MVŠO

ZÁKLADY STATISTIKY

STUDIJNÍ OPORA PRO KOMBINOVANÉ STUDIUM

Moravská vysoká škola Olomouc, o.p.s., 2024

ZÁKLADY STATISTIKY

RNDr. Jiří Fišer, Ph.D.

Autoři: RNDr. Jiří FIŠER, Ph.D.

Olomouc 2024

Obsah

Úvod

Vítejte ve světě statistiky

Vítejte ve studijní opoře pro předmět *Základy statistiky*, určené především studentům bakalářského studia ekonomicky a businessově zaměřených oborů. Skripta vás provedou základními pojmy a metodami statistiky s důrazem na jejich využití při analýze a zpracování dat v praxi.

Tato studijní opora se částečně překrývá s materiály pro navazující studium. V bakalářském studiu klademe důraz zejména na porozumění principům, správnou interpretaci výsledků a samostatné řešení typických úloh. V navazujícím studiu se témata dále rozšiřují (do hloubky i do šířky) a rozvíjejí se pokročilejší aplikace statistiky.

Struktura skript

Kapitoly jsou uspořádány tak, aby na sebe logicky navazovaly a umožnily postupné prohlubování znalostí. Každá kapitola rozvíjí dovednosti potřebné pro zvládnutí témat, která následují.

- **Kombinatorika** – Základní kombinatorické pojmy (variace, permutace, kombinace). Tyto nástroje jsou klíčové zejména pro pravděpodobnostní výpočty.
- **Pravděpodobnost jevů** – Základní principy pravděpodobnosti: klasická a geometrická pravděpodobnost, podmíněná pravděpodobnost a Bayesova věta.
- **Náhodná veličina a její rozdělení** – Pojem náhodné veličiny a rozdělení pravděpodobnosti; diskrétní a spojité rozdělení a jejich základní charakteristiky.
- **Základní typy rozdělení pravděpodobnosti** – Vybraná rozdělení často používaná v praxi: binomické, hypergeometrické, Poissonovo a normální rozdělení (včetně typických situací, kde je použít).
- **Náhodný vektor** – Více náhodných veličin současně: sdružené rozdělení, podmíněná rozdělení, kovariance a korelace (základ pro analýzu vztahů mezi veličinami).
- **Statistický soubor a jeho analýza** – Zpracování dat: třídění, tabulky četností, grafy, charakteristiky polohy a variability.
- **Regresní a korelační analýza** – Analýza vztahů mezi proměnnými: korelace a jednoduchá regrese jako nástroje pro popis a predikci.
- **Časové řady** – Základy analýzy dat v čase; jednoduché postupy pro popis trendu a sezónnosti.
- **Induktivní statistika** – Odhady parametrů, intervaly spolehlivosti a testování hypotéz; závěry o populaci na základě výběru.

- **Využití statistických softwarů** – Základní práce se softwarem (zejména MS Excel, dále R a Wolfram Alpha) pro výpočty a prezentaci výsledků.

Každá kapitola obsahuje teoretický výklad i praktické příklady. Cílem je, abyste nejen zvládli výpočty, ale především rozuměli významu a interpretaci získaných výsledků.

Co vás v kapitolách čeká

Každá kapitola začíná stručným uvedením tématu a cíli, kterých byste měli po jejím prostudování dosáhnout. Dále kapitoly obvykle obsahují:

- **Teoretický výklad** – Vysvětlení pojmů, metod a postupů včetně podmínek jejich použití.
- **Řešené příklady** – Typické úlohy s postupem řešení.
- **Rámečky** – Zvýraznění klíčových poznatků a shrnutí postupů.
- **Shrnutí** – Rekapitulace hlavních bodů kapitoly.
- **Kontrolní otázky a příklady** – Úlohy pro ověření porozumění. U vybraných příkladů jsou uvedeny výsledky v hranatých závorkách pro rychlou kontrolu.

Praktická aplikace a význam softwaru

Statistika je v ekonomické a manažerské praxi nepostradatelným nástrojem. Ve skriptech proto klademe důraz nejen na teorii, ale i na její praktické využití: výběr vhodné metody, správný výpočet a především interpretaci výsledků v kontextu úlohy.

V průběhu studia zjistíte, že statistický software (zejména MS Excel) výrazně usnadňuje výpočty a práci s daty. Pokud zvládnete i základy prostředí R, rozšíříte své možnosti analýzy dat a zvýšíte efektivitu i kontrolu nad postupem výpočtu.

Motivace a podpora

Cílem skript je pomoci vám osvojit si statistiku jako praktický jazyk pro práci s daty. Učte se postupně: nejprve porozumět zadání, zvolit vhodný postup, provést výpočet a na závěr výsledek smysluplně interpretovat. Chyby jsou přirozenou součástí učení; důležité je umět je rozpoznat a opravit.

Věříme, že pro vás budou tato skripta užitečným průvodcem a oporou při studiu i při řešení praktických úloh.

Kapitola 1

Kombinatorika

Po prostudování této kapitoly budete umět:

- rozlišovat mezi variacemi, kombinacemi a permutacemi (s opakováním i bez opakování),
- rozpoznat, kdy v úloze záleží na pořadí a kdy nikoli,
- rozlišovat situace s opakováním a bez opakování,
- řešit typové úlohy s využitím pravidla součinu a pravidla součtu (příp. principu inkluze a exkluze).

Klíčová slova:

Kombinatorika, faktoriál, kombinační číslo, variace bez opakování, variace s opakováním, kombinace bez opakování, kombinace s opakováním, permutace bez opakování, permutace s opakováním, pravidlo součinu, pravidlo součtu, princip inkluze a exkluze.

Náhled kapitoly

Kombinatorika se zabývá počítáním počtu možností, jak *vybrat* nebo *uspořádat* prvky z dané množiny. V této kapitole zavedeme a procvičíme tři základní typy úloh:

- **permutace** (uspořádání všech prvků),
- **variace** (uspořádání vybraných prvků),
- **kombinace** (výběr bez ohledu na pořadí).

U každého typu budeme rozlišovat, zda se prvky mohou *opakovat* (výběr s opakováním), nebo nikoli (výběr bez opakování). Základním vodítkem při volbě metody bude odpověď na dvě otázky: *Záleží na pořadí?* a *Je povoleno opakování?* Důraz bude kladen na řešení typových úloh, které tvoří přirozený základ pro následující kapitolu o pravděpodobnosti.

Cíle kapitoly

Po prostudování této kapitoly byste měli být schopni:

- rozhodnout, zda je daná situace permutace, variace, nebo kombinace,
- rozlišit úlohy s opakováním a bez opakování,
- správně zvolit a použít odpovídající vzorec a výsledek interpretovat,
- řešit typové úlohy s využitím pravidla součinu a pravidla součtu (příp. principu inkluze a exkluze).

Časová náročnost

Doporučený čas na zvládnutí kapitoly je přibližně 3–4 hodiny: přečtení výkladu, průběžné řešení ukázkových příkladů a samostatné procvičení na úlohách na konci kapitoly. Uvedený odhad předpokládá, že cílem není pouze dosadit do vzorce, ale také umět správně rozpoznat typ úlohy.

1.1 Základní pojmy a vlastnosti

Co je to kombinatorika?

Definice 1.1. Kombinatorika je část matematiky, která se zabývá *počítáním* počtu možností, jak z dané množiny prvků

- prvky **vybrat** (výběr) nebo
- prvky **uspořádat** (uspořádání),

přičemž rozhodujícími otázkami bývá, zda **záleží na pořadí** a zda je **povoleno opakování** prvků.

Kombinatorika se v základních úlohách nejčastěji opírá o tři pojmy:

- **Permutace** – uspořádání všech prvků (pořadí rozhoduje).
- **Variace** – uspořádání vybraných k prvků z n (pořadí rozhoduje).
- **Kombinace** – výběr k prvků z n bez ohledu na pořadí (pořadí nerozhoduje).

Kombinatorika je důležitým základem zejména pro teorii pravděpodobnosti a statistiku; využití má také v informatice, optimalizaci a kryptografii.

Kombinatorické pravidlo součinu

Definice 1.2. (Kombinatorické) pravidlo součinu říká: lze-li určitý postup rozdělit na k po sobě jdoucích kroků tak, že v i-tém kroku existuje n_i možností (pro $i = 1, \ldots, k$), pak celkový počet možností je

$$n_1 \cdot n_2 \cdot \cdots \cdot n_k.$$

Příklad 1.3. V restauraci jsou na výběr 3 druhy předkrmů, 4 druhy hlavních jídel a 2 druhy dezertů. Kolika způsoby lze sestavit menu (předkrm, hlavní jídlo, dezert)?

Řešení: V každém chodu volíme nezávisle jednu možnost, proto použijeme pravidlo součinu:

$$3 \cdot 4 \cdot 2 = 24.$$

Menu lze sestavit 24 způsoby. □

Kombinatorické pravidlo součtu

Definice 1.4. (Kombinatorické) pravidlo součtu říká: lze-li volbu provést *buď* jedním z n_1 způsobů *nebo* jedním z n_2 způsobů a tyto možnosti jsou **vzájemně neslučitelné** (tj. nelze je realizovat současně), potom celkový počet možností je

$$n_1 + n_2.$$

Příklad 1.5. V knihovně je 5 beletristických knih a 3 odborné knihy. Kolik různých knih si můžete vybrat, pokud si můžete vzít právě jednu knihu: buď beletrii, nebo odbornou?

Řešení: Možnosti výběru jsou neslučitelné (vybírá se právě jedna kniha), proto platí:

$$5 + 3 = 8.$$

Vybrat lze 8 různých knih. □

Princip inkluze a exkluze

Definice 1.6. Princip inkluze a exkluze slouží k určení počtu prvků ve sjednocení množin $A_1, \ldots, A_n$. Platí

$$|A_1 \cup A_2 \cup \cdots \cup A_n| = \sum_{i=1}^{n} |A_i| - \sum_{1 \le i < j \le n} |A_i \cap A_j| + \sum_{1 \le i < j < k \le n} |A_i \cap A_j \cap A_k| - \cdots + (-1)^{n+1} |A_1 \cap A_2 \cap \cdots \cap A_n|.$$

Vzorec střídavě přičítá velikosti jednotlivých množin a odčítá velikosti jejich průniků, aby se prvky započítané vícekrát korigovaly.

Speciální případ pro $n = 2$

Definice 1.7. Pro dvě množiny A a B platí

$$|A \cup B| = |A| + |B| - |A \cap B|.$$

Příklad 1.8. Ve třídě je 30 studentů. Kurz matematiky navštěvuje 15 studentů, kurz fyziky 10 studentů a oba kurzy 5 studentů. Kolik studentů navštěvuje alespoň jeden z těchto kurzů?

Řešení: Označme M množinu studentů navštěvujících matematiku a F množinu studentů navštěvujících fyziku. Potom

$$|M \cup F| = |M| + |F| - |M \cap F| = 15 + 10 - 5 = 20.$$

Alespoň jeden z kurzů navštěvuje 20 studentů. □

Speciální případ pro $n = 3$

Definice 1.9. Pro tři množiny A, B a C platí

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|.$$

Příklad 1.10. V knihovně (oddělení matematiky, fyziky a informatiky) je určitý počet knih. 40 z nich obsahuje kapitoly o matematice, 25 o fyzice a 35 o informatice. Dále 10 knih je současně o matematice i fyzice, 15 o matematice i informatice, 5 o fyzice i informatice a 3 knihy pokrývají všechny tři oblasti. Kolik knih je v oddělení celkem (předpokládejme, že jiné knihy v oddělení nejsou)?

Řešení: Označme M, F, I množiny knih podle toho, zda obsahují kapitoly o matematice, fyzice a informatice. Použijeme vzorec pro tři množiny:

$$|M \cup F \cup I| = |M| + |F| + |I| - |M \cap F| - |M \cap I| - |F \cap I| + |M \cap F \cap I|.$$

Dosadíme:

$$|M \cup F \cup I| = 40 + 25 + 35 - 10 - 15 - 5 + 3 = 73.$$

V oddělení je 73 knih. □

Faktoriál

Definice 1.11. Faktoriál nezáporného celého čísla n (značíme $n!$) je definován takto:

$$n! = \begin{cases} 1, & n = 0, \\ 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n, & n \in \mathbb{N},\ n \geq 1. \end{cases}$$

Příklad 1.12. Vypočtěte hodnotu 5!.

Řešení:

$$5! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120.$$

□

Faktoriál se používá zejména v kombinatorice (např. při výpočtu počtu permutací, variací a kombinací). Hodnota $n!$ roste s n velmi rychle, proto se ve výpočtech často pracuje se zkracováním výrazů s faktoriály.

1.2 Variace

Variace jsou *uspořádané* výběry z dané množiny prvků. Budeme rozlišovat dvě situace:

- **bez opakování** – každý prvek lze vybrat nejvýše jednou,
- **s opakováním** – prvky lze vybírat opakovaně.

1.2.1 Variace bez opakování

Příklad 1.13. Vypište všechny uspořádané dvojice ze základní množiny prvků $\{1, a, B\}$, pokud se prvky nemohou opakovat. Kolik jich je?

Řešení: Jde o „variace druhé třídy ze tří prvků bez opakování“ (též „2-prvkové variace ze tří prvků bez opakování“). Vypíšeme všechny možnosti:

$$(1,a),\ (a,1),\ (1,B),\ (B,1),\ (a,B),\ (B,a).$$

Celkem tedy dostáváme 6 uspořádaných dvojic. □

Při větších hodnotách n a k je vypisování všech možností nepraktické. Proto odvodíme vzorec pro počet variací.

Definice 1.14. Variace bez opakování jsou uspořádané k-prvkové výběry z n prvků, přičemž každý prvek může být vybrán nejvýše jednou. Počet variací k-té třídy z n prvků (bez opakování) je

$$V_k(n) = \frac{n!}{(n-k)!} = \underbrace{n\,(n-1)\cdots(n-k+1)}_{k \text{ činitelů}}.$$

Zde platí $0 \le k \le n$.

Příklad 1.15. Kolik různých uspořádaných trojic lze vybrat z množiny $\{1, 2, 3, 4, 5\}$, pokud se prvky nemohou opakovat?

Řešení: Jde o variace třetí třídy z pěti prvků bez opakování:

$$V_3(5) = \frac{5!}{(5-3)!} = \frac{5!}{2!} = \frac{120}{2} = 60, \qquad \text{příp.} \qquad V_3(5) = 5 \cdot 4 \cdot 3 = 60.$$

□

Příklad 1.16. Kolika způsoby lze obsadit první tři místa v závodě s 10 účastníky, pokud se o umístění nelze dělit?

Řešení: Pořadí (1., 2., 3. místo) je rozhodující a každý účastník může obsadit nejvýše jedno místo, proto použijeme variace bez opakování:

$$V_3(10) = \frac{10!}{(10-3)!} = 10 \cdot 9 \cdot 8 = 720.$$

□

1.2.2 Variace s opakováním

Definice 1.17. Variace s opakováním jsou uspořádané k-prvkové výběry z n prvků, přičemž prvky lze vybírat opakovaně. Počet variací k-té třídy z n prvků s opakováním je

$$V_k^*(n) = n^k = \underbrace{n \cdot n \cdot \cdots \cdot n}_{k \text{ činitelů}}.$$

Zde platí $k \geq 0$ a $n \geq 1$.

Příklad 1.18. Kolik různých trojciferných čísel lze vytvořit pomocí cifer $1, 2, 3, 4, 5$, pokud se cifry mohou opakovat?

Řešení: Na každé ze tří pozic lze zvolit jednu z 5 cifer, opakování je dovoleno, proto:

$$V_3^*(5) = 5^3 = 125.$$

□

Příklad 1.19. Kolik různých čtyřmístných PIN kódů lze vytvořit, pokud každé místo může obsahovat cifru od 0 do 9 a cifry se mohou opakovat?

Řešení: Jde o variace s opakováním, kde $n = 10$ a $k = 4$:

$$V_4^*(10) = 10^4 = 10\,000.$$

□

Příklad 1.20. Kolik různých značek lze vytvořit v Morseově abecedě, pokud se sestavují z teček a čárek do skupin o délce 1 až 3?

Řešení: Základní množina má $n = 2$ znaky (tečka a čárka) a opakování je dovoleno. Počet značek délky k je $V_k^*(2) = 2^k$. Protože délky 1, 2 a 3 představují neslučitelné případy, použijeme pravidlo součtu:

$$V_1^*(2) + V_2^*(2) + V_3^*(2) = 2^1 + 2^2 + 2^3 = 2 + 4 + 8 = 14.$$

□

1.3 Permutace

Permutace jsou *uspořádání všech* prvků dané množiny. Jde o speciální případ variací, kdy vybíráme $k = n$ prvků, takže pořadí vždy rozhoduje. Budeme rozlišovat permutace **bez opakování** (všechny prvky jsou různé) a **s opakováním** (některé prvky se opakují a jsou nerozlišitelné).

1.3.1 Permutace bez opakování

Definice 1.21. Permutace bez opakování jsou uspořádání všech n navzájem různých prvků. Počet permutací je

$$P(n) = n!.$$

Příklad 1.22. Vypište všechny permutace množiny prvků $\{1, a, B\}$ a ověřte, že jejich počet odpovídá vzorci.

Řešení: Vypíšeme všechny možnosti uspořádání tří různých prvků:

$$(1, a, B),\ (1, B, a),\ (a, 1, B),\ (a, B, 1),\ (B, 1, a),\ (B, a, 1).$$

Celkem je permutací 6, což odpovídá $P(3) = 3! = 6$. □

Příklad 1.23. Kolika způsoby lze uspořádat 6 různých knih na polici?

Řešení: Jde o permutace šesti prvků:

$$P(6) = 6! = 720.$$

□

1.3.2 Permutace s opakováním

Definice 1.24. Permutace s opakováním nastávají tehdy, když v souboru n prvků se některé prvky opakují a jsou *nerozeznatelné*. Nechť existuje k typů prvků a i-tý typ se opakuje n_i-krát, kde

$$n = n_1 + n_2 + \cdots + n_k.$$

Počet různých uspořádání je

$$P^*_{n_1,n_2,\ldots,n_k}(n) = \frac{n!}{n_1!\, n_2! \cdots n_k!}.$$

Vzorec zohledňuje, že prohození dvou stejných prvků nevytváří nové uspořádání.

Příklad 1.25. Vypište všechny permutace multmnožiny $\{1, a, a\}$ a ověřte, že jejich počet odpovídá vzorci.

Řešení: Rozlišitelná uspořádání jsou:

$$(1, a, a),\ (a, 1, a),\ (a, a, 1).$$

Celkem jsou 3. Zde je $n = 3$, prvek 1 se vyskytuje jednou ($n_1 = 1$) a prvek a dvakrát ($n_2 = 2$), proto

$$P^*_{1,2}(3) = \frac{3!}{1!\,2!} = \frac{6}{2} = 3.$$

□

Příklad 1.26. Kolik různých šesticiferných čísel lze vytvořit z číslic $1, 1, 2, 2, 2, 3$?

Řešení: Máme $n = 6$ číslic, přičemž 1 se opakuje dvakrát, 2 třikrát a 3 jednou, tedy $(n_1, n_2, n_3) = (2, 3, 1)$:

$$P^*_{2,3,1}(6) = \frac{6!}{2!\,3!\,1!} = \frac{720}{2 \cdot 6} = 60.$$

□

Příklad 1.27 (Uspořádání písmen ve slově)**.** Kolik různých uspořádání písmen lze vytvořit ze všech deseti písmen slova *STATISTIKA*?

Řešení: Ve slově *STATISTIKA* je $n = 10$ písmen. Počty opakování jsou:

$$S : 2, \quad T : 3, \quad A : 2, \quad I : 2, \quad K : 1.$$

Proto

$$P^*_{2,3,2,2,1}(10) = \frac{10!}{2!\,3!\,2!\,2!\,1!} = \frac{3\,628\,800}{2 \cdot 6 \cdot 2 \cdot 2} = \frac{3\,628\,800}{48} = 75\,600.$$

Celkem lze vytvořit 75 600 různých uspořádání. □

Příklad 1.28 (Tvorba řad korálků)**.** Máme 8 korálků, z nichž 4 jsou červené, 3 modré a 1 zelený. Kolik různých *řad* (lineárních uspořádání) korálků lze vytvořit, pokud korálky stejné barvy nerozlišujeme?

Řešení: Jde o permutace s opakováním: $n = 8$, počty opakování jsou $(4, 3, 1)$, tedy

$$P^*_{4,3,1}(8) = \frac{8!}{4!\,3!\,1!} = \frac{40\,320}{24 \cdot 6} = 280.$$

□

1.4 Kombinace

Kombinace jsou výběry prvků z dané množiny, při kterých *nezáleží na pořadí*. Budeme rozlišovat kombinace **bez opakování** (každý prvek lze vybrat nejvýše jednou) a kombinace **s opakováním** (prvky lze vybírat opakovaně).

Kombinační číslo

Definice 1.29. Kombinační číslo (binomický koeficient) $\binom{n}{k}$ udává počet způsobů, jak vybrat k prvků z n různých prvků **bez opakování** a **bez ohledu na pořadí**. Pro $0 \leq k \leq n$ platí

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}.$$

Příklad 1.30. Vypočítejte kombinační číslo $\binom{7}{3}$.

Řešení: Použijeme vzorec a vhodně zkrátíme:

$$\binom{7}{3} = \frac{7!}{3!\,4!} = \frac{7 \cdot 6 \cdot 5 \cdot 4!}{(3 \cdot 2 \cdot 1)\,4!} = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1} = \frac{210}{6} = 35.$$

□

1.4.1 Kombinace bez opakování

Definice 1.31. Kombinace bez opakování je výběr k prvků z n různých prvků, kde na pořadí nezáleží a každý prvek lze vybrat nejvýše jednou. Počet takových výběrů je

$$C_k(n) = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}, \qquad 0 \leq k \leq n.$$

Pozn.: V literatuře se často používá místo $C_k(n)$ přímo zápis $\binom{n}{k}$.

Příklad 1.32. Najděte všechny kombinace druhé třídy bez opakování z množiny $M = \{1, 2, 3, 4, 5\}$.

Řešení: Počet kombinací je

$$C_2(5) = \binom{5}{2} = \frac{5!}{2!\,3!} = \frac{5 \cdot 4}{2} = 10.$$

Jednotlivé dvojice (bez ohledu na pořadí) jsou:

$$\{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{2,3\}, \{2,4\}, \{2,5\}, \{3,4\}, \{3,5\}, \{4,5\}.$$

□

Příklad 1.33. Kolik různých pětičlenných týmů lze vybrat ze skupiny 12 studentů?

Řešení: Pořadí členů týmu nerozhoduje, proto použijeme kombinace bez opakování:

$$C_5(12) = \binom{12}{5} = \frac{12!}{5!\,7!} = \frac{12 \cdot 11 \cdot 10 \cdot 9 \cdot 8}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 792.$$

□

Příklad 1.34. Kolika způsoby lze vybrat 3 knihy z police, která obsahuje 7 různých knih?

Řešení: Pořadí vybraných knih nerozhoduje:

$$C_3(7) = \binom{7}{3} = 35.$$

□

Příklad 1.35. Kolika způsoby lze sestavit výbor složený ze 4 mužů a 3 žen, pokud máme k dispozici 8 mužů a 5 žen?

Řešení: Nejprve vybereme 4 muže z 8:

$$\binom{8}{4} = 70.$$

Poté vybereme 3 ženy z 5:

$$\binom{5}{3} = 10.$$

Podle pravidla součinu je celkový počet možností

$$\binom{8}{4}\binom{5}{3} = 70 \cdot 10 = 700.$$

□

1.4.2 Kombinace s opakováním

Definice 1.36. Kombinace s opakováním jsou výběry k prvků z n různých prvků, kde *nezáleží na pořadí* a *opakování je dovoleno*. Jinými slovy: vybíráme k prvků tak, že stejný prvek může být vybrán i vícekrát.

Počet kombinací s opakováním k-té třídy z n prvků je

$$C_k^*(n) = \binom{n+k-1}{k} = \binom{n+k-1}{n-1} = \frac{(n+k-1)!}{k!\,(n-1)!}, \qquad n \geq 1,\ k \geq 0.$$

Příklad 1.37. Najděte všechny kombinace druhé třídy s opakováním z množiny $M = \{1, 2, 3, 4, 5\}$.

Řešení: Zde je $n = 5$ a $k = 2$, proto

$$C_2^*(5) = \binom{5+2-1}{2} = \binom{6}{2} = 15.$$

Jednotlivé kombinace (bez pořadí, s možností opakování) jsou:

$\{1,1\}, \{1,2\}, \{1,3\}, \{1,4\}, \{1,5\}, \{2,2\}, \{2,3\}, \{2,4\}, \{2,5\}, \{3,3\}, \{3,4\}, \{3,5\}, \{4,4\}, \{4,5\}, \{5,5\}.$

Celkem tedy existuje 15 kombinací druhé třídy s opakováním. □

Příklad 1.38. Kolika způsoby lze vybrat 4 bonbóny ze 3 různých druhů, pokud nezáleží na pořadí a bonbóny se mohou opakovat?

Řešení: Jde o kombinace s opakováním ($n = 3$, $k = 4$):

$$C_4^*(3) = \binom{3+4-1}{4} = \binom{6}{4} = 15.$$

□

Příklad 1.39. Kolika způsoby lze rozdělit 10 jablek mezi 3 děti, pokud každé dítě může dostat libovolný počet jablek?

Řešení: Označme x_1, x_2, x_3 počet jablek pro jednotlivé děti. Hledáme počet řešení v nezáporných celých číslech rovnice

$$x_1 + x_2 + x_3 = 10.$$

To je ekvivalentní kombinacím s opakováním ($n = 3$, $k = 10$), tedy

$$C_{10}^*(3) = \binom{3+10-1}{10} = \binom{12}{10} = 66.$$

□

Příklad 1.40. Kolika způsoby lze rozdělit 8 identických bonbónů mezi 4 děti?

Řešení: Analogicky hledáme počet řešení v nezáporných celých číslech rovnice $x_1+x_2+x_3+x_4 = 8$. Proto

$$C_8^*(4) = \binom{4+8-1}{8} = \binom{11}{8} = 165.$$

□

Příklad 1.41. Kolika způsoby lze vybrat 6 květin z 5 druhů, pokud se mohou opakovat?

Řešení: Jde o kombinace s opakováním ($n = 5$, $k = 6$):

$$C_6^*(5) = \binom{5+6-1}{6} = \binom{10}{6} = 210.$$

□

Příklad 1.42. Zjistěte, kolik existuje různých kvádrů, pro něž platí, že délka každé hrany je přirozené číslo z intervalu $[2; 5]$, přičemž nezáleží na pořadí stran.

Řešení: Délky hran kvádru můžeme popsat trojicí (a, b, c), kde $a, b, c \in \{2, 3, 4, 5\}$ a nezáleží na pořadí (tj. trojice $(2, 3, 5)$ je totéž co $(5, 3, 2)$). Jde tedy o výběr 3 prvků z 4 hodnot s opakováním: $n = 4$, $k = 3$.

$$C_3^*(4) = \binom{4+3-1}{3} = \binom{6}{3} = 20.$$

Celkem existuje 20 různých kvádrů. □

1.4.3 Souhrnné příklady

Příklad 1.43. Jsou dány cifry $1, 2, 3, 4, 5$. Cifry nelze opakovat. Kolik je možno vytvořit z těchto cifer čísel, která jsou:

- a) pětimístná, sudá,
- b) pětimístná, končící dvojčíslím 21,
- c) pětimístná, menší než 30 000,
- d) trojmístná, lichá,
- e) čtyřmístná, větší než 2 000,
- f) dvojmístná nebo trojmístná.

Řešení: **ad a)** Pětimístné sudé číslo musí končit cifrou 2 nebo 4 (2 možnosti). Zbylé čtyři pozice vyplníme permutací zbývajících čtyř cifer:

$$2 \cdot P(4) = 2 \cdot 4! = 2 \cdot 24 = 48.$$

ad b) Číslo má tvar XXX21. Na první tři pozice lze dosadit libovolné uspořádání tří zbývajících cifer:

$$P(3) = 3! = 6.$$

ad c) Podmínka „menší než 30 000“ znamená, že první cifra je 1 nebo 2 (2 možnosti). Zbylé čtyři pozice vyplníme permutací zbývajících čtyř cifer:

$$2 \cdot P(4) = 48.$$

ad d) Trojmístné liché číslo musí končit cifrou 1, 3 nebo 5 (3 možnosti). Zbylé dvě pozice obsadíme dvěma různými ciframi ze zbývajících čtyř, přičemž pořadí rozhoduje (variace bez opakování):

$$3 \cdot V_2(4) = 3 \cdot (4 \cdot 3) = 36.$$

ad e) Čtyřmístné číslo větší než 2 000 má tisíce 2, 3, 4 nebo 5 (4 možnosti). Zbylé tři pozice obsadíme třemi různými ciframi ze zbývajících čtyř, pořadí rozhoduje:

$$4 \cdot V_3(4) = 4 \cdot (4 \cdot 3 \cdot 2) = 96.$$

ad f) Hledáme počet dvojmístných *nebo* trojmístných čísel (neslučitelné případy), proto použijeme pravidlo součtu:

$$V_2(5) + V_3(5) = (5 \cdot 4) + (5 \cdot 4 \cdot 3) = 20 + 60 = 80.$$

□

Příklad 1.44. Kolik různých státních poznávacích značek tvaru `4M9 XX-XX` existuje s alespoň dvěma trojkami? (Na místech `X` mohou být jen číslice.)

Řešení: Na čtyřech pozicích `X` počítáme řetězce číslic s alespoň dvěma trojkami, tj. s právě 2, 3 nebo 4 trojkami. Označme x_r počet značek s právě r trojkami.

4 trojky: jediná možnost `33-33`, tedy

$$x_4 = 1.$$

3 trojky: zvolíme pozici, na které *není* trojka (4 možnosti). Na zbývající pozici lze dát jednu z 9 číslic $\{0, 1, 2, 4, 5, 6, 7, 8, 9\}$:

$$x_3 = \binom{4}{1} \cdot 9 = 4 \cdot 9 = 36.$$

(Pozn.: ekvivalentně $x_3 = \frac{4!}{3!\,1!} \cdot 9$.)

2 trojky: nejprve zvolíme, na kterých 2 pozicích jsou trojky: $\binom{4}{2} = 6$ možností. Zbylé dvě pozice vyplníme libovolnými číslicemi z množiny 9 možností, přičemž opakování je dovoleno a pořadí pozic je dáno (variace s opakováním):

$$x_2 = \binom{4}{2} \cdot 9^2 = 6 \cdot 81 = 486.$$

Celkový počet požadovaných značek je

$$x = x_2 + x_3 + x_4 = 486 + 36 + 1 = 523.$$

□

Σ

V této kapitole jsme se seznámili se základními pojmy kombinatoriky, tj. s metodami pro počítání počtu možností *výběru* a *uspořádání* prvků. Klíčovým krokem při řešení úloh bylo vždy rozhodnout, zda **záleží na pořadí** a zda je **dovoleno opakování**.

Probrali jsme tři základní typy úloh:

- **Variace** – uspořádané výběry k prvků z n (pořadí rozhoduje), a to *bez opakování* i *s opakováním*.
- **Permutace** – uspořádání všech n prvků (speciální případ variací pro $k = n$), opět *bez opakování* i *s opakováním*.
- **Kombinace** – výběry k prvků z n bez ohledu na pořadí (pořadí nerozhoduje), *bez opakování* i *s opakováním*.

Dále jsme používali základní principy pro počítání počtu možností:

- **Pravidlo součinu** – pro postupy složené z několika po sobě jdoucích kroků (násobení počtu možností v jednotlivých krocích).
- **Pravidlo součtu** – pro volbu z několika *vzájemně neslučitelných* možností (sčítání počtu možností).
- **Princip inkluze a exkluze** – pro výpočet počtu prvků ve sjednocení množin se zohledněním průniků.

Cílem kapitoly bylo, abyste uměli správně rozpoznat typ úlohy, zvolit odpovídající postup a výsledek interpretovat.

1. Státní poznávací značku tvoří dvě písmena, tři číslice a další dvě písmena (formát `AAXXXAA`, kde `A` je písmeno a `X` číslice). Kolik různých značek lze vytvořit, pokud můžeme vybírat z 25 písmen a 10 číslic? [390 625 000]
2. Kolik různých šestimístných čísel lze sestavit z cifer 1, 2 a 3, pokud se cifry mohou opakovat? [729]
3. V MHD se kdysi používaly lístky s devíti čtverečky označenými čísly 1 až 9. Po nastoupení cestující zasunul lístek do strojku, který prodírkoval tři nebo čtyři z nich (specificky pro dané vozidlo a den). Kolik je různých způsobů prodírkování lístku? [210]
4. Kolika způsoby mohou sedět v kině sedm kamarádů (A, B, C, D, E, F, G) na sedadlech 1 až 7 tak, aby kamarád B seděl na sedadle č. 4 a kamarád G na sedadle č. 2? [120]
5. Do tanečního kroužku přišlo 24 chlapců a 15 dívek. Kolik různých párů lze vytvořit, pokud pár tvoří vždy dvojice chlapec–dívka? [360]
6. Ve třídě je 20 žáků. Kolika způsoby lze vybrat dvojici pro týdenní službu? [190]
7. Kolik hráčů se zúčastnilo turnaje ve stolním tenise, pokud se ve dvouhře odehrálo 21 utkání a každý hráč hrál s každým právě jednou? [7]
8. Ve třídě je 20 dívek a 15 chlapců. Kolik různých pětičlenných hlídek na branné závody lze vytvořit, pokud v každé hlídce mají být 3 dívky a 2 chlapci? [119 700]
9. Hokejové družstvo má 20 hráčů: 13 útočníků, 5 obránců a 2 brankáře. Kolik různých sestav může trenér vytvořit, pokud sestava má obsahovat 3 útočníky, 2 obránce a 1 brankáře? [5 720]
10. Učitel má k dispozici 20 aritmetických a 30 geometrických úloh. Na písemné práci mají být dvě aritmetické a tři geometrické úlohy. Kolik má učitel možností k vytvoření písemné práce? [771 400]
11. Ze 7 mužů a 4 žen máme vytvořit 6člennou skupinu, ve které mají být 3 ženy. Kolika způsoby lze takovou skupinu vytvořit? [140]
12. Učitel má vybrat na recitační soutěž tři studenty ze třídy 3.A a dva studenty ze třídy 3.B. V 3.A je 22 studentů a v 3.B je 17 studentů. Kolik má učitel možností výběru? [209 440]
13. Kolik existuje způsobů, jak uspořádat sedadla pro kamarády A, B, C, D a E tak, aby kamarád A seděl vedle kamaráda C? [48]

14. Latinská abeceda má 26 písmen. Kolik různých 6písmenných „slov“ lze vytvořit, pokud se písmena mohou opakovat? [308 915 776]
15. Státní poznávací značka tvoří 7 znaků. Na prvních třech pozicích může být číslice nebo písmeno, na zbývajících čtyřech jen číslice. Kolik různých značek lze vytvořit, pokud použijeme 28 písmen a 10 číslic? [548 720 000]
16. Na hodině tělesné výchovy stojí v řadě 5 dívek, z nichž dvě jsou sestry. Kolika způsoby lze rozestavit dívky tak, aby sestry stály vedle sebe? [48]

Literatura k tématu:

[1] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[2] CALDA, E., DUPAČ, V. (2008). Matematika pro gymnázia: Kombinatorika, pravděpodobnost, statistika (5. vydání, dotisk 2011). Praha: Prometheus. ISBN 978-80-7196-365-3.

Kapitola 2

Pravděpodobnost jevů

Po prostudování této kapitoly budete umět:

- objasnit pojmy náhodný pokus, náhodný jev, operace s jevy a jejich použití,
- představit klasickou a geometrickou pravděpodobnost,
- řešit typové úlohy z oblasti pravděpodobnosti včetně podmíněné pravděpodobnosti, nezávislosti a Bayesovy věty.

Klíčová slova:

Náhodný pokus, náhodný jev, klasická pravděpodobnost, geometrická pravděpodobnost, operace s jevy, podmíněná pravděpodobnost, nezávislé jevy, úplná pravděpodobnost, Bayesova věta.

Náhled kapitoly

V této kapitole se zaměříme na základní pojmy a pravidla teorie pravděpodobnosti, která tvoří výchozí rámec pro následné statistické metody. Nejprve zavedeme pojmy *náhodný pokus* a *náhodný jev* a ukážeme si, jak s jevy pracovat pomocí základních operací (sjednocení, průnik, doplněk). Poté představíme *klasickou* a *geometrickou* pravděpodobnost a procvičíme je na typových příkladech.

Dále se budeme věnovat *podmíněné pravděpodobnosti* a pojmu *nezávislosti jevů*, které umožňují analyzovat složitější situace. Kapitolu uzavřeme pravidlem *úplné pravděpodobnosti* a *Bayesovou větou*, jež jsou klíčové pro řadu aplikací (např. aktualizace pravděpodobností na základě nové informace).

Cíle kapitoly

Po prostudování této kapitoly byste měli být schopni:

- definovat náhodný pokus a náhodný jev a pracovat s operacemi s jevy,
- používat klasickou a geometrickou pravděpodobnost v typových úlohách,
- vypočítat podmíněnou pravděpodobnost a rozhodnout o nezávislosti jevů,
- aplikovat pravidlo úplné pravděpodobnosti a Bayesovu větu.

Časová náročnost

Doporučený čas na zvládnutí kapitoly je přibližně 4–5 hodin (výklad + průběžné řešení příkladů + samostatné procvičení).

2.1 Základní pojmy

Definice 2.1. Náhodný pokus je opakovatelný proces, jehož výsledek nelze předem jednoznačně určit, i když jsou podmínky pokusu stejné. Množinu všech možných výsledků náhodného pokusu nazýváme **prostor elementárních jevů** a označujeme ji Ω.

Například při hodu hrací kostkou je $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Definice 2.2. Náhodný jev je podmnožina prostoru elementárních jevů, tedy $A \subseteq \Omega$. Řekneme, že jev A **nastal**, právě když výsledek náhodného pokusu patří do A.

Například při hodu kostkou může být jev A „padne sudé číslo“, tedy $A = \{2, 4, 6\}$.

Druhy náhodných jevů

Definice 2.3. Nechť $A, B \subseteq \Omega$ jsou náhodné jevy.

- **Jev jistý** je jev, který nastane vždy. Platí $A = \Omega$ a jeho pravděpodobnost je
$$P(\Omega) = 1.$$
- **Jev nemožný** je jev, který nikdy nenastane. Platí $A = \emptyset$ a jeho pravděpodobnost je
$$P(\emptyset) = 0.$$
- **Jev elementární** je jev, který obsahuje právě jeden výsledek, tj. má tvar $\{\omega\}$ pro nějaké $\omega \in \Omega$.
- **Jev složený** je jev, který obsahuje alespoň dva výsledky.
- **Doplněk jevu** A (opačný jev) je jev
$$A^c = \Omega \setminus A,$$
tj. nastane právě tehdy, když jev A nenastane.
- **Neslučitelné (disjunktní) jevy** A a B jsou takové, že nemohou nastat současně, tedy
$$A \cap B = \emptyset.$$
- **Slučitelné jevy** A a B jsou takové, že mohou nastat současně, tedy
$$A \cap B \neq \emptyset.$$

2.2 Klasická pravděpodobnost

Definice 2.4. Nechť náhodný pokus má konečný prostor elementárních jevů Ω a nechť všechny elementární výsledky jsou stejně pravděpodobné (rovnoměrný model). Potom **klasická pravděpodobnost** jevu A je
$$P(A) = \frac{\text{počet příznivých výsledků}}{\text{celkový počet možných výsledků}}.$$

Pozn.: Pokud si prostor výsledků zapisujeme jako množinu, pak „počet prvků množiny“ se značí $|A|$ a $|\Omega|$ a lze psát také $P(A) = \frac{|A|}{|\Omega|}$.

Kdy lze použít klasickou pravděpodobnost?

- Ω je **konečná** a její prvky (elementární jevy) jsou jednoznačně určeny.
- Všechny elementární jevy jsou **stejně pravděpodobné** (např. férová kostka, férová mince).

Pozn.: Nezávislost opakovaných pokusů není předpokladem samotného vzorce $P(A) = |A|/|\Omega|$; je důležitá až při modelování *více* pokusů (např. dva hody kostkou).

Příklad 2.5. Hod hrací kostkou je klasickým příkladem náhodného pokusu. Popište prostor elementárních jevů a uveďte příklady jevů.

Řešení: Náhodný pokus: hod hrací kostkou.

Prostor elementárních jevů je

$$\Omega = \{1, 2, 3, 4, 5, 6\}.$$

Příklady náhodných jevů:

- $A = \{1, 3, 5\}$: „padne liché číslo“,
- $B = \{4, 5, 6\}$: „padne číslo ≥ 4“,
- $\emptyset$: „padne číslo > 6“ (jev nemožný),
- Ω: „padne číslo mezi 1 a 6“ (jev jistý),
- jevy „padne sudé číslo“ a „padne liché číslo“ jsou neslučitelné, protože jejich průnik je prázdný.

□

Příklad 2.6. Při hodu kostkou určete pravděpodobnost jevů:

- a) A: „padne číslo 5“,
- b) B: „padne číslo ≤ 2“.

Řešení: Protože všechny výsledky jsou stejně pravděpodobné a $|\Omega| = 6$, dostáváme:

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

□

Příklad 2.7. S jakou pravděpodobností padne při hodu dvěma hracími kostkami součet:

- a) 6,

- b) menší než 7?

Řešení: Uvažujme uspořádané dvojice (i, j), kde i je výsledek na první kostce a j na druhé. Platí $|\Omega| = 6 \cdot 6 = 36$.

ad a) Součet 6 nastane pro pět dvojic:

$$(1,5), (2,4), (3,3), (4,2), (5,1).$$

Proto

$$P(\text{součet } 6) = \frac{5}{36}.$$

ad b) Součet menší než 7 znamená součet $2, 3, 4, 5$ nebo 6. Počty možností jsou postupně $1, 2, 3, 4, 5$, celkem tedy $1 + 2 + 3 + 4 + 5 = 15$ příznivých dvojic. Proto

$$P(\text{součet } < 7) = \frac{15}{36} = \frac{5}{12}.$$

□

Příklad 2.8. V cele předběžného zadržení sedí vedle sebe 10 podezřelých, z toho 3 ženy. Jaká je pravděpodobnost, že všechny tři ženy sedí vedle sebe?

Řešení: Uvažujme všechna možná uspořádání 10 různých osob v řadě. Celkový počet uspořádání je

$$n = 10!.$$

Aby všechny tři ženy seděly vedle sebe, budeme je chápat jako jeden „blok“. Pak máme celkem 8 objektů (blok žen + 7 mužů), které lze uspořádat v řadě

$$8!$$

způsoby. Uvnitř bloku se ženy mohou prohodit

$$3!$$

způsoby. Počet příznivých uspořádání je tedy

$$m = 8! \cdot 3!.$$

Hledaná pravděpodobnost je

$$P = \frac{m}{n} = \frac{8! \cdot 3!}{10!} = \frac{6}{10 \cdot 9} = \frac{1}{15}.$$

□

Příklad 2.9. Stanovte pravděpodobnost jevu, že z 10 náhodně vytažených bridžových karet budou alespoň 3 esa. (V balíčku je 52 karet, z toho 4 esa.)

Řešení: Označme A jev „vytáhneme alespoň 3 esa“. To znamená „vytáhneme právě 3 esa“ nebo „vytáhneme právě 4 esa“. Tyto případy jsou neslučitelné, proto

$$P(A) = P(A_3) + P(A_4),$$

kde A_3 je jev „právě 3 esa“ a A_4 je jev „právě 4 esa“.

Celkový počet výběrů 10 karet z 52 je $\binom{52}{10}$. Dále:

- pro A_3 vybíráme 3 esa ze 4 a zbylých 7 karet z 48 ne-es,
- pro A_4 vybíráme všechna 4 esa a zbylých 6 karet z 48 ne-es.

Proto

$$P(A_3) = \frac{\binom{4}{3}\binom{48}{7}}{\binom{52}{10}}, \qquad P(A_4) = \frac{\binom{4}{4}\binom{48}{6}}{\binom{52}{10}},$$

a tedy

$$P(A) = \frac{\binom{4}{3}\binom{48}{7} + \binom{4}{4}\binom{48}{6}}{\binom{52}{10}}.$$

□

Příklad 2.10. Při slosování sportky je z osudí vylosováno 6 čísel ze 49. Poté je ze zbývajících 43 čísel vylosováno dodatkové číslo. Při správném tipování:

- a) šesti čísel získává sázející výhru 1. pořadí,
- b) pěti čísel a dodatkového čísla (5+1) získává sázející výhru 2. pořadí,
- c) pěti čísel získává sázející výhru 3. pořadí,
- d) čtyř čísel získává sázející výhru 4. pořadí,
- e) tří čísel získává sázející výhru 5. pořadí.

Vypočítejte pravděpodobnosti, se kterými při vsazeném jednom sloupci vyhrajete v 1. tahu výhry a)–e).

Řešení: V jednom sloupci tipujeme **6** čísel. Základní počet všech možných šestic je

$$\binom{49}{6} = 13\,983\,816.$$

ad a) (6 správných) Jediný příznivý případ je, že tipovaná šestice je přesně vylosovaná:

$$P(6) = \frac{1}{\binom{49}{6}}.$$

ad b) (5+1) Tipujeme 5 čísel z vylosované šestice a zároveň tipujeme dodatkové číslo. To lze provést

$$\binom{6}{5} \cdot \binom{1}{1} = 6$$

způsoby, proto

$$P(5+1) = \frac{\binom{6}{5}\binom{1}{1}}{\binom{49}{6}} = \frac{6}{\binom{49}{6}}.$$

ad c) (5 správných, bez dodatkového) Tipujeme 5 čísel z vylosované šestice a šesté tipované číslo musí být z ostatních 43 čísel, která nejsou vylosována v hlavní šestici ani jako dodatkové:

$$P(5) = \frac{\binom{6}{5}\binom{43}{1}}{\binom{49}{6}} = \frac{258}{\binom{49}{6}}.$$

ad d) (4 správná) Tipujeme 4 čísla z vylosované šestice a zbývající 2 tipovaná čísla volíme z oněch 43 nevylosovaných čísel:

$$P(4) = \frac{\binom{6}{4}\binom{43}{2}}{\binom{49}{6}}.$$

ad e) (3 správná) Tipujeme 3 čísla z vylosované šestice a zbývající 3 tipovaná čísla volíme z 43 nevylosovaných čísel:

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

□

2.3 Geometrická pravděpodobnost

Definice 2.11. Geometrická pravděpodobnost je model, ve kterém jsou všechny výsledky náhodného pokusu *rovnoměrně rozloženy* v nějaké geometrické oblasti (např. na úsečce, v rovině nebo v prostoru). Pravděpodobnost jevu A se pak určuje jako poměr *míry* příznivé části k míře celé oblasti:

$$P(A) = \frac{\text{délka / plocha / objem příznivé části}}{\text{délka / plocha / objem celé oblasti}}.$$

Používáme ji typicky tehdy, když výsledek pokusu závisí na spojité veličině (čas, poloha bodu, úhel apod.).

Příklad 2.12. Jaká je pravděpodobnost, že meteorit dopadne na pevninu, víme-li, že pevnina má rozlohu 149 milionů km^2 a moře 361 milionů km^2?

Řešení: Celková plocha (pevnina + moře) je

$$S = 149 + 361 = 510 \text{ milionů km}^2.$$

Pravděpodobnost dopadu na pevninu určíme jako poměr ploch:

$$P(\text{pevnina}) = \frac{149}{510} \approx 0{,}2922.$$

□

Příklad 2.13. Je dán kruh o poloměru 10 cm. Uvnitř je vyznačena kruhová oblast o poloměru 5 cm. Jaká je pravděpodobnost, že náhodně zvolený bod z většího kruhu padne do menšího kruhu?

Řešení: Plocha většího kruhu je

$$S_{\text{větší}} = \pi \cdot 10^2 = 100\pi \text{ cm}^2,$$

plocha menšího kruhu je

$$S_{\text{menší}} = \pi \cdot 5^2 = 25\pi \text{ cm}^2.$$

Hledaná pravděpodobnost je poměr ploch:

$$P = \frac{S_{\text{menší}}}{S_{\text{větší}}} = \frac{25\pi}{100\pi} = 0{,}25.$$

□

Příklad 2.14. Dva známí se domluví, že se sejdou na určitém místě mezi 15:00 a 16:00. Každý z nich po příchodu čeká nejvýše 20 minut. Jaká je pravděpodobnost, že se setkají?

Řešení: Označme x čas (v minutách po 15:00), kdy přijde první osoba, a y čas příchodu druhé osoby. Předpokládáme rovnoměrné a nezávislé příchody, tedy (x, y) je rovnoměrně rozložen v čtverci $[0, 60] \times [0, 60]$.

Setkají se právě tehdy, když

$$|x - y| \leq 20.$$

Celková plocha čtverce je

$$S_\Omega = 60 \cdot 60 = 3600.$$

Nevyhovující oblasti tvoří dva shodné pravoúhlé trojúhelníky v rozích čtverce (nad přímkou $y = x + 20$ a pod přímkou $y = x - 20$). Každý má odvěsny délky 40, tedy obsah

$$S_{\text{troj}} = \frac{1}{2} \cdot 40 \cdot 40 = 800.$$

Celková nevyhovující plocha je $2 \cdot 800 = 1600$, a proto příznivá plocha je

$$S_A = 3600 - 1600 = 2000.$$

Hledaná pravděpodobnost je

$$P(\text{setkají se}) = \frac{S_A}{S_\Omega} = \frac{2000}{3600} = \frac{5}{9} \approx 0{,}5556.$$

□

2.4 Statistická pravděpodobnost

Definice 2.15. Statistická pravděpodobnost (frekventistické pojetí) vychází z relativní četnosti výskytu jevu při opakování téhož náhodného pokusu. Označme n počet provedených pokusů a $N_n(A)$ počet pokusů, ve kterých nastal jev A. **Relativní četnost** jevu A je

$$f_n(A) = \frac{N_n(A)}{n}.$$

Je-li možné uvažovat dlouhou řadu pokusů za stejných podmínek, pak pravděpodobnost jevu A chápeme jako limitu relativní četnosti:

$$P(A) = \lim_{n\to\infty} f_n(A) = \lim_{n\to\infty} \frac{N_n(A)}{n}.$$

V praxi pracujeme s odhadem $P(A) \approx f_n(A)$ pro velké n.

Kdy má statistická pravděpodobnost smysl?

- pokus lze **opakovat** za (přibližně) stejných podmínek,
- jednotlivá opakování lze považovat za **nezávislá** a **stejně rozdělená** (i.i.d. model),
- pro dostatečně velké n se relativní četnosti **stabilizují** (zákon velkých čísel).

Statistická pravděpodobnost je vhodná tehdy, když máme k dispozici data z opakovaných pozorování a chceme na jejich základě odhadnout pravděpodobnosti jevů.

Poznámka k diskrétním a spojitým situacím

- **Diskrétní situace:** Jevy často odpovídají konkrétním hodnotám (např. „padne 6“). Pravděpodobnosti lze odhadovat relativními četnostmi jednotlivých hodnot.
- **Spojitá situace:** Pro spojitou náhodnou veličinu je pro každou konkrétní hodnotu typicky $P(X = x) = 0$. Odhady proto děláme pro **intervaly** (např. $P(170 \le X < 175)$) pomocí četností v intervalech; při jemnějším dělení intervalů pak přecházíme k pojmu **hustoty pravděpodobnosti**.

Příklad 2.16 (spojitý případ)**.** Sledujme dobu, po kterou se zákazníci zdržují v obchodě. Čas pobytu byl zaznamenán a rozdělen do intervalů o délce 5 minut. Data o četnostech pro jednotlivé intervaly shrnuje tabulka:

Určete statistické pravděpodobnosti pro jednotlivé intervaly.

Řešení: Celkem bylo sledováno $n = 200$ zákazníků. Statistické pravděpodobnosti odhadneme

Tab. 1: Četnosti doby pobytu zákazníků v obchodě (intervaly 5 minut)

Interval (min)	**Četnost**
$\langle 0; 5)$	77
$\langle 5; 10)$	83
$\langle 10; 15)$	25
$\langle 15; 20)$	15
Celkem	200

relativními četnostmi:

$$P(\langle 0; 5)) \approx \frac{77}{200} = 0{,}385, \qquad P(\langle 5; 10)) \approx \frac{83}{200} = 0{,}415,$$

$$P(\langle 10; 15)) \approx \frac{25}{200} = 0{,}125, \qquad P(\langle 15; 20)) \approx \frac{15}{200} = 0{,}075.$$

Odhady tvoří rozdělení pravděpodobnosti na zvolených intervalech (součet je 1). □

2.5 Podmíněná pravděpodobnost a nezávislé jevy

Podmíněná pravděpodobnost

Definice 2.17. Podmíněná pravděpodobnost je pravděpodobnost jevu A za předpokladu, že nastal jev B. Označuje se $P(A \mid B)$ a je definována jako:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{pokud } P(B) > 0.$$

Tento koncept je užitečný v mnoha praktických situacích, například při odhadu pravděpodobnosti úspěchu produktu na trhu, pokud víme, že byl úspěšný v podobném segmentu.

Nezávislé jevy

Definice 2.18. Nezávislé jevy jsou takové jevy, jejichž výskyt jeden druhého neovlivňuje. To znamená, že pravděpodobnost výskytu jednoho jevu neovlivňuje pravděpodobnost výskytu druhého jevu. Pokud jsou dva jevy A a B nezávislé, pak platí následující rovnost:

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

Tato rovnost říká, že pravděpodobnost současného výskytu jevů A a B (jejich průniku) je součinem pravděpodobností jednotlivých jevů. Nezávislost je důležitý koncept, který se často vyskytuje v reálných situacích, například při opakovaných náhodných pokusech, jako je házení kostkou nebo mincí. V těchto případech výsledek jednoho hodu neovlivňuje výsledek následujících hodů, a proto jsou tyto pokusy nezávislé.

Skupinově nezávislé jevy

Definice 2.19. Jevy A, B a C jsou **skupinově nezávislé**, jestliže platí následující podmínky:

- **Nezávislost po dvou:** Každá dvojice jevů musí být nezávislá, což znamená, že pro všechny dvojice jevů platí:

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

$$P(A \cap C) = P(A) \cdot P(C),$$

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

- **Nezávislost po třech:** Pro tři jevy zároveň musí platit, že průnik všech tří jevů odpovídá součinu jejich pravděpodobností:

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

Pokud jsou splněny všechny tyto podmínky, říkáme, že jevy A, B a C jsou skupinově nezávislé. Tato vlastnost je klíčová v situacích, kde analyzujeme souběh více nezávislých jevů, a je využívána v pravděpodobnostních modelech, jako je například rozklad nezávislých náhodných veličin.

Příklad 2.20 (mini-příklad)**.** Z balíčku 52 karet vytáhneme jednu kartu. Nechť

$$A = \{\text{karta je eso}\}, \qquad B = \{\text{karta je piková}\}.$$

Určete $P(A \mid B)$.

Řešení: Platí $P(A) = \frac{4}{52}$, $P(B) = \frac{13}{52}$ a $P(A \cap B) = \frac{1}{52}$ (pikové eso je právě jedno). Proto

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{52}}{\frac{13}{52}} = \frac{1}{13}.$$

□

Příklad 2.21. Házíme dvěma férovými mincemi. Určete pravděpodobnost jevu:

- A: padne líc a rub (v libovolném pořadí),
- B: na první minci padne líc.

Určete pravděpodobnost jevu A za předpokladu, že nastal jev B.

Řešení: Možné výsledky hodu dvěma mincemi (uspořádané dvojice) jsou:

1. mince	**2. mince**
LÍC	LÍC
LÍC	RUB
RUB	LÍC
RUB	RUB

Nejprve určíme pravděpodobnosti potřebné pro podmínění. Jev B nastane ve dvou ze čtyř stejně pravděpodobných výsledků, tedy

$$P(B) = \frac{2}{4} = \frac{1}{2}.$$

Jev $A \cap B$ znamená: na první minci je líc a zároveň padne líc i rub, takže na druhé minci musí být rub. To je právě jeden výsledek ze čtyř, tedy

$$P(A \cap B) = \frac{1}{4}.$$

Podle definice podmíněné pravděpodobnosti:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{4}}{\frac{1}{2}} = \frac{1}{2}.$$

□

Příklad 2.22. Studenti při zkoušení mohou dostat tři otázky. První student je připraven pouze na 1. otázku, druhý pouze na 2. otázku, třetí pouze na 3. otázku a čtvrtý je připraven na všechny tři otázky. Náhodně vybereme jednoho studenta. Uvažujme jevy:

- A_1: vybraný student dokáže zodpovědět 1. otázku,
- A_2: vybraný student dokáže zodpovědět 2. otázku,
- A_3: vybraný student dokáže zodpovědět 3. otázku.

Ukažte, že jevy A_1, A_2, A_3 jsou po dvou nezávislé, ale nejsou vzájemně nezávislé.

Řešení: Označme studenty (1), (2), (3), (4) podle zadání; každý je vybrán se stejnou pravděpodobností 1/4.

Jednotlivé jevy. Jev A_1 nastane, pokud byl vybrán student (1) nebo (4), tedy

$$P(A_1) = \frac{2}{4} = \frac{1}{2}.$$

Analogicky $P(A_2) = P(A_3) = \frac{1}{2}$.

Průniky dvojic. Jev $A_1 \cap A_2$ nastane právě tehdy, když byl vybrán student (4) (jen ten umí obě otázky), tedy

$$P(A_1 \cap A_2) = \frac{1}{4}.$$

Stejně platí

$$P(A_1 \cap A_3) = P(A_2 \cap A_3) = \frac{1}{4}.$$

Proto pro každou dvojici $i \neq j$ dostáváme

$$P(A_i \cap A_j) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A_i)P(A_j),$$

a jevy jsou **po dvou nezávislé**.

Průnik trojice. Jev $A_1 \cap A_2 \cap A_3$ opět nastane pouze tehdy, když byl vybrán student (4), tedy

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{4}.$$

Kdyby byly jevy vzájemně nezávislé, muselo by platit

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}.$$

Protože $\frac{1}{4} \neq \frac{1}{8}$, jevy A_1, A_2, A_3 **nejsou vzájemně nezávislé**. □

2.6 Úplná pravděpodobnost a Bayesova věta

Úplná pravděpodobnost

Definice 2.23. Nechť $B_1, \ldots, B_n$ tvoří **rozklad prostoru** Ω, tj.

$$B_i \cap B_j = \emptyset \ (i \neq j), \qquad \bigcup_{i=1}^{n} B_i = \Omega, \qquad \text{a} \qquad P(B_i) > 0 \text{ pro všechna } i.$$

Potom pro libovolný jev A platí **zákon úplné pravděpodobnosti**

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i)\, P(A \mid B_i).$$

Poznámka 2.24. Smysl vzorce: jev A může nastat v různých „scénářích“ $B_1, \ldots, B_n$. Celková pravděpodobnost $P(A)$ je vážený průměr podmíněných pravděpodobností $P(A \mid B_i)$ s vahami $P(B_i)$.

Příklad 2.25. V obchodě jsou tři pokladny. Na pokladně 1 dojde k chybě v účtování s pravděpodobností 0,1, na pokladně 2 s pravděpodobností 0,05 a na pokladně 3 s pravděpodobností 0,2. Pravděpodobnosti, že zákazník bude odbaven pokladnami 1, 2 a 3, jsou postupně 0,3, 0,25 a 0,45. Jaká je pravděpodobnost, že zákazník opouštějící obchod má chybný účet?

Řešení: Označme A jev „došlo k chybě v účtování“ a H_i jev „zákazník byl odbaven na i-té pokladně“, $i = 1, 2, 3$. Jevy H_1, H_2, H_3 tvoří rozklad prostoru (zákazník projde právě jednou pokladnou), proto použijeme zákon úplné pravděpodobnosti:

$$P(A) = \sum_{i=1}^{3} P(H_i)\, P(A \mid H_i).$$

Dosadíme:

$$P(A) = 0{,}3 \cdot 0{,}1 + 0{,}25 \cdot 0{,}05 + 0{,}45 \cdot 0{,}2.$$

$$P(A) = 0{,}03 + 0{,}0125 + 0{,}09 = 0{,}1325.$$

Pravděpodobnost chybného účtu je tedy 0,1325 (tj. přibližně 13,25 %). □

Bayesova věta

Definice 2.26. Nechť $B_1, \ldots, B_n$ tvoří rozklad prostoru Ω (tj. $B_i \cap B_j = \emptyset$ pro $i \neq j$, $\bigcup_{i=1}^{n} B_i = \Omega$ a $P(B_i) > 0$). Potom pro libovolný jev A s $P(A) > 0$ platí **Bayesova věta**:

$$P(B_i \mid A) = \frac{P(A \mid B_i)\, P(B_i)}{\sum_{j=1}^{n} P(A \mid B_j)\, P(B_j)}.$$

Jmenovatel je celková pravděpodobnost jevu A, tj. podle zákona úplné pravděpodobnosti

$$P(A) = \sum_{j=1}^{n} P(A \mid B_j)\, P(B_j).$$

Poznámka 2.27. Bayesova věta „obrací podmínku“: z pravděpodobnosti důsledku při dané příčině $P(A \mid B_i)$ a z apriorní pravděpodobnosti příčiny $P(B_i)$ určíme aposteriorní pravděpodobnost příčiny po pozorování důsledku $P(B_i \mid A)$.

Příklad 2.28 (Bayesova věta)**.** V obchodě jsou tři pokladny. Pravděpodobnost chyby v účtování je na pokladnách 1, 2, 3 postupně 0,1, 0,05 a 0,2. Pravděpodobnosti odbavení zákazníků pokladnami 1, 2, 3 jsou 0,3, 0,25 a 0,45. Pokud dojde k chybě v účtování, jaká je pravděpodobnost, že k ní došlo na třetí pokladně?

Řešení: Označme A jev „došlo k chybě“ a H_i jev „zákazník byl odbaven na i-té pokladně“, $i = 1, 2, 3$. Hledáme $P(H_3 \mid A)$.

Nejprve určíme $P(A)$ zákonem úplné pravděpodobnosti:

$$P(A) = 0{,}3 \cdot 0{,}1 + 0{,}25 \cdot 0{,}05 + 0{,}45 \cdot 0{,}2 = 0{,}1325.$$

Pak použijeme Bayesovu větu:

$$P(H_3 \mid A) = \frac{P(A \mid H_3)\, P(H_3)}{P(A)} = \frac{0{,}2 \cdot 0{,}45}{0{,}1325} = \frac{0{,}09}{0{,}1325} \approx 0{,}6792.$$

Pravděpodobnost, že chyba vznikla na třetí pokladně, je přibližně 67,92 %. □

Příklad 2.29 (Pozitivní lékařský test)**.** Prevalence výskytu AIDS v populaci je 0,6 %. Test má senzitivitu 99,9 % (tj. je pozitivní s pravděpodobností 0,999, je-li osoba nakažená) a specificitu 99 % (tj. je negativní s pravděpodobností 0,99, je-li osoba zdravá). Jaká je pravděpodobnost, že osoba s pozitivním testem má skutečně AIDS?

Řešení: Označme:

- A: osoba má AIDS, tedy $P(A) = 0{,}006$,
- $\overline{A}$: osoba nemá AIDS, tedy $P(\overline{A}) = 0{,}994$,
- T^+: test je pozitivní.

Ze zadání:

$$P(T^+ \mid A) = 0{,}999, \qquad P(T^+ \mid \overline{A}) = 1 - 0{,}99 = 0{,}01.$$

Použijeme Bayesovu větu:

$$P(A \mid T^+) = \frac{P(T^+ \mid A)\,P(A)}{P(T^+ \mid A)\,P(A) + P(T^+ \mid \overline{A})\,P(\overline{A})}.$$

Dosadíme:

$$P(A \mid T^+) = \frac{0{,}999 \cdot 0{,}006}{0{,}999 \cdot 0{,}006 + 0{,}01 \cdot 0{,}994} = \frac{0{,}005994}{0{,}015934} \approx 0{,}376.$$

Pravděpodobnost, že osoba s pozitivním testem má skutečně AIDS, je přibližně 37,6 %. □

Pozor (typická chyba / base-rate fallacy): Vysoká senzitivita a specificita ještě neznamenají, že $P(A \mid T^+)$ bude blízko 1. Výsledek výrazně závisí na prevalenci $P(A)$: je-li nemoc vzácná, mohou falešně pozitivní výsledky tvořit velkou část všech pozitivních testů.

Interpretace (test na vzácné onemocnění): Uvažujme 10 000 náhodně vybraných osob. Při prevalenci 0,6 % očekáváme asi $0{,}006 \cdot 10\,000 = 60$ nakažených a 9 940 zdravých. Z nakažených bude test pozitivní přibližně u $0{,}999 \cdot 60 \approx 60$ osob, zatímco ze zdravých bude falešně pozitivních asi $0{,}01 \cdot 9\,940 \approx 99$ osob. Celkem tedy bude pozitivních zhruba $60 + 99 = 159$ osob, z nichž nakažených je asi 60, takže

$$P(A \mid T^+) \approx \frac{60}{159} \approx 0{,}38,$$

což odpovídá vypočtené hodnotě 0,376.

2.7 Opakované pokusy

Definice 2.30. Opakované pokusy jsou situace, kdy tentýž náhodný pokus provádíme vícekrát za stejných podmínek. Zajímá nás zejména rozdělení počtu výskytů určitého jevu v n opakováních.

2.7.1 Nezávislé pokusy

Definice 2.31. Nezávislé opakované pokusy jsou takové, v nichž výsledek jednoho pokusu neovlivňuje výsledky dalších pokusů. V každém pokusu má sledovaný jev (např. „úspěch“) stejnou pravděpodobnost.

Poznámka 2.32. Typickým příkladem je opakovaný hod férovou mincí nebo kostkou. V praxi se s nezávislými pokusy setkáme např. při testování shodně vyrobených kusů (každý testovaný kus je jiný exemplář) nebo při opakovaném náhodném výběru.

Definice 2.33 (Bernoulliho schéma (binomické rozdělení))**.** Mějme n nezávislých pokusů, v nichž může nastat jev A („úspěch“) s pravděpodobností p; označme $q = 1 - p$. Nechť X je počet úspěchů v n pokusech. Potom X má **binomické rozdělení** a pro $k = 0, 1, \ldots, n$ platí

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

Nejpravděpodobnější počet úspěchů (modus). Nejpravděpodobnější hodnota k splňuje

$$(n+1)p - 1 \leq k \leq (n+1)p.$$

Je-li $(n+1)p$ celé číslo, existují dvě nejpravděpodobnější hodnoty: $k = (n+1)p - 1$ a $k = (n+1)p$; jinak je modus jednoznačný a platí $k = \lfloor (n+1)p \rfloor$.

Příklad 2.34. Házíme šestkrát férovou hrací kostkou. Vypočtěte pravděpodobnost, že šestka padne právě dvakrát.

Řešení: Jde o Bernoulliho schéma s parametry $n = 6$, $p = \frac{1}{6}$ („úspěch“ = „padne šestka“) a $k = 2$.

$$P(X = 2) = \binom{6}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^4.$$

Numericky:

$$P(X = 2) = 15 \cdot \frac{1}{36} \cdot \frac{625}{1296} = \frac{9375}{46656} \approx 0{,}2009.$$

Pravděpodobnost, že šestka padne právě dvakrát, je tedy přibližně 0,2009. □

Příklad 2.35. Sportovní střelec zasáhne cíl při každém výstřelu s pravděpodobností $p = 0{,}8$. Vypočtěte pravděpodobnost, že při 5 výstřelech budou v cíli:

1. právě 2 zásahy,
2. nejvýše jeden zásah,
3. alespoň 2 zásahy.

Řešení: Počet zásahů označme X. Při nezávislých výstřelech platí $X \sim \mathrm{Bi}(n = 5, p = 0{,}8)$, tedy

$$P(X = k) = \binom{5}{k}(0{,}8)^k(0{,}2)^{5-k}.$$

1. Pravděpodobnost právě 2 zásahů:

$$P(X = 2) = \binom{5}{2}(0{,}8)^2(0{,}2)^3 = 10 \cdot 0{,}64 \cdot 0{,}008 = 0{,}0512.$$

2. Pravděpodobnost nejvýše jednoho zásahu:

$$\begin{aligned} P(X \leq 1) &= P(X = 0) + P(X = 1), \\ P(X = 0) &= \binom{5}{0}(0{,}8)^0(0{,}2)^5 = (0{,}2)^5 = 0{,}00032, \\ P(X = 1) &= \binom{5}{1}(0{,}8)^1(0{,}2)^4 = 5 \cdot 0{,}8 \cdot 0{,}0016 = 0{,}0064, \\ P(X \leq 1) &= 0{,}00032 + 0{,}0064 = 0{,}00672. \end{aligned}$$

3. Pravděpodobnost alespoň dvou zásahů:

$$P(X \geq 2) = 1 - P(X \leq 1) = 1 - 0{,}00672 = 0{,}99328.$$

□

Příklad 2.36. Pravděpodobnost, že náhodně vybraný student bude znát učivo, je $p = 0{,}05$. Jaká je pravděpodobnost, že mezi dvaceti vybranými studenty bude:

- a) právě 5 znalých studentů,
- b) nejvýše 2 znalí studenti,
- c) alespoň jeden znalý student?

Řešení: Označme X počet znalých studentů mezi $n = 20$ náhodně vybranými. Předpokládáme nezávislost a stejnou pravděpodobnost znalosti, tedy $X \sim \mathrm{Bi}(20, 0{,}05)$ a

$$P(X = k) = \binom{20}{k}(0{,}05)^k(0{,}95)^{20-k}.$$

a) Pravděpodobnost, že budou právě 5 znalí:

$$P(X=5) = \binom{20}{5}(0{,}05)^5(0{,}95)^{15}$$
$$= 15504 \cdot 0{,}0000003125 \cdot 0{,}463291 \approx 0{,}002245.$$

b) Pravděpodobnost, že budou nejvýše 2 znalí:

$$P(X \le 2) = P(X=0) + P(X=1) + P(X=2),$$
$$P(X=0) = (0{,}95)^{20} \approx 0{,}358486,$$
$$P(X=1) = \binom{20}{1}(0{,}05)(0{,}95)^{19} = 1 \cdot (0{,}95)^{19} \approx 0{,}377354,$$
$$P(X=2) = \binom{20}{2}(0{,}05)^2(0{,}95)^{18} = 190 \cdot 0{,}0025 \cdot (0{,}95)^{18} \approx 0{,}188677,$$
$$P(X \le 2) \approx 0{,}358486 + 0{,}377354 + 0{,}188677 = 0{,}924516.$$

c) Pravděpodobnost, že bude alespoň jeden znalý:

$$P(X \ge 1) = 1 - P(X=0) = 1 - (0{,}95)^{20} \approx 1 - 0{,}358486 = 0{,}641514.$$

□

2.7.2 Závislé pokusy

Definice 2.37. Závislé opakované pokusy jsou takové, v nichž výsledek jednoho pokusu mění pravděpodobnosti v pokusech následujících. Typicky se to děje tehdy, když po provedení pokusu dojde ke změně podmínek (např. změna složení urny po výběru bez vracení).

Poznámka 2.38. Nejčastějším modelem závislých opakovaných pokusů v základním kurzu je **výběr bez vracení**. Počet „úspěchů“ ve výběru pak má **hypergeometrické rozdělení**.

Definice 2.39 (Výběr bez vracení (hypergeometrické rozdělení))**.** Mějme soubor N prvků, z nichž M má sledovanou vlastnost („úspěch“) a $N-M$ ji nemá („neúspěch“). Náhodně vybereme bez vracení n prvků. Označme X počet vybraných prvků se sledovanou vlastností. Potom pro $k = 0, 1, \ldots, n$ (přesněji pro ta k, pro něž má výraz smysl) platí

$$P(X=k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}.$$

Příklad 2.40. V osudí jsou 2 bílé a 3 černé koule. Určete pravděpodobnost toho, že:

- a) vytáhneme naráz 3 koule a budou 2 černé a 1 bílá,
- b) vytáhneme po jedné bez vracení 2 černé a 1 bílou (v libovolném pořadí).

Řešení: V obou případech jde o tentýž výběr bez vracení, jen jinak popsaný.

ad a) Naráz vybíráme $n = 3$ koule z $N = 5$, přičemž „úspěch“ definujeme jako „černá koule“. Tedy $M = 3$ a chceme $k = 2$:

$$P(X = 2) = \frac{\binom{3}{2}\binom{2}{1}}{\binom{5}{3}} = \frac{3 \cdot 2}{10} = 0{,}6.$$

ad b) Při postupném výběru bez vracení a požadavku „2 černé a 1 bílá v libovolném pořadí“ dostaneme stejnou pravděpodobnost jako v bodě a). Např. pro konkrétní pořadí ČBČ platí

$$P(\text{ČBČ}) = \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{2}{3} = \frac{1}{5}.$$

Stejnou pravděpodobnost mají i pořadí ČČB a BČČ, takže

$$P(\text{2 černé a 1 bílá}) = 3 \cdot \frac{1}{5} = 0{,}6.$$

□

Příklad 2.41. Mezi 15 výrobky je 5 zmetků. Vybereme 3 výrobky. Jaká je pravděpodobnost, že právě jeden z nich je vadný, jestliže:

- a) vybereme všechny 3 najednou,
- b) vybíráme po jednom bez vracení?

Řešení: Opět jde v obou případech o tentýž výběr bez vracení. Označme X počet vadných kusů ve výběru. Máme $N = 15$, $M = 5$ (vadné), $n = 3$ a chceme $k = 1$:

$$P(X = 1) = \frac{\binom{5}{1}\binom{10}{2}}{\binom{15}{3}} = \frac{5 \cdot 45}{455} = \frac{225}{455} = \frac{45}{91} \approx 0{,}4945.$$

ad a) Výsledek je přímo uveden výše.

ad b) Při postupném výběru bez vracení lze stejně dojít součtem přes pořadí (V = vadný, D = dobrý):

$$P(\text{VDD}) = \frac{5}{15} \cdot \frac{10}{14} \cdot \frac{9}{13}, \quad P(\text{DVD}) = \frac{10}{15} \cdot \frac{5}{14} \cdot \frac{9}{13}, \quad P(\text{DDV}) = \frac{10}{15} \cdot \frac{9}{14} \cdot \frac{5}{13},$$

a tedy $P(X = 1) = P(\text{VDD}) + P(\text{DVD}) + P(\text{DDV})$, což dá stejný výsledek $\frac{45}{91}$. □

2.8 Souhrnné příklady

Příklad 2.42. Mějme pět vstupenek po 100 Kč, tři vstupenky po 300 Kč a dvě vstupenky po 500 Kč. Náhodně vybereme tři vstupenky (bez vracení). Určete pravděpodobnost toho, že:

- a) alespoň dvě z těchto vstupenek mají stejnou hodnotu,
- b) všechny tři vstupenky stojí dohromady 700 Kč.

Řešení: Celkem je $N = 10$ vstupenek a vybíráme $n = 3$, takže počet všech stejně pravděpodobných výběrů je $\binom{10}{3}$.

ad a) Řešíme přes opačný jev. Opačný jev k „alespoň dvě mají stejnou hodnotu“ je „všechny tři mají různé hodnoty“, tj. jedna za 100 Kč, jedna za 300 Kč a jedna za 500 Kč. Počet takových výběrů je $\binom{5}{1}\binom{3}{1}\binom{2}{1}$, tedy

$$P(\text{všechny různé}) = \frac{\binom{5}{1}\binom{3}{1}\binom{2}{1}}{\binom{10}{3}}.$$

Proto

$$P(\text{alespoň dvě stejné}) = 1 - P(\text{všechny různé}) = 1 - \frac{\binom{5}{1}\binom{3}{1}\binom{2}{1}}{\binom{10}{3}} = 1 - \frac{30}{120} = \frac{3}{4} = 0{,}75.$$

ad b) Součet 700 Kč může nastat jen ve dvou typech výběrů:

- (100, 300, 300): $\binom{5}{1}\binom{3}{2}$,
- (100, 100, 500): $\binom{5}{2}\binom{2}{1}$.

Tedy

$$P(\text{celkem 700 Kč}) = \frac{\binom{5}{1}\binom{3}{2} + \binom{5}{2}\binom{2}{1}}{\binom{10}{3}} = \frac{15 + 20}{120} = \frac{7}{24} \approx 0{,}2917.$$

□

Příklad 2.43. Z celkové produkce závodu jsou 4 % zmetků a z dobrých výrobků je 75 % standardních. Určete pravděpodobnost, že náhodně vybraný výrobek je standardní.

Řešení: Označme:

$$A = \{\text{výrobek je dobrý (není zmetek)}\}, \qquad B = \{\text{výrobek je standardní}\}.$$

Zadání říká, že $P(A) = 0{,}96$ a $P(B \mid A) = 0{,}75$. Standardní výrobek musí být dobrý, tedy $B \subseteq A$ a platí

$$P(B) = P(A \cap B) = P(A)\,P(B \mid A) = 0{,}96 \cdot 0{,}75 = 0{,}72.$$

□

Příklad 2.44. Z výrobků určitého druhu dosahuje 95 % předepsanou kvalitu. V určitém závodě, který vyrábí 80 % celkové produkce, má předepsanou kvalitu 98 % výrobků. Mějme náhodně vybraný výrobek předepsané kvality. Jaká je pravděpodobnost, že byl vyroben ve výše uvedeném závodě?

Řešení: Označme:

$$A = \{\text{výrobek je ze zmíněného závodu}\}, \qquad B = \{\text{výrobek je předepsané kvality}\}.$$

Hledáme $P(A \mid B)$. Známe

$$P(A) = 0{,}8, \quad P(\overline{A}) = 0{,}2, \quad P(B \mid A) = 0{,}98.$$

Dále je dáno, že celkově platí $P(B) = 0{,}95$. (To je klíčový údaj; bez něj nelze $P(A \mid B)$ určit.) Použijeme Bayesovu větu:

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)} = \frac{0{,}98 \cdot 0{,}8}{0{,}95} = \frac{0{,}784}{0{,}95} \approx 0{,}8253.$$

□

Σ

V této kapitole jsme zavedli základní pojmy teorie pravděpodobnosti a ukázali jsme jejich použití na typových úlohách. Pracovali jsme s modely, ve kterých pravděpodobnost vyjadřuje míru nejistoty výsledku náhodného pokusu, a naučili jsme se rozlišovat situace s konečným i spojitým prostorem výsledků.

- **Náhodný pokus** – opakovatelný proces, jehož výsledek nelze předem jistě určit (např. hod kostkou, losování). Množinu všech možných výsledků nazýváme **prostor elementárních jevů** Ω.
- **Náhodný jev** – podmnožina Ω; jev A nastane právě tehdy, když výsledek pokusu patří do A. Rozlišili jsme jev jistý, nemožný, elementární a složený a uvedli základní vztahy mezi jevy (doplněk, průnik, sjednocení, neslučitelnost).
- **Klasická pravděpodobnost** – v konečném prostoru Ω se stejně pravděpodobnými elementárními jevy platí

  $$P(A) = \frac{\text{počet příznivých výsledků}}{\text{počet všech výsledků}}.$$

  Typicky např. při hodu férovou kostkou je $P(\{\text{padne } 6\}) = \frac{1}{6}$.
- **Geometrická pravděpodobnost** – v „kontinuálním“ modelu určíme pravděpodobnost jako poměr délek/obsahů/objemů, např.

  $$P(A) = \frac{\text{velikost příznivé oblasti}}{\text{velikost celé oblasti}}.$$

- **Statistická (frekvenční) pravděpodobnost** – pravděpodobnost jevu interpretujeme jako limitu relativní četnosti v dlouhé řadě opakování pokusu; v praxi ji odhadujeme z dat.

- **Podmíněná pravděpodobnost** – pravděpodobnost jevu A za podmínky, že nastal jev B,
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0.$$
- **Nezávislost jevů** – jevy A a B jsou nezávislé, jestliže
$$P(A \cap B) = P(A)\, P(B),$$
a upozornili jsme na rozdíl mezi nezávislostí po dvou a vzájemnou (skupinovou) nezávislostí.
- **Zákon úplné pravděpodobnosti a Bayesova věta** – použili jsme rozklad prostoru na disjunktní případy a vypočítali pravděpodobnosti „zpětně“ (pravděpodobnost příčiny při známém důsledku).
- **Opakované pokusy** – pro nezávislé opakované dichotomické pokusy jsme uvedli **Bernoulliho schéma** (binomický vzorec) a pro výběr bez vracení **hypergeometrické rozdělení**.

Získané pojmy a vzorce tvoří základ pro následující kapitoly: umožňují jednak správně modelovat náhodné situace, jednak přesně interpretovat výsledné pravděpodobnosti v kontextu daného problému.

1. Máme 230 výrobků, mezi nimiž je 20 nekvalitních. Vybereme 15 výrobků bez vracení. Jaká je pravděpodobnost, že mezi 15 vybranými bude právě 10 dobrých (a tedy 5 nekvalitních)? [0,00448]
2. Pacienta lze kontrolovat v čase od 7 do 20 hodin. Vycházky má od 13 do 15 hodin. Jaká je pravděpodobnost, že při náhodně zvolené kontrole v intervalu $\langle 7; 20\rangle$ bude pacient doma k zastižení? [11/13]
3. Dva sportovní střelci střílejí nezávisle na sebe do jednoho terče (každý jednou). Pravděpodobnost zásahu prvního střelce je 0,8, druhého 0,4. Při střelbě byl v terči právě jeden zásah. Jaká je pravděpodobnost, že terč zasáhl první střelec? [0,857]
4. Pravděpodobnost výhry hráče v jedné partii je 0,6. Určete nejpravděpodobnější počet výher hráče v deseti odehraných partiích. [6]
5. Série 100 výrobků je kontrolována náhodným výběrem 5 kusů bez vracení. Série je považována za „špatnou“, je-li alespoň jeden z pěti vybraných výrobků vadný. Vypočtěte pravděpodobnost, že série bude vyhodnocena jako špatná, víme-li, že obsahuje 5 % vadných výrobků. [0,230]
6. V telefonním seznamu náhodně vybereme jedno šestimístné číslo (může začínat nulou) a předpokládáme, že v seznamu jsou použita všechna šestimístná čísla. Jaká je pravděpodobnost, že číslo:
   a. neobsahuje číslici 0? [0,53144]
   b. obsahuje alespoň jednu číslici 3? [0,46856]
   c. obsahuje právě jednu číslici 3? [0,35429]

Literatura k tématu:

[1] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: https://homel.vsb.cz/~oti73/cdpast1/

[2] CALDA, E., DUPAČ, V. (2008). Matematika pro gymnázia: Kombinatorika, pravděpodobnost, statistika (5. vydání, dotisk 2011). Praha: Prometheus. ISBN 978-80-7196-365-3.

[3] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 3

Náhodná veličina

Po prostudování této kapitoly budete umět:

- rozlišovat mezi diskrétními a spojitými náhodnými veličinami,
- chápat rozdíl mezi pravděpodobnostní funkcí (pro diskrétní veličiny) a hustotou pravděpodobnosti (pro spojité veličiny),
- vypočítat střední hodnotu, rozptyl a směrodatnou odchylku pro různá rozdělení náhodných veličin,
- chápat význam distribuční funkce a umět ji interpretovat pro různé typy náhodných veličin,
- sestavit pravděpodobnostní funkci, hustotu pravděpodobnosti a distribuční funkci a graficky je znázornit.

Klíčová slova:

Diskrétní rozdělení, spojité rozdělení, pravděpodobnostní funkce, distribuční funkce, hustota pravděpodobnosti, střední hodnota, rozptyl, šikmost a špičatost.

Náhled kapitoly

V této kapitole navážeme na pojem náhodného jevu z předchozí části a zavedeme si klíčový pojem teorie pravděpodobnosti – **náhodnou veličinu**. Ta nám umožňuje převést abstraktní výsledky náhodných pokusů do světa čísel. Dále se podíváme, jak je možné pomocí rozdělení pravděpodobnosti určit, s jakou šancí budou tyto číselné hodnoty nastávat.

Kapitola se zaměřuje na rozlišení diskrétních a spojitých náhodných veličin a na způsoby výpočtu jejich základních charakteristik: střední hodnoty, rozptylu a směrodatné odchylky.

Cíle kapitoly

Cílem této kapitoly je prohloubení základů teorie pravděpodobnosti a upevnění poznatků o náhodných veličinách a jejich rozděleních, které budou nezbytným předpokladem pro metody induktivní statistiky v následujících kapitolách.

Časová náročnost

Doporučený čas na zvládnutí kapitoly je přibližně 4 až 5 hodin. Tento čas zahrnuje čtení textu, pochopení základních pojmů a principů, řešení ukázkových příkladů a samostatné procvičení výpočtů základních charakteristik.

Náhodný jev a náhodná veličina

Definice 3.1. Náhodný jev je událost, která může, ale nemusí nastat v rámci nějakého pokusu nebo procesu. Můžeme si ho představit jako výsledek experimentu, který závisí na náhodě. Pravděpodobnost je míra, která kvantifikuje možnost, že k danému náhodnému jevu dojde, a pohybuje se v rozmezí od 0 (jev nemožný) do 1 (jev jistý). Například pravděpodobnost, že při hodu kostkou padne číslo 6, je $\frac{1}{6}$, protože existuje 6 možných výsledků a každý má stejnou šanci nastat.

Definice 3.2. Náhodná veličina je proměnná, která může nabývat různých (reálných) hodnot v závislosti na výsledku náhodného pokusu. Například při hodu kostkou může náhodná veličina X představující výsledek hodu nabývat hodnot $1, 2, 3, 4, 5$ nebo 6. Každý z těchto výsledků je výsledek náhodného procesu.

Náhodné veličiny slouží především k tomu, abychom abstraktním výsledkům náhodných pokusů (např. „padne líc“) přiřadili konkrétní číselné hodnoty, se kterými lze dále matematicky a statisticky pracovat.

Příklady náhodných veličin mohou být:

- Počet líců při deseti hodech mincí.
- Počet zákazníků, kteří navštíví obchod v určitém dni.
- Výška náhodně vybraného člověka z populace.
- Doba, za kterou přijede autobus na zastávku.
- Výsledek hodu dvěma kostkami (součet bodů).
- Počet vadných kusů ve výrobní sérii 100 produktů.

Tyto příklady ukazují různé typy náhodných veličin – některé jsou *diskrétní* (počet líců, počet zákazníků), jiné *spojité* (výška člověka, čas čekání).

Rozdělení pravděpodobnosti

Rozdělení pravděpodobnosti popisuje, jak jsou pravděpodobnosti jednotlivých možných výsledků náhodné veličiny rozloženy. Například u hodu (férovou) kostkou mají všechny výsledky (hodnoty 1 až 6) stejnou pravděpodobnost, tedy $\frac{1}{6}$. V praxi však ne vždy všechny výsledky mají stejnou pravděpodobnost. Rozdělení pravděpodobnosti tedy udává, s jakou pravděpodobností různé hodnoty náhodné veličiny nastanou.

Rozdělení pravděpodobnosti nám tedy poskytuje obraz o tom, jak často můžeme očekávat jednotlivé výsledky náhodného pokusu.

V závislosti na typu náhodné veličiny rozlišujeme dvě hlavní kategorie: **diskrétní** a **spojité** náhodné veličiny.

3.1 Rozdělení pravděpodobnosti diskrétní náhodné veličiny

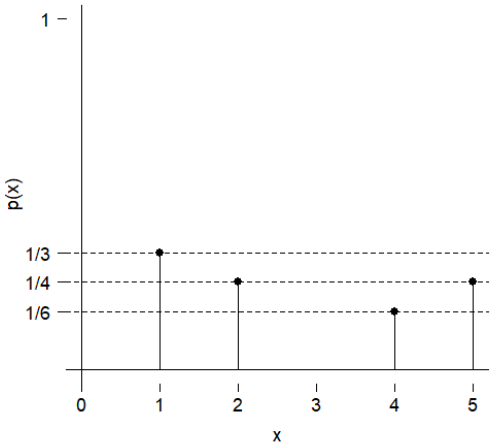
Diskrétní náhodná veličina nabývá pouze konečného nebo spočetně nekonečného množství možných hodnot. Příkladem diskrétní náhodné veličiny je počet vadných výrobků v sérii nebo počet zákazníků přicházejících do obchodu za jeden den. Diskrétní náhodná veličina je jednoznačně určena posloupností reálných čísel $\{x_n\}$ a posloupností pravděpodobností $\{p_n = P(X = x_n)\}$.

Příklad 3.3. Diskrétní náhodná veličina X nabývá hodnot $M = \{1, 2, 4, 5\}$ s pravděpodobnostmi $p(k) = P(X = k)$, kde

$$p(1) = \frac{1}{3},\ p(2) = \frac{1}{4},\ p(4) = \frac{1}{6},\ p(5) = \frac{1}{4} \text{ a } p(x) = 0 \text{ jinak.}$$

Zapisujeme také pomocí tabulky či obrázku:

k	1	2	4	5
$P(X=k)$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{4}$



Definice 3.4. Diskrétní náhodné veličiny mají svou **pravděpodobnostní funkci**, která přiřazuje každé hodnotě náhodné veličiny určitou pravděpodobnost $P(X = x_i) = p_i$, pro $x_i \in M$, kde x_i je možná hodnota diskrétní náhodné veličiny X, a p_i je pravděpodobnost, že X nabude hodnoty x_i.

Vlastnosti pravděpodobnostní funkce:

- $p(x) \geq 0 \quad \forall x \in \mathbb{R}$,
- $\sum_{x \in M} p(x) = 1$,
- Výpočet pravděpodobnosti (jevu B):

$$P(X \in B) = \sum_{x_i \in B} P(X = x_i) = \sum_{x_i \in B} p(x_i)$$

(součet pravděpodobností všech výsledků, které patří do množiny B; protože nenulové pravděpodobnosti jsou pouze pro hodnoty z množiny M, sčítáme reálně jen na průniku $B \cap M$.)

Definice 3.5 (Distribuční funkce)**. Distribuční funkce** náhodné veličiny X je reálná funkce $F : \mathbb{R} \to \langle 0; 1 \rangle$ definovaná vztahem

$$F(x) = P(X \leq x), \quad x \in \mathbb{R}.$$

Příklad 3.6 (distribuční funkce diskrétní náhodné veličiny)**.** Diskrétní náhodná veličina X nabývá hodnot $M = \{1, 2, 4, 5\}$ s pravděpodobnostmi $p(k) = P(X = k)$, kde $p(1) = \frac{1}{3}$, $p(2) = \frac{1}{4}$, $p(4) = \frac{1}{6}$, $p(5) = \frac{1}{4}$ a $p(x) = 0$ jinak.

Určete příslušnou distribuční funkci.

Řešení: Vycházíme z toho, že distribuční funkce je „zajímavá“ jen v bodech, kde je pravděpodobnostní funkce kladná. V těchto bodech dochází u distribuční funkce ke skokovému růstu právě o hodnotu pravděpodobnostní funkce v tomto bodě. Mezi těmito body je konstantní. Praktické je tedy vypočítat hodnoty F v těchto bodech a připsat je do již známé tabulky pro pravděpodobnostní funkci:

k	1	2	4	5
$P(X=k)$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{4}$
$F(k)=\sum_{k_i\le k} P(X=k_i)$	$\frac{1}{3}$	$\frac{1}{3}+\frac{1}{4}=\frac{7}{12}$	$\frac{7}{12}+\frac{1}{6}=\frac{3}{4}$	$\frac{3}{4}+\frac{1}{4}=1$

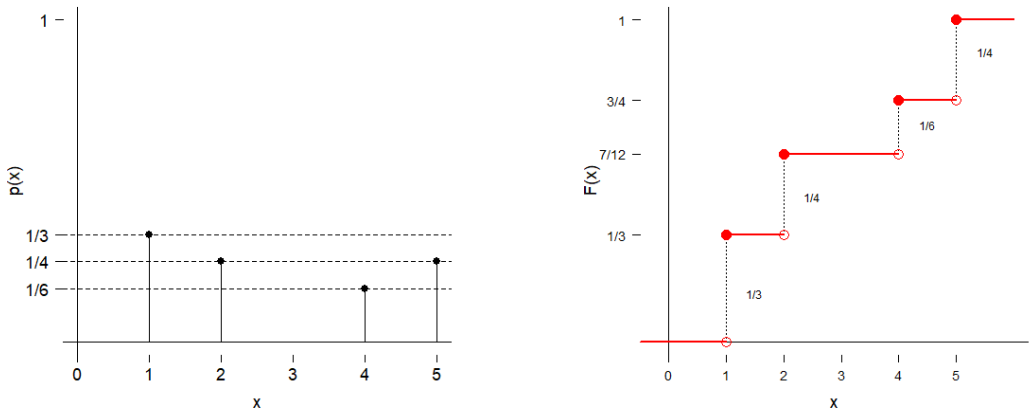
Dále F můžeme zapsat na jednotlivých intervalech, které nám pokryjí celé $\mathbb{R}$:

x	$(-\infty;1)$	$\langle 1;2)$	$\langle 2;4)$	$\langle 4;5)$	$\langle 5;\infty)$
$F(x)$	0	$\frac{1}{3}$	$\frac{7}{12}$	$\frac{3}{4}$	1

A nakonec i takto:

$$F(x)=\begin{cases}0 & x<1,\\ \frac{1}{3} & 1\le x<2,\\ \frac{7}{12} & 2\le x<4,\\ \frac{3}{4} & 4\le x<5,\\ 1 & x\ge 5.\end{cases}$$

Nejnázornější stejně budou grafy na obrázku 1.



Obr. 1: Pravděpodobnostní a distribuční funkce k příkladu 3.6

□

Z příkladu 3.6 sice můžeme odpozorovat některé vlastnosti distribuční funkce, ale raději si je zde vypíšeme:

Vlastnosti distribuční funkce:

- $F(x) \in \langle 0; 1 \rangle$,
- F je neklesající,
- F je zprava spojitá,
- F je definovaná na $\mathbb{R}$,
- $\lim\limits_{x \to -\infty} F(x) = 0, \quad \lim\limits_{x \to \infty} F(x) = 1,$
- $P(X = x_0) = F(x_0) - \lim\limits_{x \to x_0^-} F(x)$ (výška skoku v bodě x_0).

Příklad 3.7. V osudí je 5 bílých a 7 červených míčků. Náhodná veličina X představuje počet bílých míčků mezi pěti vybranými. Vytvořte pravděpodobnostní a distribuční funkci této náhodné veličiny.

Řešení: Náhodná veličina X nabývá hodnot $\{0, 1, 2, 3, 4, 5\}$. Z teorie pravděpodobnosti víme, že se jedná o výběr bez vracení (závislé pokusy). Můžeme tedy sestavit pravděpodobnostní funkci pro jednotlivé hodnoty X pomocí vzorce pro hypergeometrické rozdělení:

$$P(X = x) = \frac{\binom{5}{x}\binom{7}{5-x}}{\binom{12}{5}}.$$

Na základě této funkce vytvoříme tabulku pravděpodobností:

x_i	0	1	2	3	4	5
p_i	$\frac{21}{792}$	$\frac{175}{792}$	$\frac{350}{792}$	$\frac{210}{792}$	$\frac{35}{792}$	$\frac{1}{792}$

Pravděpodobnostní funkci lze graficky znázornit pomocí bodového nebo úsečkového (hůlkového) diagramu.

Distribuční funkce $F(x)$ bude mít skoky v bodech $0, 1, 2, 3, 4, 5$. Hodnoty funkce $F(x)$ v těchto bodech jsou určeny jako součet všech předcházejících pravděpodobností p_i:

$$F(x_i) = P(X \leq x_i).$$

Tabulka pro hodnoty distribuční funkce ve skocích:

x_i	0	1	2	3	4	5
$F(x_i)$	$\frac{21}{792}$	$\frac{196}{792}$	$\frac{546}{792}$	$\frac{756}{792}$	$\frac{791}{792}$	1

Graf distribuční funkce u diskrétní náhodné veličiny tvoří **schodovitý diagram** (funkce je po částech konstantní a v bodech x_i má skoky). □

3.2 Rozdělení pravděpodobnosti spojité náhodné veličiny

Spojitá náhodná veličina nabývá hodnot z nějakého intervalu reálných čísel. Příkladem může být výška náhodně vybraného člověka nebo doba, kterou zákazník stráví v obchodě. Spojité náhodné veličiny nemají konkrétní pravděpodobnosti pro jednotlivé hodnoty (pravděpodobnostní funkci), ale místo toho pracují s tzv. **hustotou pravděpodobnosti**, která určuje pravděpodobnost, že náhodná veličina nabyde hodnoty z určitého intervalu.

Definice 3.8. Náhodná veličina X s distribuční funkcí F se nazývá **spojitá**, jestliže existuje nezáporná funkce $f\colon \mathbb{R} \to \mathbb{R}$ taková, že

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t, \quad \forall x \in \mathbb{R}.$$

Funkce $f(x)$ se nazývá **hustota** (rozdělení pravděpodobností) náhodné veličiny X.

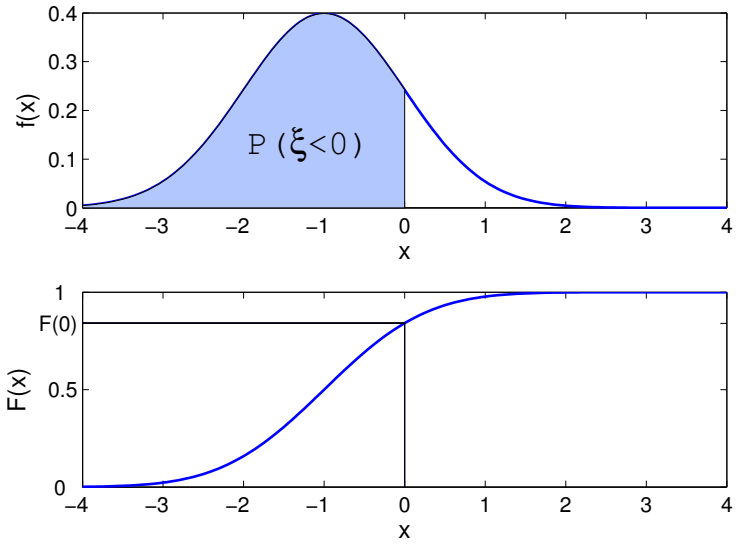
Vlastnosti hustoty:

- $f(x) \geq 0$,
- $\int_{-\infty}^{\infty} f(t)\,\mathrm{d}t = 1 \quad \Rightarrow$ plocha pod křivkou hustoty vyjadřuje celkovou pravděpodobnost,
- $f(x) = F'(x)$ v každém bodě x, kde F' existuje,
- $P(a \leq X \leq b) = F(b) - F(a) = \int_a^b f(t)\,\mathrm{d}t$,
- $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$,
- $P(X \in B) = \int_B f(t)\,\mathrm{d}t$.

Výpočet pravděpodobností pomocí $F(x)$ a $f(x)$ na nekonečném intervalu:

$$P(X < 0) = F(0) = \int_{-\infty}^{0} f(t)\,\mathrm{d}t.$$

Toto je znázorněno na obrázku 2.

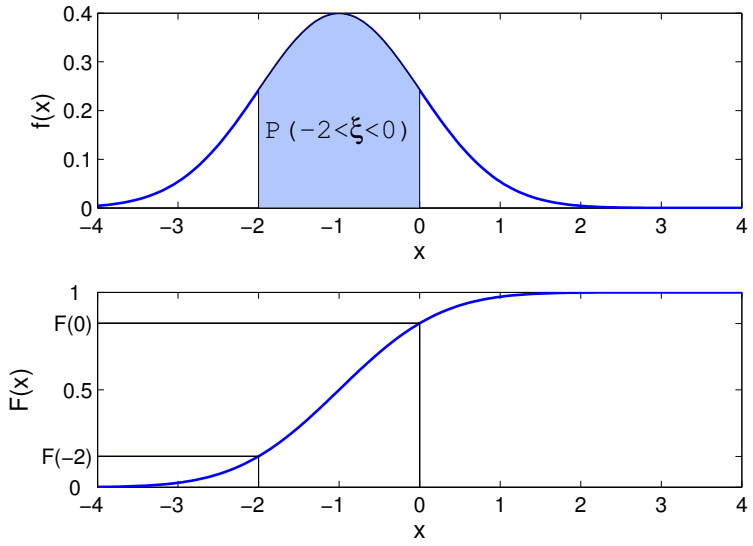


Obr. 2: Výpočet pravděpodobností na nekonečném intervalu

Výpočet pravděpodobností pomocí $F(x)$ a $f(x)$ na konečném intervalu:

$$P(-2 < X < 0) = F(0) - F(-2) = \int_{-2}^{0} f(t)\,\mathrm{d}t.$$

Toto je znázorněno na obrázku 3.



Obr. 3: Výpočet pravděpodobností na konečném intervalu

Příklad 3.9. Náhodná veličina X je dána distribuční funkcí:

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x^2}{4}, & 0 \le x < 2, \\ 1, & x \ge 2. \end{cases}$$

Určete hustotu pravděpodobnosti $f(x)$, znázorněte graficky $F(x)$ a $f(x)$, a vypočtěte $P(0{,}4 \le X < 1{,}6)$.

Řešení: Hustotu pravděpodobnosti $f(x)$ získáme derivací distribuční funkce $F(x)$:

$$f(x) = \begin{cases} 0, & x < 0, \\ \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{x^2}{4}\right) = \frac{x}{2}, & 0 \leq x < 2, \\ 0, & x \geq 2. \end{cases}$$

Graf distribuční funkce $F(x)$ a hustoty pravděpodobnosti $f(x)$ by vypadal následovně:

- Distribuční funkce $F(x)$: Kvadratický nárůst od 0 do 1 v intervalu $0 \leq x < 2$.
- Hustota pravděpodobnosti $f(x)$: Lineární funkce rostoucí od 0 do 1 v intervalu $0 \leq x < 2$.

Pravděpodobnost $P(0{,}4 \leq X < 1{,}6)$ vypočítáme jako:

$$P(0{,}4 \leq X < 1{,}6) = F(1{,}6) - F(0{,}4) = \frac{1{,}6^2}{4} - \frac{0{,}4^2}{4} = \frac{2{,}56}{4} - \frac{0{,}16}{4} = \frac{2{,}4}{4} = 0{,}6.$$

□

Příklad 3.10. Hustota pravděpodobnosti náhodné veličiny X má tvar:

$$f(x) = \begin{cases} a \cdot x, & 0 \leq x \leq 2, \\ 0, & \text{jinak.} \end{cases}$$

Určete koeficient a, distribuční funkci $F(x)$ a vypočtěte $P(0 \leq X \leq 1)$.

Řešení: Nejdříve určíme koeficient a. Platí, že integrál hustoty pravděpodobnosti přes celý definiční obor musí být roven 1:

$$\int_0^2 a \cdot x \,\mathrm{d}x = 1.$$

Po integraci dostáváme:

$$a \cdot \int_0^2 x \,\mathrm{d}x = a \cdot \left[\frac{x^2}{2}\right]_0^2 = a \cdot \frac{4}{2} = 2a.$$

Z toho plyne, že $2a = 1$, tedy $a = \frac{1}{2}$.

Distribuční funkci $F(x)$ získáme integrací hustoty pravděpodobnosti:

$$F(x) = \begin{cases} 0, & x < 0, \\ \int_0^x \frac{1}{2} \cdot t \,\mathrm{d}t = \frac{1}{2} \cdot \frac{x^2}{2} = \frac{x^2}{4}, & 0 \leq x \leq 2, \\ 1, & x > 2. \end{cases}$$

Nyní vypočítáme pravděpodobnost $P(0 \leq X \leq 1)$:

$$P(0 \leq X \leq 1) = F(1) - F(0) = \frac{1^2}{4} - 0 = \frac{1}{4} = 0{,}25.$$

□

Příklad 3.11. Určete konstanty A a B tak, aby funkce $F(x) = A + B \cdot \arctan(x)$ definovaná pro všechna reálná čísla byla distribuční funkcí rozložení náhodné veličiny.

Řešení: Aby funkce $F(x)$ byla distribuční funkcí, musí splňovat následující podmínky:

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

Z první podmínky plyne:

$$\lim_{x\to-\infty} (A + B \cdot \arctan(x)) = A + B \cdot \left(-\frac{\pi}{2}\right) = 0.$$

Z toho vyplývá, že $A = \frac{B\pi}{2}$.

Z druhé podmínky plyne:

$$\lim_{x\to\infty} (A + B \cdot \arctan(x)) = A + B \cdot \frac{\pi}{2} = 1.$$

Dosazením $A = \frac{B\pi}{2}$ dostáváme:

$$\frac{B\pi}{2} + B \cdot \frac{\pi}{2} = 1 \quad \Rightarrow \quad B\pi = 1 \quad \Rightarrow \quad B = \frac{1}{\pi}.$$

Tedy $A = \frac{1}{2}$.

Distribuční funkce má tedy tvar:

$$F(x) = \frac{1}{2} + \frac{1}{\pi} \cdot \arctan(x).$$

□

3.3 Číselné charakteristiky náhodné veličiny

Pravděpodobnostní funkce nebo hustota nám dávají kompletní a detailní obraz o chování náhodné veličiny. V praxi ale často potřebujeme tento složitý obraz shrnout do několika málo srozumitelných čísel, abychom mohli různá data rychle porovnávat. Zajímá nás především:

- **Kde je „střed“?** (Jaký výsledek můžeme v průměru očekávat?)
- **Jak velký je „rozptyl“?** (Jak moc hodnoty kolísají kolem tohoto středu?)
- **Jaký je „tvar“?** (Je rozdělení symetrické, nebo je vychýlené na jednu stranu?)

Střední hodnota

Střední hodnota není nic jiného než **teoretický průměr**. Udává hodnotu, kolem které by se ustálil průměr výsledků, kdybychom náhodný pokus opakovali mnohokrát (ideálně nekonečněkrát).

U diskrétní veličiny ji spočítáme jako klasický vážený průměr, kde vahami jsou pravděpodobnosti jednotlivých výsledků. U spojité veličiny nahradíme sumu integrálem.

Definice 3.12. Střední hodnota (očekávaná hodnota, z angl. *Expected value*) diskrétní náhodné veličiny X je definována jako:

$$E(X) = \sum_i x_i \cdot P(X = x_i) = \sum_i x_i \cdot p_i.$$

Definice 3.13. Střední hodnota spojité náhodné veličiny X je definována jako integrál z hodnot náhodné veličiny vážených její hustotou pravděpodobnosti:

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

Modus (Nejčastější hodnota)

Kromě průměru nás často zajímá i to, jaká hodnota je v daném rozdělení ta vůbec nejtypičtější.

Definice 3.14. Modus je hodnota náhodné veličiny, která má nejvyšší pravděpodobnost výskytu. Značíme ji $Mo(X)$.

- U diskrétní veličiny je to hodnota x, pro kterou je $P(X = x)$ největší.
- U spojité veličiny je to hodnota (bod na ose x), kde má hustota $f(x)$ svůj nejvyšší vrchol.

Rozptyl a směrodatná odchylka

Představte si dva střelce: oba mají průměrný zásah přesně do středu terče (stejná střední hodnota). První střelec ale trefuje desítky a devítky, zatímco druhý střídá okraje terče (jedničky a dvojky) s náhodnými trefami do středu. Potřebujeme tedy míru, která nám ukáže, jak moc se výsledky **odchylují od průměru**.

Definice 3.15. Rozptyl (z angl. *Variance*) měří průměrnou čtvercovou odchylku hodnot od střední hodnoty. Pro diskrétní veličinu:

$$D(X) = \mathrm{Var}(X) = \sum_i (x_i - E(X))^2 \cdot p_i.$$

Pro spojitou veličinu:

$$D(X) = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - E(X))^2 \cdot f(x)\,\mathrm{d}x = E(X^2) - [E(X)]^2.$$

Protože se rozptyl počítá v „kvadrátech“ (na druhou), vycházejí nám nepraktické jednotky – pokud měříme výšku v cm, rozptyl vyjde v cm^2. Proto v praxi drtivou většinu času používáme směrodatnou odchylku, která nás odmocněním vrátí do původních jednotek.

Definice 3.16. Směrodatná odchylka je druhá odmocnina z rozptylu:

$$\sigma(X) = \sqrt{D(X)}.$$

Poskytuje nám přirozené měřítko toho, jak daleko od středu (v původních jednotkách) můžeme typicky očekávat další hodnoty.

Koeficienty šikmosti a špičatosti

Zatímco střední hodnota a rozptyl řeší polohu a šířku rozdělení, koeficienty šikmosti a špičatosti popisují jeho **tvar**.

Definice 3.17 (Koeficient šikmosti náhodné veličiny X)**.**

$$\gamma_1 = \frac{E\left[(X - E(X))^3\right]}{(\sigma(X))^3}.$$

Interpretace šikmosti (γ_1):

- $\gamma_1 = 0$: Rozdělení je symetrické (např. dokonalý zvonovitý tvar). Platí, že $E(X) \approx Mo(X)$.
- $\gamma_1 > 0$: Rozdělení je protáhlé napravo (tzv. pravostranná asymetrie). Typickým příkladem jsou **mzdy** – většina lidí má průměrnou či podprůměrnou mzdu (kopec vlevo), ale malá skupina extrémně bohatých táhne dlouhý „ocas“ grafu doprava. Platí: $Mo(X) < E(X)$.
- $\gamma_1 < 0$: Rozdělení je protáhlé nalevo (ocas grafu směřuje doleva).

Definice 3.18 (Koeficient špičatosti náhodné veličiny X)**.**

$$\gamma_2 = \frac{E\left[(X - E(X))^4\right]}{(\sigma(X))^4} - 3.$$

Interpretace špičatosti (γ_2): Špičatost měří, jak velká část pravděpodobnosti je koncentrována blízko středu v porovnání s „chvosty“ (okraji) rozdělení.

- $\gamma_2 = 0$: Normální (Gaussovo) rozdělení. (Proto je ve vzorci odečtena trojka, aby normální rozdělení vyšlo jako etalon s nulou).
- $\gamma_2 > 0$: Rozdělení je „špičatější“ s tlustšími okraji. Hodnoty jsou buď silně nahuštěné u středu, nebo naopak obsahují extrémní odchylky.
- $\gamma_2 < 0$: Rozdělení je „plošší“. Hodnoty jsou rovnoměrněji rozprostřeny do šířky, extrémy se nevyskytují tak často.

Příklad 3.19. Náhodná veličina X je dána tabulkou:

x_i	1	2	3	4
p_i	0,3	0,1	0,4	?

Určete její základní číselné charakteristiky.

Řešení: Nejprve zjistíme chybějící hodnotu pravděpodobnosti p_4:

$$p_4 = 1 - (p_1 + p_2 + p_3) = 1 - (0{,}3 + 0{,}1 + 0{,}4) = 0{,}2.$$

Nyní vypočítáme jednotlivé číselné charakteristiky. Použijeme následující tabulku:

x_i	**1**	**2**	**3**	**4**	$\sum$
p_i	0,3	0,1	0,4	0,2	**1,0**
$x_i \cdot p_i$	0,3	0,2	1,2	0,8	**2,5**
$x_i^2 \cdot p_i$	0,3	0,4	3,6	3,2	**7,5**

- Střední hodnota (průměr): $E(X) = 2{,}5$
- Rozptyl: $D(X) = E(X^2) - [E(X)]^2 = 7{,}5 - (2{,}5)^2 = 7{,}5 - 6{,}25 = 1{,}25$
- Směrodatná odchylka: $\sigma(X) = \sqrt{1{,}25} \approx 1{,}118$

□

Příklad 3.20. Náhodná veličina X má hustotu pravděpodobnosti:

$$f(x) = \begin{cases} 2x, & 0 \leq x \leq 1, \\ 0, & \text{jinak.} \end{cases}$$

Určete její základní číselné charakteristiky (střední hodnotu, rozptyl a směrodatnou odchylku).

Řešení: Nejprve vypočítáme střední hodnotu a následně rozptyl pomocí prvního a druhého obecného momentu:

1. Střední hodnota ($E(X)$):

$$E(X) = \int_0^1 x \cdot f(x)\,\mathrm{d}x = \int_0^1 2x^2\,\mathrm{d}x = 2\left[\frac{x^3}{3}\right]_0^1 = \frac{2}{3}.$$

2. Rozptyl ($D(X)$): Nejprve určíme očekávanou hodnotu $E(X^2)$:

$$E(X^2) = \int_0^1 x^2 \cdot f(x)\,\mathrm{d}x = \int_0^1 2x^3\,\mathrm{d}x = 2\left[\frac{x^4}{4}\right]_0^1 = \frac{1}{2}.$$

Rozptyl se pak vypočítá ze vztahu $D(X) = E(X^2) - [E(X)]^2$:

$$D(X) = \frac{1}{2} - \left(\frac{2}{3}\right)^2 = \frac{1}{2} - \frac{4}{9} = \frac{9-8}{18} = \frac{1}{18}.$$

3. Směrodatná odchylka ($\sigma(X)$):

$$\sigma(X) = \sqrt{D(X)} = \sqrt{\frac{1}{18}} = \frac{1}{3\sqrt{2}} \approx 0{,}236.$$

Výsledné základní číselné charakteristiky jsou:

- Střední hodnota: $E(X) = \frac{2}{3}$,
- Rozptyl: $D(X) = \frac{1}{18}$,
- Směrodatná odchylka: $\sigma(X) \approx 0{,}236$.

□

3.4 Kvantilové charakteristiky náhodné veličiny

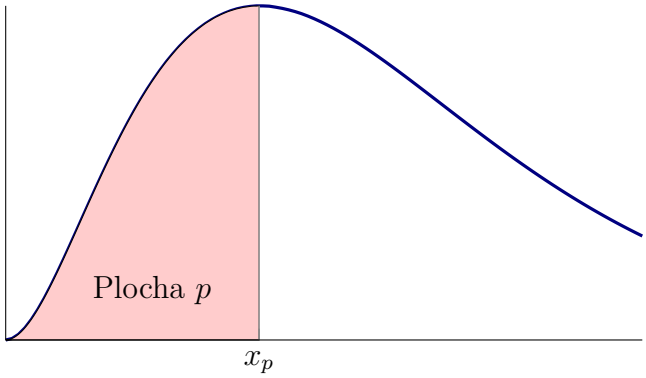
Kvantily spojitých rozdělení

Definice

Definice 3.21. Kvantil spojitého rozdělení je hodnota na ose x (viz obrázek 4), která rozděluje oblast pod hustotou pravděpodobnosti na dvě části s přesně danou plochou. Pro p-kvantil x_p platí, že plocha pod křivkou hustoty vlevo od x_p je rovna p, tj.

$$P(X \leq x_p) = F(x_p) = \int_{-\infty}^{x_p} f(t)\,\mathrm{d}t = p,$$

kde $p \in (0; 1)$.



Obr. 4: Znázornění hustoty a p-kvantilu x_p pro spojité rozdělení pravděpodobnosti (viz definici 3.21)

Speciální kvantily:

- **Medián** ($x_{0,5}$) je 50%-kvantil. Rozděluje celkovou pravděpodobnost na dvě stejné poloviny – jedna polovina hodnot leží pod mediánem, druhá polovina nad ním.
- **Kvartily** jsou kvantily, které rozdělují data na čtvrtiny. První (dolní) kvartil (Q_1) je 25%-kvantil, druhý kvartil je medián (Q_2) a třetí (horní) kvartil (Q_3) je 75%-kvantil.
- **Decily** rozdělují rozdělení na desetiny. Například první decil (D_1) je 10%-kvantil, pátý decil (D_5) odpovídá mediánu, a devátý decil (D_9) je 90%-kvantil.
- **Percentily** rozdělují rozdělení na 100 částí. Například první percentil (P_1) je 1%-kvantil, padesátý percentil (P_{50}) odpovídá mediánu a devadesátý devátý percentil (P_{99}) je 99%-kvantil.

Další běžně používané kvantily mohou zahrnovat **tercily** (dělí rozdělení na třetiny) a **kvintily** (dělí rozdělení na pětiny).

Speciálním případem kvantilu je **kritická hodnota**, používaná při statistických testech. Ta označuje mezní hodnotu, která odděluje oblast zamítnutí a nezamítnutí nulové hypotézy (viz kapitolu *Testování statistických hypotéz*).

Určování kvantilů

Kvantily se určují z tabulek nebo se pohodlně počítají pomocí softwaru. My budeme většinou používat excelovské funkce, jako jsou:

- pro **normální rozdělení** funkce `NORM.INV(p; μ; σ)`,
- pro **Studentovo rozdělení** funkce `T.INV(p; ν)` a
- pro **F-rozdělení** funkce `F.INV(p; ν1; ν2)`.

Všechny mají v názvu `INV`. Tím se poukazuje na to, že jde vlastně o inverzní funkci k distribuční funkci daného rozdělení:

$$F(x_p) = p \quad \Longleftrightarrow \quad F^{-1}(p) = x_p,$$

tedy zatímco F k zadané hodnotě x_p na ose x vypočte pravděpodobnost p, tak F^{-1} (tedy inverze k F) vypočte k zadané pravděpodobnosti p hodnotu kvantilu x_p na ose x.

Příklad 3.22. Určete první decil $x_{0,1}$ a třetí kvartil $x_{0,75}$ pro náhodnou veličinu X s hustotou pravděpodobnosti:

$$f(x) = \begin{cases} \frac{1}{2}, & 0 \leq x \leq 2, \\ 0, & \text{jinak.} \end{cases}$$

Řešení: Hustota pravděpodobnosti $f(x)$ je konstantní v intervalu $0 \leq x \leq 2$. Distribuční funkce $F(x)$ je určena jako integrál hustoty:

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x}{2}, & 0 \leq x \leq 2, \\ 1, & x > 2. \end{cases}$$

Decil $x_{0,1}$ je hodnota, pro kterou platí $F(x_{0,1}) = 0{,}1$. Hledáme tedy:

$$\frac{x_{0,1}}{2} = 0{,}1 \quad \Rightarrow \quad x_{0,1} = 0{,}2.$$

Třetí kvartil $x_{0,75}$ je hodnota, pro kterou platí $F(x_{0,75}) = 0{,}75$:

$$\frac{x_{0,75}}{2} = 0{,}75 \quad \Rightarrow \quad x_{0,75} = 1{,}5.$$

Výsledné hodnoty jsou:

- První decil: $x_{0,1} = 0{,}2$

- Třetí kvartil: $x_{0,75} = 1{,}5$

□

Příklad 3.23. Náhodná veličina X má hustotu pravděpodobnosti:

$$f(x) = \begin{cases} \frac{1}{2}x^2 e^{-x}, & 0 < x < \infty, \\ 0, & \text{jinak.} \end{cases}$$

Určete modus.

Řešení: Modus je hodnota, ve které hustota pravděpodobnosti $f(x)$ dosahuje svého maxima. Nejprve spočítáme první derivaci funkce $f(x)$ pro $x > 0$:

$$f'(x) = \frac{1}{2} \cdot \left(2xe^{-x} - x^2 e^{-x}\right) = \frac{1}{2}xe^{-x}(2 - x).$$

Poté položíme derivaci rovnu nule:

$$\frac{1}{2}xe^{-x}(2 - x) = 0.$$

Vzhledem k tomu, že $e^{-x} > 0$ pro všechna x, má tato rovnice kořeny $x = 0$ a $x = 2$. Jelikož hledáme maximum na intervalu $x > 0$, uvažujeme pouze stacionární bod $x = 2$.

Ověříme, že se jedná skutečně o maximum, a to výpočtem druhé derivace (derivujeme výraz $\frac{1}{2}e^{-x}(2x - x^2)$):

$$f''(x) = \frac{1}{2}\left[-e^{-x}(2x - x^2) + e^{-x}(2 - 2x)\right] = \frac{1}{2}e^{-x}(x^2 - 4x + 2).$$

Dosadíme náš stacionární bod $x = 2$:

$$f''(2) = \frac{1}{2}e^{-2}(2^2 - 4 \cdot 2 + 2) = \frac{1}{2}e^{-2}(4 - 8 + 2) = \frac{1}{2}e^{-2}(-2) = -e^{-2}.$$

Protože hodnota druhé derivace je záporná ($-e^{-2} < 0$), jedná se v tomto bodě o lokální maximum.

Výsledný modus je Mo = 2. □

Σ

Tato kapitola se zaměřuje na **náhodné veličiny** a jejich základní charakteristiky. Náhodné veličiny jsou proměnné, které nabývají různých číselných hodnot v závislosti na výsledku náhodného pokusu. Kapitola vysvětluje rozdíl mezi **diskrétními** a **spojitými** náhodnými veličinami a ukazuje, že zatímco diskrétní veličiny popisujeme **pravděpodobnostní funkcí**, u spojitých využíváme **hustotu pravděpodobnosti**. Společným nástrojem pro oba typy je pak **distribuční funkce**, která představuje kumulativní pravděpodobnost.

Hlavními číselnými charakteristikami náhodných veličin jsou **střední hodnota** a **rozptyl**, které poskytují informace o teoretické průměrné hodnotě veličiny a o tom, jak moc

se jednotlivé hodnoty od tohoto průměru odchylují. V kapitole jsou vysvětleny i další charakteristiky, jako **šikmost** a **špičatost**, které popisují asymetrii a celkový tvar rozdělení. Důležitou mírou polohy jsou také **kvantily** (např. medián, kvartily či decily).

Pro **diskrétní náhodné veličiny** jsou uvedeny postupy výpočtu střední hodnoty a rozptylu na základě vážených součtů. U **spojitých náhodných veličin** se k určení těchto charakteristik používají **integrály**.

1. Co je to náhodná veličina?
2. Jaký je rozdíl mezi diskrétní a spojitou náhodnou veličinou?
3. Jakým způsobem se vyjadřuje pravděpodobnostní funkce pro diskrétní náhodnou veličinu?
4. Co je to distribuční funkce a jaký má význam?
5. Jak se počítá střední hodnota pro diskrétní náhodnou veličinu?
6. Jaký je vztah mezi pravděpodobnostní funkcí (resp. hustotou pravděpodobnosti) a distribuční funkcí?
7. Co je to rozptyl a jak se počítá pro náhodnou veličinu?
8. Jaký je význam charakteristik šikmosti a špičatosti pro popis náhodné veličiny?
9. Náhodná veličina X nabývá hodnot $1, 2, 3, 4$ s pravděpodobnostmi $0{,}1; 0{,}2; 0{,}3; 0{,}4$. Vypočítejte střední hodnotu a rozptyl veličiny X. [Střední hodnota: 3,0; Rozptyl: 1,0]
10. Pro spojitou náhodnou veličinu X je dána hustota pravděpodobnosti $f(x) = 2x$ pro $x \in \langle 0; 1\rangle$ a $f(x) = 0$ jinak. Vypočítejte střední hodnotu a rozptyl této veličiny. [Střední hodnota: $\frac{2}{3}$; Rozptyl: $\frac{1}{18}$]
11. Představte si hod kostkou, kde náhodná veličina X udává počet padlých bodů. Sestrojte pravděpodobnostní a distribuční funkci této náhodné veličiny. [Pravděpodobnostní funkce: $P(X = k) = \frac{1}{6}$ pro $k = 1, 2, 3, 4, 5, 6$; Distribuční funkce: $F(x) = 0$ pro $x < 1$, $F(x) = \frac{k}{6}$ pro $k \leq x < k+1$ (kde $k \in \{1, 2, 3, 4, 5\}$) a $F(x) = 1$ pro $x \geq 6$]
12. Hustota pravděpodobnosti náhodné veličiny X má tvar:

$$f(x) = \begin{cases} 0, & \text{pro } x < 1, \\ x - \frac{1}{2}, & \text{pro } 1 \leq x < 2, \\ 0, & \text{pro } x \geq 2. \end{cases}$$

    Určete distribuční funkci. [Distribuční funkce $F(x)$ je dána: $F(x) = 0$ pro $x < 1$, $F(x) = \frac{x^2}{2} - \frac{x}{2}$ pro $1 \leq x < 2$, $F(x) = 1$ pro $x \geq 2$]
13. Náhodná veličina X je určena tabulkou:

X	-2	0	2	4	6
$P(X = x_i)$	0,1	?	0,2	0,3	0,2

    Určete hodnotu pravděpodobnosti pro $X = 0$, distribuční funkci a pravděpodobnost jevu, že náhodná veličina nabude kladných hodnot. [Pravděpodobnost pro $X = 0$: 0,2; Distribuční funkce nabývá ve skocích hodnot $F(-2) = 0{,}1$, $F(0) = 0{,}3$, $F(2) = 0{,}5$, $F(4) = 0{,}8$, $F(6) = 1$; Pravděpodobnost kladných hodnot: 0,7]

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 4

Základní typy rozdělení pravděpodobnosti diskrétní náhodné veličiny

Po prostudování této kapitoly budete umět:

- rozpoznat situace, kdy je vhodné k modelování použít binomické, Poissonovo nebo hypergeometrické rozdělení,
- vypočítat pravděpodobnosti a další charakteristiky u konkrétních diskrétních rozdělení,
- aplikovat poznatky na modelování situací z reálného života pomocí těchto rozdělení,
- pomocí excelovských funkcí vypočíst hodnoty pravděpodobnostních a distribučních funkcí.

Klíčová slova:

Diskrétní náhodná veličina, binomické rozdělení, hypergeometrické rozdělení, Poissonovo rozdělení, pravděpodobnostní funkce, distribuční funkce.

Náhled kapitoly

V této kapitole se zaměříme na základní typy rozdělení pravděpodobnosti, které se používají u diskrétních náhodných veličin. Probereme binomické, hypergeometrické a Poissonovo rozdělení. Ukážeme si, jak každé z nich funguje a kdy se používá. Důraz bude kladen nejen na teorii, ale především na praktické příklady, které ukáží, jak tato rozdělení použít při řešení reálných problémů i umělých modelových situací. Tato rozdělení tvoří nezbytný základ pro mnoho aplikací statistiky a pravděpodobnosti v praxi.

Cíle kapitoly

Cílem je pochopit různé typy rozdělení pravděpodobnosti u diskrétních náhodných veličin s ohledem na jejich využití při modelování reálných procesů.

Časová náročnost

Na tuto kapitolu si vyhraďte přibližně 3 hodiny. Tento čas zahrnuje jak studium teorie, tak procvičování příkladů a praktických aplikací, které vám pomohou lépe pochopit chování a využití probíraných rozdělení.

4.1 Binomické rozdělení

Definice

Definice 4.1. Binomické rozdělení $\mathrm{Bi}(n;p)$ modeluje počet úspěchů v pevně daném počtu nezávislých pokusů, kde každý pokus má dva možné výsledky (úspěch nebo neúspěch) a pravděpodobnost úspěchu je ve všech pokusech konstantní.

Pravděpodobnost k úspěchů z n pokusů je dána vzorcem:

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

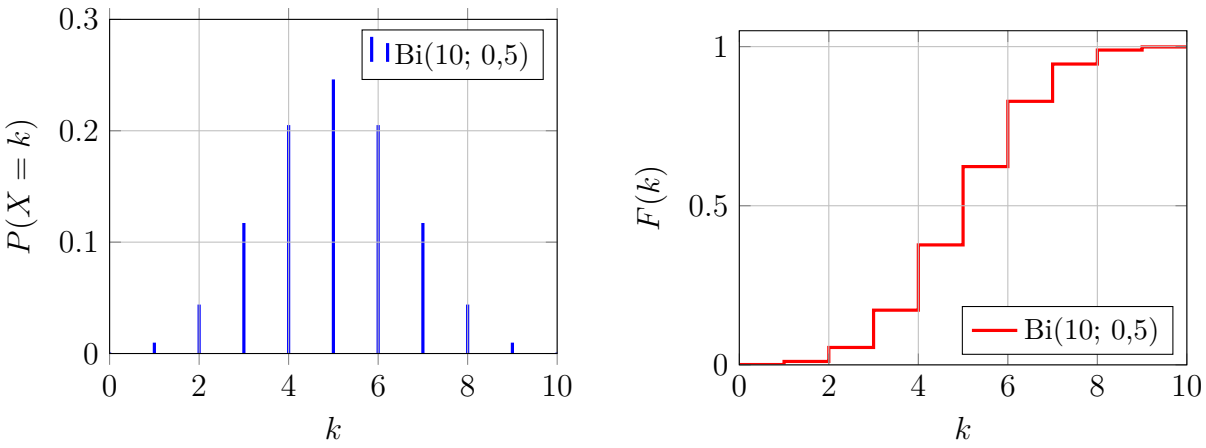
kde n je celkový počet pokusů, k je počet úspěchů ($k = 0, 1, \ldots, n$), p je pravděpodobnost úspěchu v každém jednotlivém pokusu a $\binom{n}{k}$ je kombinační číslo.

Základní číselné charakteristiky

- **Střední hodnota:** $E(X) = n \cdot p$
- **Rozptyl:** $D(X) = n \cdot p \cdot (1 - p)$

Grafy pravděpodobnostní a distribuční funkce

Grafy pravděpodobnostní funkce (PDF) a distribuční funkce (CDF) pro binomické rozdělení s počtem pokusů $n = 10$ a pravděpodobností úspěchu $p = 0{,}5$ jsou na obrázku 5.



Obr. 5: Pravděpodobnostní a distribuční funkce binomického rozdělení pro $n = 10$ a $p = 0{,}5$

Excelovské funkce

Pro práci s binomickým rozdělením lze v Excelu využít následující funkce:

- **Pravděpodobnostní funkce:** Funkce `BINOM.DIST(k; n; p; NEPRAVDA)` vrací pravděpodobnost přesně k úspěchů z n pokusů.

- **Distribuční funkce:** Funkce `BINOM.DIST(k; n; p; PRAVDA)` vrací kumulativní pravděpodobnost, tedy pravděpodobnost, že nastane nejvýše k úspěchů (tj. $0, 1, \ldots, k$ úspěchů).

4.2 Hypergeometrické rozdělení

Definice

Definice 4.2. Hypergeometrické rozdělení $\mathrm{Hg}(N; M; n)$ modeluje počet úspěchů při náhodném výběru n objektů z celkové populace N, kde přesně M objektů z této populace představuje úspěch. Výběr probíhá **bez vracení** (vybraný objekt se nevrací zpět, čímž se mění pravděpodobnost v dalším tahu).

Pravděpodobnost právě k úspěchů je dána vzorcem:

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},$$

kde N je velikost populace, M je celkový počet úspěšných objektů v populaci, n je počet vybíraných objektů (velikost vzorku) a k je počet úspěchů ve vzorku.

Základní číselné charakteristiky

- **Střední hodnota:** $E(X) = n\frac{M}{N}$
- **Rozptyl:** $D(X) = n\frac{M}{N}\left(1 - \frac{M}{N}\right)\frac{N-n}{N-1}$

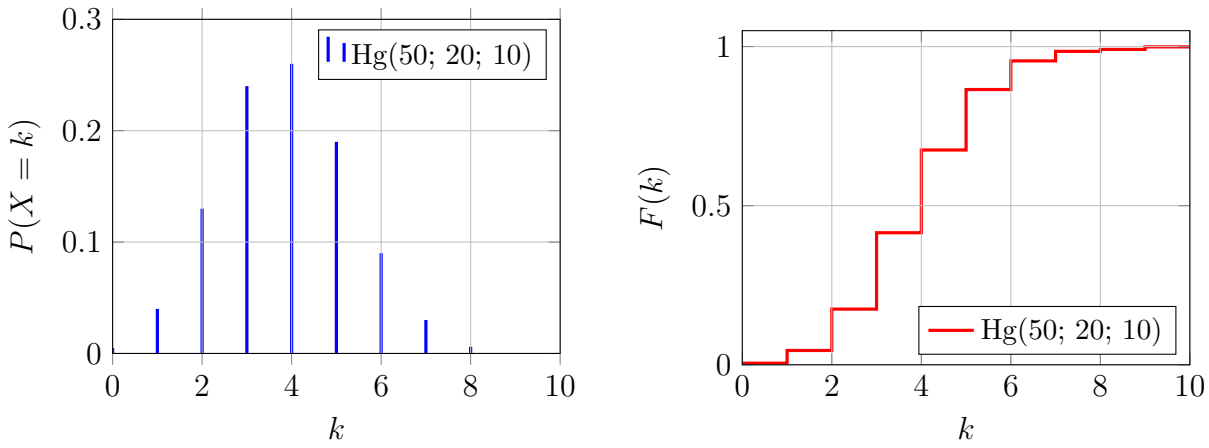
Grafy pravděpodobnostní a distribuční funkce

Grafy pravděpodobnostní funkce (PDF) a distribuční funkce (CDF) pro hypergeometrické rozdělení s parametry $N = 50$, $M = 20$, $n = 10$ jsou na obrázku 6.

Excelovské funkce

Pro práci s hypergeometrickým rozdělením lze v Excelu použít následující funkce:

- **Pravděpodobnostní funkce:** Funkce `HYPGEOM.DIST(k; n; M; N; NEPRAVDA)` vrací pravděpodobnost přesně k úspěchů.
- **Distribuční funkce:** Funkce `HYPGEOM.DIST(k; n; M; N; PRAVDA)` vrací kumulativní pravděpodobnost, tedy pravděpodobnost nejvýše k úspěchů.



Obr. 6: Pravděpodobnostní a distribuční funkce hypergeometrického rozdělení pro $N = 50$, $M = 20$ a $n = 10$

4.3 Poissonovo rozdělení

Kdy použít Poissonovo rozdělení?

Na rozdíl od binomického nebo hypergeometrického rozdělení, kde máme pevně daný celkový počet pokusů n (a tedy nemůžeme mít více úspěchů než n), u Poissonova rozdělení **neexistuje horní hranice** možného počtu událostí ($k = 0, 1, 2, \dots$).

Tento model se používá pro situace, kdy počítáme výskyt (často poměrně vzácných) událostí, které nastávají náhodně v nějakém **spojitém kontinuu** – typicky v čase, na určité ploše nebo v určitém objemu.

Typické příklady Poissonova rozdělení:

- Počet zákazníků, kteří přijdou k pokladně během jedné hodiny.
- Počet tiskových chyb na jedné stránce knihy.
- Počet dopravních nehod na určité křižovatce za měsíc.
- Počet kazů na 100 metrech vyrobené látky.

Další velmi důležitou vlastností je, že Poissonovo rozdělení skvěle funguje jako **aproximace binomického rozdělení** pro situace, kdy máme obrovský počet pokusů ($n \to \infty$), ale pravděpodobnost úspěchu v jednom pokusu je mizivě malá ($p \to 0$). V takovém případě se výpočet obrovských kombinačních čísel nahradí mnohem jednodušším Poissonovým rozdělením, kde stačí položit střední hodnotu $\lambda = n \cdot p$.

Definice

Definice 4.3. Poissonovo rozdělení Po(λ) modeluje počet událostí, které nastanou v pevně daném čase nebo prostoru, za předpokladu, že tyto události nastávají nezávisle na sobě s konstantní střední intenzitou (průměrem) λ.

Pravděpodobnost, že v daném intervalu nastane právě k událostí, je dána vzorcem:

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

kde λ je očekávaný (průměrný) počet událostí v daném intervalu, k je sledovaný počet událostí a e je Eulerovo číslo.

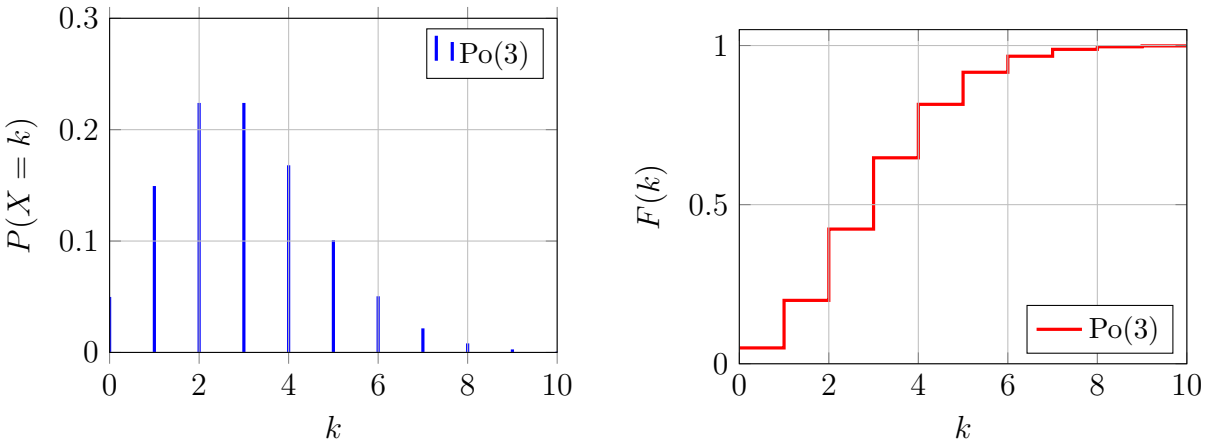
Základní číselné charakteristiky

- **Střední hodnota:** $E(X) = \lambda$
- **Rozptyl:** $D(X) = \lambda$

Poznámka: Poissonovo rozdělení je unikátní tím, že se jeho střední hodnota rovná rozptylu.

Grafy pravděpodobnostní a distribuční funkce

Grafy pravděpodobnostní funkce (PDF) a distribuční funkce (CDF) pro Poissonovo rozdělení s parametrem $\lambda = 3$ jsou na obrázku 7.



Obr. 7: Pravděpodobnostní a distribuční funkce Poissonova rozdělení pro $\lambda = 3$

Excelovské funkce

Pro práci s Poissonovým rozdělením lze v Excelu použít následující funkce:

- **Pravděpodobnostní funkce:** Funkce `POISSON.DIST(k; λ; NEPRAVDA)` vrací pravděpodobnost, že nastane přesně k událostí.
- **Distribuční funkce:** Funkce `POISSON.DIST(k; λ; PRAVDA)` vrací kumulativní pravděpodobnost, tedy že nastane nejvýše k událostí.

4.4 Některá další diskrétní rozdělení

Než se pustíme do složitějších výpočtů, zmíníme pro úplnost ještě dvě velmi jednoduchá rozdělení, se kterými se v praxi (a často i v běžném životě) setkáváme zcela intuitivně.

1. Alternativní rozdělení $Alt(p)$
Popisuje ten vůbec nejjednodušší náhodný pokus, který má pouze **dva možné výsledky** – typicky úspěch, nebo neúspěch. Uvažujme například hod mincí.

- Výsledkem je náhodná veličina X, která nabývá pouze hodnot 1 (úspěch, např. padne líc) s pravděpodobností p, nebo 0 (neúspěch, padne rub) s pravděpodobností $1 - p$.

Poznámka: Binomické rozdělení není nic jiného než součet n nezávislých alternativních rozdělení.

2. Diskrétní rovnoměrné rozdělení $R(n)$
Popisuje situaci, kdy má všech n možných výsledků náhodného pokusu **zcela stejnou pravděpodobnost**.

- Typickým příkladem je hod klasickou šestistěnnou kostkou. Prostor možných výsledků je $M = \{1, 2, 3, 4, 5, 6\}$. Každé číslo má pravděpodobnost přesně $\frac{1}{6}$. Modelujeme jej jako $R(6)$.

4.5 Řešené příklady

Binomické rozdělení

Příklad 4.4 (Binomické rozdělení)**.** Student má potíže s ranním vstáváním. Proto někdy zaspí a nestihne přednášku, která začíná již v 9 hodin. Pravděpodobnost, že zaspí, je 0,3. V semestru je 12 přednášek, což znamená 12 nezávislých pokusů dorazit na přednášku včas. Nalezněte pravděpodobnost, že student nestihne přednášku v důsledku zaspání v polovině nebo více případů.

Řešení: Jedná se o binomické rozdělení $Bi(12;0{,}3)$ s parametry $n = 12$ a $p = 0{,}3$. Hledaná pravděpodobnost (zaspí v 6 a více případech) je:

$$P(X \geq 6) = 1 - P(X \leq 5).$$

Tuto pravděpodobnost lze snadno vypočítat pomocí distribuční funkce binomického rozdělení, například pomocí funkce `BINOM.DIST` v Excelu:

$$P(X \geq 6) = 1 - \texttt{BINOM.DIST(5; 12; 0,3; PRAVDA)} \approx 1 - 0{,}8822 = 0{,}1178.$$

Pravděpodobnost, že zaspí polovinu a více přednášek, je zhruba 11,8 %. □

Příklad 4.5 (Binomické rozdělení)**.** V obchodě probíhá reklamní akce pro zákazníky. Z dlouhodobých statistik je známo, že šance na výhru reklamního dárku je pro každého zákazníka 5 % (tedy $p = 0{,}05$) a výsledky jednotlivých zákazníků jsou na sobě nezávislé. Jaká je pravděpodobnost, že z 20 nově příchozích zákazníků alespoň 2 vyhrají?

Řešení: Tento problém modelujeme jako binomické rozdělení $Bi(20;0{,}05)$ s parametry $n = 20$ a $p = 0{,}05$. Hledáme pravděpodobnost:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)].$$

Pravděpodobnosti pro 0 a 1 výherce (lze spočítat dosazením do vzorce nebo v Excelu jako `BINOM.DIST(k; 20; 0,05; NEPRAVDA)`) jsou:

$$P(X = 0) = 0{,}3585 \quad \text{a} \quad P(X = 1) = 0{,}3773.$$

Proto:

$$P(X \geq 2) = 1 - (0{,}3585 + 0{,}3773) = 1 - 0{,}7358 = 0{,}2642.$$

Pravděpodobnost, že vyhrají alespoň 2 zákazníci z 20, je přibližně 26,4 %. □

Poissonovo rozdělení

Příklad 4.6 (Poissonovo rozdělení)**.** Předpokládejme, že realitní makléř jedná v průměru s pěti zákazníky za den. Zjistěte, jaká je pravděpodobnost, že počet zákazníků makléře za jeden den bude větší než 4.

Řešení: Náhodná veličina X – počet zákazníků – splňuje kritéria pro Poissonovo rozdělení $Po(\lambda)$ s průměrem $\lambda = 5$. Hledáme pravděpodobnost, že $X > 4$:

$$P(X > 4) = 1 - P(X \leq 4).$$

Tuto pravděpodobnost lze vypočítat pomocí kumulativní funkce `POISSON.DIST` v Excelu:

$$P(X > 4) = 1 - \texttt{POISSON.DIST(4; 5; PRAVDA)} \approx 1 - 0{,}4405 = 0{,}5595.$$

Pravděpodobnost, že bude jednat s více než 4 zákazníky, je necelých 56 %. □

Příklad 4.7 (Poissonovo rozdělení)**.** V průměru přistanou na místním letišti během jedné hodiny 3 letadla. Jaká je pravděpodobnost, že během jedné hodiny přistanou přesně 2 letadla?

Řešení: Náhodná veličina X – počet přistání – splňuje kritéria pro Poissonovo rozdělení $Po(\lambda)$ s parametrem $\lambda = 3$. Hledaná pravděpodobnost je:

$$P(X = 2) = \frac{3^2 e^{-3}}{2!} \approx 0{,}2240.$$

Tuto pravděpodobnost lze případně snadno vypočítat i pomocí funkce `POISSON.DIST(2; 3; NEPRAVDA)` v Excelu. □

Hypergeometrické rozdělení

Příklad 4.8 (Hypergeometrické rozdělení)**.** Mezi stovkou výrobků je 20 zmetků. Vybereme deset výrobků a sledujeme počet zmetků mezi vybranými.

Řešení: V tomto případě má náhodná veličina X (počet vybraných zmetků) hypergeometrické rozdělení $Hg(100; 20; 10)$. Pravděpodobnostní funkce je dána vztahem:

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}},$$

kde $N = 100$, $M = 20$, $n = 10$ a k je počet zmetků mezi vybranými výrobky.

Například pravděpodobnost, že mezi deseti vybranými výrobky budou přesně 3 zmetky, lze vypočítat jako $P(X = 3)$:

$$P(X = 3) = \frac{\binom{20}{3}\binom{80}{7}}{\binom{100}{10}} \approx 0{,}2092.$$

Tuto pravděpodobnost snadno získáme i v Excelu pomocí funkce:

`HYPGEOM.DIST(3; 10; 20; 100; NEPRAVDA).`

□

Příklad 4.9 (Hypergeometrické rozdělení)**.** V krabici je 20 kuliček, z nichž 8 je červených a 12 modrých. Náhodně vybereme 5 kuliček bez vracení. Jaká je pravděpodobnost, že vybereme přesně 3 červené kuličky?

Řešení: Tento problém modelujeme jako hypergeometrické rozdělení $Hg(N; M; n)$, kde úspěchem je vytažení červené kuličky. Parametry jsou:

$$N = 20, \quad M = 8, \quad n = 5.$$

Hledaná pravděpodobnost je:

$$P(X = 3) = \frac{\binom{8}{3}\binom{12}{2}}{\binom{20}{5}}.$$

Po dosazení hodnot (kombinačních čísel) dostáváme:

$$P(X = 3) = \frac{56 \cdot 66}{15504} = \frac{3696}{15504} \approx 0{,}2384.$$

Pravděpodobnost vytažení přesně 3 červených kuliček je tedy zhruba 23,8 %.

I tento výpočet lze snadno provést pomocí funkce v Excelu: `HYPGEOM.DIST(3; 5; 8; 20; NEPRAVDA).` □

V této kapitole byla představena základní diskrétní rozdělení pravděpodobnosti, která se velmi často vyskytují v praxi (některým jsme se věnovali podrobněji, dalším jen v rámci úvodních příkladů). Seznámili jsme se s charakteristikami a použitím těchto rozdělení:

- **Alternativní rozdělení** $Alt(p)$ – Popisuje náhodný experiment se dvěma možnými výsledky (úspěch/neúspěch). Využívá se např. při modelování jednoho hodu mincí.
- **Rovnoměrné rozdělení** $R(n)$ – Předpokládá, že všech n možných výsledků má stejnou pravděpodobnost. Používá se např. při hodu spravedlivou kostkou.
- **Binomické rozdělení** $Bi(n; p)$ – Popisuje počet úspěchů při pevně daném počtu n nezávislých pokusů, kde každý pokus má stejnou pravděpodobnost úspěchu p. Příkladem je situace, kdy se sleduje počet ranních zaspání studenta během celého semestru.
- **Poissonovo rozdělení** $Po(\lambda)$ – Používá se k modelování počtu výskytů událostí v pevném časovém nebo prostorovém intervalu, kde není shora omezen počet pokusů. V praxi jde např. o modelování počtu zákazníků přicházejících k realitnímu makléři.
- **Hypergeometrické rozdělení** $Hg(N; M; n)$ – Popisuje pravděpodobnost určitého počtu úspěchů při výběru n objektů **bez vracení** z konečné populace N. Příkladem je sledování počtu vadných výrobků při jednorázovém náhodném výběru vzorku z výrobní dávky.

?

1. Jaké jsou základní číselné charakteristiky binomického rozdělení?
2. Jak vypadá pravděpodobnostní funkce binomického rozdělení pro $n = 10$ a $p = 0{,}5$?
3. Co modeluje Poissonovo rozdělení?
4. Jaký je vzorec pro pravděpodobnost, že Poissonova náhodná veličina X nabude hodnoty k, pokud má parametr λ?
5. Jaký je vztah mezi střední hodnotou a rozptylem u Poissonova rozdělení?
6. Jaké typické aplikace má Poissonovo rozdělení v reálném světě?
7. Co modeluje hypergeometrické rozdělení?
8. Jaký je rozdíl mezi binomickým a hypergeometrickým rozdělením z hlediska způsobu výběru?
9. V dodávce 80 polotovarů je 8 (tj. 10 %) vadných. Náhodně vybereme (najednou, tj. „bez vracení“) 5 kusů polotovarů k další kompletaci. Jaká je pravděpodobnost, že mezi vybranými prvky bude maximálně jeden vadný? [0,9246]
10. Ve skladišti závodu je 5 000 výrobků stejného typu. Pravděpodobnost toho, že daný výrobek nevydrží kontrolní zapojení (je vadný), je 0,1 %. Najděte pravděpodobnost, že z výrobků na skladě více než dva nevydrží kontrolní zapojení. [0,8753]
11. Korektura 500 stránek obsahuje celkem 500 nalezených tiskových chyb. Najděte pravděpodobnost toho, že na jedné náhodně vybrané stránce jsou nejméně tři chyby. [0,0803]
12. Najděte pravděpodobnost toho, že mezi 200 náhodně vybranými výrobky se vyskytnou více než tři zmetky, když v průměru je zmetkovitost výroby těchto výrobků 1 %. [0,1429 pomocí Poissonovy aproximace, resp. 0,1420 při přesném výpočtu binomickým rozdělením]

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 5

Základní typy rozdělení pravděpodobnosti spojité náhodné veličiny

Po prostudování této kapitoly budete umět:

- vyjmenovat základní spojitá rozdělení pravděpodobnosti i s jejich důležitými vlastnostmi,
- vypočítat základní charakteristiky daných typů rozdělení pravděpodobnosti,
- pomocí excelovských funkcí vypočíst hodnoty hustoty pravděpodobnosti a distribučních funkcí spojitých rozdělení,
- pomocí excelovských funkcí vypočíst kvantily spojitých rozdělení.

Klíčová slova:

Spojitá náhodná veličina, rovnoměrné rozdělení, exponenciální rozdělení, normální rozdělení, hustota pravděpodobnosti, distribuční funkce, střední hodnota, rozptyl, kvantil.

Náhled kapitoly

Tato kapitola se zaměřuje na základní typy rozdělení pravděpodobnosti pro spojité náhodné veličiny. Seznámíme se s rozděleními, jako je rovnoměrné, exponenciální a normální rozdělení. Každé z těchto rozdělení má specifické vlastnosti a používá se v různých situacích při modelování náhodných jevů. Kromě teoretického popisu si také ukážeme, jak tato rozdělení aplikovat v praxi a jak vypočítat pravděpodobnosti, kvantily a další charakteristiky. V kapitole jsou uvedeny příklady, které demonstrují užití spojitých rozdělení v reálných situacích.

Cíle kapitoly

Cílem je pochopit a rozlišovat základní typy rozdělení pravděpodobnosti pro spojité náhodné veličiny a aplikovat tyto poznatky při řešení úloh z praxe.

Časová náročnost

Pro tuto kapitolu doporučujeme vyčlenit přibližně 3 hodiny. Tento čas zahrnuje jak studium teoretických částí, tak procvičování praktických příkladů a aplikací.

5.1 Normální rozdělení

Kde se s ním setkáme a proč je tak důležité?

Normální (Gaussovo) rozdělení je nejdůležitějším rozdělením v celé statistice. Nese název „normální“, protože velmi dobře popisuje chování mnoha veličin v přírodě i ve společnosti za „normálních“ okolností – tedy tam, kde na výsledek působí velké množství drobných, vzájemně nezávislých a náhodných vlivů.

Typické příklady normálního rozdělení:

- Výška a hmotnost dospělých lidí v určité populaci.
- Hodnoty IQ v populaci.
- Chyby při fyzikálních měřeních (nepřesnost přístroje a pozorovatele).
- Rozměry součástek sjíždějících z výrobní linky.

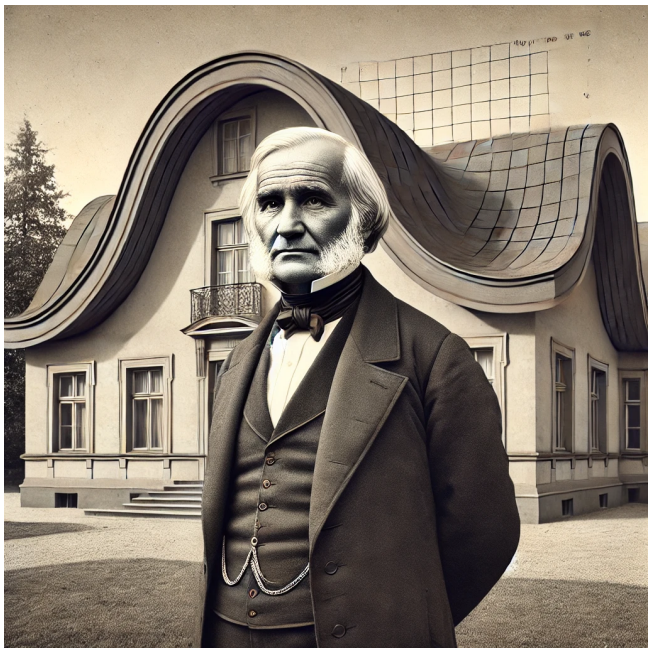
Definice

Definice 5.1. Normální rozdělení $N(\mu; \sigma^2)$ je spojité rozdělení pravděpodobnosti, které je symetrické kolem své střední hodnoty μ a má typický zvonovitý tvar (tzv. Gaussova křivka). Je jednoznačně určeno dvěma parametry: střední hodnotou μ (určuje polohu vrcholu) a směrodatnou odchylkou σ (určuje šířku a zploštění zvonu).

Hustota normálního rozdělení je dána vzorcem:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

kde μ je střední hodnota a σ^2 je rozptyl.



Obr. 8: Jeden z hrdých otců normálního rozdělení (vytvořeno pomocí ChatGPT, OpenAI)

Základní číselné charakteristiky a Empirické pravidlo

- **Střední hodnota:** $E(X) = \mu$ (Je to zároveň i modus a medián).
- **Rozptyl:** $D(X) = \sigma^2$
- **Symetrie:** Rozdělení je dokonale symetrické, koeficient šikmosti $\gamma_1 = 0$.

Velmi užitečnou pomůckou pro rychlou představu o datech s normálním rozdělením je tzv. **Pravidlo tří sigma (Empirické pravidlo 68–95–99,7 %)**. Říká nám, kolik procent všech hodnot leží v určitých vzdálenostech od průměru:

- Přibližně **68,3 %** hodnot leží v intervalu $\langle \mu - \sigma; \mu + \sigma \rangle$.
- Přibližně **95,5 %** hodnot leží v intervalu $\langle \mu - 2\sigma; \mu + 2\sigma \rangle$.
- Přibližně **99,7 %** hodnot leží v intervalu $\langle \mu - 3\sigma; \mu + 3\sigma \rangle$ (téměř všechny hodnoty).

Normované (standardizované) normální rozdělení

Pro usnadnění výpočtů a možnost používat statistické tabulky se zavádí speciální případ. Pokud má veličina střední hodnotu $\mu = 0$ a rozptyl $\sigma^2 = 1$, hovoříme o **normovaném normálním rozdělení** a značíme jej $N(0; 1)$.

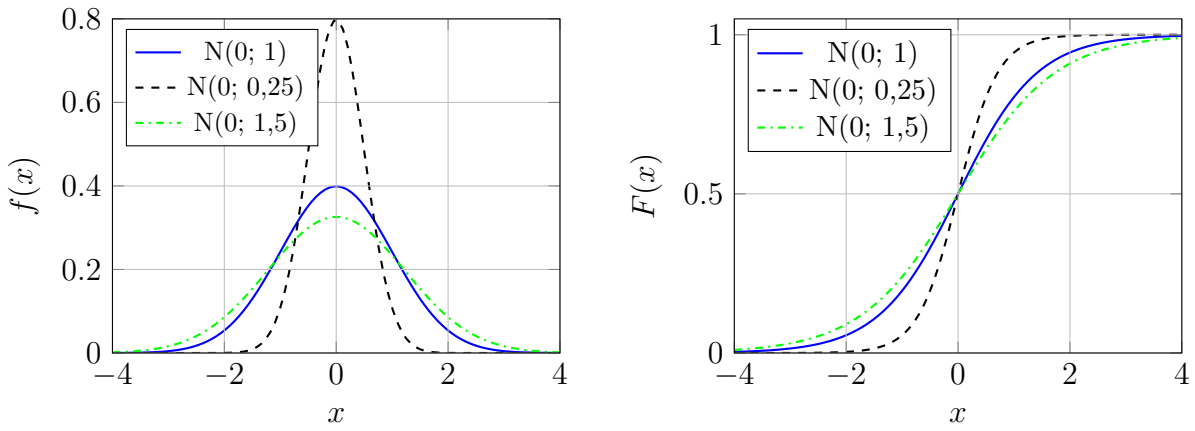
Jakoukoliv náhodnou veličinu X s normálním rozdělením $N(\mu; \sigma^2)$ můžeme jednoduše převést (standardizovat) na normovanou veličinu Z pomocí transformace:

$$Z = \frac{X - \mu}{\sigma}.$$

Veličina Z nám pak udává, o kolik směrodatných odchylek se původní hodnota X liší od průměru.

Grafy hustot a distribučních funkcí

Grafy znázorňující hustoty a distribuční funkce normálního rozdělení pro různé hodnoty μ a σ^2 jsou uvedeny na obrázcích 9 a 10.

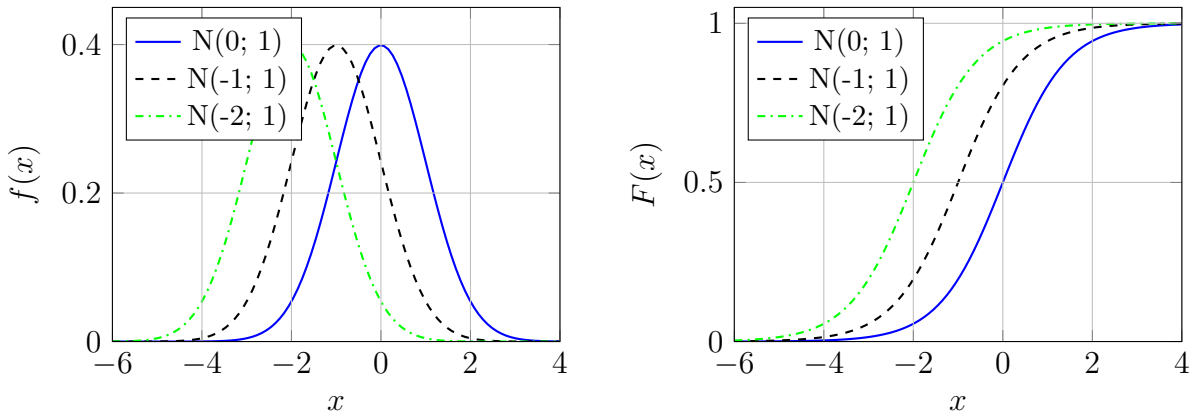


Obr. 9: Grafy hustot a distribučních funkcí normálního rozdělení s různými rozptyly

Excelovské funkce

Pro práci s normálním rozdělením lze v Excelu použít následující funkce:

- **Distribuční funkce (CDF):** Funkce `NORM.DIST(x; μ; σ; PRAVDA)` vrací hodnotu distribuční funkce (kumulativní pravděpodobnost, tedy plochu pod křivkou od $-\infty$ po x).
- **Kvantilová funkce:** Funkce `NORM.INV(p; μ; σ)` vrací kvantil pro danou pravděpodobnost p.



Obr. 10: Grafy hustot a distribučních funkcí normálního rozdělení s různými středními hodnotami

- **Hustota (PDF):** Funkce `NORM.DIST(x; μ; σ; NEPRAVDA)` vrací hodnotu hustoty. *(Pozor, u spojitých rozdělení se nejedná o pravděpodobnost! Používá se převážně k vykreslování grafů.)*

Pro práci s **normovaným normálním rozdělením** ($\mu = 0$, $\sigma = 1$) lze použít specializované zkrácené funkce:

- **Distribuční funkce (CDF):** `NORM.S.DIST(x; PRAVDA)`
- **Kvantilová funkce:** `NORM.S.INV(p)`
- **Hustota (PDF):** `NORM.S.DIST(x; NEPRAVDA)`

5.2 Rovnoměrné rozdělení

Kde se s ním setkáme?

Rovnoměrné rozdělení je vůbec nejjednodušším spojitým modelem. Používáme ho v situacích, kdy víme, že hodnota leží v určitém intervalu, ale nemáme absolutně žádný důvod předpokládat, že by se koncentrovala kolem nějakého středu (jako u normálního rozdělení) nebo na kraji. Zkrátka „všechno je stejně možné“.

Typické příklady spojitého rovnoměrného rozdělení:

- **Chyby ze zaokrouhlování:** Pokud zaokrouhlujeme čísla na celé koruny, chyba zaokrouhlení se rovnoměrně rozkládá v intervalu $\langle -0{,}5; 0{,}5 \rangle$.
- **Doba čekání:** Přijdete-li na zastávku tramvaje, u které neznáte jízdní řád, a víte jen, že jezdí přesně každých 10 minut. Vaše doba čekání je rovnoměrně rozdělena v intervalu $\langle 0; 10 \rangle$ minut.
- Generátory náhodných čísel v počítačích.

Definice

Definice 5.2. Rovnoměrné rozdělení $U(a; b)$ je spojité rozdělení pravděpodobnosti, kde každá hodnota z intervalu $\langle a; b \rangle$ má zcela stejnou šanci na výskyt (přesněji řečeno: hustota pravděpodobnosti je konstantní). Je určeno dvěma parametry: dolní mezí a a horní mezí b.

Hustota rovnoměrného rozdělení je dána vzorcem:

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{pro } a \leq x \leq b, \\ 0 & \text{jinak.} \end{cases}$$

Základní číselné charakteristiky

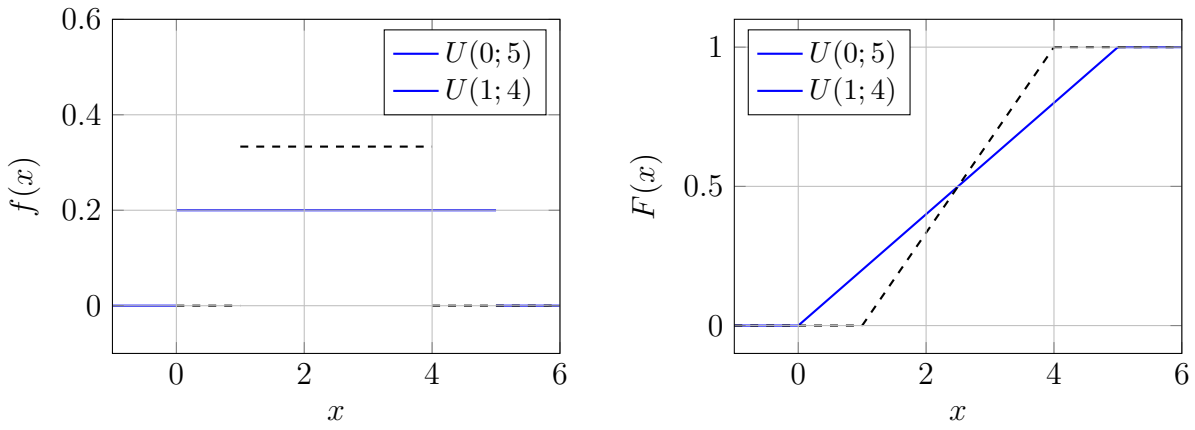
- **Střední hodnota:** $E(X) = \frac{a+b}{2}$ (Leží přesně uprostřed intervalu).
- **Rozptyl:** $D(X) = \frac{(b-a)^2}{12}$
- **Symetrie:** Rovnoměrné rozdělení je dokonale symetrické kolem své střední hodnoty ($\gamma_1 = 0$).

Grafy hustoty a distribuční funkce

Grafy hustoty a distribuční funkce rovnoměrného rozdělení pro různé hodnoty a a b jsou uvedeny na obrázku 11.

Výpočty v Excelu

Na rozdíl od normálního rozdělení **nemá** standardní Excel pro rovnoměrné rozdělení žádnou předpřipravenou funkci (jako např. `UNIFORM.DIST`). Není ale vůbec potřeba, protože vzorce jsou



Obr. 11: Grafy hustot a distribučních funkcí rovnoměrného rozdělení (různé parametry a a b)

triviální a zadáváme je do buněk pomocí obyčejné aritmetiky:

- **Hustota (PDF):** Zapíšeme vzorec `=1/(b-a)`.
- **Distribuční funkce (CDF):** Pro hodnoty x uvnitř intervalu $\langle a; b\rangle$ počítáme pravděpodobnost $P(X \leq x)$ jako `=(x-a)/(b-a)`.
- **Kvantilová funkce:** Pokud chceme najít hodnotu x pro zadanou pravděpodobnost p, otočíme vzorec: `=a + p*(b-a)`.

5.3 Exponenciální rozdělení

Kde se s ním setkáme a jak souvisí s Poissonovým rozdělením?

Exponenciální rozdělení úzce souvisí s Poissonovým rozdělením z předchozí kapitoly. Zatímco Poissonovo rozdělení nám říká, *kolik událostí* nastane za určitý čas (např. kolik zákazníků přijde do obchodu za hodinu), exponenciální rozdělení modeluje **dobu čekání mezi těmito jednotlivými událostmi** (např. jak dlouho budeme čekat, než do obchodu vejde další zákazník).

Typické příklady exponenciálního rozdělení:

- Doba mezi příjezdy dvou po sobě jdoucích autobusů na zastávku.
- Doba obsluhy jednoho zákazníka u pokladny nebo na lince podpory.
- Doba bezporuchového chodu (životnost) určitých elektronických součástek nebo žárovek.
- Časový rozestup mezi dvěma dopravními nehodami na daném úseku dálnice.

Definice

Definice 5.3. Exponenciální rozdělení $Exp(\lambda)$ je spojité rozdělení pravděpodobnosti, které se používá k modelování doby čekání na výskyt určité náhodné události. Parametr λ představuje **intenzitu** výskytu událostí (průměrný počet událostí za jednotku času).

Hustota exponenciálního rozdělení je dána vzorcem:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{pro } x \geq 0, \\ 0 & \text{pro } x < 0, \end{cases}$$

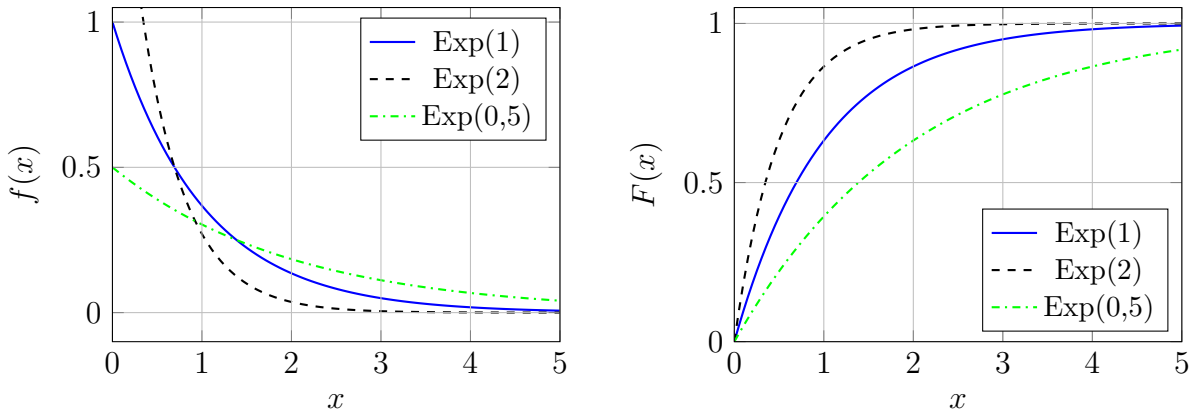
kde $\lambda > 0$ je parametr rychlosti (intenzity) a e je Eulerovo číslo.

Základní číselné charakteristiky

- **Střední hodnota:** $E(X) = \frac{1}{\lambda}$ (Tedy průměrná doba čekání).
- **Rozptyl:** $D(X) = \frac{1}{\lambda^2}$
- **Asymetrie:** Exponenciální rozdělení je silně asymetrické (pravostranná asymetrie), má dlouhý chvost směrem k vyšším hodnotám na ose x.

Grafy hustoty a distribuční funkce

Grafy hustoty a distribuční funkce exponenciálního rozdělení pro různé hodnoty λ jsou uvedeny na obrázku 12.



Obr. 12: Grafy hustot a distribučních funkcí exponenciálního rozdělení pro různé parametry λ

Výpočty v Excelu

Pro práci s exponenciálním rozdělením v Excelu můžeme použít následující postupy:

- **Distribuční funkce (CDF):** Funkce `EXPON.DIST(x; λ; PRAVDA)` vrací hodnotu distribuční funkce (pravděpodobnost, že čekání bude kratší nebo rovno x).
- **Hustota pravděpodobnosti (PDF):** Funkce `EXPON.DIST(x; λ; NEPRAVDA)` vrací hodnotu hustoty. *(Připomínáme: u spojitých rozdělení PDF neudává pravděpodobnost, funkce slouží spíše ke kreslení grafů.)*
- **Kvantilová funkce:** Excel pro exponenciální rozdělení standardně **nemá** inverzní funkci (typu `EXPON.INV`). Kvantil pro zadanou pravděpodobnost p a parametr λ se proto jednoduše spočítá inverzním vzorcem pomocí logaritmu: `=-LN(1-p)/λ`.

5.4 Řešené příklady

Příklad 5.4 (Rovnoměrné rozdělení $U(a;b)$)**.** Tramvajová linka číslo 8 odjíždí v dopoledních hodinách ze zastávky každých 10 minut. Vypočtěte pravděpodobnost, že na ni budete dopoledne čekat déle než 7 minut.

Řešení: Doba čekání je náhodná veličina X, která má rovnoměrné rozdělení pravděpodobnosti – v našem případě $U(0;10)$. Pro rovnoměrné rozdělení $U(a;b)$ platí:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b, \\ 0, & \text{jinak.} \end{cases}$$

V našem případě $a = 0$ a $b = 10$, takže hustota pravděpodobnosti je:

$$f(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10, \\ 0, & \text{jinak.} \end{cases}$$

Distribuční funkce $F(x)$ je:

$$F(x) = \begin{cases} 0, & x < 0, \\ \frac{x}{10}, & 0 \le x \le 10, \\ 1, & x > 10. \end{cases}$$

Pravděpodobnost, že budeme čekat déle než 7 minut, spočítáme pomocí distribuční funkce jako doplněk:

$$P(X > 7) = 1 - P(X \le 7) = 1 - F(7) = 1 - \frac{7}{10} = 0{,}3.$$

□

Příklad 5.5 (Exponenciální rozdělení $Exp(\lambda)$)**.** Doba čekání hosta na pivo je v restauraci U Lva průměrně 5 minut. Předpokládáme, že se řídí exponenciálním rozdělením. Určete:

1. hustotu pravděpodobnosti náhodné veličiny, která je dána dobou čekání na pivo,
2. pravděpodobnost, že budeme čekat na pivo déle než 12 minut,
3. dobu čekání, během které bude zákazník obsloužen s pravděpodobností 0,9.

Řešení: 1. Hustota pravděpodobnosti pro exponenciální rozdělení $Exp(\lambda)$ je dána vztahem:

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

V našem případě je střední doba čekání $E(X) = \frac{1}{\lambda} = 5$, takže intenzita $\lambda = \frac{1}{5} = 0{,}2$. Hustota pravděpodobnosti tedy je:

$$f(x) = \begin{cases} 0{,}2e^{-0{,}2x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

2. Příslušná distribuční funkce $F(x)$ je tvaru:

$$F(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-0{,}2x}, & x \geq 0. \end{cases}$$

Pravděpodobnost, že budeme čekat déle než 12 minut, je:

$$P(X > 12) = 1 - F(12) = 1 - \left(1 - e^{-0{,}2 \cdot 12}\right) = e^{-2{,}4} \approx 0{,}0907.$$

3. Hledáme kvantil, tedy dobu čekání t, při které bude zákazník obsloužen s pravděpodobností 0,9:

$$P(X \leq t) = 1 - e^{-0{,}2t} = 0{,}9.$$

Z toho plyne:

$$e^{-0{,}2t} = 0{,}1 \quad \Rightarrow \quad -0{,}2t = \ln(0{,}1) \quad \Rightarrow \quad t = \frac{\ln(0{,}1)}{-0{,}2} \approx 11{,}51.$$

Zákazník bude obsloužen s pravděpodobností 90 % do zhruba 11,5 minuty.

□

Příklad 5.6 (Normální rozdělení $N(\mu; \sigma^2)$)**.** Jaká je pravděpodobnost, že náhodná veličina X, která má rozdělení $N(10; 9)$, nabude hodnoty:

1. menší než 16,
2. větší než 10,
3. v mezích od 7 do 22?

Řešení: V našem případě je $\mu = 10$ a rozptyl $\sigma^2 = 9$. Směrodatná odchylka je tedy $\sigma = \sqrt{9} = 3$. Výpočty provedeme s využitím distribuční funkce $F(x) = P(X \leq x)$, jejíž hodnoty snadno získáme v Excelu pomocí funkce `NORM.DIST`.

1. Pravděpodobnost, že $X < 16$, odpovídá přímo hodnotě distribuční funkce:

$$P(X < 16) = F(16) \approx \texttt{NORM.DIST(16; 10; 3; PRAVDA)} \approx 0{,}9772.$$

2. Pravděpodobnost, že $X > 10$, spočítáme jako doplněk do jedničky. Vzhledem k tomu, že hodnota 10 je přesně střední hodnota (a rozdělení je symetrické), výsledek musí být polovina:

$$P(X > 10) = 1 - F(10) = 1 - \texttt{NORM.DIST(10; 10; 3; PRAVDA)} = 1 - 0{,}5 = 0{,}5.$$

3. Pravděpodobnost, že X nabude hodnoty v intervalu $7 < X < 22$, určíme jako rozdíl hodnot distribuční funkce v horní a dolní mezi intervalu:

$$P(7 < X < 22) = F(22) - F(7)$$

Pomocí Excelu:

$$\texttt{NORM.DIST(22; 10; 3; PRAVDA)} - \texttt{NORM.DIST(7; 10; 3; PRAVDA)} \approx 0{,}9999 - 0{,}1587 = 0{,}8412.$$

□

V této kapitole jsme se zabývali základními spojitými rozděleními pravděpodobnosti, která se hojně používají v praxi. Seznámili jsme se s jejich vlastnostmi, praktickým použitím a s metodami výpočtu pravděpodobností a charakteristik.

- **Rovnoměrné rozdělení** $U(a; b)$ – Tento typ rozdělení se používá tehdy, když má náhodná veličina stejnou pravděpodobnost výskytu v každém bodě intervalu $\langle a; b\rangle$. V této kapitole jsme si ukázali, jak vypočítat pravděpodobnosti a distribuční funkci rovnoměrně rozdělené náhodné veličiny a jaké jsou její základní charakteristiky (střední hodnota, rozptyl).
- **Exponenciální rozdělení** $Exp(\lambda)$ – Exponenciální rozdělení se používá při modelování času mezi událostmi v procesech, které se vyskytují s konstantní intenzitou. V praxi může jít například o dobu čekání na obsluhu. Zabývali jsme se výpočtem pravděpodobností, distribuční funkcí a určením časových intervalů, v nichž nastanou události s danou pravděpodobností (kvantily).
- **Normální rozdělení** $N(\mu; \sigma^2)$ – Toto rozdělení, často označované jako Gaussovo, je jedním z nejdůležitějších rozdělení vůbec. Modeluje mnohé reálné procesy a veličiny v přírodě i společnosti. V kapitole jsme si ukázali, jak pomocí normálního rozdělení odhadnout pravděpodobnosti pro různé intervaly hodnot, jak vypočítat hodnoty distribuční funkce a jak využít software (Excel) při výpočtech.

V této kapitole jsme se zaměřili také na aplikace těchto rozdělení ve formě řešených příkladů, které zahrnovaly výpočty pravděpodobností a interpretaci získaných výsledků. Naučili jsme se rozlišovat situace, kdy je vhodné použít jednotlivé typy spojitých rozdělení, a získali jsme praktické dovednosti pro jejich nasazení.

Kapitola poskytuje pevný základ pro pochopení spojitých náhodných veličin a jejich rozložení, což je klíčové pro analýzu a modelování reálných dat v nejrůznějších oblastech od ekonomie po strojové učení.

1. Jaké jsou hlavní rozdíly mezi spojitým a diskrétním rozdělením pravděpodobnosti? Uveďte příklady spojitých rozdělení.
2. Co je to distribuční funkce náhodné veličiny a jaký je její význam? Jaký tvar má distribuční funkce pro rovnoměrné rozdělení?
3. Vysvětlete, co rozumíme pod termínem hustota pravděpodobnosti. Jaká je hustota pravděpodobnosti pro exponenciální rozdělení?
4. Jaké jsou základní charakteristiky normálního rozdělení $N(\mu;\sigma^2)$? Proč je toto rozdělení tak důležité v teorii pravděpodobnosti a statistice?
5. Jaké jsou aplikace exponenciálního rozdělení v praxi? Vysvětlete, v jakých situacích je vhodné jej použít.
6. K čemu se používá rovnoměrné rozdělení? Jak se vypočítá střední hodnota a rozptyl rovnoměrně rozdělené náhodné veličiny?
7. Jaké vlastnosti musí mít data, aby bylo možné použít normální rozdělení pro jejich modelování a analýzu?
8. Jaké jsou klíčové rozdíly mezi pravděpodobnostní funkcí (u diskrétních veličin) a hustotou pravděpodobnosti (u spojitých veličin)? Jakou hodnotu pravděpodobnosti $P(X = x)$ má spojitá veličina v jednom konkrétním bodě?
9. Co rozumíme pod pojmem střední hodnota náhodné veličiny? Jak se liší střední hodnota mezi rovnoměrným, exponenciálním a normálním rozdělením?
10. Jaký je vztah mezi intenzitou λ v exponenciálním rozdělení a střední dobou čekání na událost?
11. Náhodná veličina X má normované normální rozdělení $N(0;1)$. Určete:
    a. $P(X < 2{,}31)$ [0,9896]
    b. $P(X < -1{,}1)$ [0,1357]
    c. $P(-0{,}41 < X < 2{,}92)$ [0,6573]
12. Váha v uhelných skladech váží s chybou, jejíž střední hodnota je 30 kg, přičemž váha v průměru ukazuje méně (tedy $\mu = -30$ kg). Náhodné chyby mají normální rozdělení pravděpodobnosti se směrodatnou odchylkou $\sigma = 100$ kg. Jaká je pravděpodobnost, že chyba zjištěné váhy nepřekročí v absolutní hodnotě 90 kg? [0,6106]
13. Uvažujme rovnoměrně rozdělenou náhodnou veličinu X na intervalu $\langle 2;10\rangle$. Vypočtěte:
    a. Střední hodnotu a rozptyl. [Střední hodnota: 6, Rozptyl: 5,33]
    b. $P(X > 7)$ [0,375]
14. Čas mezi událostmi je modelován exponenciálním rozdělením s intenzitou $\lambda = 0{,}5$. Jaká je pravděpodobnost, že čas mezi dvěma událostmi bude menší než 3 minuty? [0,7769]

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: https://homel.vsb.cz/~oti73/cdpast1/

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 6

Náhodný vektor

Po prostudování této kapitoly budete umět:

- určit hustotu pravděpodobnosti a distribuční funkci náhodného vektoru,
- vypočítat marginální funkce náhodného vektoru a charakteristiky náhodného vektoru – kovarianci a koeficient korelace.

Klíčová slova:

Náhodný vektor, hustota pravděpodobnosti, distribuční funkce, kovariance, koeficient korelace.

Náhled kapitoly

V této kapitole se zaměříme na pojem náhodného vektoru, což je rozšíření náhodné veličiny na případ dvou nebo více veličin současně. Probereme základní vlastnosti náhodného vektoru, společné a marginální rozdělení, a ukážeme si, jak lze analyzovat závislosti mezi jednotlivými složkami vektoru. Dále se budeme věnovat výpočtu číselných charakteristik, jako je střední hodnota, kovariance a koeficient korelace, a jejich významu při práci s vícerozměrnými daty.

Cíle kapitoly

Cílem je formálně pochopit, jak pracovat s více náhodnými veličinami současně a jakými nástroji lze měřit lineární závislost mezi nimi.

Časová náročnost

Pro zvládnutí této kapitoly doporučujeme věnovat přibližně 3 hodiny studiu teorie, výpočtu charakteristik náhodného vektoru a řešení praktických příkladů.

6.1 Dvourozměrný náhodný vektor

Náhodný vektor představuje rozšíření pojmu náhodné veličiny na případ dvou a více náhodných veličin současně. Popisuje pravděpodobnostní chování více veličin a umožňuje analyzovat jejich společnou distribuci a závislosti mezi nimi. V této kapitole se zaměříme na případ dvourozměrného náhodného vektoru.

Definice 6.1 (Náhodný vektor)**.** Náhodný vektor (X, Y) je uspořádaná dvojice náhodných veličin. Pro popis jeho pravděpodobnostní struktury se využívá společná pravděpodobnostní funkce $p(x, y)$ u diskrétních veličin nebo hustota pravděpodobnosti $f(x, y)$ u spojitých veličin.

Definice 6.2 (Společná pravděpodobnostní funkce a hustota pravděpodobnosti)**.** V případě diskrétních veličin je společná pravděpodobnostní funkce $p(x, y) = P(X = x, Y = y)$ definována jako pravděpodobnost, že $X = x$ a $Y = y$. U spojitých veličin je společná hustota pravděpodobnosti $f(x, y)$ definována tak, že pro pravděpodobnost výskytu v dané oblasti platí:

$$P(X \in \langle x_1; x_2\rangle, Y \in \langle y_1; y_2\rangle) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x, y)\,\mathrm{d}y\,\mathrm{d}x.$$

Definice 6.3 (Marginální rozdělení)**.** Marginální rozdělení popisuje pravděpodobnostní chování jednotlivých složek náhodného vektoru (jakoby izolovaně). U diskrétních veličin získáme marginální pravděpodobnosti $p_1(x)$ a $p_2(y)$ sečtením přes druhou proměnnou:

$$p_1(x) = \sum_y p(x,y), \quad p_2(y) = \sum_x p(x,y).$$

Pro spojité veličiny získáme marginální hustoty $f_1(x)$ a $f_2(y)$ integrací:

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y, \quad f_2(y) = \int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}x.$$

Definice 6.4 (Distribuční funkce)**.** Distribuční funkce náhodného vektoru $F(x,y)$ je definována jako:

$$F(x,y) = P(X \leq x, Y \leq y).$$

U spojitých veličin platí:

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,\mathrm{d}v\,\mathrm{d}u.$$

Definice 6.5 (Podmíněné rozdělení)**.** Podmíněná pravděpodobnost $p(x \mid y)$ je definována jako podíl společné a marginální pravděpodobnosti:

$$p(x \mid y) = \frac{p(x,y)}{p_2(y)} \quad \text{pro} \quad p_2(y) > 0.$$

Pro spojité veličiny je podmíněná hustota definována obdobně:

$$f(x \mid y) = \frac{f(x,y)}{f_2(y)} \quad \text{pro} \quad f_2(y) > 0.$$

Definice 6.6 (Číselné charakteristiky náhodného vektoru)**.** Mezi základní charakteristiky náhodného vektoru (X,Y) patří střední hodnota, rozptyl a kovariance:

$$E(X) = \sum_x x \cdot p_1(x) \quad \text{(diskrétní)} \quad \text{nebo} \quad E(X) = \int_{-\infty}^{\infty} x \cdot f_1(x)\,\mathrm{d}x \quad \text{(spojité)}.$$

Kovariance $Cov(X,Y)$ vyjadřuje míru společné variability obou veličin a počítá se jako:

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y).$$

Definice 6.7 (Koeficient korelace)**.** Koeficient korelace $\rho(X,Y)$ vyjadřuje míru lineární závislosti mezi veličinami X a Y. Je definován vztahem:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y},$$

kde σ_X a σ_Y jsou směrodatné odchylky veličin X a Y. Hodnota $\rho(X,Y)$ leží vždy v intervalu $\langle -1; 1\rangle$.

6.2 Řešené příklady

Příklad 6.8. Najděte konstantu c, tak aby funkce:

$$f(x,y) = \begin{cases} c\frac{x^2}{1+y^2}, & 2 \le x \le 3,\ 0 \le y \le 1, \\ 0, & \text{jinak}, \end{cases}$$

byla hustotou pravděpodobnosti nějakého spojitého náhodného vektoru (X,Y).

Řešení: Aby byla funkce $f(x,y)$ korektní hustotou pravděpodobnosti, musí být její integrál přes celý definiční obor roven 1:

$$\int_0^1 \int_2^3 c\frac{x^2}{1+y^2}\,\mathrm{d}x\,\mathrm{d}y = 1.$$

Integrál lze rozdělit na součin dvou jednorozměrných integrálů:

$$\int_0^1 \int_2^3 \frac{x^2}{1+y^2}\,\mathrm{d}x\,\mathrm{d}y = \left(\int_0^1 \frac{1}{1+y^2}\,\mathrm{d}y\right) \cdot \left(\int_2^3 x^2\,\mathrm{d}x\right).$$

Vnitřní integrál je:

$$\int_2^3 x^2\,\mathrm{d}x = \left[\frac{x^3}{3}\right]_2^3 = \frac{27}{3} - \frac{8}{3} = \frac{19}{3}.$$

Vnější integrál je:

$$\int_0^1 \frac{1}{1+y^2}\,\mathrm{d}y = [\arctan(y)]_0^1 = \arctan(1) - \arctan(0) = \frac{\pi}{4}.$$

Celková hodnota integrálu je tedy:

$$\int_0^1 \int_2^3 \frac{x^2}{1+y^2}\,\mathrm{d}x\,\mathrm{d}y = \frac{\pi}{4} \cdot \frac{19}{3} = \frac{19\pi}{12}.$$

Z podmínky pro hustotu vyplývá:

$$c \cdot \frac{19\pi}{12} = 1 \quad \Rightarrow \quad c = \frac{12}{19\pi}.$$

Hustota pravděpodobnosti má tedy výsledný tvar:

$$f(x,y) = \begin{cases} \frac{12}{19\pi} \cdot \frac{x^2}{1+y^2}, & 2 \le x \le 3,\ 0 \le y \le 1, \\ 0, & \text{jinak}. \end{cases}$$

□

Příklad 6.9. Studenti z jedné studijní skupiny byli na zkoušce z matematiky a fyziky s těmito výsledky (první hodnota v uspořádané dvojici označuje známku z matematiky, druhá z fyziky):

$$(1,1),(1,2),(1,3),(2,2),(2,3),(2,3),(3,2),(3,2),(3,3),(3,3),(3,3),(3,3),(3,3),$$
$$(3,4),(3,4),(4,3),(4,3),(4,4),(4,4),(4,4).$$

1. Vytvořte pravděpodobnostní tabulku náhodného vektoru (X, Y), kde X je výsledek z matematiky a Y z fyziky.
2. Určete marginální pravděpodobnostní funkce $p_1(x)$ a $p_2(y)$.
3. Určete hodnoty distribuční funkce $F(x, y)$.
4. Určete podmíněné pravděpodobnosti $p(x \mid y)$.

Řešení: Celkový počet studentů je $n = 20$.

1. **Pravděpodobnostní tabulka** pro náhodný vektor (X, Y):

$X\backslash Y$	1	2	3	4
1	0,05	0,05	0,05	0
2	0	0,05	0,10	0
3	0	0,10	0,25	0,10
4	0	0	0,10	0,15

   Čísla v buňkách vznikla jako relativní četnosti dvojic, např. dvojice $(3, 2)$ je v datech dvakrát, takže $p(3, 2) = \frac{2}{20} = 0{,}10$.

2. **Marginální pravděpodobnostní funkce** $p_1(x)$ a $p_2(y)$:

$X\backslash Y$	1	2	3	4	$p_1(x)$
1	0,05	0,05	0,05	0	0,15
2	0	0,05	0,10	0	0,15
3	0	0,10	0,25	0,10	0,45
4	0	0	0,10	0,15	0,25
$p_2(y)$	0,05	0,20	0,50	0,25	1,00

   Marginální pravděpodobnosti $p_1(x)$ jsou zkrátka řádkové součty a $p_2(y)$ sloupcové součty.

3. **Distribuční funkce** $F(x, y)$:

   Distribuční funkce $F(x, y)$ je součtem všech pravděpodobností pro pole vlevo nahoře od daného bodu ($X \leq x$ a $Y \leq y$). Příklad výpočtu pro $F(3, 3)$:

$$F(3,3) = P(X \leq 3, Y \leq 3) = 0{,}05 + \cdots + 0{,}25 = 0{,}65.$$

Tabulka hodnot distribuční funkce:

$X\backslash Y$	1	2	3	4
1	0,05	0,10	0,15	0,15
2	0,05	0,15	0,30	0,30
3	0,05	0,25	0,65	0,75
4	0,05	0,25	0,75	1,00

4. **Podmíněné pravděpodobnosti** $p(x \mid y)$:

Získáme je dělením hodnot uvnitř tabulky příslušnou marginální pravděpodobností sloupce $p_2(y)$:

$$p(x \mid y) = \frac{p(x,y)}{p_2(y)}.$$

Např. $p(3 \mid 3) = \frac{0,25}{0,50} = 0,50$. Tabulka podmíněných pravděpodobností:

$X\backslash Y$	1	2	3	4
1	1,00	0,25	0,10	0,00
2	0,00	0,25	0,20	0,00
3	0,00	0,50	0,50	0,40
4	0,00	0,00	0,20	0,60

□

Příklad 6.10. Určete číselné charakteristiky náhodného vektoru (X, Y), který je zadán tabulkou:

$Y\backslash X$	2	3	6
1	0,15	0,20	0,10
3	0,20	0,05	0,30

Řešení: Budeme postupovat krok za krokem:

1. Střední hodnota $E(X)$ **a** $E(Y)$**:** Vypočítají se jako vážený průměr s vahami odpovídajícími marginálním pravděpodobnostem:

$$E(X) = 2 \cdot (0,15 + 0,20) + 3 \cdot (0,20 + 0,05) + 6 \cdot (0,10 + 0,30) = 2 \cdot 0,35 + 3 \cdot 0,25 + 6 \cdot 0,40 = 3,85.$$

$$E(Y) = 1 \cdot (0,15 + 0,20 + 0,10) + 3 \cdot (0,20 + 0,05 + 0,30) = 1 \cdot 0,45 + 3 \cdot 0,55 = 2,10.$$

2. Rozptyl $D(X)$ **a** $D(Y)$**:**

$$D(X) = \sum_i \sum_j (x_i - E(X))^2 \cdot p(x_i, y_j)$$

$$D(X) = (2 - 3,85)^2 \cdot 0,35 + (3 - 3,85)^2 \cdot 0,25 + (6 - 3,85)^2 \cdot 0,40 =$$
$$= (-1,85)^2 \cdot 0,35 + (-0,85)^2 \cdot 0,25 + 2,15^2 \cdot 0,40 = 1,197875 + 0,180625 + 1,849 = 3,2275.$$

$$D(Y) = (1-2{,}10)^2 \cdot 0{,}45 + (3-2{,}10)^2 \cdot 0{,}55 = (-1{,}10)^2 \cdot 0{,}45 + 0{,}90^2 \cdot 0{,}55 = 0{,}99.$$

3. Kovariance $Cov(X,Y)$:

$$Cov(X,Y) = \sum_i \sum_j (x_i - E(X)) \cdot (y_j - E(Y)) \cdot p(x_i, y_j).$$

Rychlejší výpočet je přes vztah $Cov(X,Y) = E(XY) - E(X)E(Y)$. Očekávaná hodnota součinu je:

$$E(XY) = (2\cdot 1\cdot 0{,}15) + (3\cdot 1\cdot 0{,}20) + (6\cdot 1\cdot 0{,}10) + (2\cdot 3\cdot 0{,}20) + (3\cdot 3\cdot 0{,}05) + (6\cdot 3\cdot 0{,}30) = 8{,}55.$$

Kovariance je tedy:

$$Cov(X,Y) = 8{,}55 - (3{,}85 \cdot 2{,}10) = 8{,}55 - 8{,}085 = 0{,}465.$$

4. Koeficient korelace $\rho(X,Y)$:

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{D(X)\cdot D(Y)}} = \frac{0{,}465}{\sqrt{3{,}2275 \cdot 0{,}99}} \approx \frac{0{,}465}{1{,}7875} \approx 0{,}26.$$

Jedná se o poměrně slabou pozitivní lineární závislost mezi náhodnými veličinami X a Y. □

Příklad 6.11. Vypočtěte střední hodnotu náhodné veličiny X náhodného vektoru, který je určen hustotou pravděpodobnosti:

$$f(x,y) = \begin{cases} \frac{1}{2}\sin(x+y), & 0 \le x \le \frac{\pi}{2},\ 0 \le y \le \frac{\pi}{2}, \\ 0, & \text{jinak.} \end{cases}$$

Řešení: Střední hodnota složky X je dána dvojným integrálem:

$$E(X) = \int_0^{\frac{\pi}{2}} \int_0^{\frac{\pi}{2}} x \cdot \frac{1}{2}\sin(x+y)\,\mathrm{d}x\,\mathrm{d}y.$$

Nejprve integrujeme podle proměnné x (metodou per partes, kde $u = x$, $v' = \sin(x+y)$):

$$\int x\sin(x+y)\,\mathrm{d}x = -x\cos(x+y) + \sin(x+y).$$

Po dosazení mezí 0 a $\frac{\pi}{2}$ pro vnitřní integrál máme:

$$[-x\cos(x+y) + \sin(x+y)]_0^{\frac{\pi}{2}} = -\frac{\pi}{2}\cos\left(\frac{\pi}{2}+y\right) + \sin\left(\frac{\pi}{2}+y\right) - \sin(y).$$

S využitím goniometrických vzorců $\cos(\frac{\pi}{2}+y) = -\sin(y)$ a $\sin(\frac{\pi}{2}+y) = \cos(y)$ obdržíme vnitřní část jako:

$$\frac{\pi}{2}\sin(y) + \cos(y) - \sin(y).$$

Tento výsledek nyní integrujeme podle y na intervalu od 0 do $\frac{\pi}{2}$ a nezapomeneme vynásobit konstantou $\frac{1}{2}$ stojící před integrálem:

$$E(X) = \frac{1}{2}\int_0^{\frac{\pi}{2}} \left(\left(\frac{\pi}{2} - 1\right)\sin(y) + \cos(y)\right)\,\mathrm{d}y.$$

Základní integrály dají $[-\cos(y)]$ a $[\sin(y)]$, po dosazení mezí:

$$E(X) = \frac{1}{2}\left[\left(\frac{\pi}{2}-1\right)\cdot(0-(-1)) + (1-0)\right] = \frac{1}{2}\left(\frac{\pi}{2} - 1 + 1\right) = \frac{\pi}{4}.$$

□

V této kapitole jsme se seznámili s konceptem náhodného vektoru, který představuje rozšíření pojmu náhodné veličiny na případ dvou a více náhodných veličin současně. Náhodný vektor popisuje pravděpodobnostní chování více veličin a umožňuje analyzovat jejich společnou distribuci a závislosti mezi nimi.

V této kapitole jsme rovněž řešili praktické příklady, ve kterých jsme aplikovali výše uvedené koncepty. Náhodný vektor je důležitým nástrojem při analýze dat, kde je třeba zkoumat více proměnných současně a jejich vzájemné vztahy. Tato kapitola poskytuje základní porozumění tomu, jak tyto závislosti modelovat a analyzovat.

- **Definice náhodného vektoru:** Náhodný vektor (X, Y) je uspořádaná dvojice náhodných veličin. Pro popis jeho pravděpodobnostní struktury se využívá společná pravděpodobnostní funkce (u diskrétních veličin) nebo hustota pravděpodobnosti (u spojitých veličin).
- **Společná pravděpodobnostní funkce a hustota pravděpodobnosti:** V případě diskrétních veličin (X, Y) je společná pravděpodobnostní funkce $p(x, y)$ definována jako pravděpodobnost, že $X = x$ a $Y = y$. U spojitých veličin je obdobně definována společná hustota pravděpodobnosti $f(x, y)$.
- **Marginální rozdělení:** Marginální rozdělení $p_1(x)$ a $p_2(y)$ popisuje pravděpodobnostní chování jednotlivých složek náhodného vektoru. Získává se součtem (u diskrétních veličin) nebo integrací (u spojitých veličin) přes všechny hodnoty druhé veličiny.
- **Distribuční funkce:** Distribuční funkce náhodného vektoru $F(x, y)$ je definována jako pravděpodobnost, že $X \leq x$ a $Y \leq y$.
- **Podmíněné rozdělení:** Podmíněné rozdělení $p(x \mid y)$ popisuje pravděpodobnost, že náhodná veličina X nabude hodnoty x, pokud je známo, že $Y = y$. Pro spojité veličiny se obdobně definuje podmíněná hustota pravděpodobnosti.
- **Číselné charakteristiky náhodného vektoru:** Mezi základní číselné charakteristiky patří střední hodnota, rozptyl, kovariance a koeficient korelace. Tyto charakteristiky umožňují popsat závislosti mezi složkami náhodného vektoru a míru jejich vzájemné závislosti.
- **Koeficient korelace:** Koeficient korelace $\rho(X, Y)$ udává míru lineární závislosti mezi veličinami X a Y. Hodnota ρ se pohybuje v intervalu $\langle -1; 1 \rangle$, kde hodnoty blízké 1 nebo -1 indikují silnou pozitivní, resp. negativní závislost, zatímco hodnoty blízké 0 indikují slabou nebo žádnou závislost.

1. Co je to dvourozměrný náhodný vektor a jak se liší od jednorozměrné náhodné veličiny?
2. Jak je definována společná pravděpodobnostní funkce dvou náhodných veličin X a Y?
3. Vysvětlete rozdíl mezi marginálním a podmíněným rozdělením náhodného vektoru.
4. Jak se vypočítá marginální rozdělení z dvourozměrného náhodného vektoru?
5. Jaká je definice kovariance a co vyjadřuje o závislosti mezi náhodnými veličinami X a Y?
6. Co vyjadřuje koeficient korelace a v jakém intervalu se jeho hodnota pohybuje?

7. Jaký je vztah mezi kovariancí a koeficientem korelace pro náhodný vektor (X, Y)?
8. Uveďte příklad praktického využití dvourozměrného náhodného vektoru v ekonomii nebo managementu.
9. Náhodný vektor (X, Y) má pravděpodobnostní funkci zadanou tabulkou:

$X \backslash Y$	1	2	3
-1	0,15	0,05	0,10
0	0,10	0,10	0,15
1	0,05	0,10	0,20

   Určete:

   a. $P(X = 0, Y = 3)$ [0,15]
   b. $P(X < 0{,}5, Y < 2{,}5)$ [0,40]
   c. $P(X > 0, Y > 2{,}5)$ [0,20]
   d. marginální rozdělení $P(X)$ $[P(X = -1) = 0{,}30, P(X = 0) = 0{,}35, P(X = 1) = 0{,}35]$
   e. marginální rozdělení $P(Y)$ $[P(Y = 1) = 0{,}30, P(Y = 2) = 0{,}25, P(Y = 3) = 0{,}45]$

10. Pro náhodný vektor daný následující tabulkou vypočtěte koeficient korelace:

$X \backslash Y$	1	0
1	0,05	0,01
0	0,02	0,92

    [Koeficient korelace $\rho(X, Y) \approx 0{,}7558$]

Literatura k tématu:

[1] ANDĚL, J. Statistické metody. 5. vyd. Praha: Matfyzpress, 2019. ISBN 978-80-7378-381-5.

[2] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[3] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[4] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[5] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 7

Statistický soubor s jedním argumentem

Po prostudování této kapitoly budete umět:

- určit základní popisné charakteristiky statistického souboru s jedním argumentem (viz klíčová slova),
- využít k těmto výpočtům statistický software (Excel).

Klíčová slova:

Základní soubor, statistická jednotka, četnosti, grafické znázornění četností, aritmetický průměr, modus, kvantily, medián, kvartily, decily, percentily, rozptyl, směrodatná odchylka.

Náhled kapitoly

V předchozích kapitolách jsme se věnovali spíše teoretickým modelům, zde se dostáváme k práci s daty. Tato kapitola se zaměřuje na základní popisné statistiky statistického souboru s jedním argumentem (s jednou proměnnou). Probereme různé druhy četností, jejich tabulkové a grafické znázorňování, dále různé míry polohy a variability dat. Prostě vše, co nám umožní mít ucelenější představu o rozložení dat. V následující kapitole tyto prostředky rozšíříme na dvourozměrný případ, kde nám k popisu jednotlivých proměnných přibude i jejich vzájemný vztah.

Cíle kapitoly

Cílem této kapitoly je získat základní potřebné dovednosti při práci s jednoduchými daty z pohledu popisné statistiky, tedy umět provádět potřebné výpočty a chápat jejich výsledky.

Časová náročnost

Pro tuto kapitolu doporučujeme vyčlenit přibližně 3 hodiny, které zahrnují jak studium teoretických částí, tak procvičování praktických příkladů a aplikací.

7.1 Základní pojmy a vlastnosti

Pravděpodobnost vs. statistika

Pravděpodobnost je matematický model reality. Jedná se o idealizovaný, abstraktní model, který pracuje s jednou nebo více náhodnými veličinami, jejichž rozdělení je známé. Z podstaty věci je tento model nepozorovatelný – představuje pouze naši abstrakci skutečnosti.

- Pravděpodobnost se zabývá náhodnými veličinami a jejich rozdělením.
- Jejím cílem je popsat, jak by se náhodné veličiny mohly chovat v určitém modelu.
- Pravděpodobnostní modely jsou používány v mnoha oblastech pro predikci nejistých jevů.

Statistika naopak vychází z pozorování (měření) hodnot konkrétních veličin. Statistika zkoumá jevy na rozsáhlém souboru dat a činí o nich závěry pomocí statistické indukce. Výsledky z malého vzorku jsou zobecňovány na rozsáhlejší populaci.

- Statistika používá odhady, protože žádný konečný výběr nemůže poskytovat úplnou informaci o rozdělení náhodných veličin v populaci.
- Statistika hledá pravidelnosti a souvislosti v datech a zobecňuje výsledky na širší soubor, než byl ten, ze kterého byly odvozeny.
- Vychází z reálných dat, na jejichž základě činí závěry o celkové populaci.

Příklady aplikací statistiky:

- Mají lidé, kteří pravidelně cvičí, lepší zdravotní ukazatele než ti, kteří necvičí?
- Je průměrná výše příjmů v určité oblasti závislá na vzdělání obyvatel?
- Jaká je pravděpodobnost, že nový produkt na trhu uspěje na základě výsledků z testovacího vzorku?

Data

Data představují klíčový prvek statistických analýz. Jedná se o pozorování, která provádíme za účelem zodpovězení položených otázek.

- **Matematicky**: data jsou realizací náhodné veličiny. Jedná se tedy o konkrétní hodnoty, které náhodná veličina může nabýt při experimentu nebo měření.
- **Datové tabulky**: Data jsou často organizována ve formě tabulek, kde řádky představují jednotlivá pozorování, zatímco sloupce odpovídají měřeným proměnným.
  - Řádky: Pozorování se týkají nezávislých subjektů náhodného výběru, jako jsou osoby, experimenty nebo jednotky sledování.
  - Sloupce: Každý sloupec odpovídá určité měřené veličině, například věk, pohlaví, výška, váha apod.
- **Software**: Pro správu a zpracování dat se používá řada softwarových nástrojů. Nejčastěji jsou využívány databázové systémy nebo tabulkové procesory, jako je Excel.
- **Statistický software**: K analýze dat slouží specializované statistické programy, jako jsou SAS, Statistica, SPSS, R nebo Python.

Ve statistice hraje správná organizace a správa dat zásadní roli, protože dobře strukturovaná data umožňují efektivnější analýzu a zajišťují správnost výsledků.

Popisná statistika

Popisná statistika představuje základní část statistické analýzy. Jejím cílem je sumarizovat a jednoduše popsat data, která máme k dispozici.

- **Pojmový aparát statistiky**: Zahrnuje základní statistické pojmy, jako jsou průměr, medián, rozptyl, směrodatná odchylka, kvartily a další.
- **Základní nástroj analýzy dat**: Pomocí popisných statistik můžeme rychle získat přehled o základních vlastnostech dat. Například průměr poskytuje informaci o střední hodnotě souboru, zatímco rozptyl nám řekne, jak jsou data rozložena kolem této hodnoty.
- **Prostředky pro prezentaci dat a výsledků**: Popisná statistika je často doprovázena vizuálními nástroji, jako jsou grafy, tabulky a diagramy, které umožňují efektivní prezentaci dat a usnadňují jejich interpretaci.

Příkladem aplikace popisné statistiky může být analýza průměrných platů v různých regionech, kde nás může zajímat nejen střední hodnota platu, ale také rozptyl a medián, abychom lépe porozuměli rozložení příjmů v dané populaci.

Základní pojmy ve statistice

Pro práci se statistickými daty je důležité nejprve pochopit několik základních pojmů:

Definice 7.1. Statistická jednotka je objekt, který chceme zkoumat. Může se jednat o osoby, domácnosti, firmy, organismy, obce, kraje atd. Každá statistická jednotka je nositelem určité vlastnosti, která nás zajímá a kterou zkoumáme.

Definice 7.2. Statistický soubor je množina statistických jednotek, které jsou předmětem našeho zkoumání:

- **Základní soubor**: Množina všech statistických jednotek, jejichž vlastnosti chceme poznat. Tento soubor zahrnuje veškeré objekty, které odpovídají naší studii, např. všechny domácnosti v určitém kraji.
- **Výběrový soubor**: Množina skutečně vyšetřovaných statistických jednotek, které jsou náhodně vybrány ze základního souboru. Tento výběr by měl být reprezentativní pro celou populaci.

Definice 7.3. Statistický znak je vlastnost, která je zjišťována na každé statistické jednotce. Tato vlastnost je v rámci statistiky považována za náhodnou veličinu. Mezi běžné statistické znaky patří např. pohlaví, věk, výška, hmotnost, počet dětí, barva očí, dopravní prostředek, počet úrazů, jméno.

Definice 7.4. Rozsah souboru (často označován jako n) představuje počet zkoumaných statistických jednotek v daném souboru.

Typy statistických znaků

Statistické znaky se dělí do několika kategorií podle svého charakteru:

- **Kvalitativní znaky** (někdy nazývané kategorické): Jedná se o slovní nebo kategoriální znaky, které nemohou být vyjádřeny numericky. Příkladem jsou pohlaví, barva očí nebo dopravní prostředek, který statistická jednotka používá.
- **Kvantitativní znaky** (číselné, numerické):
  - **Spojité znaky**: Mohou nabývat jakékoli hodnoty na určitých intervalech, např. výška, hmotnost nebo věk. Tyto znaky mohou být měřeny s libovolnou přesností.
  - **Diskrétní znaky**: Nabývají pouze určitých konkrétních hodnot, např. počet dětí nebo počet úrazů. Tyto znaky mají omezený počet možných hodnot.
- **Alternativní znaky**: Tyto znaky mohou nabývat pouze dvou hodnot, např. zda osoba kouří či nikoli, nebo zda byl test úspěšný či neúspěšný.
- **Množné znaky**: Jedná se o znaky, které mohou nabývat tří a více hodnot, např. dopravní prostředek (auto, kolo, autobus).

Jednorozměrný statistický soubor

V jednorozměrném statistickém souboru se zabýváme pouze jedním statistickým znakem X a jeho hodnotami v rámci výběrového souboru.

Označení:

- $\{\varepsilon_1, \ldots, \varepsilon_n\}$ výběrový soubor: Každá ε_i je statistická jednotka.
- X: statistický znak, který zkoumáme na každé statistické jednotce.
- x_j: hodnota znaku X na objektu ε_j, kde $j = 1, \ldots, n$.
- $(x_1, \ldots, x_n)$: datový soubor, který obsahuje hodnoty znaku X pro všechny jednotky.
- $(x_{(1)}, \ldots, x_{(n)})$: uspořádaný datový soubor, tj. $x_{(1)} \leq \cdots \leq x_{(n)}$.
- $(x_{[1]}, \ldots, x_{[r]})$: vektor variant znaku X, tj. různé hodnoty, které znak X nabývá, kde $x_{[i]} \neq x_{[j]}$ pro $i \neq j$.

Jednorozměrný statistický soubor nám umožňuje analyzovat hodnoty určitého znaku v rámci výběrového souboru a zjišťovat jejich rozložení.

7.2 Rozložení četností

Kde se s ním setkáme v praxi?
Rozložení četností je ten vůbec nejzákladnější nástroj pro práci s daty. Setkáme se s ním všude, kde potřebujeme z nepřehledné hromady surových dat získat rychlý přehled – ať už jde o přehled známek studentů z písemky, analýzu počtu prodaných kusů zboží v jednotlivých dnech, nebo rozdělení velikostí bot prodaných v e-shopu.

Rozložení četností slouží ke zpřehlednění datového souboru. Při této analýze sledujeme, kolikrát se jednotlivé hodnoty nebo intervaly hodnot vyskytují v našem výběrovém souboru.

- **Bodové rozložení četností**: Používá se pro diskrétní znaky s malým počtem variant, kdy četnost přiřazujeme jednotlivým variantám (hodnotám).
- **Intervalové rozložení četností**: Používá se pro diskrétní znaky s velkým počtem variant nebo pro spojité znaky, kdy četnost přiřazujeme třídícím intervalům.

Bodové rozložení četností

Bodové rozložení četností se vztahuje k jednotlivým hodnotám diskrétního znaku a zahrnuje následující typy četností:

Definice 7.5. (Absolutní) četnost varianty $x_{[j]}$: označována jako n_j, představuje počet výskytů hodnoty $x_{[j]}$ ve výběrovém souboru.

Definice 7.6. Relativní četnost varianty $x_{[j]}$: označována jako $p_j = \frac{n_j}{n}$, kde n je celkový počet pozorování. Relativní četnost můžeme chápat jako empirickou pravděpodobnost.

Definice 7.7. (Absolutní) kumulativní četnost prvních j variant: označována jako $N_j = n_1 + \cdots + n_j$, představuje součet četností prvních j variant.

Definice 7.8. Relativní kumulativní četnost prvních j variant: označována jako $F_j = \frac{N_j}{n} = p_1 + \cdots + p_j$, představuje kumulativní relativní četnost, což je suma relativních četností až po j-tou variantu.

Definice 7.9. Empirická distribuční funkce pro bodové rozložení četností je definována následovně:

$$F(x) = \begin{cases} 0 & \text{pro } x < x_{[1]} \\ F_j & \text{pro } x_{[j]} \leq x < x_{[j+1]}, \quad j = 1, \ldots, r-1 \\ 1 & \text{pro } x \geq x_{[r]} \end{cases}$$

Tato funkce zachycuje rozložení četností ve výběrovém souboru a zobrazuje kumulativní pravděpodobnost dosažení určité hodnoty.

Příklad 7.10 (Bodové rozložení četností)**.** Při zápočtu ze statistiky se studenti podrobili testu, ve kterém mohli získat 0 až 15 bodů. Výsledky testu jsou následující:

5, 10, 6, 7, 0, 2, 2, 4, 8, 10, 12, 15, 0, 0, 4, 2, 7, 10, 15, 0, 6, 5, 6, 9, 8, 7, 10, 12, 6, 0.

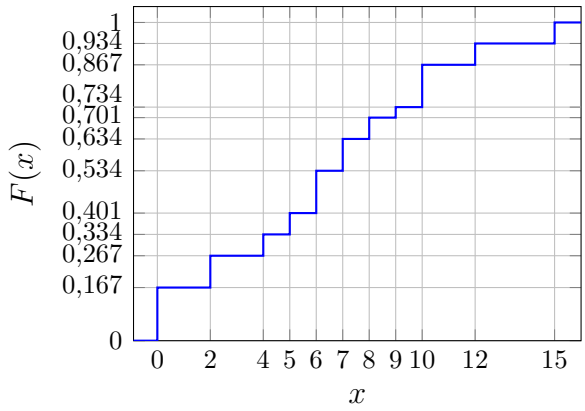
Vytvořte tabulku rozložení bodových četností (absolutních, relativních a kumulativních relativních) a nakreslete graf empirické distribuční funkce.

Řešení: Bodové rozložení četností je zobrazeno v tabulce 2 a graf empirické distribuční funkce na obrázku 13. □

Tento příklad ilustruje základní práci s bodovým rozložením četností, které umožňuje zjistit, kolik studentů dosáhlo určitého výsledku v testu a jak se tyto výsledky kumulují v rámci celého souboru.

Tab. 2: Bodové rozložení četností výsledků testu z příkladu 7.10

Body	n_j	p_j (%)	F_j (%)
0	5	16,7	16,7
2	3	10,0	26,7
4	2	6,7	33,4
5	2	6,7	40,1
6	4	13,3	53,4
7	3	10,0	63,4
8	2	6,7	70,1
9	1	3,3	73,4
10	4	13,3	86,7
12	2	6,7	93,4
15	2	6,7	100,0
Celkem	30	100,0	-



Obr. 13: Graf empirické distribuční funkce pro bodové rozložení četností z příkladu 7.10

Intervalové rozložení četností

Od bodového se liší tím, že na počátku celkový interval (rozsah) hodnot rozdělíme na menší podintervaly (rozsahy) a následně četnosti přiřazujeme celým těmto podintervalům. Po tomto kroku již vše funguje jako u bodových četností. Ukažme si to na následujícím příkladu.

Příklad 7.11 (Intervalové rozložení četností)**.** U 70 žen byla změřena hladina hemoglobinu s přesností 0,1 g/100 ml. Výsledky jsou následující:

10,2; 13,7; 10,4; 14,9; 11,5; 12,0; 11,0; 13,3; 12,9; 12,1; 9,4; 13,2; 10,8; 11,7; 10,5; 13,7; 11,8; 14,1; 10,3; 13,6; 12,1; 12,9; 11,4; 12,7; 10,6; 11,4; 11,9; 9,3; 13,3; 14,6; 11,2; 11,7; 10,9; 10,4; 12,0; 12,9; 11,1; 10,2; 11,6; 12,5; 13,4; 12,1; 9,7; 11,3; 10,9; 14,7; 10,8; 13,3; 11,9; 11,4; 12,5; 13,0; 11,6; 13,4; 12,3; 11,0; 14,6; 11,1; 13,5; 10,9; 13,1; 11,8; 12,2; 8,5; 10,1; 10,7; 11,3; 12,8; 13,9; 15,2.

Vytvořte tabulku rozložení intervalových četností (absolutních, relativních a kumulativních relativních).

Řešení: Intervalové rozložení četností je zobrazeno v tabulce 3. □

Tab. 3: Intervalové rozložení četností hladiny hemoglobinu u žen z příkladu 7.11

Hladina hemoglobinu v g/100 ml	n_j	p_j (%)	F_j (%)
8,0 – 8,9	1	1,4	1,4
9,0 – 9,9	3	4,3	5,7
10,0 – 10,9	14	20,0	25,7
11,0 – 11,9	19	27,1	52,9
12,0 – 12,9	14	20,0	72,9
13,0 – 13,9	13	18,6	91,4
14,0 – 14,9	5	7,1	98,6
15,0 – 15,9	1	1,4	100,0
Celkem	70	100,0	-

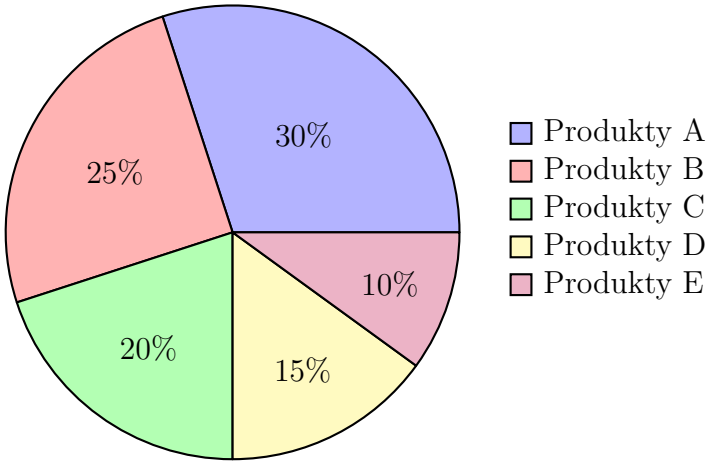
Tento příklad ilustruje základní práci s intervalovým rozložením četností, které nám umožňuje zjistit rozložení hodnot v rámci měřeného souboru a sledovat kumulativní četnosti pro jednotlivé intervaly.

7.2.1 Grafické znázornění četností

Znázorňujeme relativní a absolutní četnosti nebo relativní a absolutní kumulativní četnosti.

Koláčový graf

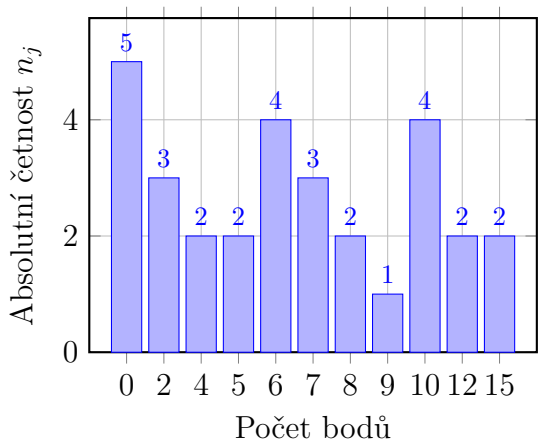
Koláčový graf se používá pro zobrazení absolutních i relativních četností, ale v obou případech vypadá stejně. Liší se jen popiskami (absolutními nebo relativními, ale mohou tam být i obě). Na obrázku 14 je příklad koláčového grafu, který zobrazuje rozložení prodeje různých kategorií produktů ve firmě.



Obr. 14: Koláčový graf rozložení prodeje produktů ve firmě

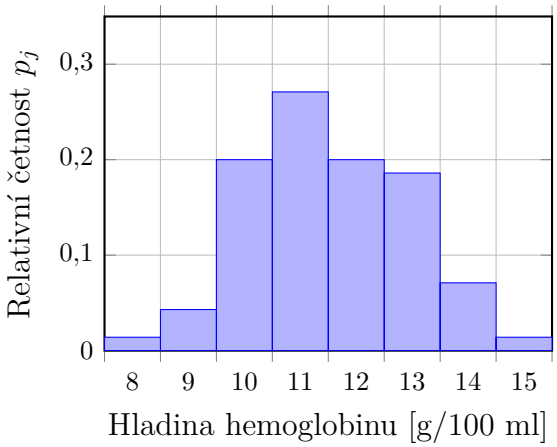
Histogram (sloupcový graf)

Histogram je sloupcový graf, který používáme pro znázornění rozložení četností. U bodového rozložení četností přiřadíme hodnotě $x_{[j]}$ obdélník, jehož výška je úměrná zjištěné četnosti. Na obrázku 15 je histogram výsledků testu ze statistiky z příkladu 7.10.



Obr. 15: Histogram absolutních četností výsledků testu ze statistiky z příkladu 7.10

Histogram pro hladinu hemoglobinu (v g/100 ml) z příkladu 7.11 je na obrázku 16. Každý sloupec pokrývá celý rozsah daného intervalu.



Obr. 16: Histogram relativních četností hladiny hemoglobinu z příkladu 7.11

7.3 Charakteristiky polohy a variability

Kde se s nimi setkáme v praxi?
Zatímco rozložení četností (tabulky a grafy) nám dává detailní pohled na celá data, v praxi často potřebujeme soubor popsat a porovnat jen pomocí několika málo čísel. Například při hodnocení platů ve firmě nás zajímá *průměrný* plat, nebo ještě lépe *medián* (typický plat očištěný o extrémně vysoké odměny managementu). Dále nás zajímá, jak moc se platy od tohoto středu liší – jsou všichni placeni zhruba stejně, nebo jsou mezi platy propastné rozdíly? K tomu právě slouží charakteristiky polohy (kde je střed dat) a variability (jak moc jsou data rozptýlená).

Charakteristiky polohy a variability jsou základními nástroji pro popis rozložení dat.

Mezi **charakteristiky polohy** patří například aritmetický průměr, medián, modus a výběrové kvantily. Tyto charakteristiky poskytují informace o střední hodnotě dat a jejich umístění na číselné ose.

Charakteristiky variability zahrnují mj. rozptyl, směrodatnou odchylku, rozpětí a interkvartilové rozpětí. Tyto charakteristiky popisují, jak jsou data rozptýlena kolem střední hodnoty. Společně tyto charakteristiky umožňují komplexní popis a analýzu statistických dat.

Míry polohy

Míry polohy, nebo také charakteristiky centrální tendence, popisují střední hodnotu dat a poskytují přehled o tom, kde se data nejvíce koncentrují. Mezi nejdůležitější charakteristiky patří:

- **Aritmetický průměr** ($\overline{x}$) – Nejběžnější charakteristika centrální tendence, která se počítá jako podíl součtu všech hodnot a jejich počtu:

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

kde n je celkový počet hodnot a x_i jsou jednotlivé hodnoty.

- **Medián** ($\tilde{x}$) – Střední hodnota uspořádaných dat. U lichého počtu hodnot je medián prostřední hodnota, u sudého počtu hodnot je medián průměr dvou prostředních hodnot. Medián je velmi vhodný pro data s odlehlými (extrémními) hodnotami, protože jimi není na rozdíl od průměru ovlivněn.
- **Modus** ($\hat{x}$) – Hodnota, která se v datech vyskytuje nejčastěji. V některých případech mohou data mít více než jeden modus, což se označuje jako vícemodální (multimodální) rozložení.
- **Harmonický průměr** ($\overline{x}_{\text{harm}}$) – Je vhodný pro průměrování poměrových veličin, jako je například výpočet průměrné rychlosti na stejně dlouhých úsecích:

$$\overline{x}_{\text{harm}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

- **Geometrický průměr** ($\overline{x}_{\text{geom}}$) – Používá se pro data, která se vztahují k růstu nebo procentním změnám (např. průměrný meziroční koeficient růstu):

$$\overline{x}_{\text{geom}} = \sqrt[n]{\prod_{i=1}^{n} x_i} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

- **Výběrové kvantily** – Hodnoty, které dělí uspořádaný datový soubor na daný počet stejně velkých částí. Kvantil na úrovni α (kde $\alpha \in (0;1)$) odděluje dolních $100\alpha\,\%$ hodnot od zbylých horních $100(1-\alpha)\,\%$. Nejčastěji používané kvantily jsou:
  - **První kvartil** (0,25-kvantil) – Hodnota, pod kterou leží 25 % dat.
  - **Medián** (0,50-kvantil) – Hodnota, pod kterou leží 50 % dat.
  - **Třetí kvartil** (0,75-kvantil) – Hodnota, pod kterou leží 75 % dat.

  Výběrové kvantily se obvykle určí z uspořádaných dat jako hodnoty, které odpovídají pozicím $P = \alpha(n+1)$, kde α je zvolená hladina kvantilu a n je počet pozorování. Pokud pozice není celé číslo, používá se k přesnému výpočtu lineární interpolace mezi dvěma sousedními hodnotami (je dobré vědět, že různé softwary jako Excel nebo R mohou používat mírně odlišné vzorce pro výpočet interpolace).

Aritmetický průměr

Pozorování $x_1, \ldots, x_n$ představují hodnoty znaku zjištěné na jednotlivých statistických jednotkách z nesetříděného datového souboru. Aritmetický průměr je základní mírou polohy, která se počítá jako součet všech pozorování dělený jejich počtem.

Definice 7.12. Aritmetický průměr (nesetříděného) souboru:

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Definice 7.13. Aritmetický průměr z rozložení četností (tzv. vážený průměr, kde vahami jsou absolutní četnosti):

$$\overline{x} = \frac{x_{[1]}n_1 + \cdots + x_{[r]}n_r}{n_1 + \cdots + n_r} = \frac{1}{n}\sum_{j=1}^{r} x_{[j]}n_j,$$

kde $x_{[j]}$ jsou jednotlivé varianty znaku a n_j jsou jejich absolutní četnosti (přičemž $\sum n_j = n$).

Definice 7.14. Vážený aritmetický průměr:
Pokud je soubor rozdělen do s dílčích skupin (podsložek), které mají své vlastní dílčí průměry $\overline{x}_i$ a rozsahy n_i, můžeme celkový průměr vypočítat takto:

$$\overline{x} = \frac{\sum_{i=1}^{s} \overline{x}_i n_i}{n_1 + \cdots + n_s} = \frac{1}{n}\sum_{i=1}^{s} \overline{x}_i n_i$$

Tento vzorec se používá například při výpočtech, kdy jednotlivé části souboru mají různé velikosti (váhy), které je třeba zohlednit při výpočtu celkového průměru.

Vhodné a nevhodné využití aritmetického průměru

Aritmetický průměr je velmi užitečná míra centrální tendence v situacích, kdy jsou data rovnoměrně rozložena a nejsou ovlivněna extrémními hodnotami.

Vhodné využití:

- Aritmetický průměr je vhodný pro soubory dat, které mají symetrické rozdělení (například normální rozdělení), protože průměr zde dobře reprezentuje skutečný střed dat.
- Používá se ve statistikách výkonu, výzkumu nebo finanční analýze, kde jsou hodnoty vyvážené a nemají extrémní odchylky.

Nevhodné využití:

- Aritmetický průměr je nevhodný pro soubory dat, které mají výrazně asymetrické (sešikmené) rozdělení nebo obsahují odlehlé (extrémní) hodnoty. V těchto případech může průměr znatelně zkreslovat představu o datech. Například u příjmů, kde několik málo osob má velmi vysoké příjmy, bude aritmetický průměr vyšší než příjem většiny populace.
- Průměr také nelze smysluplně použít v situacích, kde jsou data kvalitativní (mají nominální nebo ordinální povahu, například jména, pohlaví nebo úroveň vzdělání). Zde matematická operace sčítání postrádá smysl.

V těchto případech je vhodnější použít jiné míry polohy. U asymetrických nebo extrémními hodnotami zatížených dat volíme **medián**, který lépe vystihuje „typickou“ hodnotu. Pro nominální data je pak jedinou smysluplnou charakteristikou **modus**.

Výběrové kvantily

Definice 7.15. Mějme setříděný soubor, tedy hodnoty dat jsou uspořádané vzestupně: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, kde indexy označují pořadí hodnot v setříděném souboru.

Výběrový α-kvantil (kde $\alpha \in (0;1)$) je hodnota, která rozděluje seřazený datový soubor na dvě části tak, že:

- alespoň $100\alpha\,\%$ všech dat je menších nebo rovných $\tilde{x}_\alpha$,
- alespoň $100(1-\alpha)\,\%$ všech dat je větších nebo rovných $\tilde{x}_\alpha$.

Určení výběrového α-kvantilu z dat

Postup určení výběrového α-kvantilu závisí na tom, zda hodnota αn (kde n je celkový počet pozorování) je přirozené číslo, nebo nikoliv:

- Pokud je $\alpha n = c$, kde c je přirozené číslo, pak výběrový α-kvantil je průměr hodnot na pozicích $x_{(c)}$ a $x_{(c+1)}$:
$$\tilde{x}_\alpha = \frac{x_{(c)} + x_{(c+1)}}{2}.$$
- Pokud αn není přirozené číslo, zaokrouhlujeme αn nahoru na nejbližší vyšší přirozené číslo c a položíme:
$$\tilde{x}_\alpha = x_{(c)}.$$

Pozor na výpočet v softwaru (Excel)!
Výše uvedený postup je klasický "papírový"algoritmus založený na krokové funkci. Pokud ale k výpočtu použijete Excel (funkce `PERCENTIL.INC` nebo `KVARTIL.INC`), pravděpodobně dostanete mírně odlišné číslo. Excel totiž k výpočtu používá spojitou *lineární interpolaci* mezi hodnotami. Z didaktického hlediska jsou správně oba přístupy, u rozsáhlých datových souborů se jejich výsledky prakticky neliší.

Pojmenované kvantily

Některé z často používaných kvantilů mají svá specifická jména:

- **Medián** (0,50-kvantil) – Hodnota, která dělí data na dvě stejně velké části, tedy 50 % dat je menší nebo rovno této hodnotě a 50 % je větší nebo rovno.
- **Kvartily** – Speciální kvantily, které dělí data na čtyři stejné části:

  - **První kvartil** (0,25-kvantil) – Hodnota, pod kterou leží 25 % dat.
  - **Druhý kvartil** (0,50-kvantil) – Medián.
  - **Třetí kvartil** (0,75-kvantil) – Hodnota, pod kterou leží 75 % dat.
- **Decily** – Kvantily, které dělí data na deset stejných částí:
  - **První decil** (0,10-kvantil) – Hodnota, pod kterou leží 10 % dat.
  - **Druhý decil** (0,20-kvantil) – Hodnota, pod kterou leží 20 % dat, atd.
  - **Devátý decil** (0,90-kvantil) – Hodnota, pod kterou leží 90 % dat.
- **Percentily** – Kvantily, které dělí data na sto stejných částí:
  - **První percentil** (0,01-kvantil) – Hodnota, pod kterou leží 1 % dat.
  - **Pátý percentil** (0,05-kvantil) – Hodnota, pod kterou leží 5 % dat.
  - **Devadesátý pátý percentil** (0,95-kvantil) – Hodnota, pod kterou leží 95 % dat.

Medián jako speciální případ výběrového kvantilu

Medián je speciálním případem výběrového kvantilu pro $\alpha = 0{,}5$. Tento kvantil rozdělí data na dvě stejně velké části.

Případ lichého n:
Pro lichý počet pozorování n není hodnota $n \times 0{,}5$ přirozené číslo. Proto podle obecného postupu výpočtu kvantilu zaokrouhlíme $n \times 0{,}5$ nahoru na nejbližší celé číslo, což určí pořadí mediánu:

$$\tilde{x}_{0,5} = x_{\left(\frac{n+1}{2}\right)}.$$

Tento vzorec plyne z obecného pravidla zaokrouhlení kvantilu nahoru, kdy medián je hodnota na pozici $\frac{n+1}{2}$.

Příklad 7.16 (Výpočet mediánu pro lichý počet hodnot)**.** Mějme soubor o lichém počtu hodnot $n = 7$, seřazených jako $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(7)}$. Určete medián.

Řešení: Hodnota $n \times 0{,}5 = 7 \times 0{,}5 = 3{,}5$. Tuto hodnotu zaokrouhlíme nahoru na 4 (což odpovídá vzorci $\frac{7+1}{2} = 4$). Medián bude hodnota na čtvrté pozici, tedy $\tilde{x}_{0,5} = x_{(4)}$. □

Případ sudého n:
Pro sudý počet pozorování n je hodnota $n \times 0{,}5$ přirozené číslo. Proto medián, stejně jako obecný kvantil pro přirozené hodnoty $n \times \alpha$, bude průměrem dvou hodnot na pozicích:

$$\tilde{x}_{0,5} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}.$$

Příklad 7.17 (Výpočet mediánu pro sudý počet hodnot)**.** Mějme soubor o sudém počtu hodnot $n = 8$, seřazených jako $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(8)}$. Určete medián.

Řešení: Hodnota $n \times 0{,}5 = 8 \times 0{,}5 = 4$. Jedná se o celé číslo, takže medián je průměrem hodnot na 4. a 5. pozici:

$$\tilde{x}_{0,5} = \frac{x_{(4)} + x_{(5)}}{2}.$$

□

Tímto způsobem medián vyplývá jako speciální případ obecného výpočtu výběrového kvantilu, kde pro liché n postupujeme zaokrouhlením nahoru a pro sudé n použijeme průměr dvou středních hodnot:

Definice 7.18. Mějme uspořádaný datový soubor. Potom **medián** definujeme takto:

$$\tilde{x}_{0,5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{pro liché } n, \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{pro sudé } n. \end{cases}$$

Příklad 7.19 (Výpočet kvantilů – n sudé)**.** Ve výrobě se v posledním půl roce v jednotlivých měsících vyskytl následující počet úrazů: 1, 3, 2, 4, 2, 4. Určete medián, první (dolní) kvartil $\tilde{x}_{0,25}$ a třetí (horní) kvartil $\tilde{x}_{0,75}$ počtu úrazů za měsíc.

Řešení: Nejprve data uspořádáme vzestupně podle velikosti:

$$1, 2, 2, 3, 4, 4$$

Rozsah souboru je $n = 6$.

- **Medián:** Jelikož $\alpha n = 0{,}5 \cdot 6 = 3$ je celé číslo, medián je průměrem hodnot na 3. a 4. pozici:
$$\tilde{x}_{0,5} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{2+3}{2} = 2{,}5$$
- **První kvartil:** $\alpha n = 6 \cdot 0{,}25 = 1{,}5$. Výsledek není celé číslo, zaokrouhlujeme nahoru na 2. pozici: $\tilde{x}_{0,25} = x_{(2)} = 2$.
- **Třetí kvartil:** $\alpha n = 6 \cdot 0{,}75 = 4{,}5$. Zaokrouhlujeme nahoru na 5. pozici: $\tilde{x}_{0,75} = x_{(5)} = 4$.

□

Příklad 7.20 (Výpočet kvantilů – n liché)**.** Ve výrobě se v posledním půl roce v jednotlivých měsících vyskytl následující počet úrazů: 1, 3, 2, 4, 2, 4, 1. Určete medián, první a třetí kvartil počtu úrazů za měsíc.

Řešení: Data uspořádáme vzestupně:

$$1, 1, 2, 2, 3, 4, 4$$

Rozsah souboru je $n = 7$.

- **Medián:** $\alpha n = 7 \cdot 0{,}5 = 3{,}5 \Rightarrow$ zaokrouhlujeme nahoru na 4. pozici:
$$\tilde{x}_{0,5} = x_{(4)} = 2$$
- **První kvartil:** $\alpha n = 7 \cdot 0{,}25 = 1{,}75 \Rightarrow$ zaokrouhlujeme nahoru na 2. pozici: $\tilde{x}_{0,25} = x_{(2)} = 1$.

- **Třetí kvartil:** $\alpha n = 7 \cdot 0{,}75 = 5{,}25 \Rightarrow$ zaokrouhlujeme nahoru na 6. pozici: $\tilde{x}_{0{,}75} = x_{(6)} = 4$.

□

Příklad 7.21 (Kvantily z tabulky četností)**.** Uvažujme data x daná následující tabulkou rozložení četností:

$x_{[j]}$	1	2	3	4
n_j	10	12	6	3

Určete první decil $\tilde{x}_{0{,}1}$, první kvartil a třetí kvartil.

Řešení: Celkový rozsah souboru je $n = 10 + 12 + 6 + 3 = 31$. Pro snadnější určení pozic si můžeme představit (nebo vypsat pomocí kumulativních četností) pořadí hodnot:

- 1. až 10. hodnota jsou 1,
- 11. až 22. hodnota jsou 2,
- 23. až 28. hodnota jsou 3,
- 29. až 31. hodnota jsou 4.

Výpočet jednotlivých kvantilů:

- **První decil (0,10-kvantil):** $\alpha n = 31 \cdot 0{,}1 = 3{,}1 \Rightarrow$ hledáme $x_{(4)}$. Podle seznamu je $\tilde{x}_{0{,}1} = 1$.
- **První kvartil (0,25-kvantil):** $\alpha n = 31 \cdot 0{,}25 = 7{,}75 \Rightarrow$ hledáme $x_{(8)}$. Podle seznamu je $\tilde{x}_{0{,}25} = 1$.
- **Třetí kvartil (0,75-kvantil):** $\alpha n = 31 \cdot 0{,}75 = 23{,}25 \Rightarrow$ hledáme $x_{(24)}$. Tato pozice spadá do třetí skupiny, tedy $\tilde{x}_{0{,}75} = 3$.

□

Využití výběrových kvantilů

Výběrové kvantily mají široké využití v různých oborech statistiky a aplikovaných věd. Zde jsou uvedeny některé praktické příklady využití kvantilů:

- **Hladina cholesterolu v krvi**
  Jakou hladinu cholesterolu v krvi nepřekročí 90 % zdravé populace České republiky? Výběrový 0,90-kvantil (90. percentil) by zde představoval referenční hodnotu pro stanovení diagnostických limitů, která se běžně využívá v klinické praxi. Podobně jsou stanoveny referenční hodnoty pro další ukazatele krevního obrazu, například hladinu cukru, triglyceridů nebo krevní tlak.

- **Délka lišek**
  Jakou délku nepřekročí 95 % lišek? Zde můžeme využít výběrový 0,05-kvantil a 0,95-kvantil k určení rozmezí, ve kterém se nachází většina jedinců dané populace. Například pokud délka lišek spadá do rozmezí 58–90 cm, můžeme říci, že pouze 5 % lišek je delších než 90 cm a pouze 5 % lišek je kratších než 58 cm. Tyto kvantily pomáhají určit, které jedince považujeme za „typické“ a které za extrémní.

- **Stoletá voda**
  Jak definovat pojem stoletá voda? Výběrový 0,99-kvantil se často používá v hydrologii k definici stoleté vody. Jde o takovou výši maximálního ročního průtoku, která je v průměru překročena pouze jednou za sto let (tedy v 1 % případů). Tato hodnota je zásadní pro projektování protipovodňových opatření a infrastruktury.

- **Požadavky na kapitál pojišťoven**
  Jakou výši kapitálu musí pojišťovny v EU držet, aby minimalizovaly riziko úpadku? Regulace Solvency II vyžaduje, aby pojišťovny držely kapitál na úrovni odpovídající 0,995-kvantilu možných finančních ztrát v ročním horizontu. To znamená, že pojišťovna musí být schopna pokrýt rizika v 99,5 % případů a pouze v 0,5 % situací (extrémně nepříznivý vývoj) může dojít k ohrožení její solventnosti.

- **Testování (SCIO testy, srovnávací zkoušky)**
  Při hodnocení výsledků plošných testů se často využívá tzv. *percentil*. Pokud student dosáhne 75. percentilu (tedy 0,75-kvantilu), znamená to, že dopadl lépe než nebo stejně jako 75 % všech ostatních účastníků. Na základě těchto kvantilů mohou školy identifikovat 25 % nejúspěšnějších (nad horním kvartilem) nebo naopak studenty vyžadující zvýšenou podporu (pod dolním kvartilem).

- **Percentilové grafy v pediatrii**
  Kvantily jsou základem růstových grafů, které pediatři používají ke sledování vývoje dítěte (výška, váha, obvod hlavy). Pokud se křivka vývoje dítěte drží stabilně kolem určitého kvantilu, je jeho vývoj považován za přirozený, i když je dítě například drobnější než průměr.

Shrnutí

Výběrové kvantily jsou univerzálním nástrojem, který se využívá v mnoha oblastech – od medicíny a biologie přes hydrologii až po finance a školství. Pomáhají nám stanovit normy (referenční meze), identifikovat extrémní hodnoty a objektivně porovnávat jednotlivce s celou populací.

Míry variability

Míry absolutní variability

Míry absolutní variability popisují rozsah rozptýlenosti dat v původních jednotkách (např. v Kč, metrech apod.).

Definice 7.22.

- **Variační obor** $\langle x_{(1)}; x_{(n)} \rangle$ – Interval vymezený nejmenší a největší hodnotou.
- **Variační rozpětí** $R = x_{(n)} - x_{(1)}$ – Rozdíl mezi největší a nejmenší hodnotou.
- **Kvartilové rozpětí** $R_Q = \tilde{x}_{0,75} - \tilde{x}_{0,25}$ – Rozdíl mezi třetím a prvním kvartilem (šířka „krabice“ v boxplotu).
- **Kvartilová odchylka** $\frac{R_Q}{2}$ – Polovina kvartilového rozpětí.

Definice 7.23. Rozptyl $D(X)$ – (V literatuře také s^2). Pro neseskupená data jej počítáme jako průměrnou čtvercovou odchylku od aritmetického průměru (ve výběrové verzi dělíme $n-1$):

$$D(X) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$$

Pro data vyjádřená pomocí četností (seskupená) je rozptyl definován jako:

$$D(X) = \frac{1}{n-1} \sum_{j=1}^{r} (x_{[j]} - \overline{x})^2 n_j,$$

kde n_j je absolutní četnost varianty $x_{[j]}$.

Definice 7.24. Směrodatná odchylka – Druhá odmocnina z rozptylu. Udává, o kolik se hodnoty v průměru odchylují od aritmetického průměru v původních jednotkách:

$$s_x = \sqrt{D(X)}$$

Míry relativní variability

Míry relativní variability jsou bezrozměrná čísla. Používají se k porovnání variability mezi různými soubory, které mají odlišné jednotky nebo výrazně odlišné úrovně průměrů.

Definice 7.25. Variační koeficient:

$$V_x = \frac{s_x}{\overline{x}}$$

Obvykle se vyjadřuje v procentech ($V_x \cdot 100\,\%$). Pokud je $V_x > 0{,}5$ ($50\,\%$), považujeme soubor za silně rozptýlený (nehomogenní).

Relativní kvartilová odchylka:

$$Q_r = \frac{\tilde{x}_{0,75} - \tilde{x}_{0,25}}{\tilde{x}_{0,75} + \tilde{x}_{0,25}}$$

Příklad 7.26 (Porovnání variability platů)**.** Ve dvou firmách byly zkoumány měsíční platy zaměstnanců (v tisících Kč). Firma A: 25, 28, 30, 32, 35. Firma B: 20, 22, 24, 26, 80. Porovnejte variabilitu platů pomocí směrodatné odchylky a variačního koeficientu.

Řešení: 1. Aritmetické průměry:

$$\overline{x}_A = \frac{25+28+30+32+35}{5} = 30, \quad \overline{x}_B = \frac{20+22+24+26+80}{5} = 34{,}4$$

2. Rozptyly $D(X)$:

$$D(X)_A = \frac{(25-30)^2+(28-30)^2+(30-30)^2+(32-30)^2+(35-30)^2}{4} = \frac{25+4+0+4+25}{4} = 14{,}5$$

$$D(X)_B = \frac{(20-34{,}4)^2+(22-34{,}4)^2+(24-34{,}4)^2+(26-34{,}4)^2+(80-34{,}4)^2}{4} = \frac{2619{,}2}{4} = 654{,}8$$

3. Směrodatné odchylky:

$$s_A = \sqrt{14{,}5} \approx 3{,}81, \quad s_B = \sqrt{654{,}8} \approx 25{,}59$$

4. Variační koeficienty:

$$V_A = \frac{3{,}81}{30} \approx 0{,}127 \text{ (tj. } 12{,}7\,\%) \quad V_B = \frac{25{,}59}{34{,}4} \approx 0{,}744 \text{ (tj. } 74{,}4\,\%)$$

Závěr: Variabilita ve firmě A je nízká ($12{,}7\,\%$), platová struktura je zde vyrovnaná. Ve firmě B je variabilita extrémní ($74{,}4\,\%$), což je způsobeno odlehlou hodnotou 80 tisíc Kč. Směrodatná odchylka ve firmě B je téměř sedmkrát vyšší než ve firmě A. □

7.4 Míry tvaru rozdělení

Kromě charakteristik polohy a variability existují i charakteristiky, které popisují tvar rozdělení dat (zda je rozdělení symetrické, či nikoliv, a jak moc je „špičaté“).

Definice 7.27. Výběrová šikmost (skewness) měří asymetrii rozdělení dat:

$$\gamma_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right)^3$$

- $\gamma_1 = 0$: Rozdělení je symetrické (např. normální rozdělení).
- $\gamma_1 > 0$: Pozitivní šikmost (sešikmení doprava) – většina hodnot je nahlucena vlevo, vpravo je dlouhý „ocas“.
- $\gamma_1 < 0$: Negativní šikmost (sešikmení doleva) – většina hodnot je vpravo, vlevo je dlouhý „ocas“.

[Image of positive and negative skewness]

Definice 7.28. Výběrová špičatost (kurtosis, exces) měří „koncentraci“ dat kolem středu v porovnání s normálním rozdělením:

$$\gamma_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- $\gamma_2 = 0$: Špičatost odpovídá normálnímu rozdělení.
- $\gamma_2 > 0$: Špičatější rozdělení (více hodnot je blízko průměru a zároveň v extrémních koncích).
- $\gamma_2 < 0$: Plošší rozdělení (hodnoty jsou rozprostřeny rovnoměrněji).

Příklad 7.29 (Výpočet šikmosti a špičatosti)**.** V následující tabulce jsou uvedeny hodnoty datového souboru: 2, 3, 5, 7, 8, 10. Spočítejte výběrovou šikmost a špičatost tohoto souboru.

Řešení: 1. Aritmetický průměr a směrodatná odchylka:

$$\overline{x} = \frac{2+3+5+7+8+10}{6} = \frac{35}{6} \approx 5{,}833$$

$$s_x = \sqrt{\frac{1}{5}\sum (x_i - 5{,}833)^2} \approx \sqrt{\frac{46{,}833}{5}} \approx 3{,}061$$

2. Výběrová šikmost (γ_1): Dosadíme do vzorce pro $n = 6$:

$$\gamma_1 = \frac{6}{5 \cdot 4} \sum_{i=1}^{6} \left(\frac{x_i - 5{,}833}{3{,}061} \right)^3 \approx 0{,}3 \cdot (-1{,}312 - 0{,}944 - 0{,}028 + 0{,}163 + 0{,}376 + 1{,}084) \approx 0{,}198$$

Rozdělení má mírnou pozitivní šikmost.

3. Výběrová špičatost (γ_2): Dosadíme do vzorce pro $n = 6$:

$$\gamma_2 = \frac{6 \cdot 7}{5 \cdot 4 \cdot 3} \sum_{i=1}^{6} \left(\frac{x_i - 5{,}833}{3{,}061} \right)^4 - \frac{3 \cdot 25}{4 \cdot 3} \approx 0{,}7 \cdot 5{,}575 - 6{,}25 \approx -2{,}347$$

Rozdělení je výrazně plošší než normální rozdělení (což je u takto malého rovnoměrně rozloženého souboru očekávané). □

7.5 Řešené příklady

Příklad 7.30 (Četnosti statistického souboru)**.** Určete relativní, kumulativní a relativní kumulativní četnosti dat z tabulky:

$x_{[j]}$	**0**	**1**	**2**	**3**	**4**
n_j	7	44	56	30	12

Řešení: Nejprve vypočítáme celkový rozsah souboru n:

$$n = \sum_{j=1}^{5} n_j = 7 + 44 + 56 + 30 + 12 = 149.$$

Relativní četnosti $p_j = \frac{n_j}{n}$ se vypočítají jako podíl absolutní četnosti a celkového počtu prvků:

$x_{[j]}$	**0**	**1**	**2**	**3**	**4**	$\sum$
n_j	7	44	56	30	12	149
p_j	0,047	0,295	0,376	0,201	0,081	1,000

Nyní určíme kumulativní četnosti $N_j = \sum_{k=1}^{j} n_k$ a relativní kumulativní četnosti $F_j = \frac{N_j}{n}$:

$x_{[j]}$	**0**	**1**	**2**	**3**	**4**
N_j	7	51	107	137	149
F_j	0,047	0,342	0,718	0,919	1,000

□

Příklad 7.31 (Charakteristiky statistického souboru)**.** Vypočtěte modus, kvartily, aritmetický průměr, rozptyl, směrodatnou odchylku, šikmost a špičatost variační řady dané tabulkou:

$x_{[j]}$	**0**	**1**	**2**	**3**	**4**
n_j	7	44	51	30	12

Řešení: Celkový počet prvků je $n = 7 + 44 + 51 + 30 + 12 = 144$.

1. Modus: Hodnota s nejvyšší četností ($n_3 = 51$) je $\mathrm{Mo} = 2$.

2. Kvartily:

- **První kvartil** ($\alpha = 0{,}25$)**:** $\alpha n = 0{,}25 \cdot 144 = 36$. Pozice je celé číslo, bereme průměr 36. a 37. hodnoty. Obě leží v druhé skupině ($N_1 = 7, N_2 = 51$), tedy $\tilde{x}_{0{,}25} = 1$.
- **Medián** ($\alpha = 0{,}50$)**:** $\alpha n = 0{,}5 \cdot 144 = 72$. Průměr 72. a 73. hodnoty. Obě leží ve třetí skupině ($N_2 = 51, N_3 = 102$), tedy $\tilde{x}_{0{,}5} = 2$.
- **Třetí kvartil** ($\alpha = 0{,}75$)**:** $\alpha n = 0{,}75 \cdot 144 = 108$. Průměr 108. a 109. hodnoty. Obě leží ve čtvrté skupině ($N_3 = 102, N_4 = 132$), tedy $\tilde{x}_{0{,}75} = 3$.

3. Aritmetický průměr:

$$\overline{x} = \frac{0 \cdot 7 + 1 \cdot 44 + 2 \cdot 51 + 3 \cdot 30 + 4 \cdot 12}{144} = \frac{284}{144} \approx 1{,}972.$$

4. Rozptyl $D(X)$**:**

$$D(X) = \frac{\sum(x_{[j]} - \overline{x})^2 \cdot n_j}{n} \approx \frac{(0 - 1{,}972)^2 \cdot 7 + \cdots + (4 - 1{,}972)^2 \cdot 12}{144} \approx \frac{149{,}889}{144} \approx 1{,}041.$$

5. Směrodatná odchylka s_x**:**

$$s_x = \sqrt{D(X)} \approx \sqrt{1{,}041} \approx 1{,}020.$$

6. Šikmost γ_1**:**

$$\gamma_1 = \frac{\sum(x_{[j]} - \overline{x})^3 \cdot n_j}{n \cdot s_x^3} \approx 0{,}252.$$

Kladná hodnota naznačuje, že rozdělení je mírně sešikmené doprava.

7. Špičatost γ_2**:**

$$\gamma_2 = \frac{\sum(x_{[j]} - \overline{x})^4 \cdot n_j}{n \cdot s_x^4} - 3 \approx 2{,}446 - 3 \approx -0{,}554.$$

Záporná hodnota značí, že rozdělení je o něco plošší než normální rozdělení. □

Σ

V této kapitole jsme prozkoumali základní charakteristiky jednorozměrného statistického souboru. Zaměřili jsme se na popisné statistiky, které nám umožňují stručně a jasně popsat vlastnosti datového souboru.

- **Aritmetický průměr** ($\overline{x}$) popisuje „střední“ hodnotu v souboru, je však citlivý na extrémní hodnoty.
- **Medián** ($\tilde{x}_{0{,}5}$) rozděluje uspořádaný soubor na dvě stejně velké části a poskytuje dobrou představu o typické hodnotě i v asymetrických souborech.
- **Modus** ($\hat{x}$) je nejčastěji se vyskytující hodnota v souboru.
- **Rozptyl** $D(X)$ **a směrodatná odchylka** s_x jsou míry variability, které udávají, jak moc jsou hodnoty rozptýleny kolem průměru.
- **Šikmost** (γ_1) hodnotí asymetrii (sešikmení) rozložení, zatímco **špičatost** (γ_2) popisuje koncentraci dat kolem středu v porovnání s normálním rozdělením.

Ukázali jsme si, jak tyto charakteristiky vypočítat a interpretovat, což je klíčové pro správné pochopení dat v praktických aplikacích.

1. Co je to aritmetický průměr a jak se vypočítá? Uveďte rozdíl mezi prostým a váženým průměrem.
2. Jaký je rozdíl mezi mediánem a aritmetickým průměrem?
3. Kdy je didakticky i statisticky vhodnější použít k popisu středu dat medián místo průměru?
4. Co vyjadřuje rozptyl $D(X)$ a jaký má matematický vztah ke směrodatné odchylce s_x?
5. Jaký význam má šikmost a špičatost (exces) při analýze rozložení dat? Nakreslete, jak vypadá pozitivně zešikmené rozdělení.
6. Jak se vypočítá relativní četnost p_j a kumulativní relativní četnost F_j? Co vyjadřuje hodnota $F_j = 1$?
7. Co jsou to kvartily a jak se určí jejich pozice v datovém souboru o rozsahu n?
8. Určete medián a průměr měsíční spotřeby elektrické energie (kWh) v bytech z následujících údajů:
   169, 108, 26, 43, 114, 68, 35, 183, 103, 266, 74, 205, 62, 230, 85, 487, 120, 148, 91, 18, 58, 96, 295, 42, 137.
   [103; 130,52]
9. Zkoušky životnosti žárovek daly následující výsledky (v hodinách):
   606, 1249, 267, 44, 510, 340, 109, 1957, 463, 801, 1082, 169, 233, 1734, 1458, 80, 1023, 2736, 917, 459.
   Určete průměrnou dobu životnosti žárovek a jejich výběrový rozptyl.
   [811,85; 519 375,9]

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 8

Statistický soubor se dvěma argumenty

Po prostudování této kapitoly budete umět:

- určit základní charakteristiky dvourozměrného statistického souboru,
- vypočítat střední hodnotu, rozptyl a kovarianci pro dvourozměrný soubor,
- využít vhodné grafické nástroje pro vizualizaci dvourozměrných dat,
- interpretovat výsledky analýzy závislosti mezi dvěma znaky.

Klíčová slova:

Dvourozměrný soubor, aritmetický průměr, kovariance, rozptyl, směrodatná odchylka, kontingenční tabulka, bodový graf.

Tab. 4: Ukázka dvourozměrného statistického souboru

Statistická jednotka	Znak X (Výška v cm)	Znak Y (Hmotnost v kg)
1	170	65
2	165	70
3	180	80
4	175	75
5	160	60

Náhled kapitoly

Zde přímo navazujeme na předchozí kapitolu, její látku rozšíříme na případ dvou proměnných. Novinkou budou pojmy specifické pro tento dvojrozměrný případ, například kontingenční tabulky, bodové grafy a kovariance, které popisují vztahy dvojice proměnných. Pokročilejší metody, jako jsou regrese a korelace, si necháme až na další kapitoly.

Cíle kapitoly

Cílem této kapitoly je získat povědomí o rozdílu mezi jednorozměrným a dvojrozměrným případem a nachystat si pojem kovariance pro další kapitolu.

Časová náročnost

Pro tuto kapitolu doporučujeme vyčlenit přibližně 2 hodiny, které zahrnují jak studium teoretických částí, tak procvičování praktických příkladů a aplikací.

Úvod

Dvourozměrný statistický soubor se skládá z dvojic hodnot (argumentů), kde každý argument představuje hodnotu jiného statistického znaku měřeného na stejných statistických jednotkách. Tento typ souboru je používán k analýze vztahů mezi dvěma různými proměnnými, například výškou a hmotností osob, věkem a platem zaměstnanců, apod.

Každá statistická jednotka je tedy charakterizována dvojicí hodnot, které spolu mohou nebo nemusí být nějakým způsobem závislé. Dvourozměrný statistický soubor nám umožňuje analyzovat nejen vlastnosti jednotlivých znaků samostatně, ale i vztah mezi nimi.

Příklad dvourozměrného statistického souboru je v tabulce 4:

V tomto příkladu je znak X výška v centimetrech a znak Y hmotnost v kilogramech. Každý řádek představuje jednu statistickou jednotku (například jednu osobu), na které jsou měřeny oba znaky současně.

8.1 Základní pojmy

- **Statistická jednotka**: Objekt, na kterém jsou měřeny oba znaky. Může to být osoba, firma, stroj apod. Každá statistická jednotka má přiřazenou dvojici hodnot – jednu pro každý znak.
- **Znak** X: První proměnná, která je měřena na všech statistických jednotkách. Například výška osob nebo věk zaměstnanců.
- **Znak** Y: Druhá proměnná, která je rovněž měřena na stejných statistických jednotkách jako znak X. Například hmotnost osob nebo plat zaměstnanců.
- **Dvojice hodnot**: Každá statistická jednotka má přiřazenou dvojici hodnot (x_i, y_i), kde x_i je hodnota znaku X a y_i je hodnota znaku Y pro i-tou statistickou jednotku.
- **Statistický soubor**: Množina všech dvojic hodnot $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, kde n je počet statistických jednotek.
- **Rozsah souboru**: Počet statistických jednotek v souboru, označovaný jako n. V dvourozměrném souboru je rozsah stejný pro oba znaky, protože oba znaky jsou měřeny na stejných jednotkách.

Můžeme se vrátit k tabulce 4, kde jsou statistickými jednotkami jednotlivé osoby, znakem X je výška a znakem Y je hmotnost. Rozsah souboru $n = 5$.

8.2 Tabulkové a grafické zobrazení dvourozměrných dat

Při práci s dvourozměrným statistickým souborem je důležité umět data správně zobrazit. Existují různé způsoby, jak data vizualizovat a interpretovat. Mezi nejběžnější metody patří kontingenční tabulky a bodové grafy.

Kontingenční tabulky

Kontingenční tabulky se používají pro dvourozměrné soubory s diskrétními znaky. Tabulka obsahuje četnosti výskytu jednotlivých kombinací hodnot znaků X a Y. Tyto tabulky poskytují přehled o tom, jak často se různé kombinace hodnot vyskytují ve statistickém souboru.

- **Řádky tabulky** představují jednotlivé kategorie znaku X.
- **Sloupce tabulky** představují jednotlivé kategorie znaku Y.
- **Buňky tabulky** obsahují absolutní četnosti kombinací hodnot X a Y.

Tab. 5: Ukázka kontingenční tabulky

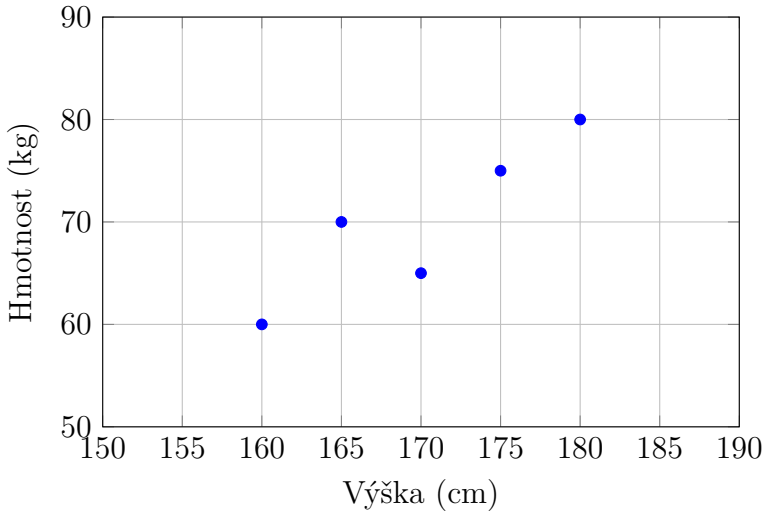
	Y_1	Y_2	Y_3
X_1	5	7	3
X_2	8	12	4
X_3	6	2	9

Příklad kontingenční tabulky je v tabulce 5, kde jsou zobrazeny četnosti kombinací hodnot X a Y. Například hodnota 5 znamená, že kombinace X_1 a Y_1 se vyskytuje pětkrát.

Kontingenční tabulky jsou užitečné pro analýzu závislosti mezi dvěma diskrétními znaky. Mohou být základem pro další metody analýzy, jako je například výpočet podmíněných pravděpodobností nebo chi-kvadrát test závislosti.

Bodové grafy

Bodové grafy (scatter plots) se používají pro dvourozměrné soubory, kde oba znaky nabývají spojitých hodnot. Na ose x je vynášen znak X a na ose y znak Y. Každá dvojice hodnot (x_i, y_i) se zobrazuje jako bod v rovině.



Obr. 17: Ukázka bodového grafu

Příklad bodového grafu je na obrázku 17. Každý bod v grafu představuje jednu statistickou jednotku a její hodnoty znaků X a Y. Například bod na souřadnicích (160, 60) odpovídá jednotce s výškou 160 cm a hmotností 60 kg.

Bodové grafy umožňují vizuálně analyzovat vztah mezi dvěma znaky. Pokud jsou body uspořádány podél určité linie nebo křivky, může to naznačovat nějaký druh závislosti mezi znaky X a Y. Tyto grafy jsou základním nástrojem pro identifikaci vzorů a závislostí v datech.

Interpretace grafických zobrazení

Grafická zobrazení nám pomáhají lépe pochopit vztah mezi dvěma znaky. V případě bodového grafu může například kladná korelace znamenat, že vyšší hodnoty znaku X jsou často doprovázeny vyššími hodnotami znaku Y. Naopak záporná korelace by znamenala, že vyšší hodnoty jednoho znaku jsou spojeny s nižšími hodnotami druhého.

Kontingenční tabulky nám umožňují odhalit závislosti mezi kategoriemi dvou znaků. Pokud se některé kombinace kategorií vyskytují mnohem častěji než jiné, může to naznačovat silnou závislost mezi znaky.

Tabulkové a grafické metody jsou důležité nástroje pro první krok analýzy dvourozměrných statistických souborů, protože poskytují vizuální a kvantitativní přehled o datech.

8.3 Míry polohy a variability pro dvourozměrný soubor

8.3.1 Míry polohy

Podobně jako u jednorozměrného statistického souboru, můžeme i u dvourozměrného souboru vypočítat míry polohy pro oba znaky X a Y. Tyto míry zahrnují aritmetický průměr, medián a modus.

Aritmetický průměr

Pro každý znak zvlášť můžeme vypočítat aritmetický průměr, který udává střední hodnotu daného znaku v souboru.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Zde $\overline{X}$ je průměrná hodnota znaku X a $\overline{Y}$ je průměrná hodnota znaku Y. Výpočty probíhají stejným způsobem jako v jednorozměrném souboru.

Příklad 8.1. Pro dvourozměrný statistický soubor z předchozího příkladu (výška a hmotnost osob) bychom vypočítali průměrnou výšku a hmotnost následovně:

$$\overline{x} = \frac{170 + 165 + 180 + 175 + 160}{5} = 170 \text{ cm},$$

$$\overline{y} = \frac{65 + 70 + 80 + 75 + 60}{5} = 70 \text{ kg}.$$

Podobným způsobem by se vypočítaly mediány a modus pro oba znaky. □

8.3.2 Míry variability a kovariance

Míry variability pro dvourozměrný statistický soubor jsou obdobné jako u jednorozměrného souboru, přičemž jsou vypočítávány zvlášť pro každý znak X a Y.

Rozptyl a směrodatná odchylka

Rozptyl a směrodatná odchylka se pro dvourozměrný soubor počítají obdobně jako v jednorozměrném případě, zvlášť pro každý znak:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2, \quad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Směrodatná odchylka je druhá odmocnina rozptylu:

$$s_X = \sqrt{s_X^2}, \quad s_Y = \sqrt{s_Y^2}.$$

Podrobnosti o rozptylu a směrodatné odchylce byly probrány v předchozí kapitole o jednorozměrném statistickém souboru.

Kovariance

Kovariance měří míru vzájemné závislosti mezi dvěma znaky X a Y. Je-li kovariance kladná, znamená to, že se vysoké hodnoty znaku X pojí s vysokými hodnotami znaku Y. Záporná kovariance naopak naznačuje, že vyšší hodnoty jednoho znaku se pojí s nižšími hodnotami druhého znaku.

Definice 8.2. Kovariance se vypočítá podle vzorce:

$$\mathrm{Cov}(X, Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Pokud jsou hodnoty X a Y nezávislé, je jejich kovariance blízká nule.

Příklad 8.3. Uvažujme opět dvourozměrný statistický soubor (výška a hmotnost osob) (tabulka 4). Vypočteme kovarianci.

Řešení: Nejprve vypočítáme průměry:

$$\overline{x} = 170, \quad \overline{y} = 70.$$

Poté vypočítáme kovarianci:

$$\mathrm{Cov}(X,Y) = \frac{1}{5-1}[(170-170)(65-70)+(165-170)(70-70)+\cdots+(160-170)(60-70)] = 50.$$

Tato kladná hodnota kovariance naznačuje, že mezi výškou a hmotností existuje pozitivní vztah — vyšší osoby mají obecně vyšší hmotnost. □

8.4 Řešené příklady

Příklad 8.4. Vypočítejte základní číselné charakteristiky dvourozměrného statistického souboru. Tabulka uvádí hodnoty X a Y pro jednotlivá pozorování:

$x\backslash y$	20	30	40	50	60	70	80
250	19	5					
350	23	116	11				
450	1	41	98	9			
550		4	32	65	7		
650		1	4	21	46	3	
750			1	2	11	13	1
850					1	3	2

Řešení: Pro řešení vypočítáme:

1. Průměry:

$$\overline{x} = \frac{1}{540} \cdot 259800 \approx 481{,}1, \quad \overline{y} = \frac{1}{540} \cdot 22030 \approx 40{,}80.$$

2. Rozptyly:

$$s_X^2 = \frac{1}{540} \cdot 134490000 - 481{,}1^2 \approx 17587{,}65, \quad s_Y^2 = \frac{1}{540} \cdot 989900 - 40{,}8^2 \approx 168{,}81.$$

3. Směrodatné odchylky:

$$s_X \approx 132{,}62, \quad s_Y \approx 12{,}99.$$

4. Kovariance:

$$\mathrm{Cov}(X,Y) = \frac{1}{540} \cdot 11427500 - 481{,}1 \cdot 40{,}8 \approx 1534{,}49.$$

□

Příklad 8.5. Vypočítejte číselné charakteristiky dvourozměrného statistického souboru, který je zadán tabulkou:

x	27	31	87	93	114	124	190	193	250	254	264	272	308	324
y	28	21	71	36	30	43	54	54	59	25	82	22	38	22

371	372	440	442	502	503	506	522	556	620	624
56	63	46	24	33	40	41	28	53	38	66

Řešení: Výpočty provedeme pomocí Excelu:

1. Průměry:
$$\overline{x} = \frac{7989}{25} \approx 319{,}56, \quad \overline{y} = \frac{1073}{25} \approx 42{,}92.$$

2. Rozptyly:
$$s_X^2 = \frac{3371599}{25} - 319{,}56^2 \approx 32745{,}37, \quad s_Y^2 = \frac{52945}{25} - 42{,}92^2 \approx 275{,}67.$$

3. Směrodatné odchylky:
$$s_X \approx 180{,}96, \quad s_Y \approx 16{,}60.$$

4. Kovariance:
$$\mathrm{Cov}(X, Y) = \frac{349250}{25} - 319{,}56 \cdot 42{,}92 \approx 254{,}48.$$

□

Σ

V této kapitole jsme se seznámili s dvourozměrným statistickým souborem, který analyzuje dvojice hodnot (x_i, y_i) pro každou statistickou jednotku. Pro oba znaky jsme vypočítali základní míry polohy (průměr, medián, modus) a variability (rozptyl, směrodatná odchylka).

Představili jsme kovarianci jako nástroj k měření závislosti mezi dvěma znaky, kde kladná kovariance ukazuje na pozitivní vztah a záporná na negativní.

Kromě výpočtů jsme se věnovali kontingenčním tabulkám pro diskrétní znaky a bodovým grafům pro spojité znaky, které umožňují vizuální analýzu vztahů mezi znaky.

Tato kapitola připravuje základ pro další analýzy závislostí mezi dvěma znaky, které budou následovat v příštích kapitolách.

8.5 Kontrolní otázky

1. Jaký je rozdíl mezi jednorozměrným a dvourozměrným statistickým souborem?
2. Jak vypočítáme aritmetický průměr pro dvourozměrný statistický soubor?
3. Co znamená kovariance a jaký má význam při analýze dvourozměrného souboru?
4. Jaká je interpretace kladné a záporné hodnoty kovariance?
5. Jaký grafický nástroj lze použít pro vizualizaci dvourozměrného statistického souboru, kde oba znaky jsou spojité?
6. Jak funguje kontingenční tabulka a kdy ji použijeme?
7. Jaký je vztah mezi rozptylem a směrodatnou odchylkou pro jednotlivé znaky v dvourozměrném statistickém souboru?
8. Proč používáme bodový graf (scatter plot) při analýze dvourozměrných dat a co nám ukazuje o závislosti mezi znaky X a Y?
9. U 130 zákrsků bylo zjištěno stáří stromu v letech (argument X) a sklizeň v jistém roce v kg (argument Y). Podle údajů v tabulce určete kovarianci.

$X\backslash Y$	4	5	6	7	8	9	10	11
3	6	0	0	0	0	0	0	0
4	0	5	10	2	0	0	0	0
5	0	0	0	2	8	3	0	0
6	0	0	0	0	0	12	10	0
7	0	0	0	0	0	8	15	4
8	0	0	0	0	4	16	8	0
9	0	3	12	2	0	0	0	0

[$\mathrm{Cov}(X,Y) \approx 1{,}12$]

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 9

Regresní a korelační analýza

Po prostudování této kapitoly budete umět:

- vysvětlit, co korelační koeficient popisuje a jaké jsou jeho varianty,
- vypočítat Pearsonův korelační koeficient na základě zadaných dat.
- interpretovat výsledky korelační analýzy,
- používat Excel nebo jiný statistický software k výpočtu korelačních koeficientů,
- odhadovat parametry lineárního regresního modelu,
- aplikovat lineární regresi na reálná data,
- používat Excel a modul `Analýza dat - Regrese` pro výpočty.

Klíčová slova:

Korelační koeficient, statistická závislost, lineární vztah, ineární regrese, regresní analýza, regresní koeficienty, Excel, modul Analýza dat.

Náhled kapitoly

V této kapitole navážeme na předchozí kapitolu, kde jsme zkoumali vztah dvou statistických znaků. Zde se seznámíme s dvěma pokročilejšími metodami analýzy těchto závislostí.

Korelační analýza slouží k měření síly a směru lineárního vztahu mezi dvěma proměnnými. Probereme různé varianty korelačních koeficientů a jejich využití v praxi, zejména Pearsonův korelační koeficient, který je nejčastěji používán. Ukážeme si také omezení tohoto koeficientu a situace, kdy je vhodné použít alternativní metody.

Metoda **lineární regrese** umožňuje odhadnout vztah mezi závislou a nezávislou proměnnou pomocí přímky (případně i jiné křivky).

Obě metody se naučíme provádět i v Excelu.

Cíle kapitoly

Cílem této kapitoly je praktické seznámení s dvěma metodami, korelační a regresní analýzou, které nám umožňují studovat vztah (závislost) dvou statistických znaků.

Odhad času potřebného ke studiu

Odhaduje se, že studium této kapitoly zabere přibližně 3 hodiny. Tento čas zahrnuje čtení textu, pochopení teoretických konceptů a řešení příkladů (i v Excelu).

9.1 Princip korelační analýzy

Co je to korelační koeficient?

Korelační koeficient je statistická míra, která určuje sílu a směr vztahu mezi dvěma proměnnými. Pearsonův korelační koeficient, označovaný jako r, měří lineární vztah mezi dvěma spojitými proměnnými a nabývá hodnot mezi -1 a 1. Pokud je $r = 1$, jedná se o perfektní pozitivní lineární vztah, pokud $r = -1$, jedná se o perfektní negativní lineární vztah, a pokud $r = 0$, neexistuje žádná lineární závislost mezi proměnnými.

Výpočet korelačního koeficientu

Definice 9.1. Pearsonův korelační koeficient je definován vztahem:

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y} = \frac{\sum(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \cdot \sum(y_i - \overline{y})^2}}$$

kde x_i a y_i jsou jednotlivé hodnoty obou proměnných, a $\overline{x}$ a $\overline{y}$ jsou jejich průměry.

Řešené příklady

Příklad 9.2. Mějme data o prodejích produktů ve dvou různých regionech. Vypočítejte Pearsonův korelační koeficient a určete, zda mezi těmito proměnnými existuje lineární vztah.

Prodeje (Region 1)	10	15	20	25	30
Prodeje (Region 2)	12	18	25	24	28

Řešení: Nejprve vypočítáme průměry $\overline{x} = 20$ a $\overline{y} = 21.4$. Poté provedeme výpočet Pearsonova korelačního koeficientu podle výše uvedeného vzorce. Korelační koeficient $r \approx 0.88$, což ukazuje na silnou pozitivní lineární závislost mezi prodeji v obou regionech.

Excel: Korelační koeficient lze spočítat pomocí funkce `CORREL(array1, array2)` v Excelu.

Příklad 9.3. Mějme data o počtu zákazníků navštěvujících obchod a průměrné denní tržby. Vypočítejte korelační koeficient a určete, zda existuje lineární závislost.

Počet zákazníků	50	60	70	80	90
Denní tržby (v tis. Kč)	20	25	30	28	35

Řešení: Vypočítáme průměry $\overline{x} = 70$ a $\overline{y} = 27.6$. Pomocí vzorce pro korelační koeficient získáme $r \approx 0.91$, což značí velmi silnou pozitivní lineární závislost mezi počtem zákazníků a tržbami.

Excel: Pomocí funkce `CORREL(array1, array2)` lze získat stejný výsledek. □

Příklad 9.4. Zde jsou data pro prodej dvou produktů v různých týdnech. Určete, zda mezi prodejem těchto produktů existuje lineární vztah.

Prodeje produktu A	100	105	110	95	115	90	120	85	125	80
Prodeje produktu B	200	180	205	185	190	185	190	195	200	190

Řešení: Průměry pro produkt A a produkt B jsou $\overline{x} = 102.5$ a $\overline{y} = 192$. Po výpočtu korelačního koeficientu dostaneme $r \approx 0.08$, což naznačuje velmi slabou nebo žádnou lineární závislost mezi prodeji těchto produktů.

Excel: Výpočet pomocí `CORREL(array1, array2)` v Excelu také ukazuje, že korelace je blízká nule, tedy nevýznamná. □

Historie a varianty korelačních koeficientů

Historie korelačních koeficientů sahá až do 19. století, kdy Francis Galton poprvé navrhl metody pro kvantifikaci statistických vztahů mezi proměnnými. Na jeho práci navázal Karl Pearson, který formalizoval a popularizoval Pearsonův korelační koeficient.

V průběhu času byly vyvinuty další varianty korelačních koeficientů pro specifické účely:

- Spearmanův korelační koeficient (Spearman's rho): Používá se, pokud data nejsou normálně rozložena nebo vykazují monotónní, nikoli lineární vztah.
- Kendallův tau: Měří sílu vztahu mezi pořadím hodnot a používá se zejména u malých souborů dat.
- Point-biserial correlation: Využívá se pro měření korelace mezi spojitou a binární proměnnou.

Každý z těchto korelačních koeficientů má své specifické aplikace a závisí na typu dat, které jsou analyzovány. Korelační analýza našla využití v mnoha oblastech, včetně psychologie, ekonomie, marketingu a biostatistiky.

Kdy je korelační koeficient vhodný?

Korelační koeficient popisuje sílu a směr lineárního vztahu mezi dvěma spojitými proměnnými. Jeho použití je vhodné, pokud jsou splněny následující podmínky:

- Obě proměnné mají přibližně normální rozložení.
- Vztah mezi proměnnými je lineární.
- Nejsou přítomny výrazné odlehlé hodnoty, které by ovlivnily výsledek.

Použití Pearsonova korelačního koeficientu je nevhodné, pokud vztah mezi proměnnými není lineární nebo pokud se jedná o ordinální data, u nichž je vhodnější použít Spearmanův korelační koeficient nebo Kendallův tau.

Praktické cvičení

Mějte následující data pro dva produkty a určete, zda existuje lineární závislost mezi jejich prodeji:

Prodeje produktu A	5	10	15	20	25
Prodeje produktu B	8	12	17	22	24

Spočítejte korelační koeficient pomocí výše uvedeného vzorce nebo pomocí Excelu (`CORREL(array1, array2)`). Na základě výsledku určete, zda mezi těmito proměnnými existuje lineární závislost.

9.2 Princip lineární regrese

Úvodní příklad

Představte si, že jste ekonomický analytik ve společnosti, která chce předpovědět tržby na základě výdajů na reklamu. Máte k dispozici následující data z posledních 10 měsíců (tabulka 6).

Tab. 6: Ukázková data pro lineární regresi

Měsíc	1	2	3	4	5	6	7	8	9	10
Reklama (tis. Kč)	20	25	30	35	40	45	50	55	60	65
Tržby (tis. Kč)	200	220	250	280	310	330	360	390	420	450

Cílem je zjistit, jak silný je vztah mezi výdaji na reklamu a tržbami, a vytvořit model, který umožní předpovědět tržby při různých úrovních výdajů na reklamu.

Formulace problému

- **Závislá proměnná** (Y): Tržby (tis. Kč).
- **Nezávislá proměnná** (X): Výdaje na reklamu (tis. Kč).

Cíl analýzy

Pomocí lineární regrese odhadnout vztah mezi výdaji na reklamu a tržbami a posoudit, zda je tento vztah statisticky významný.

Co je to lineární regrese?

Lineární regrese je statistická metoda používaná k modelování vztahu mezi závislou proměnnou a jednou nebo více nezávislými proměnnými. V případě jednoduché lineární regrese se jedná o vztah mezi dvěma proměnnými, který je modelován pomocí přímky.

Regresní model

Lineární regresní model lze vyjádřit rovnicí:

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

kde:

- Y je závislá proměnná,
- X je nezávislá proměnná,
- β_0 je absolutní člen (intercept),
- β_1 je směrnice přímky (sklon),
- ε je náhodná chyba (reziduální složka).

Metoda nejmenších čtverců

Parametry β_0 a β_1 jsou odhadnuty pomocí *metody nejmenších čtverců*, která minimalizuje součet čtverců odchylek mezi skutečnými hodnotami Y a predikovanými hodnotami $\hat{Y}$:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Odhady parametrů

Odhady parametrů $\hat{\beta}_0$ a $\hat{\beta}_1$ lze vypočítat pomocí vzorců:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2},$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x},$$

kde $\overline{x}$ a $\overline{y}$ jsou průměry X a Y.

Předpoklady lineární regrese

Aby byly odhady parametrů platné, musí být splněny následující předpoklady:

- **Linearita**: Vztah mezi X a Y je lineární.
- **Homoskedasticita**: Rozptyl náhodné složky ε je konstantní pro všechna X.
- **Nezávislost**: Hodnoty náhodné složky ε jsou nezávislé.
- **Normalita**: Náhodná složka ε je normálně rozložena.

Historické poznámky

Metoda lineární regrese byla poprvé formálně představena anglickým statistikem **Sir Francis Galtonem** v 19. století při studiu dědičnosti výšky mezi rodiči a dětmi. Termín *regrese* pochází z Galtonova pozorování, že extrémní hodnoty mají tendenci "regresovat" k průměru v následující generaci.

Později **Karl Pearson** a **Ronald A. Fisher** rozvinuli matematické základy regresní analýzy a metodu nejmenších čtverců, která je dnes standardním nástrojem v statistice a ekonometrice.

Odhad parametrů a interpretace

Výpočet odhadů

Pomocí výše uvedených vzorců lze spočítat odhady $\hat{\beta}_0$ a $\hat{\beta}_1$ na základě dostupných dat.

Interpretace parametrů

- **Směrnice přímky** ($\hat{\beta}_1$): Udává změnu v závislé proměnné Y při jednotkové změně nezávislé proměnné X.
- **Absolutní člen** ($\hat{\beta}_0$): Hodnota závislé proměnné Y, když nezávislá proměnná X je nulová.

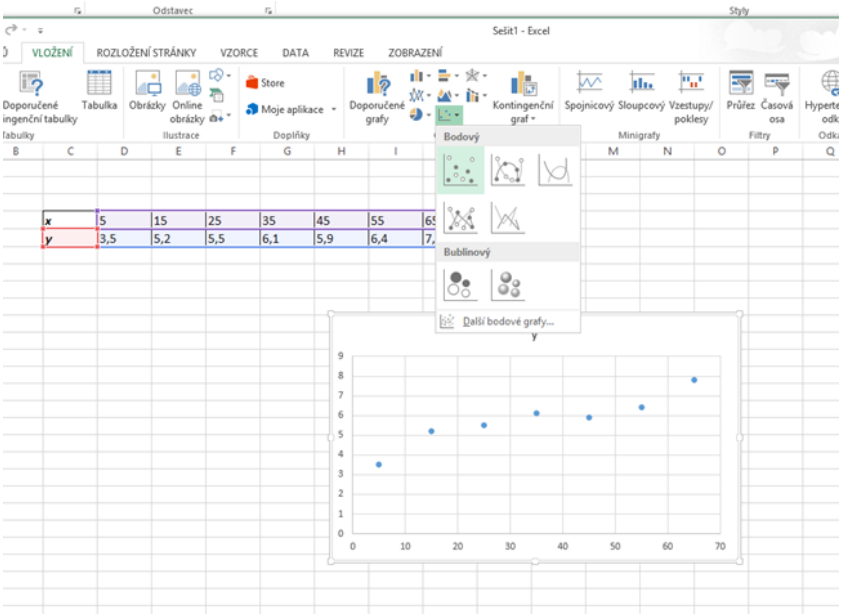
9.3 Řešené příklady

Příklad 9.5. Vyrovnejte data v tabulce regresní přímkou:

x	5	15	25	35	45	55	65
y	3,5	5,2	5,5	6,1	5,9	6,4	7,8

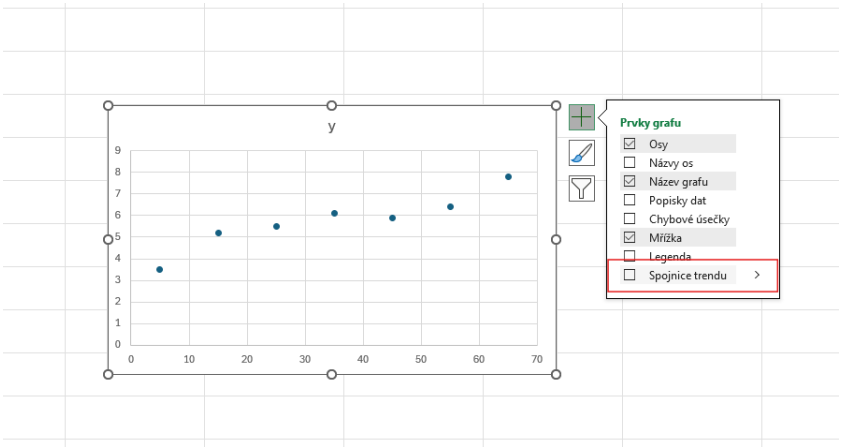
Řešení: Ukážeme, jak by se tato úloha řešila v Excelu:

1. Nejdříve označíme data a klikneme na *Vložit Graf*, přičemž vybereme typ grafu *XY bodový* (obrázek 18).



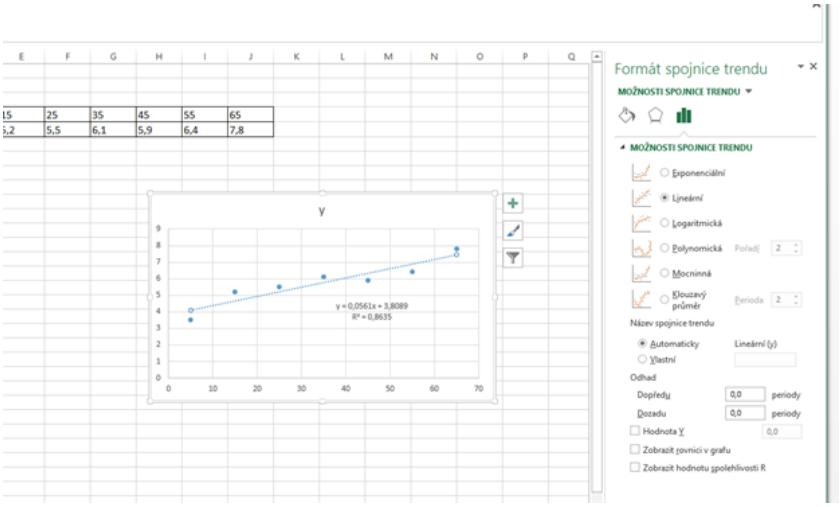
Obr. 18: Vložení bodového grafu

2. Máme-li aktivní okno grafu, v nabídce + vybereme možnost *Spojnice trendu* (obrázek 19).



Obr. 19: Přidání spojnice trendu

3. V rámci volby můžete volit i jiné křivky než přímku, a také vložit rovnici přímky přímo do grafu (obrázek 20):



Obr. 20: Nastavení lineární regrese

4. Výsledkem je rovnice regrese $y = 0{,}0561 \cdot x + 3{,}8089$. Z grafu vidíme, že rovnice dobře vystihuje závislost proměnných.

Řešení bez použití Excelu:

Pro výpočet regresní přímky použijeme vzorce:

$$y = \hat{\beta}_1 \cdot x + \hat{\beta}_0,$$

kde:

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2},$$

$$\hat{\beta}_0 = \frac{\sum y_i - a \sum x_i}{n}.$$

Pro naše data:

$$\sum x_i = 5 + 15 + 25 + 35 + 45 + 55 + 65 = 245,$$

$$\sum y_i = 3{,}5 + 5{,}2 + 5{,}5 + 6{,}1 + 5{,}9 + 6{,}4 + 7{,}8 = 40{,}4,$$

$$\sum x_i^2 = 5^2 + 15^2 + 25^2 + 35^2 + 45^2 + 55^2 + 65^2 = 8575,$$

$$\sum x_i y_i = 5 \cdot 3{,}5 + 15 \cdot 5{,}2 + 25 \cdot 5{,}5 + 35 \cdot 6{,}1 + 45 \cdot 5{,}9 + 55 \cdot 6{,}4 + 65 \cdot 7{,}8 = 1601{,}5.$$

Dosadíme do vzorců:

$$\hat{\beta}_1 = \frac{7 \cdot 1601{,}5 - 245 \cdot 40{,}4}{7 \cdot 8575 - 245^2} = 0{,}0561,$$

$$\hat{\beta}_0 = \frac{40{,}4 - 0{,}0561 \cdot 245}{7} = 3{,}8089.$$

Rovnice regresní přímky je tedy:

$$y = 0{,}0561 \cdot x + 3{,}8089.$$

□

Příklad 9.6. Použijte data z úvodního příkladu (tabulka 6) a odhadněte lineární regresní model pro vztah mezi výdaji na reklamu a tržbami. Určete odhady parametrů $\hat{\beta}_0$ a $\hat{\beta}_1$.

Řešení: **Krok 1: Výpočet průměrů**

$$\overline{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{20 + 25 + \cdots + 65}{10} = 42{,}5,$$

$$\overline{y} = \frac{\sum_{i=1}^{10} y_i}{10} = \frac{200 + 220 + \cdots + 450}{10} = 321.$$

Krok 2: Výpočet odhadu $\hat{\beta}_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{10}(x_i - \overline{x})^2}.$$

Spočítáme jednotlivé sumy:

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i y_i) - n\overline{x}\overline{y},$$
$$\sum (x_i - \overline{x})^2 = \sum x_i^2 - n\overline{x}^2.$$

Výpočty:

Vytvoříme tabulku pro výpočty (část výpočtů):

i	X_i	Y_i	X_iY_i	X_i^2
1	20	200	4 000	400
2	25	220	5 500	625
3	30	250	7 500	900
4	35	280	9 800	1 225
5	40	310	12 400	1 600
6	45	330	14 850	2 025
7	50	360	18 000	2 500
8	55	390	21 450	3 025
9	60	420	25 200	3 600
10	65	450	29 250	4 225
$\sum$	425	3 210	147 950	20 125

A tedy

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\overline{x}\overline{y}}{\sum x_i^2 - n\overline{x}^2} = \frac{147\,950 - 10 \cdot 42{,}5 \cdot 321}{20\,125 - 10 \cdot (42{,}5)^2} \approx 5{,}5882.$$

Výpočet $\hat{\beta}_0$:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 321 - 5{,}5882 \cdot 42{,}5 = 321 - 237{,}5 = 83{,}5.$$

Regresní rovnice:

$$\hat{Y} = 5{,}5882X + 83{,}5.$$

Výpočty v Excelu: Kromě postupu přímo v Excelu, jak jsme si to předvedli v předchozím příkladu, můžeme použít i pokročilejší modul `Analýza dat - Regrese`:

Postup:

1. Vložíme data do dvou sloupců: X (Reklama) a Y (Tržby).
2. Spustíme *Analýza dat* a vybereme *Regrese*.
3. Nastavíme vstupní rozsahy pro závislou a nezávislou proměnnou.
4. Zvolíme výstupní oblast a případně další možnosti (např. reziduální grafy).

Výstupem bude tabulka s odhady parametrů, ale také jejich směrodatnými chybami, hodnotami t-statistik a P-hodnotami.

Interpretace výsledků z Excelu:

Výsledky mohou vypadat například takto:

Parametr	**Odhad**	**Směr. chyba**	t	P**-hodnota**
$\hat{\beta}_0$	83,5	5,0	16,7	0,0000
$\hat{\beta}_1$	5,5882	0,2	27,9	0,0000

Rozhodnutí:

Protože P-hodnota pro $\hat{\beta}_1$ je mnohem menší než $\alpha = 0{,}05$, zamítáme nulovou hypotézu H_0 : $\beta_1 = 0$. Regresní koeficient $\hat{\beta}_1$ je tedy statisticky významný.

□

Σ

V této kapitole jsme se zabývali korelační a regresní analýzou, která slouží k analýze závislostí mezi dvěma kvantitativními znaky. Korelace hodnotí sílu a směr lineárního vztahu mezi dvěma proměnnými pomocí korelačního koeficientu r_{XY}. Pozitivní korelace značí, že s růstem jedné proměnné roste i druhá, zatímco negativní korelace ukazuje opačný vztah.

Regresní analýza pak umožňuje vyjádřit tento vztah pomocí matematického modelu. Nejčastěji se používá lineární regresní model, který popisuje vztah mezi závisle proměnnou

Y a nezávislou proměnnou X pomocí přímky. Parametry modelu, jako je směrnice a průsečík, jsou odhadovány metodou nejmenších čtverců.

V rámci kapitoly jsme si ukázali, jak tyto metody aplikovat na konkrétní data, jak interpretovat výsledky korelace a regrese. Důležitou součástí byla také vizualizace dat pomocí bodových grafů a regresních přímek.

1. Co je korelační koeficient a jaká je jeho interpretace?
2. Jaký je rozdíl mezi korelační a regresní analýzou?
3. Jak se vypočítá koeficient korelace r_{XY} mezi dvěma proměnnými?
4. Co znamená hodnota korelačního koeficientu blízká 1, 0 nebo -1?
5. Co je to lineární regrese a k čemu slouží?
6. Jak se odhadují parametry lineárního regresního modelu?
7. Co vyjadřuje směrnice a průsečík regresní přímky?
8. Jaké grafické nástroje se používají k vizualizaci výsledků korelační a regresní analýzy?
9. Uvažujme následující data, která představují počet hodin fyzického cvičení za týden a spotřebu kalorií (v tisících) pěti osob:

Osoba	Hodiny cvičení za týden (X)	Spotřeba kalorií (Y, v tisících)
1	3	2,2
2	5	2,8
3	7	3,1
4	8	3,5
5	10	4,0

Vypočítejte korelační koeficient mezi počtem hodin cvičení a spotřebou kalorií a interpretujte výsledek. $[r = 0{,}98]$

10. V následující tabulce jsou uvedeny hodnoty proměnných X a Y, kde X představuje počet hodin studia a Y dosažené skóre v testu:

Osoba	Hodiny studia (X)	Skóre (Y)
1	2	50
2	3	55
3	4	60
4	5	60
5	6	70

Určete parametry lineární regresní přímky pro závislost skóre na počtu hodin studia (vztah mezi X a Y) a napište rovnici regresní přímky. $[Y = 2X + 51]$

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 10

Časové řady

Po prostudování této kapitoly budete umět:

- definovat a vysvětlit základní pojmy časových řad,
- popsat klíčové složky časových řad, jako jsou trend, sezónnost a náhodná složka,
- rozlišit mezi stacionárními a nestacionárními časovými řadami,
- interpretovat grafickou analýzu časových řad.

Klíčová slova:

Časová řada, trend, sezónnost, cykličnost, stacionarita, grafická analýza.

Náhled kapitoly

V této kapitole se seznámíme s konceptem časových řad a jejich základními charakteristikami. Časové řady představují posloupnost hodnot sledovaných (většinou) v pravidelných časových intervalech. Tyto řady se používají k analýze dat v mnoha oblastech, jako jsou ekonomie, finance a další disciplíny. Probereme základní složky časových řad, jako jsou trend, sezónnost, cyklické jevy a náhodné výkyvy. Naučíme se, jak tyto složky rozlišit a interpretovat pomocí grafických metod.

Cíle kapitoly

Cílem této kapitoly je představit časové řady jako důležitý nástroj pro analýzu dat sledovaných v čase. Studenti se naučí rozpoznávat základní složky časových řad, pochopí rozdíl mezi stacionárními a nestacionárními řadami a budou schopni provést základní grafickou analýzu.

Odhad času potřebného ke studiu

Odhaduje se, že studium této kapitoly zabere přibližně 2 hodiny. Tento čas zahrnuje čtení textu, pochopení teoretických konceptů a interpretaci grafických analýz časových řad.

Úvod

Definice 10.1. Časové řady představují posloupnost hodnot, které jsou zaznamenávány v pravidelných nebo nepravidelných časových intervalech. Každá hodnota časové řady odpovídá určitému okamžiku nebo časovému úseku. Tento typ dat umožňuje analyzovat změny proměnné v čase a může odhalit různé vzorce chování proměny dat, jako jsou trendy (růst nebo pokles ve větším časovém měřítku) nebo sezónní výkyvy.

Příkladem časové řady může být vývoj ceny akcií na burze, počet prodaných výrobků v obchodě za jednotlivé měsíce nebo denní teplota zaznamenaná meteorologickou stanicí.

Kde se časové řady využívají?

Časové řady se využívají v mnoha oblastech, kde je třeba analyzovat a předvídat vývoj veličin v čase. Mezi nejčastější aplikace patří:

- *Ekonomie a finance*: Analýza vývoje cen akcií, kurzů měn, inflace nebo nezaměstnanosti.
- *Marketing*: Předpovědi poptávky, prodejních trendů, či sezónních výkyvů v tržbách.
- *Meteorologie*: Analýza teplotních změn, srážkových úhrnů nebo předpovědi počasí na základě historických dat.
- *Výrobní procesy*: Monitoring a analýza výkonnosti výrobních zařízení v čase, sledování kvality nebo optimalizace výrobních kapacit.

Díky těmto aplikacím je možné provádět analýzy, které pomáhají organizacím předvídat budoucí vývoj a lépe plánovat své aktivity.

10.1 Základní pojmy časových řad

Pozorování a časová osa

Časová řada je posloupnost hodnot určité veličiny, které jsou měřeny nebo zaznamenávány v nějakých (většinou pravidelných) časových intervalech.

Definice 10.2. Každá časová řada má dvě klíčové složky:

- *Časová osa*: Zahrnuje jednotlivé časové body (např. dny, měsíce, roky), ve kterých jsou hodnoty proměnné zaznamenány.
- *Hodnoty proměnné*: Reprezentují sledovanou veličinu (např. teplotu, cenu akcií, prodeje).

Časové řady jsou důležité pro zkoumání změn a trendů v průběhu času, což nám potenciálně umožňuje predikovat budoucí hodnoty na základě předchozích dat.

Trend, sezónnost, cykličnost a náhodná složka

Definice 10.3. Časovou řadu můžeme rozložit na několik základních složek:

- *Trend*: Dlouhodobý směr vývoje časové řady, který může být vzestupný, sestupný nebo konstantní. Představuje systematickou změnu hodnot v čase.
- *Sezónnost*: Krátkodobé pravidelné fluktuace, které se opakují v určitém časovém období (např. roční období, měsíční prodeje).
- *Cykličnost*: Dlouhodobé nepravidelné výkyvy, které nejsou striktně periodické, ale mohou souviset s ekonomickými nebo jinými cykly.
- *Náhodná složka*: Nepravidelné, nepředvídatelné výkyvy, které nelze vysvětlit trendem, sezónností ani cykličností. Tato složka představuje vlivy, které nejsou systematické a mohou být způsobeny různými náhodnými faktory.

Rozklad časové řady na tyto složky nám umožňuje lépe pochopit její strukturu a provádět analýzy, které jsou užitečné například při modelování a predikci.

10.2 Typy časových řad

Deterministické a stochastické časové řady

Definice 10.4. Časové řady můžeme rozdělit do dvou základních kategorií:

- *Deterministické časové řady*: U těchto řad je budoucí vývoj plně určen předchozími hodnotami. Neobsahují žádnou náhodnou složku a jsou často popsány jednoduchými matematickými funkcemi, například lineárním nebo exponenciálním trendem.
- *Stochastické časové řady*: Tyto řady obsahují náhodnou složku, což znamená, že jejich budoucí vývoj není zcela předvídatelný. Příkladem je fluktuace na finančních trzích, kde se vývoj ceny akcie v čase nedá přesně určit.

Rozlišení mezi deterministickými a stochastickými řadami je klíčové pro výběr vhodných metod analýzy a předpovědí.

Stacionární a nestacionární časové řady

Definice 10.5. Další důležité dělení časových řad je na stacionární a nestacionární:

- *Stacionární časové řady*: Časová řada je stacionární, pokud její statistické vlastnosti (např. průměr a rozptyl) zůstávají v čase konstantní. To znamená, že v průběhu času nepozorujeme žádný výrazný trend ani změny v kolísání hodnot. Stacionární časové řady jsou často jednodušší na analýzu a modelování.
- *Nestacionární časové řady*: V těchto řadách dochází ke změnám v čase, například k růstu nebo poklesu průměru, změnám v rozptylu nebo výskytu sezónních výkyvů. Pro analýzu nestacionárních časových řad je obvykle nutné aplikovat metody, které tyto změny zohlední, například diferenciaci.

Stacionarita je důležitý koncept, protože mnoho statistických metod předpokládá, že časová řada je stacionární. Pokud není, je třeba použít vhodné transformace, které pomohou dosáhnout stacionarity.

10.3 Analýza časových řad

Grafická analýza časových řad

Jedním z prvních kroků při analýze časové řady je vizuální zkoumání jejích vlastností pomocí grafů. Grafická analýza časových řad nám umožňuje identifikovat základní složky časové řady, jako jsou trend, sezónnost nebo náhodné výkyvy.

Definice 10.6. Mezi nejčastěji používané grafické nástroje patří:

- *Časový graf*: Zobrazuje hodnoty časové řady na vertikální ose a časové body na horizontální ose. Tento graf je ideální pro identifikaci dlouhodobých trendů a sezónních výkyvů.
- *Sezónní diagram*: Používá se k vizualizaci opakujících se sezónních vzorců. Umožňuje snadno rozpoznat, zda má časová řada pravidelné sezónní fluktuace v průběhu jednotlivých období (například různé měsíce nebo roční období).
- *Bodový diagram (scatter plot)*: Může být použit ke zkoumání závislosti mezi hodnotami časové řady v různých časových intervalech. Tento graf může odhalit autokorelaci (závislost mezi hodnotami v různých časech).

Grafická analýza poskytuje rychlý přehled o struktuře časové řady a je často prvním krokem před aplikací pokročilejších analytických metod.

Rozklad časové řady

Pro lepší pochopení struktury časové řady je často užitečné rozložit ji na jednotlivé složky: trend, sezónnost a náhodnou složku. Tento rozklad umožňuje oddělit systematické vlivy od náhodných výkyvů, což usnadňuje interpretaci a předpovědi.

Definice 10.7. Rozklad časové řady lze provést pomocí několika metod, například:

- *Additivní model*: Předpokládá, že časová řada je součtem trendu, sezónnosti a náhodné složky. Tento model je vhodný, pokud amplituda sezónních výkyvů zůstává konstantní v čase.
- *Multiplikativní model*: Předpokládá, že časová řada je součinem trendu, sezónnosti a náhodné složky. Tento model je vhodný, pokud se amplituda sezónních výkyvů mění s velikostí časové řady (například větší pro vyšší hodnoty časové řady).

Rozklad časové řady nám umožňuje lépe porozumět jejím jednotlivým složkám a případně predikovat budoucí hodnoty na základě trendů a sezónních vzorců.

10.4 Charakteristiky časových řad

Charakteristiky časových řad

Při analýze časových řad se používají základní charakteristiky růstu, které nám umožňují kvantifikovat změny hodnot mezi jednotlivými časovými body.

Definice 10.8. Mezi hlavní charakteristiky patří:

- *Absolutní přírůstky* (diference): Rozdíl mezi hodnotami časové řady ve dvou po sobě jdoucích obdobích. Absolutní přírůstek Δx_t pro období t je dán vztahem:

$$\Delta x_t = x_t - x_{t-1},$$

  kde x_t je hodnota časové řady v období t a x_{t-1} je hodnota v předchozím období.

- *Koeficienty růstu*: Poměr mezi hodnotou časové řady v období t a hodnotou v předchozím období $t-1$. Koeficient růstu k_t je dán vztahem:

$$k_t = \frac{x_t}{x_{t-1}}.$$

  Tento koeficient nám ukazuje relativní změnu hodnot mezi dvěma obdobími.

Průměrné charakteristiky

Pro získání obecnějšího obrazu o vývoji časové řady v delším období používáme

Definice 10.9. průměrné charakteristiky:

- *Průměrný absolutní přírůstek*: Jedná se o průměr všech absolutních přírůstků časové řady a vypočítá se jako:

$$\text{Průměrný přírůstek} = \frac{\sum_{t=2}^{n} \Delta x_t}{n-1}$$

  kde n je počet období.

- *Průměrný koeficient růstu*: Tento koeficient vyjadřuje průměrnou relativní změnu časové řady v průběhu několika období. Vypočítá se jako geometrický průměr koeficientů růstu:

$$\text{Průměrný koeficient růstu} = \left(\prod_{t=2}^{n} k_t \right)^{\frac{1}{n-1}}$$

Tyto průměrné charakteristiky poskytují přehled o celkovém trendu časové řady.

Aplikace v praxi

Charakteristiky růstu lze využít k analýze změn v různých oblastech, jako je produkce, prodej nebo zásoby. Například pomocí průměrného absolutního přírůstku lze sledovat, jak se postupně mění objem výroby v továrně, a průměrný koeficient růstu nám může ukázat, zda růst prodeje vykazuje stabilní tempo nebo kolísá mezi obdobími.

10.5 Řešené příklady

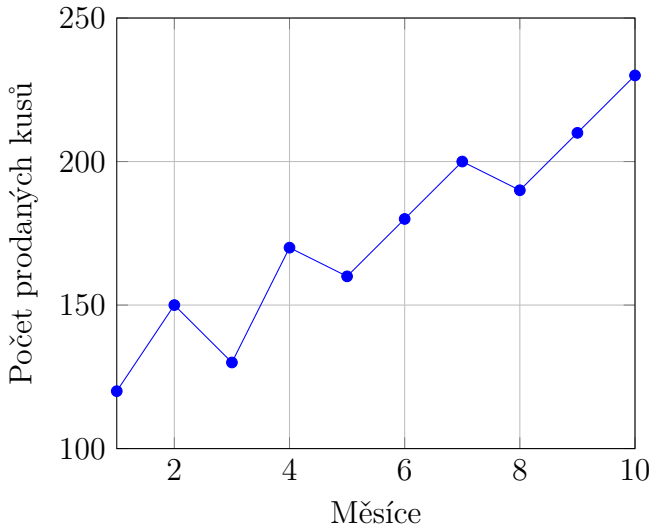
Příklad 10.10. Mějme následující časovou řadu, která představuje počet prodaných kusů určitého produktu v obchodě za posledních 10 měsíců:

$$(120, 150, 130, 170, 160, 180, 200, 190, 210, 230)$$

Vaším úkolem je:

1. Vykreslit časový graf této časové řady.
2. Identifikovat, zda časová řada obsahuje trend.

Řešení: 1. Pro vykreslení časového grafu použijeme hodnoty z časové řady na vertikální ose a čas (v měsících) na horizontální ose. Graf ukazuje, jak se počet prodaných kusů mění v čase.



2. Z časového grafu můžeme vidět, že počet prodaných kusů má obecně rostoucí trend. Ne v každém měsíci se počet prodaných kusů zvyšuje, ale celkově je jasný pozitivní růst. Tato časová řada tedy obsahuje trend.

□

Příklad 10.11. Určete elementární charakteristiky růstu časové řady sledující výrobu plynu v letech 1980 - 1985.

rok	**1980**	**1981**	**1982**	**1983**	**1984**	**1985**
výroba (m^3)	1286	1363	1393	1495	1571	1610

Řešení: Řešení:

rok	**výroba (m^3)** y_t	**absolutní přírůstky**	**koeficienty růstu**
1980	1286		
1981	1363	77	1,060
1982	1393	30	1,022
1983	1495	102	1,073
1984	1571	76	1,051
1985	1610	39	1,025

Průměrný absolutní přírůstek:

$$\overline{\Delta} = \frac{\sum \Delta y_t}{n-1} = \frac{(y_2 - y_1) + (y_3 - y_2) + \cdots + (y_n - y_{n-1})}{n-1} = \frac{y_n - y_1}{n-1} = \frac{1610 - 1286}{5} = 64{,}8$$

Průměrný koeficient růstu:

$$\overline{k} = \sqrt[n-1]{k_1 \cdot k_2 \cdot \cdots \cdot k_{n-1}} = \sqrt[5]{\frac{y_2}{y_1} \cdot \frac{y_3}{y_2} \cdot \frac{y_4}{y_3} \cdot \cdots \cdot \frac{y_n}{y_{n-1}}} = \sqrt[5]{\frac{1610}{1286}} \approx 1{,}046$$

□

10.6 Softwarová analýza časových řad

V předchozích dvou příkladech jsme si předvedli jen velmi základní výpočty.

Pro pokročilejší analýzu časových řad lze využít různé softwarové nástroje, které nabízejí specializované funkce a metody:

- *Excel*: Excel umožňuje provádět základní analýzu časových řad, jako je vykreslování časových grafů nebo výpočet klouzavých průměrů. Pro pokročilejší analýzy je možné použít doplněk `Analýza dat`, který zahrnuje funkce pro regresní analýzu nebo sezónní dekompozici.
- *R*: Ve statistické softwaru R jsou k dispozici speciální balíčky, jako například `forecast` nebo `tseries`, které poskytují nástroje pro modelování časových řad, jako jsou ARIMA modely, exponenciální vyrovnávání a testy stacionarity. R je velmi flexibilní a široce využívaný pro komplexní analýzy.
- *Wolfram Alpha*: Wolfram Alpha je interaktivní nástroj, který umožňuje provádět základní analýzu časových řad, jako je vykreslení grafů nebo výpočet trendů. Méně se hodí pro komplexní statistické modely, ale je užitečný pro rychlé vizualizace a základní výpočty.

Použití konkrétního softwaru závisí na potřebách analýzy – Excel je vhodný pro jednodušší úlohy a rychlou vizualizaci, zatímco R poskytuje nástroje pro pokročilé statistické modely, a Wolfram Alpha nabízí snadno přístupnou platformu pro základní výpočty.

Příklad 10.12. Ukázka grafických výstupů při analýze časové řady počtu cestujících. Data jsou součástí instalace softwaru R.

Řešení: Nejprve uvedeme programový kód, který nám v R, mimo jiné, vytvoří zmíněné grafické výstupy:

```
# Načtení datasetu AirPassengers
data("AirPassengers")
# Základní informace o datasetu
summary(AirPassengers)
plot(AirPassengers, main="Počet cestujících v letecké dopravě (1949-1960)",
     ylab="Počet cestujících", xlab="Rok", col="blue")
# Decompose časové řady (rozklad na trend, sezónnost a náhodnou složku)
decomposed <- decompose(AirPassengers)
plot(decomposed, col="darkred")
# Autokorelační graf
acf(AirPassengers, main="Autokorelační funkce pro AirPassengers")
# ARIMA model pro předpověď
library(forecast)
model <- auto.arima(AirPassengers)
forecasted <- forecast(model, h=24)
# Graf předpovědi
plot(forecasted, main="Předpověď počtu cestujících na příští 2 roky", col="green")
# Výstup modelu
summary(model)
```

Pokračujeme ukázkou grafů.

- Na obrázku 21 na straně 156 je znázorněna časová řada počtu cestujících.
- Na obrázku 22 na straně 156 je provedena tzv. dekompozice (rozklad) časové řady na trendovou, sezónní a náhodnou složku.
- Na obrázku 23 na straně 156 je ukázka předpovědi.

□

V této kapitole jsme se věnovali časovým řadám, které popisují vývoj veličin v čase. Hlavní body zahrnují:

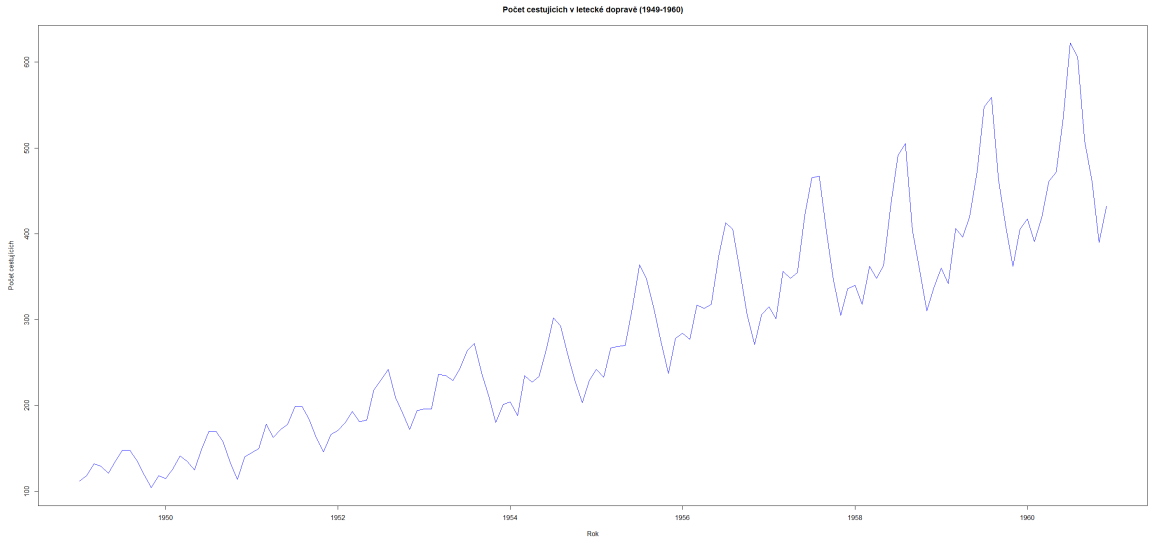
- **Základní pojmy**: Probrali jsme časovou osu, hodnoty proměnných a základní složky časové řady, jako jsou trend, sezónnost a náhodné výkyvy.
- **Typy časových řad**: Rozdělili jsme časové řady na deterministické a stochastické, stacionární a nestacionární.
- **Charakteristiky růstu**: Představili jsme absolutní přírůstky, koeficienty růstu a jejich průměrné hodnoty jako nástroje pro kvantifikaci změn časové řady.

Kapitola poskytuje jen velmi základní nástroje pro analýzu časových řad v různých oborech.

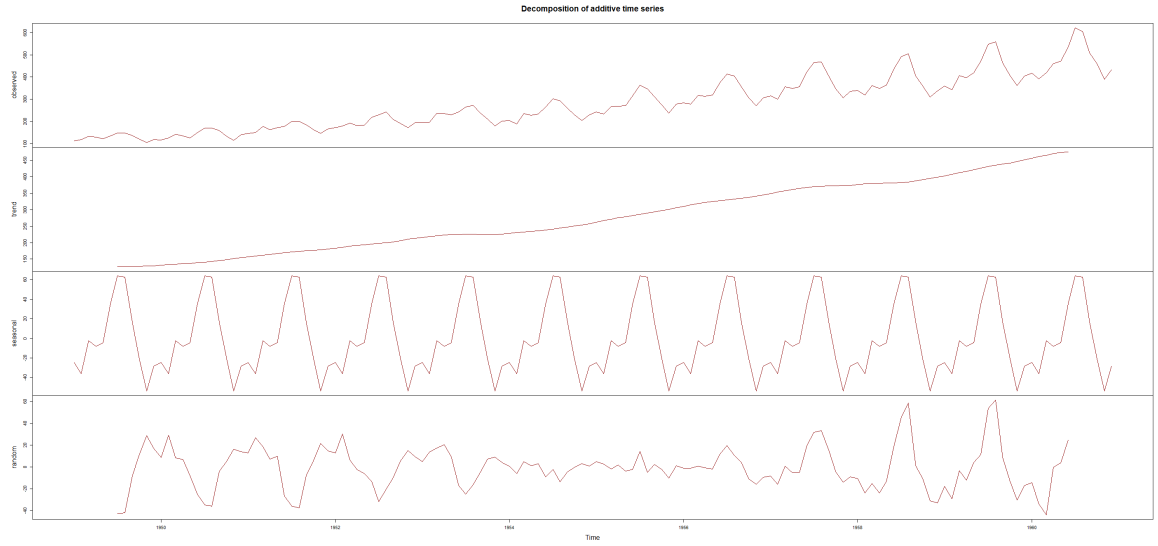
1. Jaké jsou základní složky časové řady? Uveďte příklady každé z nich.
2. Jaký je rozdíl mezi stacionární a nestacionární časovou řadou?
3. Jaký je význam průměrného absolutního přírůstku a průměrného koeficientu růstu v analýze časových řad?
4. V jakých situacích byste použili multiplicativní model namísto aditivního modelu pro rozklad časové řady?
5. Vysvětlete, jak lze využít Excel, R nebo Wolfram Alpha pro analýzu časových řad. Jaké jsou hlavní rozdíly mezi těmito nástroji?
6. Majitel prodejny evidoval čtvrtletně objem prodeje ovocných kompotů a jejich zásoby na počátku čtvrtletí.

čtvrtletí	prodej ks	zásoby ks
I.	560	220
II.	480	210
III.	520	215
IV.	550	200

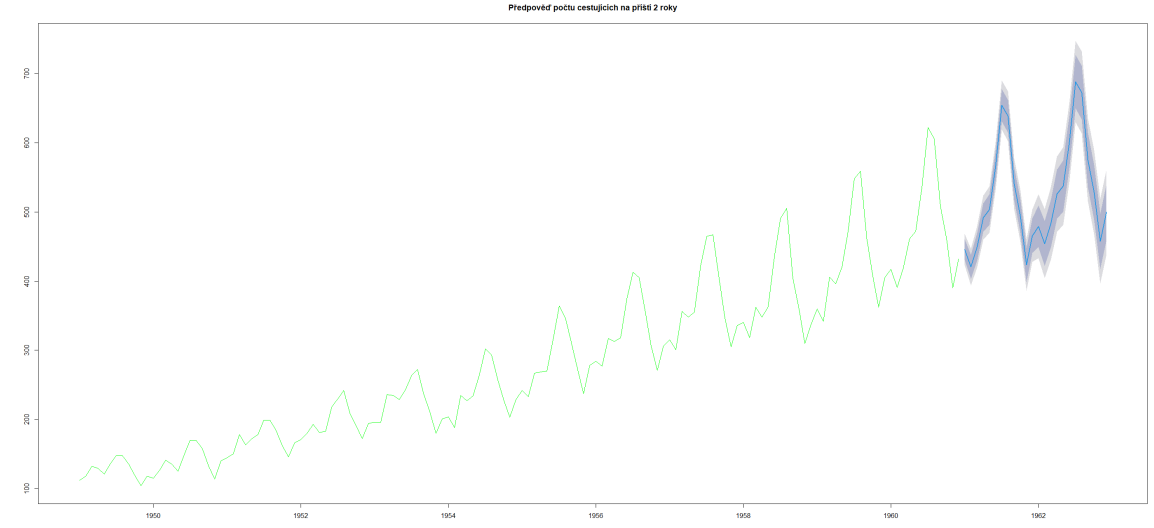
Na konci 4. čtvrtletí bylo v zásobě 150 ovocných kompotů. Vypočtěte průměrný čtvrtletní prodej a průměrnou čtvrtletní zásobu ovocných kompotů. [527,5, 199]



Obr. 21: Graf časové řady z příkladu 10.12



Obr. 22: Dekompozice časové řady z příkladu 10.12



Obr. 23: Graf předpovědi časové řady z příkladu 10.12

7. Časová řada následujících hodnot představuje počet prodaných kusů elektroniky v obchodě za posledních 12 měsíců:

$$(120, 130, 110, 150, 140, 160, 170, 165, 180, 175, 190, 185)$$

   a. Vypočtěte absolutní přírůstky pro každý měsíc.
   b. Vypočtěte koeficient růstu pro každý měsíc.
   c. Určete průměrný absolutní přírůstek a průměrný koeficient růstu.

   [. . . , . . . , 7,27, 1,0217]

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 11

Induktivní statistika

Po prostudování této kapitoly budete umět:

- určit bodový odhad zvolených parametrů,
- určit intervalový odhad (interval spolehlivosti) střední hodnoty a rozptylu při zvolené hladině spolehlivosti,
- použít nástroje Excelu a R pro výpočty bodových a intervalových odhadů v praktických příkladech.

Klíčová slova:

Bodový odhad, intervalový odhad, střední hodnota, rozptyl, Excel, R.

Náhled kapitoly

V této kapitole se budeme věnovat základním nástrojům induktivní statistiky, kterými jsou bodové a intervalové odhady. Tyto odhady umožňují na základě výběrových dat vyvodit závěry o základním souboru, což je klíčová součást statistické analýzy. Naučíme se, jak vypočítat bodový a intervalový odhad střední hodnoty (průměru) a rozptylu, a to jak teoreticky, tak i prakticky s využitím programů Excel a R.

Cíle kapitoly

Cílem této kapitoly je pochopit hlavní myšlenku induktivní statistiky a naučit se odhadovat parametry základního souboru pomocí bodových a intervalových odhadů.

Odhad času potřebného ke studiu

Studium této kapitoly by mělo zabrat přibližně 2 hodiny. Tento čas zahrnuje prostudování teorie, porozumění odhadovým metodám a zvládnutí praktických výpočtů v Excelu a R.

Úvod

Zopakujme si, že statistika je obor, který se zabývá sběrem, analýzou a interpretací hromadných pozorování a výsledků opakovaných pokusů. Je rozdělena na dva hlavní typy:

- **Deskriptivní (popisná) statistika**: Zaměřuje se na uspořádání datových souborů, jejich popis a účelnou sumarizaci.
- **Induktivní statistika**: Pomocí empirických poznatků umožňuje vytvářet vědecky odůvodněné obecné závěry. Tento přístup je založen na teorii pravděpodobnosti.

Stejně jako statistika, i lidské myšlení lze rozdělit na různé typy podle způsobu uvažování. Mezi nejvýznamnější typy patří:

Deduktivní myšlení

Deduktivní myšlení je proces, při kterém vyvozujeme závěry z obecných zákonitostí nebo pravidel. Z obecných principů vytváříme specifické závěry, které se uplatňují v jednotlivých případech. Deduktivní myšlení zajišťuje přesné a logické usuzování.

Příklad: Všichni lidé jsou smrtelní. Sokrates je člověk. Tudíž Sokrates je smrtelný.

Induktivní myšlení

Induktivní myšlení vychází z konkrétních pozorování jednotlivých případů a zobecňuje je do obecných závěrů. Na rozdíl od dedukce, indukce často pracuje s nejistotou, protože závěry jsou ovlivněny subjektivními postoji a mají omezenou platnost.

Příklad: Každé ráno, kdy jsem pozoroval východ slunce, slunce skutečně vyšlo. Proto mohu induktivně usoudit, že slunce vyjde i zítra ráno.

Další typy myšlení

- **Abduktivní myšlení**: Vyvozování nejpravděpodobnějšího vysvětlení na základě dostupných informací. Často se používá při řešení neúplných problémů, kde se snažíme najít nejlepší hypotézu.

  Příklad: „Zem je mokrá, pravděpodobně pršelo.“

- **Kreativní myšlení**: Schopnost generovat nové a originální nápady nebo řešení. Zaměřuje se na netradiční přístupy k řešení problémů.

  Příklad: „Namísto tradičního reklamačního procesu navrhneme zcela nový způsob zákaznického servisu pomocí umělé inteligence.“

- **Kritické myšlení**: Proces systematického hodnocení a zkoumání informací, argumentů a důkazů. Cílem je dospět ke správným závěrům založeným na logice a důkazech.

  Příklad: „Tento článek tvrdí, že určité potraviny jsou škodlivé, ale podívejme se na důkazy a ověřme, zda to podporují i jiné studie.“

Statistická indukce je proces, při kterém pomocí statistických metod dokážeme vytvářet obecné závěry z dostupných dat. Jejich spolehlivost lze kvantifikovat pomocí pravděpodobnosti. Základem statistické indukce je práce s **výběrem** a **základním souborem**.

Základní soubor (populace)

Základní soubor, někdy označován jako populace, je množina všech prvků, které jsou předmětem zkoumání. Tento soubor může být:

- **Konečný**: Např. počet obyvatel v určité zemi.
- **Nekonečný**: Hypotetický soubor, který je ideální a v realitě neexistuje.

Prvky základního souboru mají různé vlastnosti, nazývané **znaky**. Tyto znaky dělíme na:

- **Kvalitativní**:
  - *Nominální*: Vlastnosti, které lze pouze pojmenovat (např. barva očí).
  - *Ordinální*: Vlastnosti, které lze uspořádat (např. spokojenost zákazníků na škále 1 až 5).
- **Kvantitativní**:
  - *Diskrétní*: Hodnoty mohou nabývat pouze určitých hodnot (např. počet dětí v rodině).
  - *Spojité*: Hodnoty mohou nabývat jakékoliv hodnoty v daném intervalu (např. výška člověka).

Výběr

Výběr je část základního souboru, kterou zkoumáme a na základě které usuzujeme na celou populaci. Aby byl výběr reprezentativní, musí odpovídat vlastnostem celého základního souboru. Pokud není výběr reprezentativní, jedná se o selektivní výběr.

Metody výběru:

- **Náhodný výběr**: Prvky vybíráme náhodně, například losováním nebo pomocí tabulek náhodných čísel.
- **Mechanický (systematický) výběr**: Prvky vybíráme podle pevně stanoveného pravidla (např. každý třetí prvek).
- **Oblastní (stratifikovaný) výběr**: Základní soubor je rozdělen na homogenní oblasti, ze kterých jsou prvky vybírány náhodně.
- **Skupinový výběr**: Používá se pro velké populace, kdy vybíráme celé skupiny prvků (např. domácnosti nebo rodiny).
- **Vícestupňový výběr**: Prvky jsou vybírány postupně z různých úrovní hierarchie (např. město – domácnost – osoba).

11.1 Odhady v induktivní statistice

V oblasti induktivní statistiky se nejčastěji zaměřujeme na odhadování parametrů základního souboru na základě výběrových dat. Mezi hlavní parametry, které odhadujeme, patří:

- **Průměr (střední hodnota)**: Odhadujeme střední hodnotu populace na základě průměru ve výběru.
- **Rozptyl**: Odhadujeme rozptyl populace na základě výběrového rozptylu.
- **Proporce**: Odhady podílů určité charakteristiky v populaci (např. podíl lidí s určitým názorem).

Zde se konkrétně zaměříme na **bodový** a **intervalový** odhad průměru (střední hodnoty) a rozptylu.

11.1.1 Bodový a intervalový odhad průměru (střední hodnoty)

Bodový odhad průměru

Definice 11.1. Bodový odhad průměru vyjadřuje nejlepší odhad skutečné střední hodnoty populace na základě výběrového průměru. Bodový odhad střední hodnoty μ se vypočítá jako:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

kde x_i jsou jednotlivé hodnoty z výběru a n je počet pozorování.

Praktický výpočet v Excelu:

V Excelu můžete bodový odhad průměru vypočítat pomocí funkce `PRŮMĚR`:

```
=PRŮMĚR(A1:A10),
```

kde rozsah buněk `A1:A10` obsahuje hodnoty výběru.

Praktický výpočet v R:

V R můžete bodový odhad průměru spočítat funkcí `mean()`:

```
mean(data),
```

kde `data` je vektor obsahující hodnoty výběru.

Intervalový odhad průměru

Definice 11.2. Intervalový odhad poskytuje rozsah hodnot, ve kterém se s určitou pravděpodobností nachází skutečný průměr populace. Intervalový odhad pro střední hodnotu μ s danou hladinou spolehlivosti $1 - \alpha$ se vypočítá jako:

$$\left\langle \hat{\mu} - u_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \hat{\mu} + u_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right\rangle,$$

kde $u_{1-\alpha/2}$ je kvantil normálního rozdělení pro zvolenou hladinu spolehlivosti, σ je směrodatná odchylka populace (případně odhad ze vzorku) a n je velikost výběru.

Praktický výpočet v Excelu:

Intervalový odhad průměru lze v Excelu vypočítat pomocí následujícího postupu:

1. Výpočet průměru: `=PRŮMĚR(A1:A10)`
2. Výpočet směrodatné odchylky: `=SMODCH.VÝBĚR.S(A1:A10)`
3. Výpočet velikosti výběru: `=POČET(A1:A10)`
4. K výpočtu kvantilu normálního rozdělení použijeme funkci `NORM.INV` nebo `NORM.S.INV`, např. pro hladinu spolehlivosti 95%: `=NORM.S.INV(0,975)`
5. Intervalový odhad pak získáme jako průměr $\pm\, u_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$.

Praktický výpočet v R:

V R můžeme intervalový odhad průměru vypočítat pomocí kombinace funkcí:

```
mean(data) + c(-1, 1) * qnorm(0.975) * sd(data)/sqrt(length(data))
```

11.1.2 Bodový a intervalový odhad rozptylu

Bodový odhad rozptylu

Definice 11.3. Bodový odhad rozptylu vyjadřuje nejlepší odhad skutečného rozptylu populace na základě výběrového rozptylu. Bodový odhad rozptylu σ^2 se vypočítá jako:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2,$$

kde $\hat{\mu}$ je průměr výběru a x_i jsou jednotlivé hodnoty z výběru.

Praktický výpočet v Excelu:

V Excelu můžete bodový odhad rozptylu vypočítat pomocí funkce `VAR.S`:

```
=VAR.S(A1:A10)
```

Praktický výpočet v R:

V R můžete bodový odhad rozptylu vypočítat funkcí `var()`:

```
var(data)
```

Intervalový odhad rozptylu

Definice 11.4. Intervalový odhad rozptylu lze vypočítat s využitím χ^2 rozdělení, které se používá pro odhady rozptylu. Intervalový odhad rozptylu s hladinou spolehlivosti $1-\alpha$ se vypočítá jako:

$$\left\langle \frac{(n-1)\cdot\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)\cdot\hat{\sigma}^2}{\chi^2_{\alpha/2,n-1}} \right\rangle,$$

kde $\chi^2_{\alpha/2,n-1}$ je kvantil χ^2 rozdělení.

Praktický výpočet v Excelu:

Intervalový odhad rozptylu můžete vypočítat pomocí následujících kroků:

1. Výpočet rozptylu: `=VAR.S(A1:A10)`
2. Výpočet velikosti výběru: `=POČET(A1:A10)`
3. K výpočtu kvantilu χ^2 rozdělení použijte funkci `CHISQ.INV`, např.: `=CHISQ.INV(0,975; n-1)`
4. Intervalový odhad rozptylu pak získáme dosazením o vzorce pro interval.

Praktický výpočet v R:

V R můžeme intervalový odhad rozptylu vypočítat pomocí následujícího kódu:

```
n <- length(data)

var(data) * (n-1) / qchisq(c(0.975, 0.025), n-1)
```

Tento výpočet nám poskytne dolní a horní hranici intervalového odhadu rozptylu.

11.2 Řešené příklady

Příklad 11.5. Při měření průměru vačkového hřídele na 250 součástkách bylo zjištěno, že výběrový průměr činí $\bar{x}_p = 995{,}6$ a výběrová disperze $s^2 = 134{,}7$. Předpokládáme, že soubor má normální rozdělení. Určete interval spolehlivosti pro střední hodnotu základního souboru při hladině významnosti $\alpha = 0{,}05$.

Řešení: Pro odhad střední hodnoty základního souboru μ na základě výběrových dat se používá interval spolehlivosti ve tvaru:

$$\langle \bar{x}_p - \Delta; \bar{x}_p + \Delta \rangle ,$$

kde $\bar{x}_p$ je výběrový průměr, Δ je tzv. mezní chyba odhadu a určuje se podle vztahu:

$$\Delta = \frac{s}{\sqrt{n}} \cdot u_{1-\frac{\alpha}{2}}.$$

V tomto výrazu:

- s je směrodatná odchylka výběru,
- n je počet pozorování (v našem případě $n = 250$),

- $u_{1-\frac{\alpha}{2}}$ je kritická hodnota normálního rozdělení odpovídající zvolené hladině významnosti α.

Pro hladinu významnosti $\alpha = 0{,}05$ je hodnota $u_{1-\frac{\alpha}{2}} = \text{NORM.S.INV}(0{,}975) \approx 1{,}96$.

Nyní vypočítáme mezní chybu odhadu Δ:

$$\Delta = \frac{\sqrt{134{,}7}}{\sqrt{250}} \cdot 1{,}96 \approx 1{,}441558.$$

Intervalový odhad střední hodnoty μ je tedy:

$$\langle \bar{x}_p - \Delta; \bar{x}_p + \Delta \rangle = \langle 995{,}6 - 1{,}441558; 995{,}6 + 1{,}441558 \rangle = \langle 994{,}1584; 997{,}0416 \rangle .$$

Z toho plyne, že s 95 % spolehlivostí lze tvrdit, že skutečná střední hodnota průměru vačkového hřídele leží v intervalu $\langle 994{,}1584; 997{,}0416 \rangle$. □

Příklad 11.6. Určete oboustranný konfidenční interval rozptylu normálně rozloženého základního souboru pro hladiny spolehlivosti 0,90, 0,95 a 0,99, když u výběru s rozsahem $n = 12$ byl zjištěn rozptyl $s^2 = 0{,}64$. Posuďte získané výsledky.

Řešení: Pro výpočet konfidenčního intervalu pro rozptyl σ^2 normálně rozloženého základního souboru použijeme vztah:

$$\frac{n \cdot s^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \leq \sigma^2 \leq \frac{n \cdot s^2}{\chi^2_{\frac{\alpha}{2}}(n-1)},$$

kde

- $n = 12$ je rozsah výběru,
- $s^2 = 0{,}64$ je výběrový rozptyl,
- $\chi^2_{1-\frac{\alpha}{2}}(n-1)$ a $\chi^2_{\frac{\alpha}{2}}(n-1)$ jsou kritické hodnoty Pearsonova rozdělení s $n - 1 = 11$ stupni volnosti.

1. Případ: Hladina spolehlivosti 0,90

Pro hladinu spolehlivosti $1 - \alpha = 0{,}90$ je $\alpha = 0{,}10$. Kritické hodnoty jsou:

$$\chi^2_{0{,}05}(11) = \text{CHIINV}(0{,}05; 11) \approx 19{,}675,$$

$$\chi^2_{0{,}95}(11) = \text{CHIINV}(0{,}95; 11) \approx 4{,}575.$$

Dosazením do vztahu:

$$\frac{12 \cdot 0{,}64}{19{,}675} \leq \sigma^2 \leq \frac{12 \cdot 0{,}64}{4{,}575},$$

$$0{,}390 \leq \sigma^2 \leq 1{,}678.$$

2. Případ: Hladina spolehlivosti 0,95

Pro hladinu spolehlivosti $1 - \alpha = 0{,}95$ je $\alpha = 0{,}05$. Kritické hodnoty jsou:

$$\chi^2_{0{,}025}(11) = \text{CHIINV}(0{,}025; 11) \approx 22{,}362,$$

$$\chi^2_{0,975}(11) = \text{CHIINV}(0{,}975; 11) \approx 3{,}816.$$

Dosazením do vztahu:

$$\frac{12 \cdot 0{,}64}{22{,}362} \leq \sigma^2 \leq \frac{12 \cdot 0{,}64}{3{,}816},$$
$$0{,}343 \leq \sigma^2 \leq 2{,}012.$$

3. Případ: Hladina spolehlivosti 0,99

Pro hladinu spolehlivosti $1 - \alpha = 0{,}99$ je $\alpha = 0{,}01$. Kritické hodnoty jsou:

$$\chi^2_{0,005}(11) = \text{CHIINV}(0{,}005; 11) \approx 26{,}757,$$
$$\chi^2_{0,995}(11) = \text{CHIINV}(0{,}995; 11) \approx 2{,}603.$$

Dosazením do vztahu:

$$\frac{12 \cdot 0{,}64}{26{,}757} \leq \sigma^2 \leq \frac{12 \cdot 0{,}64}{2{,}603},$$
$$0{,}287 \leq \sigma^2 \leq 2{,}952.$$

Z výsledků vidíme, že s rostoucí hladinou spolehlivosti se konfidenční interval rozšiřuje. □

V této kapitole jsme se věnovali základním metodám induktivní statistiky, zejména bodovým a intervalovým odhadům, které jsou klíčovými nástroji pro usuzování o parametrech základního souboru na základě výběrových dat.

V kapitole byl kladen důraz na praktické využití těchto metod v Excelu a R, kde byly představeny konkrétní funkce pro výpočet bodových a intervalových odhadů, jako např. `PRŮMĚR`, `SMODCH.VÝBĚR`, `NORM.S.INV` a `CHISQ.INV` v Excelu a `mean()`, `var()`, `qnorm()` a `qchisq()` v R.

1. Vysvětlete, co je bodový odhad a jak se liší od intervalového odhadu.
2. Jaké jsou hlavní složky při výpočtu intervalového odhadu pro střední hodnotu? Vysvětlete, co znamená hladina spolehlivosti.
3. Kdy použijete pro intervalový odhad průměru normální rozdělení a kdy Studentovo t-rozdělení?
4. Jaký je rozdíl mezi intervalovým odhadem pro střední hodnotu a intervalovým odhadem pro rozptyl? Uveďte vzorce a vysvětlete jednotlivé členy.
5. Co znamená, že intervalový odhad má hladinu spolehlivosti 95 %? Může být tento odhad vždy správný?
6. Jaký je význam kritické hodnoty v kontextu intervalových odhadů? Jak se liší kritické hodnoty pro různé hladiny spolehlivosti?
7. Jakou roli hraje velikost výběru při výpočtu intervalových odhadů? Jak se mění šířka intervalu s rostoucím počtem pozorování?

8. Byla měřena délka trvání určitého procesu. Z 12 měření byla zjištěna střední doba trvání procesu 44 s a směrodatná odchylka 4 s. Sestrojte 90 % a 95 % interval spolehlivosti pro očekávanou délku procesu za předpokladu normálního rozdělení. [90 % CI: (42,02; 45,98), 95 % CI: (41,45; 46,55)]
9. Při měření kapacity sady kondenzátorů bylo provedeno 10 měření s výsledky: 152, 156, 148, 153, 150, 156, 140, 155, 145, 148. Odhadněte interval spolehlivosti pro kapacitu těchto kondenzátorů se spolehlivostí a) 90 %, b) 95 %. [90 % CI: (146,41; 154,19), 95 % CI: (145,58; 155,02)]
10. Určete intervalový odhad s 90 % spolehlivostí střední hodnoty a směrodatné odchylky pro následující hodnoty: 606, 1249, 267, 44, 510, 340, 109, 1957, 463, 801, 1086, 169, 233, 1734, 1458, 80, 1023, 2736, 917, 459. [90 % CI: (487,87; 1224,73)]
11. Vzorek 20 studentů měl průměrnou dobu studia na zkoušku 5,6 hodiny se směrodatnou odchylkou 1,2 hodiny. Určete 95 % interval spolehlivosti pro průměrnou dobu studia celé populace studentů, pokud předpokládáme, že délka studia má normální rozdělení. [95 % CI: (5,00; 6,20)]
12. Po změření výšky 30 osob bylo zjištěno, že průměrná výška je 172 cm a směrodatná odchylka je 5 cm. Sestrojte 99 % interval spolehlivosti pro průměrnou výšku celé populace. [99 % CI: (169,49; 174,51)]
13. Při experimentu s délkou životnosti určitého druhu baterie bylo zaznamenáno 15 hodnot. Výběrový průměr životnosti je 500 hodin a směrodatná odchylka je 40 hodin. Určete 90 % interval spolehlivosti pro očekávanou délku životnosti baterií. [90 % CI: (481,80; 518,20)]
14. Vzorek 25 produktů měl výběrový rozptyl 0,36. Určete interval spolehlivosti pro rozptyl populace na hladině významnosti 0,05. [95 % CI: (0,219; 0,693)]

Literatura k tématu:

[1] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6. ISBN 978-80-867-3208-8.

[2] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[3] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4. Dostupné z: `https://homel.vsb.cz/~oti73/cdpast1/`

[4] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Kapitola 12

Využití softwaru při řešení statistických úloh

Po prostudování této kapitoly budete umět:

- využít software pro řešení vybraných statistických úloh,
- načíst data z externích zdrojů do Excelu,
- analyzovat rozsáhlejší data v Excelu.

Klíčová slova:

MS Excel, Wolfram Alpha, R, statistické úlohy, data z internetu, analýza dat.

Náhled kapitoly

Tato kapitola se zaměřuje na využití softwarových nástrojů pro řešení statistických úloh. Nejprve shrneme hlavní funkce a možnosti Excelu, který jsme používali v předchozích kapitolách. Dále se podíváme na Wolfram Alpha, což je výkonný výpočetní nástroj, vhodný pro rychlé teoretické výpočty, ale méně vhodný pro práci s rozsáhlými datovými sadami. Nakonec se informativně seznámíme s R, jehož hlavní výhoda spočívá v pokročilých analýzách dat, které však vyžadují instalaci softwaru a základní znalost programování. Kapitolu zakončíme praktickými příklady s reálnými daty z internetu.

Cíle kapitoly

Cílem této kapitoly je naučit studenty využívat různé nástroje pro statistické výpočty. Studenti si zopakují základní funkce Excelu, naučí se používat Wolfram Alpha k řešení menších úloh a získají základní povědomí o možnostech softwaru R. Důraz bude kladen na praktické aplikace těchto nástrojů při analýze aktuálních dat dostupných na internetu.

Odhad času potřebného ke studiu

Studium této kapitoly zabere přibližně 3 hodiny, včetně času na procvičování. Tento čas zahrnuje shrnutí práce s Excelem, seznámení s Wolfram Alpha, informativní přehled o R a řešení praktických příkladů.

Úvod

V této kapitole se budeme věnovat třem hlavním nástrojům pro řešení statistických úloh: MS Excel, Wolfram Alpha a R. Každý z nich má své výhody i omezení. Excel je široce dostupný a praktický nástroj, Wolfram Alpha umožňuje rychlé teoretické výpočty a R je silný nástroj pro pokročilé analýzy, ale vyžaduje určité programovací dovednosti. Zaměříme se především na Excel a Wolfram Alpha, zatímco R bude představen spíše informativně.

Excel jsme již používali v předchozích kapitolách, proto zde shrneme jeho hlavní funkce a ukážeme, jak je aplikovat na aktuální data. Wolfram Alpha bude vysvětlen více od začátku, jelikož jsme s ním zatím moc nepracovali. U složitějších analýz, jako je regresní analýza nebo práce s rozsáhlými daty, doporučujeme používat R, které však vyžaduje programování.

V této kapitole se také podíváme, jak stáhnout aktuální data z veřejně dostupných zdrojů, například z ČNB , a jak tato data analyzovat pomocí Excelu.

12.1 Shrnutí práce s MS Excel

V této sekci si shrneme hlavní statistické funkce a možnosti, které jsme v Excelu již používali v předchozích kapitolách. Excel je nástroj široce dostupný a je ideální pro základní statistické úlohy, zejména pro práci s menšími datovými soubory a pro vizualizaci dat.

Základní statistické funkce v Excelu

- **PRŮMĚR** – slouží k výpočtu průměrné hodnoty datové sady.
  `=PRŮMĚR(A1:A10)`
- **SMODCH.VÝBĚR.S** – vypočítá směrodatnou odchylku výběru.
  `=SMODCH.VÝBĚR.S(A1:A10)`
- **VAR.S** – slouží k výpočtu výběrového rozptylu dat.
  `=VAR.S(A1:A10)`
- **COVARIANCE.P** – vypočítá kovarianci mezi dvěma datovými sadami.
  `=COVARIANCE.P(A1:A10;B1:B10)`
- **CORREL** – slouží k výpočtu korelačního koeficientu mezi dvěma proměnnými.
  `=CORREL(A1:A10;B1:B10)`

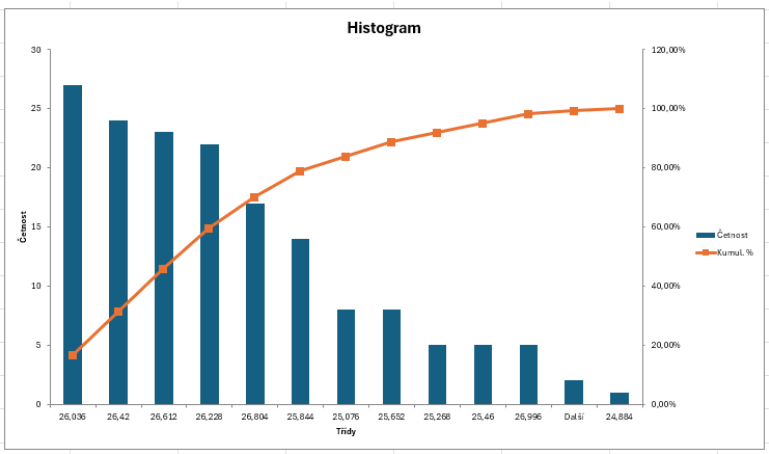
Modul Analýza dat

Excel obsahuje modul **Analýza dat**, který poskytuje více pokročilých statistických nástrojů:

- **Popisná statistika** – zobrazí základní souhrnné statistiky jako je průměr, medián, směrodatná odchylka a rozptyl.
- **Histogram** – vizualizace dat v podobě rozdělení četností. Ukázka je na obrázku 24.
- **Regresní analýza** – nástroj pro výpočet regresních koeficientů a analýzu vztahu mezi proměnnými.
- **Korelační matice** – výpočet korelačních koeficientů mezi více proměnnými.

Postup pro použití modulu Analýza dat:

1. Aktivujte modul `Analýza dat` (Pokud není modul aktivní, přidejte jej přes Možnosti Excelu → Doplňky → Analytické nástroje).
2. Zvolte požadovanou metodu analýzy, např. `Popisná statistika`.
3. Vyberte oblast dat, na kterých chcete analýzu provést, a potvrďte volbu.
4. Výsledky se zobrazí v novém listu nebo ve zvoleném rozsahu buněk.



Obr. 24: Ukázka histogramu (četnosti a kumulativní relativní četnosti) z modulu Analýza dat

Tento modul je velmi užitečný pro provádění rychlých analýz a statistických výpočtů, které by jinak vyžadovaly více manuálních kroků.

Grafické zpracování dat

Excel také poskytuje nástroje pro vytváření vizuálních reprezentací dat:

- **Sloupcové grafy** – vhodné pro vizualizaci kategoriálních dat.
- **Spojnicové grafy** – ideální pro znázornění časových řad.
- **Bodové grafy** – často používané při regresní a korelační analýze.
- **Histogram** – pro znázornění rozdělení četností.

Všechny tyto grafy lze snadno vytvořit prostřednictvím nástroje `Vložit` → `Grafy`. Vizualizace dat je důležitou součástí analýzy, protože poskytuje okamžitý náhled na strukturu a rozdělení dat.

Ilustrativní příklad

Příklad 12.1. Máme následující data o výnosech akcií za poslední 10 dnů: 3, 5, 2, 7, 6, 8, 4, 7, 9, 5. Pomocí Excelu vypočítejte průměr, směrodatnou odchylku a vytvořte histogram těchto dat.

Řešení: V Excelu použijeme následující funkce:

- Průměr: `=PRŮMĚR(A1:A10)` = 5,6.
- Směrodatná odchylka (výběrová): `=SMODCH.VÝBĚR.S(A1:A10)` = 2,059.
- Histogram vytvoříme pomocí modulu `Analýza dat` → `Histogram`, kde zvolíme intervaly a četnosti.

□

12.2 Představení Wolfram Alpha a R

V minulé sekci jsme shrnuli základní práci s excelovskými funkcemi a moduly. V této části se podíváme na další softwarové nástroje, konkrétně Wolfram Alpha a R.

12.2.1 Srovnání R a Wolfram Alpha

1) Licencování:

- **R:** Otevřený software, zdarma, licencován pod GPL. Komunita neustále přidává nové balíčky. Je nutné ho nainstalovat na lokální počítač, ale existují i některé online služby pro běh R.
- **Wolfram Alpha:** Komerční produkt. Základní verze je zdarma, pokročilé funkce vyžadují předplatné Wolfram Alpha Pro. Dostupný interaktivně online, bez potřeby instalace.

2) Použití pro statistické výpočty:

- **R:** Pokrývá širokou škálu statistických výpočtů, od základních po pokročilé metody (na co si člověk vzpomene).
- **Wolfram Alpha:** Umožňuje základní statistické výpočty, rychlé a snadné použití, vhodné pro rychlé dotazy.

3) Šířka záběru:

- **R:** Zaměřeno hlavně na statistiku a analýzu dat. Lze rozšířit o balíčky pro různé oblasti (text mining, geografická data, strojové učení).
- **Wolfram Alpha:** Pokrývá širokou škálu oborů (matematika, další vědy, ekonomie), ale s omezenými možnostmi pro pokročilé statistické analýzy.

12.2.2 Základní příkazy ve Wolfram Alpha

Wolfram Alpha umožňuje provádět různé typy výpočtů, jako je výpočet průměru, směrodatné odchylky, rozptylu a mnoho dalších. Zde jsou některé základní příkazy, které lze zadat přímo

do vyhledávacího pole Wolfram Alpha:

- **Mean of {data}** – vypočítá průměr datové sady.

  `Mean of {3, 5, 2, 7, 6, 8, 4, 7, 9, 5}` → `5.6`

- **Standard deviation of {data}** – vypočítá směrodatnou odchylku datové sady.

  `Standard deviation of {3, 5, 2, 7, 6, 8, 4, 7, 9, 5}` → `2.059`

- **Variance of {data}** – vypočítá rozptyl datové sady.

  `Variance of {3, 5, 2, 7, 6, 8, 4, 7, 9, 5}` → `4.24`

- **Correlation between {data1} and {data2}** – vypočítá korelační koeficient mezi dvěma sadami dat.

  `Correlation between {3, 5, 2} and {7, 8, 4}` → `0.866`

Po zadání do vyhledávače Wolfram Alpha systém automaticky provede výpočet. Výsledky jsou doplněny o další související informace, jako jsou grafy nebo dodatečné statistické hodnoty.

Ilustrativní příklady

Příklad 12.2 (Regresní analýza ve Wolfram Alpha)**.** Zadejte `linear regression of {(1,2), (2,3), (3,5)}`.

Řešení: Po zadání Wolfram Alpha vypočítá regresní přímku ve tvaru $y = ax + b$, kde a je směrnice a b průsečík.

Výstup: $y = 1.5x + 0.5$

Wolfram Alpha rovněž poskytne graf a hodnotu koeficientu determinace (R^2), což je užitečné pro hodnocení kvality modelu. □

Příklad 12.3. Vyzkoušejte ve Wolfram Alpha následující příkazy a prozkoumejte jejich výstupy:

- {10, 12, 8, 14, 11, 9, 15, 13}
- five number summary {20, 25, 18, 30, 22, 19, 28, 30, 24}
- variance {20, 25, 18, 30, 22, 19, 28, 30, 24}
- median {20, 25, 18, 30, 22, 19, 28, 30, 24}
- poisson distribution
- normal distribution, mean=0, sd=2

- Student t, 17 degrees of freedom

Wolfram Alpha nám poskytuje okamžité výsledky, které lze použít pro další analýzu nebo kontrolu správnosti našich výpočtů. V následující sekci se podíváme na informativní přehled o využití softwaru R.

12.2.3 Použití R pro statistické úlohy

R je volně dostupný programovací jazyk zaměřený na statistické výpočty a datovou analýzu. I když jeho využití není v tomto kurzu klíčové, stojí za to jej zmínit jako výkonný nástroj pro složitější úlohy, které mohou být mimo možnosti Excelu nebo Wolfram Alpha. V této části si ukážeme několik základních funkcí v R, které se používají pro statistické úlohy, a to spíše informativně, bez nutnosti provádět výpočty během výuky.

Základní příkazy v R pro statistické výpočty

R nabízí širokou škálu funkcí, které jsou velmi užitečné při řešení statistických úloh. Zde je přehled některých základních příkazů:

- `mean()` – vypočítá průměr zadaných dat.

  `mean(c(3, 5, 2, 7, 6, 8, 4, 7, 9, 5))` → 5,6.

- `sd()` – vypočítá výběrovou směrodatnou odchylku zadaných dat.

  `sd(c(3, 5, 2, 7, 6, 8, 4, 7, 9, 5))` → 2,22.

- `var()` – vypočítá výběrový rozptyl zadaných dat.

  `var(c(3, 5, 2, 7, 6, 8, 4, 7, 9, 5))` → 4,93.

- `cor()` – vypočítá korelační koeficient mezi dvěma sadami dat.

  `cor(c(3, 5, 2), c(7, 8, 4))` → 0,891.

- `lm()` – provádí lineární regresi.

  ```
  lm(y ~x, data = dataframe)
  ```

  Tato funkce provede lineární regresní analýzu mezi proměnnými `x` a `y` v datovém rámci `dataframe`.

Výhody a nevýhody R

Výhody:

- R je zdarma a otevřený software, který je snadno dostupný.
- Nabízí širokou škálu funkcí a knihoven pro různé statistické metody, od jednoduchých výpočtů po složité modelování.
- Je vhodný pro analýzu velkých datových sad, které by byly v Excelu obtížně zpracovatelné.
- Možnost vytvářet pokročilé vizualizace a grafy přímo z dat (pomocí programového kódu).

Nevýhody:

- R vyžaduje určitou znalost programování, což může být pro začínající studenty obtížné. Ovšem tuto nevýhodu lze do značné míry potlačit s asistencí AI.
- Pro mnoho uživatelů je Excel jednodušší a intuitivnější, zejména pro menší a jednodušší úlohy.

Ilustrativní příklad

Příklad 12.4. Zvažte následující data o cenách produktů v obchodech: {10, 12, 8, 14, 11, 9, 15, 13}. Pomocí R vypočítejte průměr, směrodatnou odchylku a rozptyl. Napište příkazy a uveďte, co každý z nich dělá.

Řešení:
- Průměr: `mean(c(10, 12, 8, 14, 11, 9, 15, 13))` = 11,5.
- Směrodatná odchylka (výběrová): `sd(c(10, 12, 8, 14, 11, 9, 15, 13))` = 2,44.
- Rozptyl (výběrový): `var(c(10, 12, 8, 14, 11, 9, 15, 13))` = 6.

□

12.3 Analýza dat z externích zdrojů

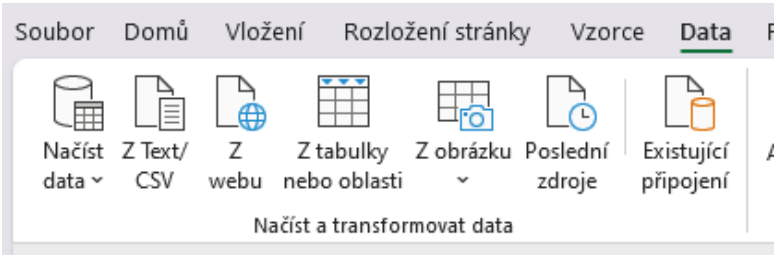
V této sekci se zaměříme na příklady rozsáhlejších statistických úloh, které zahrnují stahování dat z internetu, jejich zpracování v Excelu, grafické znázornění a následné výpočty popisných statistik a korelace. Zaměříme se na reálná data z ČNB (kurzy měn) a akciových trhů.

Kde hledat statistická data na internetu?

Existuje mnoho dostupných zdrojů, ze kterých lze stahovat reálná statistická data. Klasicky ve formě souborů, například ve formátu csv, nebo přímým napojením. Mezi ty české patří například Český statistický úřad (czso.cz) a ČNB (cnb.cz). Z těch zahraničních například Eurostat (ec.europa.eu/eurostat) a Světová banka (data.worldbank.org), případně Yahoo Finance (finance.yahoo.com) a Google Finance (google.com/finance).

Načítání dat z vnějších zdrojů do Excelu

V Excelu existuje několik možností, jak načítat a transformovat data z různých externích zdrojů. Tyto možnosti umožňují zpracovávat data nejen ze souborů na lokálním disku, ale také z online zdrojů s aktuálními informacemi. Mezi základní možnosti patří (viz obrázek 25:



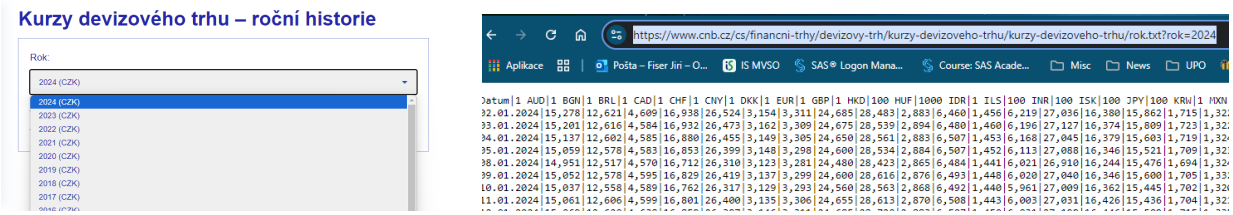
Obr. 25: Excel: Skupina Načíst a transformovat data na kartě Data

- **Načítání z Text/CSV**
  Pomocí této funkce lze načíst data z textových souborů (.txt) nebo souborů CSV (.csv). Jedná se o jednoduchý způsob, jak dostat strukturovaná data do Excelu.
- **Načítání z webu**
  Tato možnost umožňuje přímé načtení dat z webové stránky. Excel si z webu stáhne tabulková data a umožní je dále zpracovávat. To je zvláště užitečné pro načítání kurzů měn, cen akcií nebo jiných finančních dat, která se pravidelně aktualizují.
- **Načítání z tabulky nebo oblasti**
  Tento nástroj umožňuje načítat data přímo z jiných tabulek v Excelu nebo z definovaných oblastí buněk. Hodí se při práci s velkými datovými sadami rozdělenými do více souborů.
- **Načítání z obrázku**
  Excel dokáže načítat data přímo z obrázků, což je užitečné pro digitalizaci dat v tištěných tabulkách nebo grafech. Stačí nahrát obrázek a Excel rozpozná strukturu dat.
- **Načítání z webových API a online zdrojů**
  Excel umožňuje načítání dat z online zdrojů pomocí webových API. Tato funkce je klíčová pro práci s aktuálními daty, například z finančních trhů, online databází nebo jiných služeb poskytujících aktualizované informace. Pomocí rozhraní API lze získat přístup k datům, která se pravidelně aktualizují, což je ideální pro tvorbu reportů nebo analýz založených na živých datech.
- **Poslední zdroje**
  V této části Excelu je možné rychle znovu načíst data z posledních použitých zdrojů. To usnadňuje opakované aktualizace dat z těchto zdrojů.
- **Existující připojení**
  Tato funkce umožňuje správu a opětovné využití dříve nastavených připojení k datovým zdrojům, jako jsou databáze, webové služby nebo další Excelové soubory.

Načítání dat z online zdrojů je pro analýzy v Excelu zásadní, zejména pokud pracujeme s dynamickými daty, která se často mění. Pomocí těchto nástrojů je možné zajistit, že naše tabulky budou obsahovat aktuální a relevantní informace pro daný účel.

Ilustrativní příklad

Příklad 12.5 (Načtení a analýza tabulky kurzů měn z ČNB). 1. Na stránkách ČNB najděte údaje „Kurzy devizového trhu - roční historie“ a vyberte rok 2024 (obrázek 26)[1] .



Obr. 26: ČNB: Kurzy devizového trhu - roční historie – zadání roku 2024

2. Zkopírujte odkaz a použijte jej v `Excelu -> Data -> (Načíst a transformovat data) Z webu` (obrázek 27).

	A	B	C	D	E	F	G
1	Datum	1 AUD	1 BGN	1 BRL	1 CAD	1 CHF	1 CN
2	02.01.2024	15,278	12,621	4,609	16,938	26,524	3,
3	03.01.2024	15,201	12,616	4,584	16,932	26,473	3,
4	04.01.2024	15,137	12,602	4,585	16,88	26,455	3,
5	05.01.2024	15,059	12,578	4,583	16,853	26,399	3,
6	08.01.2024	14,951	12,517	4,57	16,712	26,31	3,
7	09.01.2024	15,052	12,578	4,595	16,829	26,419	3,
8	10.01.2024	15,037	12,558	4,589	16,762	26,317	3,
9	11.01.2024	15,061	12,606	4,599	16,801	26,4	3,

Obr. 27: Načtení dat z ČNB do Excelu pomocí volby Data -> Z webu

3. Pomocí volby `Analýza dat -> Popisná statistika` vypočtěte popisné statistiky pro všechny měny (na zvláštní list).

4. Pomocí volby `Analýza dat -> Korelace` vypočtěte korelační koeficienty pro všechny dvojice měn (na zvláštní list).

5. Pomocí podmíněného formátování korelační koeficienty obarvěte podle velikosti. Zvlášť zvýrazněte hodnoty větší než 0,9. (obrázek 28).

6. Jak si vysvětlujete tak vysokou pozitivní lineární korelaci?

7. Vyberte jednu dvojici z předchozího bodu a vytvořte pro ni bodový graf.

[1] AUD – Australský dolar, BGN – Bulharské leva, BRL – Brazilský real, CAD – Kanadský dolar, CHF – Švýcarský frank, CNY – Čínský jüan, DKK – Dánská koruna, EUR – Euro, GBP – Britská libra, HKD – Hongkongský dolar, HUF – Maďarský forint (kurz za 100 jednotek), IDR – Indonéská rupie (kurz za 1000 jednotek), ILS – Izraelský nový šekel, INR – Indická rupie (kurz za 100 jednotek), ISK – Islandská koruna (kurz za 100 jednotek), JPY – Japonský jen (kurz za 100 jednotek), KRW – Jihokorejský won (kurz za 100 jednotek), MXN – Mexické peso, MYR – Malajsijský ringgit, NOK – Norská koruna, NZD – Novozélandský dolar, PHP – Filipínské peso (kurz za 100 jednotek), PLN – Polský zlotý, RON – Rumunský lei, SEK – Švédská koruna, SGD – Singapurský dolar, THB – Thajský baht (kurz za 100 jednotek), TRY – Turecká lira (kurz za 100 jednotek), USD – Americký dolar, XDR – Speciální práva čerpání (měna používaná MMF), ZAR – Jihoafrický rand.

	A	B	C	D	E	F	
1		1 AUD	1 BGN	1 BRL	1 CAD	1 CHF	1
2	1 AUD	1					
3	1 BGN	0,56662	1				
4	1 BRL	-0,26862	-0,03034	1			
5	1 CAD	0,27497	0,54962	0,70436	1		
6	1 CHF	0,1362	0,47057	-0,2287	0,11274	1	
7	1 CNY	0,489	0,83009	0,20362	0,71783	0,42913	
8	1 DKK	0,56791	0,99647	-0,00012	0,57281	0,46649	
9	1 EUR	0,57007	0,99703	-0,02245	0,55534	0,4615	
10	1 GBP	0,74454	0,80486	-0,45514	0,15803	0,49149	
11	1 HKD	0,45092	0,67855	0,30423	0,7807	0,02264	

Obr. 28: Podmíněné formát tabulky korelačních koeficientů

12.3.1 Excelovské nástroje pro analýzu akcií

Využití datového typu Akcie v Excelu

Datový typ **Akcie** umožňuje získávat aktuální finanční údaje o veřejně obchodovaných společnostech. Pro jeho použití stačí zadat **název společnosti** nebo její **ticker** (např. "AAPL"pro Apple) do buňky, následně zvolit z karty *Data* možnost *Akcie*. Excel poté poskytne aktuální údaje jako *cena*, *tržní kapitalizace*, *P/E ratio* atd., ale i samotný ticker. Tyto údaje se automaticky aktualizují (minimálně při každém otevření souboru).

Získaný ticker lze následně využít ve funkci **STOCKHISTORY** pro načtení historických dat obchodování dané akcie.

Použití funkce STOCKHISTORY

Syntaxe je následující:

```
=STOCKHISTORY("ticker"; "start_date"; "end_date"; [interval]; [headers];
[property0]; [property1]; ...)
```

Příklad použití pro načtení denních uzavíracích cen akcií Microsoftu za září 2024:

```
=STOCKHISTORY("MSFT"; "2024-09-01"; "2024-09-30"; 0; 1; 0; 5)
```

Tento vzorec vrátí tabulku obsahující data a uzavírací ceny pro každý obchodní den v uvedeném období. Funkce **STOCKHISTORY** je vhodná pro analýzu historických finančních dat a sledování časových řad.

Ilustrativní příklady

Příklad 12.6 (Analýza uzavíracích cen akcií firem NVIDIA a Intel)**.** Pomocí datového typu `Akcie` zjistěte tickery firem **NVIDIA** a **Intel**.

Pomocí funkce `STOCKHISTORY` načtěte uzavírací denní ceny jejich akcií v období od 1. srpna 2024 do 30. září 2024.

Tyto dvě časové řady graficky znázorněte, vypočtěte pro ně základní popisné statistiky a proveďte jejich korelační analýzu.

Řešení: **1. Tickery**

Nejprve získáme tickery společností NVIDIA a Intel pomocí datového typu **Akcie**:

- Do buněk vložíme názvy společností (NVIDIA, Intel).
- Označíme buňky s názvy a na kartě *Data* zvolíme možnost **Akcie**. Excel automaticky přiřadí k názvům společností jejich tickery.
- NVIDIA má ticker **NVDA**, Intel **INTC**.

2. Zisk historických uzavíracích cen

Pro získání denních uzavíracích cen akcií obou společností v období od 1. srpna 2024 do 30. září 2024 použijeme následující funkce:

```
=STOCKHISTORY("NVDA"; "2024-08-01"; "2024-09-30"; 0; 1; 0; 1)
=STOCKHISTORY("INTC"; "2024-08-01"; "2024-09-30"; 0; 1; 0; 1)
```

Experimujte s tímto zápisem tak, abyste získali tabulku o třech sloupcích: datum, ceny NVIDIA, ceny Intel. Tato funkce načte uzavírací ceny pro každý obchodní den v uvedeném období. Získané datové řady budou použity pro další analýzu.

3. Grafické znázornění časových řad

Po získání uzavíracích cen vytvoříme spojnicový graf, který vizuálně znázorní vývoj uzavíracích cen akcií NVIDIA a Intel:

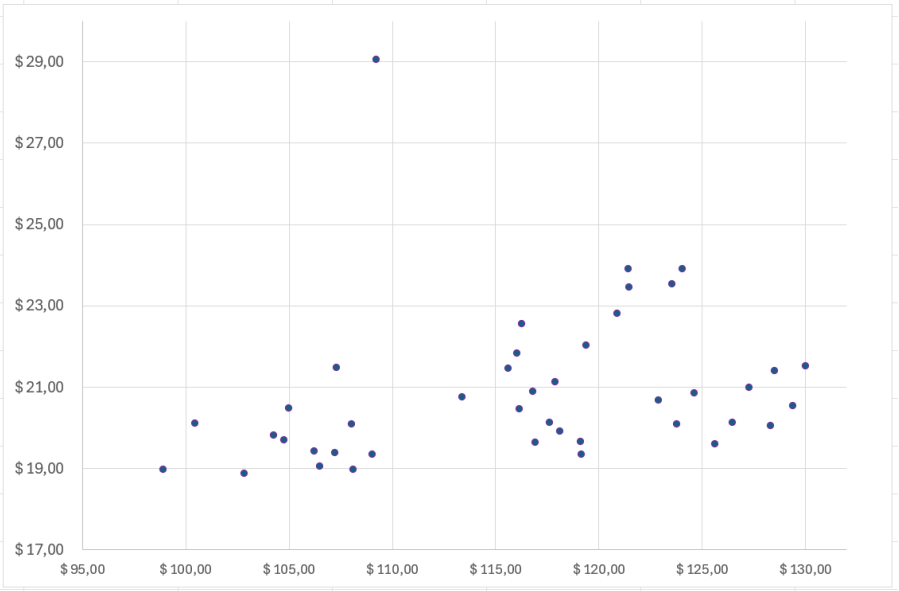
- Označíme sloupce s daty (datum, uzavírací ceny NVIDIA a Intel).
- Na kartě *Vložení* zvolíme typ grafu **Spojnicový graf**.
- Excel vygeneruje graf, který zobrazí vývoj cen akcií obou společností v průběhu sledovaného období.

4. Korelační analýza

Pro určení míry lineární závislosti mezi cenami akcií NVIDIA a Intel použijeme funkci **CORREL**. Vzorec pro výpočet korelačního koeficientu mezi dvěma časovými řadami uzavíracích cen je následující:

```
=CORREL(B2:B45, C2:C45)\approx 0{,}249.
```

Funkce vrátila korelační koeficient o hodnotě 0,249, který popisujeme jako slabou pozitivní korelaci. Mějme ale na paměti, že korelační koeficient popisuje jen lineární závislost, a tak je vždy užitečné si celkový obraz doplnit obrázkem. V tomto případě je bodový graf na obrázku 29. Můžeme na něm zaznamenat jednu odlehlou hodnotu (v takovém případě bychom měli prověřit, zda nejde o chybnou hodnotu, resp. zjistit, jak mohla nastat). Na obrázku je znatelný drobný nárůst vertikálních hodnot (souřadnic) při růstu horizontálních hodnot. Uvědomme si také, že v tomto typu grafu není zachycena časová složka dat.



Obr. 29: Bodový graf cen akcií NVIDIA (horizontální osa) a Intel (vertikální osa) z příkladu 12.6

□

Příklad 12.7 (Analýza maximálního rozdílu mezi maximálními a minimálními denními cenami)**.** Zvolte si tři firmy. Získejte jejich tickery a maximální a minimální denní ceny za jedno roční období, končící na konci předminulého měsíce (vzhledem ke dni, kdy příklad počítáte).

Následně pro každou akcii vypočtěte denní rozdíly mezi maximální a minimální cenou. Poté najděte pro každou firmu nejvyšší hodnotu těchto denních rozdílů (tzv. maximální denní rozpětí) a tyto tři hodnoty porovnejte.

Protože ceny akcií mohou být velmi rozdílné, je nutné výsledky porovnávat relativně. Nejprve pro každou akcii spočítejte tzv. *průměrnou denní cenu* jako průměr maximální a minimální ceny pro každý den. Z těchto průměrů vypočtěte jejich průměrnou hodnotu za celé období.

Nakonec relativně porovnejte maximální denní rozpětí s touto průměrnou cenou (v procentech). Toto procentuální vyjádření vám umožní porovnat, která akcie vykazuje největší cenové výkyvy vzhledem ke své průměrné ceně.

12.3.2 Načítání externích statistických dat v R

Ač Exel lze dobře použít pro import aktuálních finančních a dalších statistických dat, tak ten, kdo ovládá práci v R má situaci mnohem pohodlnější.

R nabízí několik balíčků, které usnadňují přímé načítání aktuálních statistických a finančních dat z externích zdrojů. Mezi nejpoužívanější patří

- `quantmod`, který umožňuje získávat data o cenách akcií, měnových kurzech a dalších finančních údajích z Yahoo Finance a FRED.
- Balíček `wbstats` poskytuje přístup k datům Světové banky, včetně ukazatelů inflace, HDP a dalších makroekonomických dat.
- Pro evropská data lze použít balíček `eurostat`, který umožňuje stahovat data o ekonomických a sociálních ukazatelích v rámci členských států EU.
- Kromě toho balíček `fredr` poskytuje přístup k bohaté databázi ekonomických ukazatelů FRED.

Tyto nástroje v R umožňují rychlé a efektivní načítání aktuálních dat pro další analýzu. Samozřejmě, samostatná data nestačí, je třeba nejprve nastudovat jejich strukturu, označení a význam.

V této kapitole jsme se věnovali statistické analýze z pohledu použitého softwaru, přirozeně s největším důrazem na MS Excel, ale prošli jsme i možnosti Wolfram Alpha a R. Zaměřili jsme se na výpočty základních statistik, korelační analýzu a tvorbu grafických výstupů. Ukázali jsme také, jakým způsobem lze data načítat do Excelu z externích zdrojů a jak je následně zpracovat.

Wolfram Alpha byl představen jako jednoduchý nástroj pro rychlé výpočty pravděpodobností a dalších základních statistických úloh, kdy není třeba složitého programování.

R bylo popsáno jako pokročilý nástroj pro statistickou analýzu, který je vhodný pro práci s rozsáhlými datovými soubory, jejich vizualizaci a modelování, a umožňuje přímé načítání externích dat z různých statistických zdrojů, jako jsou například Světová banka nebo Eurostat.

1. Jaké zdroje lze využít pro stahování statistických dat z internetu?
2. Jaké jsou základní kroky pro načtení externích dat do Excelu?
3. Popište postup pro vytvoření grafu časových řad v Excelu.
4. Jaké funkce v Excelu použijete pro výpočet průměru, mediánu a směrodatné odchylky?
5. Co je Pearsonův korelační koeficient a jak se v Excelu vypočítá?
6. Kdy je vhodné použít Wolfram Alpha pro statistické výpočty? Uveďte příklady.
7. Jakým způsobem lze analyzovat a znázornit data z akciových trhů?

8. Stáhněte data o inflaci z webu Českého statistického úřadu (`https://www.czso.cz`) za posledních 10 let. Načtěte tato data do Excelu, analyzujte je pomocí grafu časové řady a vypočítejte základní statistiky (průměr, medián, směrodatná odchylka, minimum, maximum).
9. Získejte data o cenách akcií tří ropných společností za tři roky (začátek a konec si zvolte sami) pomocí funkce STOCKHISTORY. Vypočtěte jejich popisné statistiky. Vytvořte graf s těmito třemi časovými řadami. Proveďte jejich korelační analýzu včetně bodových grafů. Komentujte výsledky (největší podobnosti a rozdíly).

Literatura k tématu:

[1] PRAŽSKÁ BURZA CENNÝCH PAPÍRŮ. Dostupné z: `https://www.pse.cz/`.

[2] YAHOO FINANCE. Dostupné z: `https://finance.yahoo.com/`.

[3] MICROSOFT EXCEL. Podpora pro statistické funkce. Dostupné z: `https://support.microsoft.com/excel`.

[4] WOLFRAM ALPHA. Online nástroj pro výpočty. Dostupné z: `https://www.wolframalpha.com/`.

[5] R CORE TEAM. (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Dostupné z: `https://cran.r-project.org/manuals.html`.

[6] ČESKÁ NÁRODNÍ BANKA (ČNB) - Data. Česká národní banka. (2023). Data a statistiky. Dostupné z: `https://www.cnb.cz/cs/statistika/`.

[7] EUROSTAT. Statistiky Evropské unie. Dostupné z: `https://ec.europa.eu/eurostat`.

[8] SVĚTOVÁ BANKA. (2023). Data Světové banky. Dostupné z: `https://data.worldbank.org/`.

[9] ČESKÝ STATISTICKÝ ÚŘAD (ČSÚ). Data a statistiky České republiky. Dostupné z: `https://www.czso.cz/`.

Seznam literatury a použitých zdrojů

[1] ANDĚL, J. Statistické metody. 5. vyd. Praha: Matfyzpress, 2019. ISBN 978-80-7378-381-5.

[2] CALDA, E., DUPAČ, V. (2008). Matematika pro gymnázia: Kombinatorika, pravděpodobnost, statistika (5. vydání, dotisk 2011). Praha: Prometheus. ISBN 978-80-7196-365-3.

[3] HANSEN, B. Probability and Statistics for Economists. Princeton University Press, 2022. ISBN 9780691236148.

[4] HENDL, J. Základy matematiky, logiky a statistiky pro sociologii a ostatní společenské vědy v příkladech. 3. vyd., Karolinum, 20232. ISBN 978-80-246-5400-3.

[5] HINDLS, R. Statistika pro ekonomy. 8. vyd. Praha: Professional Publishing, 2007. ISBN 978-80-869-4643-6.

[6] HONG, Y. Probability and Statistics for Economists. World Scientific, 2017. ISBN 9789813228818.

[7] JANÁČEK, J. Statistika jednoduše. Grada, 2022. ISBN 978-80-271-1738-3.

[8] KELLER, G. Statistics for Management and Economics. 12th ed., Cengage Learning, 2022. ISBN 9780357714393.

[9] MAREK, L. Statistika v příkladech. 2. vyd. Praha: Kamil Mařík – Professional Publishing, 2015. ISBN 978-80-743-1153-6.

[10] NEUBAUER, J. a SEDLAČÍK, M. Základy statistiky: Aplikace v technických a ekonomických oborech - 3., rozšířené vydání. Grada, 2021. ISBN 978-80-271-3421-2.

[11] OPENAI. Asistovaná příprava studijní opory pomocí ChatGPT. OpenAI. Dostupné na `https://chat.openai.com`, 2024.

[12] OTIPKA, P., ŠMAJSTRLA, V. Pravděpodobnost a statistika [online]. 1. vydání. Ostrava: VŠB-TU Ostrava, 2007 [cit. 2024-09-09]. ISBN 80-248-1194-4.

[13] ŘEZANKOVÁ, H. a kol. Úvod do statistiky. 2. dotisk 1. vyd., Oeconomica, nakladatelství VŠE, 2019. ISBN 9788024523019.

[14] ZVÁRA, K. a ŠTĚPÁN, J. Pravděpodobnost a matematická statistika. Matfyzpress, 2019. ISBN 978-80-7378-388-4.

Seznam obrázků

Seznam tabulek