# PRAVDĚPODOBNOST A STATISTIKA
## aneb Krátký průvodce skripty [1] a [2]

**Použitá literatura:**

**[1]:** J.Reif, Z.Kobeda: Úvod do pravděpodobnosti a spolehlivosti, ZČU Plzeň, 2004 (2. vyd.)

**[2]:** J.Reif: Metody matematické statistiky, ZČU Plzeň, 2004 (2. vyd.)

**Náhodné jevy** jsou podmnožiny množiny všech možných výsledků nějakého pokusu (podrobněji [1], str. 9-10).

**Nejčastějšími operacemi s jevy jsou**:

- sjednocení dvou jevů: $A \cup B$ ($A$ nebo $B$)
- průnik dvou jevů: $A \cap B$ ($A$ a $B$)
- negace jevu $A$: $\overline{A}$ (jev opačný k $A$, neboli doplňkový k $A$)

**Neslučitelné (disjunktní) jevy**:
Jevy $A$, $B$ se nazývají neslučitelné, je-li $A \cap B = \emptyset$.

**Pravděpodobnost** $P(A)$ jevu $A$ je limita relativní četnosti jevu $A$, zvětšujeme-li počet pokusů $n \to \infty$ ([1], str. 13). Např. $P(A) = 0,25$ tedy znamená, že při velkém množství pokusů nastane jev $A$ přibližně ve 25% případů.

Za určitých dosti specifických podmínek sice lze pravděpodobnost jevu počítat pomocí tzv. „klasické definice pravděpodobnosti“ ([1], str. 13), při praktických aplikacích však (přibližnou) pravděpodobnost stanovujeme právě pomocí relativní četnosti výskytu jevu při velkém množství pokusů (např. pravděpodobnost vzniku zmetku při neměnném výrobním postupu). Z takto stanovených pravděpodobností jevů pak umíme spočítat pravděpodobnosti jevů z nich odvozených. Používáme k tomu následující pravidla.

**Základní pravidla pro pravděpodobnost**:

1) Vždy $0 \leq P(A) \leq 1$.
2) $P(\overline{A}) = 1 - P(A)$.
3) Jsou-li jevy $A$, $B$ neslučitelné, pak
$$P(A \cup B) = P(A) + P(B)$$
4) Jsou-li jevy $A$, $B$ dva tzv. nezávislé jevy, pak
$$P(A \cap B) = P(A) \cdot P(B)$$

Poslední vztah je vlastně definicí nezávislosti dvou jevů. Podobně pravděpodobnost průniku většího počtu jevů počítáme součinem, pokud jde o jevy nezávislé. Pro jevy, které jsou závislé, tento vztah neplatí.

**Podmíněná pravděpodobnost**:
Jsou-li $A$, $B$ dva jevy a $P(B) > 0$, pak tzv. podmíněná pravděpodobnost jevu $A$ za předpokladu, že nastal jev $B$, se značí $P(A|B)$ a spočteme ji podle vzorce

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

Pokud jevy $A$, $B$ jsou závislé (tj. nejsou nezávislé), pak $P(A \cap B) \neq P(A) \cdot P(B)$ a odtud plyne, že $P(A|B) \neq P(A)$. V takovém případě může být $P(A|B)$ vyšší nebo nižší než $P(A)$.

**Náhodná veličina** je funkce, která náhodným jevům přiřazuje čísla. Náhodná veličina, která může nabývat jen hodnot z nějaké konečné množiny nebo jen celočíselných hodnot, se nazývá diskrétní veličinou. Náhodná veličina, která může nabývat všech hodnot v nějakém intervalu, se nazývá spojitou veličinou.

**Distribuční funkce** náhodné veličiny $X$ je funkce $F(x)$ reálné proměnné $x$, která každému $x \in (-\infty, +\infty)$ přiřadí pravděpodobnost jevu $X \leq x$, viz [1], str. 24.

Příklad: Je-li $X$ = počet ok při hodu obvyklou hrací kostkou, pak $F(-5) = 0$, $F\left(\frac{1}{2}\right) = 0$, $F(1) = \frac{1}{6}$, $F(2) = \frac{2}{6}$, $F(2,8) = \frac{2}{6}$, $F(3) = \frac{3}{6}$, $F(7,5) = 1$. Jde o nespojitou po částech konstantní funkci, což platí vždy, jde-li o náhodnou veličinu diskrétní. Graf funkce $F(x)$ je obdobou obrázku v [1], str. 25, bodů nespojitosti je ovšem v tomto případě šest.

Je-li $X$ spojitá náhodná veličina, která může nabývat všech hodnot v intervalu $(a, b)$, pak její distribuční funkce $F(x)$ je spojitá (nemá „skoky“) a je rostoucí na intervalu $(a, b)$. Z definice distribuční funkce $F(x)$ plyne, že pro libovolnou náhodnou veličinu je $F(x)$ neklesající funkcí a vždy platí

$$0 \leq F(x) \leq 1 \quad .$$

## DISKRÉTNÍ NÁHODNÁ VELIČINA

Důležitými charakteristikami náhodné veličiny $X$ jsou:

1) střední hodnota $E(X)$;
2) rozptyl $D(X)$, někdy značený $\sigma^2(X)$;
3) směrodatná odchylka $\sigma(X)$.

Pro způsob jejich výpočtu v případě diskrétní veličiny $X$, která může nabývat jen konečně mnoha hodnot, viz [1], str. 29, příklad 2.5. Výpočet střední hodnoty veličiny, která může nabývat hodnot z nějaké nekonečné množiny, vyžaduje hlubší znalosti matematické analýzy.

Základní diskrétní náhodné veličiny jsou popsány v [1], str. 32-38. Budeme pro ně používat označení:

$Bi(n,p)$ (tzv. binomické rozdělení s parametry $n$, $p$)

$A(p)$ (tzv. alternativní rozdělení s parametrem $p$)

$H(N,K,n)$ (tzv. hypergeometrické rozdělení s parametry $N$, $K$, $n$)

$Po(\lambda)$ (tzv. Poissonovo rozdělení s parametrem $\lambda$)

Hypergeometrické rozdělení se často aproximuje jednodušším binomickým rozdělením, viz [1], str. 35.

Binomické rozdělení lze za určitých podmínek aproximovat Poissonovým rozdělením, viz [1], str. 38. Tato aproximace je velmi užitečná, protože Poissonovo rozdělení má pouze jediný parametr, a proto lze pro toto rozdělení snadno zhotovit tabulky, viz např.[1], str. 104.

## SPOJITÁ NÁHODNÁ VELIČINA

### Hustota pravděpodobnosti spojité náh. veličiny

Hustota pravděpodobnosti se značí $f(x)$ a definuje se pouze pro náhodné veličiny spojité. Ten, kdo zná pojmy derivace a integrálu, si může zapamatovat, že hustota pravděpodobnosti $f(x)$ je derivací distribuční funkce $F(x)$ a $F(x)$ je tedy integrálem k $f(x)$ s vhodně zvolenou integrační konstantou.

Význam hustoty pravděpodobnosti lze ozřejmit následujícím popisem. Nechť náhodná veličina $X$ nabývá hodnot z intervalu $(a,b)$. Rozdělíme tento interval na menší intervaly (tzv. třídy), uskutečníme velký počet experimentů a budeme zaznamenávat počty výskytů veličiny v jednotlivých třídách (tzv. třídní četnosti). Dělíme-li třídní četnosti počtem experimentů, získáme tzv. relativní třídní četnosti, jejich součet je zřejmě 1. Tyto relativní četnosti přehledně graficky znázorníme pomocí tzv. histogramu, viz [2], str. 26. Výšky sloupců volíme tak, aby jejich obsahy byly rovny relativním třídním četnostem, tedy součet obsahů jejich sloupců je 1.

Zvětšujme nyní počet experimentů a zároveň volme stále jemnější dělení intervalu $(a,b)$, tj. větší počet tříd. Za určitých poměrně obecných předpokladů budou histogramy konvergovat k nějaké funkci $f(x)$, která se nazývá hustotou pravděpodobnosti veličiny $X$. Tato funkce je nezáporná a mezi ní a osou $x$ je celková plocha 1.

Protože jsme předpokládali, že veličina mohla nabývat pouze hodnot z intervalu $(a,b)$, je vně intervalu $(a,b)$ hustota pravděpodobnosti nulová. Pro některé náhodné veličiny se předpokládá, že mohou nabývat všech hodnot z intervalu $(-\infty,+\infty)$, a pro takové veličiny bude hustota pravděpodobnosti nenulová na celé reálné ose.

Je-li $-\infty \le x_1 < x_2 \le +\infty$, pak pravděpodobnost, že veličina $X$ padne do intervalu $(x_1,x_2)$, je rovna obsahu plochy omezené zdola osou $x$, shora hustotou $f(x)$ a ze stran svislými přímkami $x = x_1$ a $x = x_2$.

Analogické tvrzení platí, použijeme-li místo $(x_1,x_2)$ uzavřený interval $[x_1,x_2]$ Tyto poučky však platí jen pro spojitou náhodnou veličinu, protože pro diskrétní veličinu není hustota pravděpodobnosti definována.

### „Paradox“ spojité náhodné veličiny

Je-li $X$ spojitá náhodná veličina a $x_0$ reálné číslo, pak $P(X = x_0) = 0$. Graficky to lze zdůvodnit tak, že plošný obsah pod hustotou pravděpodobnosti $f(x)$ v mezích od $x_1 = x_0$ do $x_2 = x_0$ je nulový, neboť plošný útvar „degeneroval“ na úsečku kolmou k ose $x$.

### Výpočet pravděpododnosti jevu $\alpha < X < \beta$

Je-li $X$ spojitá náhodná veličina, pak pro libovolná reálná čísla $\alpha$, $\beta$ splňující nerovnost $\alpha \leq \beta$ platí

$$P(\alpha \leq X \leq \beta) = F(\beta) - F(\alpha) \ ,$$

kde $F$ je distribuční funkce veličiny $X$.

V důsledku výše popsaného „paradoxu“ spojité náhodné veličiny můžeme použít stejný vzorec také pro otevřený interval $(\alpha, \beta)$ nebo interval kombinovaný, viz [1], str. 39, Věta 2.9 - vlastnost 5).

### Základní spojité náhodné veličiny

**1) Rovnoměrné rozdělení pravděpodobnosti** na intervalu $(a, b)$. Pro tuto veličinu se používá symbolické označení $R(a, b)$. Hustota pravděpodobnosti je rovna konstantě $1/(b - a\,)$ na intervalu $(a, b)$ a nule vně tohoto intervalu. Všechny důležité informace lze pro toto rozdělení najít v [1], str. 40-41. K výpočtu pravděpodobnosti, že tato veličina patří do nějakého intervalu, nepotřebujeme znát v tomto případě distribuční funkci. Jak vyplývá z postupu v příkladu 2.10 v [1], str. 40, výpočet lze v tomto případě převést na výpočet obsahu určitého obdélníka.

**2) Exponenciální rozdělení pravděpodobnosti** s parametrem $\delta > 0$, viz [1], str. 41-42, budeme symbolicky označovat $Exp(\delta)$. Doporučujeme samostatně vyřešit příklady z [1], 2.9.7 a) a 2.9.8, výsledky jsou uvedeny v [1].

**3) Normální rozdělení pravděpodobnosti** s parametry $\mu$ a $\sigma^2$, viz [1], str. 43-44, budeme symbolicky označovat $N(\mu,\, \sigma^2\,)$. Chyby při měření se nejčastěji řídí přibližně normálním rozdělením. Normální rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$ se nazývá normální normované rozdělení. Při výpočtech s tímto rozdělením se zpravidla neobejdeme bez tabulky pro distribuční funkci $\Phi$ normálního normovaného rozdělení, kterou lze najít v [1], str. 105. Čtenáři doporučujeme samostatně vyřešit příklad 2.9.10 z [1].

### Centrální limitní věta

Mějme $n$ nezávislých náhodných veličin $X_1, \ldots, X_n$, které mají stejné rozdělení pravděpodobnosti se střední hodnotou $\mu_0$ a rozptylem $\sigma_0^2$. Označme

$$\mathcal{S} = \sum_{i=1}^{n} X_i \, ,$$

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

Podle tzv. centrální limitní věty mají pro velké $n$ veličiny $\mathcal{S}$ a $\overline{X}$ přibližně normální rozdělení s následujícími parametry:

$$\mathcal{S} \approx N(n\mu_0,\, n\sigma_0^2) \quad ,$$

$$\overline{X} \approx N\left(\mu_0,\, \frac{\sigma_0^2}{n}\right) .$$

Odtud je např. vidět, že výběrový průměr má stejnou střední hodnotu jako původní veličiny, ale jeho rozptyl je $n$-kráte menší. To je důvodem, proč se při odhadu střední hodnoty nějaké veličiny vyplatí provést více nezávislých měření a hodnoty průměrovat.

Čtenáři doporučujeme k prostudování [1], příklad 2.16 na str. 52 a cvičení 2.9.13.

### Kvantily spojitých rozdělení

Velmi důležitým pojmem je pojem kvantilu spojité náhodné veličiny, viz [1], str. 53-54. Např. 90 % (devadesáti-procentní) kvantil je číslo označované $x_{0,90}$ takové, že veličina je s pravděpodobností 0,90 menší než $x_{0,90}$ a s pravděpodobností 0,10 větší než $x_{0,90}$ (hodnoty přesně $x_{0,90}$ nabývá spojitá veličina s pravděpodobností 0). K procvičení tohoto pojmu jsou v [1] určeny příklady 2.9.7 b), 2.9.9 a 2.9.14. Kvantily normálního nornovaného rozdělení se zpravidla značí $u_p$ a pro vybrané pravděpodobnosti $p$ je lze najít v [1], str. 105, Tabulka 3.

### Kovariance a korelace

Vzájemný vztah dvou náhodných veličin $X_1$, $X_2$ lze do určité míry charakterizovat pomocí kovariance $\mathrm{cov}(X_1, X_2)$ a korelačního koeficientu $\varrho(X_1, X_2)$ těchto veličin. Označme $\mu_1$, $\mu_2$ střední hodnoty a $\sigma_1$, $\sigma_2$ směrodatné odchylky veličin $X_1$, $X_2$. Kovariance veličin $X_1$, $X_2$ je definována vztahem

$$\mathrm{cov}\,(X_1, X_2) = E([X_1 - \mu_1] \cdot [X_2 - \mu_2])\,.$$

Z definice je vidět, že je-li nadprůměrná hodnota $X_1$ obvykle doprovázena nadprůměrnou hodnotou $X_2$ a podprůměrná hodnota $X_1$ je obvykle doprovázena podprůměrnou hodnotou $X_2$, pak kovariance těchto dvou veličin je kladná (tyto veličiny se ovlivňují v „kladném smyslu“). Tak např. mezi výškou a váhou osob je kladná kovariance. Jestliže nadprůměrná hodnota $X_1$ je zpravidla doprovázena podprůměrnou hodnotou $X_2$ a podprůměrná hodnota $X_1$ je doprovázena nadprůměrnou hodnotou $X_2$, pak kovariance těchto dvou veličin je záporná.

Kovarianci lze také počítat pomocí tzv. výpočetního tvaru

$$\mathrm{cov}\,(X_1, X_2) = E(X_1\, X_2) \,-\, \mu_1\, \mu_2\,.$$

Čtenáři doporučujeme k pozornosti příklad 3.1 v [1], str. 59.

Korelační koeficient $\varrho(X_1, X_2)$ je definován vztahem,

$$\varrho(X_1, X_2) = \frac{\mathrm{cov}(X_1, X_2)}{\sigma_1 \sigma_2} .$$

Z definice je vidět, že korelační koeficient má stejné znaménko jako kovariance. Lze dokázat, že vždy platí

$$-1 \leq \varrho(X_1, X_2) \leq 1 .$$

Tak např. korelační koeficient mezi výškou a váhou osob je přibližně 0,4. Je-li $\varrho(X_1, X_2) = 0$, pak veličiny $X_1$, $X_2$ se nazývají nekorelované. Platí implikace, že jsou-li dvě veličiny nezávislé, pak jsou nekorelované.

**Odhady parametrů**

Provedeme-li $n$ nezávislých pokusů, při kterých sledujeme určitou náhodnou veličinu $X$, pak její zjištěné hodnoty $x_1, \ldots, x_n$ nazýváme náhodným výběrem (přesněji, náhodným výběrem z rozdělení dané náhodné veličiny). Počet $n$ se nazývá rozsahem náhodného výběru. Číslo

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

se nazývá výběrový (aritmetický) průměr a

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

se nazývá výběrová směrodatná odchylka.

Hodnoty $\bar{x}$ a $s^2$ jsou tzv. bodovými odhady střední hodnoty a rozptylu veličiny $X$ a pro $n \to \infty$ k nim v jistém smyslu konvergují,

$$\lim_{n\to\infty} \bar{x} = E(X) ,$$

$$\lim_{n\to\infty} s^2 = D(X) .$$

Tak např. při vytrvalém házení standardní hrací kostkou se bude průměr z počtu ok blížit k hodnotě 3,5.

V některých případech nás zajímá nikoliv bodový, ale tzv. intervalový odhad nějakého parametru. V takovém případě hledáme interval $(a, b)$ takový, aby sledovaný parametr ležel v tomto intervalu s předem zvolenou pravděpodobností $p$ (zpravidla se volí $p = 0,90$ nebo $p = 0,95$ nebo $p = 0,99$). Říkáme pak, že interval $(a, b)$ je $100\,p$-procentním intervalem spolehlivosti pro daný parametr. Číslo $p$ se v obecných vzorcích zpravidla píše ve tvaru $p = 1 - \alpha$, kde $\alpha$ je malé číslo (zpravidla $\alpha = 0,10$ nebo $\alpha = 0,05$ nebo $\alpha = 0,01$).

Tak např. je-li $x_1, \ldots, x_n$ náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$ a $n \geq 10$, pak přibližným $100\,(1-\alpha)\%$-ním intervalem spolehlivosti pro střední hodnotu $\mu$ je interval

$$\bar{x} - u_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + u_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} ,$$

kde $\bar{x}$ je výběrový průměr, $s$ je výběrová směrodatná odchylka a $u_{1-\frac{\alpha}{2}}$ je $100\,(1-\frac{\alpha}{2})\,\%$-ní kvantil rozdělení $N(0,1)$.

Tak např. přibližný 90%-ní interval spolehlivosti pro střední hodnotu $\mu$ má pro velké $n$ tvar

$$\bar{x} - u_{0,95}\,\frac{s}{\sqrt{n}} < \mu < \bar{x} + u_{0,95}\,\frac{s}{\sqrt{n}} ,$$

kde podle Tabulky 3 v [1], str. 105, je $u_{0,95} = 1,645$.

Pro malá $n$ je třeba místo kvantilů rozdělení $N(0,1)$ používat kvantily tzv. $t$-rozdělení (Studentova rozdělení); případné zájemce o tuto problematiku odkazujeme na [2], odst.3.3.3.

### Výběrový korelační koeficient

Zajímá-li nás vztah mezi dvěma veličinami $X, Y$ a provedeme-li $n$ nezávislých pokusů, máme k dispozici dvojice $(x_i, y_i)$ pro $i = 1, \ldots, n$. Pomocí těchto údajů lze spočítat tzv. výběrový korelační koeficient $r$, (vzorec si nebudeme uvádět, zájemce jej nalezne v literatuře), pomocí kterého odhadujeme korelační koeficient $\varrho(X, Y)$. Platí, že pro $n \to \infty$ výběrový korelační koeficient konverguje (v jistém smyslu) k $\varrho(X, Y)$. Vždy platí

$$-1 \leq r \leq 1 .$$

Jaké mohou být hodnoty výběrového korelačního koeficientu $r$ při různých rozmístěních bodů $(x_i, y_i)$ v rovině ukazuje obrázek ve [2], str. 84.

### Regresní funkce, metoda nejmenších čtverců

Regresní funkcí se míní závislost střední hodnoty nějaké náhodné veličiny (kterou nazýváme vysvětlovanou veličinou) na jiné nebo několika jiných veličinách (ty se pak nazývají vysvětlujícími veličinami). Pro jednoduchost uvažujme pouze jednu vysvětlující veličinu, kterou označme $x$, vysvětlovanou veličinu označme $y$. Mějme k dispozici $n$ dvojic $(x_i, y_i)$, kde $i = 1, \ldots, n$, které jsme získali $n$-násobným nezávislým opakováním pokusu. Příkladem může být výška a váha u $n$ náhodně vybraných mužů, cílem je přibližně popsat závislost váhy na výšce dospělých mužů.

V případě jedné vysvětlující veličiny je nejčastěji používaným modelem regresní funkce regresní přímka

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (i = 1, \ldots, n) ,$$

kde $\varepsilon_i$ $(i = 1, \ldots, n)$ jsou nějaké náhodné odchylky a $\beta_0$, $\beta_1$ jsou tzv. regresní koeficienty. Náhodné odchylky $\varepsilon_i$ způsobují, že i při platnosti výše uvedeného modelu nebudou body $(x_i, y_i)$ ležet přesně na přímce, ale pouze v její blízkosti, viz obrázek ve [2], str. 91. Regresní

koeficienty $\beta_0$, $\beta_1$ neznáme a chceme je určit tak, aby přímka $y = \beta_0 + \beta_1 x$ „co nejlépe“ vystihovala polohu bodů $(x_i, y_i)$ v rovině. Na obrázku ve [2] str. 91 jsou kromě bodů $(x_i, y_i)$ zakresleny také body $(x_i, \hat{y}_i)$, kde

$$\hat{y}_i = \beta_0 + \beta_1 x_i \quad (i = 1, \ldots, n)$$

jsou tzv. očekávané hodnoty; abychom tyto očekávané hodnoty mohli vyčíslit, musíme za neznámé regresní koeficienty $\beta_0$, $\beta_1$ dosadit jejich odhady. V našem případě, kdy modelem regresní funkce je přímka, leží body $(x_i, \hat{y}_i)$ přesně na přímce. Za „nejlepší“ volbu odhadů koeficientů $\beta_0$, $\beta_1$ se obvykle považuje taková, při které je minimalizován součet

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad .$$

Protože v anglickém jazyce se druhá mocnina vyjadřuje slovem „square“, vznikl doslovným překladem pojmenování tohoto postupu název „metoda nejmenších čtverců“.

Metodou nejmenších čtverců můžeme počítat „nejlepší“ koeficienty také v jiných modelech regresních funkcí. Na základě polohy bodů $(x_i, y_i)$ v rovině se např. můžeme rozhodnout, zda použijeme parabolu

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad (i = 1, \ldots, n)$$

nebo hyperbolu

$$y_i = \frac{\beta_0}{x_i} + \varepsilon_i \quad (i = 1, \ldots, n)$$

nebo nějakou jinou regresní funkci.

Zatímco volba typu regresní funkce je na našem rozhodnutí (a je tedy do určité míry subjektivní záležitostí), výpočet „nejlepších“ koeficientů pro zvolený typ regresní funkce se v praktických aplikacích téměř vždy provádí popsanou metodou nejmenších čtverců a používá se počítačových programů. Např. produkt EXCEL umožňuje aplikovat metodu nejmenších čtverců za účelem odhadu až šestnácti neznámých regresních koeficientů, viz [2], str. 236.

**Testování hypotéz**

Čtenáři doporučujeme k prostudování z textu [2] odstavec 4.1, úvodní pojednání odstavce 4.10, odst.4.10.1, příklad na str. 63 a odstavce 4.11.1 a 4.12. Samostatně by měl čtenář zvládnout cvičení 4.15.1 až 4.15.5 a 4.15.20 až 4.15.22.