získal také titul RNDr. Titul CSc. a docent získal v oboru Statistika na Vysoké škole ekonomické v Praze. Na VŠE vyučuje předměty věnované teorii pravděpodobnosti a matematické statistice na všech stupních studia. Na Fakultě informatiky a statistiky VŠE působil ve vedoucích pozicích, např. jako vedoucí vědeckých pracovišť, vedoucí katedry a také jako její děkan. Je autorem či spoluautorem několika knih a více než stovky odborných článků. Je členem řady tuzemských i zahraničních statistických společností.

Doc. Ing. Iva Pecáková, CSc. absolvovala obor Statistika na Vysoké škole ekonomické v Praze. Od roku 1980 zde působí jako pedagožka; získala titul kandidátky věd a v roce 2005 se ve stejném oboru habilitovala. Kromě vedení nejrůznějších statistických kurzů na všech stupních studia na VŠE v Praze učila i na dalších školách či pracovištích, dlouhodoběji např. na Fakultě sociálních věd Univerzity Karlovy. Ve svojí pedagogické, vědecko-výzkumné a publikační činnosti se zaměřuje na jednorozměrnou i vícerozměrnou analýzu průřezových dat, především analýzu datových souborů z terénních průzkumů, v nichž převládají kategoriální proměnné.

Prof. Ing. Hana Řezanková, CSc. je absolventkou Vysoké školy ekonomické v Praze, kde v roce 1988 získala též titul kandidáta věd v oboru Aplikace matematiky v ekonomii. Zde také od roku 1990 působí na katedře statistiky a pravděpodobnosti V roce 1994 se habilitovala v oboru Statistika a v roce 2008 byla ve stejném oboru jmenována profesorkou. Je garantkou doktorského studia oboru Statistika a od roku 2017 i předsedkyní oborové rady pro doktorské studium v tomto oboru. Prof. Řezanková je dlouholetou členkou výboru České statistické společnosti, v letech 2013-2017 zastávala funkci předsedkyně. V pedagogické a vědecko-výzkumné činnosti se zaměřuje zejména na analýzu kategoriálních dat a na shlukovou analýzu.

Ing. Karel Hełman, Ph.D. vystudoval statistiku na Vysoké škole ekonomické v Praze, kde pedagogicky působí na katedře statistiky a pravděpodobnosti od roku 2005. Titul Ph.D. získal v roce 2012. V pedagogické i vědeckovýzkumné činnosti se zabývá regresní analýzou a analýzou časových řad. Zajímá se o aspekty popisného přístupu ke statistickému modelování, má blízko i ke zkoumání poznatků o environmentální a klimatologické problematice. V letech 2005-2014 působil v Českém hydrometeorologickém ústavu, kde byl v řešitelském týmu mezinárodního projektu zaměřeného na zkoumání městského tepelného ostrova.

Ing. Pavel Zimmermann, Ph.D. absolvoval obor Statistické a pojistné inženýrství na Vysoké škole ekonomické v Praze. Doktorské studium absolvoval tamtéž a stal se pedagogem katedry statistiky a pravděpodobnosti VŠE. Odborně se zaměřuje na aplikace statistiky v oblasti pojistné matematiky, zejména v neživotním pojištění. Odpovídá za rozvoj pojistně matematických předmětů katedry. Pracoval jako risk manažer a vývojář rizikových modelů v nadnárodních společnostech. Jako držitel osvědčení o způsobilosti vykonávat aktuárskou činnost působí rovněž jako externí konzultant v několika pojišťovnách a poradenských společnostech.

1 POPISNÁ STATISTIKA

Každá disciplína má své základní stavební kameny. Vytváří pojmy, vtiskává jim pevný obsah, propojuje je a přisuzuje jim určité vlastnosti. Na těchto kamenech vyrůstá celá složitá stavba vědecké disciplíny, v našem případě statistiky. Ta je orientována na práci s daty. Sbírá je, zpracovává, využívá nejrůznější analytické postupy a metody a snaží se na základě toho připravovat půdu pro kvalifikovaná rozhodnutí.

Základní pojmy je potřeba si osvojit. Patří k nim takové kategorie jako hromadný jev, statistická proměnná (neboli znak), statistický soubor, statistická jednotka. Údaje někde vznikají a obvykle se opakují, takže mluvíme o hromadném jevu. Údaje se dále zpracovávají, zpřehledňují a koncentrují, někdy do tabulek, jindy do grafické podoby (často se hovoří o vizualizaci dat). To vše pomáhá k rychlé orientaci uživatelů statistických dat, ale na druhé straně je třeba si uvědomit, že jde pouze o prvotní popis, o první přiblížení se podstatě problému. I to však může být velmi důležité. Proto tato úvodní kapitola je neodmyslitelným začátkem statistické práce.

Vypovídací schopnost tabulek a grafů je ale omezená a jejich použitelnost pro hlubší analýzy malá. Proto statistika buduje složitější aparát, který údaje, jež jsme primárně zachytili v tabulkách a grafech, přibližuje v charakteristikách, které dovolují více zobecňovat a které se snaží pod povrchem hledat nové informace a souvislosti. Cesta k tomu vede přes ucelený systém těchto statistických charakteristik, které měří úroveň hodnot ve zkoumaném statistickém souboru (jaká je např. výše průměrné mzdy zaměstnance či jaká je úroveň mzdy, které nejvýše dosahuje právě polovina zaměstnanců v Česku), dále měnlivost hodnot v souboru (zda jsou mzdy našich zaměstnanců diferencovanější než mzdy zaměstnanců třeba na Slovensku) a další důležité vlastnosti, které poznáte při čtení této kapitoly.

1.1 Význam a pojetí moderní statistiky

Poznání stále pronikavěji zasahuje do všech stránek našeho života. Neexistuje oblast lidské činnosti, jež by nebyla touto skutečností zasažena a ovlivněna. Množství znalostí, objem dat a rozsah informací o dění kolem nás se zvyšují značnou rychlostí. Orientace v oborech lidské činnosti a v jejich výsledcích je stále náročnější. Vzrůstá i cena informací, protože ty se staly zbožím stejně běžným jako třeba automobil. Řada informací mívá charakter údajů o hromadných jevech. Jejich zkoumání a vyhodnocování se stalo předmětnou náplní praktické i teoretické statistiky.

Počátky statistiky nacházíme už ve starověkých říších. Tehdy se jednalo o různé soupisy obyvatelstva, nejčastěji pro daňové a vojenské účely. Se skutečnými statistickými analýzami událostí, které tehdejší svět charakterizovaly, se setkáváme až v 17. století a spojujeme je zejména se jmény Johna Graunta (1620–1674) a Williama Pettyho (1623–1687). Slovo statistika jako pojmenování vědního oboru vzniklo v polovině 18. století (Gottfried Achenwall, 1719–1772). Další vývoj statistiky se již opíral

o základy teorie pravděpodobnosti, položené Blaisem Pascalem, Jacobem Bernoullim, Karlem F. Gaussem a mnoha dalšími učenci.

Od počátku 20. století se statistika rychle rozvíjela již jako samostatná vědní disciplína. K její další akceleraci od 70. let minulého století nepochybně přispěl také rychlý rozvoj výpočetní techniky a vznik statistického výpočetního prostředí, bez kterého je dnes moderní statistická analýza nemyslitelná.

Bez nadsázky lze tvrdit, že v současnosti již neexistuje vědní obor, který by nepracoval s hromadnými daty a nevyužíval k jejich vyhodnocování statistické metody. Mezi obory, které zcela běžně a přirozeně aplikují tyto metody, patří medicína, fyzika, biologie a další přírodní i technické disciplíny. Významné místo zaujímá statistika i ve sféře analýzy sociálněekonomických jevů. Její uplatnění v hospodářské oblasti je dnes dokonce tak rozsáhlé, že se setkáváme s řadou speciálních disciplín a partií statistiky, které se postupem doby vytvořily právě na základě potřeb ekonomické teorie i praxe.

Slovu statistika bývá dáván nejrůznější význam. Jednou jsou tak nazývány vyplněné statistické výkazy či dotazníky, nejrůznější číselné údaje uveřejněné ve sdělovacích prostředcích, jindy organizace určitého statistického zjišťování nebo zpracování jeho výsledků, často také nejrůznější poznatky o metodách, které statistika používá. Slovo statistika lze tedy chápat nejméně ve trojím pojetí. Jednak jako číselné údaje o hromadných jevech, dále jako praktickou činnost, spočívající ve sběru, zpracování a vyhodnocování údajů, a konečně jako teoretickou disciplínu, jež se zabývá metodami, sloužícími k odhalování zákonitostí při působení podstatných, relativně stálých činitelů na hromadné jevy, tj. jevy, vyskytující se v masovém měřítku u velkého počtu jedinců (prvků).

Ve všech pojetích statistiky se tedy setkáváme s **hromadnými jevy**, tj. s takovými skutečnostmi, jež se vyskytují mnohokrát a mohou se znovu a znovu opakovat. Číselné údaje o jednotlivostech, popř. o několika málo skutečnostech, nejsou předmětem zájmu, a tedy ani statistikou. K získání určitých poznatků a vyslovení závěrů o zkoumaných jevech tedy nestačí jednotlivá pozorování, nutná jsou pozorování hromadná.

Význam statistiky v současném světě – a v ekonomice zejména – je bezesporu značný. Správný a cílevědomý chod ekonomiky v zájmu maximalizace její efektivnosti je nemyslitelný bez kvalitní informační soustavy. Mimořádně významná role přitom přísluší statistice, která poskytuje soustavu číselných informací o národním hospodářství jako celku i o jeho subsystémech. Úspěšná realizace změn v ekonomice je nemyslitelná bez kvalitní statistiky a bez schopnosti ekonomů nemajících specializované statistické vzdělání rozumět statistice alespoň na nezbytné úrovni. Stejně tak je statistika důležitou podporou v procesu tvorby manažerských rozhodování, analýz trhu, v řízení jakosti či v přejímce, v pojistné a finanční matematice apod.

V minulosti se statistika někdy ztotožňovala s pouhým zjišťováním, evidencí, sumarizací a publikací údajů. V současné době se na každém kroku ukazuje, že bez statistiky není možné provádět praktickou hospodářskou politiku a přijímat důležitá rozhodnutí ani na úrovni národního hospodářství, ani na úrovni firmy. Obsah statistiky,

periodicita i formy prezentace údajů by měly mnohem více než doposud respektovat potřeby kvalifikovaných rozhodovacích procesů, statistické údaje by se měly stát východiskem rozhodování na všech stupních národního hospodářství v různých horizontech – od rozhodování operativního až k rozhodování s dlouhodobými důsledky. Statistiku tedy nelze ztotožňovat, jak se to někdy děje, s pouhým elementárním zpracováním údajů. Moderní statistika je nejen praktická činnost, ale i náročná metodologická disciplína.

1.2 Základní statistické pojmy

Ve statistickém zkoumání nás zajímají, jak jsme uvedli, hromadné jevy a procesy, tj. jevy a procesy, vyskytující se u velkého množství prvků. Tyto prvky nazýváme **statistickými jednotkami** a jsou elementárními jednotkami statistického zjišťování. Mohou to být fyzické osoby (např. zaměstnanci podniku při zkoumání mezd), právnické osoby (např. nefinanční podniky, finanční instituce, provozy, obchodní jednotky atd. při zkoumání hospodářského výsledku), věci (např. účetní doklady), události (dopravní nehody) apod.

Vlastnosti statistických jednotek vyjadřují **statistické znaky**. Jestliže např. zvolíme za statistickou jednotku pracovníka určité firmy, můžeme tuto jednotku charakterizovat znaky, jako jsou např. mzda, věk, pohlaví, nejvyšší dosažené vzdělání, počet let praxe aj. Statistické znaky obvykle označujeme písmeny z konce abecedy, to znamená x, y, z apod.

Statistické znaky můžeme rozdělit podle několika kritérií, zejména podle toho, jak lze vyjádřit varianty (obměny) hodnot těchto znaků. Např. počet členů domácnosti lze vyjádřit číselnou formou. Naproti tomu typ vlastnictví bytu, ve kterém domácnost bydlí, je prezentován slovní definicí (může tedy jít např. o byt družstevní, v osobním vlastnictví, nájemní atd.). Podobná členění nejsou samoúčelná, protože při používání konkrétních statistických postupů a metod budeme většinou jinak přistupovat k číselným údajům, jinak k údajům verbálním.

Lze-li varianty hodnot znaku vyjádřit číselně, jde o **znaky kvantitativní** (počet členů domácnosti, spotřeba elektřiny, výše nájemného aj.). Můžeme-li je vyjádřit slovně, jde o **znaky kvalitativní** (druh vlastnictví bytu, místo trvalého pobytu, typ vystudované střední školy) – často hovoříme o tomto typu znaku také jako o **znaku kategoriálním** (též někdy nominálním nebo slovním).

Jestliže může statistický znak nabývat pouze dvou variant hodnot, hovoříme o **znaku alternativním** (sledovaný jev nastal, nebo nenastal). Znak, který může mít více než dvě varianty, nazveme **znakem množným** (nejvyšší dosažené vzdělání – základní, střední, vysokoškolské). S tímto členěním se setkáváme obvykle jen u znaků kvalitativních.

V souvislosti s pojmem statistický znak se můžeme rovněž setkat s pojmem **znak měřitelný** (metrický) a **znak pořadový** (ordinální). V případě měřitelného znaku lze

jeho hodnoty porovnávat rozdílem nebo poměrem (porovnání měsíční mzdy dvou pracovníků). Naproti tomu pořadový znak odráží pouze odstupňované pořadí variant znaku, zatímco jejich srovnávání podílem u něj prakticky nelze rozumně interpretovat (např. nejvyšší dosažené vzdělání – stěží prohlásíme, že vysokoškolák je dvakrát více vzdělán než středoškolák, nepochybujeme však o tom, že jeho vzdělání je vzděláním vyšším). Zvláštním případem pořadového znaku je znak, jehož hodnoty jsou tvořeny jednotlivými stupni možných odpovědí podle předem zadané škály. Se škálami se často setkáváme v marketingovém výzkumu nebo při výzkumech veřejného mínění.

Kvantitativní statistické znaky někdy též rozlišujeme na **diskrétní** (někdy říkáme též **nespojité**) a **spojité**. Diskrétní statistické znaky nabývají pouze některých číselných hodnot, nejčastěji přirozených čísel nebo celých nezáporných čísel (počet členů domácnosti, počet chyb na účetním dokladu, počet zmetků v sérii). Spojité statistické znaky mohou naopak nabývat v rámci určitého intervalu libovolných hodnot (spotřeba elektřiny, doba čekání na příchod zákazníka k obslužné jednotce, výše tržeb, kurz koruny k euru).

Na závěr ještě řekněme, že mnohdy se místo pojmu statistický znak používá termín **statistická proměnná**. Oba tyto názvy však můžeme považovat za zcela rovnocenné – v praxi jde spíše o zvyk uživatelů. Pojem proměnná je běžný zejména v anglické terminologii statistického softwaru („variable“).

Množinu všech statistických jednotek, u nichž zkoumáme příslušné statistické znaky, nazýváme **statistickým souborem**. Zjišťujeme-li u každé statistické jednotky jenom jeden statistický znak, nazýváme tento soubor **souborem jednorozměrným**. Zjišťujeme-li u každé statistické jednotky dva nebo více znaků a zkoumáme-li současně jejich vzájemné vztahy, hovoříme o statistických **souborech dvourozměrných**, resp. o **souborech vícerozměrných**.

Statistický soubor všech jednotek, který je vlastním předmětem sledování a o němž chceme provádět závěry, se nazývá **základním souborem** (mnohdy též **cílovou populací**). Jeho rozsah může být konečný i nekonečný, zpravidla je ale značný. Proto se velmi často z úsporných, časových i technických důvodů provádí **šetření výběrové**, kdy se ze základního souboru určitým způsobem vyberou pouze některé jednotky. Tím vznikne **výběrový soubor**. Z něho získané výsledky potom mohou za určitých podmínek sloužit k provádění úsudků o základním souboru.

Na volbě statistických jednotek a vhodném výběru statistických znaků, pomocí nichž chceme sledovat vlastnosti statistického souboru, závisí úspěch a výsledky veškeré další práce. Proto je třeba správnému vymezení statistického souboru a správné volbě statistické jednotky a statistických znaků věnovat náležitou pozornost. Statistická jednotka i zjišťované znaky přitom musejí být přesně vymezeny z hlediska časového, prostorového a věcného. Jinak řečeno, obsah zjišťovaného znaku musí být přesně definován.

1.3 Prvky statistického zkoumání

Práci s daty lze zpravidla rozdělit do několika etap. Jde o etapu pořizování dat (šetření), prvotního zpracování zjištěných údajů a konečně o etapu jejich statistického vyhodnocování (analýzy). I když nejdůležitější je nepochybně fáze třetí – analytická, v níž probíhá analýza statistických údajů pomocí vhodných metod, je nezbytným předpokladem její úspěšnosti, aby byly správně provedeny rovněž etapy předchozí. Sebelepší statistická analýza nemůže poskytnout hodnotné výsledky, jsou-li nekvalitní data, či jsou-li data sice kvalitní, ale pro analýzu jsou chybně připravena. Celý problém chyb v této fázi statistické práce akceleruje ještě i ten fakt, že v současné moderní době často musíme pracovat s obrovskými objemy dat („big data“). Věnujme proto v této části alespoň stručně pozornost statistickému zjišťování a primárnímu zpracování dat.

V mnoha oblastech našeho života i společnosti, ve které žijeme, jsou data získávána přímým **pozorováním** stanovených objektů či jevů, ať už je realizováno smyslovými orgány, nebo za pomoci přístrojů. Nejjednodušším výsledkem pozorování je zjištění, že určitý jev nastal či nenastal (výrobek je vadný či vadný není), případně zjištění četnosti výskytu nějakého jevu (celkový počet vad výrobků v dané výrobní sérii apod.). Při pozorování hmotnosti výrobků či emisí automobilů však naše smysly poskytují jen nepřesné výsledky a provádíme proto měření. Pozorovatelné jevy přitom porovnáváme s obecně přijatou jednotkou, a to obvykle za pomoci nějakého měřicího nástroje či přístroje, který zahrnuje měřicí stupnici. Změření pak charakterizuje zkoumanou veličinu přesněji než vágní kvalitativní údaje, může být porovnáváno a případně opakováno a výsledek lze kvantitativně analyzovat.

V ekonomických disciplínách se mnohé jevy týkají osob nebo jejich skupin. Možnosti pozorování takových jevů jsou často omezené nebo dokonce historicky zcela neopakovatelné. Výsledky zjišťování mohou být nepřesné a použití přístrojů, pokud vůbec připadá v úvahu, nepraktické či neekonomické (například při zjišťování věku osoby). Potom je účelné (pokud je to realizovatelné) pozorování nahradit studiem dokumentů (věk, vzdělání, délka pracovní neschopnosti, výše daní atd.). Skutečnosti nacházející se pouze v lidském vědomí či podvědomí pak nedovolují pozorování vůbec a dají se zjistit pouze **dotazováním**.

Problémem jevů dotazovaných – oproti jevům pozorovatelným – je možnost značného odchýlení zjištěného údaje od skutečnosti, tedy charakteristická je jeho nízká validita (tato okolnost se někdy vyjadřuje termínem „měkká“ data). Vzhledem k charakteru ekonomických disciplín a s ohledem na velký podíl subjektivních postojů v ekonomii je tedy dotazování a studium dokumentů velmi často používanou metodou zjišťování statistických údajů.

V tržní ekonomice hrají významnou roli informace o existujícím trhu, o jeho sociálněekonomickém prostředí, o spotřebiteli, o jeho postojích a spotřebním chování a o jeho změnách. Politický vývoj zase motivuje zájem o veřejné mínění, názory, postoje i chování různých skupin obyvatelstva, jež se týkají vztahů ke společenským

událostem (např. volby), k činnosti různých organizací. Získaná data zde odrážejí hodnocení těch či oněch společenských, hospodářských a kulturních fenoménů. Nemalý počet soukromých firem a agentur, vlastně v podstatě celé ekonomické odvětví, se tak dnes zabývá sběrem těchto dat na základě výběrových šetření.

Výzkumy trhu a veřejného mínění jsou realizovány telefonickým dotazováním, dotazováním prostřednictvím internetu i na sociálních sítích. Avšak mnoho cenných dat o motivacích a chování zákazníků mohou dnes firmy získávat i pouhým „odezíráním" našich individuálních aktivit na internetu, např. toho, jaké stránky a s jakým obsahem prohlížíme, jaký druh zboží a služeb na webových stránkách častěji vyhledáváme, co na webu preferujeme apod.

Nositelem primární informace v takových výzkumech je konkrétní osoba a základní soubor tvoří obvykle značně rozsáhlá skupina osob, tedy populace. Její vyčerpávající prošetření by vyžadovalo nemalé finanční i časové náklady. Již v první polovině minulého století se po rozsáhlé debatě vědci i uživatelé dat shodli na relevantní metodě pro získání vzorku populace, kterou je **pravděpodobnostní** (náhodný) **výběr**. Moderní statistika totiž disponuje postupy, které v takovém případě umožňují provádět pro základní soubor zobecnění skutečností zjištěných pouze v jeho vzorku, tedy v malé části základního souboru. Tedy provádět projekci poznatků získaných z výběrového souboru do souboru základního (někdy říkáme též do cílové populace).

Zobecnění poznatků získaných v rámci kvantitativního průzkumu vzorku na populaci lze realizovat pouze na základě **reprezentativního vzorku**. S málokterým pojmem se však zachází tak vágně, jako s reprezentativností. Používá se ho ve smyslu, že „data jsou v pořádku", že ve výběru nebyly favorizovány žádné jednotky ani jejich skupiny, a to ať už vědomě nebo nevědomě, vzorek je věrnou zmenšeninou populace, má tedy stejné vlastnosti – míněno v tom smyslu, že má ve sledovaných znacích stejné proporce jako populace, tvoří ho pro populaci typické jednotky, ale přitom pokrývá celou populaci apod.

Ideálním východiskem pro realizaci pravděpodobnostního výběrového šetření je existence **opory výběru**, tedy seznamu všech jednotek tvořících cílovou populaci. Pořízení adekvátní a především aktuální opory může být dosti náročné (i s ohledem na platnost zákona na ochranu osobních dat). Zejména informace o obyvatelstvu vzhledem k jeho přirozené reprodukci i k jeho migraci rychle zastarávají, ale rychle se mění také informace o ekonomických subjektech.

Jako opora výběru mohou sloužit například volební seznamy povinně vedené obcemi a městy, registry obyvatel, přidělené IČO, databáze nemovitostí (například České pošty), SIPO (tj. „Soustředěné inkaso plateb obyvatelstva" u České pošty), registry sčítacích obvodů odrážejících územní organizaci platnou v době sčítání lidu, domů a bytů, registr ekonomických subjektů ČSÚ, firemní databanky, telefonní seznamy, evidence návštěvníků, ale také účetní záznamy všech provedených účetních operací pro účely auditu účetní závěrky atd. U internetových populací opora výběru představuje seznam jednotek včetně kontaktu (e-mailové adresy). To si ovšem lze

představit pouze u homogenních populací s vysokým pokrytím (univerzity, státní instituce, velké firmy, významné nadace aj.).

Výše uvedené chápání reprezentativnosti i úskalí spojená s realizací pravděpodobnostních výběrů (různého typu) stojí za vznikem **nepravděpodobnostního výběrového postupu**, který je ovšem ve výzkumné praxi nejrozšířenější a který je znám jako **kvótní výběr**. Shodu vzorku se základním souborem mají zajistit kvótní znaky, což jsou jednoduché, snadno identifikovatelné a z hlediska souboru klíčové vlastnosti (například pohlaví, věk, vzdělání či druh ekonomické aktivity). Podle hodnot kvótních znaků je základní soubor vlastně rozdělen do skupin a dodržení shodné struktury vzorku znamená, že počet vybíraných osob je úměrný velikosti skupiny. Za předpokladu dodržení náhodnosti volby jednotek by se tedy jednalo o složitější typ pravděpodobnostního výběru. O samotné volbě jednotky však rozhodují i nenáhodné a obtížně kontrolovatelné mechanismy (například osoba realizující výběr – tazatel). Dosažení shodné struktury vzhledem k několika kvótním znakům pak nemusí nutně znamenat shodnou strukturu i z hlediska znaků ostatních, včetně zjišťovaných. Podrobněji se k problematice typů výběru vrátíme ještě jednou, a to v části 3.1.

Ve snaze co nejlépe se přiblížit idejím pravděpodobnostního výběru jsou v praxi vytvářeny **panely**. Jde o opakované šetření, které je prováděno na stejném souboru osob v delším časovém horizontu (tento soubor osob se pak nazývá panel a je většinou vytvořen jako reprezentativní vzorek určité populace). Pomiňme zde zcela panely vzniknuvší nepravděpodobnostním samovýběrem dobrovolníků například na základě pozvání k účasti na často navštěvované webové stránce. I když jsou následně zohledňovány demografické či jiné charakteristiky účastníků takového panelu, samotný mechanismus jeho vzniku neumožňuje objektivně definovat populaci, na kterou by z něho bylo usuzováno. Vytvoření, řízení a udržování panelu je nákladné. Je to však možnost, jak se alespoň do určité míry přiblížit požadavkům zobecňování výběrové informace na populaci. V praxi pro vytvoření panelu bývá obvykle použit kvótní výběr, často se skutečně velkým počtem kvótních znaků.

Útvary státní statistické služby (především Český statistický úřad, ale také specializovaná pracoviště jednotlivých ministerstev, České národní banky aj.) shromažďují údaje pro vytváření statistických informací o sociálním, ekonomickém, demografickém a ekologickém vývoji celé České republiky. Získávány jsou prostřednictvím výkazů (on-line, interaktivní či papírový formulář). Poskytnutí dat je pro jednotky zařazené do konkrétního šetření povinné a plyne ze zákona o státní statistické službě. Zjišťování jsou realizována jako zjišťování plošná (tj. v celém základním souboru), mnohem častěji však jako výběrová (tedy jen ve vzorku).

Bez ohledu na postup pořizování je třeba získaná data přehledně zaznamenávat. Jak jsme už zmínili, nejrůznější formuláře používané k tomuto účelu dnes mají většinou elektronickou podobu a také do samotného procesu zjišťování údajů vstupuje stále více výpočetní a komunikační technika. Už jen samotný proces sběru dat může ovlivnit

jejich kvalitu, spolehlivost i nezávislost. Použití moderních technologií ale kvalitní data ještě nezaručuje, mnohem podstatnější je důkladná příprava každého šetření.

V první řadě je nutné co nejpřesněji definovat zkoumaný statistický znak, statistickou jednotku (například domácnost apod.) a stanovit základní statistický soubor (cílovou populaci), například všechny domácnosti v České republice. Pokud nehodláme realizovat vyčerpávající zjišťování, tj. šetření u všech jednotek v základním souboru, je třeba ze základního souboru pořídit vzorek (výběrový soubor). Způsob výběru jednotek i velikost vzorku přitom významným způsobem ovlivňují možnosti a kvalitu úsudků o základním souboru, které následně na základě výběrových dat provádíme.

Informace o statistické jednotce poskytuje **zpravodajská jednotka**; je třeba ji určit. Zpravodajská jednotka nemusí být totožná s jednotkou statistickou. Tak například údaje o odpracovaných či neodpracovaných hodinách zaměstnanců (statistické jednotky) budeme získávat spíše od podniků (zpravodajské jednotky) – to proto, že zjišťování přímým dotazem u jednotlivých pracovníků by bylo značně pracné a nákladné. Údaje o domácnosti (statistická jednotka) bude poskytovat její **přednosta**, který je zpravodajskou jednotkou (většinou jde o osobu v domácnosti, která činí zásadní rozhodnutí o velkých investicích; pokud nelze jednoznačně rozhodnout, která osoba v domácnosti to je, používá se osoba, která do domácnosti přináší nejvíce peněz; setkat se ale můžeme i s jinými definicemi tohoto pojmu) apod.

Z **časového hlediska** mohou být údaje dvojího druhu: buď se zjišťují za časový interval (objem produkce, počet zákazníků), nebo k určitému okamžiku (počet členů domácnosti). Pro zjišťování údajů prvého druhu musí být stanovena **rozhodná doba** (např. objem produkce za říjen 2018, počet zákazníků obsloužených za určený den), pro zjišťování údajů druhého typu potom **rozhodný okamžik** (např. počet pracovníků k 31. 12. 2018).

Výsledkem šetření, ať již použijeme jakoukoliv metodu, je velké množství údajů, které je třeba určitým způsobem zpracovat, utřídit a shrnout, připravit pro statistickou analýzu. **Zpracování statistických dat** by mělo vždy začínat jejich formální a logickou kontrolou a odstraněním případných úmyslných (tedy nějakým způsobem motivovaných) chyb nebo chyb neúmyslných (překlepů, chyb způsobených neznalostí nebo nezkušeností apod.). Je nutné posoudit, zda obdržené údaje mohou být vyjádřením zjišťované skutečnosti, například zda odpovídají logicky možným mezím, v nichž se mohou pohybovat (např. zda údaje, které mají udávat podíl žen ve firmě v procentech, nepřekračují 100 %, zda počet členů domácnosti není vyjádřen příliš velkým číslem či naopak nulou apod.).

Logická kontrola může být svým způsobem náročná na věcnou znalost jevů, jež jsou předmětem zjišťování. Může dále zahrnovat i prověření správnosti provedených výpočtů (například součtů nebo podílů). Mnohé kontrolní úkony jsou součástí počítačových programů na zpracování statistických dat, často si takové algoritmy pořizovatelé dat připravují rovněž sami.

Důležitým krokem ve fázi kontroly statistických dat je dále rozhodnutí o tom, jak budou vyjádřeny (kódovány) chybějící hodnoty. Zejména v případě jejich vysokého výskytu může být touto skutečností významně ovlivněna představa, kterou si na základě dat vytváříme o základním souboru jednotek. Pokud pro chybějící hodnoty používáme určité symbolické číselné vyjádření, je též nutné zajistit, aby tyto hodnoty nebyly zahrnovány do následně realizovaných výpočtů (například aby do výpočtu průměrného počtu pracovníků firmy nebyly zahrnovány chybějící hodnoty, vyjádřené např. symbolickým označením „99999“).

Součástí základního statistického zpracování dat je dále jejich třídění, sestavování tabulek či grafů, výpočet jednoduchých číselných charakteristik (například průměrů). Tyto činnosti jsou víceméně jednoduché, ostatně software pro jejich realizaci je dnes dosažitelný prakticky pro každého (např. MS Excel). Úskalí však často spočívá ve volbě vhodných nástrojů, adekvátních vlastnostem dat a charakteru sledovaného jevu. Ne každá tabulka či graf, ne každá číselná charakteristika je vhodná pro každý datový soubor. V dalším textu se proto problematikou elementárního zpracování statistických údajů budeme podrobněji zabývat.

1.4 Elementární zpracování statistických údajů

Výsledkem statistického šetření je zpravidla velké množství číselných údajů, které jsou nepřehledné, takže ani zkušený pracovník z nich mnoho nevyčte. Aby vynikly charakteristické rysy a zákonitosti analyzovaného souboru a aby se údaje staly přehlednými, musíme je setřídit. **Tříděním** tedy rozumíme rozdělení jednotek souboru do skupin tak, aby co nejlépe vynikly charakteristické vlastnosti zkoumaných jevů. Tříděním dosáhneme kromě uspořádání údajů do přehledné formy také jejich zhuštění. Provádíme-li třídění jenom podle jednoho statistického znaku (jedné vlastnosti), hovoříme o jednostupňovém třídění. Provádíme-li třídění podle více statistických znaků najednou, jedná se o třídění vícestupňové.

1.4.1 Rozdělení četností

Povšimněme si jednostupňového třídění, kdy sledujeme u statistických jednotek pouze jeden kvantitativní statistický znak. V tomto případě třídění provádíme tak, že uspořádáme údaje o sledovaném kvantitativním znaku do rostoucí posloupnosti. Ke každé variantě hodnoty statistického znaku pak přiřadíme počty příslušných statistických jednotek, které nazýváme **četnostmi**. Rozdělení četností se ale používá často i u znaků kategoriálních (např. typ vlastnictví bytu); zde se pochopitelně uspořádání údajů do rostoucí posloupnosti neprovádí a základem je jen prostý výčet vyskytujících se variant znaku.

Pokud vše uspořádáme do tabulky, vznikne **tabulka rozdělení četností**. Podává informaci o četnostech výskytu jednotlivých variant hodnot znaku v souboru. Označíme-li jednotlivé varianty hodnot nespojitého znaku symbolem x_i, $i = 1, 2, ..., k$, a jim

odpovídající **absolutní četnosti** n_i, $i = 1, 2, ..., k$, lze rozdělení četností vyjádřit způsobem, uvedeným v tabulce 1.1 (symbol n označuje rozsah celého souboru).

Jestliže chceme mezi sebou porovnávat výskyt stejných variant hodnot znaku v souborech různého rozsahu a dospět tak ke snazší interpretaci výsledků, je vhodné převést absolutní četnosti na **relativní četnosti**. Relativní četnosti p_i získáme jako podíly jednotlivých absolutních četností a celkového rozsahu souboru

$$p_i = \frac{n_i}{\sum_{i=1}^{k} n_i} \tag{1.1}$$

přičemž platí

$$\sum_{i=1}^{k} p_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i = \frac{1}{n} n = 1. \tag{1.2}$$

Tab. 1.1 Schéma tabulky rozdělení četností

Varianta hodnot znaku x_i	Četnost		Kumulativní četnost	
	absolutní n_i	relativní p_i	absolutní	relativní
x_1	n_1	p_1	n_1	p_1
x_2	n_2	p_2	$n_1 + n_2$	$p_1 + p_2$
...	...	...	...	...
x_k	n_k	p_k	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$
Součet	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$	×	×

Relativní četnosti jsou uvedeny ve třetím sloupci tabulky 1.1. Označení „×" v řádku „Součet" u posledních dvou sloupců tabulky označuje, že hodnota součtu nemá smysl.

Kromě uvedených dvou forem vyjádření rozdělení četností (absolutní, relativní) konstruujeme někdy rovněž rozdělení **kumulativních absolutních četností** a **kumulativních relativních četností**. Podávají informaci o tom, kolik jednotek souboru, resp. jaká poměrná část souboru má variantu hodnoty znaku menší nebo rovnou určité dané hodnotě. Kumulativní absolutní i relativní četnosti jsou uvedeny v posledních dvou sloupcích tabulky 1.1.

Příklad 1.1

Z personálního útvaru fakulty určité univerzity jsme získali údaje o aktuálním zařazení 75 pedagogů do jednotlivých tarifních tříd a o počtu let, které každý pedagog odpracoval. Všechny zjištěné údaje jsou souhrnně uvedeny v tabulce 1.2.

Tab. 1.2 Údaje k příkladu 1.1

(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	9	4	20	11	16	39	10	11	58	13	31
2	13	38	21	10	13	40	10	13	59	11	14
3	9	1	22	9	3	41	10	14	60	12	25
4	12	18	23	12	26	42	11	17	61	11	19
5	12	16	24	10	13	43	11	16	62	11	14
6	13	42	25	10	13	44	10	9	63	10	9
7	13	29	26	11	26	45	11	13	64	10	8
8	13	32	27	11	29	46	11	13	65	13	35
9	9	3	28	10	9	47	12	25	66	9	2
10	10	7	29	11	11	48	12	21	67	12	18
11	13	44	30	12	34	49	11	11	68	10	9
12	10	11	31	11	24	50	13	34	69	10	8
13	10	9	32	11	22	51	10	7	70	11	16
14	9	2	33	11	22	52	11	12	71	12	27
15	13	32	34	11	15	53	12	21	72	9	3
16	9	3	35	10	6	54	11	11	73	11	24
17	11	16	36	9	3	55	10	6	74	10	14
18	9	2	37	13	31	56	11	14	75	11	18
19	12	24	38	12	24	57	11	16			

Legenda: (1) = pořadové číslo pedagoga
(2) = tarifní zařazení pedagoga do platové třídy
(3) = počet odpracovaných let

Údaje, tak jak jsou uvedeny v tabulce 1.2, jsou značně nepřehledné, zabírají mnoho místa, nejsou použitelné ke zjednodušující a zpřehledňující interpretaci či pro grafické zobrazení a ani zkušený pracovník z nich nezíská základní informace o rozdělení pedagogů podle tarifních tříd.

Uspořádáme proto tyto údaje do tabulky rozdělení četností. Ze třetího sloupce tabulky 1.3 např. vidíme, že nejvíce pedagogů – konkrétně 24 pedagogů, tj. 32,1 % – má právě 11. tarifní třídu. Z posledního sloupce téže tabulky je vidět, že 70,7 % bylo zařazeno nanejvýš do 11. tarifní třídy, 86,7 % má tarifní třídu 12 nebo nižší apod.

Tab. 1.3 Tabulka rozdělení četností dat z příkladu 1.1

Tarifní třída x_i	Počet pracovníků n_i	Relativní četnosti p_i	Kumulativní absolutní četnosti	Kumulativní relativní četnosti
9	10	0,133	10	0,133
10	19	0,253	29	0,386
11	24	0,321	53	0,707
12	12	0,160	65	0,867
13	10	0,133	75	1,000
Součet	75	1,000	×	×

■

Tabulka rozdělení četností je vhodným prostředkem zpracování údajů diskrétního (nespojitého) statistického znaku, který nabývá pouze menšího počtu obměn (variant) hodnot. Máme-li však k dispozici údaje o spojitém statistickém znaku či o statistickém znaku, jenž je sice diskrétní, ale z věcného hlediska může nabývat velkého počtu nejrůznějších obměn (v naší ukázce je to znak „počet odpracovaných let"), pak je použití tabulky rozdělení četností nevýhodné až sporné a její vypovídací možnosti jsou velmi malé. Proto konstruujeme **intervalové rozdělení četností**, v němž **variační rozpětí** souboru R (tímto termínem rozumíme rozdíl mezi maximální a minimální zjištěnou hodnotou znaku) rozdělíme na určitý počet stejně širokých intervalů, a potom zjistíme počty jednotek, patřících do těchto intervalů.

Při sestavování intervalového rozdělení četností je třeba především vyřešit otázku, do jakého počtu stejně širokých intervalů sledované hodnoty roztřídit. Při určování počtu intervalů se snažíme potlačit náhodné kolísání četností, ale zároveň nesmíme setřít charakteristické rysy rozdělení. Na stanovení počtu intervalů a jejich délku neexistuje jednotný názor ani obecný předpis. Obecně lze konstatovat, že počet intervalů nemá být ani příliš malý, ani příliš velký. Příliš malý počet intervalů vede k velmi hrubému, zjednodušenému pohledu na rozdělení četností, příliš velký počet intervalů vede pak k tomu, že se rozdělení četností stává nepřehledným a nevyniknou zákonitosti charakteristické pro daný soubor.

Počet intervalů má mít tedy hlavně určitou interpretační logiku vzhledem k věcné podstatě analyzovaných údajů a celkové vyznění konstrukce intervalového rozdělení četností má být pro uživatele maximálně přehledné a srozumitelné.

Druhý problém při konstrukci intervalového rozdělení četností je spojen s otázkou, jak určovat hranici intervalů, aby bylo možné jednotlivé hodnoty znaku zařadit do příslušných intervalů jednoznačně. Konkrétní hodnota znaku totiž může ležet na hranici některého z intervalů; zde se lze setkat s vícero doporučeními, jak hraniční hodnoty znaku do intervalů zařazovat, ale vždy musí platit, že ať už se rozhodneme jakkoli, musí být naše řešení tohoto problému transparentní a uživateli zcela srozumitelné.

Příklad 1.2

Setřídíme údaje o počtu odpracovaných let v souboru 75 pedagogů, uvedené v tabulce 1.2, do tabulky intervalového rozdělení četností. Soubor přitom roztřídíme do pěti intervalů s jednotnou délkou 10 let. Je zřejmé, že jsme při konstrukci rozdělení mohli postupovat i jinak, tj. mohli jsme volit intervaly i kratší nebo delší. Výsledné intervalové rozdělení je uvedeno v tabulce 1.4.

Tab. 1.4 Tabulka intervalového rozdělení četností

Počet odpracovaných let	Počet pedagogů n_i	Relativní četnosti p_i	Kumulativní absolutní četnosti	Kumulativní relativní četnosti
1–10	21	0,280	21	0,280
11–20	28	0,373	49	0,653
21–30	16	0,213	65	0,866
31–40	8	0,107	73	0,973
41–50	2	0,027	75	1,000
Součet	75	1,000	×	×

Z tabulky 1.4 lze vyčíst, že 28 pedagogů (37,3 %) má odpracováno 11 až 20 let; 65 pedagogů (86,6 %) nepracuje déle než 30 let (tj. pracuje nanejvýš 30 roků) atd.

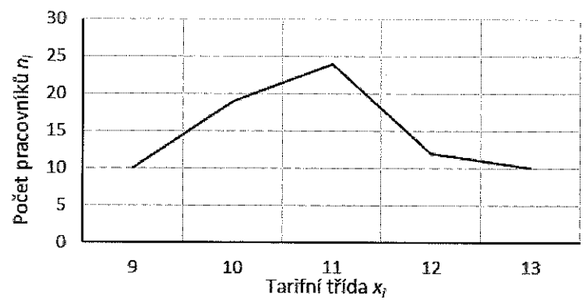
■

1.4.2 Statistické grafy

Vedle statistických tabulek je účinnou formou zobrazování údajů graf. Grafy dávají rychlou a přehlednou představu o tendencích a charakteristických rysech analyzovaných jevů a jsou i vhodným prostředkem popularizace statistických výsledků. Z hlediska konstrukce se grafy dělí do různých skupin. O některých z nich se zmíníme.

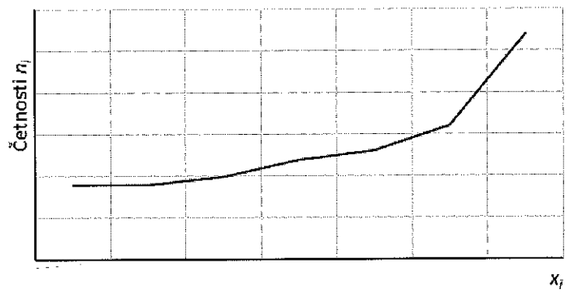
Spojnicové a sloupcové grafy

Rozdělení četností kvantitativních (nebo pořadových) znaků lze graficky znázornit **polygonem četností**. Ten pro diskrétní znak vznikne tak, že v pravoúhlém souřadném systému spojíme úsečkami body o souřadnicích (x_i, n_i), $i = 1, 2, ..., k$, kde x_i jsou hodnoty znaku a n_i jim odpovídající absolutní četnosti (lze použít i relativní četnosti p_i). Pro údaje z tabulky 1.3 je polygon absolutních četností uveden na obrázku 1.1.



Obr. 1.1 Polygon četností

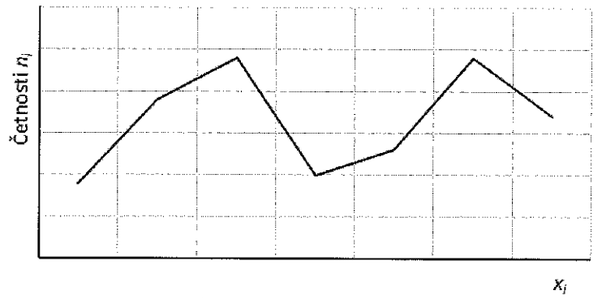
Pohled na tabulku 1.3 (tabulka rozdělení četností) a obrázek 1.1 (polygon četností) ukazuje, že četnosti jednotlivých variant od minimální hodnoty tarifní třídy x_{min} (= 9) s absolutní četností 10 do maximální hodnoty x_{max} (= 13) s absolutní četností také 10 jsou v souboru 75 pedagogů dost různé. Vidíme, že největší četnost má 11. tarifní třída (24 pracovníků), v níž je vrchol rozdělení četností. Nejčetnější variantu hodnoty znaku nazýváme **modus** (značíme jej $\hat{x}$, v příkladu $\hat{x} = 11$). Poloha vrcholu je u rozdělení četností důležitá. Důležitý je však i celkový tvar rozdělení četností. Dle počtu vrcholů rozlišujeme jednovrcholová a vícevrcholová rozdělení četností. Jednovrcholová (unimodální) rozdělení četností se vyskytují nejčastěji. V zásadě existují dva typy těchto rozdělení. Jeden je takový, že modus se nachází mezi minimální a maximální hodnotou proměnné. S tímto typem o jednom vrcholu uprostřed se setkáváme nejčastěji (viz obrázek 1.1). Dalším typem jednovrcholového rozdělení je rozdělení „J", u něhož má největší četnost minimální či maximální hodnota proměnné (obrázek 1.2).



Obr. 1.2 Schéma rozdělení „J"

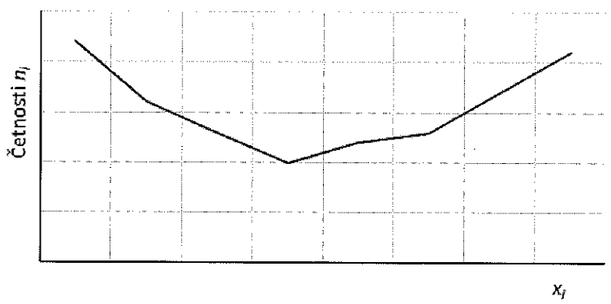
Je-li modus chápán obecněji jako obměna znaku s největší četností vzhledem k nejbližšímu okolí, pak modální obměna znaku je každá varianta, jejíž četnost je větší než četnosti dvou sousedních obměn. Vícevrcholové rozdělení četností má tedy více než jeden modus. Nazývá se proto též vícemodálním neboli multimodálním rozdělením. Častým případem multimodálního rozdělení bývá rozdělení četností, které má dva vrcholy. Mluvíme pak o bimodálním rozdělení četností. Jeho nejčastější případ je na obrázku 1.3.

Zvláštním případem bimodálního rozdělení je rozdělení „U" (viz obrázek 1.4), které má vrcholy na dvou okrajích. V rozděleních typu „U" je pro charakterizování polohy rozdělení četností důležitá varianta znaku s nejmenší četností; ta se nazývá **antimodus**.



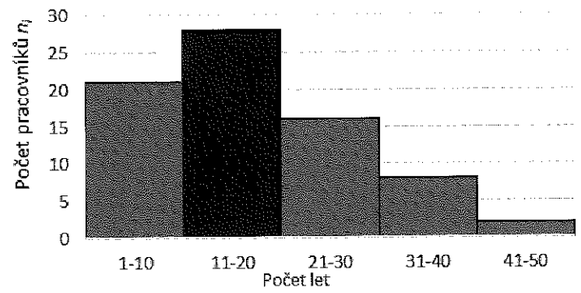
Obr. 1.3 Schéma bimodálního rozdělení

Větší počet vrcholů rozdělení četností má většinou původ v nestejnorodosti zkoumaného statistického souboru, z něhož je v takovém případě možné (a většinou i nezbytné) vytvořit vhodným roztříděním údajů tolik statistických souborů, kolik mělo původní rozdělení četností vrcholů. Takto lze převést vícevrcholové rozdělení četností na obvyklá jednovrcholová rozdělení četností. Třídění údajů v souboru podle podobných vlastností do podskupin (neboli do strat, tedy tzv. stratifikace) je často nezbytnou primární fází práce s daty a vede ke zlepšení kvality výsledků použitých statistických metod.



Obr. 1.4 Schéma rozdělení „U"

Pro grafické vyjádření intervalového rozdělení četností se nejčastěji používá histogram četností. Je to sloupcový graf tvořený obdélníky, jejichž základny mají délku zvolených intervalů a jejichž výšky odpovídají příslušným absolutním četnostem n_i jednotlivých intervalů (ev. relativním četnostem p_i). Histogram intervalového rozdělení četností z tabulky 1.4 je na obrázku 1.5. Četnosti modálního intervalu (tj. intervalu s největší četností $n_i = 28$) jsou na obrázku zabarveny černě.



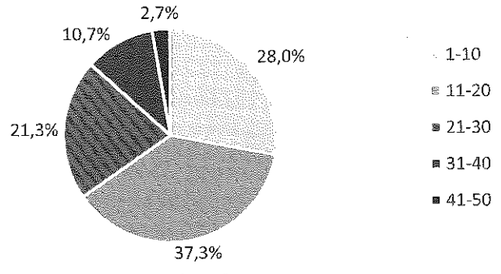
Obr. 1.5 Histogram rozdělení četností

Bodové grafy

Bodové grafy používají jako vizuální prostředky body umísťované v souřadnicové soustavě. Slouží ke znázorňování závislosti mezi dvěma kvantitativními znaky (popř. ke znázorňování průběhu časové řady, viz kapitola 5). Vodorovná osa je stupnicí pro hodnoty kvantitativního znaku x_i (vysvětlující proměnná), svislá osa je určena pro vynášení hodnot druhého kvantitativního znaku y_i (vysvětlovaná proměnná). S bodovými grafy se setkáme později zejména v části o regresní a korelační analýze (kapitola 4).

Výsečové grafy

K vyjádření struktury statistického souboru vzhledem k diskrétnímu (a též k pořadovému a kategoriálnímu) znaku je velmi oblíbený výsečový graf. Relativní četnosti variant hodnot znaku zde znázorníme pomocí výsečí kruhu, které získáme rozdělením středového úhlu 360° úměrně k podílům jednotlivých částí zobrazovaného jevu (podíly často vyjadřujeme v procentech). Např. pro znak „počet odpracovaných let", překódovaný do intervalů a uvedený v příkladu 1.1, bude výsečový graf vypadat tak, jak je na obrázku 1.6.

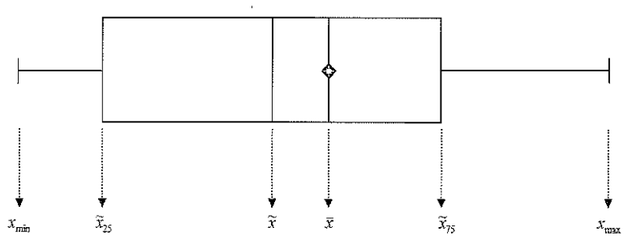


Obr. 1.6 Výsečový graf

Krabičkový graf

Krabičkový graf slouží ke znázornění vztahu krajních hodnot souboru (minimum, maximum), průměru a kvartilů (o těchto pojmech pojednáme podrobně v následujících částech 1.5 a 1.6). Podobu krabičkového grafu přibližuje obrázek 1.7.

Často je vhodná a účelná souběžná vizuální analýza více krabičkových grafů najednou, konstruovaných buď pro různé statistické znaky jednoho statistického souboru, nebo pro jeden statistický znak sledovaný v různých statistických souborech. To umožňuje rychlé a přehledné porovnávání analyzovaných vlastností pomocí důležitých hodnot znaku v souboru.



Obr. 1.7 Znázornění významu jednotlivých bodů v krabičkovém grafu

1.5 Kvantily

Kvantil je hodnota, která rozděluje hodnoty určitého statistického znaku v souboru na dvě části – jedna obsahuje ty hodnoty, které jsou menší (nebo stejné) než tento kvantil, druhá část naopak obsahuje hodnoty, které jsou větší (nebo stejné) než kvantil.

Kvantily jsou poměrně oblíbenými a hojně používanými charakteristikami pro svou srozumitelnou intepretaci. S pojmem kvantilu se setkáváme v řadě běžných situací, jak napoví příklad 1.3.

Příklad 1.3

Při hodnocení životní úrovně se můžeme setkat s novinovou zprávou: „V zemi XY se pod hranicí životního minima, které činí 10 399 (měnových jednotek této země) měsíčně, nachází 15 % obyvatel". V tomto případě měsíční částka 10 399 představuje patnáctiprocentní kvantil hodnot statistického znaku výše měsíčního příjmu v souboru obyvatel dané země. Pokud by se třeba za dva roky pod touto hranicí nacházelo už jen 12 % obyvatel (za předpokladu, že se hranice životního minima ekonomicky nezměnila), šlo by nyní o kvantil dvanáctiprocentní.

■

Můžeme tedy upřesnit, že kvantil je hodnota v souboru určená tak, že hodnoty, které jsou menší (a stejné), tvoří určitou stanovenou část rozsahu statistického souboru,

např. 1, 15, 50, 90 % apod., zatímco hodnoty, které jsou větší (nebo stejné), tvoří zbývající část rozsahu souboru, tj. např. 99, 85, 50, 10 % atd. Kvantil proměnné x, který odděluje $100 \cdot p$ % nižších hodnot proměnné x (kde p je relativní četnost menších hodnot) od $(1-p) \cdot 100$ % větších hodnot proměnné x, nazýváme ho $100p$% kvantilem proměnné x a označujeme ho jako $\tilde{x}_{100p}$.

Tak např. $\tilde{x}_{50}$ je 50% kvantil, který člení statistický soubor na dvě stejně četné poloviny. Nazývá se **medián** neboli **prostřední hodnota** a označuje se zpravidla jen $\tilde{x}$. Při lichém rozsahu souboru je medián jednoduše vždy hodnota konkrétní prostřední statistické jednotky souboru (předtím ovšem nesmíme opomenout hodnoty statistického znaku uspořádat podle velikosti od nejmenší k největší a až teprve v tomto uspořádání hledat medián). Při sudém rozsahu souboru však medián leží mezi dvěma prostředními statistickými jednotkami, proto z hodnot znaku těchto dvou jednotek obvykle stanovíme průměr (střed) a až ten bude označen jako medián.

Mezi další nejčastěji používané kvantily patří **kvartily**, **decily** a **percentily**. Kvartily jsou hodnoty, které dělí uspořádaný statistický soubor na čtyři části, přičemž každá část obsahuje 25 % jednotek. Kvartily jsou celkem tři, dolní kvartil $\tilde{x}_{25}$ odděluje čtvrtinu nejnižších hodnot znaku. Prostřední kvartil – samozřejmě je to medián $\tilde{x}$ – rozděluje hodnoty znaku na dvě stejné (stejně početné) části, z nichž každá obsahuje 50 % jednotek. Horní kvartil $\tilde{x}_{75}$ odděluje 75 % nejnižších hodnot znaku od zbývajících 25 % hodnot znaku.

Obdobně jsou definovány decily $\tilde{x}_{10}$, $\tilde{x}_{20}$, ..., $\tilde{x}_{90}$, které dělí soubor na 10 stejně obsazených částí (decilů je devět), a percentily $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_{99}$, které rozdělují soubor na 100 stejně obsazených částí (percentilů je 99). Výpočet kvantilů je jednoduchý, jak ukazuje příklad 1.4.

Příklad 1.4

Uvažujme pracoviště se sedmi pracovníky, kteří měli za čtvrtletí tyto počty dnů absence, seřazené již podle velikosti: 0, 0, 2, 3, 5, 8, 66. Podle definice je medián tohoto souboru rovný 3 (dny), protože tato hodnota stojí přesně uprostřed seřazených hodnot. Charakteristickou vlastností mediánu je, že ho neovlivňují extrémní hodnoty. Takovou hodnotou je zde číslo 66 (v tomto případě to např. může být dlouhodobě nemocný pracovník). Kdyby tento pracovník byl nemocen místo 66 dnů jen třeba pět dnů, medián by se nezměnil. Teprve kdyby jeho absence byla nižší než tři dny, došlo by k posunu mediánu.

Díky uvedené extrémní hodnotě lze v tomto případě očekávat značný rozdíl proti aritmetickému průměru (viz dále část 1.6)

$$\bar{x} = \frac{0+0+2+3+5+8+66}{7} = 12,$$

což je čtyřnásobek mediánu. V tomto případě, kdy v souboru existuje odlehlá, tj. výrazně extrémní hodnota, medián lépe charakterizuje úroveň absencí než aritmetický průměr.

■

Pokud bychom chtěli vypočítat ještě např. dolní kvartil $\tilde{x}_{25}$ a zároveň rovněž horní kvartil $\tilde{x}_{75}$, tedy kdybychom chtěli soubor pracovníků rozdělit na čtvrtiny, postupovali bychom tak, že dolním kvartilem bude hodnota počtu absencí pracovníka, jehož pořadové číslo bude $(n+1)/4$, tj. v našem konkrétním příkladu 1.4 půjde o pořadové číslo $(7+1)/4 = 8/4 = 2$, to značí, že je to druhý pracovník, který ve sledovaném čtvrtletí měl 0 dnů absencí. Pro horní kvartil použijeme analogicky formuli $3 \cdot (n+1)/4 = 6$, tedy půjde o šestého pracovníka ve výše uspořádané řadě; počet jeho absencí ve sledovaném čtvrtletí činil osm dnů.

Poznamenejme ještě, že poměrně často se stane, že pořadové číslo, např. $(n+1)/4$ apod., nebude číslem celým. Kdyby počet pracovníků sledovaného pracoviště byl třeba osm (a ne sedm), potom by odpovídající pořadové číslo pro určení dolního kvartilu bylo $(8+1)/4 = 9/4 = 2{,}25$. Jako dolní kvartil by pak bylo nutné stanovit některou z hodnot počtu absencí, která leží mezi počtem absencí pracovníků s pořadovými čísly 2 a 3. Zde se postupuje jednoduše tak, že se obvykle vezme střed mezi dvěma hodnotami, které představují počet absencí 2. a 3. pracovníka, tj. vezme se střed z počtu jejich absencí, tedy $(0+2)/2 = 1$ absence.

1.6 Míry úrovně (polohy)

Jak jsme viděli v předcházejícím výkladu, poskytuje rozdělení četností sice užitečnou informaci a přehled o základní struktuře zkoumaného statistického souboru, ale popisovat a zejména porovnávat několik souborů pouze pomocí tabulek rozdělení četností by bylo značně pracné a těžkopádné. Z těchto důvodů se snažíme shrnout informaci obsaženou ve zjištěných údajích o kvantitativním statistickém znaku a vyjádřit ji v koncentrované formě pomocí určitých charakteristik, kterých by nebylo příliš mnoho, ale které by charakterizovaly základní rysy zkoumaného souboru a které by měly zcela srozumitelnou interpretaci.

Při popisu statistických souborů nás zajímá především **úroveň** (poloha) rozdělení četností a **variabilita** rozdělení. Méně často se zaměřujeme na dvě další vlastnosti, šikmost (asymetrii) a špičatost (exces) rozdělení. O těch se zmiňovat nebudeme.

Za základní vlastnost rozdělení lze považovat jeho úroveň. Měří se pomocí různých druhů středních hodnot, což jsou jednoduché číselné charakteristiky, pomocí kterých můžeme nahradit a zobecnit hodnoty souboru. Počítají-li se střední hodnoty ze všech jednotek statistického souboru, nazývají se **průměry**. Nejdůležitější z nich jsou **průměr aritmetický, harmonický, geometrický** a **kvadratický**.

Do další skupiny středních hodnot lze zařadit ty střední hodnoty, které jsou založeny pouze na některých vybraných hodnotách souboru. Nejdůležitější z nich jsou medián a modus. Již z této stručné charakteristiky vyplývá, že průměry, založené na velikosti hodnoty znaku u všech statistických jednotek, představují většinou výstižnější charakteristiku polohy než střední hodnoty ostatní.

Proto považujeme za správné, aby příslušný průměr byl jakožto charakteristika polohy spočítán vždy, modus (pokud jej vůbec použít lze) a zejména medián pak jen jako střední hodnoty doplňkové, pokud jich bude k charakteristice souboru zapotřebí. Pojmům medián a modus jsme se věnovali už výše, nyní přiblížíme průměry.

Průměry

Výrazně převažujícím druhem průměru, který má uplatnění při řešení téměř všech úloh statistiky, je **průměr aritmetický**. Ze zjištěných hodnot $x_1, x_2, ..., x_n$ (které nemusí být uspořádány), kde n je celkový počet pozorování, lze **prostý aritmetický průměr**, který značíme $\bar{x}$, vypočítat jako

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} . \qquad (1.3)$$

Příklad 1.5

Máme k dispozici údaje o počtech výrobků vyrobených za směnu v souboru 15 pracovníků určitého provozu: 7, 4, 8, 7, 5, 5, 9, 7, 6, 5, 8, 6, 7, 5, 4. Průměrný počet výrobků vyrobených za směnu jedním pracovníkem získáme podle (1.3) tak, že jednotlivé počty výrobků sečteme a součet vydělíme počtem pracovníků, tj.

$$\bar{x} = \frac{93}{15} = 6,2 .$$

Počet výrobků, vyrobených za směnu v průměru jedním pracovníkem, je tedy 6,2. ■

Jsou-li zjištěné hodnoty statistického znaku uspořádány do tabulky rozdělení četností, používáme pro výpočet aritmetického průměru vzorec

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + ... + x_k n_k}{n_1 + n_2 + ... + n_k} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} . \qquad (1.4)$$

Četnosti $n_1, n_2, ..., n_k$ udávají váhu, která je přisuzována jednotlivým variantám hodnot znaku $x_1, x_2, ..., x_k$. Aritmetický průměr, počítaný podle vzorce (1.4), se proto nazývá **vážený aritmetický průměr**.

Příklad 1.6

Údaje z předcházejícího příkladu 1.5 o počtech výrobků, vyrobených za směnu v souboru 15 pracovníků určitého provozu, sestavíme do tabulky rozdělení četností a z této tabulky vypočteme znovu průměrný počet výrobků vyrobených za směnu jedním pracovníkem. Tentokrát však přirozeně použijeme vážený aritmetický průměr v podobě (1.4). Údaje i výpočty, potřebné pro dosazení do vzorce (1.4), jsou uvedeny v tabulce 1.5.

Tab. 1.5 Výpočet váženého aritmetického průměru

Počet výrobků vyrobených za směnu jedním pracovníkem x_i	Počet pracovníků, kteří vyrobili uvedený počet výrobků n_i	$x_i n_i$
4	2	8
5	4	20
6	2	12
7	4	28
8	2	16
9	1	9
Součet	15	93

Podle (1.4) snadno určíme, že průměrný počet výrobků, vyrobených za směnu jedním pracovníkem, je opět

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{93}{15} = 6,2 .$$

■

Aritmetický průměr má řadu vlastností, z nichž některé mají význam teoretický, jiné se dají s výhodou použít při jeho výpočtu. Uveďme si je:

1. součet odchylek jednotlivých hodnot od průměru je nulový,
2. aritmetický průměr konstanty je roven této konstantě,
3. přičteme-li ke všem hodnotám znaku konstantu, zvýší se o tuto konstantu také původní aritmetický průměr,
4. násobíme-li jednotlivé hodnoty znaku konstantou, je touto konstantou násoben i původní průměr,
5. násobíme-li váhy (četnosti hodnot) při výpočtu aritmetického průměru konstantou, původní průměr se nezmění,
6. je-li statistický soubor rozdělen do k dílčích podsouborů (skupin), v nichž známe skupinové průměry $\bar{x}_1, \bar{x}_2, ..., \bar{x}_k$ a zároveň počty pozorování $n_1, n_2, ..., n_k$, pak

průměr celého souboru je váženým aritmetickým průměrem těchto skupinových průměrů, kde vahami jsou četnosti těchto podsouborů. Platí tedy

$$\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}_i n_i}{\sum_{i=1}^{k} n_i}. \tag{1.5}$$

Vztahu (1.5) lze rovněž využít k výpočtu aritmetického průměru z intervalového rozdělení četností za situace, kdy neznáme průměry v jednotlivých skupinách. V tomto případě ovšem nemůžeme proto aritmetický průměr vypočítat přesně, nýbrž jej pouze odhadujeme s větší nebo menší přesností. Pokud jsou všechny intervaly ohraničeny, potom předpokládáme, že průměr v každém intervalu je roven jeho středu a jednotlivé intervaly, jak již bylo řečeno, nahrazujeme jejich středy. Výpočet pak provádíme stejně jako výpočet váženého aritmetického průměru v rozdělení četností, přičemž středy intervalů považujeme vlastně za skupinové průměry. Při popsaném způsobu výpočtu se můžeme dopustit chyby, jejíž maximum je rovno polovině průměrné délky intervalu.

Někdy bývají krajní intervaly (dolní nebo horní, resp. někdy i oba) neuzavřené. V takovém případě se buď interval považuje za stejně široký jako bezprostředně následující (předcházející) interval a pak se nahradí středem takto uměle odhadnuté šířky intervalu, nebo se logicky určí minimální (maximální) hodnota znaku v souboru a ta pak slouží za odhad hranice intervalu, který se opět potom nahradí svým středem. Pokud si však nevíme rady s otevřenými intervaly, je vhodnější použít jako míry polohy medián.

Aritmetický průměr ale není průměrem jediným. Máme-li statistický soubor o rozsahu n pozorování $x_1, x_2, \ldots, x_n$, je **harmonický průměr** z těchto hodnot definován jako podíl počtu pozorování a součtu převrácených hodnot znaku, tj.

$$\bar{x}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}. \tag{1.6}$$

Průměr ve tvaru (1.6) se nazývá **prostý harmonický průměr**. V případě, že máme údaje již setříděny do podoby tabulky rozdělení četností, počítáme **vážený harmonický průměr**

$$\bar{x}_H = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}. \tag{1.7}$$

Příklad 1.7

Z údajů v tabulce 1.6 vypočítáme vážený harmonický průměr. Zatím ukážeme pouze jednoduchý výpočet bez konkrétní aplikace. Tu uvidíme v 6. kapitole, konkrétně půjde o použití váženého harmonického průměru při výpočtu Paascheho indexu.

Tab. 1.6 Výpočet váženého harmonického průměru

x_i	n_i	n_i / x_i
4	5	1,25
6	10	1,67
8	12	1,50
11	15	1,36
12	8	0,67
Součet	50	6,45

Dosadíme-li do vzorce (1.7), dostaneme

$$\bar{x}_H = \frac{50}{6,45} = 7,75.$$

■

Praktické používání harmonického průměru ve statistické praxi je vcelku omezené. Ve větším rozsahu se s ním setkáme např. při výpočtu průměrových tvarů souhrnných indexů (viz kapitola 6, která pojednává o metodách statistického srovnávání).

Dalším typem průměru je **geometrický průměr**. Máme-li statistický soubor o rozsahu n pozorování $x_1, x_2, \ldots, x_n$, je **prostý geometrický průměr** definován jako n-tá odmocnina ze součinu hodnot těchto znaků

$$\bar{x}_G = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}. \tag{1.8}$$

Podobně jako v předchozích případech lze zapsat i vzorec **váženého geometrického průměru**

$$\bar{x}_G = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_k^{n_k}}, \text{ kde } n = \sum_{i=1}^{k} n_i. \tag{1.9}$$

Příklad 1.8

Vypočtěme geometrický průměr z čísel 3, 5, 6, 6, 7, 10, 12. Řešením je

$$\bar{x}_G = \sqrt[7]{3 \cdot 5 \cdot 6 \cdot 6 \cdot 7 \cdot 10 \cdot 12} = \sqrt[7]{453\,600} = 6,43.$$

■

Aplikace geometrického průměru je v běžné ekonomické praxi minimální. Setkáváme se s ním zejména při výpočtu průměrného koeficientu růstu hodnot v časové řadě a v teorii indexů (viz kapitoly 5 a 6).

Nakonec se ještě zmíníme o **prostém**, resp. **váženém kvadratickém průměru**. Prostý tvar má podobu

$$\overline{x}_K = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}, \tag{1.10}$$

resp. u váženého tvaru jde o výraz

$$\overline{x}_K = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{\sum_{i=1}^{k} n_i}}. \tag{1.11}$$

Kvadratický průměr má opět jen minimální praktickou konkrétní ekonomickou aplikaci. Jak si však budete moci povšimnout dále v souvislosti s měrami variability, je jedna z těchto měr – směrodatná odchylka – vlastně kvadratickým průměrem odchylek jednotlivých hodnot od jejich aritmetického průměru. To bude mít poměrně značný význam pro pochopení interpretace směrodatné odchylky.

Jsou-li průměry počítány z kladných hodnot znaku x, pak mezi uvedenými čtyřmi typy průměrů platí relace nerovnosti

$$\overline{x}_H \le \overline{x}_G \le \overline{x} \le \overline{x}_K . \tag{1.12}$$

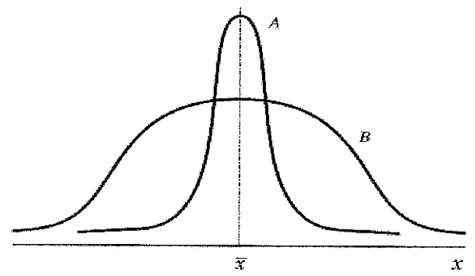
Znaménko rovnosti v (1.12) platí ovšem pouze v případě, kdy jsou všechny hodnoty znaku x v souboru stejné. Jakmile se alespoň jedna odlišuje, platí už relace ostré nerovnosti „$<$“.

V realitě se však nestává, že by pro konkrétní reálná statistická data mělo smysl počítat více než jeden typ z uvedených průměrů. Platí totiž pravidlo, že počítat ten či onen průměr má smysl jen tehdy, má-li věcně smysl příslušná operace s hodnotami znaku x_i. Tak např. má-li věcně smysl tyto hodnoty sčítat, použijeme aritmetický průměr. Má-li věcně smysl počítat převrácené hodnoty znaku x_i, půjde o harmonický průměr atp. V jednom a tomtéž „datovém“ případě nemívá už smysl jiná operace, takže použití více typů průměrů pro stejná data není de facto reálné.

Jedná-li se o obecné vlastnosti uvedených průměrů, jsou většinou analogické s vlastnostmi aritmetického průměru tak, jak jsme je uvedli pod příkladem 1.6 a tabulkou 1.5.

1.7 Míry variability

Střední hodnoty, o kterých jsme mluvili v předcházejícím odstavci 1.6, v sobě shrnují informaci pouze o jedné vlastnosti rozdělení četností – o úrovni hodnot (tedy o poloze hodnot na číselné ose). Často se však setkáváme se situací, že různá rozdělení četností budou mít shodnou polohu, ale přesto se budou od sebe výrazně lišit dalšími vlastnostmi, jak je schematicky znázorněno na obrázku 1.8.



Obr. 1.8 Rozdělení lišící se variabilitou

Hodnoty statistického souboru A jsou těsněji koncentrovány okolo aritmetického průměru, než je tomu u hodnot souboru B. Lze tedy říci, že v souboru A vystihuje průměr úroveň souboru lépe než v souboru B.

Měření variability má v případě kvantitativního znaku význam při posuzování vypovídací schopnosti aritmetického průměru. Obecně lze říci, že vypovídací schopnost aritmetického průměru je tím větší, čím je variabilita sledovaného znaku menší a naopak. Význam měření variability se však neomezuje jen na posouzení vypovídacích schopností středních hodnot. Protože rozšiřují informace o statistickém souboru, setkáme se s nimi téměř ve všech oblastech statistiky. Dokonce se dá říci, že v případě měr variability jde o jedny z nejdůležitějších charakteristik ve statistice vůbec – v řadě statistických disciplín totiž zkoumáme intenzitu odlišností údajů a analyzujeme význam a roli faktorů, jež tyto odlišnosti způsobují. K tomu všemu ovšem nejprve potřebujeme umět popsat právě proměnlivost obměn hodnot statistického znaku.

Měr variability existuje celá řada. Zmíníme se jenom o těch nejdůležitějších, které doznaly praktického použití. Míry, které charakterizují proměnlivost statistického souboru v absolutní velikosti, nazýváme **mírami absolutní variability**. Míry tohoto typu vyjadřují variabilitu ve stejných nebo odvozených měrných jednotkách, v nichž je vyjadřován sledovaný znak. V případě, že porovnáváme variabilitu souborů lišících se svojí úrovní, používáme **míry relativní variability**, které měří variabilitu v poměru k úrovni sledovaného znaku v souboru. Míry relativní variability jsou již bezrozměrná čísla, což také dovoluje porovnávat variabilitu statistických znaků lišících se měrnou jednotkou.

1.7.1 Míry absolutní variability

Nejjednodušší, ale také nejhrubší mírou variability kvantitativního znaku, je **variační rozpětí**. Definuje se jako rozdíl největší a nejmenší hodnoty znaku

$$R = x_{max} - x_{min}. \tag{1.13}$$

Jeho předností je snadnost a rychlost výpočtu i jednoduchá interpretace. Krajní hodnoty řady pozorování, na nichž je variační rozpětí založeno, mohou ovšem být dost nahodilé. Případné extrémní vlivy se projeví především právě na těchto hodnotách. Výskyt jediné extrémní hodnoty znaku tak vyvolává značnou velikost variačního rozpětí. Další nevýhodou je, že tato míra neříká nic o variabilitě hodnot uvnitř variačního rozpětí.

Ve většině případů však dávají jak statistická teorie, tak i statistická praxe přednost takovým mírám variability, jejichž velikost je závislá na všech hodnotách statistického souboru. Z nich je nejvýznamnější taková míra variability, která měří variabilitu hodnot kolem aritmetického průměru a zároveň také variabilitu ve smyslu vzájemných odchylek jednotlivých hodnot znaku. Taková míra se nazývá **rozptyl**. Rozptyl je definován jako průměr čtverců odchylek jednotlivých hodnot znaku x_i od jejich aritmetického průměru, tedy

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}. \tag{1.14}$$

Pro praktické výpočty (zejména v situaci, kdy nemáme k dispozici počítač) není vzorec (1.14) příliš vhodný. Jednoduchou úpravou se dá ale převést do formy vhodnější pro praktické výpočty. Čitatel ve vzorci (1.14) lze totiž vyjádřit takto

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \sum x_i^2 - 2\bar{x}\sum x_i + \bar{x}\sum x_i =$$

$$= \sum x_i^2 - \bar{x}\sum x_i = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2.$$

Můžeme potom používat tzv. **výpočtový tvar rozptylu**, který je samozřejmě zcela ekvivalentní se vzorcem (1.14)

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2 = \overline{x^2} - \bar{x}^2. \tag{1.15}$$

Příklad 1.9

V tabulce 1.7 jsou údaje o výši částek (v Kč), uvedených na vystavených fakturách ve dvou různých odděleních obchodní společnosti. Porovnáme variabilitu fakturovaných částek v obou těchto odděleních společnosti výpočtem rozptylu ve tvaru (1.15). Potřebné výpočty jsou rovněž uvedeny v tabulce 1.7.

Dosadíme-li ze součtového řádku tabulky 1.7 do vzorce (1.15), dostaneme pro první oddělení

$$s_x^2 = \frac{100\,220\,862}{12} - \left(\frac{34\,246}{12}\right)^2 = 207\,375{,}708\,(\text{Kč}^2)$$

a pro druhé oddělení

$$s_y^2 = \frac{112\,820\,518}{12} - \left(\frac{36\,500}{12}\right)^2 = 149\,971{,}694\,(\text{Kč}^2).$$

V oddělení 2 je tak nižší variabilita částek na fakturách než v oddělení 1. Všimněme si ale ještě důležité okolnosti, a sice že měrnou jednotkou vypočteného rozptylu je druhá mocnina koruny, což může stěžovat intepretaci rozptylu obecně jako takového.

Tab. 1.7 Výpočet rozptylu

Faktura	Oddělení 1 x_i	Faktura	Oddělení 2 y_i	x_i^2	y_i^2
1	2 400	13	2 613	5 760 000	6 827 769
2	2 134	14	2 496	4 553 956	6 230 016
3	2 407	15	2 736	5 793 649	7 485 696
4	2 445	16	2 676	5 978 025	7 160 967
5	2 984	17	3 093	8 375 236	9 566 649
6	3 354	18	3 537	11 249 316	12 510 369
7	3 515	19	3 622	12 355 225	13 118 884
8	3 515	20	3 561	12 355 225	12 680 721
9	3 225	21	3 385	10 400 625	11 458 225
10	3 063	22	3 155	9 381 969	9 954 025
11	2 694	23	2 788	7 257 636	7 772 944
12	2 600	24	2 838	6 760 000	8 054 244
Součet	34 246	Součet	36 500	100 220 862	112 820 518

■

Určitou nevýhodou rozptylu z interpretačního hlediska je, že je vždy vyjádřen ve čtvercích použité měrné jednotky – třeba v našem příkladu (výše částek na fakturách) je to tedy „(Kč)^2“. Proto se často variabilita popisuje pomocí kladně vzaté odmocniny z rozptylu, která se nazývá **směrodatná odchylka**

$$s_x = \sqrt{s_x^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}. \tag{1.16}$$

Její výhodou je to, že je uvedena ve stejných měrných jednotkách jako zkoumaný statistický znak (tedy nikoli ve druhé mocnině měrné jednotky, jak je tomu u rozptylu) a na rozdíl od rozptylu ji lze snáze interpretovat; povšimněme si totiž ještě výrazu (1.16). Porovnáme-li jej se vzorcem (1.10), popř. (1.11), můžeme vidět, že v případě směrodatné odchylky se vlastně jedná o kvadratický průměr odchylek hodnot x_i od

jejich průměru $\bar{x}$, jak jsme již výše uvedli. Tak např. směrodatná odchylka částek na fakturách v 1. oddělení z příkladu 1.9

$$s_x = \sqrt{207\ 375{,}708} = 455{,}385 \text{ (Kč)}$$

udává, že průměrná („průměrná" ve smyslu kvadratického průměru) odchylka hodnot od průměru činí přibližně 455 Kč.

Rozptyl z tabulky rozdělení četností určujeme ve váženém tvaru

$$s_x^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{\sum_{i=1}^{k} n_i}, \tag{1.17}$$

který lze analogicky s (1.15) upravit do výpočtového tvaru

$$s_x^2 = \frac{\sum_{i=1}^{k} x_i^2 n_i - \bar{x}\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{\sum_{i=1}^{k} n_i} - \left(\frac{\sum_{i=1}^{k} x_i n_i}{\sum_{i=1}^{k} n_i}\right)^2. \tag{1.18}$$

Příklad 1.10

Obchodní řetězec má 33 prodejen; počty pracovníků v nich ukazuje intervalové rozdělení četností (viz tabulka 1.8). Variabilitu tohoto rozdělení budeme charakterizovat výpočtem rozptylu ve tvaru (1.18). Potřebné výpočty jsou uvedeny rovněž v tabulce 1.8. Pracovat budeme se středy intervalů x_i, takže výsledek bude jen přibližný.

Tab. 1.8 Výpočet rozptylu ve váženém tvaru

Počet pracovníků	Počet prodejen	x_i	$x_i n_i$	$x_i^2 n_i$
1 – 5	9	3	27	81
6–10	8	8	64	512
11–15	8	13	104	1 352
16–20	5	18	90	1 620
21–25	2	23	46	1 058
26–30	1	28	28	784
Součet	33	x	359	5 407

Dosadíme-li údaje ze součtového řádku tabulky 1.8 do vzorce (1.18), dostaneme přibližnou hodnotu rozptylu počtu pracovníků

$$s_x^2 = \frac{5\,407}{33} - \left(\frac{359}{33}\right)^2 = 45{,}5.$$

Odtud směrodatná odchylka $s_x = 6{,}75$; tedy průměrná („průměrná" ve smyslu kvadratického průměru) odchylka hodnot od průměru činí přibližně 6,75 pracovníka.

■

Rozptyl má tyto vlastnosti:

1. rozptyl konstanty je roven nule,
2. přičteme-li ke všem hodnotám znaku konstantu, původní velikost rozptylu se nezmění,
3. násobíme-li všechny hodnoty znaku konstantou, původní rozptyl je násoben čtvercem této konstanty,
4. rozptyl s_z^2 součtu (rozdílu) dvou proměnných $z_i = x_i \pm y_i$ je roven součtu rozptylů obou proměnných zvětšenému (+) nebo zmenšenému (–) o dvojnásobek tzv. **kovariance**, tj.

$$s_z^2 = s_{x\pm y}^2 = s_x^2 + s_y^2 \pm 2s_{xy}.$$

Výraz

$$s_{xy} = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y} = \overline{xy} - \bar{x}\,\bar{y} \tag{1.19}$$

se nazývá **kovariance** proměnných x a y a charakterizuje vzájemnou závislost obou proměnných. Její význam poznáme níže ve 4. kapitole,

5. předpokládejme, že statistický soubor o rozsahu n statistických jednotek je rozdělen do k dílčích podsouborů (skupin), ve kterých známe skupinové rozptyly s_{ix}^2, dále skupinové průměry $\bar{x}_i$ a rovněž četnosti i-tého podsouboru n_i. Schéma této situace je zaznamenáno v tabulce 1.9.

Pak rozptyl celého souboru je dán součtem rozptylu skupinových průměrů a průměru ze skupinových rozptylů

$$s_x^2 = s_{\bar{x}}^2 + \overline{s^2}, \tag{1.20}$$

kde

$$s_x^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2}{\sum_{i=1}^{k} n_i} \tag{1.21}$$

je celkový rozptyl souboru,

$$s_{\bar{x}}^2 = \frac{\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 n_i}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k}\bar{x}_i^2 n_i}{\sum_{i=1}^{k} n_i} - \left(\frac{\sum_{i=1}^{k}\bar{x}_i n_i}{\sum_{i=1}^{k} n_i}\right)^2 \tag{1.22}$$

je rozptyl skupinových průměrů (tzv. meziskupinová variabilita) a konečně

$$\overline{s^2} = \frac{\sum_{i=1}^{k} s_{ix}^2 n_i}{\sum_{i=1}^{k} n_i} \tag{1.23}$$

značí průměr ze skupinových rozptylů (tzv. vnitroskupinová variabilita).

Tab. 1.9 Schéma vztahu celkového a dílčích rozptylů

Dílčí soubor č.	Hodnoty znaku x_{ij}	Dílčí průměry $\bar{x}_i$	Dílčí rozptyly s_{ix}^2	Dílčí četnosti n_i
1	$x_{11}, x_{12}, \ldots, x_{1j}, \ldots, x_{1n_1}$	$\bar{x}_1$	s_{1x}^2	n_1
2	$x_{21}, x_{22}, \ldots, x_{2j}, \ldots, x_{2n_2}$	$\bar{x}_1$	s_{2x}^2	n_1
...	...	...	...	...
i	$x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in_i}$	$\bar{x}_i$	s_{ix}^2	n_i
...	...	...	...	...
k	$x_{k1}, x_{k2}, \ldots, x_{kj}, \ldots, x_{kn_k}$	$\bar{x}_k$	s_{kx}^2	n_k
Součet	×	×	×	n

Vztah (1.20) tedy umožňuje výpočet celkového rozptylu s_x^2 i bez znalosti původních hodnot znaku. Tento rozklad dále dovoluje rovněž hodnotit, jak dalece je celkový rozptyl ovlivněn variabilitou hodnot uvnitř skupin (složka $\overline{s^2}$) a jak dalece variabilitou mezi skupinami (složka $s_{\bar{x}}^2$). Jinak řečeno, zkoumáme otázku, zda velikost celkového rozptylu je více závislá na nestejnosti jednotlivých hodnot znaku uvnitř skupin nebo naopak závisí spíše na relativně významnější odlišnosti skupinových průměrů. S důležitým použitím tohoto rozkladu rozptylu se setkáme později ve 4. kapitole (v části o analýze rozptylu).

Příklad 1.11

V tabulce 1.10 jsou uvedeny průměrné doby doručení zásilky kurýrními službami v určitém městě (v minutách) a zároveň další údaje, charakterizující tyto kurýrní služby. Předpokládáme, že podmínky doručování jsou pro všechny firmy ve městě srovnatelné. Na trhu těchto služeb působí čtyři doručovatelské firmy, které ve sledovaném období tří dnů uskutečnily celkem 131 doručení. Z údajů tabulky 1.10 – jedná se o průměrnou dobu doručení $\bar{x}_i$ (v minutách) u každé z firem, rozptyl těchto časů pro každou firmu s_{ix}^2 a konečně o počet doručení n_i – vypočteme na základě vztahu (1.20) celkovou variabilitu doby doručení. Všechny potřebné údaje a následné výpočty jsou rovněž uvedeny v tabulce 1.10.

Dosadíme-li ze součtového řádku tabulky 1.10 do vzorců (1.22) a (1.23), dostaneme

$$s_x^2 = \left[\frac{177\,231}{131} - \left(\frac{4\,803}{131}\right)^2\right] + \frac{3\,012{,}8}{131} = 8{,}650\,5 + 22{,}998\,5 = 31{,}649,$$

přičemž na této celkové variabilitě se větší měrou podílí vnitroskupinová variabilita (22,998 5), tj. nerovnoměrnost časů jednotlivých doručení uvnitř kurýrních firem (vyjádřená rozptylem časů v těchto firmách s_{ix}^2).

Tab. 1.10 Výpočet celkového rozptylu z dílčích rozptylů

Firma	$\bar{x}_i$	s_{ix}^2	n_i	$\bar{x}_i n_i$	$\bar{x}_i^2 n_i$	$s_{ix}^2 n_i$
A	33	15,5	32	1 056	34 848	496,0
B	35	22,0	34	1 190	41 650	748,0
C	38	24,0	36	1 368	51 984	864,0
D	41	31,2	29	1 189	48 749	904,8
Součet	×	×	131	4 803	177 231	3 012,8

■

Mezi další míry absolutní variability, které se někdy v praxi používají, patří především kvantilové odchylky, které jsou definovány jako aritmetický průměr kladných odchylek sousedních kvantilů. Např. **kvartilová odchylka** je tedy

$$Q = \frac{(\tilde{x}_{75} - \tilde{x}) + (\tilde{x} - \tilde{x}_{25})}{2} = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{2}. \tag{1.24}$$

Podobně bychom konstruovali např. i decilovou či percentilovou odchylku apod.

Základem vztahu (1.24) je v čitateli rozdíl horního a dolního kvartilu. Tento rozdíl kvartilů se nazývá **kvartilové rozpětí** (porovnejte jej – viz výše vzorec (1.13) – s analogickým pojmem variační rozpětí). Podobné míry se ale většinou používají pouze doplňkově.

Příklad 1.12

Ze statistického šetření o výši měsíčních mezd v Česku v roce 2018 jsme zjistili, že dolní kvartil $\tilde{x}_{25} = 18\,399$ Kč a horní kvartil $\tilde{x}_{75} = 50\,485$ Kč. Kvartilová odchylka tedy je (50 485 – 18 399)/2 = 16 043 Kč, kvartilové rozpětí 50 485 – 18 399 = 32 086 Kč. Interval ⟨18 399 ; 50 485⟩, tj. vzdálenost od dolního k hornímu kvartilu, informuje o tom, v jakých mezích se pohybuje prostředních 50 % měsíčních mezd v Česku. Tato znalost může mít pro svou vypovídací schopnost větší cenu, než má v této souvislosti obvykle prezentovaný aritmetický průměr (průměrná měsíční mzda ve stejném období činila 29 613 Kč). Průměr totiž nevypovídá nic o různosti hodnot mezd uvnitř souboru. Proto v některých zemích se spíše než průměrná mzda publikuje kvartilové rozpětí.

■

Obdobně můžeme definovat i **decilovou** a **percentilovou odchylku**, resp. **decilové** a **percentilové rozpětí.** Obecnou nevýhodou všech kvantilových měr variability je to, že nezachycují variabilitu všech hodnot znaku a vzhledem k jejich konstrukci je nelze hlouběji analyzovat a rozkládat. S ohledem na tyto skutečnosti i na vlastnosti, které zařazují rozptyl a směrodatnou odchylku do širších souvislostí s ostatními statistickými charakteristikami, lze považovat rozptyl a směrodatnou odchylku za nejdůležitější míry variability pro kvantitativní znak.

1.7.2 Míry relativní variability

Charakteristiky variability uvedené v předcházejícím odstavci jsou vyjádřeny ve stejných (nebo odvozených) měrných jednotkách, jako jsou hodnoty analyzovaného znaku či jejich průměr. Měří variabilitu absolutně. Diskutabilní je však srovnávat pomocí těchto absolutních měr variabilitu statistického znaku u dvou či více souborů, které se výrazně liší úrovní hodnot znaku, i když měrné jednotky budou u všech souborů shodné. Srovnávat absolutně variabilitu statistických znaků vyjádřených dokonce v různých měrných jednotkách není však možné vůbec. V uvedených případech proto používáme charakteristiky relativní variability, které vliv úrovně či vliv měrné jednotky vylučují tím, že charakteristiky absolutní variability dávají do poměru k průměru nebo k mediánu. Tak vznikne bezrozměrná charakteristika.

Nejznámější mírou relativní variability je **variační koeficient**

$$v_x = \frac{s_x}{\bar{x}}, \tag{1.25}$$

který je definován jako poměr směrodatné odchylky a aritmetického průměru. Variační koeficient je bezrozměrné číslo, jeho stonásobek ($100v_x$) udává variabilitu v procentech. Podle velmi orientačního pravidla variační koeficient vyšší než 50 % je už projevem značné nesourodosti statistického souboru. Možnost interpretovat variační koeficient v procentech ale dost často uživatele svádí k chápání jeho definičního oboru od 0 % do 100 %, resp. od 0 do 1. Je to zásadní omyl. Vzhledem k tomu, že nelze najít žádné omezení na číselné ose zprava pro velikost směrodatné odchylky, může obecně být směrodatná odchylka i větší než aritmetický průměr stejných hodnot, a pak tedy variační koeficient je větší než 1 (resp. větší než 100 %). A navíc variační koeficient není strukturální ukazatel (čitatel zlomku není poměrnou součástí jmenovatele), takže k vymezení definičního oboru od 0 % do 100 % není žádný důvod.

Příklad 1.13

Uvažujme denní produkci ve dvou firmách. Produkce firmy A se vykazuje v tisících kusů, produkce firmy B v tunách. Na základě údajů v tabulce 1.11 máme posoudit, ve které z firem byla během sledovaného období 10 dnů výroba rovnoměrnější.

Tab. 1.11 Zadání a pomocné výpočty pro variační koeficient

Den i	Produkce firmy A (tis. ks) x_i	Produkce firmy B (tuny) y_i	x_i^2	y_i^2
1	1	6	1	36
2	2	6	4	36
3	2	5	4	25
4	3	8	9	64
5	2	9	4	81
6	4	4	16	16
7	2	4	4	16
8	1	6	1	36
9	2	5	4	25
10	4	7	16	49
Součet	23	60	63	384

Protože výrobky jsou v různých měrných jednotkách, nelze k řešení úlohy použít měr absolutní variability. Použijeme proto variační koeficient (1.25). Nejprve vypočteme podle (1.15) rozptyly a podle (1.16) směrodatné odchylky

$$s_x^2 = \frac{63}{10} - 2{,}3^2 = 1{,}00, \qquad s_x = 1{,}00$$

a dále

$$s_y^2 = \frac{384}{10} - 6^2 = 2{,}40, \qquad s_y = 1{,}55.$$

Dosazením do (1.25) dostaneme variační koeficienty

$$v_x = \frac{s_x}{\bar{x}} = \frac{1}{2{,}3} = 0{,}43\ (= 43\ \%),$$

$$v_y = \frac{s_y}{\bar{y}} = \frac{1{,}55}{6} = 0{,}26\ (= 26\ \%).$$

Porovnání obou variačních koeficientů ukazuje, že relativně rovnoměrnější (tj. méně variabilní) je v dané dekádě výroba ve firmě B.

■

Vedle variačního koeficientu se k měření relativní variability používají i některé další míry, např. relativní kvartilová odchylka

$$Q_{rel} = \frac{\tilde{x}_{75} - \tilde{x}_{25}}{\tilde{x}_{75} + \tilde{x}_{25}}, \tag{1.26}$$

analogicky relativní decilová a relativní percentilová odchylka. Jejich význam je značně omezený a používají se pouze jako hrubý ukazatel míry variability.

1.7.3 Variabilita pořadového a kategoriálního statistického znaku

Řada statistických dat, pořízených např. při marketingových průzkumech i jinde, je charakteristická poměrně nízkým zastoupením měřitelných kvantitativních znaků. Například pocity respondentů a jejich spotřebitelské postoje k určitému výrobku nebo službě jsou těžko vyjádřitelné měřitelným způsobem. Jindy je velmi nesnadné a často až nemístné klást přímé otázky třeba na mzdu respondentů, na výši rodinných výdajů, na výši úspor apod. Do nitra respondentových názorů, zkušeností a postojů se v takových případech snažíme vstoupit spíše „opisem". Ten může být někdy cíleně rámován použitím vhodně konstruované škály odpovědí, ale může podle okolností být také až extrémně volný (zde jde o tzv. otevřené odpovědi, o volný popis pocitů apod.). Podobně je běžné, že odpovědi respondentů mají povahu slovního (kategoriálního) znaku (jakou univerzitu respondent vystudoval, jakou politickou stranu by volil apod.).

Pro statistickou analýzu je nutné tyto neměřitelné odpovědi určitým způsobem kvantifikovat. U kategoriálního znaku tento požadavek obvykle nahrazujeme tím, že místo s kategoriemi pracujeme s jejich četnostmi výskytu. U pořadových znaků, majících nejčastěji podobu odpovědí na číselné škále, lze postupovat podobně, tj. analyzovat především četnosti výskytu jednotlivých variant odpovědí na škále. Častým požadavkem je pak určit variabilitu odpovědí. Jak je patrno, použití rozptylu (1.14) je buď sporné, nebo nemožné (např. u slovních odpovědí nelze určit průměr, takže o rozptylu je nesmysl vůbec uvažovat).

Pro určení **variability pořadového znaku**, definovaného v podobě lineární bodové stupnice, kdy předpokládáme stejné vzdálenosti mezi variantami hodnot (např. škála ve tvaru uvedeném v příkladu 1.14), můžeme použít **poměrový koeficient diferenciace** P_D. Jde o podíl čtyřnásobku rozptylu odpovědí k druhé mocnině šířky škály R

$$P_D = \frac{4s_x^2}{R^2}. \tag{1.27}$$

Protože jeho hodnota se vždy pohybuje od 0 do 1, tedy v intervalu $\langle 0; 1\rangle$, lze pomocí něho porovnávat proměnlivost (nestejnost) odpovědí respondentů v jednotlivých otázkách bez ohledu na to, jak „široká" (tj. kolikabodová) stupnice je pro odpověď respondentům nabízena. Pro úplnost uveďme, že základem vzorce (1.27) je poměr s_x^2/R^2; tento poměr se pohybuje vždy jen v mezích $\langle 0; 0{,}25\rangle$. Vynásobení konstantou „4" posouvá tyto meze na hodnoty v intervalu $\langle 0; 1\rangle$, což usnadňuje interpretaci získaného výsledku (stejný postup lze aplikovat i v případě kvantitativní proměnné).

Za prahovou úroveň diferenciace názorů se považují hodnoty P_D zhruba v pásmu 0 až 0,10. Pokud totiž odpovědi ve škále na určitou otázku takto málo diferencují postoje respondentů, nastoluje to vážný problém. Tím je okolnost, zda je hodnocené hledisko v daném šetření vůbec hlediskem relevantním. To pak ale vybízí k revizi celkového designu průzkumu, popř. k jeho korekci (otázka nemusela být dobře položena, postoj v ní zkoumaný se může respondentům jevit jako nevýznamný nebo nemusí mít vztah k předmětné problematice apod.).

Příklad 1.14

Personální oddělení uspořádalo mezi svými 50 zaměstnanci anketu, kde se pracovníci mohli vyjádřit k úrovni nabízeného stravování. Byla jim nabídnuta pětibodová číselná škála. Četnosti jejich odpovědí a zároveň i výpočet poměrového koeficientu diferenciace jsou uvedeny v tabulce 1.12.

Tab. 1.12 Výpočet poměrového koeficientu diferenciace

Škála odpovědí	x_i	Počet odpovědí n_i	$x_i n_i$	$(x_i - \bar{x})^2 n_i$
1 (= velmi spokojen)	1	6	6	24,48
2 (= spíše spokojen)	2	10	20	10,40
3 (= napůl spokojen)	3	17	51	0,01
4 (= spíše nespokojen)	4	11	44	10,56
5 (= velmi nespokojen)	5	6	30	23,52
Součet	x	50	151	68,97

Průměr odpovědí je

$$\bar{x} = \frac{151}{50} = 3{,}02,$$

rozptyl

$$s_x^2 = \frac{68{,}97}{50} = 1{,}38.$$

Odtud poměrový koeficient diferenciace (při $R = 5 - 1 = 4$)

$$P_D = \frac{4 \cdot 1{,}38}{16} = 0{,}345.$$

Výsledek tedy svědčí o spíše menší variabilitě názorů pracovníků. ■

Význam poměrového koeficientu diferenciace ovšem více vynikne až tehdy, pokud k hodnocení respondentům nabídneme i jinou, vícebodovou stupnici (škálu). Provedeme tedy stejný průzkum jako v příkladu 1.14 jen s tím rozdílem, že tentokrát zaměstnanci mohli hodnotit úroveň stravování nikoli na pětibodové stupnici, nýbrž na stupnici desetibodové (1 = nejlepší hodnocení, 10 = nejhorší hodnocení). Výsledné odpovědi jsou v tabulce 1.13.

Tab. 1.13 Výpočet poměrového koeficientu diferenciace – desetibodová stupnice

Škála odpovědí	x_i	Počet odpovědí n_i	$x_i n_i$	$(x_i - \bar{x})^2 n_i$
1 = nejlepší	1	3	3	61,83
2	2	3	6	37,59
3	3	4	12	25,81
4	4	6	24	14,23
5	5	8	40	2,33
6	6	9	54	1,90
7	7	7	49	14,92
8	8	4	32	24,21
9	9	3	27	35,91
10 = nejhorší	10	3	30	59,67
Součet	x	50	277	278,42

Průměr odpovědí je

$$\bar{x} = \frac{\sum_{j=1}^{k} x_i n_i}{n} = \frac{277}{50} = 5{,}54,$$

rozptyl

$$s_x^2 = \frac{\sum_{j=1}^{k} (x_i - \bar{x})^2 n_i}{n} = \frac{278{,}42}{50} = 5{,}568\,4.$$

Odtud poměrový koeficient diferenciace

$$P_D = \frac{4 s_x^2}{R^2} = \frac{4 \cdot 5{,}568\,4}{(10-1)^2} = \frac{22{,}273\,6}{81} = 0{,}275\,0.$$

Výsledek svědčí pro spíše nižší variabilitu odpovědí na položenou otázku. ■

Hodnoty pořadového znaku však často bývají vyjádřeny pouze slovně, např. vzdělání můžeme označit odlišnými kategoriemi, tj. jako základní, středoškolské, vysokoškolské apod. (neoznačujeme je pořadovými čísly). V tom případě nelze spočítat průměr, a tudíž ani rozptyl a z něj odvozené charakteristiky, a tak ani variační rozpětí.

V takovém případě se k výpočtu využívají kumulativní četnosti jednotlivých kategorií pořadového znaku. Lze použít **ordinální rozptyl** ve tvaru

$$2 \cdot \sum_{i=1}^{k-1} \left[c_i (1 - c_i) \right],$$

kde c_i jsou kumulativní relativní četnosti a k je počet kategorií.

Ordinální rozptyl nabývá hodnot z intervalu $\langle 0; (k-1)/2 \rangle$. Svého maxima dosahuje právě tehdy, když u 50 % objektů nabývá sledovaná proměnná hodnoty x_1 a u zbylých 50 % objektů hodnoty x_k za předpokladu, že proměnná obsahuje odpovědi respondentů na uzavřenou otázku s více než dvěma nabídkami a že bereme v úvahu četnosti všech možných kategorií, tj. četnosti kategorií kromě první a poslední jsou nula. Kumulativní četnosti c_1 až c_{k-1} se rovnají 1/2, $c_k = 1$, takže potom hodnota ordinálního rozptylu je $2 \cdot (k-1) \cdot 1/2 \cdot 1/2 = (k-1)/2$.

Dělením ordinálního rozptylu touto maximální možnou hodnotou získáme **normalizovaný ordinální rozptyl**, který nabývá hodnot z intervalu od 0 do 1 a jehož výpočet provedeme podle vzorce

$$\frac{4 \cdot \sum_{i=1}^{k-1} \left[c_i (1 - c_i) \right]}{k-1}. \qquad (1.28)$$

Příklad 1.15

V nefinančním podniku pracuje celkem 50 zaměstnanců. Nejvyšší dosaženou úroveň jejich vzdělání přinášejí data uvedená v tabulce 1.14. Určíme z nich ordinální rozptyl a dále rovněž normalizovaný ordinální rozptyl. Pro možnost srovnání toho, jaké vlastnosti mají jednotlivé charakteristiky, jsme záměrně použili stejné četnosti jako v příkladu 1.14.

Tab. 1.14 Výpočet ordinálního rozptylu

Vzdělání	Počet odpovědí n_i	p_i	c_i	$c_i(1 - c_i)$
Základní	6	0,12	0,12	0,105 6
Středoškolské bez maturity	10	0,20	0,32	0,217 6
Středoškolské s maturitou	17	0,34	0,66	0,224 4
Bakalářské	11	0,22	0,88	0,105 6
Magisterské	6	0,12	1,00	0,000 0
Součet	50	1,00	×	0,653 2

Hodnota ordinálního rozptylu je $2 \cdot 0{,}653\,2 = 1{,}306\,4$ a normalizovaná hodnota

$$\frac{4 \cdot 0{,}653\,2}{5-1} = \frac{2{,}612\,8}{4} = 0{,}653\,2.$$

Přestože hodnoty četností byly stejné jako v tabulce 1.12, pomocí ordinálního rozptylu je v intervalu od 0 do 1 variabilita hodnocena jako větší. ■

Základem pro zkoumání **variability kategoriálního znaku** je zjištění koncentrace. Jako míru koncentrace pro nominální proměnné můžeme použít

a) relativní četnost modální kategorie, tj. $p_{Mo} = n_{Mo}/n$, kde n je celkový rozsah souboru),

b) součet druhých mocnin relativních četností, tj. $\sum_{i=1}^{k} p_i^2$, kde k je počet kategorií.

Budeme uvažovat dva extrémní případy. V prvním bude nenulová četnost jen u jedné kategorie, což značí, že ostatní kategorie nejsou ve sledovaném souboru zastoupeny. Pak

$$p_{Mo} = 1 \text{ a } \sum_{i=1}^{k} p_i^2 = 1.$$

Ve druhém případě budou kategorie v souboru rovnoměrně zastoupeny, takže $p_i = 1/k$ (pro $i = 1, 2, \ldots, k$).

Potom též $p_{Mo} = 1/k$ a $\sum_{i=1}^{k} p_i^2 = 1/k$, neboť

$$\sum_{i=1}^{k} p_i^2 = \sum_{i=1}^{k} \left(\frac{1}{k}\right)^2 = k\frac{1}{k^2} = \frac{1}{k}.$$

Jako míry variability slouží

a) **variační poměr** θ, který spočteme podle vzorce

$$\theta = 1 - p_{Mo} = 1 - n_{Mo}/n, \tag{1.29}$$

b) **nominální rozptyl** *nomvar* (Giniho koeficient), vyjadřující relativní počet všech dvojic, které nejsou ve stejné kategorii. Ten počítáme podle vzorce

$$nomvar = 1 - \sum_{i=1}^{k} p_i^2 = \sum_{i=1}^{k-1} \left[p_i (1 - p_i) \right] \tag{1.30}$$

nebo

$$nomvar = 1 - \sum_{i=1}^{k} \left(\frac{n_i}{n}\right)^2 = \frac{n^2 - \sum_{i=1}^{k} n_i^2}{n^2}, \tag{1.31}$$

c) **entropie** H, která je dána vzorcem

$$H = -\sum_{i=1}^{k} p_i \ln p_i, \tag{1.32}$$

je-li pro všechna i relativní četnost $p_i > 0$. V případě, že $p_i = 0$, se pro dané i položí příslušný sčítanec roven nule. Platí, že

$$p_{Mo} \in \langle 1/k; 1 \rangle \text{ a } \sum_{i=1}^{k} p_i^2 \in \langle 1/k; 1 \rangle,$$

tudíž $\theta \in \langle 0; (k-1)/k \rangle$ a rovněž $nomvar \in \langle 0; (k-1)/k \rangle$. Dále $H \in \langle 0; \ln k \rangle$. Pokud míra variability nabude hodnoty nula, hovoříme o nulovém rozptýlení, čili o úplné homogenitě. Platí, že čím vyšší je hodnota, která charakterizuje variabilitu, tím vyšší je heterogenita souboru (maximální variabilita nastává v případě, kdy jsou všechny kategorie rovnoměrně zastoupeny).

Míry variability můžeme vyjadřovat rovněž pomocí hodnot z intervalu od 0 do 1, čehož dosáhneme tím, že hodnotu vypočítanou podle některého z výše uvedených vzorců dělíme maximálně možnou hodnotou, tj. $(k-1)/k$, resp. $\ln k$. Používány jsou míry **normalizovaný nominální rozptyl** *norm. nomvar* $= k \cdot nomvar/(k-1)$ a dále **normalizovaná entropie** $H^* = H/\ln k$.

Příklad 1.16

Při marketingovém průzkumu mezi 50 náhodně vybranými potenciálními zájemci o služby cestovní kanceláře byla těmto respondentům kladena otázka, o jaký typ zájezdu by měli zájem, resp. jakou náplň pobytu by ve svých požadavcích preferovali. Jejich odpovědi jsou uvedeny v tabulce 1.15. Určíme variabilitu odpovědí potenciálních zákazníků podle vztahů (1.29) až (1.32).

Tab. 1.15 Výpočet nominálního rozptylu a entropie

Typ zájezdu	Počet odpovědí n_i	p_i	$p_i(1-p_i)$	$p_i \ln p_i$
Pobytový	16	0,32	0,217 6	−0,364 6
Pobytový s výlety	14	0,28	0,201 6	−0,356 4
Poznávací	10	0,20	0,160 0	−0,321 9
Poznávací s turistikou	6	0,12	0,105 6	−0,254 4
Zaměřený na sport	4	0,08	0,073 6	−0,202 1
Součet	50	1,00	0,758 4	−1,499 4

Hodnota variačního poměru je $\theta = 1 - 0{,}32 = 0{,}68$. Hodnota nominálního rozptylu je z uvedených dat 0,758 4 a normalizovaná hodnota

$$norm.\ nomvar = \frac{5 \cdot 0{,}758\,4}{5-1} = \frac{3{,}792}{4} = 0{,}948.$$

Hodnota entropie je v uvedeném příkladu 1,499 4 a normalizovaná hodnota

$$H^* = \frac{1{,}499\,4}{\ln 5} = \frac{1{,}499\,4}{1{,}609\,4} = 0{,}932.$$

Z výsledků vidíme, že v postojích zákazníků je patrná poměrně velká pestrost (variabilita) jejich požadavků.

■