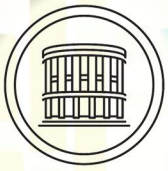
Eva Capková

Analýza dát pre mediamatikov
časť I.



UNIVERZITA
KOMENSKÉHO
V BRATISLAVE

Eva Capková

Analýza dát pre mediamatikov

Časť I.

2025

Univerzita Komenského v Bratislave

Publikácia je financovaná z grantu „Kreatívna analýza dát: komplexná platforma pre inováciu výučby predmetu analýza dát s využitím multimediálnych technológií“ KEGA 069UK-4/2023

Univerzita Komenského v Bratislave, Fakulta sociálnych a ekonomických vied,
Ústav mediamatiky

Recenzenti
Ivana Pobočíková
Katedra aplikovanej matematiky, Žilinská univerzita v Žiline
Naďa Krivoňáková
Ústav informatizácie, automatizácie a matematiky, Slovenská technická univerzita v Bratislave

Ilustrácia na obálke
Juraj Grečnár

https://stella.uniba.sk/texty/FSEV/EC_mediamatika_analyza_dat_1.pdf

Vydavateľ
Univerzita Komenského v Bratislave

ISBN 978-80-223-6235-1 (online)

Predslov

Táto učebnica je určená predovšetkým pre študentov v študijnom programe mediamatika ako pomôcka na zvládnutie povinného predmetu Analýza dát, ktorý je v odporúčanom študijnom pláne zaradený do 2. semestra štúdia, ale môže byť nápomocná pre študentov aj iných študijných programov predovšetkým v študijnom odbore mediálne a komunikačné štúdiá, ako aj v iných spoločensko vedných odboroch. Ide o predmet, ktorý dlhodobo figuruje na popredných priečkach v náročnosti pre študentov a táto učebnica si kladie za cieľ pomôcť pochopeniu základných princípov analýzy dát tak, aby sa objavovanie zákonitostí stalo viac zábavnou detektívkou ako nudným strašiakom.

Učebnica je koncipovaná ako pomôcka, ktorá je súčasťou portálu s názvom DATAINAK. Portál vznikol ako výsledok projektu KEGA č. 069UK-4/2023 s názvom Kreatívna analýza dát: komplexná platforma pre inováciu výučby predmetu Analýza dát s využitím multimediálnych technológií, na ktorom sa okrem tejto učebnice nachádzajú prednášky k predmetu, cvičné testy, kahoot kvízy a ukážky študentských prác.

Táto učebnica je prvou časťou väčšieho celku, v ktorej sú v troch kapitolách podrobne rozobrané základy nevyhnutné pre zvládnutie náročnejších tém. V prvej kapitole je dôraz kladený na motivovanie čitateľa tým, že má možnosť zoznámiť sa s množstvom aplikácií analýzy dát v mediamatickej praxi, konkrétne v oblasti online obchodovania, dizajnovania webových stránok, sociálnych sietí a ďalších. Druhá kapitola sa venuje rôznym typom dát a premenných, s ktorými sa pri analýze stretávame. Opísané sú aj metódy náhodného výberu vzorky, ktorý je nutnou podmienkou pre správnu interpretáciu výsledkov získaných z neskoršej analýzy. V poslednej kapitole prvej časti učebnice Analýza dát pre mediamatikov sú predstavené základy teórie pravdepodobnosti, bez ktorých sa žiadne štatistické metódy nezaobídu.

Ako pracovať s online učebnicou?

Okrem samotného textu sa v učebnici nachádzajú aj **príklady**. Ide o príklady, úlohy, či otázky, nad ktorými sa oplatí zamyslieť v kontexte predchádzajúceho textu a odpovedaním na otázky tak môžete lepšie pochopiť teoretické koncepty. Odpovede na otázky z týchto príkladov nájdete v poznámkach pod čiarou**.** Na konci podkapitol sa nachádzajú **cvičenia,** na ktorých si môžete precvičiť, či ste teoretické základy pochopili správne. Riešenie cvičení sa nachádza na webe DATAINAK v multimediálnej forme.

Obsah

Zoznam obrázkov a tabuliek

1. ÚVOD DO ANALÝZY DÁT

Úlohou analýzy dát je systematicky spracúvať, skúmať a interpretovať dáta s cieľom získať spoľahlivé informácie, ktoré umožňujú porozumieť sledovaným javom, overovať hypotézy, odhaľovať vzorce a podporovať kvalifikované rozhodovanie. Inak povedané analýza dát znamená pozerať sa na dáta tak, aby sme z nich dostali zrozumiteľné informácie. Pomáha nám zistiť, čo sa deje, prečo sa to deje a ako sa môžeme lepšie rozhodovať.

Analýza dát v mediamatike znamená zbieranie a skúmanie údajov napríklad o tom, ako ľudia používajú médiá, ako funguje komunikácia, pomáha zistiť, čo publikum zaujíma, ktoré články alebo kampane sú úspešnejšie, prípadne aký majú médiá vplyv na spoločnosť.

Prvá kapitola sa venuje úvodnému predstaveniu problematiky analýzy dát, pričom sú opísané konkrétne situácie, s ktorými je možné sa stretnúť v reálnej mediamatickej praxi. Pozornosť je venovaná aj experimentu, ako jednej z výskumných metód, ktorá má svoje uplatnenie napr. v podobe A/B testovania aj mimo laboratórnych podmienok. Prvá kapitola ďalej predstavuje aj „surovinu“, s ktorou budeme pracovať: dáta a premenné a ich typológiu. Dôležitou súčasťou prvej kapitoly je opis výberu vzorky z populácie, pričom sa zameriavame na opis rôznych typov náhodného výberu, ktorý je nutným predpokladom pre správnu interpretáciu získaných výsledkov.

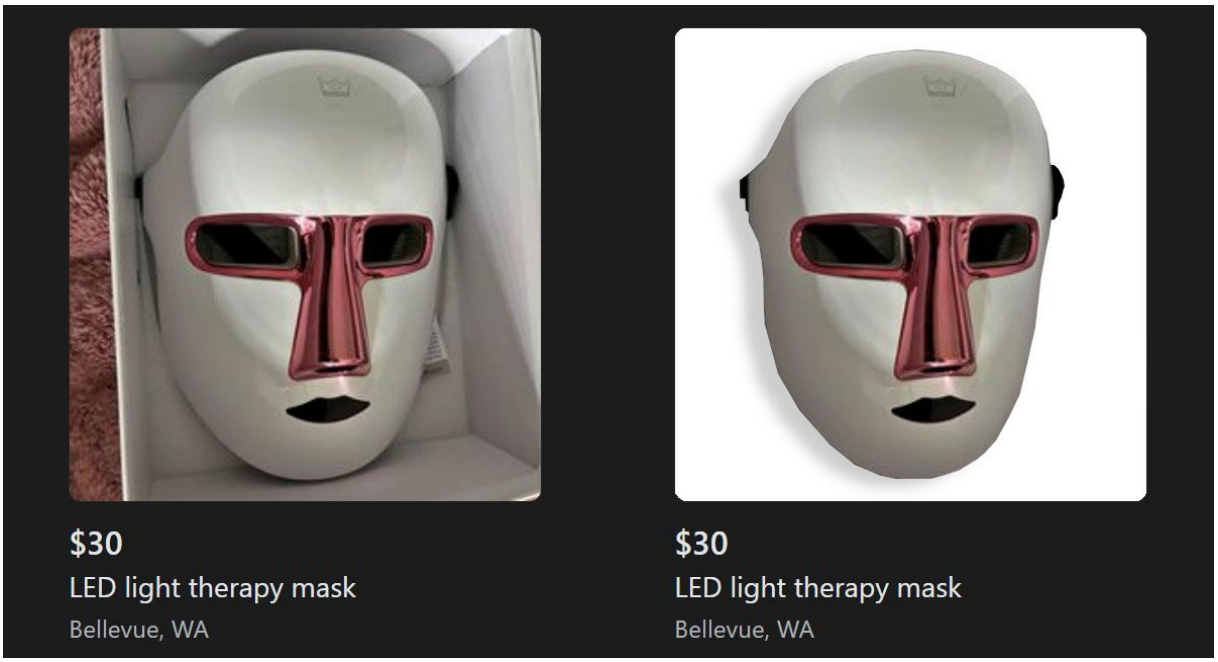
1.1. Analýza dát v praxi

V časti 1.1 sú predstavené konkrétne aplikácie analýzy dát v mediamatickej praxi, aby čitateľ získal lepšiu predstavu o tom, kde všade môže analýza dát pomôcť a s akými výsledkami. Čitateľ sa zoznámi aj s riadeným (kontrolovaným) experimentom, tzv. A/B testom a experimentom, ktorý bol použitý v medicínskej praxi a kde si jeho využitie vieme všetci dobre predstaviť.

Napriek tomu, že na prvý pohľad sa môže javiť využitie analýzy dát v mediamatickom kontexte len ako okrajová záležitosť a že mediamatik alebo mediamatička sa bez analýzy dát spokojne

zaobídu, opak je pravdou. Analýza dát je z pohľadu mediamatiky predovšetkým o pochopení interakcií medzi médiami, technológiami a ľuďmi – či už z obchodného, kultúrneho alebo spoločenského hľadiska. Preto je využiteľná pri tvorbe a optimalizácii mediálnych obsahov napríklad v zmysle analýzy preferencií publika (sledovanie aké typy článkov, videí, podcastov alebo grafických prvkov majú väčší dosah a odozvu), v marketingu zasa poslúži pri segmentácii a poznaní publika (rozdelenie používateľov podľa demografie, správania, záujmov), či optimalizácii mediálnych rozpočtov – kde sa oplatí investovať viac (sociálne siete, bannery, video reklamy). Nezastupiteľné miesto má aj v UX dizajne a analýze používateľského správania, kedy pomocou A/B testov rozhodneme, ktoré verzie dizajnu vedú k lepším výsledkom z pohľadu použiteľnosti, či ekonomickej návratnosti.

Jedným zo zaujímavých príkladov uplatnenia analýzy dát v mediamatickom kontexte je situácia z roku 2020 v spoločnosti Facebook (teraz Meta), keď sa spoločnosť snažila presadiť elektronický obchod na Marketplace, ktorý vnímali ako prirodzené rozšírenie svojho ekosystému online nakupovania. Spoločnosť si najala produktových lídrov z gigantov ako eBay, Walmart a Amazon a v tom čase sa iniciatíva zdala byť predurčená na úspech. Tímy neúnavne pracovali na vytvorení elegantného, moderného rozhrania s plynulým procesom platby, porovnávaním cien a silnou politikou vrátenia tovaru. Všetko týkajúce sa realizácie bolo dobre premyslené. Odborníci vykonali prvý riadený experiment, tzv, A/B test a ten vykreslil úplne iný obraz. Všetky kľúčové ukazovatele zaznamenali pokles – konverzné pomery, miery preklikov, zobrazenia pri prehliadaní a dokonca aj denný počet aktívnych používateľov. Používatelia sa nezapájali do e-commerce ponúk tak, ako to robili v prípade tradičných ponúk na Marketplace. Tento neúspech bol šokujúci. Vedúci pracovníci prirodzene spochybnili realizáciu – možno bolo niečo v neporiadku s UX, výberom produktov alebo cenovou stratégiou. Následný experiment však poskytol nepopierateľný dôkaz, že samotná myšlienka bola chybná. Druhý A/B test sa zameral na jednu jednoduchú zmenu dizajnu: pozadie obrázkov produktov (Obrázok 1). Tradičné ponuky na Marketplace obsahovali kontextové pozadia – predmety umiestnené na stoloch, pohovkách alebo vonkajších priestoroch. Ponuky v e-shope mali čisté, profesionálne, biele pozadia – rovnako ako Amazon alebo Walmart, (Sun, 2020).



Obrázok 1 Pozadie obrázkov produktov: vľavo kontextové pozadie, vpravo čisté pozadie (Sun, 2020)

V rámci experimentu sa používateľom náhodne zobrazovala buď verzia s bielym pozadím alebo verzia s kontextovým pozadím. Výsledok bol taký, že biele pozadie výrazne znížilo záujem a konverzie. Používatelia si ho spájali s bežným zážitkom z elektronického obchodu, nie s hľadaním výhodných ponúk a peer-to-peer charakterom Marketplace. Tento jeden A/B test zmenil všetko. Namiesto propagovania elektronického obchodovania sa Facebook Marketplace zameral na dopravu pre lokálne transakcie medzi jednotlivcami, čo viedlo k 26-násobnému nárastu online transakcií v priebehu nasledujúcich šiestich mesiacov. (Sun, 2020)

Druhým príkladom systematického využitia analýzy dát je proces A/B testovania v spoločnosti Booking,com. V súčasnosti je spoločnosťou vykonávaných viac ako 1 000 súbežných experimentov s rôznymi produktmi a cieľovými skupinami, čo jej umožňuje rýchlo overovať nápady a implementácie a zároveň zhromažďovať poznatky o správaní zákazníkov. Jedným zo spôsobov je A/B testovanie, keď rozdelia cieľovú skupinu na dve časti. Jedna polovica si ponechá aktuálnu verziu produktu, zatiaľ čo druhá polovica uvidí aktualizovanú verziu. Porovnávajú správanie používateľov a ak sa používatelia novej verzie správajú podľa očakávaní, tak

experiment fungoval podľa predpokladov a zmeny sú zavedené pre všetkých ostatných. Riaditeľ oddelenia experimentovania v spoločnosti Booking,com Lukas Vermeer uvádza:

> „Neexperimentujeme preto, že radi robíme experimenty, ale preto, že experimentovanie je skvelý spôsob, ako sa uistiť, že keď si myslíme, že niečo opravujeme, skutočne to opravujeme. Zmeny sú neustále, musíme neustále aktualizovať naše produkty, aby boli lepšie, ale musíme sa tiež uistiť, že tieto zmeny skutočne fungujú.“ (Vermeer, 2019)

Spoločnosťou Booking.com sa inšpiroval aj slovenský portál Profesia.sk, ktorý je na Slovensku známy ako líder v online ponuke pracovných miest. Ich ambíciou je ponúkať naozaj kvalitné online služby a neustále ich zlepšovať. Webová stránka je orientovaná na návštevníkov a jej súčasná podoba je výsledkom kontinuálneho procesu optimalizácie. V Profesii funguje tzv. CRO (conversion rate optimalization) tím, ktorý má na starosti vylepšovanie konverzií (cieľov) na stránke, ktorého súčasťou je aj A/B testovanie. V prípade Profesie je jedna z najdôležitejších konverzií počet reakcií na pracovnú ponuku. Testujú sa rôzne variácie textov a štylistiky, skúšajú sa rôzne vizuálne prvky, call to action tlačidlá či „vyskakovacie“ okná. To všetko má vplyv na to, ako sa používatelia na stránke správajú. (Vidová, 2018)

Ďalším príkladom spojenia mediamatickej praxe s vedeckými postupmi pri analýze dát je americký mediálny vydavateľ Upworthy.com, ktorý vykonával náhodné testovanie každého článku, ktorý publikoval. Každý experiment testoval variácie v „balíku“ nadpisov a obrázkov a zaznamenával, koľko náhodne vybraných divákov si vybralo každú variáciu. Hoci žiadny z týchto testov nebol navrhnutý tak, aby odpovedal na vedecké otázky, vedci môžu posunúť poznatky vpred prostredníctvom dátovej analýzy desiatok tisíc experimentov, ktoré Upworthy vykonal. Tento archív zaznamenával podnety a výsledky každého A/B testu, ktorý Upworthy vykonal v období od 24. januára 2013 do 30. apríla 2015. Archív obsahuje celkovo 32 487 experimentov, 150 817 experimentálnych ramien a 538 272 878 priradení účastníkov a je dostupný pre vedeckú obec. (Matias, a iní, 2021)

1.1.1. Ukážka použitia analýzy dát z medicínskeho prostredia

Použitie analýzy dát a špeciálne experimentu, ako výskumnej metódy je známe z medicíny a preto v učebnici uvádzame príklad práve z tohto prostredia. Išlo o experiment, ktorý skúmal účinnosť stentov pri liečbe pacientov s rizikom mozgovej príhody. Stenty sú zariadenia vložené do ciev, ktoré pomáhajú pri zotavovaní pacientov po srdcových príhodách a znižujú riziko ďalšieho infarktu alebo úmrtia. Mnohí lekári dúfali, že podobné výhody budú mať aj pacienti s rizikom mozgovej príhody. Začneme napísaním hlavnej výskumnej otázky, na ktorú výskumníci chcú získať odpoveď: „Znižuje používanie stentov riziko mozgovej príhody?“
Vedci, ktorí si položili túto otázku, uskutočnili experiment so 451 rizikovými pacientmi. Každý dobrovoľný pacient bol náhodne zaradený do jednej z dvoch skupín:

- **liečebná (alebo experimentálna) skupina,** v ktorej pacienti dostávali stent a medikamentóznu liečbu. Lekársky manažment zahŕňal lieky, manažment rizikových faktorov a pomoc pri úprave životného štýlu;
- **kontrolná skupina,** v ktorej pacienti dostávali rovnaký lekársky manažment ako pacienti v liečebnej skupine, ale nedostávali stenty.

Výskumníci náhodne zaradili všetkých 451 pacientov do skupín nasledovne: 224 pacientov do liečebnej skupiny a 227 do kontrolnej skupiny. V tejto štúdii poskytuje kontrolná skupina referenčný bod, na základe ktorého môžeme merať medicínsky vplyv stentov v liečebnej skupine. Výskumníci skúmali účinok stentov v dvoch časových bodoch: 30 dní po zaradení a 365 dní po zaradení. Výsledky u 5 konkrétnych pacientov sú zhrnuté v Tabuľka 1. Výsledky pacientov sú zaznamenané ako "cievna mozgová príhoda" alebo "bez príhody", čo predstavuje, či pacient na konci časového obdobia mal alebo nemal cievnu mozgovú príhodu.

Tabuľka 1 Výsledok experimentu pre 5 pacientov

Pacient	Skupina	0-30 dní	0-365 dní
1	liečebná	bez príhody	bez príhody
2	liečebná	cievna mozgová príhoda	cievna mozgová príhoda
3	liečebná	bez príhody	bez príhody
...	...	...	...

450	kontrolná	bez príhody	bez príhody
451	kontrolná	bez príhody	bez príhody

Zohľadnenie údajov od každého pacienta osobitne sa javí zdĺhavou a ťažkopádnou cestou k zodpovedaniu pôvodnej výskumnej otázky. Namiesto toho nám vykonanie analýzy dát umožňuje zohľadniť všetky údaje naraz. V Tabuľka 2 sú zaznamenané údaje užitočnejším spôsobom. V tejto tabuľke môžeme rýchlo vidieť, čo sa dialo počas celého výskumu. Napríklad, ak chceme zistiť počet pacientov v liečenej skupine. ktorí mali do 30 dní od začiatku experimentu cievnu mozgovú príhodu, pozrieme sa na ľavej strane tabuľky na priesečník liečebnej skupiny a cievnej mozgovej príhody:

Tabuľka 2 Počty pacientov počas výskumu v oboch skupinách

Skupina	**0-30 dní**		**0-365 dní**	
	cievna mozgová príhoda	bez príhody	cievna mozgová príhoda	bez príhody
liečebná	33	191	45	179
kontrolná	13	214	28	199
Spolu	46	405	73	378

Príklad 1.1

Z 224 pacientov v liečebnej skupine malo do konca prvého roka 45 cievnu mozgovú príhodu. Vypočítajte, koľko percent pacientov v liečebnej skupine malo do konca roka cievnu mozgovú príhodu.[1]

Z tabuľky môžeme vypočítať aj základné štatistické charakteristiky, ktoré nazývame aj opisné štatistiky. **Opisná štatistika je jedno číslo, ktoré sumarizuje veľké množstvo údajov.** Napríklad primárne výsledky štúdie po prvom roku by mohli byť opísané dvoma opisnými štatistikami: percentuálny podiel ľudí, ktorí mali cievnu mozgovú príhodu v liečenej a kontrolnej skupine:

[1] Liečebná skupina obsahuje 224 pacientov a z nich 45 prekonalo počas prvých 365 dní cievnu mozgovú príhodu. Preto percentuálny podiel je 45/224 = 0,20 = 20%.

Percentuálny podiel pacientov s cievnou mozgovou príhodou v liečebnej skupine: 45/224= 0,20 = **20%.**

Percentuálny podiel osôb, ktoré mali mozgovú príhodu v kontrolnej skupine: 28/227 = 0,12 = **12%.**

Tieto dve opisné štatistiky sú užitočné pri hľadaní rozdielov v skupinách a čaká nás prekvapenie: až o 8% viac pacientov v liečenej skupine malo počas prvého roka mozgovú príhodu! To je dôležité z dvoch dôvodov. Po prvé, je to v rozpore s tým, čo lekári očakávali, a to, že stenty znížia počet mozgových príhod, Po druhé, vedie to k štatistickej otázke: ukazujú údaje "skutočný" rozdiel medzi skupinami? Je možné, že 8 %-ný rozdiel v štúdii o stentoch je spôsobený prirodzenou odchýlkou? Čím väčší rozdiel však pozorujeme (pre konkrétnu veľkosť vzorky), tým menej je vierohodné, že rozdiel je spôsobený náhodou. V skutočnosti sa teda pýtame nasledovné: je rozdiel taký veľký, že by sme mali odmietnuť názor, že bol spôsobený náhodou?

Aj keď ešte nemáme k dispozícii štatistické nástroje, aby sme mohli túto otázku úplne vyriešiť sami, môžeme pochopiť závery publikovanej analýzy: v tejto štúdii pacientov s cievnou mozgovou príhodou boli presvedčivé dôkazy o tom, že stenty nielenže nepomáhajú, ale dokonca poškodzujú pacientov.

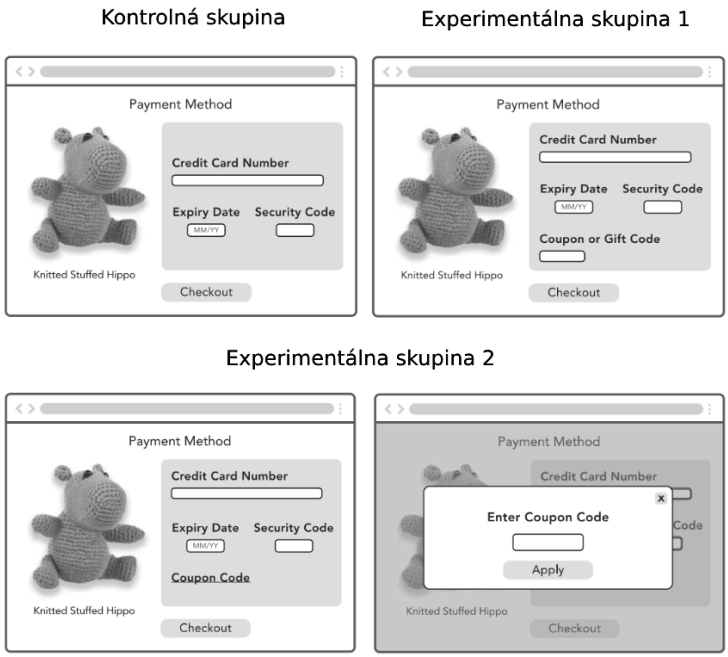
Buďte opatrní: Výsledky tejto štúdie nezovšeobecňujte na všetkých pacientov a všetky stenty. Táto štúdia sa zaoberala pacientmi s veľmi špecifickými charakteristikami, ktorí sa dobrovoľne zapojili do tejto štúdie a ktorí nemusia byť reprezentatívni pre všetkých pacientov s cievnou mozgovou príhodou. Okrem toho existuje mnoho typov stentov a v tejto štúdii sa posudzoval len jeden typ.

1.1.2. Ukážka použitia analýzy dát z prostredia online nakupovania

Naším konkrétnym príkladom je fiktívna online obchodná stránka, ktorá predáva háčkované hračky. Existuje široká škála zmien, ktoré môžeme na webovej stránke testovať: zavedenie novej funkcie, zmena používateľského rozhrania, zmena back-endu atď.

V našom príklade chce marketingové oddelenie zvýšiť predaj zasielaním propagačných e-mailov, ktoré obsahujú kupónový kód na zľavy na hračky. Táto zmena je potenciálnou zmenou obchodného modelu, pretože spoločnosť doteraz kupóny neponúkala. Zamestnanec spoločnosti

však nedávno čítal o tom, že istá spoločnosť utrpela významnú stratu príjmov po pridaní kupónového kódu a tiež sa dočítal, že *odstránenie* kupónových kódov je pozitívnym trendom. Vzhľadom na tieto externé údaje existujú obavy, že pridanie poľa pre kupónový kód na stránku platby zníži tržby, aj keď nebudú k dispozícii žiadne kupóny, t. j. len skutočnosť, že používatelia uvidia toto pole, ich spomalí a donúti ich hľadať kódy alebo dokonca opustiť stránku.
Chceme vyhodnotiť vplyv jednoduchého pridania poľa pre kupónový kód. Môžeme použiť prístup falošných dverí alebo namaľovaných dverí – analógia spočíva v tom, že postavíme falošné dvere alebo ich namaľujeme na stenu a sledujeme, koľko ľudí sa ich pokúsi otvoriť. V tomto prípade implementujeme triviálnu zmenu pridania poľa pre kupónový kód na stránku platby. Neimplementujeme skutočný systém kupónových kódov, pretože žiadne kódy nie sú k dispozícii. Nech už používateľ zadá čokoľvek, systém oznámi: „Neplatný kód kupónu“. Naším cieľom je jednoducho posúdiť vplyv na tržby pridaním tohto poľa pre kód kupónu a vyhodnotiť obavy, že to bude ľudí odrádzať od dokončenia nákupu. Keďže ide o jednoduchú zmenu, otestujeme dve implementácie používateľského rozhrania. Tento jednoduchý A/B test je kľúčovým krokom pri posudzovaní realizovateľnosti nového obchodného modelu.
V rámci nášho experimentu pridávame na stránku platby pole pre kód kupónu a testujeme dve rôzne používateľské rozhrania, ako je znázornené na **Chyba! Nenašiel sa žiaden zdroj odkazov.** a radi by sme vyhodnotili vplyv (ak nejaký bude) na tržby. Naša otázka znie: „Znižuje pridanie poľa pre kód kupónu tržby?“ alebo presnejšie : „Znižuje pridanie poľa pre kód kupónu na stránku platby tržby?“



Obrázok 2 Pôvodná stránka platby = kontrolná skupina (vľavo hore), vložené pole pre kupón = experimentálna skupina 1 (vpravo hore), vyskakovacie okno pre kupón = experimentálna skupina 2 (dole)

Jednoduchý A/B test bol navrhnutý tak, že používatelia boli náhodne rozdelení do troch skupín (Obrázok 2):

- **kontrolná skupina (34 % používateľov)** videla pôvodnú verziu stránky,
- **experimentálna skupina 1 (33 %používateľov)** videla variant 1, ktorý mala kupónové pole priamo na stránke pod údajmi o platbe a
- **experimentálna skupina 2 (zvyšných 33 % používateľov)** videla variant 2, v ktorom sa zobrazovalo kupónové pole vo forme vyskakovacieho okna.

Experiment prebiehal počas jedného týždňa, aby zachytil rôzne správanie počas pracovných dní aj víkendu, pričom cieľovou metrikou bol príjem na používateľa, ktorý začal proces nákupu. Po vyhodnotení sa ukázalo, že kontrolná skupina mala priemerný príjem 3,21 € na používateľa, zatiaľ čo priemerný príjem v experimentálnej skupine 1 dosiahol 3,12 € a v experimentálnej skupine 2 dokonca len 2,96 €. Rozdiel medzi experimentálnou skupinou 1 a kontrolnou skupinou je teda –0,09 € a rozdiel medzi experimentálnou skupinou 2 a kontrolnou skupinou je až – 0,25 €.

Príklad 1.2

Vypočítajte, aký relatívny pokles príjmu bol zaznamenaný v experimentálnej skupine 1 a aký v experimentálnej skupine 2. [2]

Relatívny pokles príjmu v experimentálnych skupinách si môžeme predstaviť tak, že ak by obchod zarobil za týždeň 100 € pri pôvodnom nastavení bez kupónov, tak pri variante 1 by príjmy klesli na úroveň 97,2 € (= 100 € – 2,8 €) a pri variante 2 až na 92,2 € (= 100 € – 7,8 €). Dôležitá je pre nás otázka, či je takýto pokles príjmu len výsledkom náhody a pri inej vzorke používateľov stránky by bol zaznamenaný vzostup, a nie pokles príjmu. V nasledujúcich kapitolách si predstavíme, ako zistiť nasledujúci záver A/B testovania: rozdiely boli štatisticky významné a analýza ukázala, že zavedením kupónov na zľavu by poklesli príjmy spoločnosti nech by už sa spoločnosť rozhodla pre ktorýkoľvek variant.

1.2. Dáta

Efektívna organizácia a opis údajov je prvým krokom pri väčšine analýz. V tejto časti je predstavená *dátová matica* na organizáciu údajov, ako aj terminológia o rôznych typoch dát, ktorá sa bude používať v tejto učebnici.

1.2.1. Pozorovania, premenné a matice údajov

V Tabuľka 3 sú zobrazené riadky 1, 2, 3 a 50 súboru údajov pre 50 náhodne vybraných úverov ponúkaných prostredníctvom spoločnosti Lending Club, ktorá sprostredkúva pôžičky medzi jednotlivcami. Tieto pozorovania sa budú označovať ako súbor údajov **pôžička50**.
Každý riadok v tabuľke predstavuje jeden úver. Formálny názov pre riadok je **prípad** alebo **jednotka pozorovania,** prípadne **štatistická jednotka**. Stĺpce predstavujú charakteristiky,

[2] Rozdiel príjmu medzi experimentálnou skupinou 1 a kontrolnou skupinou je 3,12 – 3,21 = – 0,09. Potrebujeme zistiť, aký podiel príjmu z kontrolnej skupiny (3,21 €) tvorí rozdiel 0,09 €: 0,09/3,21 ≈ 0,028 = 2,8%, Relatívny pokles príjmu pre experimentálnu skupinu 1 je preto –2,8%. Analogicky rozdiel príjmu medzi experimentálnou skupinou 2 a kontrolnou skupinou je 2,96 – 3,21 = – 0,25. Tento rozdiel (0,25 €) tvorí 7,8% príjmu z kontrolnej skupiny (3,21 €): 0,25/3,21 ≈ 0,078 = 7,8%. Relatívny pokles príjmu pre experimentálnu skupinu 2 je preto –7,8%.

nazývané **premenné** pre každý z úverov. Napríklad prvý riadok predstavuje úver 22 000 € s úrokovou sadzbou 10,90 %, pričom dlžník má sídlo v New Jersey (NJ) a príjem 59 000 €.

Príklad 1.3

Aký je stupeň prvej pôžičky v Tabuľka 3? Aký je stav vlastníctva domu dlžníka v prípade tohto prvého úveru?[3]

V praxi je obzvlášť dôležité klásť vysvetľujúce otázky, aby sa zabezpečilo, že sa pochopia dôležité aspekty údajov. Napríklad je vždy dôležité uistiť sa, že vieme, čo jednotlivé premenné znamenajú a aké sú ich merné jednotky. Popisy premenných **pôžičky50** sú uvedené v Tabuľka 4.

Tabuľka 3 Riadky dátového súboru **pôžička50**

ID	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	22 000	10,90	60	B	NJ	59 000	prenájom
2	6 000	9,92	36	B	CA	60 000	prenájom
3	25 000	26,30	36	E	SC	75 000	hypotéka
…	…	…	…	…	…	…	…
50	15 000	6,08	36	A	TX	77 500	hypotéka

Tabuľka 4 Premenné a ich opis pre dátový súbor **pôžička50**

Premenná	Popis
loan_amount	Výška poskytnutej pôžičky v amerických dolároch,
interest_rate	Úroková sadzba na pôžičku, vyjadrená ako ročná percentuálna sadzba,
term	Dĺžka trvania pôžičky, vždy uvedená ako celé číslo mesiacov,

[3] Hodnotenie úveru je B a dlžník si prenajíma svoje bydlisko.

grade	Hodnotenie pôžičky (A až G), ktoré vyjadruje kvalitu a pravdepodobnosť splatenia,
state	Americký štát, v ktorom dlžník býva,
total_income	Celkový príjem dlžníka vrátane druhého príjmu, v amerických dolároch,
homeownership	Informácia o tom, či osoba vlastní bývanie, má hypotéku, alebo býva v prenájme,

Údaje v Tabuľka 3 predstavujú **dátovú maticu**, ktorá je vhodným a bežným spôsobom usporiadania údajov, najmä ak sa údaje zbierajú v tabuľkovom procesore (napr, Microsoft Excel). Každý riadok matice údajov zodpovedá jedinečnému prípadu (jednotke pozorovania alebo štatistickej jednotke) a každý stĺpec zodpovedá premennej.
Pri zaznamenávaní údajov používajte dátovú maticu, pokiaľ nemáte veľmi dobrý dôvod na použitie inej štruktúry. Táto štruktúra umožňuje pridávať nové prípady ako riadky alebo nové premenné ako nové stĺpce.

1.2.2. Typy premenných

K dispozícii máme dátový súbor s názvom **okresyUSA** o 3 142 okresoch jednotlivých štátov v Spojených štátoch amerických (Tabuľka 5) a opis použitých premenných (Tabuľka 6). Dátový súbor poskytuje informácie o názve každého okresu, štát, v ktorom sa okres nachádza, počet obyvateľov v roku 2017, zmenu počtu obyvateľov od roku 2010 do roku 2017, mieru chudoby a šesť ďalších charakteristík. Dôležitým krokom pri každej analýze dát je porozumieť všetkým použitým premenným a vedieť ich hodnoty interpretovať.

Tabuľka 5 Dátový súbor **okresyUSA**

ID	nazov	stat	pop	pop_zmena	chudoba	vlastnictvo	viacbytove	nezam	metro	priem_vzd	priem_prij
1	Autauga	Alabama	55504	1,48	13,7	77,5	7,2	3,86	ano	stredoškolské	55317
2	Baldwin	Alabama	212628	9,19	11,8	76,7	22,6	3,99	ano	stredoškolské	52562
3	Barbour	Alabama	25270	-6,22	27,2	68,0	11,1	5,90	nie	vysokoškolské	33368

4	Bibb	Alabama	22668	0,73	15,2	82,9	6,6	4,39	ano	vysokoškolské	43404
5	Blount	Alabama	58013	0,68	15,6	82,0	3,7	4,02	ano	vysokoškolské	47412
6	Bullock	Alabama	10309	-2,28	28,5	76,9	9,9	4,93	nie	vysokoškolské	29655
7	Butler	Alabama	19825	-2,69	24,4	69,0	13,7	5,49	nie	vysokoškolské	36326
8	Calhoun	Alabama	114728	-1,51	18,6	70,7	14,3	4,93	ano	stredoškolské	43686
9	Chambers	Alabama	33713	-1,20	18,8	71,4	8,7	4,08	nie	vysokoškolské	37342
10	Cherokee	Alabama	25857	-0,60	16,1	77,5	4,3	4,05	nie	vysokoškolské	40041
...	...	...	...	...	...	...	...	...	...	...	...
3142	Weston	Wyoming	6927	-2,93	14,4	77,9	6,5	3,98	nie	stredoškolské	59605

Tabuľka 6 Opis premenných dátového súboru **okresyUSA**

Premenná	**Popis**
nazov	Názov okresu,
stat	Štát, v ktorom sa okres nachádza, alebo District of Columbia,
pop	Populácia v roku 2017,
pop_zmena	Percentuálna zmena populácie od roku 2010 do 2017, Napr. hodnota 1,48 znamená nárast populácie o 1,48 % medzi rokmi 2010 a 2017,
chudoba	Percento populácie žijúcej v chudobe,
vlastníctvo	Percento populácie, ktorá býva vo vlastnom dome alebo s vlastníkom (napr. deti s rodičmi vlastniacimi dom),
viacbytove	Percento obytných jednotiek, ktoré sú vo viacbytových stavbách (napr. bytovky),
nezam	Miera nezamestnanosti v percentách,
metro	Označuje, či okres obsahuje metropolitnú oblasť,

priem_vzd	Priemerná úroveň vzdelania – môže byť: základné, stredoškolské, vysokoškolské,
priem_prij	Priemerný príjem domácnosti v okrese – zahŕňa príjem všetkých osôb v domácnosti starších ako 15 rokov,

Dátový súbor **okresyUSA** ponúka 11 premenných (11 stĺpcov), ktorých podrobnejší opis sa nachádza v Tabuľka 6. Vezmime si na preskúmanie 4 z nich: populácia (*pop*), miera nezamestnanosti (*nezam),* štát (*stat*) a priemerná úroveň vzdelania (*priem_vzd)*. Každá z týchto premenných sa vo svojej podstate líši od ostatných troch, avšak niektoré z nich majú určité spoločné charakteristiky.

Najprv preskúmajme premennú miera nezamestnanosti (*nezam):* ide o tzv. **číselnú (numerickú, kardinálnu, kvantitatívnu)**[4] premennú, pretože môže nadobúdať širokú škálu číselných hodnôt a má zmysel tieto hodnoty sčítať, odčítať alebo vypočítať priemer. Na druhej strane, ak by sme mali premennú napríklad s telefónnymi alebo poštovými smerovacími číslami, tak by sme ju nemohli klasifikovať ako číselnú (napriek tomu, že obsahuje číselné údaje a dokonca by v názve mala slovo „číslo“), pretože priemer, súčet alebo rozdiel telefónnych čísel, či smerovacích čísel nedáva nijaký zmysel. Premenná populácia (*pop)* je tiež číselná, ale trochu iná, ako miera nezamestnanosti. Táto premenná môže nadobúdať len celé nezáporné čísla (0, 1, 2, ...), Rozdiel medzi nimi je v tom, že kým premenná *nezam* môže nadobúdať ľubovoľnú hodnotu v určitom intervale (napr, 4,02, 4,025, 4,0251...*),* pri premennej *pop* nemá zmysel hovoriť o desatinných číslach (napríklad o 2,5 obyvateľa). Obe premenné sú teda číselné, premennú *pop* nazývame **diskrétna** a premennú *nezam* nazývame **spojitá.**[5]

Premenná štát (*stat*) môže nadobúdať 51 hodnôt: Alabama, Aljaška, ... až Wyoming. Každý štát predstavuje jednu kategóriu a premenná sa nazýva **kategoriálna (kvalitatívna).** Podobne

[4] Všetky uvedené termíny pre číselnú premennú budeme v učebnici používať ako jej synonymá.

[5] Pomenovanie spojitá premenná má aj v prirodzenom jazyku opodstatnenie a intuitívne vnímame, že hodnoty spojitej premennej tvoria plynulé spektrum. Pomenovanie diskrétna premenná môže evokovať význam slova diskrétny, kedy myslíme niekoho, kto je taktne mlčanlivý, nenápadný, vie zachovať súkromie. V matematike ale slovo diskrétny znamená nespojitý, oddelený na jednotlivé časti. S trochou fantázie môžeme ako diskrétneho človeka označiť toho, kto „oddeľuje“ to, čo vie, od toho, čo prezradí – teda zachováva odstup. V tomto význame je použitý termín diskrétna premenná.

premenná priemerná úroveň vzdelania (*priem_vzd)* obyvateľov okresu môže nadobudnúť 3 hodnoty: základné, stredoškolské, vysokoškolské a rovnako ako premenná štát je kategoriálna a každá úroveň vzdelania predstavuje jednu kategóriu. Avšak v prípade premennej *priem_vzd* existuje medzi kategóriami prirodzené usporiadanie, ktoré pri premennej *stat* neexistuje. Kategoriálna premenná, ktorej kategórie môžeme usporiadať sa nazýva **ordinálna** (napr, *priem_vzd)* a kategoriálna premenná bez tohto usporiadania kategórií sa nazýva **nominálna** (napr. *stat*).[6]

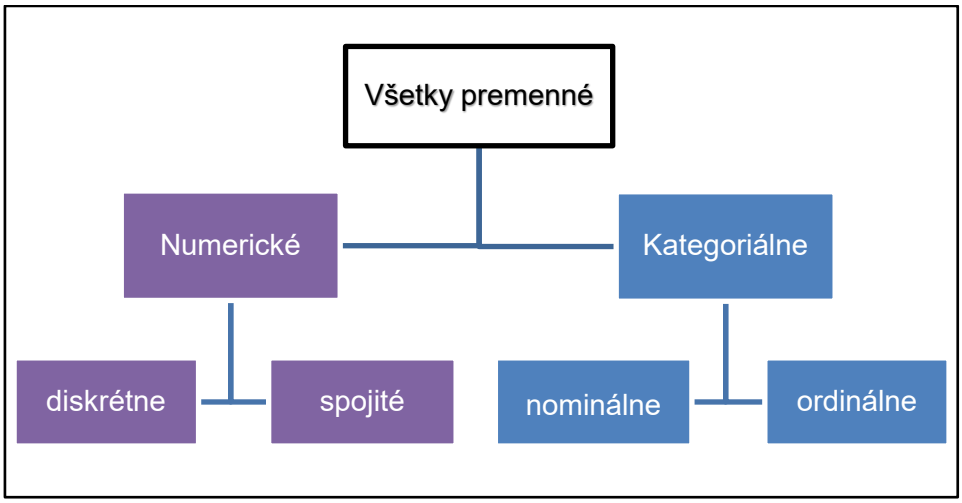
Rozdelenie všetkých premenných na jednotlivé typy je znázornené na **Chyba! Nenašiel sa žiaden zdroj odkazov.**.

Príklad 1.4

Predstavte si, že vyučujúca predmetu Analýza dát zozbierala údaje o všetkých študentoch, ktorí absolvovali tento predmet. Pri každom študentovi boli zaznamenané tri premenné: počet súrodencov, výška študenta a to, či študent už predtým absolvoval predmet Analýza dát. Pre každú premennú určte, o aký typ premennej ide.[7]

[6] Slovo nominálny pochádza z latinského „nomen“, čo znamená meno. Nominálna premenná obsahuje kategórie, ktoré majú iba názov (meno). Slovo ordinálny pochádza z latinského „ordo, ordinis“, čo znamená poriadok, usporiadanie (prípadne z angličtiny order) a preto ordinálna premenná má kategórie, ktoré sa dajú usporiadať.

[7] Počet súrodencov a výška študenta predstavujú číselné premenné. Počet súrodencov je diskrétna premenná (nie je možné, aby sme uvažovali o 2,5 súrodenca) a výška študenta je to spojitá číselná premenná. Posledná premenná rozdeľuje študentov do dvoch kategórií - tých, ktorí absolvovali a tých, ktorí neabsolvovali predmet Analýza dát – ide teda o kategoriálnu premennú, pričom kategórie nemá zmysel usporiadať, preto je to nominálna premenná.



Cvičenia:

1. V doterajšom hodnotení študentov mediamatiky boli výsledky z predmetu Analýza dát nasledujúce: známku A získalo 11% študentov; známku B získalo 14% študentov; známku C získalo 18% študentov; známku D získalo 22% študentov; známku E získalo 29% študentov a nakoniec 6% sa nepodarilo predmet úspešne ukončiť, čiže ich hodnotenie bolo FX. Aký typ premennej predstavuje známka, ktorú získali študenti?
   a) kategoriálna, ordinálna
   b) kategoriálna, nominálna
   c) numerická, spojitá
   d) numerická, diskrétna

2. Ktoré z nasledujúcich sú príkladmi kvantitatívnych/numerických/číselných premenných (viacero možností je správnych)?
   a) mesačná mzda uvádzaná v eurách
   b) miesto pobytu
   c) telefónne číslo
   d) výška skúmanej osoby udávaná v centimetroch
   e) farba očí
   f) číslo občianskeho preukazu

3. Ktoré z nasledujúcich náhodných premenných nie sú spojité? (viacero možností je správnych)
   a) dĺžka telefonického hovoru
   b) počet znakov, ktoré padli pri 100násobnom hádzaní mincou
   c) počet úmrtí pri pádoch lietadla za rok
   d) množstvo peňazí, ktoré domácnosť utratila za potraviny za rok
   e) počet základných škôl v meste

4. Ktoré z nasledujúcich premenných sú diskrétne? (viacero možností je správnych)
   a) hmotnosť dospelých jedincov v SR uvádzaná v kilogramoch
   b) vzdialenosť medzi dvoma mestami uvádzaná v kilometroch
   c) počet detí v rodine
   d) mesačný úhrn zrážok v meste uvádzaný v milimetroch na meter štvorcový
   e) mesačná mzda v eurách
   f) počet ľudí odsúdených za závažný trestný čin za rok

5. Ktoré z nasledujúcich sú príkladmi kategoriálnych premenných? (viacero možností je správnych)
   a) počet ukončených rokov školskej dochádzky
   b) ročný úhrn zrážok
   c) ukončený stupeň vzdelania
   d) ročný príjem (v eurách)
   e) akademická hodnosť (Bc., Mgr., PhD., atď)

1.3. Zásady a stratégia výberu vzorky

Prvým krokom pri realizácii výskumu je identifikácia tém alebo otázok, ktoré sa majú skúmať. Jasne stanovená výskumná otázka je nápomocná pri určovaní toho, aké objekty by sa mali skúmať a aké premenné sú dôležité. Dôležité je tiež zvážiť *spôsob* zberu údajov, aby boli spoľahlivé a pomohli dosiahnuť ciele výskumu.

1.3.1. Populácia a vzorka

Zvážme nasledujúce dve výskumné otázky:

1. Aký je priemerný čas potrebný na ukončenie štúdia študentov Univerzity Komenského za posledných 5 rokov?
2. Vedie zavedenie novej kreatívnej stratégie na sociálnych sieťach k zvýšeniu miery zapojenia používateľov (engagement rate)?

Každá výskumná otázka sa vzťahuje na cieľovú skupinu alebo **populáciu**. **Populácia** je celý súbor objektov alebo prípadov (používateľov, médií, obsahov, udalostí a pod.), o ktorých chceme pomocou dát urobiť záver.
V prvej otázke sú populáciou všetci študenti Univerzity Komenského, ktorí v priebehu posledných 5 rokov ukončili štúdium a každý takýto študent predstavuje jeden prípad. Často je príliš nákladné zbierať údaje o každom prípade v populácii. Namiesto toho sa vyberie vzorka. **Vzorka** predstavuje podskupinu prípadov a často je malým zlomkom populácie. Napríklad sa môže vybrať 100 študentov (alebo iný počet) v populácii a údaje z tejto vzorky sa môžu použiť na odhad priemerného času populácie a na zodpovedanie výskumnej otázky.

Príklad 1.5

Čo by mohlo predstavovať populáciu a jeden individuálny prípad pre druhú výskumnú otázku?[8]

1.3.2. Výber vzorky z populácie

Mohli by sme sa pokúsiť odhadnúť čas na ukončenie štúdia študentov Univerzity Komenského za posledných 5 rokov a to na základe vzorky študentov. Všetci absolventi za posledných 5 rokov predstavujú *populáciu* a absolventi, ktorí sú vybraní na preskúmanie sa spoločne nazývajú *vzorka*. Vo všeobecnosti sa vždy snažíme *náhodne* vybrať vzorku z populácie. Najzákladnejší typ náhodného výberu zodpovedá spôsobu, akým sa uskutočňujú lotérie. Napríklad pri výbere absolventov by sme mohli napísať meno každého absolventa na tombolový lístok a vylosovať

[8] Odpoveď na otázku

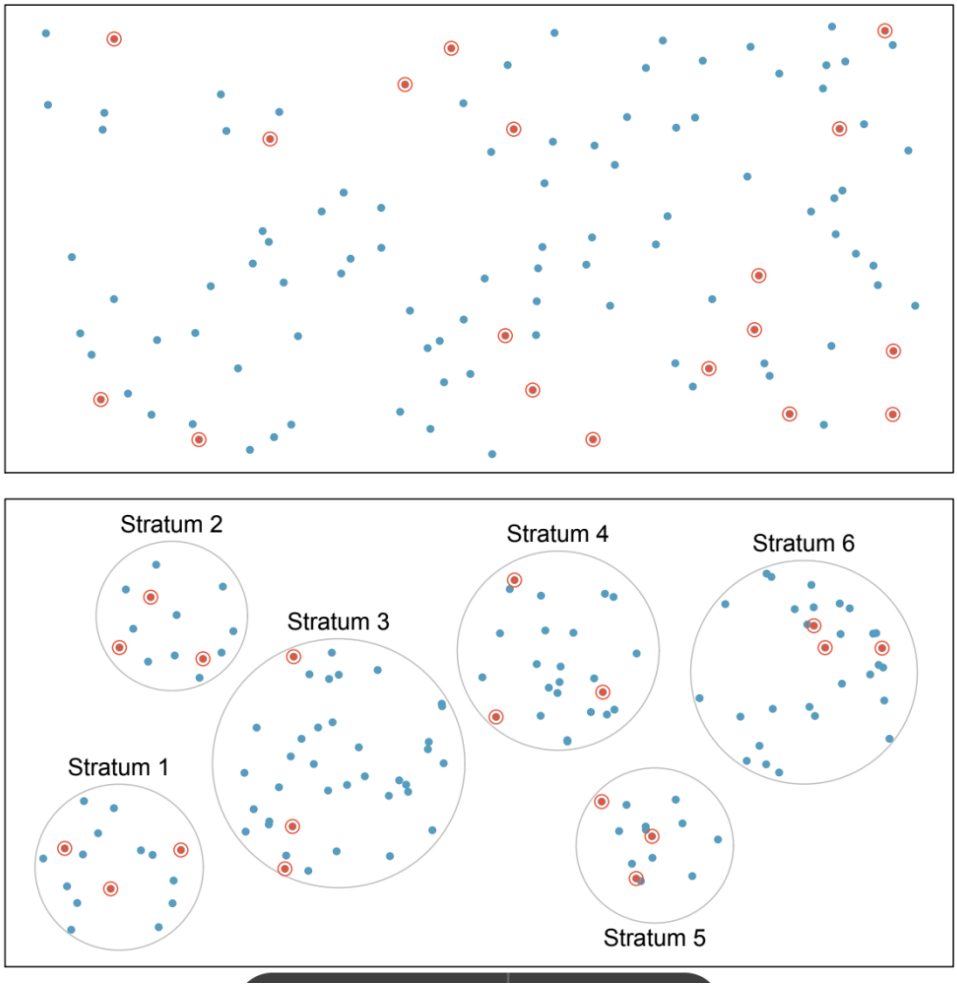
100 lístkov. Vybrané mená by predstavovali náhodnú vzorku 100 absolventov. Vzorky vyberáme náhodne, aby sme znížili pravdepodobnosť, že do nich vnesieme skreslenie. To znamená, že každý prípad v populácii má rovnakú šancu byť zaradený do populácie a medzi prípadmi vo vzorke nie je žiadna implicitná súvislosť. Takýto výber vzorky sa nazýva **jednoduchý (prostý) náhodný výber** a pomáha minimalizovať skreslenie.

Skreslenie sa však môže prejaviť aj iným spôsobom. Aj keď sa ľudia vyberajú náhodne, napr. pri prieskumoch, treba byť opatrný, ak je miera odpovedania na prieskum nízka. Ak napríklad len 30 % ľudí náhodne vybraných do prieskumu skutočne odpovedá, nie je jasné, či sú výsledky reprezentatívne pre celú populáciu.

Ďalším častým nedostatkom je **pohodlná vzorka**, kde je väčšia pravdepodobnosť, že do vzorky budú zaradení jednotlivci, ktorí sú ľahko dostupní. Ak sa napríklad politický prieskum uskutoční tak, že sa zastavia ľudia prechádzajúci sa v Bronxe, nebude to reprezentovať celý New York. Často je ťažké rozlíšiť, akú subpopuláciu reprezentuje pohodlná vzorka.

1.3.3. Metódy náhodného výberu vzorky

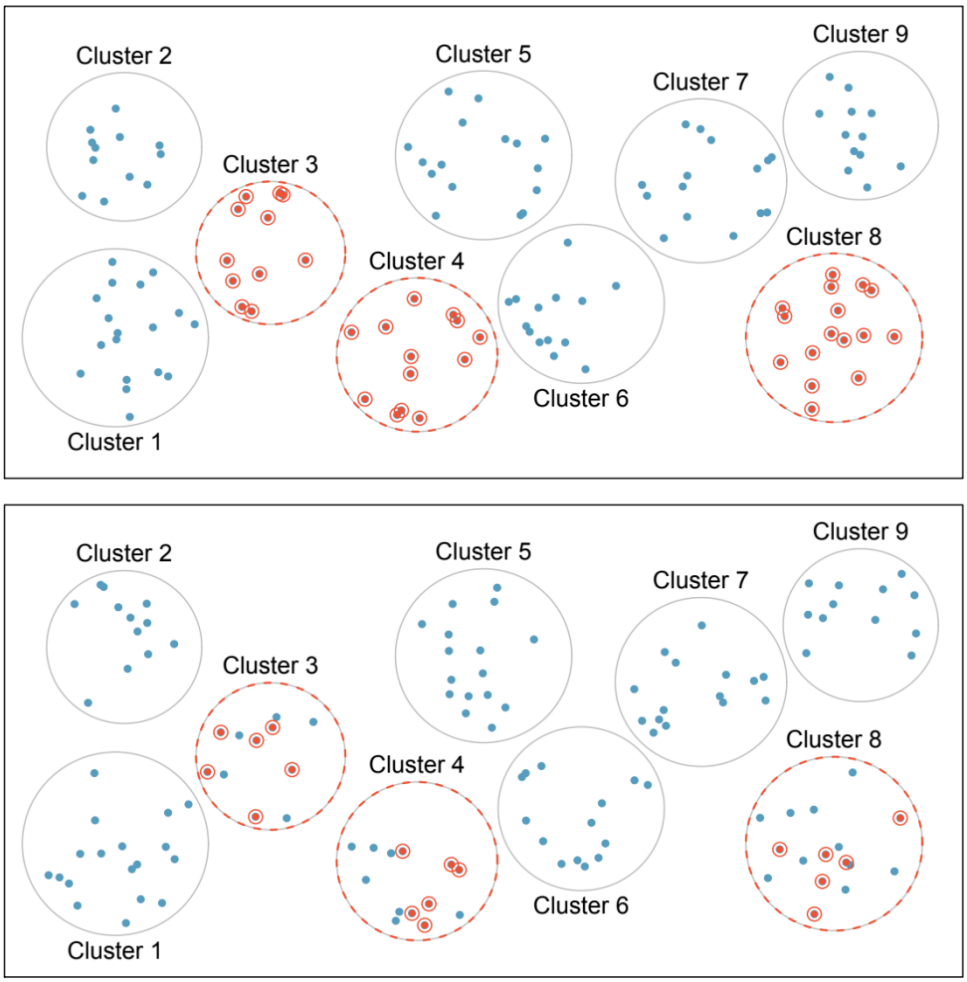
Takmer všetky štatistické metódy sú založené na pojme náhodnosti. Ak sa pozorované údaje nezbierajú v rámci náhodného výberu z populácie, tieto štatistické metódy - odhady a chyby spojené s odhadmi - nie sú spoľahlivé. V tejto časti sa zaoberáme štyrmi technikami náhodného výberu: jednoduchým, stratifikovaným, skupinovým a viacstupňovým výberom. Na sú graficky znázornené tieto techniky.



Obrázok 3 Príklad jednoduchého a stratifikovaného náhodného výberu

Jednoduchý náhodný výber je pravdepodobne najintuitívnejšou formou náhodného výberu. Povedzme, že nás zaujímajú platy hráčov Major League Baseball (MLB), kde každý hráč je členom jedného z 30 tímov ligy. Ak chceme vybrať jednoduchú náhodnú vzorku napríklad 120 hráčov baseballu a následne zistiť ich platy, môžeme napísať mená všetkých hráčov v danej sezóne na lístky (v jednej sezóne ich môžu byť stovky), hodiť lístky do dostatočne veľkej nádoby, potriasť nádobou, kým si nebudeme istí, že sú všetky mená pomiešané a potom vyťahovať lístky, kým nezískame vzorku 120 hráčov. Vo všeobecnosti sa vzorka označuje ako "jednoducho náhodná", ak každý prípad v populácii má rovnakú šancu byť zahrnutý do konečnej vzorky a informácia o tom, že prípad je zahrnutý do vzorky, neposkytuje užitočné informácie o tom, ktoré ďalšie prípady sú zahrnuté.

Stratifikovaný výber sa používa vtedy, ak je populácia rozdelená do oblastí nazývaných aj **vrstvy alebo straty**[9]. Vrstvy sa vyberajú tak, aby sa podobné prípady zoskupili a potom sa v rámci každej vrstvy použije druhá metóda výberu vzorky, zvyčajne jednoduchý náhodný výber. V príklade s platmi hráčov baseballu by mohli vrstvy predstavovať tímy, pretože niektoré tímy majú oveľa viac peňazí (až 4-násobne viac!). Potom by sme mohli náhodne vybrať 4 hráčov z každého tímu, spolu 120 hráčov.



Obrázok 4 Príklad skupinového a viacstupňového náhodného výberu

Pri **skupinovom výbere** rozdelíme populáciu do mnohých skupín, niekedy nazývané aj zhluky. Potom náhodne vyberieme pevný počet skupín a do vzorky zahrnieme všetky objekty z každej z

[9] z lat. *stratum* = vrstva

týchto skupín. **Viacstupňový výber** je podobný skupinovému výberu, ale namiesto toho, aby sme ponechali všetky objekty v každej skupine, vyberáme náhodným výberom objekty aj v rámci každej vybranej skupiny.

Predpokladajme, že nás napríklad zaujíma odhad miery malárie v husto osídlenej tropickej časti indonézskeho vidieka. Dozvedeli sme sa, že v tejto časti indonézskej džungle sa nachádza 30 dedín, z ktorých každá je viac či menej podobná tej nasledujúcej. Naším cieľom je otestovať 150 osôb na maláriu. Akú metódu výberu do vzorky treba použiť?

Jednoduchým náhodným výber by sme pravdepodobne vybrali jednotlivcov zo všetkých 30 dedín, čo by mohlo veľmi predražiť zber údajov. Stratifikovaný výber by bol problémom, pretože nie je jasné, ako by sme vytvorili vrstvy podobných jednotlivcov. Skupinový výber alebo viacstupňový výber sa však javia ako veľmi dobrý spôsob výberu. Ak by sme sa rozhodli použiť viacstupňový výber, mohli by sme náhodne vybrať polovicu dedín a potom z každej náhodne vybrať 10 osôb. To by pravdepodobne výrazne znížilo naše náklady na zber údajov v porovnaní s jednoduchým náhodným výberom a viacstupňový výber by nám stále poskytol spoľahlivé informácie, aj keď by sme museli analyzovať údaje pomocou trochu pokročilejších metód, než o akých hovoríme v tejto učebnici.

Cvičenia:

6. Vedci zhromažďovali údaje na preskúmanie vzťahu medzi látkami znečisťujúcimi ovzdušie a predčasnými pôrodmi v južnej Kalifornii. Počas štúdie bolo znečistenie ovzdušia merané monitorovacími stanicami kvality ovzdušia. Konkrétne boli zaznamenané hladiny oxidu uhoľnatého (v časticiach na milión), oxidu dusičitého a ozónu (v časticiach na sto miliónov) a hrubé častice (PM10) v µg/m3. Údaje o dĺžke tehotenstva sa zhromaždili u 143 196 pôrodov medzi rokmi 1989 a 1993 a pre každý pôrod sa počítala expozícia znečisteniu ovzdušia počas tehotenstva. Analýza naznačila, že výskyt predčasných pôrodov môže súvisieť so zvýšeným množstvom PM10 v prostredí a v menšej miere aj s koncentráciou CO.

a) Sformulujte hlavnú výskumnú otázku štúdie.

b) Čo je štatistická jednotka?

c) Aké sú premenné v štúdii? Ku každej premennej priraďte jej typ.

7. Vedci, ktorí študovali vzťah medzi čestnosťou, vekom a sebaovládaním, uskutočnili experiment na 160 deťoch vo veku od 5 do 15 rokov. Účastníci uviedli svoje vek, pohlavie a či boli jedináčikom alebo nie. Vedci požiadali každé dieťa, aby v súkromí hodili mincou

a zaznamenali si výsledok na papier (hlava alebo znak) a deťom povedali, že budú odmenené iba vtedy, ak padne znak. Zistenia štúdie možno zhrnúť takto: „Polovici detí bolo výslovne povedané, aby nepodvádzali, a ostatní nedostali žiadne výslovné pokyny. V skupine, kde neboli žiadne inštrukcie, bola rovnaká pravdepodobnosť podvádzania vo všetkých skupinách na základe charakteristík dieťaťa. V skupine, v ktorej bolo výslovne povedané, aby nepodvádzala, dievčatá mali menšiu pravdepodobnosť podvádzania, zatiaľ čo miera podvádzania u chlapcov sa nelíšila podľa veku, u dievčat sa s vekom zmenšovala."

a) Sformulujte hlavnú výskumnú otázku štúdie.

b) Čo je štatistická jednotka?

c) Aké sú premenné v štúdii? Ku každej premennej priraďte jej typ.

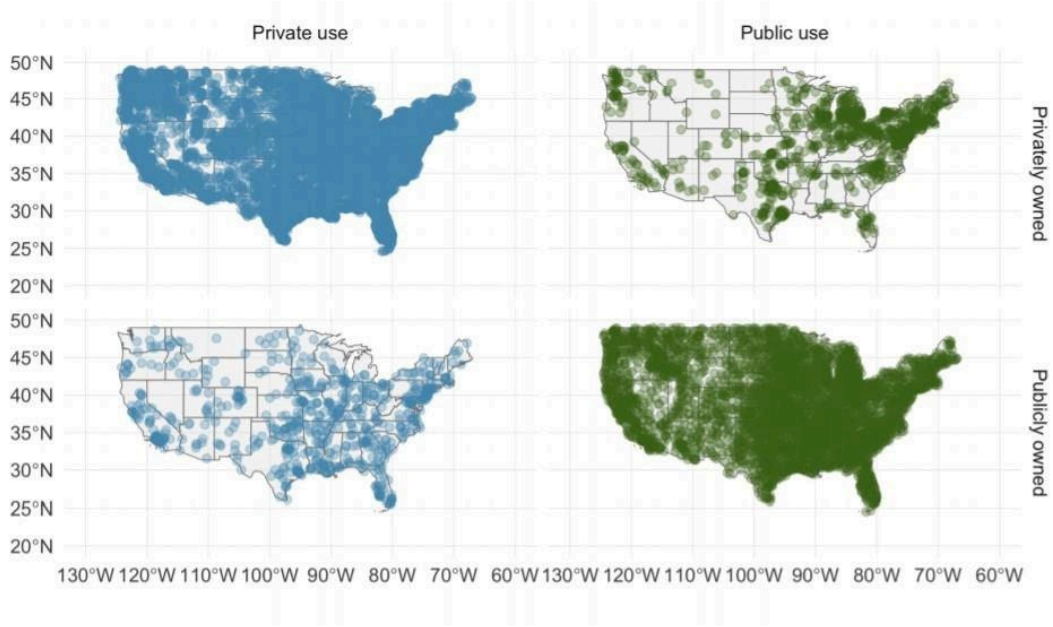
8. Pri štúdiu vzťahu medzi socioekonomickou triedou a neetickým správaním, 129 vysokoškolákov z Kalifornskej univerzity v Berkeley bolo požiadaných, aby identifikovali sami seba ako príslušníkov nízkej alebo vyššej spoločenskej triedy. V miestnosti, kde sa uskutočňoval rozhovor s vysokoškolákmi boli aj cukríky, o ktorých dostali informáciu, že sú určené pre deti, ktoré sú vo vedľajšej miestnosti, ale ak chcú, môžu sa cukríkom ponúknuť. Po splnení niekoľkých nesúvisiacich úloh, účastníci nahlásili počet cukríkov, ktoré si vzali. Štúdia zistila, že študenti, ktorí sami seba považovali za ľudí z vyššej triedy, si vzali viac cukríkov ako ostatní.

a) Sformulujte hlavnú výskumnú otázku štúdie.

b) Čo je štatistická jednotka?

c) Aké sú premenné v štúdii? Ku každej premennej priraďte jej typ.

9. Nasledujúca vizualizácia zobrazuje geografické rozmiestnenie letísk v USA. Čo je štatistická jednotka? Aké premenné boli vo vizualizácii použité? Ku každej premennej priraďte jej typ.



2. Opisná štatistika

V druhej kapitole sa venujeme tzv. opisnej alebo deskriptívnej štatistike, ktorá má za úlohu na malom priestore ideálne pomocou jedného čísla, tabuľky, či grafu opísať premenné použité vo výskume. Predstavujeme aj opisné štatistiky pre oba typy premenných, ako numerické, tak kategoriálne. Opisné štatistiky sú čísla, ktoré sumarizujú veľké množstvo údajov a s istou dávkou fantázie môžeme povedať, že vypovedajú o nich istý príbeh. Môžeme ich získať jednoduchým výpočtom na základe vzorca alebo môžeme na ich získanie použiť ľubovoľný štatistický softvér, napr. IBM SPSS Statistics, RStudio, Microsoft Excel (s doplnkom Analýza údajov), Statistica, SAS a mnoho ďalších v závislosti od odvetvia, z ktorého pochádzajú dáta. Ak je čitateľ v tejto chvíli zneistený tým, že sa v predchádzajúcom odstavci vyskytol pojem opisná štatistika hneď v dvoch rôznych významoch, tak je to v poriadku.
Opisnú štatistiku môžeme totiž chápať aj ako **oblasť štatistiky**, ktorá sa zaoberá zhrnutím, usporiadaním, popisom a vizualizáciou údajov, ale aj ako **konkrétne číslo**, ktoré sumarizuje určitú vlastnosť premennej. Rozlíšenie, v akom význame je pojem opisná štatistika použitý, bude v nasledujúcom texte zrejmý z kontextu.

2.1. Opisná štatistika pre numerickú premennú

Keďže študenti mediamatiky sú kreatívne bytosti, rozhodli sa opisné štatistiky vyskúšať na vlastnej koži. Ako numerickú premennú, ktorú chceli skúmať, si vybrali takú, ktorá opisuje stav hotovosti študentov predmetu Analýza dát v aktuálnom akademickom roku počas záhradnej slávnosti. Treba doplniť, že ide o numerickú spojitú premennú. Premennú si nazvali *peňaženka*. Počet študentov predmetu v aktuálnom akademickom roku bol 19 a každý z nich si skontroloval svoju hotovosť v peňaženke. Údaje od študentov boli zaznamenané do Tabuľka 7:

Tabuľka 7 Údaje o hotovosti pre premennú peňaženka od 19 študentov mediamatiky

študent	***peňaženka***
Peter	15,90 €
Juraj	0,00 €
Ema	5,60 €

Klaudia	11,30 €
Alexandra	13,70 €
Emília	20,30 €
Dávid	90,60 €
Pavol	30,70 €
Barbora	45,00 €
Michal	0,50 €
Michaela	2,30 €
Ladislav	4,60 €
Katarína	7,20 €
Lívia	12,40 €
Lukáš	12,50 €
Matej	17,80 €
Lucia	31,20 €
Filip	50,80 €
Slávka	23,50 €

V nasledujúcich častiach sú predstavené spôsoby opisu numerickej premennej pomocou opisných štatistík (špeciálnych čísel, ktoré sú podrobne rozoberané v časti 2.1.1), pomocou frekvenčnej tabuľky (v časti 2.1.2) a pomocou grafického znázornenia vo forme histogramu a škatuľkového grafu (v časti 2.1.3).

2.1.1. Opisné štatistiky pre numerickú premennú

Študenti sa rozhodli preskúmať premennú *peňaženka* a zistili pre premennú nasledujúce opisné štatistiky: 1. rozsah, 2. minimum a maximum, 3. variačné rozpätie, 4. modus, 5. aritmetický priemer, 6. rozptyl, 7. smerodajná odchýlka, 8. medián, 9. kvartily (prvý a tretí), 10. medzikvartilové rozpätie.

Štatistické softvéry poskytujú výsledok opisnej štatistiky (descriptive statistics) vo forme tabuľky, ktorá by vyzerala napríklad ako Tabuľka 8:

Tabuľka 8 Opisné štatistiky pre premennú peňaženka

	peňaženka
Count	19
Minimum	0

Maximum	90,6
Range	90,6
Mode	##########
Mean	20,8368421
Sample Variance	486,105789
Standard Deviation	22,0478069
Median	13,7
Quartils	5,6; 30,7
IQR	25,1

V nasledujúcom texte sú postupne vysvetlené jednotlivé opisné štatistiky. V našej učebnici sú vysvetlené vyššie uvedené, niekedy sa však používajú aj opisné štatistiky na určenie tvaru rozdelenia premennej šikmosť (skewness) a špicatosť (kurtosis), ktoré však nie sú predmetom skúmania v našej učebnici.

1. Rozsah (count) udáva počet štatistických jednotiek zaradených do vzorky, v našom prípade 19 študentov, resp. 19 hodnôt obsahu ich peňaženky; rozsah označujeme zvyčajne písmenom ***n.***

2. Minimum a maximum predstavujú najnižšiu a najvyššiu hodnotu premennej, v našom prípade je minimum 0 €, čo znamená prázdnu Jurajovu peňaženku a maximum je Dávidových 90,60 €; označenie pre minimum a maximum je ***min*** a ***max***.

3. Rozpätie (range) získame ako rozdiel medzi najvyššou a najnižšou hodnotou, pre nás 90,60 € – 0 € = 90,60 €; zaužívané označenie pre rozsah je ***R.***

4. Modus (mode) je tá hodnota, ktorá sa v súbore dát vyskytuje najčastejšie. Keďže v našom konkrétnom prípade pri zisťovaní hotovosti v peňaženkách študentov predmetu Analýza dát sa nestalo, že by dvaja študenti mali rovnakú hotovosť, nemožno žiadnu z hodnôt považovať za modus. preto hovoríme, že v tomto prípade modus neexistuje a štatistický softvér použil na označenie ##########.

5. Aritmetický priemer (mean alebo average) je hodnota, ktorá opisuje stred všetkých údajov a ktorá vznikne tak, že súčet všetkých hodnôt vydelíme ich počtom:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Aritmetický priemer sa označuje symbolom $\bar{x}$ (x s pruhom) a v našom prípade ho vypočítame tak, že spočítame postupne obsah peňaženky všetkých 19 študentov: Petra, Juraja, atď. až Slávky a vydelíme počtom študentov:

$$\bar{x} = \frac{15{,}90 + 0 + \cdots + 23{,}5}{19} = \frac{395{,}90}{19} \cong 20{,}84$$

Priemerne mali teda študenti predmetu Analýza dát v deň záhradnej slávnosti vo svojich peňaženkách približne 21 €.

Príklad 2.1

Čo by sa však stalo, keby v tento deň Dávid prišiel na záhradnú slávnosť nie s 90,60 €, ale prišiel by priamo z brigády, v ktorej dostal mesačnú výplatu a v jeho peňaženke by bolo napríklad 1000 €? Ako by zmena v peňaženke jedného študenta ovplyvnila aritmetický priemer celej skupiny študentov?[10]

V tomto prípade by sa aritmetický priemer zvýšil až na 68,70 €, čo nie veľmi vypovedá o stave peňaženiek mediamatikov v spomínaný deň. Je preto užitočné si uvedomiť, že aritmetický priemer je opisná štatistika, ktorá je ľahko ovplyvniteľná extrémnymi hodnotami.

6. a 7. Rozptyl a smerodajná odchýlka (sample variance a standard deviation) sú takzvané miery variability a informujú nás o tom, ako veľmi sa jednotlivé hodnoty v súbore údajov od seba líšia. Čím viac sa hodnoty navzájom líšia, tým väčšia bude hodnota rozptylu a smerodajnej odchýlky. Obe sú mierne komplikovanejšie na „ručný“ výpočet, preto si často pri ich určení pomáhame štatistickými softvérmi. **Smerodajná odchýlka** je jednoduchšia na pochopenie a zhruba opisuje, ako ďaleko je typické pozorovanie od priemeru alebo inak povedané získame ju ako priemernú vzdialenosť hodnôt súboru od priemeru. **Rozptyl** je potom druhou mocninou smerodajnej odchýlky (alebo smerodajná odchýlka je druhou odmocninou rozptylu) a vyjadruje tzv. priemernú štvorcovú (umocnenú na druhú) vzdialenosť od priemeru. Označenie pre smerodajnú odchýlku je s a potom rozptyl je s^2 **.** Pri

[10] V tomto prípade by sa aritmetický priemer rovnal: $\bar{x} = \frac{1305{,}30}{19} \cong 68{,}70$

výpočte však postupujeme presne naopak: najskôr vypočítame rozptyl a potom z neho vypočítame smerodajnú odchýlku ako druhú odmocninu rozptylu. Na lepšie pochopenie vypočítajme rozptyl a smerodajnú odchýlku pre našu premennú *peňaženka*:

Začneme tým, že pre každú hodnotu súboru, teda pre každého študenta, vypočítame rozdiel od aritmetického priemeru, ktorého hodnota na základe predchádzajúceho výpočtu je 20,84. Pre Petra to bude znamenať, že od jeho stavu peňaženky 15,90 € odpočítame priemer 20,84 € a získame zápornú hodnotu – 4,94 €. Pre Dávida získame ale kladnú hodnotu: 90,60 € – 20,84 € = 69,76 €. V Tabuľka 9 sú zaznamenané rozdiely od priemeru pre všetkých študentov:

Tabuľka 9 Výpočet rozdielu hodnôt a aritmetického priemeru

študent	***peňaženka***	***hodnota - priemer***
Peter	15,90 €	-4,94 €
Juraj	0,00 €	-20,84 €
Ema	5,60 €	-15,24 €
Klaudia	11,30 €	-9,54 €
Alexandra	13,70 €	-7,14 €
Emília	20,30 €	-0,54 €
Dávid	90,60 €	69,76 €
Pavol	30,70 €	9,86 €
Barbora	45,00 €	24,16 €
Michal	0,50 €	-20,34 €
Michaela	2,30 €	-18,54 €
Ladislav	4,60 €	-16,24 €
Katarína	7,20 €	-13,64 €
Lívia	12,40 €	-8,44 €
Lukáš	12,50 €	-8,34 €
Matej	17,80 €	-3,04 €
Lucia	31,20 €	10,36 €
Filip	50,80 €	29,96 €
Slávka	23,50 €	2,66 €

Niektoré hodnoty rozdielu sú kladné, niektoré záporné v závislosti od toho, či hodnota peňazí v peňaženke študenta bola väčšia alebo menšia ako priemer.

Príklad 2.2

Čo by sa stalo, keby sme chceli vypočítať priemer týchto rozdielov?[11]

Kvôli tomu, že rozdiely sú kladné aj záporné, ich priemer je 0. Tým sme nezískali nijakú užitočnú informáciu o tom, ako sa hodnoty vzdialili od priemeru. Aby sme eliminovali to, že niektoré hodnoty rozdielov sú kladné a niektoré záporné, umocníme ich na druhú, čím zo všetkých hodnôt „vyrobíme" len kladné čísla. Výsledné druhé mocniny rozdielov sú v Tabuľka 10:

Tabuľka 10 Druhé mocniny rozdielov od priemeru

študent	***peňaženka***	***hodnota - priemer***	***hodnota - priemer na druhú***
Peter	15,90 €	-4,94 €	24,37 €
Juraj	0,00 €	-20,84 €	434,17 €
Ema	5,60 €	-15,24 €	232,16 €
Klaudia	11,30 €	-9,54 €	90,95 €
Alexandra	13,70 €	-7,14 €	50,93 €
Emília	20,30 €	-0,54 €	0,29 €
Dávid	90,60 €	69,76 €	4 866,90 €
Pavol	30,70 €	9,86 €	97,28 €
Barbora	45,00 €	24,16 €	583,86 €
Michal	0,50 €	-20,34 €	413,59 €
Michaela	2,30 €	-18,54 €	343,61 €
Ladislav	4,60 €	-16,24 €	263,64 €
Katarína	7,20 €	-13,64 €	185,96 €
Lívia	12,40 €	-8,44 €	71,18 €
Lukáš	12,50 €	-8,34 €	69,50 €
Matej	17,80 €	-3,04 €	9,22 €

[11] Na výpočet aritmetického priemeru rozdielov od priemeru by sme najskôr potrebovali všetky rozdiely spočítať. Ich súčet je 0 a tým pádom aj priemer rozdielov je 0.

Lucia	31,20 €	10,36 €	107,40 €
Filip	50,80 €	29,96 €	897,79 €
Slávka	23,50 €	2,66 €	7,09 €

Následne by sme vypočítali aritmetický priemer druhých mocnín rozdielov tak, že všetky druhé mocniny sčítame (8 749,90 €) a vydelíme počtom hodnôt (19). Tento postup by bol platný pre výpočet rozptylu celej populácie, ale postup pre výpočet rozptylu vzorky je taký, že súčet druhých mocnín vydelíme počtom hodnôt mínus 1 (18). Formálny zápis postupu je nasledovný:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

Pre našu situáciu bude výpočet takýto:

$$s^2 = \frac{(15{,}9 - 20{,}84)^2 + (0 - 20{,}84)^2 + \cdots + (23{,}5 - 20{,}84)^2}{19 - 1} = \frac{8\,749{,}9}{18} \cong 486{,}11$$

Hodnota rozptylu pre premennú *peňaženka* je teda 486,11. Je však problematická jej zmysluplná interpretácia. Zmysluplnú interpretáciu získame tým, že hodnotu rozptylu odmocníme a nájdeme tak hodnotu smerodajnej odchýlky:

$$s = \sqrt{s^2} = \sqrt{486{,}11} \cong 22{,}05.$$

Číslo 22,05 vyjadruje to, ako sa priemerne všetky hodnoty premennej vzdialili od priemeru. Výpočet rozptylu a smerodajnej odchýlky by bol uvedeným postupom zdĺhavý hlavne pre väčšie súbory údajov, preto sa na ich zistenie používajú štatistické softvéry ako sme to urobili aj my v Tabuľka 8.

Príklad 2.3

Akú hodnotu by mal rozptyl a smerodajná odchýlka pre situáciu z príkladu 1.6, keď Dávid prišiel na záhradnú slávnosť so zaujímavým obnosom peňazí 1 000 €?[12]

[12] Výpočet môžeme zopakovať so zmenenou Dávidovou hodnotou a hodnota priemeru bude tentokrát 68,7. Treba ju odpočítať od každej hodnoty súboru údajov, následne umocniť na druhú, sčítať druhé mocniny (919 116,74) a vydeliť číslom 18. Získame tak hodnotu rozptylu s^2 = 51 062,04 a smerodajnej odchýlky s = 225,97. Pochopiteľne môžeme použiť štatistický softvér pre uľahčenie práce.

Porovnajme teraz hodnoty rozptylu a smerodajnej odchýlky pre dve situácie: situácia, keď Dávid prišiel na záhradnú slávnosť s obnosom 90,60 € a keď prišiel až s 1 000 € v peňaženke:

situácia 1: $s^2 = 486{,}11,\ s = 22{,}05$;

situácia 2: $s^2 = 51\,062{,}04,\ s = 225{,}97$.

Vidíme obrovský rozdiel v hodnotách ako rozptylu, tak aj smerodajnej odchýlky. Napriek tomu, že sme zmenili iba jednu hodnotu pôvodného dátového súboru, rozdiely v situáciách sú značné. Z toho plynú hneď tri závery:

- čím viac sú hodnoty od seba vzdialené, tým väčšie hodnoty rozptyl aj smerodajná odchýlka nadobúdajú;
- rozptyl aj smerodajná odchýlka (podobne ako priemer) sú veľmi „citlivé“ na extrémne odchýlené hodnoty;
- aritmetický priemer sám o sebe má len obmedzenú výpovednú hodnotu, ale doplnený napríklad o smerodajnú odchýlku poskytuje lepší obraz o tom, ako dátový súbor vyzerá aj bez toho, aby sme poznali všetky dáta.

Príklad 2.4

Aký rozptyl a smerodajnú odchýlku má dátový súbor, ktorého hodnoty sú všetky rovnaké, napr. 4, 4, 4, 4, 4?[13]

8. Medián vyjadruje stred dát, ale iným spôsobom ako aritmetický priemer.

V predchádzajúcom texte sme zistili, že aritmetický priemer (spolu s rozptylom a smerodajnou odchýlkou) sú citlivé na extrémne odchýlené hodnoty a zmena malého počtu

[13] Odpoveď môžeme získať na základe úvahy alebo na základe výpočtu. Úvaha by mohla znieť nasledovne: keďže sa hodnoty dátového súboru navzájom nijako nelíšia, čiže nie je medzi nimi nijaká variabilita (rôznorodosť), tak obe opisné štatistiky by sa mali rovnať 0. Pre potvrdenie našej úvahy vykonajme výpočet: keďže sú všetky hodnoty dátového súboru rovnaké, tak aj priemer bude to isté číslo (v našom konkrétnom prípade s dátovým súborom 4, 4, 4, 4, 4 je priemer tiež 4). Pri výpočte rozptylu je potrebné od každej hodnoty odpočítať priemer, čiže to isté číslo a vznikne 0. Druhá mocnina 0 je tiež 0, súčet všetkých 0, bude tiež 0. Ak 0 vydelíme ľubovoľným číslom (okrem 0) získame zasa 0. Takže rozptyl je skutočne 0 a druhá odmocnina z 0, čiže smerodajná odchýlka bude tak isto 0.

hodnôt, prípadne len jednej hodnoty, veľmi ovplyvní aj hodnoty priemeru, rozptylu a odchýlky. Bolo by preto užitočné, aby ak by sme mali k dispozícii aj opisnú štatistiku, ktorá by udávala stred údajov nezávislý od extrémne odchýlených hodnôt. Takouto opisnou štatistikou je medián, ktorý získame tak, že všetky hodnoty premennej usporiadame podľa veľkosti od najmenšej po najväčšiu a pozrieme sa, ktorá hodnota je presne v strede. Ak má súbor údajov rozsah *n* nepárny počet, hodnota mediánu sa nachádza na pozícii $(n + 1)/2$. V prípade párneho počtu údajov, hodnotu mediánu vypočítame ako aritmetický priemer dvoch hodnôt najbližšie k stredu dátového súboru, teda ako aritmetický priemer hodnôt na pozícii $(\frac{n}{2})$ a $(\frac{n}{2} + 1)$.

Zistime medián našej premennej *peňaženka.* Najskôr je potrebné usporiadať dáta podľa veľkosti. V Tabuľka 11 sme oproti Tabuľka 7 vložili aj stĺpec s poradovým číslom študenta, aby bolo zrejmé, že Jurajova hotovosť 0 € je najmenšia hodnota a Dávidových 90,60 € je najväčšia hodnota premennej *peňaženka*.

Tabuľka 11 Usporiadané hodnoty premennej *peňaženka*

por. číslo	***študent***	***peňaženka***
1	Juraj	0,00 €
2	Michal	0,50 €
3	Michaela	2,30 €
4	Ladislav	4,60 €
5	Ema	5,60 €
6	Katarína	7,20 €
7	Klaudia	11,30 €
8	Lívia	12,40 €
9	Lukáš	12,50 €
10	Alexandra	13,70 €
11	Peter	15,90 €
12	Matej	17,80 €
13	Emília	20,30 €
14	Slávka	23,50 €
15	Pavol	30,70 €
16	Lucia	31,20 €
17	Barbora	45,00 €
18	Filip	50,80 €

19	Dávid	90,60 €

Keďže máme k dispozícii 19 hodnôt, vieme ich rozdeliť jednou hodnotou na dve polovice: prvých 9 hodnôt predstavuje jednu polovicu (od Juraja po Lukáša), Alexandra je presne v strede a posledných 9 hodnôt (od Petra po Dávida) tvorí druhú polovicu hodnôt. Hodnota hotovosti, ktorú mala Alexandra v peňaženke: 13,70 €, predstavuje teda medián našej premennej a nachádza sa na 10. pozícii. Situácia je viditeľne znázornená v Tabuľka 12:

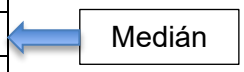
Tabuľka 12 Medián premennej *peňaženka*

por. číslo	***študent***	***peňaženka***
1	Juraj	0,00 €
2	Michal	0,50 €
3	Michaela	2,30 €
4	Ladislav	4,60 €
5	Ema	5,60 €
6	Katarína	7,20 €
7	Klaudia	11,30 €
8	Lívia	12,40 €
9	Lukáš	12,50 €
10	Alexandra	13,70 €
11	Peter	15,90 €
12	Matej	17,80 €
13	Emília	20,30 €
14	Slávka	23,50 €
15	Pavol	30,70 €
16	Lucia	31,20 €
17	Barbora	45,00 €
18	Filip	50,80 €
19	Dávid	90,60 €

Ak by na záhradnú slávnosť prišla ešte Marcela, ktorá by mala v peňaženke napríklad 25 €, tak by sme ju v usporiadanom dátovom súbore zaradili na 15. tu pozíciu tak, ako je to znázornené v Tabuľka 13. Medián sa v tomto prípade vypočíta ako aritmetický priemer hodnôt na 10. a 11. pozícii, teda $\frac{13{,}70+15{,}90}{2} = 14{,}80$.

Tabuľka 13 Usporiadané hodnoty a výpočet mediánu pre párny počet hodnôt premennej *peňaženka*

por. číslo	*študent*	*peňaženka*
1	Juraj	0,00 €
2	Michal	0,50 €
3	Michaela	2,30 €
4	Ladislav	4,60 €
5	Ema	5,60 €
6	Katarína	7,20 €
7	Klaudia	11,30 €
8	Lívia	12,40 €
9	Lukáš	12,50 €
10	Alexandra	13,70 €
11	Peter	15,90 €
12	Matej	17,80 €
13	Emília	20,30 €
14	Slávka	23,50 €
15	Marcela	25,00 €
16	Pavol	30,70 €
17	Lucia	31,20 €
18	Barbora	45,00 €
19	Filip	50,80 €
20	Dávid	90,60 €



Príklad 2.5

Aký by bol medián pre situáciu z príkladu 1.6, kedy bolo prítomných 19 študentov a Dávid prišiel na záhradnú slávnosť s 1 000 € v peňaženke?[14]

Podarilo sa nám zistiť, že hodnota mediánu nie je nijakým spôsobom ovplyvnená extrémnymi hodnotami premennej, na rozdiel od aritmetického priemeru, rozptylu a smerodajnej odchýlky. Preto sa často používa na vyjadrenie stredu hodnôt.

[14] Hodnota mediánu pre túto situáciu bude rovnaká, ako keď mal Dávid v peňaženke „len“ 90,60 €, pretože posledná hodnota, nech by sa zvýšila na akúkoľvek úroveň nijako neovplyvní to, že v strede usporiadaných hodnôt ostáva Alexandra so svojimi 13,70 €.

9. Kvartily sú hodnoty, ktoré rozdeľujú usporiadaný dátový súbor na štyri rovnako veľké časti**.** Tak ako medián rozdelí usporiadaný dátový súbor na rovnako veľké polovice, tak kvartily každú z polovíc ešte ďalej rozdelia na polovice. **Prvý kvartil**, ktorý označujeme $\mathbf{Q_1}$ tak rozdelí usporiadaný súbor dát na prvú štvrtinu a zvyšné tri štvrtiny, **tretí kvartil**, ktorý označujeme $\mathbf{Q_3}$ zasa rozdelí usporiadaný súbor dát na prvé tri štvrtiny a zvyšnú štvrtinu. Situáciu pre našu premennú *peňaženka* môžeme vidieť v Tabuľka 14. Prvý kvartil sa nachádza na 5. pozícii a má hodnotu 5,60 € a tretí kvartil je na 10. pozícii s hodnotou 30,70 €. Aj z Tabuľka 14 je zrejmé, že druhý kvartil je vlastne medián. Kvartily, analogicky ako medián nie sú ovplyvnené extrémne odchýlenými hodnotami. Výpočet kvartilov je užitočné ponechať na špecializovaný štatistický softvér, postačí porozumieť konceptu rozdelenia dátového súboru na štvrtiny.

Tabuľka 14 Prvý a tretí kvartil pre premennú *peňaženka*

por. číslo	*študent*	*peňaženka*	
1	Juraj	0,00 €	
2	Michal	0,50 €	
3	Michaela	2,30 €	
4	Ladislav	4,60 €	
5	Ema	5,60 €	← Prvý kvartil
6	Katarína	7,20 €	
7	Klaudia	11,30 €	
8	Lívia	12,40 €	
9	Lukáš	12,50 €	
10	Alexandra	13,70 €	← Medián / Tretí kvartil
11	Peter	15,90 €	
12	Matej	17,80 €	
13	Emília	20,30 €	
14	Slávka	23,50 €	
15	Pavol	30,70 €	
16	Lucia	31,20 €	
17	Barbora	45,00 €	
18	Filip	50,80 €	
19	Dávid	90,60 €	

Príklad 2.6

Môže sa stať, že by sa v nejakom dátovom súbore hodnota prvého a tretieho kvartilu rovnali?[15]

10. Medzikvartilové rozpätie (IQR = inter quartile range) je opisná štatistika, ktorá udáva vzdialenosť medzi kvartilmi a získame ju ako rozdiel tretieho a prvého kvartilu:

$$IQR = Q_3 - Q_1.$$

Pre našu premennú *peňaženka* je $IQR = 30{,}70 - 5{,}60 = 25{,}10$.

Cvičenia:

10. Dostali ste nasledujúci súbor údajov: 6, 3, 5, 2, 6, 4, 9. Ktoré z nasledujúcich tvrdení o tomto súbore údajov NIE JE pravdivé?

a) Medián je 5
b) Rozpätie je 7
c) Priemer je 6
d) Rozsah je 7

11. Ktoré z nasledujúcich tvrdení NIE JE pravdivé?

a) Medián možno použiť pri nominálnych a číselných údajoch.
b) Rozptyl sa zvyčajne používa na výpočet iných opisných štatistík.
c) Rozpätie aj priemer sú citlivé na extrémne odchýlené hodnoty.
d) Rozptyl je druhou odmocninou smerodajnej odchýlky.

12. K dispozícii máte údaje o platoch prvých desiatich najlepšie zarábajúcich tenistoch v tomto roku (2022) podľa The New York Times. Vypočítajte aritmetický priemer a smerodajnú odchýlku ročného platu týchto tenistov.

Rank	Player	Country	Points	Earnings in $
1	Rafael Nadal	Spain	6 515	2 159 790
2	Daniil Medvedev	Russia	8 435	1 649 130
3	Felix Auger-Aliassime	Canada	3 883	1 214 760
4	Denis Shapovalov	Canada	2 863	1 026 674
5	Stefanos Tsitsipas	Greece	6 565	992 345

[15] Áno, môže sa to stať. Takýto jav nastane v prípade, že by sa hodnoty premennej v druhej a tretej štvrtine rovnali, napríklad ak by dátový súbor vyzeral takto: 1, 4, 4, 4, 4, 4, 6.

6	Matteo Berrettini	Italy	4 928	954 828
7	Roberto Bautista Agut	Spain	2 585	793 391
8	Jannik Sinner	Italy	3 429	656 435
9	Pablo Carreno Busta	Spain	2 181	655 671
10	Diego Schwartzman	Argentina	2 865	628 251

13. Vypočítajte aritmetický priemer nasledujúcej množiny dát: 0,003; 0,045; 0,58; 0,687; 1,25; 10,38; 11,252; 12,001. Výsledok zaokrúhlite na tisíciny.

14. Ktorý z nasledujúcich dátových súborov má najvyššiu smerodajnú odchýlku (bez nutnosti počítať)?

a) 1, 2, 3, 4

b) 1, 1, 1, 4

c) 1, 1, 4, 4

d) 4, 4, 4, 4

e) 1, 2, 2, 4

15. Ktorá z nasledujúcich dátovým množín má aritmetický priemer 3 a smerodajnú odchýlku 1?

a) 1, 2, 3, 4, 5

b) 3, 3, 3, 3, 3

c) 2, 2, 3, 4, 4

d) 1, 1, 1, 1, 1

e) 0, 0, 3, 6, 6

16. Ktoré z nasledujúcich dátových množín majú rovnakú smerodajnú odchýlku ako dátová množina 1, 2, 3, 4, 5. (Vyriešte bez výpočtu).

a) 10, 20, 30, 40, 50

b) 0,1; 0,2; 0,3; 0,4; 0,5

c) 6, 7, 8, 9, 10

d) 1, 1, 3, 4, 4

17. Ktorá z nasledujúcich dátových množín má aritmetický priemer 15 a smerodajnú odchýlku 1?

a) 13, 14, 15, 16, 17

b) 15, 15, 15, 15, 15

c) 1, 1, 1, 1, 1

d) žiadna

e) 14, 14, 15, 16, 16

18. Porovnajte aritmetický priemer a smerodajnú odchýlku dvojíc dátových množín A a B (bez nutnosti počítať);

a) A: 3, 5, 5, 5, 8, 11, 11, 11, 13
B: 3, 5, 5, 5, 8, 11, 11, 11, 20

b) A: -20, 0, 0, 0, 15, 25, 30, 30
B: -40, 0, 0, 0, 15, 25, 30, 30

c) A: 0, 2, 4, 6, 8, 10
B: 20, 22, 24, 26, 28, 30

d) A: 100, 200, 300, 400, 500
B: 0, 50, 300, 550, 600

19. V teste zo štatistiky mali možnosť študenti získať maximálne 100 bodov. Náhodným výberom sme vybrali vzorku 20 študentov, ktorých bodové hodnotenie bolo nasledovné: 98 bodov: 2 študenti; 95 bodov: 1 študent; 92 bodov: 3 študenti; 88 bodov: 4 študenti; 87 bodov: 2 študenti; 85 bodov: 2 študenti; 79 bodov: 1 študent; 78 bodov: 2 študenti; 73 bodov: 1 študent; 72 bodov: 1 študent; 65 bodov: 1 študent. Aká je smerodajná odchýlka bodového hodnotenia vybraných študentov? Výsledok zaokrúhlite na desatiny.

20. Pracovníci na konkrétnom ťažobnom mieste majú priemerne 35 dní zaplatenej dovolenky za rok (v závislosti od odpracovaných rokov a iných okolností), ktorá je však nižšia ako celoštátny priemer. Vedúci tohto závodu je pod tlakom a miestny zväz baníkov ho núti zvýšiť počet dní platenej dovolenky každému baníkovi. To však vedúci urobiť nechce, pretože by to bolo nákladné. Namiesto toho sa rozhodne prepustiť 10 zamestnancov takým spôsobom, aby sa zvýšil priemerný počet dní dovolenky jeho zamestnancov. Ktorých zamestnancov by mal vedúci prepustiť, aby splnil svoj cieľ?

a) tých, ktorí majú najvyšší počet dní dovolenky

b) tých, ktorí majú najnižší počet dní dovolenky

c) tých, ktorí majú priemerný počet dní dovolenky (35).

21. Ktorý z nasledujúcich dátových súborov nemá medián 3?

a) 3; 3; 3; 3; 3

b) 2; 5; 3; 1; 1

c) 1; 4; 3; 4; 1

d) 1; 2; 5; 3; 4

22. Porovnajte medián a IQR dvojíc dátových množín X a Y (bez nutnosti počítať);

a) X: 3, 5, 6, 7, 9
Y: 3, 5, 6, 7, 20

b) X: 3, 5, 6, 7, 9
Y: 3, 5, 7, 8, 9

c) X: 1, 2, 3, 4, 5
Y: 6, 7, 8, 9, 10

d) X: 0, 10, 50, 60, 100
Y: 0, 100, 500, 600, 1000

23. Ktorá z nasledujúcich opisných štatistík je najmenej citlivá na extrémne odchýlené hodnoty?

a) aritmetický priemer

b) medián

c) variačné rozpätie

d) smerodajná odchýlka

24. Ktoré z nasledujúcich tvrdení sú pravdivé?

a) Medián a prvý kvartil nemôžu byť rovnaké hodnoty.

b) Maximum a minimum môžu byť rovnaké hodnoty.

c) Tretí kvartil má vždy väčšiu hodnotu ako prvý kvartil.

d) Variačné rozpätie a medzikvartilové rozpätie môžu byť rovnaké hodnoty.

25. Ktoré z nasledujúcich tvrdení **nie je** pravdivé?

a) Medián a prvý kvartil môžu byť rovnaké hodnoty.

b) Maximum a minimum môžu byť rovnaké hodnoty.

c) Prvý a tretí percentil môžu byť rovnaké hodnoty.

d) Variačné rozpätie a medzikvartitové rozpätie môžu byť rovnaké hodnoty.

e) Tretí kvartil má vždy väčšiu hodnotu ako prvý kvartil.

26. Ktoré z nasledujúcich tvrdení nie je pravdivé?

a) 50% všetkých hodnôt premennej leží medzi prvým a tretím kvartilom

b) 50% všetkých hodnôt premennej leží medzi mediánom a maximálnou hodnotou

c) 50% všetkých hodnôt premennej leží medzi mediánom a minimálnou hodnotou

d) 50% všetkých hodnôt premennej je menších alebo rovných ako medián

e) 50% všetkých hodnôt premennej je vždy menších ako aritmetický priemer

27. Ktoré z nasledujúcich tvrdení možno považovať za pravdivé?

a) Aritmetický priemer je vždy menší ako medián
b) Rozptyl je vždy väčší ako smerodajná odchýlka
c) Variačné rozpätie je vždy menšie ako medzikvartilové rozpätie
d) IQR je vždy menšie ako smerodajná odchýlka
e) Ani jedna zo zvyšných odpovedí nie je správna

2.1.2. Frekvenčná tabuľka pre numerickú premennú

Frekvenčná tabuľka je tabuľka, ktorá ukazuje ako často sa jednotlivé hodnoty v dátach vyskytujú. V prípade numerickej **diskrétnej premennej, ktorá nenadobúda veľa hodnôt**, môžeme zostrojiť tabuľku, kde v riadkoch budú hodnoty premennej a v stĺpcoch zaznamenáme početnosti: absolútnu, relatívnu, prípadne aj kumulatívnu.
Vezmime napríklad numerickú diskrétnu premennú *súrodenci*, ktorá predstavuje počet súrodencov študentov predmetu Analýza dát, ktorí sa zúčastnili záhradnej slávnosti.
Z dátového súboru v Tabuľka 15 môžeme zistiť, že napríklad Peter je jedináčik a Emília žije v rodine so štyrmi súrodencami.

Tabuľka 15 Dátový súbor pre numerickú diskrétnu premennú *súrodenci*

študent	***súrodenci***
Peter	0
Juraj	1
Ema	3
Klaudia	1
Alexandra	2
Emília	4
Dávid	2
Pavol	1
Barbora	1
Michal	1
Michaela	2
Ladislav	2
Katarína	0
Lívia	3
Lukáš	2
Matej	0
Lucia	2

Filip	1
Slávka	2

Ak by sme chceli získať obraz o dátovom súbore pomocou opisných štatistík, použili by sme štatistický softvér. Iný spôsob, ako získať prehľad, aké hodnoty a ako často sa v dátovom súbore nachádzajú, je použitie **frekvenčnej tabuľky**. Okrem samotných hodnôt premennej sa vo frekvenčnej tabuľke nachádzajú **absolútna, relatívna a kumulatívna početnosť**. Frekvenčná tabuľka pre premennú *súrodenci* so všetkými typmi početností je v Tabuľka 16.

Absolútna početnosť vyjadruje počet (uvádzaný v celých nezáporných číslach) koľkokrát sa daná hodnota v dátovom súbore vyskytla a označuje sa $\mathbf{n_i}$**.** Napríklad študentov, ktorí majú dvoch súrodencov je 7. Frekvenčná tabuľka obsahuje aj riadok, ktorý označuje celkový súčet, ktorý je v našom prípade 19.

Relatívna početnosť vyjadruje, akú časť alebo percento tvorí daná hodnota z celkového počtu. Relatívnu početnosť môžeme uvádzať v reálnych číslach z intervalu $\langle 0,1\rangle$, označujeme ju $\mathbf{p_i}$ a vypočítame ako podiel absolútnej početnosti hodnoty premennej a celkového počtu hodnôt, teda $p_i = \frac{n_i}{n}$. Relatívnu početnosť môžeme uvádzať aj percentách a v tom prípade ju označujeme $\mathbf{f_i}$ a vypočítame ju $f_i = \frac{n_i}{n}.100\%$. Relatívnu početnosť pre hodnotu 2 premennej *súrodenci* získame tak, že vydelíme absolútnu početnosť (7) celkovým počtom hodnôt (19): $p_3 = \frac{7}{19} = 0{,}368$. Relatívna početnosť vyjadrená v percentách potom bude 36,8%.

Kumulatívna početnosť vyjadruje**,** koľko údajov je menších alebo rovnakých ako daná hodnota. Označujeme ju napríklad $\mathbf{kn_i}$ v prípade, že nás zaujíma absolútna kumulatívna početnosť. Napríklad absolútna kumulatívna početnosť pre hodnotu 2 súrodencov znamená, že sa pýtame, koľko údajov je menších alebo rovných 2. Potrebujeme spočítať koľko je núl, teda jedináčikov (3), koľko jednotiek (6), teda študentov s jedným súrodencom a aj dvojok (7), teda študentov s dvomi súrodencami, čo je dohromady 16.

Tabuľka 16 Frekvenčná tabuľka pre diskrétnu numerickú premennú *súrodenci*

	$\mathbf{n_i}$	$\mathbf{p_i}$	$\mathbf{f_i}$	$\mathbf{kn_i}$
0	3	0,158	15,8%	3
1	6	0,316	31,6%	9

2	7	0,368	36,8%	16
3	2	0,105	10,5%	18
4	1	0,053	5,3%	19
spolu	19	1,00	100,0%	

Premenná *súrodenci* je síce numerická premenná, avšak je pomerne špecifická v tom, že je diskrétna a zároveň má iba 5 rôznych hodnôt (0, 1, 2, 3, 4). Ak však máme numerickú premennú, ktorá je **spojitá,** zvyčajne nadobúda veľa rôznych hodnôt. Napríklad v premennej *peňaženka* je každá hodnota zastúpená práve raz a frekvenčná tabuľka, ktorá by zaznamenávala početnosť hodnôt by nemala nijaký zmysel. V tomto prípade je vhodné zostrojiť **intervaly** a následne zisťovať, koľko hodnôt premennej sa nachádza v príslušnom intervale. **Intervaly pre spojitú premennú** možno tvoriť tak, že majú rovnakú šírku**,** určený optimálny počet intervalov podľa veľkosti vzorky alebo podľa logických hraníc daného problému.
To platí aj pre premennú peňaženka, ktorá je numerická spojitá a v tomto prípade má zmysel nastaviť intervaly, ktorých dĺžka bude 10. Intervaly, ktoré používame sú polouzavreté, čo znamená, že ľavá hranica intervalu doňho nepatrí a pravá hranica intervalu patrí. Finálna frekvenčná tabuľka (Tabuľka 17) pre premennú *peňaženka* bude nasledovná:

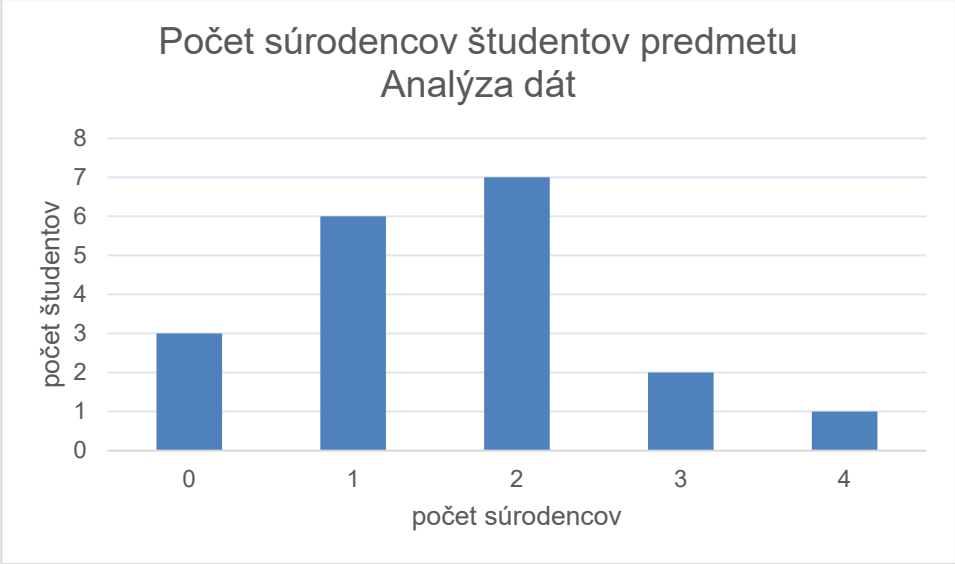
Tabuľka 17 Frekvenčná tabuľka pre spojitú numerickú premennú *peňaženka*

	ni	**pi**	**fi**	**kni**
$\langle 0, 10\rangle$	6	0,316	31,6%	6
$(10, 20\rangle$	6	0,316	31,6%	12
$(20, 30\rangle$	2	0,105	10,5%	14
$(30, 40\rangle$	2	0,105	10,5%	16
$(40, 50\rangle$	1	0,053	5,3%	17
$(50, 60\rangle$	1	0,053	5,3%	18
$(60, 70\rangle$	0	0,000	0,0%	18
$(70, 80\rangle$	0	0,000	0,0%	18
$(80, 90\rangle$	0	0,000	0,0%	18
	1	0,053	5,3%	19

(90, 100⟩				
spolu	19	1,000	100,0%	

2.1.3. Grafické znázornenie numerickej premennej

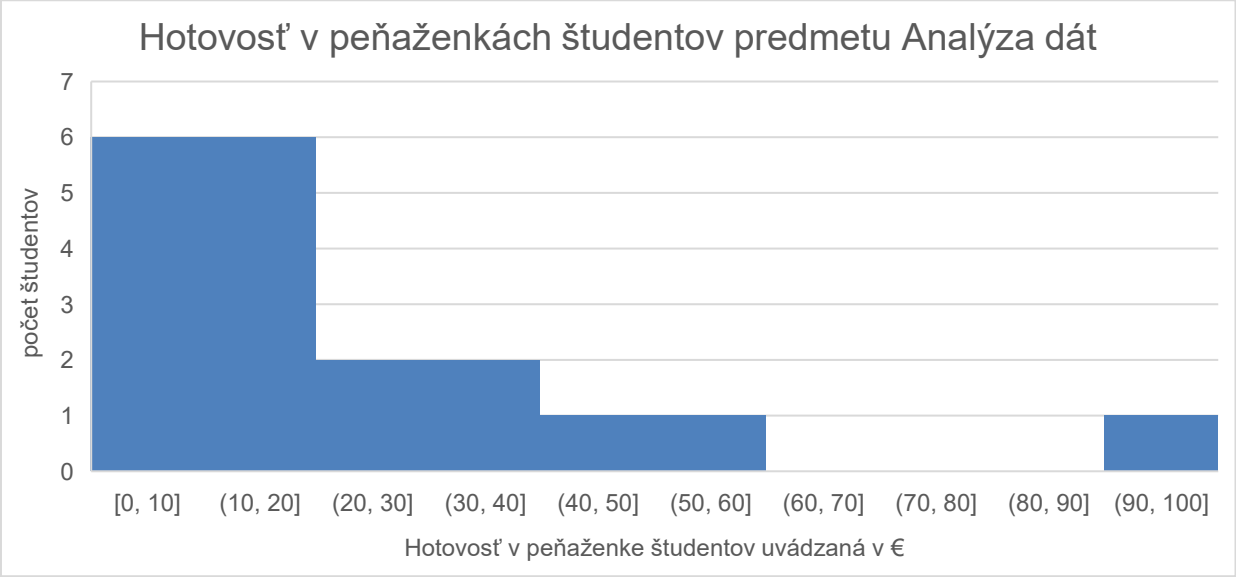
Grafické znázornenie akejkoľvek premennej vychádza z frekvenčnej tabuľky. Analogicky ako v časti 2.1.2 budeme uvažovať o dvoch typoch numerických premenných. V prvom prípade ide o **numerickú diskrétnu premennú, ktorá má relatívne malý počet hodnôt**. V tom prípade je vhodné na jej grafické zobrazenie použiť **stĺpcový graf**, kde na x-ovej osi budú hodnoty premennej a na y-ovej osi bude buď absolútna, relatívna alebo kumulatívna početnosť. Grafické znázornenie takejto premennej je na Obrázok 5, kde je znázornená premenná *súrodenci* a vychádza z frekvenčnej tabuľky tejto premennej (Tabuľka 16).



Obrázok 5 Stĺpcový graf pre absolútnu početnosť premennej *súrodenci*

V druhom prípade ide o **numerickú spojitú premennú,** kde sme jej hodnoty zaradili do intervalov. V tomto prípade použijeme na grafické znázornenie špeciálny typ grafu, ktorý je určený na grafické znázornenie tohto typu premenných. Jeho názov je **histogram,** ktorý je odvodený z gréckych slov: "histos" (ἱστός) = *stĺp, kolmá tyč, stojan* a "gramma" (γράμμα) = *záznam, kresba, grafický prejav*. Čiže histogram je graf, ktorý zobrazuje údaje pomocou stĺpcov, kde výška stĺpca predstavuje početnosť hodnôt v intervale. Intervaly používané v histograme sú polouzavreté a plynulo na seba nadväzujú a jednotlivé stĺpce v histograme

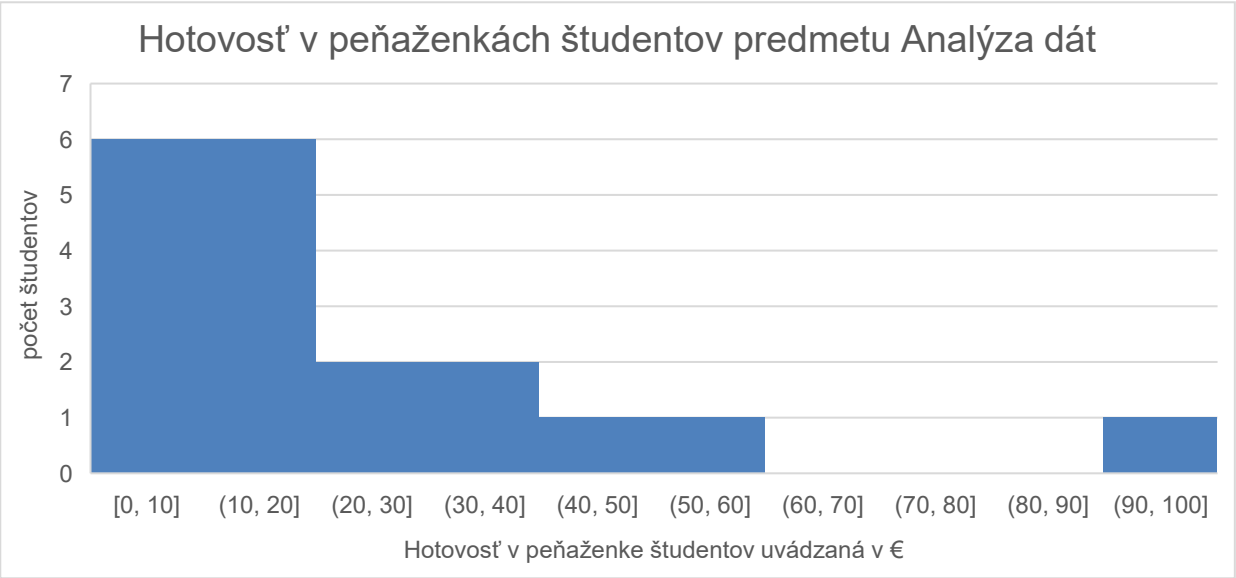
majú rovnakú vlastnosť, teda plynulo na seba nadväzujú (nie sú medzi nimi medzery). Intervaly sa v histograme nachádzajú na x-ovej osi a na y-ovej osi sa nachádza absolútna, relatívna alebo kumulatívna početnosť. Na Obrázok 6 je znázornený histogram spojitej numerickej premennej *peňaženka*.



Obrázok 6 Histogram spojitej numerickej premennej *peňaženka*

Z grafického znázornenia numerických premenných, či už ide o stĺpcový graf alebo histogram, možno vyčítať mnohé skutočnosti. Okrem samotných početností je možné jednoducho zistiť hodnotu minima, či maxima, ale aj napríklad odhadnúť aj hodnotu aritmetického priemeru a mediánu. V prípade, že graf je približne symetrický, hodnota aritmetického priemeru a mediánu je tiež približne rovnaká.

V prípade premennej *peňaženka* je zrejmé, že graf nie je symetrický a extrémne odchýlená hodnota (Dávidových 90,60 €) spôsobila, že hodnota aritmetického priemeru sa posunula smerom ku nej. Hodnota aritmetického priemeru bola 20,84 € a mediánu iba 13,70 € (na základe výpočtov v časti 2.1.1), ako je to znázornené na Obrázok 7.



Obrázok 7 Porovnanie hodnoty aritmetického priemeru a mediánu

V poradí tretím spôsobom ako graficky znázorniť hodnoty numerickej premennej je graf so slovenským názvom **škatuľkový** alebo **škatuľový graf**, často sa však používa jeho anglický ekvivalent **boxplot.** Ide o špecifickú formu grafického znázornenia, ktorá narába s tzv. päťbodovým zhrnutím alebo inak povedané s nasledujúcimi piatimi opisnými štatistikami: minimum, prvý kvartil, medián, tretí kvartil a maximum. Postup pri jeho tvorbe je nasledujúci:

1. usporiadame hodnoty premennej podľa veľkosti od najmenšej po najväčšiu
2. zistíme **päťbodové zhrnutie**
3. zistíme hodnotu medzikvartilového rozpätia **IQR**
4. nájdeme **hranice pre extrémne hodnoty**:

   dolná hranica (DH) = $Q_1 - 1{,}5.IQR$

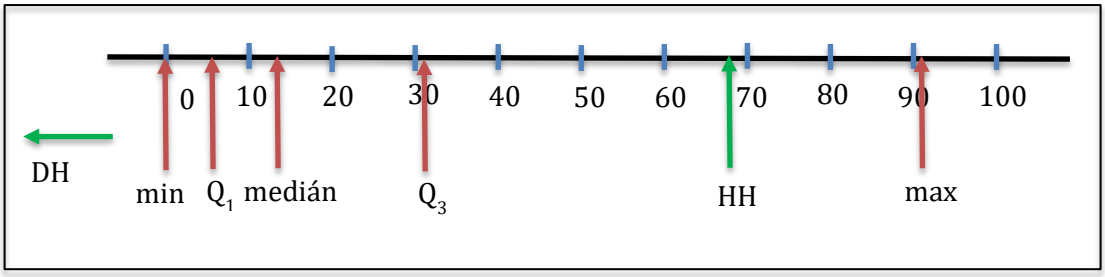
   horná hranica (HH) = $Q_3 + 1{,}5.IQR$
5. nakreslíme **číselnú os** vodorovnú alebo zvislú, na ktorej vyznačíme päťbodové zhrnutie a hranice boxplotu
6. nad číselnou osou alebo vedľa číselnej osi nakreslíme **obdĺžnik** (škatuľka alebo box), ktorého dve strany sú nad, resp. vedľa kvartilov; v obdĺžniku vyznačíme čiaru v úrovni mediánu

7. po oboch stranách obdĺžnika nakreslíme „fúzy“ (whiskers), čiže úsečky rovnobežné s číselnou osou
8. porovnáme hodnotu minima a dolnej hranice pre extrémne hodnoty, ak je minimum väčšie alebo rovné dolnej hranici, tak ľavú (dolnú) úsečku ukončíme na úrovni minima, ak je minimum menšie ako dolná hranica, tak úsečku ukončíme na úrovni dolnej hranice a všetky hodnoty premennej, ktoré sú menšie ako dolná hranica špeciálne vyznačíme, napríklad hviezdičkou;
9. analogicky postupujeme s maximom a hornou hranicou pre extrémne hodnoty: porovnáme hodnotu maxima a hornej hranice, aj je maximum menšie alebo rovné hornej hranici, tak úsečku ukončíme na úrovni maxima, ak je maximum väčšie ako horná hranica, tak úsečku ukončíme na úrovni hornej hranice a všetky hodnoty, ktoré sú väčšie ako horná hranica špeciálne vyznačíme, napríklad hviezdičkou.

Nakreslime boxplot pre numerickú spojitú premennú *peňaženka*. Budeme postupovať podľa 9 bodového návodu opísaného vyššie, aby sme objasnili pointu tohto typu grafu. Prirodzene pre praktické dôvody sa na tvorbu boxplotu používajú štatistické softvéry.

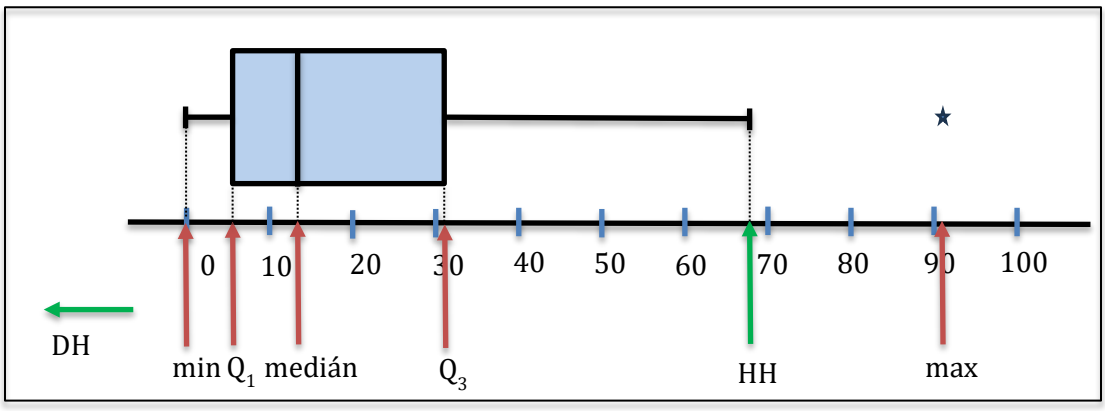
1. usporiadané hodnoty premennej sú v Tabuľka 11;
2. päťbodové zhrnutie môžeme zistiť napríklad z Tabuľka 8, kde sú uvedené všetky opisné štatistiky premennej: minimum = 0, Q_1 = 5,6, medián = 13,7, Q_3 = 30,7, maximum = 90,6;
3. $\boldsymbol{IQR = Q_3 - Q_1 = 30,7 - 5,6 = 25,1}$
4. $\boldsymbol{DH = Q_1 - 1,5.IQR = 5,6 - 1,5.25,1 = -32,5}$
   $\boldsymbol{HH = Q_3 + 1,5.IQR = 30,7 + 1,5.25,1 = 68,35}$
   dolná a horná hranica pre extrémne hodnoty vyjadruje to, že všetky hodnoty, ktoré by boli menšie ako dolná hranica alebo väčšie ako horná hranica, teda sú vzdialenejšie od kvartilov viac ako o 1,5 násobok medzikvartilového rozpätie IQR, sú naozaj extrémne a treba im venovať špeciálnu pozornosť;
5. môžeme pristúpiť ku kresleniu číselnej osi, my zvolíme vodorovnú číselnú os a tým aj vodorovný boxplot, ale rovnako dobre by fungovala aj zvislá číselná os a zvislý boxplot;

na číselnej osi vyznačíme dôležité hodnoty päťbodového zhrnutia a hraníc pre extrémne hodnoty:

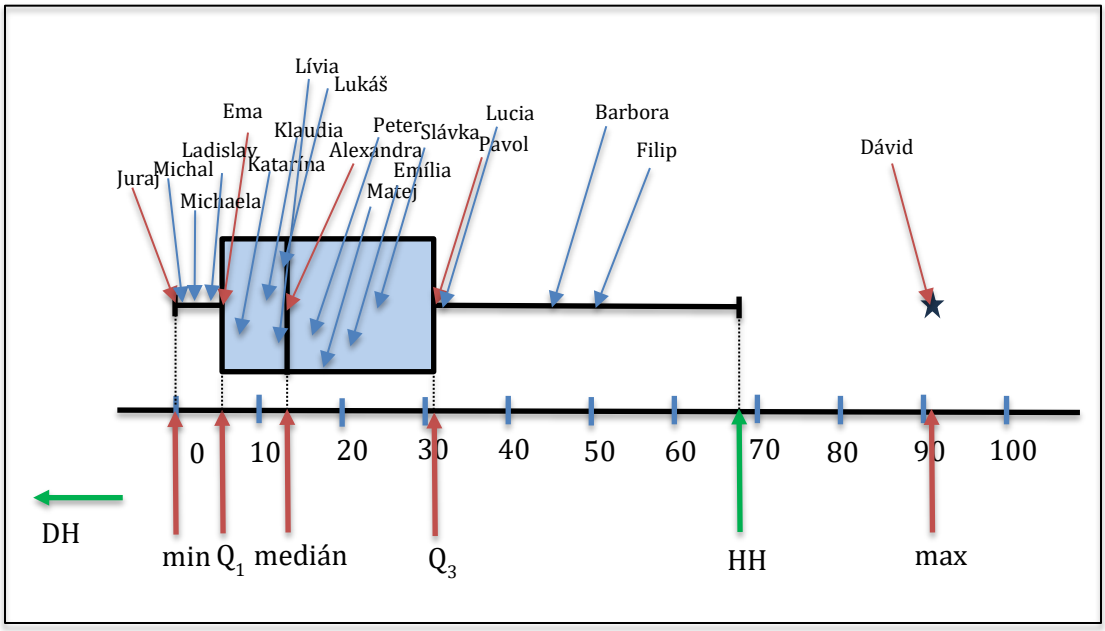


Obrázok 8 Číselná os pri tvorbe boxplotu

6. **– 9.** finálny boxplot vyzerá ako na Obrázok 9; obdĺžnik boxplotu je široký jedno IQR a jeho zvislé strany sú na úrovni kvartilov; keďže dolná hranica pre extrémne hodnoty je záporná a nachádzala by sa v našom obrázku viac vľavo, je iba naznačená zelenou šípkou; z obrázku je zrejmé, že hodnota minima je väčšia ako dolná hranica, preto je úsečka vychádzajúca z boxplotu (fúzik) ukončená na úrovni minima; keďže maximum premennej (Dávidových 90,60 €) je až za hornou hranicou, je táto hodnota špeciálna a preto sme ju označili hviezdičkou a predstavuje tzv. **extrémne odchýlenú hodnotu** (v angličtine **outlier);** v tomto prípade sme úsečku (fúzik) ukončili na úrovni hornej hranice; pre lepšie pochopenie situácie, čo vlastne boxplot zobrazuje si môžeme pomôcť Obrázok 10, na ktorom sú spolu s boxplotom znázornené aj hodnoty, resp. študentov, ktorí predstavujú jednotlivé hodnoty; ak je napríklad obdĺžnik boxplotu rozdelený mediánovou čiarou na dve časti, ktoré nie sú rovnako veľké, znamená to, že v tej menšej časti sú hodnoty „viac nahusto“, pretože v každej časti boxplotu sa nachádza 25 % všetkých hodnôt.



Obrázok 9 Boxplot pre premennú *peňaženka*



Obrázok 10 Boxplot spolu s vyznačením hodnôt premennej

Pre lepšiu predstavu, ako boxplot vznikol si môže čitateľ pozrieť video, kde študenti predmetu Analýza dát vytvorili živý boxplot a sledovať v priamom prenose všetkých 9 bodov tvorby boxplotu na záhrade Ústavu mediamatiky, Fakulty sociálnych a ekonomických vied Univerzity Komenského na ulici P. O. Hviezdoslava v Martine.

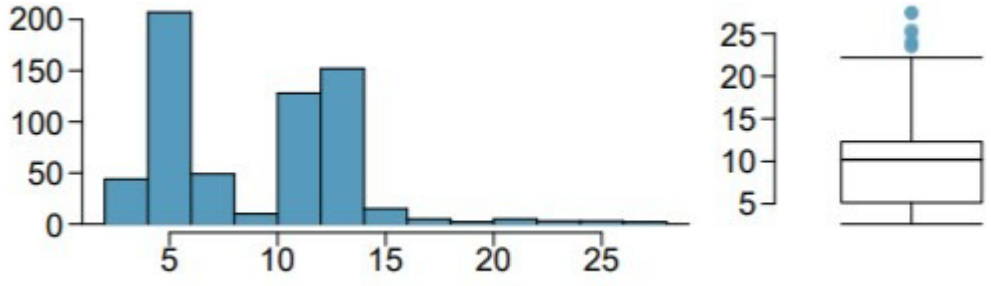
Cvičenia

28. 18 študentov získalo z 15 bodového testu zo štatistiky nasledujúce počty bodov. Určte všetky opisné štatistiky a vytvorte histogram a škatuľkový graf.

Adam	5
Božena	4
Dušan	10
Eva	4
Fero	4

Henrieta	7
Jakub	4
Janka	5
Jano	5
Jarmila	12
Karol	5
Matej	2
Milan	6
Renáta	5
Tomáš	6
Veronika	6
Viktor	3
Zlatica	3

29. Ktoré opisné štatistiky možno vyčítať aj zo škatuľkového grafu a z histogramu nie a naopak?



30. Ktorú z nasledujúcich opisných štatistík možno vyčítať aj zo škatuľkového grafu, aj z histogramu:

a) medián
b) IQR
c) tretí kvartil
d) minimum
e) prvý kvartil

31. Róbert sa pokúšal vypočítať 5 bodové zhrnutie z počtu dosiahnutých bodov na skúške a zistil nasledujúce údaje:

Minimum= 30
Maximum = 90
Prvý kvartil = 50
Tretí kvartil = 80

Medián = 85

Čo je na Róbertových výpočtoch nesprávne?

32. Tu sú výsledky dvadsiatich študentov z testu zo štatistiky: 57, 66, 69, 71, 72, 73, 74, 77, 78, 78, 79, 79, 81, 81, 82, 83, 83, 88, 89, 94. Aké je 5-bodové zhrnutie premennej? Ako vyzerá histogram a škatuľkový graf?

2.2. Opisná štatistika pre kategoriálnu premennú

Analogicky ako pre numerickú premennú študenti predmetu Analýza dát postupovali aj pri kategoriálnej premennej a rozhodli sa preskúmať to, aké ročné obdobie obľubujú. Počas záhradnej slávnosti zisťovali okrem stavu peňaženky a počtu súrodencov aj obľúbené ročné obdobie a vznikol nasledujúci súbor dát:

Tabuľka 18 Dátový súbor pre premennú ročné obdobie

študent	***ročné obdobie***
Peter	zima
Juraj	leto
Ema	leto
Klaudia	jar
Alexandra	jeseň
Emília	jar
Dávid	leto
Pavol	leto
Barbora	leto
Michal	zima
Michaela	jar
Ladislav	leto
Katarína	leto
Lívia	jeseň
Lukáš	jar
Matej	leto
Lucia	leto
Filip	zima
Slávka	leto

2.2.1. Opisné štatistiky pre kategoriálnu premennú

Pre kategoriálnu premennú, či už nominálnu alebo ordinálnu používame oproti numerickej premennej len malý počet opisných štatistík. Nejde tu dokonca len o jedno číslo, ktoré by charakterizovalo celú premennú, ako napríklad aritmetický priemer, či smerodajná odchýlka, ale zisťujeme **absolútnu a relatívnu početnosť jednotlivých kategórií**. V tomto prípade sú

teda absolútna a relatívna početnosť kategórií **opisnými štatistikami pre kategoriálnu premennú**.[16]

Keďže absolútna a relatívna početnosť sú súčasťou frekvenčnej tabuľky, venujeme sa im v nasledujúcej časti 2.2.2.

2.2.2. Frekvenčná tabuľka pre kategoriálnu premennú

Frekvenčná tabuľka pre kategoriálnu premennú obsahuje v riadkoch názvy jednotlivých kategórií, čo budú pre nás v tomto prípade hodnoty premennej. V stĺpcoch sa potom nachádzajú absolútna početnosť $\mathbf{n_i}$ a relatívna početnosť $\mathbf{p_i}$ a $\mathbf{f_i}$. Ich vysvetlenie a výpočet sa nachádzajú v časti 2.1.2. V prípade, že kategoriálna premenná je ordinálna môžeme zvážiť zaradenie aj kumulatívnej početnosti.

Frekvenčná tabuľka pre premennú *ročné obdobie* je takáto:

Tabuľka 19 Frekvenčná tabuľka pre kategoriálnu premennú

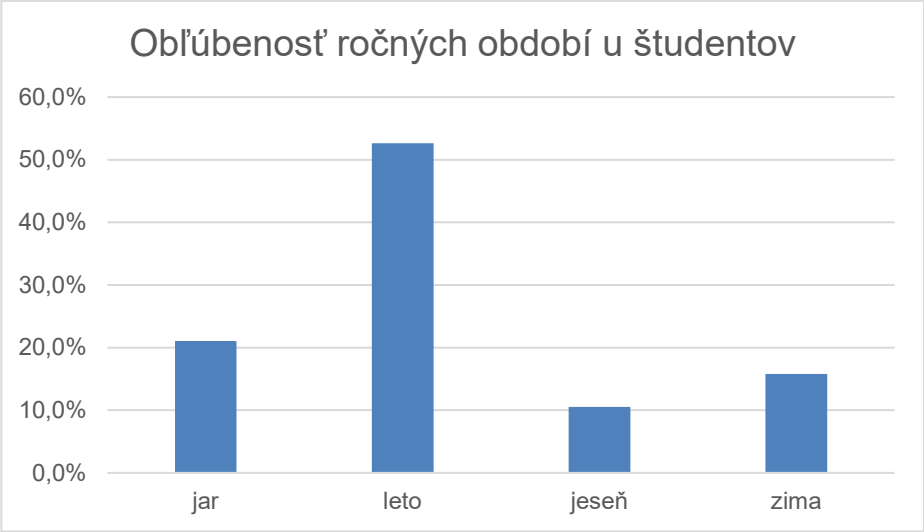
	$\mathbf{n_i}$	$\mathbf{p_i}$	$\mathbf{f_i}$
jar	4	0,211	21,1%
leto	10	0,526	52,6%
jeseň	2	0,105	10,5%
zima	3	0,158	15,8%
spolu	19	1,000	100,0%

2.2.3. Grafické znázornenie kategoriálnej premennej

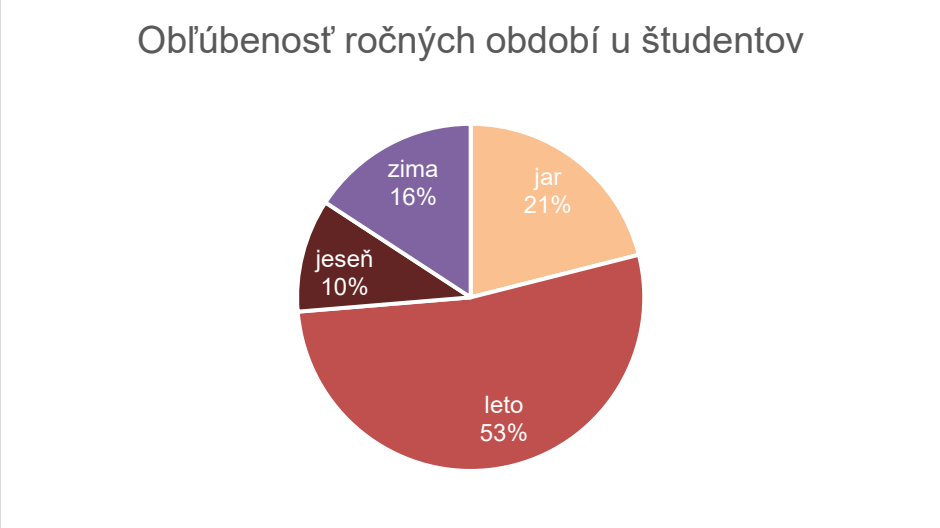
Grafické znázornenie kategoriálnej premennej bez ohľadu na to, či ide o nominálnu alebo ordinálnu premennú vychádza z frekvenčnej tabuľky a môže byť buď vo forme stĺpcového grafu alebo koláčového grafu. Špeciálne pri koláčovom grafe treba mať na pamäti, že ho používame hlavne vtedy, ak chceme zdôrazniť, akú časť celku tvorí nejaká kategória alebo kategórie a vyslovene nie je vhodný, ak je počet kategórií vyšší (podľa odporúčaní informačných dizajnérov by sme mali na koláčový graf zabudnúť, ak je počet kategórií vyšší ako 5).

[16] Existujú náročnejšie opisné štatistiky, ktoré vyjadrujú mieru variability kategórií kategoriálnej premennej, ale v našej učebnici sa im nebudeme špeciálne venovať. Ak by čitateľa táto téma zaujala, môže použiť pokročilejšie učebnice zaoberajúce sa analýzou dát a nájsť v nich informácie napríklad o indexe diverzity, Herfindahlov–Hirschmanovom indexe, či o Shannonovom index entropie.

Premennú *ročné obdobie* sme graficky zobrazili pomocou stĺpcového grafu na Obrázok 11 a pomocou koláčového grafu na Obrázok 12.



Obrázok 11 Stĺpcový graf pre kategoriálnu premennú *ročné obdobie*



Obrázok 12 Koláčový graf pre kategoriálnu premennú *ročné obdobie*

3. Základy teórie pravdepodobnosti

Pravdepodobnosť tvorí základ štatistiky a pravdepodobne čitateľ už pozná mnohé myšlienky uvedené v tejto kapitole. Je však celkom možné, že formalizácia pojmov pravdepodobnosti je pre väčšinu čitateľov nová. Táto kapitola poskytuje teoretický základ pre myšlienky v ďalších kapitolách a poskytuje cestu k ich hlbšiemu pochopeniu.
Čitateľ sa už určite stretol s otázkami typu: aká je šanca, že hodím pri hode klasickou kockou šestku? Alebo aká je šanca, že v hre Osadníci z Catanu padne na dvoch klasických kockách súčet sedem? Alebo aká je šanca, že vyhrám v lotériách Loto, Keno 10, či Joker?

3.1. Náhodný pokus, náhodný jav a priestor náhodných javov

Vo všetkých prípadoch hovoríme o náhodných javoch a pýtame sa na šancu alebo pravdepodobnosť, s akou nastanú. **Náhodný pokus** je taký pokus, ktorého výsledok závisí na náhode a ktorý môžeme za tých istých podmienok zopakovať ľubovoľne mnohokrát a **náhodný jav** je výsledok takéhoto náhodného pokusu. Náhodné javy budeme v nasledujúcom označovať veľkými písmenami A, B, atď. **Priestor náhodných javov** je množina všetkých možných výsledkov náhodného pokusu.
Napríklad hádzanie mincou je náhodný pokus, ktorý môžeme pri dodržaní približne rovnakých podmienok opakovať ľubovoľne veľakrát. Hodenie hlavy je náhodný jav, rovnako ako hodenie znaku. Priestor náhodných javov pri hádzaní mincou obsahuje dva náhodné javy: hlava a znak. Analogicky je to pri hádzaní klasickou kockou, čo predstavuje náhodný pokus za predpokladu, že kocka nie je nijakým spôsobom poškodená, či upravená. Náhodný jav je potom napríklad hodenie jednotky, šestky alebo iného čísla. Priestor náhodných javov pri hádzaní jednou klasickou kockou obsahuje šesť náhodných javov: hodenie jednotky, dvojky, trojky, štvorky, päťky a šestky.[17]

[17] Skutočnosť, že hádžeme klasickou kockou opakujeme preto, lebo existuje mnoho spoločenských a iných hier, kde sa používajú „kocky“ s iným počtom stien ako 6. Napríklad pri hre Dungeons & Dragons sa používajú kocky (správne by sme mali hovoriť mnohosteny), ktoré majú 4, 6, 8, 10, 12 a 20 stien.

3.2. Klasická definícia pravdepodobnosti

Príklad 3.1

Aká je šanca, že pri hode mincou hodíme znak?[18]

Príklad 3.2

Aká je šanca, že pri hre „Človeče nehnevaj sa“ sa mi podarí hodiť šestku?[19]

To, čo sme v predchádzajúcich príkladoch nazvali šanca, budeme nazývať **teoretická pravdepodobnosť** a budeme vyjadrovať ako pomer počtu priaznivých možností *m* a počtu všetkých možností *n*. Nech je A náhodný jav, potom jeho pravdepodobnosť P(A) vypočítame:

$$P(A) = \frac{počet\ priaznivých\ možností}{počet\ všetkých\ možností} = \frac{m}{n}$$

Ak budeme hrať spoločenskú hru Osadníci z Catanu, kde sa hádže súčasne dvomi klasickými hracími kockami, tak je nás zaujíma súčet, ktorý padol na kockách. Veľmi zaujímavá herná situácia nastáva, ak padne súčet 7 a prichádza na scénu tzv. zlodej.[20] Čo je v tomto prípade náhodný pokus, náhodný jav, priestor náhodných javov a aká je pravdepodobnosť, že nastúpi zlodej do akcie?

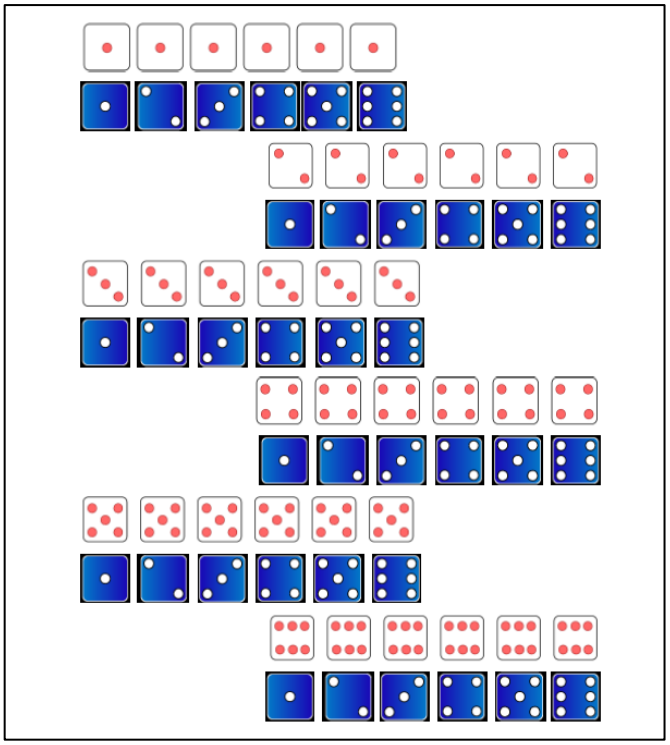
Náhodný pokus, ktorý nielen počas hry môžeme opakovať mnohokrát je hádzanie dvomi kockami, náhodný jav je hodenie súčtu 7 (označme tento náhodný jav A) a priestor náhodných javov bude obsahovať všetky možnosti pri hádzaní dvomi kockami. Pre lepšiu predstavu o všetkých možnostiach zvolíme kocky rôznych farieb (bielu a modrú), aj keď farba kocky nijako

[18] Aj bez špeciálnych teoretických základov vieme povedať, že šanca hodiť znak sa bude rovnať ½ alebo 50%.

[19] Opäť jednoduchá úloha vychádzajúca z toho, že počet všetkých možností pri hode klasickou kockou je 6 a nás zaujíma práve jedna z nich a to šestka. Preto Šanca, že hodíme šestku je $\frac{1}{6}$ alebo 0,167 (po zaokrúhlení) alebo 16,7%.

[20] Toto nie je platená reklama, ani iná, ale vrelo odporúčam zahrať si túto hru, ktorú navrhol zubný technik Klaus Teuber okrem iného s cieľom relaxu a úniku od stresu po náročnej práci.

neovplyvní priestor náhodných javov. Ak hodím na bielej 1, tak na modrej mám 6 možností hodu, ak na bielej hodím 2, tak na modrej mám opäť 6 možností hodu, atď. Všetkých možností a zároveň počet prvkov priestoru náhodných javov je 6 . 6 = 36 a najlepšiu predstavu získame z Obrázok 13. Pravdepodobnosť, že začne úradovať zlodej, teda, že hodíme súčet 7, vypočítame tak, že najskôr zistíme, koľko je priaznivých možností: $m = 6$[21] a všetkých možností $n = 36$. Pravdepodobnosť hodu 7 je: $P(A) = \frac{6}{36} = \frac{1}{6} \cong 0{,}167$ alebo inak povedané, máme 16,7%-nú šancu, že pri nejakom hode padne súčet 7 a budeme sa musieť v hre zmieriť s následkami úradovania zlodeja.



Obrázok 13 Priestor náhodných javov pri hode dvomi kockami

[21] To, že priaznivých možností je 6 zistíme buď z Obrázok 13 alebo vypísaním všetkých možností:
(1, 6), (6, 1), (2, 5), (5, 2), (3, 4), (4, 3).

Príklad 3.3

Aká je pravdepodobnosť, že pri hode klasickou kockou padne párne číslo?[22]

3.3. Disjunktné javy a výpočet ich pravdepodobnosti

Dva náhodné javy sa nazývajú **disjunktné** alebo **vzájomne sa vylučujúce**, ak nemôžu nastať oba naraz. Napríklad ak hodíme kockou, hod 1 a 2 sú disjunktné, pretože na kocke nemôžeme súčasne hodiť 1 a aj 2. Na druhej strane, hod 1 a "hod nepárneho čísla" nie sú disjunktné, pretože oba nastanú, ak je výsledkom hodu 1. Pojmy *disjunktné* a *vzájomne sa vylučujúce* sú ekvivalentné a zameniteľné.

Aká je pravdepodobnosť, že pri hode klasickou kockou hodíme 1 alebo 2?

Výpočet pravdepodobnosti vzájomne sa vylučujúcich javov je jednoduchý. Pri hode kockou sú výsledky 1 a 2 oddelené a pravdepodobnosť, že jeden z týchto výsledkov nastane, vypočítame tak, že ich jednotlivé pravdepodobnosti spočítame:

$$P\ (1 \text{ alebo } 2) = P\ (1) + P\ (2) = 1/6 + 1/6 = 1/3$$[23]

A čo pravdepodobnosť, že padne 1, 2, 3, 4, 5 alebo 6? Aj v tomto prípade sú všetky výsledky disjunktné, takže pravdepodobnosti sčítame:

$$P\ (1 \text{ alebo } 2 \text{ alebo } 3 \text{ alebo } 4 \text{ alebo } 5 \text{ alebo } 6) =$$

$$= P\ (1) + P\ (2) + P\ (3) + P\ (4) + P\ (5) + P\ (6) =$$

$$= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1$$

Pri **výpočte pravdepodobnosti disjunktných** (vzájomne sa vylučujúcich javov) sa riadime **pravidlom sčítania:**

> Nech A_1 a A_2 sú disjunktné javy, potom pravdepodobnosť, že nastane A_1 alebo A_2 (nastane aspoň jeden z nich) vypočítame ako súčet jednotlivých pravdepodobností javov:

---

[22] Pravdepodobnosť vypočítame tak, že zistíme počet priaznivých a počet všetkých možností. Počet priaznivých možností je 3 (hod 2, 4, 6) a počet všetkých možností je 6. Preto pravdepodobnosť hodu párneho čísla je $\frac{3}{6} = \frac{1}{2} = 0{,}5$

[23] Zápis pravdepodobnosti náhodných javov môžeme vyjadriť aj slovne alebo znakom, z ktorého je zrejmý kontext. Napríklad pravdepodobnosť, že pri hode kockou padne 1, môžeme zapísať takto P (pri hode kockou padne 1) alebo skrátene P (1).

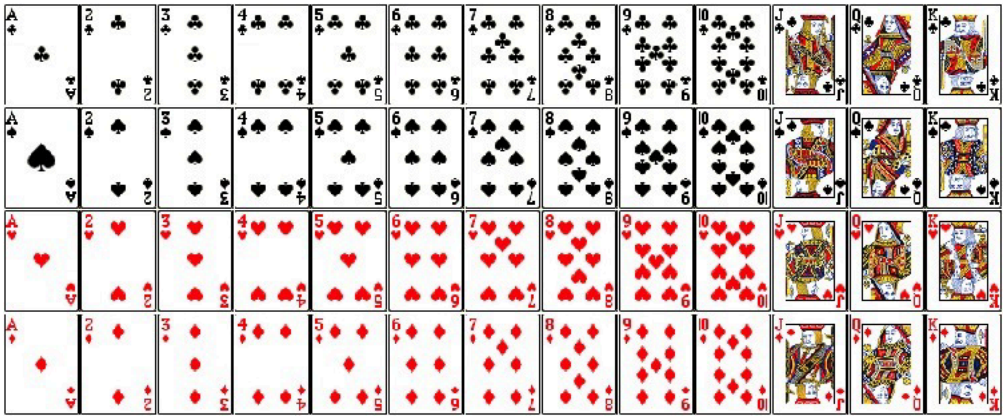
$$P(A_1\ alebo\ A_2) = P(A_1 \cup A_2) = P(A_1) + P(A_2)$$[24]

Analogicky nech A_1, A_2, ..., A_k je *k* disjunktných náhodných javov, potom pravdepodobnosť, že nastane aspoň jeden z nich sa rovná súčtu ich pravdepodobností:

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

3.4. Pravdepodobnosť javov, ktoré nie sú disjunkné

Zamyslime sa nad tým, ako by sme vypočítali pravdepodobnosť dvoch javov, ktoré nie sú disjunktné v kontexte balíčka 52 pokrových kariet. Pre tých, ktorí poker nehrávajú, je takýto balíček kariet znázornený na Obrázok 14.



Obrázok 14 Balíček 52 pokrových kariet

Príklad 3.4

Aká je pravdepodobnosť, že z balíčka dobre premiešaných 52 pokrových kariet vytiahneme srdcové eso?[25]

Príklad 3.5

[24] Formálny zápis pre logický súčet slovne vyjadrený „alebo“ je „U“.

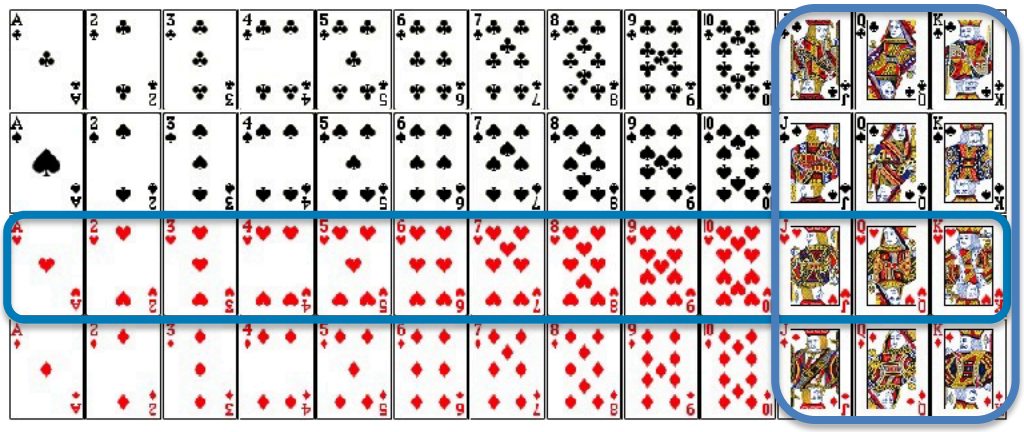
[25] Pri výpočte zistíme počet priaznivých a počet všetkých možností: počet priaznivých možností je 1 (v balíčku sa nachádza len jedno srdcové eso) a počet všetkých možností je 52, preto pravdepodobnosť, že vytiahnem srdcové eso sa bude rovnať: $P(srdcové\ eso) = \frac{1}{52} \cong 0{,}02$.

Aká je pravdepodobnosť, že náhodne vybraná karta (z balíčka 52 dobre premiešaných pokrových kariet) bude srdcová karta?[26]

Príklad 3.6

Aká je pravdepodobnosť, že náhodne vybraná karta (z balíčka 52 dobre premiešaných pokrových kariet) bude tzv. „face“ karta, teda karta označená symbolom J, Q, K?[27]

Aká je pravdepodobnosť, že náhodne vytiahnutá karta bude srdcová karta alebo „face“ karta? Hľadáme pravdepodobnosť dvoch javov: vytiahnutie srdcovej karty a vytiahnutie „face“ karty. Označme vytiahnutie srdcovej karty písmenom S a vytiahnutie „face“ karty písmenom F. Zaujíma nás teda: $P(S\ alebo\ F)$. Je dôležité si uvedomiť, či ide o disjunktné javy. Aj z Obrázok 15 je zrejmé, že nejde o disjunktné javy, pretože existujú karty, ktoré sú súčasne srdcové a súčasne „face“ karty.



Obrázok 15 Ukážka javov, ktoré nie sú disjunktné

Ak by sme pri výpočte pravdepodobnosti $P(S\ alebo\ F)$ postupovali ako pri disjunktných javoch a sčítali obidve pravdepodobnosti, tak by sme karty, ktoré sú srdcové a súčasne „face“ karty

[26] Pri výpočte zistíme počet priaznivých a počet všetkých možností: počet priaznivých možností je 13 (v balíčku sa nachádza 13 srdcových kariet) a počet všetkých možností je 52, preto pravdepodobnosť, že vytiahnem srdcovú kartu sa bude rovnať: $P(srdcová\ karta) = \frac{13}{52} = 0{,}25$. Alebo inak povedané, srdcové karty tvoria štvrtinu všetkých kariet.

[27] Pri výpočte zistíme počet priaznivých a počet všetkých možností: počet priaznivých možností je 12 (v balíčku sa nachádza 12 „face“ kariet) a počet všetkých možností je 52, preto pravdepodobnosť, že vytiahnem „face“ kartu sa bude rovnať: $P(\text{"face"}\ karta) = \frac{12}{52} \cong 0{,}23$.

(také sú 3) započítali dvakrát. Preto je potrebné, aby sme tomu zamedzili a z celkového súčtu pravdepodobností ich odpočítali:

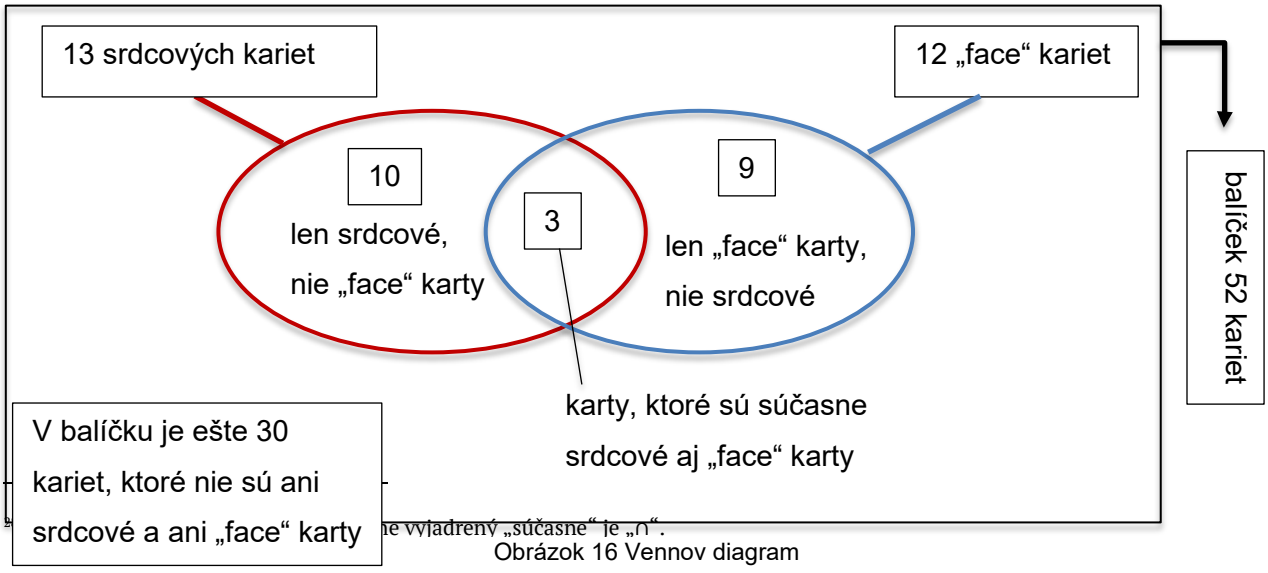
$$P(S\ alebo\ F) = P(S) + P(F) - P(S\ súčasne\ F) = \frac{13}{52} + \frac{12}{52} - \frac{3}{52} = \frac{22}{52} \cong 0{,}43$$

Všeobecné pravidlo pre sčítanie pravdepodobnosti dvoch javov, či už ide o disjunktné alebo nie, znie:

> Nech A a B sú dva náhodné javy, potom pravdepodobnosť, že nastane aspoň jeden z nich vypočítame:
>
> $$P(A\ alebo\ B) = P(A) + P(B) - P(A\ súčasne\ B)$$
>
> $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$[28]

Všimnime si, že v predchádzajúcom pravidle nie je zmienka o tom, či sú náhodné javy A a B disjunktné alebo nie. Môžeme si pravidlo formulovať takýmto spôsobom preto, lebo v prípade, že javy nie sú disjunktné, tak odpočítame ich pravdepodobnosť prienik, ktorá bude nenulová. V prípade, že javy sú disjunktné, znamená to, že ich prienik je nulový a jeho pravdepodobnosť sa tiež rovná 0 a pravidlo bude vyzerať rovnako ako pre výpočet pravdepodobnosti disjunktných javov.

Pri výpočte pravdepodobnosti dvoch javov si môžeme pomôcť Vennovým diagramom, ktorý schematicky predstavuje náhodné javy a vzťahy medzi nimi. Vennov diagram pre výpočet pravdepodobnosti, že náhodne vytiahnutá karta z balíčku 52 pokrových kariet je srdcová alebo „face“ karta bude takýto:



[28] Namiesto slovne vyjadrený „súčasne“ je „∩“.

Obrázok 16 Vennov diagram

3.5. Rozdelenie pravdepodobnosti

Rozdelenie pravdepodobnosti je tabuľka všetkých náhodných javov z priestoru náhodných javov, ktoré sú disjunktné a k nim prislúchajúcich pravdepodobností.
Vezmime náhodný pokus hádzanie dvoma kockami a bude nás zaujímať, aký súčet padne na kockách. Disjunktné náhodné javy, ktoré môžu pri tomto náhodnom pokuse nastať tvoria priestor náhodných javov a sú také, pri ktorých padne súčet 2, 3, 4, ..., 12.

Príklad 3.7
Aká je pravdepodobnosť, že pri hode dvoma kockami padne súčet 2?[29]

Príklad 3.8
Aká je pravdepodobnosť, že pri hode dvoma kockami padne súčet 3?[30]

Príklad 3.9
Aká je pravdepodobnosť, že pri hode dvoma kockami padne súčet 10?[31]

[29] Pravdepodobnosť vypočítame tak, že zistíme počet priaznivých a počet všetkých možností. Počet priaznivých možností je 1, pretože súčet 2 padne iba v prípade, že na oboch kockách padne 1 a počet všetkých možností je 36. Preto pravdepodobnosť súčtu 2 je $\frac{1}{36}$.

[30] Pravdepodobnosť vypočítame tak, že zistíme počet priaznivých a počet všetkých možností. Počet priaznivých možností je 2, pretože súčet 3 padne vtedy, ak je na jednej kocke 1 a na druhej 2 (1, 2) a naopak (2, 1) a počet všetkých možností je 36. Preto pravdepodobnosť súčtu 3 je $\frac{2}{36}$.

[31] Pravdepodobnosť vypočítame tak, že zistíme počet priaznivých a počet všetkých možností. Počet priaznivých možností je 4, pretože súčet 10 padne vtedy, ak je na jednej kocke 4 a na druhej 6 (4, 6) a naopak (6, 4) alebo na oboch kockách 5 (5, 5) a počet všetkých možností je 36. Preto pravdepodobnosť súčtu 10 je $\frac{4}{36}$.

Niektoré z pravdepodobností náhodných javov sme už vypočítali, zvyšné čitateľ určite zvládne vypočítať sám a doplniť tak tabuľku rozdelenia pravdepodobnosti pre náhodný pokus, v ktorom sledujeme súčet, ktorý padne pri hode dvomi kockami:

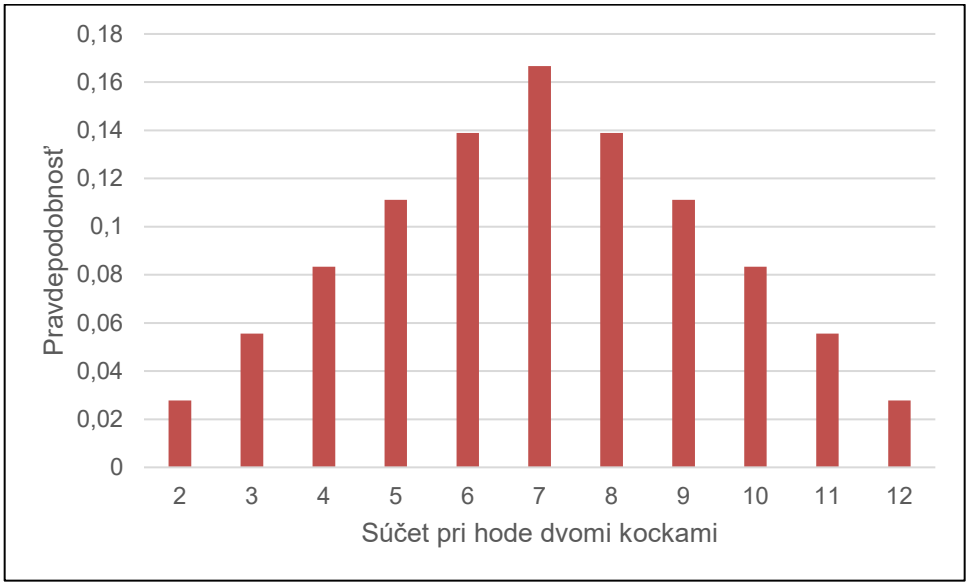
Tabuľka 20 Tabuľka rozdelenia pravdepodobnosti

Súčet	2	3	4	5	6	7	8	9	10	11	12
Pravdepodobnosť	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Grafické znázornenie rozdelenia pravdepodobnosti je na Obrázok 17.

Pre rozdelenie pravdepodobnosti platia nasledovné pravidlá:

1. náhodné javy musia byť disjunktné
2. pravdepodobnosť každého javu je hodnota z intervalu $\langle \mathbf{0}, \mathbf{1} \rangle$
3. súčet pravdepodobností je rovný 1.



Obrázok 17 Graf rozdelenia pravdepodobnosti pri hode dvomi kockami

3.6. Doplnok náhodného javu

Pri hode klasickou kockou priestor náhodných javov obsahuje 6 disjunktných náhodných javov: hodenie {1, 2, 3, 4, 5, 6}. Tento priestor označujeme S a využívame ho často vtedy, keď chceme vypočítať pravdepodobnosť toho, že sa nejaký náhodný jav nestal.
Nech D je náhodný jav, že pri hode kockou padne 2 alebo 3 a označme ho D = {2, 3}. Potom doplnok náhodného javu D reprezentuje tie náhodné javy v priestore S, ktoré nepatria do D. Označujeme ho D^C (horný index C znamená complement) a pre našu situáciu D^C = {1, 4, 5, 6}.

Príklad 3.10

Aká je pravdepodobnosť náhodného javu D a D^C? Aký je súčet týchto pravdepodobností?[32]

Doplnok náhodného javu A označujeme A^C a predstavuje všetky náhodné javy, ktoré nepatria do A. Pre pravdepodobnosti A a A^C platí:

$$P(A) + P(A^C) = 1$$
$$P(A) = 1 - P(A^C)$$

Príklad 3.11

Spomeňme si na hru Osadníci z Catanu a špeciálnu situáciu, ak pri hode dvomi kockami padne súčet 7 a do hry zasiahne zlodej. Aká je pravdepodobnosť, že zlodej nezasiahne? Resp. aká je pravdepodobnosť, že pri hode dvomi kockami nepadne súčet 7?[33]

3.7. Nezávislé javy

---

[32] D je náhodný jav, že padne 2 alebo 3 a jeho pravdepodobnosť je P(D) = $\frac{2}{6}$. D^C je náhodný jav, že padne 1 alebo 4 alebo 5 alebo 6 a jeho pravdepodobnosť je P (D^C)= $\frac{4}{6}$. Súčet pravdepodobností je:

P(D) + P (D^C) = $\frac{2}{6} + \frac{4}{6} = 1$.

[33] Pravdepodobnosť by sme mohli vypočítať tak, že by sme sčítali pravdepodobnosti, že padne iný súčet ako 7, teda pravdepodobnosť, že padne súčet 2, 3, 4, 5, 6, 8, 9, 10, 11 alebo 12, čo by bolo značne zdĺhavé. Preto využijeme vlastnosť doplnku náhodného javu, že padne 7. P(padne súčet 7) = $\frac{1}{6}$ a P(nepadne súčet 7) = 1 − P(padne súčet 7) = 1 − $\frac{1}{6} = \frac{5}{6}$.

Dva náhodné javy sú **nezávislé**, ak poznanie o výsledku jedného z nich neposkytuje žiadnu užitočnú informáciu o výsledku druhého. Napríklad hádzanie mincou a hod kockou sú dva nezávislé náhodné javy a znalosť toho, že na minci padla hlava, nepomáha určiť výsledok hodu kockou. Podobne je to aj pri hádzaní dvomi mincami alebo dvoma kockami, vždy ide o nezávislé javy. Naproti tomu vybratie dvoch kariet z balíčka 52 pokrových kariet za predpokladu, že prvú vytiahnutú kartu do balíčka nevrátime už nie sú nezávislé javy, pretože ak z balíčka vytiahneme prvú kartu, tak tým ovplyvníme pravdepodobnosť druhej karty.

Predstavme si, že hádžeme dvomi klasickými kockami, napríklad prvá z nich je biela a druhá modrá a zaujíma nás pravdepodobnosť, že na oboch súčasne padne 1. Pravdepodobnosť, že na bielej padne 1 je $\frac{1}{6}$, rovnako ako pravdepodobnosť, že na modrej padne 1 je $\frac{1}{6}$ a javy sú nezávislé (informácia o tom, ako dopadol hod na bielej kocke nám nedáva žiadnu informáciu o tom, ako dopadol hod na modrej kocke). Pravdepodobnosť, že na oboch kockách súčasne padla 1 vypočítame tak, že obe pravdepodobnosti vynásobíme.

$$\text{P(na bielej padne 1 } súčasne \text{ na modrej padne 1)} = \left(\frac{1}{6}\right).\left(\frac{1}{6}\right) = \frac{1}{36}$$

Pri **výpočte pravdepodobnosti nezávislých** javov sa riadime **pravidlom násobenia:**

> Nech A_1 a A_2 sú nezávislé javy, potom pravdepodobnosť, že nastane A_1 a súčasne A_2 vypočítame ako súčin jednotlivých pravdepodobností javov:
>
> $$P(A_1 \text{ súč}asne\ A_2) = P(A_1 \cap A_2) = P(A_1).P(A_2)$$
>
> Analogicky nech A_1, A_2, …, A_k je k nezávislých náhodných javov, potom pravdepodobnosť, že nastanú všetky súčasne sa rovná súčinu ich pravdepodobností:
>
> $$P(A_1 \cap A_2 \cap \ldots \cap A_k) = P(A_1).P(A_2).\ldots.P(A_k)$$

Príklad 3.12

Predpokladajme, že premenné pohlavie a to, či je človek pravák alebo ľavák sú nezávislé (teda vedomosť o tom, či je človek muž alebo žena nám nedá žiadnu užitočnú informáciu o tom, či je pravák alebo ľavák) a že v populácii je 9% ľavákov a 50% mužov. Aká je pravdepodobnosť, že náhodne vybraná osoba je žena praváčka?[34]

[34] Pravdepodobnosť, že náhodne vybraná osoba je muž je 50% = 0,5, rovnako ako pravdepodobnosť, že náhodne vybraná osoba je žena. Pravdepodobnosť, že vybraná osoba je ľavák je 9% = 0,09, z čoho vyplýva, že pravdepodobnosť, že vybraná osoba je pravák

Cvičenia:

33. Je pravdivé nasledujúce tvrdenie?: “Ak hádžeme mincou mnohokrát za sebou a v posledných ôsmich hodoch padol znak, potom šanca, že v nasledujúcom hode padne znak je menej ako 50%.”

34. Ak hádžeme kockou, aká je pravdepodobnosť, že padne:

a) štvorka;

b) jednotka alebo šestka

c) párne číslo alebo 6

d) nepárne číslo alebo číslo väčšie ako 3

35. Nech jav A predstavuje vytiahnutie "Face" karty (t.j. karty s označením J, Q, K) a jav B predstavuje vytiahnutie červenej karty z balíčku 52 kariet. Javy A a B sú vzájomne sa vylučujúce (disjunktné) javy. Ide o pravdivé alebo nepravdivé tvrdenie?

36. Nech jav A predstavuje vytiahnutie "Face" karty (t.j. karty s označením J, Q, K) a jav B predstavuje vytiahnutie esa z balíčku 52 kariet. Javy A a B sú disjunktné javy. Ide o pravdivé alebo nepravdivé tvrdenie?

37. V roku 2012 sa uskutočnil prieskum na vzorke 2 373 náhodne vybraných respondentov, ktorí sa mali vyjadriť akú čokoládu majú radi: horkú, mliečnu alebo oba druhy čokolád. 45% respondentov označili, že majú radi mliečnu čokoládu, 33% respondentov horkú čokoládu a 11% majú radi oba druhy čokolád. Určte:

a) či to, akú čokoládu má respondent rád, sú disjunktné javy?

b) znázornite vennov diagram aj s príslušnými relatívnymi početnosťami;

c) koľko percent respondentov majú radi horkú čokoládu, ale určite nie mliečnu?

d) koľko percent respondentov majú radi horkú alebo mliečnu čokoládu?

e) koľko percent respondentov nemajú radi ani horkú, ani mliečnu čokoládu?

38. Z dát zozbieraných zo základných škôl sa zistilo, že 25% všetkých žiakov vymeškalo v škole iba jeden deň, 15% vymeškalo 2 dni a 18% vymeškalo 3 alebo viac dní.

---

bude 1 – 0,09 = 0,91. Pravdepodobnosť, že vybraná osoba je žena praváčka vypočítame ako súčin pravdepodobností, že vybraná osoba je žena a súčasne praváčka: P(žena *súčasne* praváčka) = P (žena) . P (praváčka) = 0,5 . 0,91 = 0,455, pretože náhodné javy sú nezávislé.

a) Aká je pravdepodobnosť, že náhodne vybraný žiak nevymeškal ani jeden deň?
b) Aká je pravdepodobnosť, že náhodne vybraný žiak nevymeškal viac ako jeden deň?
c) Aká je pravdepodobnosť, že náhodne vybraný žiak vymeškal aspoň jeden deň?

39. Americká spoločnosť pre prácu s komunitami každoročne vydáva správu s aktuálnymi informáciami o živote komunít. V roku 2010 táto spoločnosť zverejnila informáciu, že 14,6% Američanov, žije pod hranicou chudoby, 20,7% Američanov rozpráva doma iným ako anglickým jazykom a 4,2% patrí do oboch kategórií, teda medzi Američanmi je 4,2% takých, ktorí žijú pod hranicou chudoby a zároveň sa doma nerozprávajú po anglicky. Na základe vennovho diagramu určte:
a) koľko percent Američanov žije pod hranicou chudoby a doma sa rozprávajú po anglicky?
b) koľko percent Američanov žije pod hranicou chudoby alebo sa doma rozprávajú inak ako po anglicky?
c) koľko percent Američanov žije nad hranicou chudoby a doma rozprávajú po anglicky?

40. Ak hádžeme dvoma kockami (jednou bielou, druhou modrou), aká je pravdepodobnosť, že padne:
a) na bielej jednotka a súčasne na modrej šestka
b) súčet 5
c) súčet aspoň 10
d) na bielej padne 2 a súčasne na modrej číslo väčšie ako 3
e) padne súčet 6 alebo na oboch kockách rovnaké číslo (napr. na bielej 1 a na modrej tiež 1)
f) padne súčet 5 alebo 7

41. Budeme hádzať 4 rôznofarebnými kockami (biela, modrá, červená, zelená). Aká je pravdepodobnosť, že:
a) padne práve jedna šestka
b) padnú práve tri šestky

42. Predstavme si, že hádžeme mincami. Aká je pravdepodobnosť, že:
a) pri hode dvoma mincami na oboch padne znak;
b) pri hode troma mincami na všetkých troch padne znak;
c) pri hode desiatimi mincami vždy padne hlava;
d) pri hode desiatimi mincami padne aspoň jeden znak?

43. Vieme, že v populácii sa nachádza 9% ľavákov Vypočítajte:
a) ak náhodne vyberieme dvoch jedincov, aká je pravdepodobnosť, že budú obaja ľaváci;

b) ak náhodne vyberieme dvoch jedincov, aká je pravdepodobnosť, že budú obaja praváci;
c) ak náhodne vyberieme piatich jedincov, aká je pravdepodobnosť, že budú všetci praváci
d) ak náhodne vyberieme piatich jedincov, aká je pravdepodobnosť, že budú všetci ľaváci
e) ak náhodne vyberieme piatich jedincov, aká je pravdepodobnosť, že nie všetci budú praváci?

44. Predpokladajme, že premenné pohlavie a to, či je človek pravák alebo ľavák sú nezávislé (teda vedomosť o tom, či je človek muž alebo žena nám nedá žiadnu užitočnú informáciu o tom, či je pravák alebo ľavák)a že v populácii je 9% ľavákov a 50% mužov. Vypočítajte:
a) aká je pravdepodobnosť, že jeden vybraný jedinec bude muž pravák;
b) aká je pravdepodobnosť, že dvaja vybraní jedinci budú muži praváci;
c) aká je pravdepodobnosť, jeden vybraný jedinec bude žena ľaváčka;
d) aká je pravdepodobnosť, že z troch jedincov budú dvaja muži praváci a jedna žena ľaváčka?

45. Predstavte si, že máte vrecko, v ktorom je 5 červených, 3 modré a 2 oranžové kocky.
a) ako prvú kocku ste z vrecúška vytiahli modrú kocku a nechali ste si ju (nevrátili ste ju naspäť do vrecúška). Aká je pravdepodobnosť, že druhá vytiahnutá kocka bude opäť modrá?
b) tentokrát ste ako prvú kocku vytiahli oranžovú kocku a nechali ste si ju (nevrátili ste ju naspäť do vrecúška). Aká je pravdepodobnosť, že druhá vytiahnutá kocka bude modrá?
c) Aká je pravdepodobnosť, že prvá aj druhá vytiahnutá kocka budú modré (prvú kocku po vytiahnutí nevrátite naspäť do vrecúška)?

46. Vo vašej zásuvke na ponožky sa nachádzajú 4 modré, 5 sivých a 3 čierne ponožky. Ráno, ešte poslepiačky, si zo zásuvky náhodne vyberiete 2 ponožky. Zistite, aká je pravdepodobnosť, že:
a) ste si vybrali 2 modré
b) ste si nevybrali ani jednu sivú.
c) ste si vybrali aspoň jednu čiernu.
d) ste si vybrali zelenú ponožku.
e) ste si vybrali pár (teda dve ponožky rovnakej farby).

47. Predstavte si triedu, v ktorej je 24 študentov a študentiek. 7 z nich má oblečené džínsy, 4 majú oblečené šortky, 8 má sukňu a zvyšok legíny. (pozn. žiadny študent ani študentka nemá na seba oblečené súčasne dva druhy oblečenia). Náhodne vyberieme 3 jedincov. Aká je pravdepodobnosť, že v trojici vybraných jedincov budú mať dvaja oblečené džínsy a jeden legíny?

Zoznam použitej literatúry

DIEZ, David M, BARR, Christopher D, ÇETINKAYA-RUNDELMine. *OpenIntro statistics*. 4. vyd. [s.l.]: Openintro, Inc, 2019. ISBN 9781943450077.

MATIAS, J. Nathan et al. The Upworthy Research Archive, a time series of 32,487 experiments in U.S. media. In: *Scientific Data* [online]. 2021, roč. 8, č. 1 [cit. 26.11.2025]. DOI: https://doi.org/10.1038/s41597-021-00934-7

SUN, Yuzheng. How Meta made me a big-time A/B testing advocate. In: *Statsig.com* [online] [cit. 26.11.2025]. Dostupné na internete: https://www.statsig.com/blog/meta-a-b-testing

VERMEER, Lukas. The role of experimentation at Booking.com. In: *https://partner.booking.com* [online] [cit. 25.11.2025]. Dostupné na internete: https://partner.booking.com/en-us/click-magazine/industry-perspectives/role-experimentation-bookingcom

VIDOVÁ, Veronika. Vďaka experimentom meníme stránku Profesia.sk k lepšiemu. In: *https://www.profesia.sk/* [online] [cit. 25.11.2025]. Dostupné na internete: https://firma.profesia.sk/aktualita/vdaka-experimentom-menime-stranku-profesia-sk-k-lepsiemu/

Eva Capková
Analýza dát pre mediamatikov
časť I.

Vydala Univerzita Komenského v Bratislave, 2025

Technická redakcia, návrh obálky: Juraj Grečnár

Rozsah 74 strán, 3 AH, 1. vydanie
vyšlo ako elektronická publikácia

ISBN 978-80-223-6235-1 (online)