rozdělení.

Obecnou metodou pro vytvoření náhodných čísel z určitého rozdělení je *metoda inverzní transformace*, která se opírá o následující tvrzení.

Nechť $F(x)$ je zleva spojitá distribuční funkce a nechť funkce $F^{-1}(x)$ je definována předpisem

$$F^{-1}(y) = \sup\{x, F(x) \leq y\} \quad \text{pro} \quad 0 < y < 1.$$

Nechť náhodná veličina Y má rovnoměrné rozdělení $R(0,1)$. Potom náhodná veličina $X = F^{-1}(Y)$ má rozdělení s distribuční funkcí $F(x)$.

Poznámka

Poznamenejme, že je-li distribuční funkce $F(x)$ rostoucí a spojitá, pak $F^{-1}(y)$ je inverzní funkce k $F(x)$.

Postup při generování náhodných čísel $X_1, \ldots, X_n$ z rozdělení s distribuční funkcí $F(x)$ metodou inverzní transformace je následující:

1. Vygenerujeme náhodná čísla $Y_1, \ldots, Y_n$ z rozdělení $R(0,1)$.
2. Proveďme transformaci $X_1 = F^{-1}(Y_1), \ldots, X_n = F^{-1}(Y_n)$.

Výhody metody inverzní transformace spočívají v její universálnosti, neboť teoreticky umožňuje generovat náhodná čísla z libovolného rozdělení. V případě, že lze snadno najít explicitní tvar $F^{-1}(y)$, je její použití jednoduché a často dává nejlepší výsledky. V případě, že výpočet hodnot funkce $F^{-1}(y)$ je velmi komplikovaný a časově náročný, používají se obvykle jiné metody.

Pro generování náhodného čísla X se spojitým rozdělením s hustotou $g(x)$ se často používá též *zamítací metoda.*

Nechť existuje reálná funkce $g_1(x)$ taková, že $g(x) \leq g_1(x)$, $x \in R^1$ a $\int_{-\infty}^{\infty} g_1(x)\,dx < \infty$. Označme $G_1 = \int_{-\infty}^{\infty} g_1(x)\,dx$. Předpokládejme, že dovedeme efektivně generovat náhodné číslo z rozdělení s hustotou $f_Z(x) = g_1(x)/G_1$. Náhodné číslo X z rozdělení s hustotou $g(x)$ získáme následujícím způsobem:

1. Generujeme číslo Z z rozdělení s hustotou $f_Z(x)$.
2. Generujeme náhodné číslo V z rozdělení $R(0,1)$.
3. Je-li $g_1(Z) \cdot V \leq g(Z)$, položíme $X = Z$. Je-li $g_1(Z) \cdot V > g(Z)$, přejdeme k bodu 1. (Dvojici (V, Z) „zamítáme".)

Poznámka

Poznamenejme, že zamítací metoda je tím efektivnější, čím je hodnota G_1 bližší jedné.

Zamítací metodu lze například použít v případě, že chceme generovat náhodné číslo X z rozdělení s hustotou $g(x)$, která je ohraničená a nulová vně ohraničeného intervalu (a, b). Označíme-li $M = \sup\{g(x), a < x < b\}$, potom zřejmě $g(x) \leq M$. Náhodné číslo X pak generujeme následovně:

1. Generujeme číslo Z z rozdělení s hustotou $R(a, b)$ tak, že generujeme náhodné číslo U z $R(0,1)$ a položíme $Z = (b-a) \cdot U + a$.
2. Generujeme náhodné číslo V z rozdělení $R(0,1)$.

3. Je-li $M \cdot V \leq g(Z)$, pak $X = Z$. Je-li $M \cdot V > g(Z)$, pak přejdeme k bodu 1. (Dvojici (V, Z) „zamítáme".)

Poznámka

Je-li součin $(b-a) \cdot M$ výrazně větší než jedna, je tato metoda neefektivní, viz shora uvedená poznámka.

43. Příklady algoritmů pro generování náhodných čísel z některých rozdělení

Algoritmů pro generování náhodných čísel z často užívaných rozdělení je známa celá řada. Do tohoto článku skript jsme vybrali některé velmi jednoduché algoritmy, které jsou proto také s oblibou často používané.

Normální rozdělení

Nejprve poznamenejme, že náhodné číslo X z rozdělení $N(\mu, \sigma^2)$ získáme z náhodného čísla Y z rozdělení $N(0,1)$ transformací

$$X = \sigma \cdot Y + \mu$$

Z tohoto důvodu se budeme dále zabývat pouze generováním náhodných čísel z $N(0,1)$.

Náhodná čísla z normálního rozdělení se nejčastěji generují pomocí centrální limitní věty aplikované na součty nezávislých stejně rozdělených náhodných veličin řídících se rozdělením $R(0,1)$. Je-li $X \sim R(0,1)$, pak zřejmě $\mathrm{E}\, X = 1/2$ a $\mathrm{Var}\, X = 1/12$. Uvažujeme-li náhodný výběr $X_1, \ldots, X_n$ z rozdělení $R(0,1)$, pak $\mathrm{E} \sum X_i = n/2$ a $\mathrm{Var} \sum X_i = n/12$. Podle centrální limitní věty přitom platí, že pro velká n má náhodná veličina

$$Z = \frac{\sum X_i - n/2}{\sqrt{n/12}}$$

přibližně normální rozdělení $N(0,1)$. Za prakticky vyhovující se považuje $n = 12$. Náhodné číslo z rozdělení $N(0,1)$ se generuje následovně:

1. Generujeme posloupnost náhodných čísel $X_1, \ldots, X_{12}$ z rozdělení $R(0,1)$.
2. Spočteme $Z = \sum_{i=1}^{12} X_i - 6$.

Je zřejmé, že takto nagenerované číslo Z se pohybuje v intervalu $\langle -6, 6 \rangle$, a tudíž je tento způsob vyhovující v případech, kdy není podstatné chování „chvostů" normálního rozdělení. Chceme-li lépe vystihnout toto chování, je třeba použít nějaký jiný generátor. Jednou z možností je například generovat náhodná čísla z $N(0,1)$ následujícím způsobem:

1. Generujeme dvě náhodná čísla U a V z rozdělení $R(0,1)$.
2. Transformací

$$X_1 = \sqrt{-2 \ln U} \cdot \sin(2\pi V) \quad \text{a} \quad X_2 = \sqrt{-2 \ln U} \cdot \cos(2\pi V)$$

získáme dvě náhodná čísla X_1 a X_2 z normálního rozdělení $N(0,1)$.

Dvojrozměrné normální rozdělení

Náhodný vektor $(X,Y)'$ z dvojrozměrného normálního rozdělení $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$, respektive $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, kde

$$\boldsymbol{\mu}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}$$

nagenerujeme následujícím způsobem:

1. Generujeme náhodný vektor $(U,V)'$ z dvojrozměrného normálního rozdělení $N(0,0,1,1,0)$, tj. jinak řečeno dvě nezávisle rozdělené náhodné veličiny z $N(0,1)$.
2. Transformací

$$\begin{aligned}W &= U,\\ Z &= \rho\cdot U+\sqrt{1-\rho^2}\cdot V,\end{aligned}$$

získáme náhodný vektor z dvojrozměrného rozdělení $N(0,0,1,1,\rho)$.

3. Transformací

$$\begin{aligned}X &= \sigma_1\cdot W+\mu_1,\\ Y &= \sigma_2\cdot Z+\mu_2,\end{aligned}$$

získáme náhodný vektor z žádaného dvojrozměrného rozdělení $N(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$.

Logaritmicko-normální rozdělení

Náhodné číslo X z rozdělení $LN(\mu,\sigma^2,x_0)$ získáme z náhodného čísla Y z normálního rozdělení $N(\mu,\sigma^2)$ transformací:

$$X=e^Y+x_0.$$

Exponenciální rozdělení

Nejprve opět poznamenejme, že náhodné číslo X z exponenciálního rozdělení $E(\delta)$ s hustotou

$$\begin{aligned}f(x;\delta) &= \frac{1}{\delta}\cdot e^{-(1/\delta)x} && \text{pro}\quad x>0,\\ &= 0 && \text{pro}\quad x\le 0\end{aligned}$$

lze získat z náhodného čísla Y z rozdělení $E(1)$ transformací:

$$X=\delta\cdot Y.$$

Náhodná čísla z exponenciálního rozdělení $E(1)$ získáváme nejčastěji metodou inverzní transformace. Distribuční funkce exponenciálního rozdělení $E(1)$ splňuje vztah

$$\begin{aligned} F(x) &= 1 - e^{-x} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0, \end{aligned}$$

a tedy $F^{-1}(y) = -\ln(1-y)$ pro $0 < y < 1$. Z tvrzení o inverzní transformaci vyplývá, že má-li náhodná veličina Y rovnoměrné rozdělení $R(0,1)$, pak náhodná veličina $F^{-1}(y) = -\ln(1-Y)$ má exponenciální rozdělení $E(1)$. Protože však náhodná veličina $1-Y$ má rovněž $R(0,1)$, postupujeme při generování náhodného čísla X z rozdělení $E(1)$ tak, že

1. generujeme náhodné číslo U z rozdělení $R(0,1)$,
2. náhodné číslo X z rozdělení $E(1)$ získáme transformací

$$X = -\ln U.$$

Gamma rozdělení

Nejprve uveďme, že náhodné číslo X z dvojparametrického rozdělení $G(\alpha, \kappa)$ s hustotou

$$\begin{aligned} f(x;\alpha,\kappa) &= \frac{\alpha^{\kappa} x^{\kappa-1} e^{-\alpha x}}{\Gamma(\kappa)} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0, \end{aligned}$$

lze získat z náhodného čísla Y z rozdělení $G(1,\kappa)$ transformací $X = Y/\alpha$.

Náhodné číslo $\widetilde{X}$ z trojparametrického gamma rozdělení, kterému se často říká též Pearsonovo rozdělení typu III s hustotou

$$\begin{aligned} f(x;\alpha,\kappa,\nu) &= \alpha^{\kappa} \frac{(x-\nu)^{\kappa-1}}{\Gamma(\kappa)} \exp\{-\alpha(x-\nu)\} && \text{pro} \quad x > \nu, \\ &= 0 && \text{pro} \quad x \leq \nu, \end{aligned}$$

$\alpha, \kappa > 0$ a $\nu \in R^1$, lze získat z náhodného čísla Y z rozdělení $G(1,\kappa)$ transformací $\widetilde{X} = Y/\alpha + \nu$.

Odtud plyne, že se stačí zabývat pouze generováním náhodných čísel z rozdělení $G(1,\kappa)$.

Jestliže hodnota parametru κ gamma rozdělení $G(1,\kappa)$ je přirozené číslo n, pak tomuto rozdělení říkáme Erlangovo. Hustota Erlangova rozdělení je tedy dána předpisem

$$\begin{aligned} f(x;\, n) &= \frac{x^{n-1} e^{-x}}{\Gamma(n)} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0, \end{aligned}$$

kde n je přirozené číslo.

Uvažujme náhodný výběr $X_1, \ldots, X_n$ z exponenciálního rozdělení $E(1)$, pak náhodná veličina $X = \sum_{i=1}^{n} X_i$ se řídí Erlangovým rozdělením s parametrem n. Na tomto tvrzení je založen algoritmus pro generování náhodného čísla X z Erlangova rozdělení s parametrem n.

1. Generujme posloupnost náhodných čísel $Y_1, \ldots, Y_n$ z rozdělení $R(0,1)$.
2. Transformací

$$X = -\ln(Y_1 \cdot \ldots \cdot Y_n)$$

získáme náhodné číslo z Erlangova rozdělení s parametrem n.

Přejděme nyní ke generování náhodných čísel z gamma rozdělení $G(1,\kappa)$, kde $\kappa > 0$. Nejprve uveďme případ, kde $0 < \kappa < 1$. Náhodné číslo z gamma rozdělení $G(1,\kappa)$, kde parametr $\kappa \in (0,1)$ lze generovat pomocí zamítací metody, přičemž položíme

$$\begin{aligned} g(x) &= \frac{x^{\kappa-1}e^{-x}}{\Gamma(\kappa)} && \text{pro} \quad x > 0, \\ &= 0 && \text{pro} \quad x \leq 0, \\ g_1(x) &= \frac{x^{-1}}{\Gamma(\kappa)} && \text{pro} \quad 0 < x \leq 1, \\ &= \frac{e^{-x}}{\Gamma(\kappa)} && \text{pro} \quad x > 1, \\ &= 0 && \text{pro} \quad x \leq 0. \end{aligned}$$

Zřejmě $g(x) \leq g_1(x)$. Dále $G_1 = \int_0^\infty g_1(x)\,dx = (1/\kappa + 1/e)/\Gamma(\kappa)$. Označme $K = 1/\kappa + 1/e$. Náhodné číslo z rozdělení s hustotou $g_1(x)/G_1$ můžeme generovat metodou inverzní transformace, neboť odpovídající distribuční funkce je

$$\begin{aligned} F_1(x) &= \frac{x^\kappa}{(K \cdot \kappa)} && \text{pro} \quad 0 < x \leq 1, \\ &= \frac{K - e^{-x}}{K} && \text{pro} \quad x > 1 \end{aligned}$$

a pro funkci $F_1^{-1}(y)$ platí:

$$\begin{aligned} F_1^{-1}(y) &= (K \cdot \kappa \cdot y)^{1/\kappa} && \text{pro} \quad 0 < y \leq (K \cdot \kappa)^{-1}, \\ &= -\ln(K - K \cdot y) && \text{pro} \quad y > (K \cdot \kappa)^{-1}. \end{aligned}$$

Algoritmus pro generování náhodného čísla X z gamma rozdělení $G(1,\kappa)$ s parametrem $\kappa \in (0,1)$ lze popsat následovně:

1. Generujeme náhodné číslo W z rozdělení $R(0,1)$ a položíme $Z = F_1^{-1}(W)$.
2. Generujeme náhodné číslo U z rozdělení $R(0,1)$.
3. Je-li $g_1(Z)\,U \leq g(Z)$, položíme $X = Z$. Je-li $g_1(Z)\,U > g(Z)$, přejdeme k bodu 1. (Dvojici (Z,U) „zamítáme“.)

Přejeme-li si generovat náhodné číslo X z gamma rozdělení $G(1,\kappa)$ s parametrem $\kappa > 1$, využijeme následující tvrzení. Nechť náhodná veličina V se řídí gamma rozdělením $G(1,\beta)$ a náhodná veličina W rozdělením $G(1,\gamma)$, přičemž jsou navzájem nezávislé, pak náhodná veličina $V+W$ se řídí gamma rozdělením $G(1,\beta+\gamma)$. Odtud plyne jeden ze způsobů generování náhodného čísla X z gamma rozdělení $G(1,\kappa)$, kde $\kappa > 1$.

1. Generujeme náhodné číslo Y z rozdělení $G(1,[\kappa])$ podle algoritmu pro generování čísla z Erlangova rozdělení.

2. Generujeme náhodné číslo Z z rozdělení $G(1,\kappa-[\kappa])$ podle shora popsaného algoritmu pro generování náhodných čísel z gamma rozdělení s parametrem menším než jedna.

3. Položíme $X = Y + Z$.

χ^2 rozdělení

Rozdělení χ^2 o n stupních volnosti je totéž rozdělení jako $G(1/2,\, n/2)$. Je-li n sudé, lze tedy použít algoritmu pro generování náhodného čísla z Erlangova rozdělení. Je-li n liché, lze použít algoritmu pro generování náhodného čísla z gamma rozdělením s obecným parametrem.

Jinou možností, jak generovat náhodná čísla z rozdělení χ^2 o n stupních volnosti, je vytvářet je jako součet n kvadrátů čísel naganerovaných z rozdělení $N(0,1)$. Algoritmus je v tomto případě následující:

1. Generujeme čísla $X_1,\ldots,X_n$ z rozdělení $N(0,1)$.

2. Náhodné číslo X z rozdělení χ^2 o n stupních volnosti vytvoříme transformací

$$X = X_1^2 + X_2^2 + \cdots + X_n^2.$$

t rozdělení

Algoritmus pro generování náhodného čísla X z t rozdělení o n stupních volnosti lze popsat takto:

1. Generujeme náhodné číslo U z rozdělení $N(0,1)$.

2. Generujeme náhodné číslo V z rozdělení χ^2 o n stupních volnosti.

3. Náhodné číslo X z t rozdělení o n stupních volnosti získáme transformací

$$X = \frac{U\sqrt{n}}{\sqrt{V}}.$$

F rozdělení

Algoritmus pro generování náhodného čísla X z rozdělení F o n_1 a n_2 stupních volnosti lze popsat následovně:

1. Generujeme náhodné číslo W z rozdělení χ^2 o n_1 stupních volnosti.
2. Generujeme náhodné číslo Z z rozdělení χ^2 o n_2 stupních volnosti.
3. Náhodné číslo X z F rozdělení o n_1 a n_2 stupních volnosti získáme transformací

$$X = \frac{W/n_1}{Z/n_2}.$$

Literatura

Anděl J. (1976). *Statistická analýza časových řad.* SNTL, Praha.

Anděl J. (1976). *Statistická analýza časových řad.* SNTL, Praha. 40

Anděl J. (1978). *Matematická statistika.* SNTL, Praha.

Antoch J. a Vorlíčková D. (1992). *Vybrané metody statistické analýzy dat.* Academia, Praha.

Augusti G., Baratta A. a Casciata F. (1984). *Probabilistic Methods in Structural Engineering.* Chapman and Hall, London. 129

Beneš V. a Dohnal G. (1993). *Pravděpodobnost a matematická statistika – doplňkové skriptum.* Vydavatelství ČVUT, Praha.

Cipra T. (1986). *Analýza časových řad s aplikacemi v ekonomii.* SNTL/ALFA, Praha.

Cramér H. (1946). *Mathematical methods of statistics.* Princeton Univ. Press, Princeton.

Egermayer F. a Boháč M. (1984). *Statistika pro techniky.* SNTL, Praha.

Hála M. a Jarušková D. (1999). *Pravděpodobnost a matematická statistika 11 – tabulky.* Vydavatelství ČVUT, Praha.

Hátle J. a Likeš J. (1972). *Základy počtu pravděpodobnosti a matematické statistiky.* SNTL/ALFA, Praha.

Jarušková D. a Hála M. (2000). *Pravděpodobnost a matematická statistika 12 – příklady.* Vydavatelství ČVUT, Praha.

Kos Z. a Zeman V. (1976). *Vodohospodářské soustavy ve Směrném vodohospodářském plánu.* Státní zemědělské nakladatelství, Praha. 129

Likeš J. a Machek J. (1982). *Teorie pravděpodobnosti.* SNTL, Praha.

Likeš J. a Machek J. (1983). *Matematická statistika.* SNTL, Praha. 96, 104

Nacházel K. (1986). *Teorie odhadu v hydrologii a ve vodním hospodářství.* Academia, Praha.

Nacházel K. (1986). *Teorie odhadu v hydrologii a ve vodním hospodářství.* Academia, Praha. 74

Rao R. C. (1978). *Lineární metody statistické indukce a jejich aplikace.* Academia, Praha.

Rényi A. (1972). *Teorie pravděpodobnosti.* Academia, Praha.

Svěšnikov A. A. (1971). *Sbírka úloh z teorie pravděpodobnosti, matematické statistiky a teorie náhodných funkcí.* SNTL, Praha.

Zvára K. (1989). *Regresní analýza.* Academia, Praha.

Prof. RNDr. Daniela Jarušková, CSc.

PRAVDĚPODOBNOST A MATEMATICKÁ STATISTIKA

Vydalo České vysoké učení technické v Praze,
Česká technika – nakladatelství ČVUT, Thákurova 1, 160 41 Praha 6
v roce 2015 jako svou 11 706. publikaci.

Vytiskla Česká technika – nakladatelství ČVUT, výroba, Zikova 4, 166 36 Praha 6
138 stran
1. dotisk 3. vydání. Náklad 100 výtisků. Rozsah 7,91 AA, 8,24 VA.

139 = VAKÁT 140 = TRAŤ