Kapitola 1

Základy pravděpodobnosti

1.1 Úvod

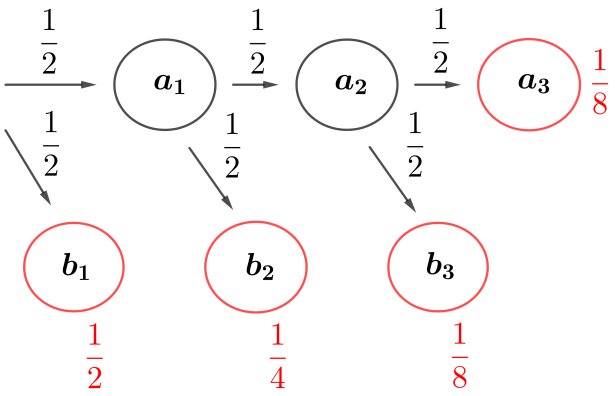
Náhodné jevy znali lidé od pradávna a využívali je nejprve zejména jako zdroj zábavy. Hrací kostky, karty i další podobné hry vedly pak k formulacím prvních pravděpodobnostních úloh. Uvedeme si jednu velmi starou úlohu, která byla zformulována v italském rukopise již v roce 1380 a kterou nejspíše přivezli do Itálie Arabové.

Příklad 1.1 *Dva hráči A a B spolu hrají sérii partií, které nemohou skončit remízou. Předpokládáme, že jednotlivé hry jsou spravedlivé, tj. každý hráč má stejnou šanci, že vyhraje. Dále předpokládejme, že jednotlivé hry jsou na sobě nezávislé, tj. v další hře nehraje žádnou roli to, jak dopadla hra předchozí. Hráči hrají o určitou částku, kterou získá ten, který první vyhraje* 6 *partií. Hráči však museli hru prerušit za stavu* $3 : 5$ *pro hráče A (tj. ve chvíli, kdy měl hráč A na kontě tři vítězství a hráč B pět vítězství). V jakém poměru si mají hráči rozdělit výhru?*

Řešení:
Uvedeme řešení, které je založené na metodě popsané v korespondenci mezi Pascalem a Fermatem. Uvažujme situaci, že budeme hrát ještě tři partie, a to bez ohledu na to, zda v jejich průběhu některý z hráčů dosáhl potřebných šesti vítězství. Označíme písmenem a vítězství hráče A a písmenem b vítězství hráče B. Pak máme následujících osm možností, jak může hra pokračovat:

$$aaa \quad aab \quad aba \quad baa \quad abb \quad bab \quad bba \quad bbb$$

Jelikož jsou všechny uvedené varianty stejně pravděpodobné, přičemž jen první varianta vede k výhře hráče A a v ostatních případech vyhraje B, je spravedlivý poměr rozdělení výhry $1 : 7$.



Obrázek 1.1: Zobrazení možných průběhů hry, kde a_i a b_i jsou jevy, že v i-tém kole vyhrál hráč A, resp. B. Červeně jsou zobrazeny situace, kdy hra končí.

Ukažme si i jiné řešení, které lze graficky zobrazit pomocí pravděpodobnostního stromu, viz obrázek 1.1.1. Označme a_i a b_i jevy, že v i-tém kole vyhrál hráč A, resp. B. Hráč A vyhraje celou hru pouze v případě tří vítězství v řadě, tedy jde o jev $a_1 \cap a_2 \cap a_3$. Pravděpodobnost tohoto jevu je $P(a_1 \cap a_2 \cap a_3) = \frac{1}{8}$. Pravděpodobnost výhry hráče B je $P(b_1) + P(a_2 \cap b_2) + P(a_1 \cap a_2 \cap b_2) = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$. Spravedlivý poměr rozdělení výhry je tedy $1 : 7$.

O tom, že pravděpodobnost byla i pro věhlasné matematiky dlouho obtížná, svědčí i následující příběh. D'Alembert (1717-1783) řešil otázku, jaká je pravděpodobnost, že při dvou hodech jednou mincí padne alespoň jednou líc. Při řešení došel k chybnému výsledku, že pravděpodobnost je $\frac{2}{3}$, neboť máme tři možnosti (2x líc, 1x líc, 0x líc), z nichž dvě odpovídají situaci, jejíž pravděpodobnost hledáme. Nesprávnost této úvahy spočívá v tom, že uvedené možnosti nejsou stejně pravděpodobné, neboť první možnost lze získat pouze kombinací líc-líc, zatímco druhá varianta nastane jak v případě kombinace líc-rub, tak v případě kombinace rub-líc, a tedy je dvakrát pravděpodobnější. Řešíme-li úlohu správně, dojdeme tedy k závěru, že hledaná pravděpodobnost je $\frac{3}{4}$, neboť máme čtyři stejně pravděpodobné možnosti (líc-líc, líc-rub, rub-líc a rub-rub), z nichž tři odpovídají situaci, jejíž pravděpodobnost hledáme.

Většina takových úloh se řeší pomocí tzv. **klasické pravděpodobnosti**, kterou známe ze střední školy pod heslem „pravděpodobnost jevu je podíl počtu příznivých situací ku počtu všech situací, které mohou nastat". K jejímu

popisu však můžeme použít i obecnou axiomatickou definici pravděpodobnosti, kterou zavedl až Kolmogorov (1903-1987) a která se používá dodnes. K tomu budeme potřebovat následující pojmy.

Mějme nějakou množinu Ω, jejíž prvky budeme značit symbolem ω_i, kde $i \in I$, I je indexová množina. Prvkům ω_i budeme říkat **elementární jevy** a jsou to všechny možné výsledky náhodného pokusu, který provádíme. Množinu Ω pak nazveme **prostor elementárních jevů**. Například při házení kostkou tedy můžeme za elementární jevy považovat počty ok, které se mohou objevit na horní stěně kostky, tj. $\Omega = \{\omega_1$ - padla jednička, $\ldots, \omega_6$ - padla šestka$\}$.

Dále zavedeme pojem **σ-algebry**.

Definice 1.1 *Nechť $\mathcal{A}$ je neprázdný systém podmnožin množiny $\Omega \neq \emptyset$ takový, že*

a) $\emptyset \in \mathcal{A}$,

b) je-li $A \in \mathcal{A}$, pak $A^c \in \mathcal{A}$, kde A^c značí doplněk množiny A do Ω.

c) jsou-li $A_i \in \mathcal{A}$, i = 1,2,..., pak $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

*Pak $\mathcal{A}$ nazýváme **σ-algebrou**.*

V případě, že Ω je prostor elementárních jevů, je množina $\mathcal{A}$ množinou jevů. Např. máme-li stejně jako výše $\Omega = \{\omega_1$ - padla jednička, $\ldots, \omega_6$ - padla šestka$\}$ a na ní σ-algebru $\mathcal{A} = \{\emptyset, \Omega, \{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\}\}$, je $\mathcal{A}$ množinou jevů {„nepadlo nic", „padlo cokoliv", „padlo liché číslo", „padlo sudé číslo"}. Poznamenejme však, že většinou množinu $\mathcal{A}$ uvažujeme jako množinu všech podmnožin Ω.

Definice 1.2 *Nechť $\Omega \neq \emptyset$ a $\mathcal{A}$ je σ-algebra definovaná na Ω. Pak **pravděpodobností** nazveme reálnou funkci P definovanou na $\mathcal{A}$, která splňuje*

a) $P(\Omega) = 1$, $P(\emptyset) = 0$,

b) $P(A) \geq 0$ pro všechna $A \in \mathcal{A}$,

c) pro každou posloupnost disjunktních jevů $\{A_n\}_{n=1}^{\infty}$ platí

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

*Trojice $(\Omega, \mathcal{A}, P)$ se nazývá **pravděpodobnostní prostor**.*

Poznámka 1.1 *Poznamenejme, že dvojice* $(\Omega, \mathcal{A})$ *tvoří* ***měřitelný prostor*** *a* P *je* ***míra****. Vlastnost* $P(\Omega) = 1$ *zajišťuje, že* P *je pravděpodobnostní míra (pravděpodobnost), obecně ale tuto vlastnost míra mít nemusí.*

A nyní již můžeme zavést výše zmíněný pojem klasické pravděpodobnosti.

Definice 1.3 *Pravděpodobnostní prostor* $(\Omega, \mathcal{A}, P)$ *nazveme* ***klasickým pravděpodobnostním prostorem****, jestliže*

i) množina $\Omega = \{\omega_1, ..., \omega_n\}$ *je konečná a všechny elementární jevy jsou stejně pravděpodobné, tj. označíme-li* $p_i = P(\omega_i),\ i = 1, ..., n$*, pravděpodobobnosti jednotlivých elementárních jevů, pak* $p_1 = \ldots = p_n = \frac{1}{n}$*,*

ii) $\mathcal{A}$ *je množina všech podmnožin* Ω*,*

iii) pravděpodobnost P *náhodného jevu* A *je rovna*

$$P(A) = \frac{n_A}{n},$$

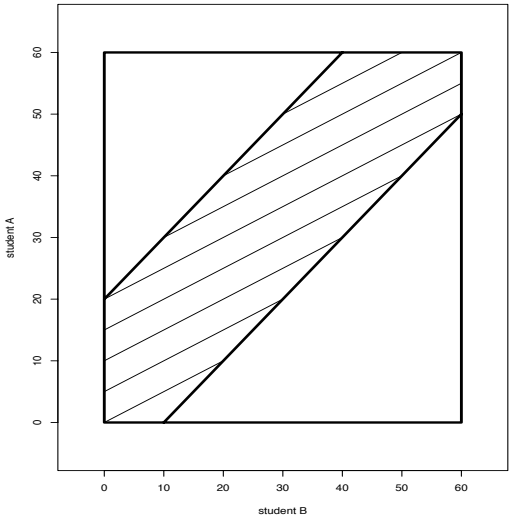
kde n_A *je počet elementárních jevů příznivých jevu* A*.*

Jak bylo vidět na D'Alembertově příběhu výše, podstatné zde je, že podmínka, aby všechny elementární jevy byly stejně pravděpodobné, je nutná a nelze ji vynechat.

Další skupinu úloh tvoří úlohy tzv. **geometrické pravděpodobnosti**. Základní myšlenka geometrické pravděpodobnosti je podobná té u klasické pravděpodobnosti, tedy že žádný z elementárních jevů z hlediska pravděpodobnosti nepreferujeme. Na rozdíl od klasické pravděpodobnosti je ale množina elementárních jevů nespočetná a lze ji vyjádřit jako nějakou omezenou podmnožinu z $\mathbb{R}^d$. Typickou ukázkou geometrické pravděpodobnosti je následující úloha.

Příklad 1.2 *Dva studenti přicházejí mezi 12:00 a 13:00 na smluvené místo. Doby příchodů obou studentů jsou náhodné, vzájemně nezávislé a žádný čas příchodu není preferovaný. Student A čeká na smluveném místě* 10 *minut a student B* 20 *minut, po této době pak oba odcházejí bez ohledu na to, zda se navzájem potkali. Jaká je pravděpodobnost, že se studenti potkají?*

Řešení:
Úlohu řešíme pomocí geometrické pravděpodobnosti. Množina všech možných jevů Ω (dvojic příchodů studentů A a B) je znázorněn na obrázku vpravo jako čtverec, v němž například bod $[10, 15]$ značí jev, že student B přišel ve 12:10 a student A přišel ve 12:15. Vyšrafovaná oblast značí množinu jevů, kdy se studenti potkají. Označme tuto oblast jako S. Úlohu řešíme tak, že hledaná pravděpodobnost se podíl šrafované oblasti S ku celé oblasti Ω. Obsah čtverce je $V(\Omega) = 60^2$, Obsah šrafované oblasti $V(S) = 60^2 - \frac{40^2}{2} - \frac{50^2}{2} = 2050$. Hledaná pravděpodobnost je tedy $P(S) = \frac{V(S)}{V(\Omega)} = \frac{2050}{3600} = 0.569444$.



Zaveďme si tedy nyní geometrickou pravděpodobnost formálně v následující definici.

Definice 1.4 ***Geometrickým pravděpodobnostním prostorem*** *nazveme pravděpodobnostní prostor* $(\Omega, \mathcal{A}, P)$ *takový, že*

i) $\Omega \subset \mathbb{R}^d$ *(obvykle* $d = 1, 2, 3$*), neboli všechny elementární jevy lze vyjádřit jako body nějaké podmnožiny* $\mathbb{R}^d$*.*

ii) $\mathcal{A} = \mathcal{B}(\Omega)$ *je Borelovská* σ*-algebra na* Ω *(formálně je to nejmenší* σ*-algebra obsahující všechny otevřené podmnožiny* Ω*, tudíž i všechny uzavřené podmnožiny a kombinace obou těchto typů, avšak zde si vystačíme s představou, že* $\mathcal{A}$ *je množina všech „rozumných" podmnožin množiny* Ω*, kde za „rozumnou" množinu považujeme takovou množinu, které lze přiřadit její délku, obsah či objem - v závislosti na dimenzi* d*).*

iii) $P(A) = \frac{\mu^d(A)}{\mu^d(\Omega)}$*, kde* μ^d *je* d*-rozměrná Lebesqueova míra. Pro naše účely postačí, pokud si pod* $\mu^1(A)$ *představíme délku množiny* A*, pod* $\mu^2(A)$ *obsah* A *a pod* $\mu^3(A)$ *objem* A*.*

V mnoha situacích si ale s klasickou ani s geometrickou pravděpodobností nevystačíme. Jsou situace, kdy všechny elementární jevy stejně pravděpodobné nejsou. Pak zavádíme např. zobecnění klasického pravděpodobnostního prostoru následovně.

Definice 1.5 *Pravděpodobnostní prostor* $(\Omega, \mathcal{A}, P)$ *se nazývá* ***obecný diskrétní****, jestliže*

i) $\Omega = \{\omega_1, \omega_2, \ldots\}$ *je konečná nebo spočetná,*

ii) $\mathcal{A}$ *je množina všech podmnožin* Ω*,*

iii) jsou dány pravděpodobnosti $P(\omega_i)$ *elementárních jevů* ω_i *splňující podmínku* $\sum_{i=1}^{\infty} P(\omega_i) = 1$ *a pravděpodobnost každého jevu* $A \in \mathcal{A}$ *je pak dána vztahem* $P(A) = \sum_{\omega_i \in A} P(\omega_i)$*.*

Ukázkou může být např. situace, kdy v případě D'Alembertova příběhu s počtem líců ve dvou hodech mincí zavedeme množinu $\Omega = \{\omega_0$ - nepadl líc, ω_1 - 1× padl líc, ω_2 - 2× padl líc$\}$. Již víme, že $P(\omega_0) = \frac{1}{4}$, $P(\omega_1) = \frac{1}{2}$ a $P(\omega_2) = \frac{1}{4}$. Označíme-li tedy A jev, že padl alespoň jeden líc, pak $P(A) = P(\omega_1) + P(\omega_2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$. Geometrický pravděpodobnostní prostor lze zobecnit analogicky tím, že jednotlivým bodům v množině Ω dáme různé váhy pomocí funkce, kterou nazýváme hustotou pravděpodobnosti. O této funkci budeme mluvit později.

V následujícím textu budeme používat následující terminologii:

1. $\emptyset$ ... jev nemožný
2. Ω ... jev jistý
3. $A \cup B$ ... sjednocení jevů A, B (jev, který nastane právě tehdy, nastane-li aspoň jeden z jevů A, B)
4. $A \cap B$ ... průnik jevů A, B (jev, který nastane právě tehdy, nastanou-li oba dva jevy současně)
5. $B - A$ ... rozdíl jevu B a A (jev, který nastane právě tehdy, když nastane jev B a zároveň nenastane jev A)
6. $A \subset B$ ... A je podjev jevu B (jev A nastane, kdykoliv nastane jev B)
7. $A^c = \Omega - A$ ... doplněk jevu A (jev, který nastane právě tehdy, když nenastane jev A)

8. $A \cap B = \emptyset$ ... jevy A, B jsou disjunktní (nemohou nastat současně)

Dále pak budeme využívat následující vlastnosti pravděpodobnosti:

1. $0 \leq P(A) \leq 1, \quad \forall A \in \mathcal{A}$,
2. P je monotónní: $A, B \in \mathcal{A}, A \subset B \Rightarrow P(A) \leq P(B)$,
3. $P(A^c) = 1 - P(A), \quad \forall A \in \mathcal{A}$,
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B), \quad \forall A, B \in \mathcal{A}$,
5. $A, B \in \mathcal{A}, A \subset B \Rightarrow P(B - A) = P(B) - P(A)$,
6. pro každou posloupnost disjunktních jevů $\{A_i\}_{i=1}^{\infty}$ takových, že $\cup_{i=1}^{\infty} A_i = \Omega$, platí $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) = 1$.

1.1.1 Podmíněná pravděpodobnost

Uvažujme následující otázku:

Příklad 1.3 *Jaká je pravděpodobnost, že při hodu ideální kostkou padne jednička, za předpokladu, že padlo liché číslo?*

Tuto úlohu lze samozřejmě řešit pomocí klasické pravděpodobnosti. My však k řešení této úlohy použijeme pojem **podmíněná pravděpodobnost**, který nyní zavedeme.

Definice 1.6 *Nechť je dán pravděpodobnostní prostor $(\Omega, \mathcal{A}, P)$ a náhodné jevy A, B, kde $P(B) > 0$.* ***Podmíněnou pravděpodobnost*** *jevu A za podmínky, že nastal jev B, definujeme vztahem*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.1}$$

Řešení:
Označme A jev, že padla jednička, a B jev, že padlo liché číslo. Jelikož $A \cap B$ je jev, že padne jednička, a sudých čísel je na kostce stejně jako lichých, pak $P(A \cap B) = \frac{1}{6}$ a $P(B) = \frac{1}{2}$. Tedy

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

Pomocí podmíněné pravděpodobnosti lze řešit i zajímavé příklady z reálného života.

Příklad 1.4 *Dne 9.10.2020 byla ve večerních zprávách zveřejněna následující zpráva: Ministr zdravotnictví Roman Prymula zvažuje možnost otestování celého národa na nemoc COVID-19. Byla přitom uvedena následující data:*

- *počet testovaných jedinců: 10 000 000,*
- *falešně negativních jedinců: 20 000,*
- *pozitivních jedinců: 180 000,*
- *falešně pozitivních jedinců: 249 000.*

Pokud by testování dopadlo dle odhadu a konkrétní pacient by byl testem označen jako pozitivní, jaká je pravděpodobnost, že je skutečně pozitivní?

Řešení:
Označme + jev, že testovanému jedinci vyšel pozitivní test, a N jev, že je testovaný jedinec skutečně nemocný. Pak $P(+) = \frac{249000+180000}{10000000} = 0.0429$ a $P(N \cap +) = \frac{180000}{10000000} = 0.018$. Tedy $P(N|+) = \frac{0.018}{0.0429} = 0.4195804$.

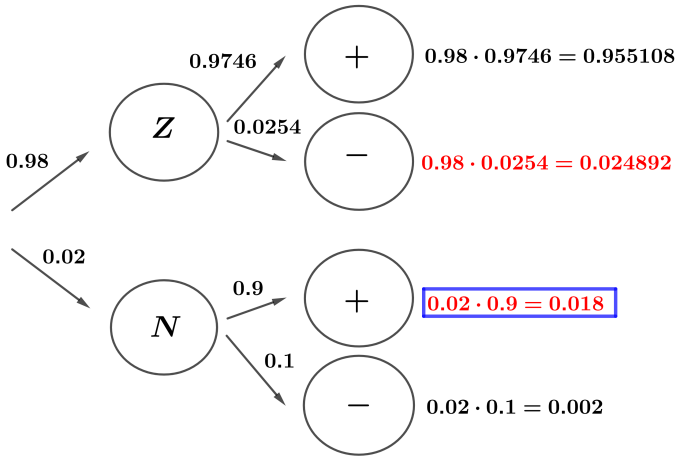
Zkusme nyní tuto úlohu přeformulovat tak, že nebudeme znát jednotlivé předpokládané počty jedinců v daných skupinách, ale budeme znát počty nemocných a úspěšnost prováděného testu.

Příklad 1.5 *Předpokládejme, že v uvedené době jsou nemocí COVID-19 nakažena* 2% *populace a prováděný test má úspěšnost* 97.46% *při testování zdravých jedinců a* 90% *při testovaní nakažených jedinců. Je-li testovaný jedinec označen testem jako pozitivní, jaká je pravděpodobnost, že je skutečně nemocný?*

Takto formulovaná úloha vede na takzvanou Bayesovu větu. Ukážeme si nejdříve různá řešení této úlohy a pak si teprve tuto větu zformulujeme.

Řešení:
Zavedeme si následující značení (stejné jako v předchozím řešení):

- N ... jev, že je testovaný jedinec nemocný,
- Z ... jev, že je testovaný jedinec zdravý,
- $+$ ... jev, že je testovaný jedinec testem označen jako pozitivní (nemocný),
- $-$ ... jev, že je testovaný jedinec testem označen jako negativní.



I. Nejdříve uvedeme grafické řešení této úlohy.

Tedy $P(N|+) = \frac{P(N\cap+)}{P(+)} = \frac{0.018}{0.024892+0.018} = \frac{0.018}{0.42892} = 0.4196587$.

II. Řešení pomocí Bayesovy věty (kterou si zde intuitivně odvodíme) spočívá v tom, že ze zadání dostaneme $P(N) = 0.02$, $P(Z) = 1 - P(N) = 0.98$, $P(+|N) = 0.9$ a $P(+|Z) = 1 - P(-|Z) = 1 - 0.9746 = 0.0254$.

Abychom mohli spočítat podmíněnou pravděpodobnost $P(N|+)$, potřebuje nejdříve určit pravděpodobnosti $P(N \cap +)$ a $P(+)$. Jelikož $P(+|N) = \frac{P(N\cap+)}{P(N)}$, pak $P(N\cap+) = P(+|N)\cdot P(N) = 0.9\cdot 0.02 = 0.018$. Podobně dostaneme $P(Z \cap +) = P(+|Z) \cdot P(Z) = 0.0254 \cdot 0.98 = 0.024892$. Pak si už stačí uvědomit, že $+ = (N \cap +) \cup (Z \cap +)$ a tyto dva jevy jsou disjunktní. Tedy $P(+) = P(N\cap+) + P(Z\cap+) = 0.018 + 0.024892 = 0.42892$. Hledaná pravděpodobnost je tedy $P(N|+) = \frac{P(N\cap+)}{P(+)} = \frac{0.018}{0.42892} = 0.4196587$.

Ve druhém řešení předchozí úlohy jsme došli ke vzorcům

$$P(+) = P(N \cap +) + P(Z \cap +) = P(+|N) \cdot P(N) + P(+|Z) \cdot P(Z),$$
$$P(N|+) = \frac{P(N \cap +)}{P(+)} = \frac{P(+|N) \cdot P(N)}{P(+|N) \cdot P(N) + P(+|Z) \cdot P(Z)}.$$

Zformulujme tyto vzorce v obecnější podobě v následujících větách (jejichž důkazy nebudeme uvádět, neboť vycházejí v postupu, který jsme již ukázali v řešení předchozího příkladu).

Věta 1.1 (O celkové pravděpodobnosti) *Nechť A_1, A_2, ... jsou náhodné jevy tvořící rozklad jevu jistého, tzn.*

$$A_i \cap A_j = \emptyset,\ \forall i \neq j \ a\ \cup_{i=1} A_i = \Omega.$$

Nechť tyto náhodné jevy mají postupně pravděpodobnosti $P(A_1), P(A_2), \ldots,$ přičemž $P(A_i) > 0,\ \forall i = 1, 2, \ldots$ Uvažujme libovolný náhodný jev B, u něhož známe podmíněné pravděpodobnosti

$$P(B|A_i),\ \forall i = 1, 2, \ldots$$

Potom

$$P(B) = \sum_{i=1} P(A_i) \cdot P(B|A_i). \tag{1.2}$$

Věta 1.2 (Bayesova věta) *Nechť jsou splněny předpoklady věty 1.1. Pak*

$$P(A_i|B) = \frac{P(B|A_i) \cdot P(A_i)}{\sum_{j=1} P(A_j) \cdot P(B|A_j)}, \quad i = 1, 2, \ldots \tag{1.3}$$

1.1.2 Nezávislost

Definice 1.7 *Náhodné jevy A a B jsou **nezávislé**, jestliže platí*

$$P(A \cap B) = P(A) \cdot P(B). \tag{1.4}$$

Pojem nezávislosti můžeme rozšířit i na skupinu náhodných jevů.

Definice 1.8 *Nechť $A_1, A_2, \ldots, A_n$ jsou náhodné jevy. Řekneme, že jsou **skupinově (totálně) nezávislé**, jestliže pro libovolnou posloupnost indexů $\{k_1, k_2, \ldots, k_r\} \subset \{1, \ldots, n\},\ r = 2, \ldots, n,$ platí*

$$P(A_{k_1} \cap A_{k_2} \cap \ldots \cap A_{k_r}) = P(A_{k_1}) \cdot P(A_{k_2}) \cdot \ldots \cdot P(A_{k_r}). \tag{1.5}$$

Definice 1.9 *Nechť $A_1, \ldots, A_n$ jsou náhodné jevy. Řekneme, že jsou **po dvou nezávislé**, jestliže jevy A_i, A_j jsou nezávislé pro všechna $i, j = 1, \ldots, n,$ $i \neq j$.*

Příklad 1.6 *Uvažujme hod jednou šestistěnnou vyváženou kostkou. Označme A jev, že na kostce padlo sudé číslo, B je jev, že padlo číslo větší než 2, a C jev, že padlo liché číslo. Jsou tyto jevy nezávislé? Jsou po dvou nezávislé? Existuje v těchto jevech dvojice jevů, které jsou nezávislé?*

Řešení:
Označme ω_i elementární jev, že padlo na kostce i ok. Pak

$$\begin{aligned}
P(A) &= P(\{\omega_2, \omega_4, \omega_6\}) = \frac{1}{2} \\
P(B) &= P(\{\omega_3, \omega_4, \omega_5, \omega_6\}) = \frac{2}{3} \\
P(C) &= P(\{\omega_1, \omega_3, \omega_5\}) = \frac{1}{2} \\
P(A \cap B) &= P(\{\omega_4, \omega_6\}) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3} = P(A) \cdot P(B) \\
P(A \cap C) &= P(\emptyset) = 0 \neq \frac{1}{2} \cdot \frac{1}{2} = P(A) \cdot P(C) \\
P(B \cap C) &= P(\{\omega_3, \omega_5\}) = \frac{1}{3} = \frac{1}{2} \cdot \frac{2}{3} = P(A) \cdot P(B).
\end{aligned}$$

Jelikož jevy A a C nejsou nezávislé, nejsou jevy A, B a C totálně nezávislé ani po dvou nezávislé. Jevy A a B však nezávislé jsou a rovněž tak jevy B a C jsou nezávislé.

1.2 Náhodná veličina

Definice 1.10 *Nechť $(\Omega, \mathcal{A}, P)$ je pravděpodobnostní prostor. Reálnou funkci X definovanou na Ω nazýváme* ***náhodnou veličinou (n.v.)****, jestliže X je* ***měřitelné zobrazení*** *$X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B})$ (tj. $\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{A}$) pro libovolnou borelovskou množinu $B \in \mathcal{B}$.*

Náhodné veličiny budeme značit velkými písmeny $X, Y, Z \ldots$ Hodnoty, kterých mohou náhodné veličiny nabývat, budeme značit malými písmeny $x, y, z \ldots$ Místo $\{\omega \in \Omega : X(\omega) \in B\}$ pak budeme zjednodušeně psát $\{X \in B\}$ a místo $\{\omega \in \Omega : X(\omega) < x\}$ budeme zjednodušeně psát $\{X < x\}$. Poznamenejme, že součty, součiny, podíly, minima a maxima náhodných veličin jsou opět náhodné veličiny; umocnění náhodné veličiny přirozeným číslem, násobení náhodné veličiny skalárem jsou také náhodné veličiny.

Definice 1.11 *Nechť X je náhodná veličina. Její* ***distribuční funkcí*** *nazýváme reálnou funkci F_X reálné proměnné x definovanou*

$$F_X(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\}), \quad x \in \mathbb{R}. \tag{1.6}$$

Distribuční funkce $F_X(x)$ náhodné veličiny X má následující vlastnosti:

1. je neklesající, tj. pro libovolné $a, b \in \mathbb{R}, a \leq b$, platí $F_X(a) \leq F_X(b)$,

2. je zprava spojitá v libovolném bodě $x \in \mathbb{R}$,

3. $\lim_{x\to-\infty} F_X(x) = 0, \lim_{x\to\infty} F_X(x) = 1$.

Diskrétní náhodná veličina

Definice 1.12 *Náhodná veličina X se nazývá **diskrétní** (nebo také říkáme, že X má **diskrétní rozdělení pravděpodobnosti**), jestliže existuje (konečná nebo spočetná) posloupnost reálných čísel $\{x_n\}$ a odpovídající posloupnost nezáporných čísel $\{p_n\} = P(X = x_n)$ taková, že $\sum_{n=1}^{\infty} p_n = 1$.*

Příklad 1.7 *Uvažujme opět hod jednou šestistěnou vyváženou kostkou. Náhodná veličina X nám udává počet ok, která na kostce padla. Označme ω_i jev, že na kostce padlo i ok. Pak*

$$X(\omega_i) = i \qquad\qquad i = 1, 2, ..., 6$$

Jiná náhodná veličina Y zase vrací jedničku, pokud padlo sudé číslo, a nulu, pokud padlo liché číslo. Pak

$$\begin{aligned} Y(\omega_i) &= 1 \qquad & i = 2, 4, 6 \\ &= 0 \qquad & i = 1, 3, 5. \end{aligned}$$

Chceme-li určit rozdělení těchto náhodných veličin, pak musíme určit příslušné pravděpodobnosti, při kterých nabývají n.v. X, resp. Y, hodnot $1, 2, ..., 6$, resp. 0 a 1. Tedy $P(X = i) = \frac{1}{6}, i = 1, ..., 6$ a $P(Y = 0) = P(Y = 1) = \frac{1}{2}$.

Poznamenejme, že rozdělení n.v. Y z předchozího příkladu se nazývá alternativní rozdělení s parametrem $p = \frac{1}{2}$.

Absolutně spojitá náhodná veličina

Definice 1.13 *Náhodná veličina X se nazývá **absolutně spojitá** (nebo také říkáme, že X má **absolutně spojité rozdělení pravděpodobnosti**), jestliže existuje nezáporná integrovatelná funkce f_X taková, že platí*

$$F_X(x) = P(X \le x) = \int_{-\infty}^{x} f_X(t)dt, \quad x \in (-\infty, \infty). \tag{1.7}$$

*Funkce f_X se nazývá **hustotou** rozdělení pravděpodobnosti.*

Poznámka 1.2 *Místo „$P(X$ má vlastnost $V) = 1$" říkáme „X má vlastnost V skoro jistě" a používáme zkratku „s.j."*

Hustota má následující vlastnosti:

1. $f_X(x) = \frac{d}{dx}F_X(x)$ s.j.,
2. $\int_{-\infty}^{\infty} f_X(x)dx = 1$,
3. $P(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$ pro libovolná reálná čísla a, b, kde $a \leq b$.

Příklad 1.8 *Pošta chodí každý den mezi* 10. *a* 12. *hodinou. Náhodná veličina X udává čas příchodu pošty. Určete distribuční funkci X, hustotu X a pravděpodobnost, že pošta přijde po půl dvanácté, za předpokladu, že příchod pošty je rovnoměrně rozdělen do intervalu* $(10, 12)$*, tj. žádný čas příchodu není preferován.*

Řešení:
Jelikož je čas příchodu rovnoměrně rozdělen na intervalu $(10, 12)$, pak hustota $f(x)$ je konstantní na tomto intervalu. Použijeme-li druhou vlastnost hustoty, dostaneme

$$\int_{10}^{12} f(x)dx = \int_{10}^{12} cdx = 2c = 1 \Rightarrow c = \frac{1}{2},$$

tedy

$$\begin{aligned} f(x) &= \frac{1}{2} && x \in (10, 12), \\ &= 0 && x \notin (10, 12). \end{aligned}$$

Pak

$$\begin{aligned} F_X(x) &= 0 && x \leq 10, \\ &= \int_{10}^{x} \frac{1}{2}dx = \frac{x}{2} - 5 && x \in (10, 12), \\ &= 1 && x \geq 12. \end{aligned}$$

Hledaná pravděpodobnost pak je $P(X > 11.5) = \int_{11.5}^{12} \frac{1}{2}dx = \frac{1}{4}$. Poznamenejme, že rozdělení n.v. X z předchozího příkladu se nazývá **rovnoměrné rozdělení** na intervalu $(10, 12)$.

1.2.1 Charakteristiky náhodných veličin

Střední hodnota

Definice 1.14 *Nechť X je náhodná veličina definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$.*

a) Je-li X diskrétní náhodná veličina nabývající reálných hodnot x_1, x_2, x_3, ..., tzn. taková, že $P(X = x_i) = p_i$, pak ***střední hodnota*** $\mathbb{E}X$ *náhodné veličiny X je tvaru*

$$\mathbb{E}X = \sum_{i=1}^{\infty} x_i \cdot p_i,$$

pokud řada konverguje.

b) Je-li X absolutně spojitá náhodná veličina s hustotou f_X, pak ***střední hodnota*** *náhodné veličiny X je*

$$\mathbb{E}X = \int_{-\infty}^{\infty} x f_X(x) dx,$$

pokud integrál existuje.

Základní vlastnosti střední hodnoty jsou následující:

1. $\mathbb{E}a = a$,
2. $\mathbb{E}(aX + bY) = a\mathbb{E}X + b\mathbb{E}Y$.

Věta 1.3 *Nechť X je náhodná veličina definovaná na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, P)$ a nechť $\phi : \mathbb{R} \to \mathbb{R}$. Pak za obdobných podmínek jako v předešlé definici je $\mathbb{E}\phi(X) = \sum_{i=1}^{\infty} \phi(x_i)p_i$ pro diskrétní náhodnou veličinu X a $\mathbb{E}\phi(X) = \int_{-\infty}^{\infty} \phi(x)f(x)dx$ pro spojitou náhodnou veličinu X.*

Rozptyl a kovariance

Definice 1.15 *Nechť X je náhodná veličina. Její* ***rozptyl*** *je definován jako*

$$\text{var } X = \mathbb{E}(X - \mathbb{E}X)^2.$$

Definice 1.16 *Nechť X, Y jsou náhodné veličiny. Jejich* ***kovariance*** *je definována jako*

$$\text{cov}(X, Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

Poznámka 1.3 *Z předešlých dvou definic vyplývá souvislost mezi rozptylem a kovariancí, a to*

$$\text{cov}(X, X) = var(X).$$

Základní vlastnosti rozptylu a kovariance jsou následující:

1. Nechť X je náhodná veličina. Pak var $X = \mathbb{E}(X^2) - (\mathbb{E}X)^2$.

2. Nechť c je konstanta. Pak var $c = 0$.

3. Nechť X je náhodná veličina a a je reálné číslo. Pak var$(aX) = a^2$var X.

4. Nechť X je náhodná veličina a c je konstanta. Pak var $(X + c) =$ var X.

5. Nechť X je náhodná veličina, která má konečnou střední hodnotu a konečný nenulový rozptyl. Nechť

$$Z = \frac{X - \mathbb{E}X}{\sqrt{\text{var } X}}.$$

   Pak $\mathbb{E}Z = 0$ a var $Z = 1$.

6. Nechť X, Y jsou náhodné veličiny. Pak var $(X + Y) =$ var $X +$ var $Y +$ $2\text{cov}(X, Y)$.

7. Nechť X, Y jsou náhodné veličiny. Pak $\text{cov}(X, Y) = \mathbb{E}(XY)$ - $\mathbb{E}X\mathbb{E}Y$.

1.2.2 Příklady diskrétních náhodných veličin

Příklad 1.9 *Uvažujme hod kostkou při hře „Člověče, nezlob se" v situaci, kdy máme všechny figurky v domečku. Zajímá nás, zda budeme moci po následujícím hodu kostkou nasadit figurku, tedy zda nám padne na kostce číslo šest. Označme ω_i jevy, že na kostce padne i ok, a zaveďme náhodnou veličinu X následujícím způsobem:*

$$X(\omega_i) = 0, \quad i = 1, ..., 5$$
$$X(\omega_6) = 1.$$

Náhodná veličina X tedy nabývá hodnoty jedna v případě úspěšného hodu a nuly v případě neúspěchu. Rozdělení této náhodné veličiny je

$$P(X = 0) = \frac{5}{6},$$
$$P(X = 1) = \frac{1}{6}.$$

Tomuto rozdělení pravděpodobnosti se říká alternativní (máme jen dva možné výsledky).

1. Alternativní rozdělení ($Alt(p)$)
má náhodná veličina X, která nabývá jen hodnot 0 a 1, a to s pravděpodobnostmi $1-p$, resp. p. Číslo p se nazývá parametr alternativního rozdělení, $0 < p < 1$.

Snadno se z definic vypočte střední hodnota $\mathbb{E}X = p$ a rozptyl $\text{var } X = p(1-p)$.

Příklad 1.10 *Uvažujme nyní situaci, že budeme házet kostkou* 10*-krát a bude nás zajímat, kolikrát nám padla šestka. Označme si X počet padlých šestek, pak*

$$P(X=k) = \binom{10}{k}\left(\frac{1}{6}\right)^k\left(\frac{5}{6}\right)^{10-k}, \quad k = 0,1,...,10.$$

Tomuto rozdělení se říká binomické.

2. Binomické rozdělení ($Bi(n,p)$)
je rozdělení náhodné veličiny X, která nám udává kolik bude úspěšných pokusů z n pokusů. Binomické rozdělení je tedy jednoznačně určeno dvěma parametry:

- přirozeným číslem n, které nám udává počet opakování náhodného pokusu,
- číslem $p \in (0,1)$, které nám udává pravděpodobnost úspěchu jednotlivých pokusů.

Rozdělení této náhodné veličiny je

$$P(X=k) = \binom{n}{k}p^k(1-p)^{n-k}, \quad k = 0,1,...,n.$$

Lze odvodit, že $\mathbb{E}X = np$ a $varX = np(1-p)$.

Příklad 1.11 *Uvažujme společnost, která má velký počet zákazníků (třeba společnost zajišťující energii pro většinu obyvatel nějakého státu). Společnost zajišťuje pro své zákazníky call centrum. Z dřívější statistiky víme, že na toto call centrum zavolá průměrně* 30 *zákazníků za hodinu. Chtěli bychom nějak odhadnout rozdělení náhodné veličiny X, která nám bude udávat počet příchozích telefonátů do call centra během jedné hodiny. Budeme uvažovat následujícím způsobem. Ze zadání víme, že $\mathbb{E}X = 30$ (průměrně volá* 30

zákazníků za hodinu). Pokud si hodinu rozdělíme na 60 *minutových intervalů a budeme předpokládat, že v každé minutě zavolá maximálně jeden zákazník, pak by* X *mělo binomické rozděleni s parametry* $n = 60$ *(počet minutových intervalů) a* $p = \frac{1}{2}$ *(pravděpodobnost, že v dané minutě zavolá jeden zákazník). Parametr* p *jsme určili z rovnosti* $30 = \mathbb{E}X = np = 60 \cdot p$*. Avšak předpoklad, že v každé minutě zavolá maximálně jeden zákazník, neodpovídá realitě, a tedy by náš model (n.v.* X*) nepopisoval situaci nejlépe.*

Zkusme tedy rozdělit hodinu na 3600 *vteřinových intervalů. Pak by předpoklad, že v každé vteřině zavolá maximálně jeden zákazník působil reálněji. V tomto případě by náhodná veličina* X *měla binomické rozdělení s parametry* $n = 3600$ *a* $p = \frac{1}{120}$*.*

Přirozenou otázkou je, zda je toto dělení časového intervalu už dostatečné, aby náš model (n.v. X*) věrně popisoval realitu. Co by se dělo, pokud bychom dále zjemňovali dělení časového intervalu? Zkusme najít limitní situaci.*

$$\begin{aligned} P(X = k) &= \lim_{n\to\infty, np=30} \binom{n}{k} p^k (1-p)^{n-k} \\ &= \lim_{n\to\infty, np=30} \frac{n(n-1)...(n-k+1)}{k!} \left(\frac{30}{n}\right)^k \cdot \left(1 - \frac{30}{n}\right)^{n-k} \\ &= \frac{30^k}{k!} e^{-30}, \quad k = 0, 1, ... \end{aligned}$$

Tomuto rozdělení se říká Poissonovo (s parametrem $\lambda = 30$*).*

3. Poissonovo rozdělení ($Po(\lambda)$**)**
je rozdělení náhodné veličiny X, která nabývá hodnot $k = 0, 1, 2, \ldots$ s pravděpodobnostmi

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}.$$

Číslo $\lambda > 0$ je parametr Poissonova rozdělení. Střední hodnota je λ a rozptyl je rovněž roven λ.

1.2.3 Příklady spojitých náhodných veličin

Příklad 1.12 *Uvažujme situaci, kdy přijdeme na zastávku tramvaje aniž bychom znaly přesný jízdní řád, ale víme, že tramvaj přijede každých deset minut. Jaké je rozdělení času, který strávíme čekáním na tramvaj? Označme* X *dobu čekání na tramvaj. Pak* X *nabývá hodnot v intervalu* $[0, 10)$ *a je to spojitá náhodná veličina (tedy má nějakou hustotu* $f_X(t)$*). Dále víme, že pravděpodobnost příjezdu tramvaje v intervalu* $(t, t+\delta t)$ *nezáleží na čase* t *ale pouze na* δt *(je-li* $(t, t + \delta t) \subset [0, 10)$*). Z toho odvodíme, že* $f_X(t) = c > 0$ *na*

intervalu $(0, 10)$ *a* $f_X(t) = 0$ *mimo interval* $(0, 10)$. *Jelikož* $\int_{\mathbb{R}} f_X(t)dt = 1$, *pak dostáváme* $c = \frac{1}{10}$. *Rozdělení takové náhodné veličiny se nazývá rovnoměrné (na intervalu* $(0, 10)$*).*

1. Rovnoměrné rozdělení na intervalu $[a, b]$ **(**$Ro(a, b)$**)**
je dáno hustotou

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & x < a, \quad x > b. \end{cases}$$

Lehce se spočte, že střední hodnota a rozptyl jsou

$$\mathbb{E}X = \frac{a+b}{2}, varX = \frac{1}{12}(b-a)^2.$$

Příklad 1.13 *Uvažujme opět situaci zákaznického call centra s průměrným počtem příchozích hovorů* 30 *hovorů za hodinu. Jak dlouho budeme v tomto centru čekat na další hovor? Označme* X_t *počet příchozích hovorů v časovém intervalu délky* t *(např.* $(0, t)$ *nebo* $(T, T + t)$*). Pak z předchozího víme, že* X *má poissonovo rozdělení s parametrem* $\lambda = \frac{t}{2}$, *kde* t *je čas udávaný v minutách. Označme* Y *čas čekání na další hovor. Přišel-li poslední hovor v čase* T, *pak*

$$F_X(t) = P(Y \leq t) = P(X_t > 0) = 1 - P(X_t = 0) = 1 - e^{-\frac{t}{2}}, \quad t > 0.$$

Ze vztahu mezi distribuční funkcí a hustotou dostaneme $f_X(t) = F'_X(t)$, *tedy* $f_X(t) = \frac{1}{2}e^{-\frac{t}{1}}, \quad t > 0$. *Rozdělení s touto hustotou se nazývá exponenciální.*

2. Exponenciální rozdělení ($Exp(\lambda)$**)**
je rozdělení s hustotou

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0 \\ 0 & \text{jinak,} \end{cases}$$

kde λ je parametrem rozdělení. Distribuční funkce je

$$F(x) = \begin{cases} 0 & \text{pro } x \leq 0 \\ 1 - e^{-\lambda x} & x > 0. \end{cases}$$

Integrací per partes se spočte střední hodnota

$$\mathbb{E}X = \int_0^\infty x\lambda e^{-\lambda x} dx = \frac{1}{\lambda}.$$

Dvojitým použitím per partes získáme

$$\mathbb{E}X^2 = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \frac{2}{\lambda^2}$$

a následně rozptyl

$$\text{var } X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{1}{\lambda^2}.$$

Poznámka 1.4 *Exponenciální rozdělení má vlastnost, že je "bez paměti". Tím myslíme, že pro všechna $T, t > 0$ platí rovnost*

$$P(X > T + t | X > t) = P(X > t).$$

Tuto rovnost můžeme slovně interpretovat třeba takto: Čekám-li třeba na příchozí telefon už deset minut a zajímá nás, jaká je pravděpodobnost, že přijde v následujících pěti minutách. Pak je tato pravděpodobnost stejná, jako že telefonát přišel v prvních pěti minutách (zde byla volba $T = 10$ a $t = 5$. Tato vlastnost se předpokládá např. u systémů hromadné obsluhy (call centrum), nebo u času poruch součástek, které se neopotřebovávají. Lze navíc ukázat, že tato vlastnost (býti bez paměti) exponenciální rozdělení přímo definuje. Přesněji, má-li mít nějaké rozdělení tuto vlastnost, pak už musí být exponenciální.

Poznámka 1.5 *Platí, že pokud doba čekání na událost má exponenciální rozdělení s parametrem λ, pak počet událostí do času t má Poisonovo rozdělení s parametrem λt.*

3. Obecné normální rozdělení ($N(\mu, \sigma^2)$)
je definováno hustotou

$$f(x) = \frac{1}{\sqrt{2\pi\ \sigma^2}}\ e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty,$$

kde μ reálné a σ^2 kladné jsou parametry.

Střední hodnota $\mathbb{E}X = \mu$ a rozptyl var $X = \sigma^2$.

Poznámka 1.6

- *Normální rozdělení má mimořádný význam v teorii pravděpodobnosti a matematické statistice, přestože se tímto rozdělením řídí přesně jen málo náhodných veličin. Takzvaná centrální limitní věta (CLV), či přesněji různé verze této věty nám zjednodušeně totiž říkají, že součet velkého počtu nezávislých náhodných veličin (o jejichž rozdělení se činí jen velmi obecné předpoklady) má přibližně normální*

rozdělení. Tím lze vysvětlit klíčovou roli tohoto rozdělení v teorii pravděpodobnosti a matematické statistice, ale i častý výskyt Gaussovy křivky (která popisuje hustotu normálního rozdělení) ve světě kolem nás. Náhodné veličiny, s nimiž se v reálném světě setkáváme, lze velmi často považovat za výslednice působení velkého počtu drobných náhodných vlivů. Pak lze očekávat, že normální rozdělení bude vhodným modelem pro takové náhodné veličiny. Nejběžnějším typem takových veličin jsou náhodné chyby (chyby měření, způsobené velkým počtem neznámých a vzájemně nezávislých příčin). Normální rozdělení je vhodným modelem pro řadu fyzikálních, technických a biologických veličin jako například tělesná výška jedinců homogenní populace, roční částka, kterou pojišťovna vyplatí za pojistné příhody atd.

- *Jelikož se s normálním rozdělením velmi často pracuje a výpočet distribuční funkce je zdlouhavý, jsou hodnoty distribuční funkce $N(0,1)$ tabelovány. Vzhledem k symetrii funkce $(\Phi(x) = 1 - \Phi(-x))$ se tabelují hodnoty Φ pouze pro nezáporné x.*

Z teoretického využití normálního rozdělení si uveďme alespoň jednu verzi centrální limitní věty. Poznamenejme jen, že pojem nezávislosti náhodných veličin, který je předpokladem v následující větě, bude zaveden v další sekci.

Věta 1.4 (Lévy-Lindebergova CLV)
Nechť X_1, X_2, ... jsou nezávislé stejně rozdělené náhodné veličiny se střední hodnotou μ a konečným rozptylem σ^2. Označme

$$Z_n = \frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n\sigma^2}} \quad n = 1, 2, \ldots$$

a označme $F_n(x)$ distribuční funkci Z_n. Potom $\lim_{n\to\infty} F_n(x) = \Phi(x)$ pro všechna $-\infty < x < \infty$, kde $\Phi(x)$ je distribuční funkce $N(0,1)$.

1.2.4 Nezávislost náhodných veličin

Definice 1.17 *Náhodné veličiny X_1, X_2 ..., X_n jsou **vzájemně nezávislé**, jestliže*

$$P(\cap_{j=1}^{r}\{\omega : X_{i_j}(\omega) < x_{i_j}\}) = \Pi_{j=1}^{r} P(\{\omega : X_{i_j}(\omega) < x_{i_j}\}) \qquad (1.8)$$

$$\forall\{i_1, i_2, \ldots, i_r\} \subset \{1, 2, \ldots, n\}, 1 \le r \le n, \forall x_{ij} \in \mathbb{R}.$$

Poznámka 1.7 *Podobně jako u náhodných jevů můžeme zde definovat nezávislost náhodných veličin X_1, X_2 ..., X_n po dvou. Definici nezávislosti po dvou bychom dostali z definice 1.17 pro $r = 2$.*

Věta 1.5 (Ověřování nezávislost náhodných veličin v praxi)

a) Nechť $\mathbb{X} = (X_1, X_2 \ldots, X_n)$ *je náhodný vektor diskrétního typu. Náhodné veličiny* $X_1, X_2 \ldots, X_n$ *jsou vzájemně nezávislé právě tehdy, když platí*

$$P(X_1 = x_1^{(i)}, \ldots, X_n = x_n^{(i)}) = \Pi_{j=1}^{n} P(X_j = x_j^{(i)})$$

pro všechna $\mathbf{x}^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)})$, $i = 1, 2, \ldots$, *kterých může* $\mathbb{X}$ *nabývat.*

b) Nechť $\mathbb{X} = (X_1, X_2 \ldots, X_n)$ *je náhodný vektor absolutně spojitého typu. Náhodné veličiny* $X_1, X_2 \ldots, X_n$ *jsou vzájemně nezávislé právě tehdy, platí-li*

$$f_{\mathbb{X}}(x_1, x_2 \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \ldots f_{X_n}(x_n), \ \forall (x_1, x_2 \ldots, x_n) \in \mathbb{R}^n.$$

Věta 1.6 *Jsou-li* X, Y *nezávislé náhodné veličiny s konečnými středními hodnotami, pak*

a) $\mathbb{E}XY = (\mathbb{E}X)(\mathbb{E}Y)$.

b) Jsou-li navíc $\mathbb{E}X^2 < \infty$ *a* $\mathbb{E}Y^2 < \infty$, *pak* $\mathrm{cov}(X, Y) = 0$.

Platí-li $\mathrm{cov}(X, Y) = 0$, pak říkáme, že náhodné veličiny jsou nekorelované. Z nekorelovanosti však obecně ještě neplyne nezávislost!

Kapitola 2

Základy statistiky

Statistika se obecně zabývá prací s daty, a to od způsobu jejich získání (sběru dat) přes základní zpracování včetně tvorby modelů (popisná statistika) a odhady různých parametrů (bodové a intervalové odhady) po ověřování modelů a jejich vlastností (testovaní hypotéz). V této kapitolce si v krátkosti projdeme popisnou statistiku, bodové a intervalové odhady, testování hypotéz.

2.1 Popisná statistika

Definice 2.1 *Náhodný vektor* $\mathbb{X} = (X_1, X_2 \ldots, X_n)$ *nezávislých, stejně rozdělených náhodných veličin, se nazývá* ***náhodný výběr***.

Poznámka 2.1 *To, co v praxi označíme za data, jsou již konkrétní hodnoty* $(x_1, x_2 \ldots, x_n)$*, které nazýváme též realizací náhodného výběru* $\mathbb{X} = (X_1, X_2 \ldots, X_n)$*. Musíme si však uvědomit, že pokud bychom sběr dat opakovali, získali bychom jiných n hodnot, tudíž je z hlediska teorie nutné na náhodný výběr nahlížet jako na vektor náhodných veličin. Stejně tak charakteristiky uvedené v následující definici jsou z teoretického hlediska náhodnými veličinami, v případě dosazení konkrétních dat se však již jedná o číselné hodnoty, které jsou vhodnými odhady teoretických charakteristik náhodných veličin, jmenovitě střední hodnoty a rozptylu.*

Definice 2.2 *Funkce*

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

náhodného výběru $\mathbb{X} = (X_1, X_2 \ldots, X_n)$ *se nazývá* ***výběrový průměr*** *a*

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

se nazývá ***výběrový rozptyl****.* $S_n = \sqrt{S_n^2}$ *je pak* ***výběrová směrodatná odchylka****.*

Definice 2.3 *Nechť F je spojitá a monotónní distribuční funkce a $0 < \beta < 1$. Pak hodnotu z_β takovou, že $F(z_\beta) = \beta$, nazýváme* ***β-kvantil*** *rozdělení s distribuční funkcí F.*

Pro náhodnou veličinu X s distribuční funkcí F a kvantily z_β pak často využíváme vlastnosti, že

$$P(z_{\alpha/2} < X < z_{1-\alpha/2}) = F(z_{1-\alpha/2}) - F(z_{\alpha/2}) = 1 - \alpha.$$

Definice 2.4 *Nechť $(x_1, x_2 \ldots, x_n)$ jsou data, tedy realizace náhodného výběru $(X_1, X_2 \ldots, X_n)$. Pak*

$$F_{emp}(x) = \frac{\#\{x_i : x_i \leq x\}}{n},$$

kde # značí počet prvků, se nazývá ***empirická distribuční funkce****.*

Definice 2.5 *Nechť $(x_1, x_2 \ldots, x_n)$ jsou data, $F_{emp}(x)$ příslušná empirická distribuční funkce a z_β značí β-kvantil náhodné veličiny s distribuční funkcí F_{emp}. Pak hodnoty $z_{1/4}$, $z_{1/2}$ a $z_{3/4}$ se nazývají* ***1.kvartil****,* ***2.kvartil*** *(též "****medián****"), resp.* ***3.kvartil****. Nejčastěji zatoupený prvek v realizaci náhodného výběru se nazývá* ***modus****.*

Poznámka 2.2 *Občas se 1.kvartil definuje jako $z^* = max(x_i : F_{emp}(x_i) \leq 1/4)$ nebo $z^{**} = min(x_i : F_{emp}(x_i) \geq 1/4)$, popř. jako $z^{***} = z^* + \frac{1}{4}(z^{**} - z^*)$. Analogicky se pak definuje i 2. a 3.kvartil. Poznamenejme, že jelikož se jedná o popisné statistiky, tedy údaje sloužící k přibližné představě o datech, a s měnícími se daty se lehce mění i hodnoty kvartilů, nejsou zmíněné rozdíly v jejich definicích zásadním problémem.*

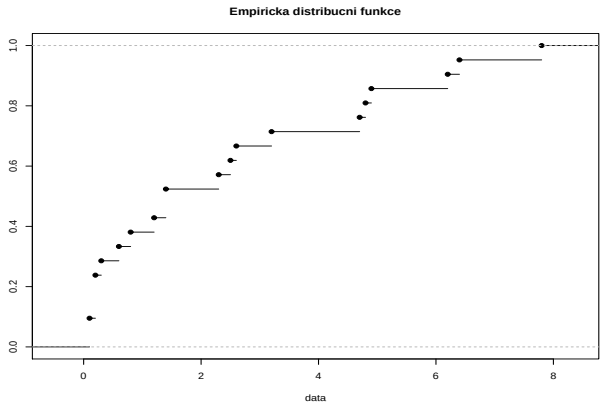
Dalšími popisnými statistikami jsou pak grafická znázornění, z nichž nejznámější je **histogram** - graf, kde do intervalů s ekvidistantními hranicemi vynášíme sloupce, jejichž výška odpovídá počtu dat spadajících do daných intervalů. Dalším grafickým nástrojem k popisu dat je tzv. **boxplot** (nebo také **krabicový graf**), v němž máme vyznačené minimum, maximum a oblast mezi 1. a 3. kvartilem s vyznačeným mediánem.

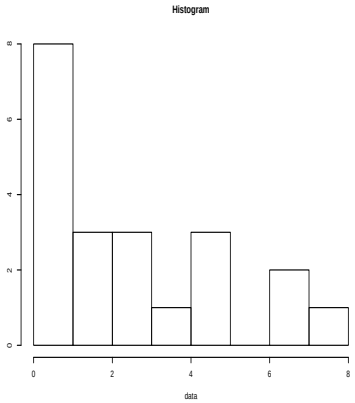
Příklad 2.1 *Sledovali jsme doby mezi příchody zákazníků (v minutách) a naměřili jsme těchto 21 hodnot:*

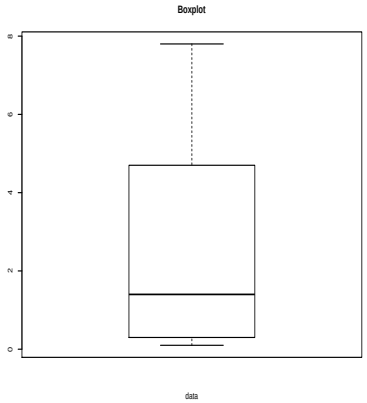
4.9, 6.2, 2.6, 0.6, 0.3, 2.3, 3.2, 1.4, 6.4, 4.8, 1.2,
2.5, 0.2, 0.2, 0.8, 0.1, 0.1, 1.4, 7.8, 0.2, 4.7.

Pro přehlednost si hodnoty seřadíme od nejmenší po největší:
***0.1**, 0.1, 0.2, 0.2, 0.2, **0.3**, 0.6, 0.8, 1.2, 1.4, **1.4**,*
*2.3, 2.5, 2.6, 3.2, **4.7**, 4.8, 4.9, 6.2, 6.4, **7.8**.*

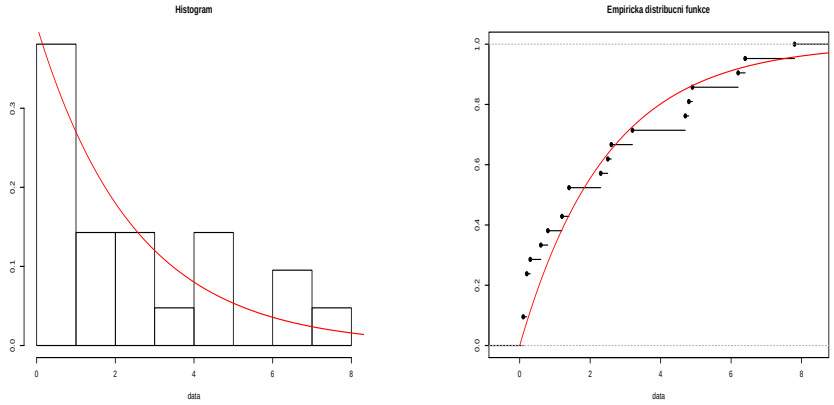
Máme zde:
výběrový průměr (pro danou realizaci) $\bar{x}_{21} = 2.471$,
výběrový rozptyl (pro danou realizaci) $S^2_{21} = 5.81$,
výběrovou směrodatnou odchylku (pro danou realizaci) $S_{21} = 2.21$,
1.kvartil = 0.3, medián (tj. 2.kvartil) = 1.4 a 3.kvartil = 4.7,
min = 0.1, max = 7.8, modus = 0.2.







Z povahy dat usuzujeme, že by data mohla mít exponenciální rozdělení. Zkusíme tedy proložit histogram (vynormovaný tak, aby plocha všech sloupců dohromady byla rovna 1) hustotou a empirickou distribuční funkci teoretickou distribuční funkcí s parametrem $\lambda = 1/\bar{x}_{21}$ *(neboť teoreticky máme pro n.v.* X *s exponenciálním rozdělením* $\mathbb{E}X = 1/\lambda \Rightarrow \lambda = 1/\mathbb{E}X$*).*



Vidíme, že obě křivky hezky kopírují příslušná rozdělení, tudíž můžeme data považovat za expoenciálně rozdělená.

2.2 Bodové a intervalové odhady

Příklad 2.2 *Uvažujme situaci, kdy v prvních deseti dnech lyžařské sezony obsloužil kiosek na sjezdovce postupně 224, 225, 209, 210, 201, 203, 239, 205, 223 a 215 zákazníků. Majitel (bývalý matfyzák) by si rád udělal model počtu obsloužených zákazníků na následující dny, aby mohl lépe plánovat nákup surovin. Jelikož jde o systém hromadné obsluhy, lze očekávat, že počty příchozích zákazníků se budou řídit Poissonovým rozdělením. Zbývá tedy odhadnou parametr tohoto rozdělení z dat. Jelikož víme, že* $\mathbb{E}X = \lambda$ *pro* $X \sim Pois(\lambda)$*, tak se nám úloha zredukovala na odhad střední hodnoty. Přirozeným odhadem střední hodnoty z dat je aritmetický průměr naměřených hodnot, tedy* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$*, kde* x_i*,* $i = 1, ..., n$ *jsou naměřené hodnoty (data). V našem případě dostaneme* $\bar{X} = 215.4$*, tedy odhad* $\hat{\lambda} = 215.4$*. Poznamenejme, že značení* $\hat{\theta}$ *se obecně používá pro označení odhadu parametru* θ*.*

To, co jsme v předchozím příkladu provedli, byl bodový odhad parametru Poissonova rozdělení. Pojďme se nyní podívat na tuto problematiku obecněji.

Definice 2.6 ***Bodový odhad*** *parametru* θ *je jakákoliv funkce náhodného výběru* $\hat{\theta}(\mathbb{X})$*, jejíž funkční předpis nezávisí na* θ*.*

Výše zmíněná definice je sice matematicky přesná, avšak pro naše potřeby příliš obecná a nepřehledná. Budeme si proto pod bodovým odhadem představovat nikoliv funkci náhodného výběru $\hat{\theta}(\mathbb{X})$, nýbrž samotné číslo $\hat{\theta}$,

které získáme z realizace $(x_1, x_2 \ldots, x_n)$ náhodného výběru $\mathbb{X}$, tedy z dat, a které co nejpřesněji popisuje hledaný parametr θ.

Předchozí definice nám sice zavádí pojem bodový odhad, ale už nám nedává žádný návod, jak v konkrétním případě bodový odhad vytvořit. Ukážeme si dvě metody, které tento problém řeší. Poznamenejme, že definice 2.6 neklade žádné požadavky na volbu funkce $\hat{\theta}(\mathbb{X})$, a tedy připouští i naprosto nesmyslné verze odhadu (např. v příkladu 2.2 bychom klidně mohli volit $\hat{\lambda} = 1000$ a s předchozí definicí by to nekolidovalo, ale prakticky by to byl nesmysl).

2.2.1 Metoda momentů

Mějme $(x_1, x_2 \ldots, x_n)$ realizaci náhodného výběru $\mathbb{X} = (X_1, X_2 \ldots, X_n)$, kde distribuční funkce F_X náhodných veličin X_i, $i = 1, \ldots, n$ závisí na parametrech $\theta_1, ..., \theta_k \in \Theta$, kde Θ je množina, z níž může parametr pocházet (např. nezáporná reálná čísla). Předpokládejme, že tzv. i-té momenty $\mathbb{E}X_1^i$ jsou konečné pro všechna $i = 1, ...k$. Tyto momenty rovněž závisejí na $\theta_1, ..., \theta_k$. Pak položením

$$\mathbb{E}X^i = m_i,$$

kde m_i je $i-$tý výběrový moment získaný jako

$$m_i = \frac{1}{n} \sum_{j=1}^{n} x_j^i$$

pro všechna $i = 1, ...k$, získáme soustavu k rovnic o k neznámých $\theta_1, ..., \theta_k$, jejímž řešením jsou odhady $\hat{\theta_1}, ..., \hat{\theta_k}$.

Poznámka 2.3 *V příkladu 2.2 jsme tuto metodu intuitivně použili. Jelikož jsme pracovali pouze s jedním parametrem λ, potřebovali jsme pouze první moment (střední hodnotu) a m_1, což je aritmetický průměr naměřených hodnot.*

2.2.2 Metoda maximální věrohodnosti

Mějme $(x_1, x_2 \ldots, x_n)$ realizaci náhodného výběru $\mathbb{X}$ z rozdělení s pravděpodobnostmi $P_\theta(X = .)$ nebo s hustotou $f_\theta(.)$, kde toto rozdělení závisí na nějakém parametru $\theta \in \Theta$. Odhad $\hat{\theta}$ je maximálně věrohodným odhadem, jestliže

$$\prod_{i=1}^{n} P_{\hat{\theta}}(X_1 = x_i) = \max_{\theta \in \Theta} \prod_{i=1}^{n} P_\theta(X = x_i)),$$

resp.

$$\prod_{i=1}^{n} f_{\hat{\theta}}(x_i) = \max_{\theta \in \Theta} \prod_{i=1}^{n} f_{\theta}(x_i).$$

Funkci $L(\theta) = \prod_{i=1}^{n} P_\theta(X_1 = x_i)$ (diskrétní případ), resp. $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$ (spojitý případ), se říká **věrohodnostní funkce**. Pro snadnější výpočet se často pracuje s **logaritmicko-věrohodnostní funkcí** $l(\theta) = \ln(L(\theta))$.

Příklad 2.3 *Vraťme se nyní v příkladu 2.2 a zkusíme ho nyní vyřešit metodou maximální věrohodnosti. Hledáme takové $\lambda > 0$, které maximalizuje hodnotu funkce $L(\lambda) = \prod_{i=1}^{n} P_\theta(X = x_i) = \prod_{i=1}^{n} (e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!})$. Pro další výpočet použijeme logaritmicko-věrohodnostní funkci, tedy budeme hledat $\lambda > 0$ takové, aby maximalizovalo hodnotu*

$$l(\theta) = \ln(L(\theta)) = \sum_{i=1}^{n} \ln(e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}) = \sum_{i=1}^{n} (-\lambda + x_i \ln(\lambda) - \ln(x_i!))$$
$$= -n\lambda - \sum_{i=1}^{n} \ln(x_i!) + \ln(\lambda) \sum_{i=1}^{n} x_i.$$

$$\frac{\partial l(\theta)}{\partial \theta)} = -n + \frac{1}{\lambda} \sum_{i=1}^{n} x_i = 0,$$

z toho dostaneme odhad $\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$. Jak vidíme, v daném příkladu nám obě metody dají stejný odhad.

Nyní si vyzkoušíme obě metody na následujícím příkladě.

Příklad 2.4 *V prvním týdnu na kolejích 17. listopadu student matfyzu při cestě na univerzitu čekal na zastávce autobusu 112 postupně 9.67, 9.25, 5.05, 4.37 a 1.70 minut. Odhadněte interval mezi příjezdy autobusu* 112*, pokud student chodil na zastávku nezávisle na jízdním řádu?*

Řešení
Jelikož student chodí na zastávku nezávisle na jízdním rádu, dá se očekávat, ze čas T, po který bude student čekat na příjezd autobusu, má rovnoměrné rozdělení $R(0, a)$, kde $a > 0$ je neznámý parametr (a je interval mezi příjezdy autobusu, tedy horní hranice doby čekání). Označme $x_i, i = 1, ...n$, naměřené hodnoty (v našem případě je $n = 5$).

- Metoda momentů:
  $\mathbb{E}X = \frac{a}{2}$, tudíž $m_1 = \bar{x} = \frac{\hat{a}}{2}$, a tedy $\hat{a} = 2\bar{x} = 2 \cdot 6.01 = 12.02$.

- Metoda maximální věrohodnosti:
  $f_X(x) = \frac{1}{a}$ pro $x \in (0, a)$ a $f_X(x) = 0$ pro $x \notin (0, a)$, tedy $L(a) = \frac{1}{a^n}$ jsou-li všechna $x_i \in (0, a)$, a $L(a) = 0$ v jiném případě. Z toho dostamene přímo odhad $\hat{a} = \max\{x_1, ..., x_n\} = 9.67$.

V tomto příkladu nám metoda maximální věrohodnosti dala jiný odhad než metoda momentů. Otázkou je, který z těchto odhadů bychom si měli vybrat, nebo přesněji, na základě jakých vlastností bychom měli mezi různými odhady volit. Je třeba si opět uvědomit, že z hlediska teorie je odhad parametru náhodná veličina, neboť je to funkce náhodného výběru, a tedy při porovnávání více odhadů a určování vlastností odhadů je třeba k nim přistupovat jako k náhodným veličinám. Máme-li nějaký odhad $\hat{\theta}$, pak nás zajímá zejména jeho střední hodnota $\mathbb{E}\hat{\theta}$ a rozptyl $var\hat{\theta}$. U střední hodnoty chceme, aby se rovnala odhadovanému parametru (tzv. **nestrannost** odhadu), u rozptylu bychom zase rádi, aby byl co možná nejmenší, tj. aby při různých měřeních vycházely odhady co nejpodobnější. Další vlastnost, kterou požadujeme, je, aby s přibývajícím počtem dat odhad konvergoval k skutečné hodnotě parametru θ (tzv. **konzistence** odhadu). Zavedeme si tedy nyní tyto vlastnosti bodového odhadu.

Definice 2.7 *Uvažujme náhodný výběr $\mathbb{X} = (x_1, ..., x_n)$ rozsahu n. Pak se bodový odhad $\hat{\theta}(\mathbb{X})$ nazývá*

- ***nestranný**, jestliže $\mathbb{E}\hat{\theta}(\mathbb{X}) = \theta$.*
- ***konzistentní**, jestliže pro libovolné $\varepsilon > 0$ platí $\lim_{n\to\infty} P(|\hat{\theta}(\mathbb{X}) - \theta| > \varepsilon) = 0$.*

Poznámka 2.4 *Konvergence použitá v definici konzistentního odhadu se nazývá **konvergence v pravděpodobnosti**. Kromě ní se v teorii pravděpodobnosti vyskytuje i silnější konvergence, a to konvergence **skoro jistě** definovaná jako $P(\lim_{n\to\infty} |\hat{\theta}(\mathbb{X}) - \theta| > \varepsilon) = 0$.*

Použijeme-li výběrový průměr $\bar{X}_n$ jako odhad střední hodnoty $\mathbb{E}X$, pak konzistenci tohoto odhadu shrnuje následující věta.

Věta 2.1 (Silný zákon velkých čísel)
Nechť $\{X_n\}_{n=1}^{\infty}$ je posloupnost nezávislých stejně rozdělených náhodných veličin s konečnou střední hodnotou $\mathbb{E}X_1 = \mu$. Pak pro $n \to \infty$ platí

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \to \mu$$

v pravděpodobnosti i skoro jistě.

2.3 Testování hypotéz

Nechť $\mathbb{X}$ je náhodný výběr z rozdělení, které závisí na parametru $\theta \in \Theta$. Tvrzení, že θ patří do nějaké množiny Θ_0, se nazývá **nulová hypotéza** (značíme $H_0 : \theta \in \Theta_0$). Na základě náhodného výběru $\mathbb{X}$ testujeme nulovou hypotézu vůči **alternativní hypotéze** $H_A : \theta \in \Theta \setminus \Theta_0$. K tomu stanovíme množinu W (tzv. **kritický obor**) tak, že H_0 zamítáme ve prospěch H_A, jestliže $\mathbb{X} \in W$, v opačném případě H_0 ve prospěch H_A nezamítáme.

Většinou testujeme $H_0 : \theta = \theta_0$, kde θ_0 je konkrétní hodnota, takže přirozenou alternativou je $H_A : \theta \neq \theta_0$. Občas však dává větší smysl testovat H_0 vůči $H_A : \theta > \theta_0$ nebo $H_A : \theta < \theta_0$, jelikož opačná situace nedává v tu chvíli praktický smysl.

Při testování mohou nastat následující situace:

- H_0 platí a test ji nezamítá $\surd$,
- H_0 neplatí a test ji zamítá $\surd$,
- H_0 platí a test ji zamítá $\rightarrow$ chyba prvního druhu,
- H_0 neplatí a test ji nezamítá $\rightarrow$ chyba druhého druhu.

Testování pak probíhá tak, že si zvolíme hodnotu α (obvykle 0.05, někdy 0.01 nebo 0.1) a kritický obor W konstruujeme tak, aby chyba prvního druhu nebyla větší než (obvykle byla rovna) α. Takové α se nazývá testovací hladina.

Ukažme si testování hypotéz na testu o střední hodnotě normálního rozdělení známém pod názvem **jednovýběrový t-test**. Nechť $\mathbb{X}$ je náhodný výběr z rozdělení $N(\mu, \sigma^2)$, kde $\sigma^2 > 0$ a ani jeden parametr není známý. Z hlubší teorie víme, že náhodná veličina $T = \frac{\bar{X}_n - \mu}{S_n}\sqrt{n}$ má tzv. **Studentovo t_{n-1}-rozdělení**, jehož kvantily lze najít ve statistických tabulkách. Testování $H_0 : \mu = \mu_0$ vůči $H_A : \mu \neq \mu_0$ pak probíhá ve dvou krocích:

1. Spočítáme tzv. testovou statistiku (hodnotu) $T_0 = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$.
2. Jestliže $|T_0| \geq t_{1-\alpha/2,n-1}$, zamítáme H_0 ve prospěch H_A, v opačném případě H_0 ve prospěch H_A nezamítáme.

Testování $H_0 : \mu = \mu_0$ vůči $H_A : \mu > \mu_0$ probíhá analogicky:

1. Spočítáme hodnotu $T_0 = \frac{\bar{X}_n - \mu_0}{S_n}\sqrt{n}$.
2. Je-li $T_0 \geq t_{1-\alpha,n-1}$, zamítáme H_0 ve prospěch H_A, v opačném případě H_0 ve prospěch H_A nezamítáme.

Nulová hypotéza $H_0 : \mu = \mu_0$ vůči $H_A : \mu < \mu_0$ je pak zamítnutá v případě, že $T_0 \leq t_{\alpha,n-1} = -t_{1-\alpha,n-1}$.

Příklad 2.5 *Výrobce tvrdí, že spotřeba jím vyráběného automobilu je 8 l/100km. Průměrná spotřeba u 49 uživatelů ale byla 8.4 l/100km. Naměřen byl dále výběrový rozptyl 2.56. Testujte na hladině 5%, zda měl výrobce pravdu.*

Řešení:
Za nulovou, resp. alternativní, hypotézu si zvolíme
H_0 : spotřeba = 8 l/100km,
H_A : spotřeba $\neq$ 8 l/100km.
Za platnosti H_0 má náhodná veličina

$$T = \frac{\bar{X}_n - 8}{\sqrt{S_n^2}}\sqrt{n}$$

rozdělení t_{48}. Vypočteme proto hodnotu

$$T_0 = \frac{8.4 - 8}{\sqrt{2.56}} \cdot 7 = 1.75.$$

Jelikož $|T_0| < t_{48;0.975} = 2.01$, hypotézu H_0 ve prospěch H_A nezamítáme.

Vzhledem k situaci (hádáme se s výrobcem, že spotřebu záměrně podhodnocuje) dává větší smysl si za nulovou, resp. alternativní, hypotézu zvolit
H_0 : spotřeba = 8 l/100km,
H_A : spotřeba > 8 l/100km.
Hodnotu $T_0 = 1.75$ pak porovnáme s kvantilem $t_{48;0.95} = 1.68$, neboť alternativě tentokrát nahrává "horních 5%". A jelikož $T_0 > t_{48;0.95} = 1.68$, hypotézu H_0 ve prospěch této H_A zamítáme.

Poznámka 2.5 *Testování hypotéz je velice široká disciplína, o níž byla napsána spousta knih. Jejich princip je však vždy stejný, jako jsme si ukázali, tj. spočítáme z dat hodnotu, tzv. testovou statistiku, u níž známe její pravděpodobnostní rozdělení za platnosti nulové hypotézy, a tuto testovou statistiku porovnáme s příslušným kvantilem. Podle výsledku porovnání pak hypotézu (ne)zamítáme.*

Kapitola 3

Regresní analýza

3.1 Základní pojmy

Uvažujme náhodný vektor $\mathbb{X} = (X_1, ..., X_r)$, náhodnou veličinu Y a předpokládejme, že Y na $\mathbb{X}$ nějakým způsobem závisí. Úkolem regresní analýzy je nalézt funkční závislost Y na $\mathbb{X}$, tj. najít takovou funkci f, že

$$Y = f_{\theta_1,...,\theta_p}(X_1, ..., X_r) + \epsilon, \tag{3.1}$$

kde tzv. chybový člen ϵ je náhodná veličina s nulovou střední hodnotou a rozptylem σ^2. Aby tato funkce dobře popisovala situaci, je třeba, aby rozptyl σ^2 nebyl příliš velký. Často navíc předpokládáme normalitu ϵ.

Uveďme si nejprve některé základní pojmy. Vztah (3.1) se nazývá **regresní model**. Funkce f se nazývá **regresní funkce**. Číslům $\theta_1, ..., \theta_p$ se říká **parametry regrese**. Náhodný vektor $\mathbb{X} = (X_1, ..., X_r)$ se nazývá **vysvětlující proměnná**. Náhodná veličina Y je pak **vysvětlovaná proměnná**.

K odhadu parametrů regrese se téměř výhradně používá metoda nejmenších čtverců, která funguje následovně. Nechť $(y_i, x_{1_i}, ..., x_{r_i}), i = 1, ..., n$ je n pozorování vektoru $(Y, X_1, ..., X_r)$. Označme

$$S(\theta_1, ..., \theta_p) = \sum_{i=1}^{n} (y_i - f(x_{1_i}, ..., x_{r_i}, \theta_1, ..., \theta_p))^2$$

Výraz $S(\theta_1, ..., \theta_p)$ se nazývá reziduální součet čtverců. Metoda spočívá v minimalizaci

$$\begin{aligned}(\hat{\theta}_1, ..., \hat{\theta}_p) &= \underset{\theta_1,...,\theta_p}{\arg\min} S(\theta_1, ..., \theta_p) \\ &= \underset{\theta_1,...,\theta_p}{\arg\min} \sum_{i=1}^{n} (y_i - f(x_{1_i}, ..., x_{r_i}, \theta_1, ..., \theta_p))^2,\end{aligned}$$

která se provádí podobně jako v případě metody maximální věrohodnosti parciální derivací podle jednotlivých parametrů.

3.2 Lineární regrese

Nejjednodušší formou regrese je případ, kdy funkce f je lineární. Obecný tvar (nazývaný vícenásobná regrese) je

$$Y = a + b_1 X_1 + ... + b_r X_r + \epsilon.$$

Často však hledáme závislost veličiny Y pouze na jediné veličině X (tj. $\mathbb{X}$ je jednorozměrný vektor). Tento případ, tj. vztah

$$Y = a + bX + \epsilon,$$

se nazývá jednoduchá regrese. Máme-li data $(y_i, x_{1_i}, ..., x_{r_i}), i = 1, ..., n$ a uvažujeme-li jednoduchou regresi, pak hodnoty

$$e_i = y_i - a - bx_i, \quad i = 1, ..., n$$

se nazývají rezidua a lze je považovat za realizace chybového členu ϵ. Reziduální součet čtverců pro jednoduchou regresi je

$$S(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

K nalezení odhadů parametrů a a b v modelu jednoduché regrese použijeme výše popsanou metodu nejmenších čtverců. Hledáme-li minimum $S(a, b)$, dostáváme rovnice

$$\frac{\partial}{\partial a} S(a, b) = -2 \sum_{i=1}^{n} (y_i - a - bx_i) = 0,$$
$$\frac{\partial}{\partial b} S(a, b) = -2 \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0,$$

z nichž dostáváme

$$\hat{b} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$
$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Pro testovaní, zda se koeficienty modelu výrazně liší od nuly, pak používáme statistiky

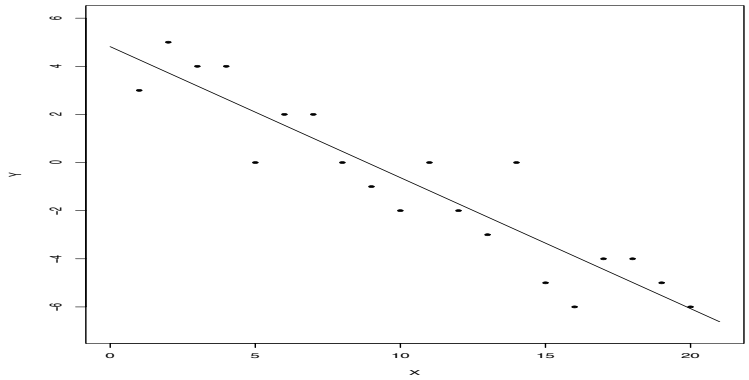
$$T_a = \frac{\hat{a}}{S_a}, \quad \text{resp.} \quad T_b = \frac{\hat{b}}{S_b},$$

které mají za platnosti hypotézy $H_0 : a = 0$, resp. $H_0 : b = 0$, t_{n-2}-rozdělení. Tedy hypotézu $H_0 : a = 0$, resp. $H_0 : b = 0$, zamítáme ve prospěch $H_A : a \neq 0$, resp. $H_A : b \neq 0$, na hladině významnosti α, pokud $|T_a| \geq t_{n-2,1-\frac{\alpha}{2}}$, resp. $|T_b| \geq t_{n-2,1-\frac{\alpha}{2}}$.

Příklad 3.1 *Mějme následující pozorování:*

x_i	1	2	3	4	5	6	7	8	9	10
y_i	3	5	4	4	0	2	2	0	-1	-2
x_i	11	12	13	14	15	16	17	18	19	20
y_i	0	-2	-3	0	-5	-6	-4	-4	-5	-6

Bodové odhady parametrů a a b jsou $\hat{a} = 4.816, \hat{b} = -0,544$ *a proložení dat regresní přímkou vypadá následovně:*



Pro ověření (ne)nulovosti koeficientů dostáváme

$$|T_a| = |7.6867| = 7.6867 > 2.1009 = t_{18,0.975}$$

a

$$|T_b| = |-10.40844| = 10.4084 > 2.1009 = t_{18,0.975},$$

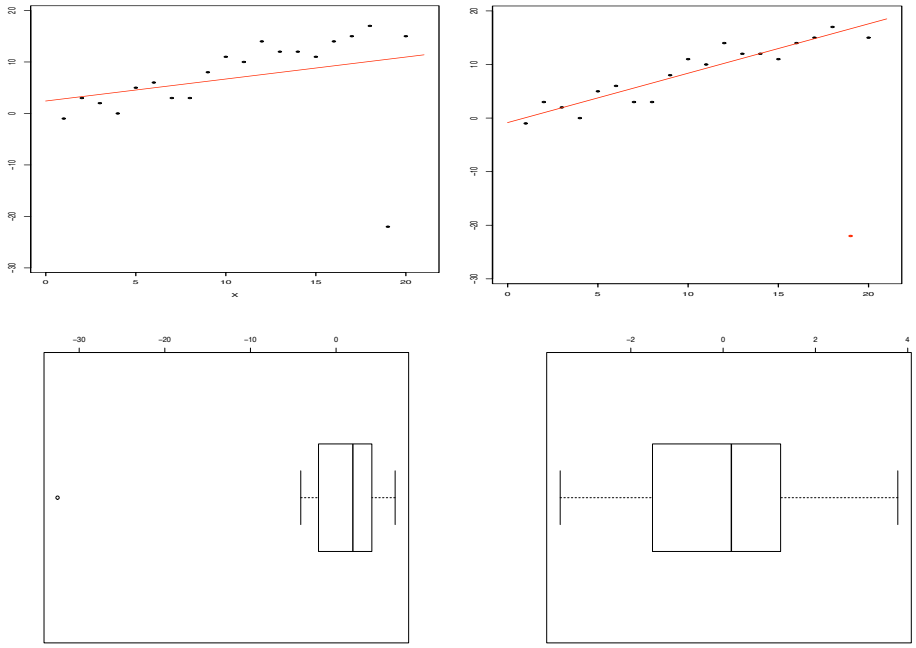
tedy zamítáme jak hypotézu $H_0 : a = 0$, *tak* $H_0 : b = 0$, *ve prospěch* $H_A : a \neq 0$, *resp.* $H_A : b \neq 0$, *na hladině 5%.*

Součástí regresní analýzy je tzv. analýza reziduí a identifikace odlehlých pozorování. Ta vychází z výše zmíněných předpokladů na chybový člen ϵ, aby byl náhodnou veličinou s nulovou střední hodnotou a normálním rozdělením.

Je tedy třeba ověřit tyto předpoklady na odhadech chybových členů $e_i = y_i - a - bx_i$. Toto ověření se nazývá analýza reziduí a používají se pro ni již různé metody, např. vykreslení histogramu, jehož vrcholy by měly připomínat Gaussovu křivku, nebo boxplotu, který by měl být symetrický s mediánem blízko hodnotě 0. Při analýze reziduí lze takto detekovat i odlehlá pozorování.

V následujícím příkladě si ukážeme, jak může jedno odlehlé pozorování ovlivnit regresní přímku a jak se graficky projeví v boxplotu.

Příklad 3.2 *Regresní přímka proložená všemi daty, daty bez odlehlého pozorování a boxploty příslušných reziduí vypadají následovně:*



Výběr vhodného regredního modelu lze pak provést např. pomocí koeficientu determinace definovaného jako

$$R^2 = \frac{\sum_{i=1}^{n}(a + bx_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Jeho hodnoty se pohybují v intervalu (0,1), přičemž větší hodnoty znamenají větší úspěšnost regrese. Pro předešlý příklad dostáváme $R^2 = 0.084$ *(špatná data), resp.* $R^2 = 0.876$ *(data bez odlehlého pozorování).*

Literatura

[1] Anděl J.: Matematika náhody. MatfyzPress, Praha, 2007.

[2] Dupač V., Hušková M.: Pravděpodobnost a matematická statistika. Karolinum, Praha, 2013.

[3] Mrkvička T., Petrášková V.: Úvod do teorie pravděpodobnosti. PF JU, České Budějovice, 2008.