evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Univerzita Hradec Králové
Ústav sociální práce

Statistické zpracování dat

Martin Kořínek

Gaudeamus 2014

Recenzovali:
Doc. PaedDr. Martina Maněnová, Ph.D.
PhDr. Martina Čierna

Publikace neprošla jazykovou úpravou.

Edice texty k sociální práci

Řada: Výzkumné metody v sociální práci - sv. 2

Studijní materiál vznikl za podpory projektu
Inovace studijních programů sociální politika a sociální práce na UHK s ohledem na potřeby trhu práce (CZ.1.07/2.2.00/28.0127)**,** který je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

ISBN 978-80-7435-399-4

Obsah

1 Úvodem

1.1 O tomto textu

Tato skripta jsou o statistice – o základech statistického myšlení a o některých vybraných metodách analýzy dat.

Proč vlastně patří statistika mezi předměty, které jsou na vysokých školách nejčastěji vyučovány? Málokdy scházejí statistické metody mezi těmi postupy, které studenti použijí při řešení problému v rámci své diplomové nebo dizertační práce. To není samoúčelné. Závěrečná práce na nejen vysoké škole je přípravou k vědecké činnosti a statistické metody jsou samozřejmou součástí vědecké metodologie.

Musíme si ale uvědomit, že výzkumné metody se uplatňují nejenom na akademických pracovištích, ale i ve všech profesích, kde (většinou vedoucí) pracovník buď potřebuje vyřešit problém nějakým systematickým způsobem, nebo musí kriticky vyhodnotit závěry výzkumu provedeného někým jiným pro svou vlastní potřebu.

Statistika hraje významnou roli ve světě ekonomiky a obchodu (a kéž by i v politickém rozhodování, v psychologickém výzkumu, pedagogice a sociologii). V některých z těchto disciplín vznikly dokonce specializace se zaměřením na statistické úlohy a problémy měření v daném oboru – např. ekonometrie, chemometrie, biometrie, psychometrie, edukometrie nebo klinická a obecná epidemiologie (a to nemluvíme o forenzní statistice).

Tento text není zaměřen na specifické koncepty a metody statistiky, jež byly vyvinuty v těchto oblastech. Spíše jde o vysvětlení (a tím pádem pokus studentů o pochopení) základu (filozofie), jenž je jednotlivým aplikacím víceméně společný.

Právě na tomto místě musím vzpomenout jenu z prvních přednášek ze statistiky, kdy nám, elévům, profesor sdělil: „Statistika není nic jiného než jiný pohled na svět, jedná se o směr filozofie.“. Význam tohoto je možno však pochopit až po zdárném prostudování statistiky jako multidisciplinárního oboru.

1.2 O statistice

Obvykle se slovem statistika často míní znázorňování číselných údajů přehlednou formou (Cyhelský, 1981). V této podobě se s ní setkáváme např. v masových médiích v souvislosti s volbami, průzkumy veřejného mínění nebo při zprávách o vývoji ekonomiky. Například Český statistický úřad předkládá, že míra nezaměstnanosti byla v minulém měsíci 10,6 %. Jak se došlo k těmto datům? Lze podobným číslům věřit? Nedošlo v průběhu získávání a zpracování dat k nějakému zkreslení?

V odborném textu se lékař dozví o novince, že pravidelné tělesné aktivity zlepšují kvalitu života a prodluží jej. Je možné závěrům věřit? Jiná informace v novinách hovoří o statistickém prokázání, že kouření způsobuje rakovinu. Co to znamená? Jak s tímto dále nakládat?

Nejen statistickým datům nemůžeme uniknout stejně tak nemůžeme přestat používat počítač. Data musí být interpretována s porozuměním. Počítačová a taktéž numerická a statistická gramotnost jako schopnost porozumět datům je důležitá pro každého (Sharma, 2005).

2 Statistika jako věda

2.1 Historie statistiky

2.1.1 Počátky popisné statistiky

Statistika je s historií spjata již odedávna. Důvody jsou zcela logické a zřejmé. Každý vládce chtěl mít přehled, jaký má majetek, kolik má k dispozici mužů do vojska či od kolika poddaných může vymáhat daně. A tak začaly první soupisy a přehledy.

Ty nejstarší písemné památky pocházející z oblasti Sumeru mají statistickou povahu – jednalo se o záznamy o časových intervalech, počtech osob a kusů domácího zvířectva a úrodě.

Nejen starověké říše, mezopotámskými městskými státy počínaje, byly finančně zcela závislé na úspěšném výběru daní, ať již se jednalo o pracovní povinnost obyvatelstva, dávky naturální či finanční. Pro tyto účely existovaly již propracované „statistické metodiky“ a z četných písemných památek je možno si udělat představu o systému tehdejší státní správy a nakládání s daty. Například v Egyptě je od roku 2850 př. n. l. pravidelně jednou za dva roky prováděn soupis dobytka, od roku 2000 př. n. l. se vybírá rovná daň z „hlavy“ (pochop – člověka), jejíž zavedení si vyžádalo aktualizované sčítání obyvatelstva (Žák, 2006).

Zvlášť zajímavý systém evidence měl starověký Řím. Census (sčítání) byl pravidelně prováděn v republikánském období; jednalo se původně o soupis nemovitého majetku, později i otroků a dobytka, na jehož základě byla vypočtena výše daně pro jednotlivé občany. Snad nejslavnějším se dnes jeví jeden z posledních pravidelných censů, provedený za císaře Augusta kolem přelomu letopočtu, který vešel do dějin jako časové určení narození Ježíše Krista.

Období raného středověku přineslo do Evropy všeobecný rozvrat. Nezapomeňme, že negramotnost byla běžným jevem i mezi panovníky. Určitá centra vzdělanosti udržovala církev, jež také byla schopna vést evidenci svého majetku a jeho změn a členové církevních řádů byli zaměstnáváni aristokracií jako „statistici“ (Kolektiv, 2011).

Středověkými evidencemi (s nadsázkou můžeme hovořit o statistických ročenkách) byly vrchnostenské urbáře, z nichž kromě evidence příjmů poddaných bylo možno vyčíst i rozsah pozemkového vlastnictví šlechty a církve, městské berní knihy a berní rejstříky, obsahující součet poplatníků spolu s jejich (nemovitým) majetkem a daněmi (výši

odvedených daní). Ve 14. století se v Evropě objevují první církevní matriky, důležitý zdroj (statistických) dat o přirozené měně obyvatelstva.

Šetření však nebyla vždy prováděna jen kvůli eráru, důvody pro statistické zjišťování byly mnohdy i zcela jiného – humánnějšího – rázu. Například za vlády císaře Rudolfa II. v roce 1583 vypukla v českých zemích epidemie moru. V jejím důsledku bylo zahájeno šetření o „zdraví populace“, které mělo zmapovat vznik a rozvoj zhoubných epidemií a umožnit přijímání důležitých protiopatření (Žák, 2006).

Je známo, že 16. století s sebou přineslo zcela odlišný pohled na svět a člověka, rozvoj filozofie a počátky moderních vědních disciplín (tudíž se jedná o období velkých společenských změn). V duchu popisné statistiky vyšlo v Benátkách roku 1562 jedno z prvních státovědných děl, O vládě a správě v různých královstvích od Francesca Sansoviny. Poznamenejme, že v roce 1589 použil Ital Girolamo Ghilini jako první termín statistika, v původním smyslu se jednalo o stav státu. Roku 1662 vyšla kniha Německý knížecí stát autora Veita Ludwiga von Seckendorff, která se během 17. a 18. století dočkala řady vydání.

V Anglii v 17. století vznikl odlišný okruh statistiky, takzvaná politická aritmetika (dnes mluvíme o demografii), která vycházela z údajů o narozeních a úmrtích a pokoušela se na jejich základě zkoumat vývoj počtu obyvatelstva v delších časových údobích (jednalo se tedy o první analýzy časových řad). Zakladatelem této disciplíny byl John Graunt (1620-1674), který jako první považoval demografické jevy za jevy hromadné. Prozkoumal poměr mezi počtem mužů a žen (v dané populaci) a stabilní poměr mezi počtem narozených chlapců a dívek. Sestavil rovněž úmrtnostní tabulky na základě zkoumání vymírání jednotlivých věkových skupin. Dalším reprezentantem anglické školy byl Edmund Halley (1656-1742), který na konci 17. století při konstrukci úmrtnostních tabulek propojil statistiku s počtem pravděpodobnosti a použil geometrický náhled na pravděpodobnostní úlohu (Halley je však nyní znám spíš jako astronom – viz Halleyova kometa) (Žák, 2006).

Další zajímavou osobností evropské statistiky byl belgický matematik, statistik (aopět astronom) Adolphe Lambert Quételet (1796-1874), který vypracoval zásady moderních sčítání lidu – poprvé byly uplatněny při belgickém sčítání lidu v roce 1846. Na základě velkého souboru dat definoval rozměry „průměrného člověka“ a odchylky jednotlivců od

tohoto středu – je tedy duchovním otcem pojmů jako průměr, střední hodnota, rozptyl a rozdělení.

Přínosem pro teorii chyb byl německý matematik Carl Friedrich Gauss (1777-1855), který zásadně přispěl k formulování normálního rozložení pravděpodobnosti (tedy rozdělení, které se jako nit táhne celou moderní statistickou). Po francouzském matematikovi a fyzikovi Simeónu Denisovi Poissonovi bylo nazváno rozdělení, které je zásadní pro nízkou pravděpodobnost jevu při značném rozsahu výběrového souboru. Neoddiskutovatelný je přínos Charlese Pearsona (1857-1936) především pro zpracování regresní a korelační analýzy, testů dobré shody a momentové charakteristiky (Kolektiv, 2011).

Časová i finanční náročnost takových šetření vedla k diskusím, zdali je opravdu třeba zkoumat celou populaci, či nebylo dostačující vybrat pouze její reprezentativní vzorek. Na základě této myšlenky se začátkem 20. století zrodila matematická statistika, disciplína, jejímž charakteristickým motivem je hledání metod, které by umožnily vytvoření závěrů o celku na základě výběru. Matematická statistika (tedy samostatný obor matematiky) si tak během prvních let 20. století vytvořila vlastní metodologii jako je analýza rozptylu, korelační počet a ověřování hypotéz.

2.1.2 Vznik teorie pravděpodobnosti, matematické statistiky

Odhlédneme-li od prací indických a čínských matematiků, kteří se řešením kombinatorických úloh zabývali již ve starověkém období, můžeme položit základy této „statistické“ disciplíny v Evropě do 16. století. Ozřejmění pozdního zájmu matematiků o náhodu může být toto (Kolektiv, 2011):

- Nebyla definována žádná souvislost mezi matematikou a náhodnými jevy: matematika je obor, v povaze kterého se nic náhodného nevyskytuje a náhoda je nahlížena buď jako projev vůle božstva, nebo je chápána jako neznalost všech příčin a kauzálních vazeb (jev, který by mohl být zevrubným zkoumáním zcela vyloučen).
- Dalším důvodem mohl být fakt, že matematické zkoumání náhodných jevů jednoduše dříve nikdo nepotřeboval – k jistě velkému potěšení falešných hráčů

– a nutnost jejich detailního popisu vyvřela až v souvislosti s mohutným rozvojem (obchodu, demografie, astronomie a fyziky).

Za „základní materiál“ teorie pravděpodobnosti lze považovat tyto dva typy problémů (původně z oblasti hazardních her):

1. První typ je v podstatě kombinatorický, jedná se o otázky, kolika způsoby může padnout jistý počet ok při házení určitým počtem hracích kostek – podobné úlohy se objevují v teorii pravděpodobnosti a jejích aplikacích i dnes.
2. Druhý typ problémů se týká úlohy o rozdělení sázky: dva hráči hrají sérii her o určitou částku s tím, že tuto sumu získá ten, který jako první dosáhne předem domluveného počtu vítězství. Série je však předčasně ukončena a hráči si chtějí částku spravedlivě rozdělit v závislosti na tom, kolik her který z nich vyhrál.

K významnému rozvoji statistiky dochází ve třicátých letech 20. století. K rozvoji matematické statistiky přispěli také ruští matematici: P. L. Čebyšev (1821-1894, definoval centrální limitní větu)), A. M. Ljapunov (1887-1918) a A. A. Markov (1856-1922), dodnes používané Markovovy řetězce. Těžiště rozvoje statistiky se do značné míry přesunulo do anglo-americké oblasti. Donald Alyner Fischer (1890-1926) vytvořil, resp. podílel se na vypracování řady statistických metod. Je považován za zakladatele teorie navrhování experimentů v biologickém a zemědělském výzkumu (Žák, 2006).

Avšak opravdovým zakladatelem moderní matematické statistiky jako disciplíny se stal britský biolog Ronald Fisher (mimo jiné Fisherův index, Fisherův f-test). Napsal učebnice Statistical Methods for Research Workers (1925) a The Design of Experiments (1935) a objevil či podstatně prohloubil celou řadu klíčových metod a pojmů jako jsou analýza rozptylu, metoda maximální věrohodnosti, diskriminační analýza, informace a plánování experimentů.

Zhruba v polovině 20. století se matematická statistika stala samostatným oborem (na pomezí čisté a aplikované matematiky). Teorii statistických dat, experimentálního designu a výběru obohatili například William Gemmell Cochran, Leslie Kish, C. R. Rao a Donald Rubin. Frank Hampel, Peter J. Huber, John Tukey a další badatelé rozvíjeli robustní metody odhadů a exaktní statistické metody, jež jsou málo citlivé k extrémním hodnotám v datovém souboru a platné i pro malé rozsahy výběrů. V návaznosti na klasické Thurstonovy a Guttmanovy práce vznikla teorie a metodologie statistického škálování, k níž přispěli například Clyde H. Coombs, Roger N. Shepard, Joseph B. Kruskal a rovněž

Lee J. Cronbach teorií reliability škál. Analýzu kategorizovaných dat obohatil Leo A. Goodman metodologií logaritmicko-lineárních modelů (Kolektiv, 2011).

Náměty z oblasti sociálních věd přineslo například rozpracování metod z okruhu faktorové analýzy, jejíž základy položili psychologové Charles Spearman a Raymond Cattell a později rozpracovali Karl G. Jöreskog a Dag Sörbom, dále teorie latentních tříd, kterou navrhl sociolog Paul F. Lazarsfeld, či teorie rozhodovacích stromů, u jejíchž kořenů stojí sociologové William A. Belson a James N. Morgan. V oblasti ekonomie a ekonometrie působili například Harold Hotelling (kanonická korelační analýza) či James Tobin (regrese cenzurovaných dat). Významná pro ekonomii je rovněž statistická analýza časových řad, k níž přispěli například George Box, Gwilym Jenkins (známá Boxova-Jenkinsova metoda) (Kořínek, 1992) či TimBollerslev, a analýza přežití, kterou rozvinuli mimo jiné Edward L. Kaplan, Paul Meier a David Cox. Použití metod matematické statistiky v průmyslu je spojeno se jmény Waltera A. Shewharta a W. EdwardseDeminga, kteří se stali zakladateli metodologie a hnutí řízení kvality (Meloun, 2006).

Nástup počítačů v polovině a především pak koncem 20. století se projevil usnadněním rozsáhlých výpočtůa umožnil i vznik takzvaných výpočetně náročných metod ve statistice, často založených na myšlence algoritmu Monte Carlo, tedy opakovaném generování náhodných jedinců ze zkoumané populace. To také umožnilo rychlý nástup bayesovských metod, které sice již dříve propagovali například Harold Jeffreys nebo Edwin Thompson Jaynes, ale které byly zhruba až do sedmdesátých let limitovány (nedostatečnou výpočetní silou počítačů a i nepřítomností numerické metodologie) (Žváček, 2007). Dalším aspektem rozšíření počítačů byl rozvoj programového vybavení pro matematickou statistiku (Kořínek, 2010). Zejména nástup univerzálních programových balíků (statictical package) od 60. let 20. století umožnil použití složitějších statistických metod pro široký okruh uživatelů.

2.1.3 Statistika v Čechách

Za zřejmě nejstarší dochovaný soupis na našem území je považován soupis majetku litoměřického kostela z roku 1058, který je součástí zakládací listiny knížete Spytihněva II. Ale důvody pro statistické zjišťování byly mnohdy i zcela jiného – humánnějšího – rázu. Například za vlády císaře Rudolfa II. v roce 1583 vypukla v českých zemích epidemie moru. V jejím důsledku bylo zahájeno šetření o „zdraví populace“, které mělo zmapovat

vznik a rozvoj zhoubných epidemií a umožnit přijímání včasných protiopatření (Český statistický úřad, 2011).

Důležité datum je 13. října 1753, kdy byl vydán patent císařovny Marie Terezie o každoročním sčítání lidu. Zdokonalení evidence obyvatel souviselo s rozsáhlou reformní činností Marie Terezie. K provedení četných reforem bylo nutné získat objektivní informace o obyvatelstvu – již tehdejší národohospodáři označovali snahu řídit někoho, aniž bychom o něm měli dostatečné údaje, za nesmyslnou a pošetilou (Žák, 2006).

Novou kapitolu v historii sčítání obyvatelstva v habsburské monarchii zahájilo sčítání provedené v roce 1754. Poprvé se konalo současně a jednotně na celém vymezeném území. Soupis mělo nejprve provést duchovenstvo podle farností, posléze bylo rozhodnuto, že paralelně se uskuteční i sčítání zajišťované vrchností a jeho obsah bude rozšířen o soupis domů a o hospodářskou charakteristiku majitele domu. Za vlády Marie Terezie došlo také k reformě evidence narozených a zemřelých. V této souvislosti byla zavedena i první jednoduchá statistická klasifikace příčin úmrtí (rozlišovaly se příčiny smrti „přirozené“ a „násilné“).

Sčítání konané v roce 1754 bylo důležitým počinem. Soupisy z 60. let 18. století, i když umožňují poprvé stanovit koncentraci osídlení podle krajů a poskytují i některé další informace o sociálním složení obyvatel, byly vcelku neúspěšné. Stoupající míra obav ze vzrůstu daní, odpor šlechty proti centralistickým snahám dvora, k němuž se postupně přidávala i církevní hierarchie, vedly k četným zkreslením. Došlo proto k další reformě. Stát neměl dostatek úředníků a soupis tedy nemohl zajistit vlastními silami. Proto bylo povoláno vojsko. Tím se od základu nejen změnila organizace soupisů, ale pronikavě i jejich obsah. Jednostranné zaměření soupisů velmi brzy přestalo vyhovovat (Český statistický úřad, 2011).

V roce 1777 byl vydán nový konskripční patent, který se s mírnými změnami a odchylkami stal základem soupisů až do roku 1851. Bylo zachyceno veškeré přítomné obyvatelstvo. Od 80. let 18. století byly na panstvích a městech (později v obcích) založeny tzv. populační knihy, v nichž byla evidována zvlášť každá rodina se všemi členy domácnosti; případné změny (úmrtí, narození atd.) byly do těchto knih vpisovány na základě ohlašovací povinnosti hlavy rodiny.

Počátky samostatného shromažďování údajů jsou u nás spojeny se jménem Josefa Antonína rytíře Rieggera (1742 – 1795). Založil organizovanou statistickou službu a byl

bezesporu prvním kvalifikovaným statistikem u nás (Cyhelský, 1981). Získal podporu císaře Josefa II a tak mohl koncipovat a organizovat statistická šetření a vytvářet pro ně metodiky. V roce 1787 začal vydávat sebraný materiál ve známých „Materialienzuraltenundneueren Statistik von Böhmen“.

Datum 30. listopadu 1856 je považován za počátek státem organizované statistiky v Českých zemích. Toho dne se uksutečnilo první zasedání Ústředního výboru pro statistiku polního a lesního hospodářství Čech jako zvláštního nově utvořeného orgánu c. k. Vlasteneckohospodářské společnosti (Český statistický úřad, 2011).

Další významnou etapu v novodobých dějinách sčítání lidu v Rakousku zahájil zákon přijatý v roce 1869. Na jeho základě bylo na začátku roku 1870 provedeno sčítání lidu, které zachytilo stav ke dni 31. 12. 1869. Zákon dále určoval, že následná statistická šetření mají zachytit stav obyvatelstva v desetiletých obdobích a to vždy k poslednímu prosinci roku končícím na 0. Samotné provedení šetření zajišťovaly obce, sčítací jednotkou se stala domácnost. Dotazníky byly buď vyplňovány majitelem domu, nebo „sčítacími komisaři“ na základě ústního sdělení sčítaných osob. Jednalo se tedy o první sčítání lidu v moderním pojetí, které vytvořilo podmínky pro porovnávání základních demografických údajů od tohoto roku až po současnost. Až díky pravidelným sčítáním můžeme získat přesný obraz o vývoji počtu obyvatel na našem území, jinak se musíme spokojit s odhady.

Je bezesporu důležité, že s výsledky sčítání 1869 byla (poprvé v širším měřítku) seznámena veřejnost a to v šestidílné publikaci. Z dnešního pohledu je tato skutečnost připadá samozřejmá, ale až do 40. let 19. století byly výsledky soupisů obyvatel pokládány stejně jako výsledky dalších statistických šetření za tajné či důvěrné (v roce 1829 byl panovníkovi předložen VersucheinerDarstellung der österreichischen Monarchie in statistischenTafeln 1828, vydaný ve100 výtiscích. Pouze šest z nich, určených pro nejbližší okolí císaře, obsahovalo i údaje o vojsku, státním rozpočtu a veškeré přehledy podle jednotlivých zemí) (Kolektiv, 2011).

Poznamenejme, že populace 18. a 19. stol. byla populací mladou. Na vesnici byl mírně vyšší podíl dětí a starších osob, zatímco ve městech se imigrací zvyšovalo množství obyvatel v reprodukčním věku. Se zlepšujícími se úmrtními poměry se postupně zvyšoval počet osob dožívajících se dospělosti a posléze i vyššího věku; pro tehdejší poměry je však stále charakteristické, že se vyššího věku dožívalo zpravidla více mužů než žen. Statistika 19. století nashromáždila velké množství údajů o rozdělení obyvatelstva do skupin podle

věku a rodinného stavu, podle mateřské řeči i profese a postavení v povolání. Dne 30. června 1870 se do Prahy dostavili účastníci první schůze „Obecní komise statistické královského hlavního města Prahy“, a zahájili tak po několikaletém úsilí o zřízení statistického úřadu systematickou statistickou činnost. Dne 6. března 1897 byl pak zřízen Zemský statistický úřad Království českého, který se stal prvním skutečně statistickým úřadem na území dnešní České republiky. Poprvé byla soustředěna všechna statistická pracoviště, která až do té doby působila v rámci různých ministerstev a dalších institucí.

V roce 1909 vyšla první „Statistická příručka království Českého“, další pak následovala v roce 1913. Zemský statistický úřad v nich podak veřejnosti výbor z důležitých statistických dat o Čechách, často s několikaletou retrospektivou a v porovnání s adekvátními údaji z Moravy, Slezska a celé monarchie. Příručky zahrnují velký okruh údajů, které jsou rozděleny do následujících 18 oddílů: výměra, rozdělení a obyvatelstvo; samospráva; volby; zdravotnictví a ústavy humanitní; chudinství; policie; zprostředkování práce; vojsko a četnictvo; kultura; školství a jiné ústavy vzdělávací; zemědělství; hornictví a hutnictví; živnosti, průmysl a obchod; doprava; úvěr; pojištění; soudnictví; finance. Koncem roku 1914 byl přijat „statistický zákon pro Moravu“, který poprvé definoval zpravodajskou povinnost. Týkal se pouze obcí a okresních silničních výborů na území Moravy. Jeho význam spočíval především v tom, že poprvé na části území dnešní ČR vznikla zpravodajská povinnost vůči statistickému úřadu – daný subjekt musel ze zákona povinně poskytnout požadovaná data (Čermák, 1968).

Devatenácté století bylo dobou prudkého rozvoje průmyslu, což pochopitelně kladlo daleko větší nároky na rozsah a kvalitu i statistického zjišťování a zpracování statistických dat. Lze říci, že právě tehdy se začala rodit současná podoba statistiky, která je z velké části zjišťováním (makro)ekonomických ukazatelů. Ostatně, o prudkém rozmachu průmyslové výroby včetně těžkého průmyslu, zejména železniční sítě, nejlépe vypovídají obsáhlé tabulky „Statistické příručky království Českého“.

Tři měsíce po vzniku samostatného Československa - přesně 28. ledna 1919 - přijalo Revoluční národní shromáždění zákon č. 49 o organizaci statistické služby. Principy tohoto zákona již tehdy odpovídaly základům, na kterých je organizována současná statistická služba ČR. V roce 1919 byl založen Státní úřad statistický (SÚS) jako nový orgán pověřený celostátními statistickými šetřeními, mezi něž patřilo i sčítání lidu jako jedno z nejdůležitějších. Úřad se v období mezi světovými válkami rozvíjel, zdokonaloval a rozšiřoval svoji činnost. K tomu přispělo i úzké spojení se statistickou teorií. Ve 20. a 30.

letech 20. století byla téměř polovina kapacity statistického úřadu věnována vědecké a teoretické činnosti (Český statistický úřad, 2011).

V knize „Československá statistika v prvním desetiletí republiky“ z roku 1928 byla funkce státní statistiky charakterizována takto: „Účelem statistické služby je podávati obraz o stavu a vývoji poměrů v celém státě, jejichž konečným cílem jest dosažení hospodářského blahobytu, mravnosti, zdatnosti a zdraví všeho obyvatelstva. Tyto snahy nemohou však býti řízeny náhodou nebo tradicionalismem a pouhým instinktem, nýbrž uvědoměle podle plánu přesně a soustavně, tedy vědecky. To však předpokládá právě znalost všech skutečností a poměrů ve státě. Zjistit tato fakta soustavně, vystihnout, co je na nich typického, jaké vztahy a vzájemné souvislosti příčinné vykazují, jakými zákonitostmi se řídí jejich vývoj, to je právě úkolem statistické služby.“

Již při vzniku SÚS bylo zřejmé, že pro rozsáhlá statistická šetření, například pro sčítání lidu, bude nutné zabezpečit potřebné strojní zařízení. Dne 1. dubna 1920 najal SÚS na zkoušku 13 děrovacích strojů a 4 třídicí stoje s počítadly od firmy Powers Accounting Machine. První prací na těchto strojích bylo zpracování materiálů o přirozené měně obyvatelstva v průběhu válečných let. Strojový park SÚS se pak rychle rozrůstal. V roce 1929 měl SÚS 6 automatických děrovacích strojů, 1 ruční děrovačku, 14 strojů třídících a 4 tabelátory - pracovalo v ní 68 zaměstnanců. V roce 1939 bylo takto zpracováváno již 19 druhů různých statistik. (V roce 1928 bylo například naděrováno 5 875 799 štítků, v roce 1939 již 6 334 816 štítků. Třídícími stroji prošlo v roce 1939 celkem 327 859 000 štítků a stroji tabelačními 19 896 000 štítků.) (Český statistický úřad, 2011)

V období 2. světové války se práce statistiky v Čechách a na Moravě omezila a odpovídala válečným podmínkám i postavení našeho území. Předseda Státního úřadu statistického dr. Jan Auerhan byl již koncem března 1939 donucen odejít do trvalé výslužby (především kvůli jeho pracím o menšinové politice). Dr. Jan Auerhan byl 6. 6. 1942 zatčen gestapem a 9. 6. 1942 zastřelen. Perzekuována byla řada dalších pracovníků úřadu. Mnozí z nich byli popraveni, jiní zemřeli v nacistických věznicích a koncentračních táborech.

Bezprostředně po skončení 2. světové války byl zřízen Státní úřad statistický s celostátní působností, s cílem obnovit poměrně věhlasnou předválečnou úroveň československé statistiky. Druhá světová válka znamenala značný zásah do národnostních struktur y českých zemí, zejména v důsledku odsunu německého obyvatelstva.

Po roce 1948 se československá statistika (především v ekonomické oblasti) zaměřovala zejména na úkoly národohospodářské evidence a kontrolu plnění plánu. (poznamenejme, že v té době se ČSÚ přezdívalo Ministerstvo informací).

V roce 1989 (po pádu komunistického režimu) se obnovily předpoklady pro budování objektivní, nestranné a nestranické státní statistické služby. K 1. 1. 1993 se vznikem ČR převzal ČSÚ všechny kompetence národního statistického úřadu. Jeho úkoly a postavení, stejně jako zásady a úkoly fungování státní statistické služby v ČR, upravil zákon č. 89/1995 Sb., o státní statistické službě, který byl novelizován k 1. 1. 2001, ve znění pozdějších předpisů. (Český statistický úřad, 2011)

Dlouhá desetiletí musel ČSÚ pracovat v provizorním sídle v Karlíně (vedle hotelu Olympia), v budovách, jež původně sloužily jako lazaret. Otázku nové budovy paradoxně vyřešila ničivá povodeň ze srpna 2002, po níž již nebylo o potřebě nového sídla úřadu sporu. Na jaře 2004 se tak jeho pracovníci mohli přesunout do nového moderního sídla české statistiky v Praze 10, poblíž stanice metra Skalka.

2.2 Vymezení statistiky

2.2.1 Statistika jako věda

V úvodu jsme si řekli, že statistika je naukou, jak získat informace (nejen z numerických) dat. Statistika nám pomáhá při přípravě a provedení výzkumu a při vyhodnocení získaných výsledků. Poskytuje prostředky a koncepty, které umožňují pracovat s výsledky tak, abychom porozuměli určitému problému.

Statistické šetření, praxe, lze rozdělit dle mnoha odborníků na několik částí – můžeme mnohdy využíti i definice poskytující moderní metody projektového řízení. Nicméně my si zde uvedeme členění dle významného didaktika statistiky Davida S. Moora (1997), který praxi statistiky dělí na tři části: získávání dat, analýzu dat a statistické usuzování (Chráska, 2009).

1. Získávání dat zahrnuje metody pro sběr dat, jež zodpoví předem danou otázku. Základní přístupy k výběru měřených objektů, k návrhu experimentů a k validizaci instrumentů pro získávání dat jsou významným příspěvkem statistiky. Touto problematikou se budeme věnovat v následujících odstavcích.

2. Analýza dat představuje organizaci dat a popis dat užitím grafů, numerických souhrnů a dalších matematicky propracovaných prostředků. Někdy se této oblasti říká popisná (deskriptivní) statistika. Poznamenejme, že počítačová revoluce vrátila popisnou (a explorační analýzu) dat do centra statistické praxe.
3. Statistické usuzování (inference) jde za sama data a usiluje o získání závěrů o širším univerzu jevů. Neprovádí jenom závěry, ale dodává k nim i zhodnocení, jak jsou tyto závěry spolehlivé. K tomu používá pravděpodobnostní pojmy. Tomuto způsobu práce s daty se říká také inferenční statistika. Metody této části patří k matematicky nejnáročnějším z celé statistiky. (Význam statistického testování hypotéz nebo používání intervalů spolehlivosti atd. je nutno posuzovat v závislosti na oprávněnosti uvedených metod, a ne podle jejich matematické rozsáhlosti.) Poznamenejme, že ne vždy tuto část – tedy statistické usuzování – využijeme. Mnohdy statistický výzkum končí bodem 2 (popisem stávajícího stavu).

2.2.2 Význam slova statistika

Slovo statistika můžeme používat nejméně ve třech (čtvrtý viz dále) významech (Hendl, 2005):

1. **Číselné údaje o jevech** – tabulky (naměřená, pozorovaná data statisticky setříděna).
2. **Praktická činnost** spočívající ve sběru, zpracování a vyhodnocování dat (vlastní proces).
3. **Teoretická disciplína**, která se zabývá metodami vyhodnocení (hromadných) jevů - složitá matematika, kterou se zabývají profesionální statistici.

Stručně bychom mohli říci, že statistika je způsob shromažďování dat, práce s daty a jejich kvantitativní vyhodnocení, tedy interpretace. Používá k tomu především metodu pozorování (měření) a popisu určité vybrané části reality (např. vybrané skupiny zdravotnických zařízení) s případným následným zobecněním (dle Moora třetí část) těchto pozorování na „celou“ realitu (všechna zdravotnická zařízení na sledovaném území).

Všimli jsme si, že jsme v minulém odstavci použili spojení „a jejich kvantitativní vyhodnocení“? Co si pod tímto máme představit? Je to jednoduché, například průměr je

kvantitativní vyhodnocení – samotný průměr, jedno jediné číslo, nám dává jistou představu o zkoumaném jevu. A také maximální či minimální hodnota – také nám přece dávají informaci o zkoumaném jevu.

A právě jsme se dostali ke čtvrtému významu slova statistika

4. Statistikou chápeme i **jedno (vypočítané) číslo**, které charakterizuje náš soubor, zkoumaný jev. (Tomuto číslu říkáme nejen statistika, ale také i v obecném pojetí charakteristika.) (Cyhelský, 1981)

A nyní jsme již jistě pochopili, co jsme mysleli tvrzením „statistika je naukou, jak získat informace (nejen z numerických) dat.“ Ano, právě statistika v posledním významu – jedno číslo, které popisuje náš zkoumaný soubor – je ta nová informace. Statistika jako věda tedy dokáže z dat získat další informace – z naměřených údajů statistika vypočítá průměr a to je ta zcela nová informace.

2.2.3 Statistické disciplíny

V této kapitole se pokusíme celou statistiku jako vědu rozčlenit do několika disciplín. Toto členění není samoúčelné, ale pomůže nám například při řešení konkrétního problému, příkladu, určit cestu, kterou se máme pustit – na základě znalostí o jednotlivých typech příkladů (o jednotlivých statistických disciplínách) si správně vybereme skupinu metod, kterými zadaný problém vyřešíme. Druhým důvodem, proč si zde výčet statistických disciplín, uvádíme je, že pochopíme, jak rozsáhlá statistiku je a co všechno s ní dokážeme řešit.

Mezi základní statistické šetření, které provádíme vždy, patří statistický popis získaných dat (Čermák, 1968). K tomu se používá **popisná** neboli **deskriptivní statistika**. Jejími výsledky jsou statistické tabulky a statistické grafy. Dalším výstupem jsou statistiky, tedy vypočítaná čísla, která charakterizují daný soubor pozorování – charakteristiky polohy a variability, případně míry šikmosti a špičatosti (Cyhelský 1981).

Míry polohy informují o „těžišti“ souboru, o jeho středu, kolem kterého jsou jednotlivá pozorování rozptýlena (Cyhelský, 1980). Konkrétně, mezi míry polohy patří všeobecně známý průměr. A velikost rozptýlení kolem tohoto středu nám zase předkládají míry variability, mezi které patří rozptyl či směrodatná odchylka. Míry šikmosti a špičatosti pak

blíže popisují, jak jsou data, pozorování, uspořádána (zdali jsou rozmístěna symetricky – napravo a nalevo od středu).

Je nutné upozornit na to, že deskriptivní statistika popisuje **jednorozměrný** soubor. Co to znamená? Vezměme si jednoduchý příklad: Sledujeme u jednotlivých studentů vybrané vysoké školy následující vlastnosti:

Výška, váha, věk, pohlaví, studovaný obor, absolvovaná střední škola, známka z matematiky na střední škole, velikost bydliště (vesnice, městys, město, velké město), počet sourozenců, vzdělání rodičů.

Jedná se o **jednorozměrný** či **vícerozměrný soubor**? Nejprve si musíme říci, že pojem rozměr se vztahuje k počtu sledovaných vlastností a především k tomu, co vlastně statisticky zjišťujeme. My sice sledujeme celou řadu (přesně 11 znaků), ale prozatím nás zajímají jednotlivé vlastnosti osamoceně. Ptáme – se jaká je průměrná výška studentů, jaký je poměr dívek a chlapců, jaká je nejčastěji absolvovaná střední škola, jaký je největší počet sourozenců atd. Takže studujeme dané vlastnosti nezávisle na ostatních. Proto deskriptivní statistika popisuje jednorozměrné statistické soubory. Na rozdíl od dalších disciplín, viz dále.

Poznamenejme, že právě popisná statistika je náplní tohoto textu.

Srovnávací analýza (**analýza rozdílnosti**) řeší další typ úlohy. Máme například údaje o výrobě kyseliny sírové a to ve dvou lokalitách (například USA a Rusko) a ve dvou časových obdobích (za rok 2000 a 2010). Jistě nás napadne otázka, ve které lokalitě byl nárůst (či pokles) výroby kyseliny sírové větší. Již ze základní školy víme, že tuto úlohu může vyřešit prostým rozdílem – odečteme hodnoty roku 2010 od hodnot z roku 2000 a to pro obě lokality zvlášť – a větší číslo ukáže na větší nárůst. Jenže nás taky napadne, že ke srovnání můžeme vedle rozdílu použíti i podíl – hodnoty z roku 2010 podělíme hodnotou roku 2000 (opět u obou lokalit zvlášť) – a opět větší číslo (v tomto případě relativní) nám ukáže na větší nárůst. Tomuto relativnímu číslu se říká index (Swoboda, 1977).

Jenže často nastávají situace, kdy rozdíl poukazuje na vítěze A (v našem případě třeba USA), kdežto podíl naopak na vítěze B (tedy Rusko). A teď otázka: která statistika – rozdíl či index – je přesnější? Kterou by měl statistik používat? Odpověď je nasnadě – statistiku musí vždy při srovnávací analýze uvádět obě čísla – jak absolutní tak i relativní. Pokud uvede pouze jedno, dopouští se zásadní chyby (v horším případě statistik vědomě správnými čísly manipuluje tak, aby výsledek vypadal „dobře“).

Abychom se nepletli, srovnávací analýza je mnohem složitější vědou a nezabývá se pouze jednoduchými měrami (rozdíly a indexy), ale formuluje metodiku jak správně porovnávat složené ukazatele. Příkladem budiž cenové indexy, které srovnávají hodnotu produkce současného období s produkcí v základním období, přičemž obě produkce jsou ohodnoceny (váženy) cenami (a opět v současném či základním období). Ale aby porovnávání bylo smysluplné a dávalo logiku, musí se použít správná kombinace vah. Pro snadnější pochopení uvedeme „výpočet"Paascheho indexu: srovnává hodnotu produkce běžného období oceněnou cenami běžného období (současné náklady na současný spotřební koš) s hodnotou této produkce (tedy běžného období) ohodnocenou cenami základního období (původními náklady na současný koš). Vedle tohoto indexu existuje celá řada dalších, například Fischerův, Montgomeryho[1] a Laspeyresův („opak" Paascheho, protože porovnává současné náklady na původní koš a původní náklady na původní koš).

Srovnávací analýza šla dokonce dál, pokusila se zkonstruovat syntetický ukazatel – ukazatel, který by přinesl jedno jediné číslo, které by v sobě zahrnovalo jak absolutní tak i relativní váhu. Teoretikové představili následující statistiku: index umocněn na rozdíl. I přes mnohé odborné statě s však tento ukazatel v praxi neujal. Nicméně uvádíme zde tento příklad, abychom pochopili, k čemu analýza rozdílnosti slouží a že se jedná o disciplínu značně rozsáhlou a nikoli jak by se na první pohled mohlo zdát triviální.

Již ze samotného názvu **analýza časových řad** jistě poznáme, k čemu nám tato oblast statistiky poslouží. Máme hodnoty jednoho ukazatele za delší časové období, například údaje o tržbách za prodej sportovního zimního oblečení v jednotlivých týdnech za posledních pět let. Jistě nás bude zajímat, jak se tyto tržby vyvíjely a jak se zřejmě v nejbližší době vyvíjet budou (abychom odhadli, kolik zimního oblečení máme mít v zásobě pro nastávající zimní sezónu). Analýza časových řad nám pomůže sestrojit model chování tohoto ukazatele a na základě tohoto modelu můžeme budoucí hodnoty vypočítat, odhadnout (Kozák, 1994).

Na tomto místě si pouze řekněme, že mezi základní metody analýzy časových řad patří dekompoziční model, který časovou řadu rozkládá na několik (přesně čtyři) složek (Cipra, 1986). První složkou je trend neboli dlouhodobá tendence časové řady. Trend může být rostoucí, klesající či konstantní (neměnný) a modelujme jej nejčastěji jednoduchými

[1]Viz CYHELSKÝ L., CYHELSKÝ P:: *Ke vzniku Montgomeryova cenového indexu před sedmdesáti lety*. InBulletin České statistické společnosti 6/2007 (ss. 509-511) ISSN 1210-8022.

křivkami (přímkou, parabolou, hyperbolou, exponenciálou, logaritmickou). Druhou složku časové řady přestavuje sezónní výkyvy (proto se nazývá sezónní složka) a modeluje změny v rámci jedné časové periody (nejčastěji v rámci jednoho roku, například kvartální či měsíční změny). Tato složka se modeluje nejčastěji pomocí indexů (porovnává výkyv daného sezónního období s průměrem). Třetí složka – cyklická – dekompozičního modelu časové složky se v mnohých případech nemodeluje, protože je podobná složce sezónní, ale s delší periodou než základní období (např. při ročních sezónních složkách se v ekonomii používá délka cyklické složky 7 či 11 let – z toho vyplývá, že tuto složku používáme u dlouhodobějších pozorování). Cyklická složka se opět nejčastěji modeluje pomocí indexů. Nyní máme model zkonstruovaný – skládá se ze složek trend, sezónní a cyklická složka. Tímto modelem vypočítáme teoretické hodnoty pro naši časovou řadu.

Je pochopitelné, že budeme chtít model nějakým způsobem ohodnotit, zjistit, zdali dobře modeluje, reprezentuje. Pro ohodnocení modelu můžeme použít velmi jednoduchý způsob – porovnáme v každém časovém okamžiku, tedy pro každé naše pozorování, údaj naměřený a údaj vypočítaný (Rumsey, 2007). Kdybychom se pohybovali pouze na úrovni hypotetické, pak při správném modelu by tyto rozdíly musely být nulové – prostě ideální model je ten, který je shodný s realitou. Jenže při měření vzniká celá řada chyb a hodnoty jsou ovlivněna mnoha dalšími neuvažovanými faktory. Z tohoto důvodu je mezi naším modelem a realitou rozdíl. Těmto rozdílům se říká rezidua a teoreticky je lze zařadit jako čtvrtou složku našeho modelu. Takže náš výsledný modul se skládá z trendu, sezónní složky, cyklické složky a náhodné složky (reziduí).

Je zcela logické, že budeme chtít, aby rezidua byla co nejmenší, aby co nejméně škodila modelu. Takový požadavek lze statisticky formulovat takto: rezidua by měla mít nulovou střední hodnotu a konstantní rozptyl. Jinými slovy, celkový vliv reziduí by měl být nulový (nulový průměr všech vychýlení) a po celé „délce“ pozorování by tato vychýlení měla být velice podobná, neměla by se s časem zvětšovat či zmenšovat.

A jako poslední poznámku u analýzy časových řad si uveďme, že existují metody, které se naopak pokouší především modelovat onu reziduální složku. Příkladem budiž Boxova-Jenkinsova metodologie (Kořínek, 1992).

Celá řada „nestatistiků“ si plete analýzu časových řad s **prognostikou**. Proč asi? No pod pojmem pronostika si představují pohled do budoucnosti, v lepším případě odhad vývoje určitého jevu. A analýzu časových řad chápou pouze jako model, který popisuje

stávající data a nic víc. Abstrahují od toho, že analýza časových řad slouží především k predikci, takže k pohledu do budoucnosti. Jaký je tedy rozdíl mezi prognostikou a analýzou časových řad?

Z předchozího je zřejmé, že analýza časových řad pracuje pouze s jedním ukazatelem, pouze s jedním sledovaným jevem a ten analyzuje v čase. Máme tedy právě dvě proměnné – onen ukazatel a čas.

Avšak u prognostiky se pokoušíme predikovat vývoj jednoho či více ukazatelů (jevů) a to na základě celé řady podmínek (Kozák, 1994). Všechny podmínky, další proměnné, jsou nedílnou součástí prognostického modelu. Navíc většina sofistikovanějších prognostických modelů nepracuje s modelem typu status quo (kdy všechny proměnné uváděné v modelu se po celou dobu modelování chovají stejně, dle daných „rovnic“), ale jedná se o modely, ve kterých se dynamicky (v závislosti na čase) seznam proměnných mění a mění se i předpis, podle kterých se proměnné chovají. Dalším zásadnějším rozdílem, který logicky vyplývá i z předešlé poznámky, je časový úsek, budoucnost, na kterou prognostické modely dokáží poměrně dobře predikovat hodnoty modelovaných jevů, ukazatelů. Predikovat vývoj pomocí prognostického modelu na 20, 50 let (a i více, například v demografii) není ojedinělé. Ale predikovat vývoj ukazatele tržba za zimní sportovní oblečení na základě modelu časové řady na příštích 20 let je absurdní, byť by bylo podloženo kdoví jak „přesným“ matematický modelem (je nám zcela jasné, že předpověď na takové dlouhé období bez zahrnutí vlivu demografického vývoje, inflace, předvídatelného sociálního, ekonomického a politického vývoje je zcela vyloučena, odkazuje nás spíše do oblasti věštění z křišťálové koule). No a právě prognostika do svých modelů uvedené vlivy zahrnuje. Navíc výsledkem prognostických modelů je predikce vývoje a to v několika variantách – minimálně se uvádí varianta optimistická (maximalistická), pesimistická (minimalistická) a nejvíce pravděpodobná (uprostřed).

Vraťme se ještě na chvilku k analýze časových řad, protože na ní lze demonstrovat další statistickou disciplínu – **regresní analýzu**. Obecně regresní analýza zkoumá závislost mezi dvěma (a více) proměnnými, mezi závislou proměnnou a proměnnou nezávislou (Meloun, 2006). Zkoumá, podle jakého předpisu je jeden ukazatel závislý na druhém. A zavzpomínejme, u analýzy časových řad, u dekompoziční metody, jsme jako první část modelovali trend časové řady. Jinými slovy pokoušeli jsme se najít jednoduchou tendenční křivku, vztah mezi sledovaným ukazatelem a časem. Nalézali jsme regresní model.

Regresní analýza nám vedle návrhu modelu, podle kterého se chová závislá proměnná na základě změny nezávislé proměnné, ještě podává informace o tom, jak tento modelovaný vztah je silný a jaký má směr (vzpomeňme si na termíny přímá a nepřímá závislost). Regresní validuje daný model – buď je obhajitelný či se musíme podívat po jiném modelu nebo po jiných nezávislých proměnných.

Standardním příkladem může být vztah mez výškou a váhou postavy. Velice zjednodušeně se dá říci, že vyšší člověk váží více. Tento model by však nebyl příliš úspěšný (regresní koeficient by daný model „nepodpořil“), protože víme, že na váhu působí i věk a především typ postavy. Na druhou stranu je asi zřejmé, že ukazatel počet lidí se zrakovými obtížemi (brýlemi) roste s věkem (přímo úměrně) a že regresní koeficient je pro tento model vysoký (tedy že model označí na relevantní, statisticky významný).

Regresní analýza tedy určuje, které námi vybrané nezávislé proměnné jsou pro danou závislou proměnnou relevantní a který model je optimální (Cyhelský, 1981). Regresní model, analýza, slouží pochopitelně vedle studia závislosti na zjištěných datech i k „predikci“, k určení jak se bude model chovat v oblasti, kde jsme data ještě nenaměřili.

U časové řady, kde modelujeme trend, je to zřejmé – pomocí regresního modelu určíme, jak se ukazatel bude chovat v blízké budoucnosti. Ale pomocí regresního modelu můžeme modelovat celou řadu proměnných – z triviálního modelu váha versus výška bychom mohli vypočítat hypotetickou váhu liliputa, tedy člověk s výškou cca 120 cm ale i giganta s výškou 240 cm (nyní nehovoříme o vhodnosti a kvalitě modelu). Stejně tak bychom mohli hypoteticky odhadnout, kolik dioptrii bude potřebovat 120letý „stařec“.

Poznamenejme, že jednou ze základních, jednodušších oblastí regresní analýzy je lineární regresní model (závislost mezi dvěma proměnnými modelovaná přímkou).

U **vícerozměrných metod** (Meloun, 2006) pracujeme s jednou závislou a celou řadou nezávislých proměnných (alternativní pojem je vysvětlovaná proměnná a vysvětlující proměnné – závislou vysvětlovanou proměnnou se snažíme popsat množinou nezávislých vysvětlujících proměnných). Příkladem může být zkoumání, kdy u studentů sledujeme následující znaky: výška, váha, věk, pohlaví, studovaný obor, absolvovaná střední škola, známka z matematiky na střední škole, velikost bydliště (vesnice, městys, město, velké město), počet sourozenců, vzdělání rodičů a známka z matematiky na studované vysoké škole. Naším úkolem je zjistit, zdali je znak známka z matematiky na studované vysoké škole ovlivněna (a jak) následujícími uvedenými znaky. Z vícerozměrné statistické analýzy

pak dostaneme model, který zahrnuje pouze relevantní proměnné. Výsledkem vícerozměrné statistické analýzy je konstatování a (číselné, kvantitativní) odůvodnění jednotlivých nezávislých proměnných – ujasnění, proč byly či nebyly do modelu zařazeny. Tato analýza také zkoumá, zdali některé původně vybrané a sledované nezávislé proměnné nejsou mezi sebou v jednoduchém vztahu (přímá či nepřímá úměra, tzv. korelace) případně úměra, závislost posunutá v čase, tzv. autokorelace (Cipra, 1986).

Další oblastí statistiky je zřejmě nejznámější (a pro nepříliš matematicky sběhlé čtenáře nejodpudivější) **pravděpodobnost** (resp. teorie pravděpodobnosti) (Hátle, 1987). Ta nám dokáže odpovědět na otázky, jaká je pravděpodobnost, že při dvou hodech kostky padnou dvě šestky, jaká je pravděpodobnost, že z balíčku karet vytáhnu srdcovou sedmu a pikovou dámu atd. Ale tím to začíná být právě zajímavé – jaká je pravděpodobnost, že označený otec je skutečným genetickým otcem (test paternity), jaká je pravděpodobnost, že daný označený pachatel je skutečným pachatelem (forenzní statistika), jaká je pravděpodobnost, že při náhodné výběru 20 ks televizorů bude alespoň jeden poškozený?

Teorie statistiky nám dává matematický aparát k tomu, abychom mohli všechny uvedené (a pochopitelně celou řadu) úloh jednoduše řešit. A aby to nebylo tak jednoduché, vedle vzorců pro standardní matematickou statistiku (součet a součin, závislé a nezávislé jevy) se používá i bayesovská statistika (která v některých oborech prožívá renesanci – například forenzní genetika).

Z teorie pravděpodobnosti se vyvinula **matematická statistika** (**statistická indukce**). S touto disciplínou se setkáváme velice často, neboť tvoří základ „moderní“ statistiky. Všechny výše uvedené metody totiž dokázaly popsat (s větším či menším úspěchem) pouze pozorovaný soubor (naměřená, zjištěná data). Ale matematická statistika dokáže odpovědět na otázku, která každého zcela jistě napadne: když je čistý průměrný měsíční příjem námi zkoumaného souboru ten a ten, jaký je asi u celé populace? S jakou pravděpodobností se čistý průměrný příjem bude pohybovat v daném rozmezí? Jaké je rozmezí tohoto přijmu, aby se do toho intervalu vešlo 90 procent populace?

A napadne nás ihned celá řada dalších otázek – je opravdu (statisticky významný) rozdíl mezi věkem dožití kuřáků a nekuřáků? Je významný rozdíl mez příjmy žen a mužů ve specifikované věkové skupině a se s daným vzdělávám? Je významný rozdíl ve spotřebě alkoholických nápojů u mladistvých mezi Čechy a Slováky?

K zodpovězení těchto otázek používá statistická indukce matematického aparátu – testování statistických hypotéz (Hátle, 1983). Na každou skupinu otázek existuje odpovídající statistický test. Jen na okraj si řekněme, že mezi tyto testy patří t-test, f-test, chí-kvadrát test a celá řada dalších. Při odpovídání na tyto otázky se můžeme, jako statistikové, dopustit dvou chyb – říká se jim chyba alfa a chyba beta (resp. chyba prvního a chyba druhého typu, druhu). Chyby prvního druhu se dopustíme, pokud odmítneme pravdivou hypotézu (odmítneme domněnku, že je rozdíl mezi příjmy mužů a žen, přestože tento rozdíl je realitou, je pravdivý). Chyba druhého druhu je v případě, že hypotéza zamítnuta není, přestože neplatí (Hanousek, 1992). (Jinými slovy se jedná se o chybné rozhodnutí při selhání testu v odmítnutí falešné hypotézy.)

Při pokusu o popis jednotlivých statistických disciplín jsme se již několikráte dotkli významu výběru jako důležitého statistického prvku. **Teorie výběrových zjišťování** nám předkládá metodologii, kterak vytvářet relevantní, statisticky správné, výběry (Čermák, 1980). Je nám jasné, že pokud chci „spočítat“ (určit) průměrný hrubý měsíční příjem ekonomicky aktivního obyvatelstva nebude zřejmě vhodné (nedostatek základních zdrojů – finance, času a pracovníků) dělat průzkum u celého základního souboru, u všech těchto ekonomicky aktivních obyvatel. Zřejmě souhlasíme s tím, že postačí vybrat si správnou skupinu a na základě informací onen průměr určit. Také je nám jasné, že pokud oslovíme „náhodně“ 4 štamgasty v restauraci, těžko získáme relevantní data, ze kterých bychom mohli určit požadovanou statistik – průměrný hrubý měsíční příjem ekonomicky aktivního obyvatelstva České republiky.

Teorie výběrových zjišťování nám pomůže s tím, jak statisticky správný výběr vytvořit (čermák, 1980). Bavíme se pak o výběrech náhodných, stratifikovaných jedno a dvou fázových atd. Ale toto je jen první část, teorie výběrových zjišťování nám nastíní, které statistiky a jak máme vypočítat, abychom z výběru mohli usuzovat na hodnotu statistiky pro celý základní soubor (populaci). Konkrétně, pokud máme statisticky korektní výběr – například správný výběr pro náš příklad průměrného hrubého měsíčního příjmu ekonomicky aktivního obyvatelstva – a to jak co do velikosti tak i struktury (poměr mužů a žen, poměr vzdělanostní struktury, poměr obyvatel z vesnice, malých měst a velkoměst, poměr jednotlivých oborů), pak si umíme představit, že průměr vypočítaný z tohoto výběru je asi dobrým odhadem pro průměr celého základního souboru, souboru všech ekonomicky aktivních obyvatel České republiky. Ale jak je to s dalšími statistikami? Například zásadní statistika popisující variabilitu, měnlivost, jednotlivých „příjmů“ (směrodatná odchylka,

resp. rozptyl) se pro výběr a základní soubor počítá podle jiného vzorce! Rozdíl je sice „nepatrný“, liší se pouze jmenovatel – pro základní soubor je ve jmenovateli počet pozorování, tedy *n*, kdežto pro výběr je ve jmenovateli hodnota *n-1*.

Teorie výběrových zjišťování nám předkládá nejen metodologii jak vytvářet statisticky korektní výběry, ale i definuje (opravené) statistiky, které musíme při výpočtu použít.

Pozorný čtenář si je vědom, že předešlé statistické disciplíny pracují s kvantitativními daty, s „počitatelnými“ údaji. Ale kolem nás je celá řada jevů, které prozatím popsujeme nikoli pomocí čísel, ale pomocí slovního vyjádření. Lze i kvalitativní údaje statisticky zpracovávat? Na tuto otázku odpovídá **kvalitativní statistika**[2]**,** jako další disciplína. Jedná se o relativně novější disciplínu, která si nicméně již v minulém století našla svoji pozici.

Uvědomme si, že mnoho kvalitativních dat lze vhodnou transformací (kódováním atd.) převést na čísla. Další možností je pokusit se kvantitativní data rozdělit do několika základních skupin (Hendl, 2005), případně přejít na čistě alternativní rozdělení (například při znaku barva očí zvolit pouze dvě skupiny – modré a ty ostatní).

Tímto jsme se pokusili pouze naznačit cesty, kterými kvalitativní statistika prošla, aby dospěla k takovým základním postupům, jako jsou kontingenční tabulky a testi chí-kvadrát, Spearmanův koeficient pořadové korelace a mnoho dalších (Sharma, 2005).

Na závěr této kapitoly si poznamenejme, že vedle výše představených více méně teoretických statistických disciplín **aplikovaná statistika** pronikla do mnoha oborů, které nesou názvy s „koncovkami“ -metrie jako biometrie, dendrometrie, ekonometrie, chemometrie a celá řada dalších. A často se setkáváme s vědami se silným statistickým základem – sociologie, psychologie, demografie aj. (např. forenzní genetika[3]).

2.3 Statistický software

Statistika se zabývá hromadnými jevy a pro analýzu shromažďuje poměrně velké množství dat. Ve své druhé etapě (viz Moore – 2.2.1) jsou data analyzována, jsou s daty prováděny statistické výpočty. Je známo, že statistické vzorce nejsou triviální a výpočty jsou značně pracné. Proto již s prvními počítači se objevily statistické programy, které

[2]DISMAN, M.: Jak se vyrábí sociologická znalost. Příručka pro uživatele. Praha, Karolinum, 2006. ISBN 80-246-0139-7

[3]LUCY, D.: *Introduction to statistics for forensic scientists.*Chichester, Wiley, 2008. ISBN 0-470-02201-9

rutinní výpočty usnadňují. S nástupem PC se statistické programy značně rozšířily[4] a v současné době můžeme tuto oblast software rozdělit do následujících skupin:

- **Statistické pakety** – jedná se o „balík“, který nabízí celou řadu různých statistických metod. Většinou jsou jednotlivé nabídky s metodami rozděleny do velmi podobných skupin, které jsme si představili k předešlé kapitole (2.2.3). Statistický paket tedy můžeme použít jako univerzální nástroj pro statistické analýzy – umožňuje pracovat se vstupními daty (obdobně jako tabulkové procesory), nabízí a pokrývá většinu statistických analýz (procedur) a výsledky analýz prezentuje formou tabulek a grafů[5].

- **Specializované programy** – jedná se o programy, které nabízejí pouze jednu statistickou metodu či metody z jedné statistické disciplíny a to většinou jednu z nejnovějších či v praxi nejméně často používanou. Takový program nabídne například pouze analýzu časových řad dle Boxovy-Jenkinsovy metody, modeluje data pouze lineární regresí, nabídne z deskriptivní statistiky pouze tvorbu Box-plot či S-L grafu.

- **Matematické systémy** – celá řada matematických programů, tedy softwarových systémů primárně určených k matematické analýze či matematickému modelování obsahuje často (spíše jako bonus) některé statistické výpočty.

- **Tabulkové procesory** – od prvního spreadsheetu[6], který byl primárně určen k snadnějším ekonomicko-finančím výpočtům se současné tabulkové procesory propracovaly v sofistikované systémy, v jejichž nabídce nalezneme celou řadu matematických a hlavně statistických funkcí[7] a to dokonce nejen z oblasti popisné statistiky, ale i z testováni hypotéz či regresního modelování[8], a také testování kontingenčních tabulek pomocí chí-kvadrát testu atd. Pro nikoli jen

[4]ŽVÁČEK J.: *Statistické výpočetní prostředí 2007*. In Informační Bulletin České Statistické Společnosti. Listopad 2007, roč. 18, č. 3, s. 1 -15. ISSN 1210-8022.
[5]KOŘÍNEK, M: Statistický programový paket SPSS a Boxova-Jenkinsova analýza časových řad. In MAA. Číslo 3, 1992. s. 74-77.
[6]KOŘÍNEK, M: Možnosti použití tabulkového procesoru QUATTRO PRO 3.0 pro statistickou analýzu dat. In Statistika. Číslo 3, 1992. s. 128-139.
[7]KOŘÍNEK, M: *Výuka statistiky a Excel 97*. In Pedagogický software ’97 (sborník přednášek). České Budějovice. 1997. s. 87-89. ISBN 80-85645-26-2.
[8]KOŘÍNEK, M.: Některé zkušenosti s využitím grafické podpory tabulkového procesoru SuperCalc. In Statistika 7/1991, ss. 308-316

úvodní seznámení se statistikou jsou tabulkové procesory asi tím nejjednodušším řešením[9].

- **Databázové systémy** – pod pojmem databáze si zjednodušeně představíme tabulku či více tabulek, která obsahují data, převážně číselná. A tato data lze také statisticky analyzovat. Proto databázové systémy nabízejí (například už v samotném jazyce SQL) celou řadu možností, jak data statisticky analyzovat a zjišťovat základní statistiky. Z tohoto důvodu je vhodné do sekce statistického software zařadit i tyto databázové systémy.
- Podpora výuky – na konec našeho seznamu statistického software si zařadíme celou řadu programů, které slouží k podpoře výuky a obecně k podpoře povědomí o statistice. Jedná se o taková řešení jako **elektronické statistické učebnice a příručky**(e-book) či celá **výuková statistická (multimediální) prostředí** doplněná o **interaktivní prvky**. Na internetu nalezneme mnoho příkladů.

A s jakými statistickými programy (ze sekce statistical package) se můžeme nejčastěji setkat[10]?

SPSS–Statistical Programs for Social Sciences (www.spss.com) – velmi oblíbený software, který pochází ještě z dob velkých sálových počítačů, sestávající se z mnoha doplňkových modulů, v současné době je již v portfoliu společnosti IBM.

NCSS–Number Cruncher Statistical Systems (www.ncss.com) – paket vzešel z univerzitního prostředí a díky licencování (pro studenty a výuku na školách se poskytuje zdarma) je hojně rozšířen i pro snadné ovládání a hlavně parametrizovatelnost statistických procedur a výstupů[11].

STATISTICA (www.statistica.cz) – opět jeden ze starších systémů, který doznal značných změn a jako jedem z mála je lokalizován (v češtině). Oproti SPSS má výhodu, že

[9] KOŘÍNEK, M.: *Excel 2002.* České Budějovice: Kopp, 2001. ISBN 80-7232-156-0.
[10]KOŘÍNEK, M: Možnosti Zkušenosti se softwarem používaným při výuce statistiky v humanitních oborech na Pedagogické fakultě Univerzity Hradec Králové. In Nové technologie ve vzdělávání (vzdělávací software a interaktivní tabule). On-line mezinárodní vědecko-odborná konference Olomouc 2010. URL: <http://www.kteiv.upol.cz/ntvv/?konf=konference2&detail-prispevku=57> [citováno 10. října 2010].
[11]TVRDÍK, J.: *STAT a NCSS z pohledu uživatele*. In Informační Bulletin České Statistické Společnosti. Prosinec 1997, roč. 8, č. 3, s. 5 -16. ISSN 1210-8022.

už v základu obsahuje velké množství statistických procedur, které lze navíc snadno parametrizovat[12].

SAS–Statistical Analysis System (www.sas.com) – jeden z dalších světoznámých paketů, který ovšem v grafických výstupech značně pokulhává za ostatními (pseudografika).

QC.Expert– dříve **ADSTAT** (www.trilobyte.cz) – jedná se o čistě český statistický paket, u jehož zrodu stáli profesor Militký (Liberec) a profesor Meloun (Pardubice). Je méně známý a rozšířený především na univerzitách. Díky společnosti Trilobyte je využíván při mnoha statistických výzkumech. K programu existuje i velice zdařilá učebnice.

Mezi další statistické pakety patří STATGRAPHICS[13][14], MINITAB, GAUSS, GENSTAT, S-PLUS a celá řada dalších[15].

[12]KLETEČKOVÁ, M.: *Statistický systém STATISTICA.* In Informační Bulletin České Statistické Společnosti. Prosinec 1998, roč. 9, č. 3, s. 9 -12. ISSN 1210-8022.
[13]KÁRNÍK, I., SVOBODA, L.: *STATGRAPHICS – studnice poznání.* In Informační Bulletin České Statistické Společnosti. Červen 1997, roč. 8, č. 1, s. 20 -22. ISSN 1210-8022.
[14] KOŘÍNEK, M: *Statistikův bedekr či STATGRAPHICS v kostce?* In MAA. Číslo 3, 1993. s. 52-53.
[15]Viz například http://en.wikipedia.org/wiki/List_of_statistical_packages

3 Úvod do statistiky

V této kapitole se seznámíme s některými základními statistickými pojmy. Jejich pochopení je nezbytné pro navazující výklad a především pro statistické šetření a zpracování dat. Přestože se jedná stále je o teoretický „textový“ popis, neměli bychom toto téma přehlédnout.

3.1.1 Základní a výběrový soubor

Základní soubor (nazývaný též **populace**) je množina všech teoreticky možných objektů (např. jedinců) v uvažované problémové situaci (Čermák, 1980). Například při volbách do parlamentu se zajímáme o všechny osoby s volebním právem v dané zemi. Při zkoumání sportovních jednot v naší republice uvažujeme všechny sportovní jednoty v našem státě. Ne vždy si můžeme populaci takto reálně představit. V mnoha situacích má populace hypotetický význam (celé obyvatelstvo naší planety či dokonce všechno tvorstvo).

Podmnožina základní populace (souboru) **se nazývá výběr (výběrový soubor)** nebo **vzorek**. Například vybereme určité množství sportovních jednot a podrobíme je zkoumání. Obvykle totiž nemáme možnost z finančních, časových či etických důvodů /obecně hovoříme o nedostatku zdrojů – čas, finance, počet statistických pracovníků) zkoumat všechny prvky základní populace. Počet prvků ve výběru nazýváme rozsah výběru (Čermák, 1980).

3.1.2 Proměnné a jejich typy

Statistická analýza se zabývá analýzou dat, která se získávají zjištěním hodnot předem definovaných proměnných, resp. znaků na definovaných objektech.

Proměnné nebo znaky jsou charakteristiky prvků základního souboru, jež mohou nabývat více hodnot. Například agenturu provádějící předvolební průzkum mohou u voliče zajímat proměnné věk, pohlaví a jakou stranu nebo kandidáta bude volit. Při zkoumání sportovních jednot nás zajímá jejich velikost, počet a typ podporovaných sportů, počet funkcionářů, trenérů a členů, ekonomické ukazatele apod. (Hendl, 2005).

Data tvoří aktuální hodnoty proměnných. Některé se zjišťují poměrně snadno - dokážeme např. změřit výšku či váhu žáků určité školní třídy. Dokážeme změřit i mnohé jejich psychologické rysy, ale s tím jsou už spojeny určité problémy. Některé proměnné se navíc v čase mění. K určení hodnot mnoha proměnných je zapotřebí značného úsilí. Například zachycení ekonomické situace sportovní jednoty si může vyžádat detailní informace z účetnictví, přičemž získání těchto dat může být časově značně náročné.

Při zkoumání dat mluvíme o rozdělení proměnné. Tímto výrazem rozumíme to, jaké hodnoty proměnná nabývá (meze hodnot – očekávaná nejmenší a největší hodnota, průměrná hodnota) a jak často se jednotlivé hodnoty vyskytují (četnosti). Pojem rozdělení proměnné pomáhá popsat statistické chování dané proměnné, daného znaku, jevu (Hátle, 1987).

3.1.2.1 Závislost proměnných

Pokud budeme statistickyanalyzovat vícerozměrný statistický soubor (viz strana 20), tedy zkoumat závislost mezi proměnnými, pak si musíme poměrné správně pojmenovat a rozčlenit. Rozlišujeme dva typy proměnných – **závislé** a **nezávislé** (Chráska, 2008).

Poznamenejme si, že někdy používáme alternativní označení pro závislou proměnnou odpovědová, kriteriální nebo cílová a pro nezávislou proměnnou pak termíny prediktor nebo explanační proměnná (Hendl, 2006).

Statistická analýza (tedy vícerozměrná) začíná definováním závisle a nezávisle proměnných a další její krokymohou případně určit existenci a další atributy (sílu, směr, významnost) jejich vztahu. Často se předpokládá mezi proměnnými příčinný vztah, což znamená, že změna nezávisle proměnné způsobuje změnu závisle proměnné bez ohledu na přítomnost jiných proměnných.

Uveďme si příklady závisle a nezávisle proměnných (dvojice):

- závislost chování (závisle proměnná) na pohlaví žáka v dané věkové kategorii (nezávisle proměnná);
- závislost průměrného prospěchu (závisle proměnná) na typu školy (nezávisle proměnná);
- závislost výkonu (např. při vzpírání) jedince (závisle proměnná) na objemu jeho svalů (nezávisle proměnná);

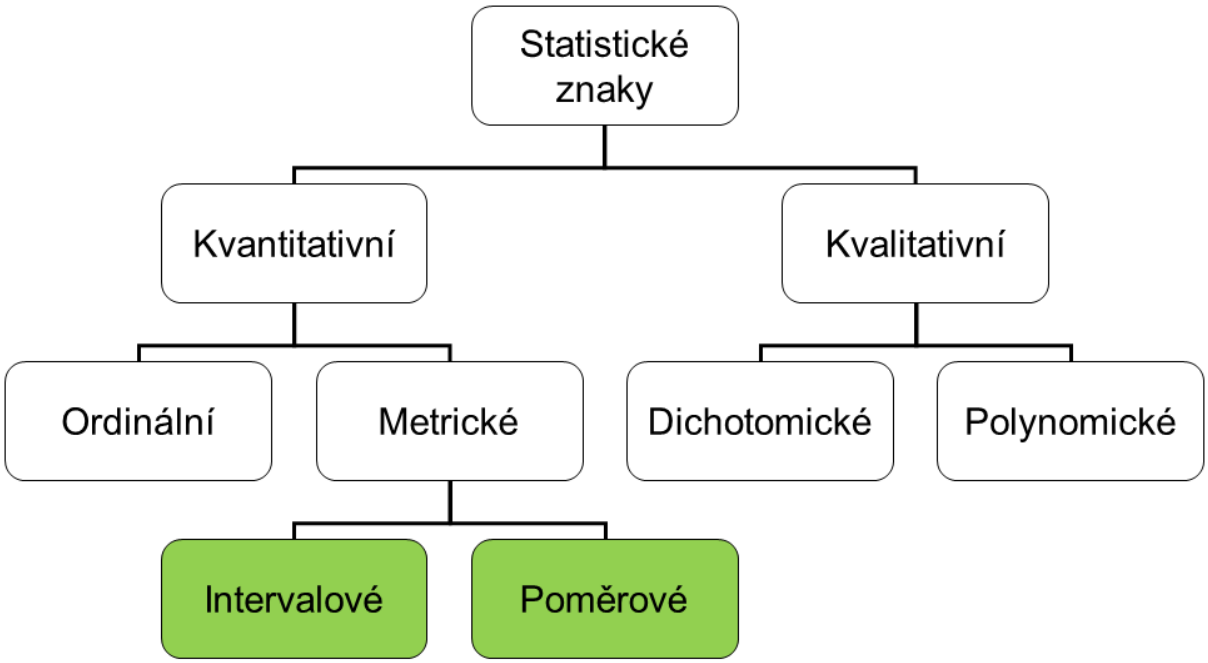
3.1.2.2 Měřítko u proměnných

Pojem měření se často ztotožňuje (zřejmě pod vlivem matematiky a fyziky) se získáváním metrických údajů. Naměřená hodnota je v nich násobkem nebo částí jednotky měření. V sociálních vědách se používá definice, která říká, že měření znamená přiřazení čísel objektům, které je ve vazbě na teorii daného oboru (Hendl, 2005).

Proměnné lze rozlišovat podle toho, co jejich hodnoty vypovídají, jakých možných variant mohou nabývat. Mluvíme pak o škále měření nebo použitém měřítku (případně doméně či oboru hodnot) (Cyhelský, 1981).

- **Kvalitativní (nominální)** měřítko znamená přiřazení, které pouze vyjadřuje, že lze rozlišit jednotlivé hodnoty. Například žáky lze rozlišit podle toho, zda umějí plavat, nebo ne. U těchto měřítek můžeme určit počet použitých kategorií (například barva očí – hnědé, zelené, šedé, modré). Jestliže rozlišuje použité měřítko jenom dvě třídy, mluvíme o binárním (alternativním, dichotomickém) znaku, jinak o polynomickém znaku.
- **Ordinální** měřítko kromě rozlišení tříd ještě vyjadřuje nějaké jejich řazení podle intenzity nebo pořadí. Data s ordinálním měřítkem lze uspořádat (třídit). Například můžeme seřadit žáky podle jejich prospěchu z matematiky, ale patří sem i bodové ohodnocení (nejen graficky někdy znázorňované jako počet hvězdiček atd.).
- **Intervalové** měřítko má vlastnosti ordinálního měřítka. Hodnota znaku u statistické jednotky může nabývat libovolné hodnoty z intervalu (například běh na 100 metrů, váha určitého výrobku a celá řada dalších)

Obě posledně uváděná měřítka společně označujeme jako **metrická**, a spolu s ordinálním měřítkem je shrnujeme do skupiny **intenzivních měřítek**.



Obrázek 1 - Rozdělení statistických znaků

Nyní si vymezíme popsaná měřítka pomocí matematických symbolů. Na dvou objektech *A* a *B* získáme měření s hodnotami x_a a x_b. Pro jednotlivá měřítka platí následující pravidla (Lamser, 1970):

- Při nominálním měřítku můžeme mít pouze vztahy x_a - x_b, nebo x_a/x_b (např. muž, žena).
- Ordinální měřítko také dovoluje vztahy $x_a>x_b$ nebo $x_a<x_b$ (např. světlý, šedý, tmavý).
- Intervalové měřítko navíc předpokládá, že může být definována velikost rozdílu, takže objekt A se liší o x_a - x_b jednotek od objektu B (např. teplota ve stupních Celsia).
- Poměrové měřítko má navíc definovanou absolutní nulu, takže má smysl říci, že A je x_a/x_b větší než B, jestliže $x_b \neq 0$ (např. teplota v Kelvinově stupnici).

3.1.2.3 Proměnné diskrétní versus spojité

Při statistickém zpracování dat hraje důležitou roli to, zda jsou hodnoty diskrétní, nebo spojité. **Spojitá** proměnná může teoreticky nabývat libovolných hodnot z určitého

intervalu (reálných) čísel. Naproti tomu **diskrétní** proměnná (neboli **kategoriální** proměnná) nabývá pouze konečného počtu hodnot (Wonnacott, 1992).

Poznamenejme však, že dle kontextu lze i s diskrétními údaji pracovat jako se spojitými, např. tepovou frekvenci můžeme považovat za spojitou proměnnou, pokud se setkáváme v analýze s velkým počtem různých hodnot. Jinými slovy, statistik chápe rozdělení proměnných na diskrétní a spojité odlišně od matematika. Pro statistika je nejdůležitější počet různých hodnot zkoumaného jevu.

Někdy uvažujeme zvlášť proměnnou pořadovou ordinální, která vznikne seřazením jedinců podle hodnot získaných změřením nějaké spojité intervalové nebo poměrové proměnné, a kategoriální ordinální proměnnou, která rozeznává jenom několik ordinálně uspořádaných kategorií, do kterých jedince nebo objekty zařazujeme (Chráska, 2008).

Upozorněme, že měřítko znaku spolurozhoduje o tom, jakou statistickou techniku použijeme při zpracování. Často se připomíná odstrašující příklad, jestliže se počítá průměr pro kvantitativně kódovaný kvalitativní znak (např. kódovali jsme žlutý - 0, zelený - 1, modrý - 2). Pak ovšem vede statistické zpracování k nesmyslným tvrzením.

3.2 Kvalita statistiky

Jeden známý vtip říká: Známe tři lži – Lež, úmyslná lež a Statistika“. Aby statistika nelhala, musíme striktně dodržovat některá pravidla. Co se týče měření, tedy prvního stupně statistiky, lze nalézt tato pravidla.

3.2.1 Objektivita

Objektivita měření znamená stupeň toho, jak jsou výsledky nezávislé na výzkumníkovi nebo měřeném jedinci ve smyslu subjektivního úmyslného nebo neúmyslného zkreslení. Při měření fyzikálních veličin v laboratoři se otázka objektivity objevuje zřídka (ačkoli porucha měřicího přístroje či špatná metodika se může vyskytnout), ale při hodnocení měření v pedagogice, sociologii nebo psychologii se objektivita musí pečlivě přezkušovat (Hendl, 2005).

3.2.2 Spolehlivost

Spolehlivost (reliabilita) měření znamená stupeň shody (konzistence) výsledků měření jedné osoby nebo jednoho objektu provedeného za stejných podmínek.

Nespolehlivost (nízká reliabilita) měření má různý původ. Jeden zdroj nespolehlivosti obvykle nazýváme **subjektivní chybou.** Zapříčiňuje ji individuální variabilita (únava, klesám zájmu, nahodilý stres, vliv proměnlivého prostředí atd.) měřeného subjektu. Pozorovací chyba je jiným zdrojem chyb. Závisí na provedení měření hodnotitelem. Také uvazujeme přístrojové chyby (např. selhání přístroje).

Existuje mnoho postupů k určení spolehlivosti měření (Hendl, 2005):

- opakovaná měření (test-retest reliabilita) - označujeme tak konzistenci neboli shodu opakovaných měření, jež jsou rozdělena do různých časů;
- měření paralelních testů - znamená shodu měření s jiným ekvivalentním měřením (například dvě verze A a B téhož testu, paralelní pokusy na dvou měřících přístrojích apod.);
- půlení testu (split-half reliabilita) - vyjadřuje, do jaké míry jsou konzistentní jednotlivé části instrumentu měření (nejčastěji se týká různých položek jednoho testu). V posledním případě se jedná o metodu posuzování internami konzistence, která nevyžaduje u jedince opakované použití měřicí procedury.

Pokud měření není spolehlivé, nemůže být ani validní.

3.2.3 Validita

Validita odkazuje na přiměřenost, smysluplnost a užitečnost konkrétních závěrů, jež se provádějí na základě výsledku měření. Validace měřicí metody je procesem k podpoře této premisy. Posuzují se provedená rozhodnutí, nikoli měřicí instrument jako takový.

Při ověřování obsahové validity zjišťujeme, do jaké míry měření reálně odráží dané vlastnosti nebo kvality. Například při tvorbě zkouškové ho testu si všímáme, zdali otázky pokrývají celou problematiku zkoušené látky.

3.3 Výběry

Jak jsme si již uvedli, mnohdy při statistickém šetření nemáme k dispozici dostatek zdrojů (finančních, časových, personálních a dalších) a proto často volíme při šetření metodu statistického výběru.

Při výběrovém šetření jde o sběr informací standardizovaným způsobem (například pomocí standardizovaného dotazníku) od skupiny lidí. Výzkumník shromažďuje data o populaci (tedy o základním souboru) pomocí konkrétní formy výběru jedinců nebo jednotek (definovaný výběrový soubor). Takové studii říkáme někdy statistické šetření nebo zjišťování. Takto provedené statistické šetření je charakterizované těmito dvěma základními vlastnostmi:

- Provádí se výběr jedinců z nějaké známé populace.
- Jedná se o sběr relativně malého množství dat ve standardizované podobě od relativně velké skupiny jedinců.

3.3.1 Druhy výběrů

Jak jsme již naznačili, jsme například v situaci, že není možné (třeba z finančních či časových důvodů) získat data od celé populace. Provedeme proto výběrové šetření. Plán výběru označuje metodu, která se použije pro výběr podmnožiny statistických jednotek ze základní populace. Nejčastěji se setkáme s těmito čtyřmi druhy výběrů (Čermák, 1980):

- Výběr na základě **dobrovolnosti** se často aplikuje v průzkumech veřejného mínění. V tomto plánu se obvykle jedná o získání odpovědí na jednu nebo několik otázek. Jedinci z populace se sami rozhodují, zda odpovědět, nebo ne. Například po televizním pořadu jsou diváci vyzváni, aby se vyjádřili k diskutované otázce. Většinu na výzvu reagují pouze vysoce motivovaní diváci, tedy Ti s krajními „opozičními“ názory (striktně se přiklánějící na jednu či druhou stranu), téměř nikdy se neozve ten, jehož názor leží „uprostřed“.
- Výběr na základě **dostupnosti**. Jedinci jsou z populace vybráni na základě dostupnosti. Například jestliže provádíme průzkum o nákupních zvycích jedinců v obchodě knihami, přičemž zvolíme výběr statistických jednotek na základě okamžité dostupnosti v této prodejně v daném okamžiku. Jiný příklad představuje genetický výzkum u pacientů, kteří se léčí s konkrétní diagnózou.

- **Kvótní výběr**. V tomto případě mají tazatelé za úkol provést rozhovor s určitým počtem jedinců v několika různých předem definovaných skupinách obyvatelstva. Za kategorie se volí např. věk, pohlaví, bydliště, profese nebo ekonomická (finanční) situace. Vychází se obvykle z demografických informací o obyvatelstvu.
- **Náhodný výběr**. Základní doporučení říká, že je pro statistické šetření nejlepší. Jedná se oteorii, která je v praxi často jen stěží či vůbec uskutečnitelná. Náhodný (či též pravděpodobnostní) výběr ze základní populace je takový, který splňuje následující podmínky:
    1. každý prvek populace má předem známou pravděpodobnost, že bude do výběru zařazen;
    2. výběr je proveden pomocí techniky, jež tyto pravděpodobnosti výběru realizuje;
    3. pravděpodobnosti výběru prvků se uvažují při zpracování získaných dat.

První tři popsané techniky výběru nejsou zcela optimální, protože získaná data mohou být zkreslena. Toto zkreslení je systematickou chybou. Prostý náhodný výběr eliminuje výběrové zkreslení, protože všechny podmnožiny daného rozsahu mají stejnou šanci, že budou vybrány pro pozorování, dotazování nebo měření.

Proto se v následujícím soustředíme právě na prostý náhodný výběr.

3.3.1.1 Prostý náhodný výběr

Základní typ pravděpodobnostního výběru je prostý náhodný výběr - pravděpodobnostní výběr, kdy každý prvek základního souboru má stejnou pravděpodobnost, že bude vybrán.

Poznamenejme si, že jednou z mnoha možností jak realizovat náhodný výběr, je například pomocí očíslování všech statistických jednotek základního souboru. K samotnému výběru pak lze použít generátor náhodných čísel (počítačový program či losovací zařízení), s jehož pomocí vybereme statistické jednotky (vytvoříme náhodný výběr).

3.3.1.2 Ekvivalent prostého náhodného výběru

Stává se, že prostý náhodný výběr je neproveditelný nebo nákladný, hlavně tehdy, když je základní populace značně rozsáhlá. Proto se používají některé alternativní náhradní metody výběru, jež využívají ve výběru náhodný mechanismus (Čermák, 1980).

- **Stratifikovaný náhodný výběr**. Jestliže populace obsahuje další podskupiny, je možné rozdělit populaci do těchto skupin a provést prostý náhodný výběr pro každou skupinu. Tyto podskupiny se nazývají strata nebo vrstvy. Jednotlivé skupiny musíme volit tak, aby byly homogenní (populaci rozdělíme dle dosaženého vzdělání na základní, středoškolské, středoškolské s maturitou, vysokoškolské typu bakalářského atd.) Výsledky pro všechny skupin pak tvoří výběr. Tato technika je vhodná, jestliže populaci lze stratifikovat podle pohlaví, věku nebo demografických parametrů a výzkumník chce zajistit reprezentaci každé podskupiny.
- **Vícestupňový shlukový výběr** se často používá pro získání informací o veřejném mínění, ale tato technika je základem i microcenzu (testu pro sčítání lidu). Při této variantě postupujeme následujících způsobem:: 1. vybere se náhodně vzorek okresů; 2. z takto vybraných okresů se v každém okresu vybere náhodně určitý počet lokalit (měst a vesnic); 3. pro takto vybraná města se vybere náhodně vzorek jejich ulic; 4. z vybraných ulic se náhodně vyberou domácnosti, ve kterých se provede dotazování. V každé vrstvě shluků se provádí náhodný výběr. Tato vícestupňová procedura vypadá velmi komplikovaně, ale ve skutečnosti je velmi efektivní a méně nákladná než prostý náhodný výběr.
- **Systematický výběr.** Tato metoda začíná soupisem a očíslováním statistických jednotek. Po té se rozhodneme, jak z tohoto seznamu systematicky vybírat prvky (například se bude vybírat vždy jeden prvek ze sta). Nejdříve si zvolíme náhodně prvek z první stovky (tato statistická jednotka má své pořadové číslo, například 65). V dalším kroku k tomuto číslu postupně přičítáváme číslo 100 a prvky s takto získaným pořadovým číslem zařazujeme d výběru (tedy 65, 165, 265 atd.). Aby byl systematický výběr validní, musíme zajistit náhodnost číslování prvků (tedy že pořadí nebude závislé na nějaké, byť skryté, vlastnosti).

3.4 Problémy výběrových šetření

Na konec kapitoly o výběrech se pozastavme nad vybranými otázkami, které nám pomohou odstranit zkreslení výběrových šetření (Hendl, 2005).

- Kdo realizoval výzkum? Výzkum preferencí zákazníků může provádět např. jeden obchodní řetězec. Neovlivní to získané výsledky?
- Jaký byl základní soubor? Koho jsme se ptali?
- Jaká byla opora výběru? Pokud opora výběru nezahrnuje celou populaci, výběr bude zkreslen stejným způsobem, jako je zkreslena opora výběru. Například pokud použijeme jako oporu výběru očíslované jedince z registru adres (trvalé bydliště), může jít o zkreslenou oporu výběru, neboť statistická jednotka se může pohybovat především na „přechodné“ či úplně jiné adrese (může se jednat o efekt „nepokrytí" - některé skupiny nebyly z nějakého důvodu zahrnuty do výběru).
- Jakou metodou byl vzorek vybrán? Byl použit náhodný výběr?
- Jak byl vzorek rozsáhlý (poměr základního a výběrového souboru)? Je důležité vědět rozsah výběru a dosažené odhady chyb v podobě intervalů spolehlivosti.
- Jaká byla návratnost? Kolik se nám vrátilo vyplněných dotazníků?
- Jak byli jedinci kontaktováni? Telefonicky, poštou nebo osobním rozhovoremdoma?
- Kdy bylo šetření provedeno? Stalo se tak po nějaké významné události (politické, ekonomické, přírodní), která mohla ovlivnit odpovědi?
- Došlo ke zkreslení odpovědí? Je dáno špatným sestavením dotazníku nebo subjektem dotazování, jenž úmyslně dává falešné odpovědi. Dotazovaný také může být ovlivněn negativně tazatelem.
- Jaké otázky a v jakém pořadí se kladly? Závisí na tazateli.

4 Základní statistické vyjadřovací prostředky

Statistika – a statistické zpracování dat – používá dva základní prostředky, kterými prezentujeme statistické výsledky – tabulky a grafy. V druhé fázi tedy při analýze dat, se právě pomocí tabulek, grafů a souhrnů překládají statistické výsledky (srovnej Moor viz kapitola 2.2.1).

Náplní této kapitoly je popis statistické tabulky a grafů a to tak, abychom jim nejen správně rozuměli, ale abychom je byli schopni korektně sestavovat.

4.1 Tabulka versus graf

Nejprve si položme standardní otázku, zdali pro popis statistických dat (pozorování) a výsledků je příhodnější tabulka či graf. Odpověď je shodná s tou, kterou jsme si uvedli v případě porovnávání údajů – je vhodnější absolutní či relativní srovnání (jinými slovy rozdíl či poměr dvou hodnot)?

Statistický popis dat a výsledků musí být vždy doplněn jak tabulkou, tak i grafy (Maněnová, 2012). A tyto dva vyjadřovací prostředky musejí být posléze slovně okomentovány. Tabulka a graf bez statistického vysvětlení (okomentování, popisu) mohou být totiž nestatisticky nepřesně pochopeny. Výsledek je pak tristní – statistika lže.

Pokud máme porovnat oba statistické vyjadřovací prostředky, pak tabulka nám předkládá přesná čísla, přesně naměřené či pozorovaní hodnoty. Jedná se o prvotní, primární statistický prostředek. Na druhou stranu tabulky mohou být nepřehledné a zvláště pro statistiky nečitelné. Jen zkušený statistik může v tabulce rozpoznat z uvedených hodnot nějaké souvislosti.

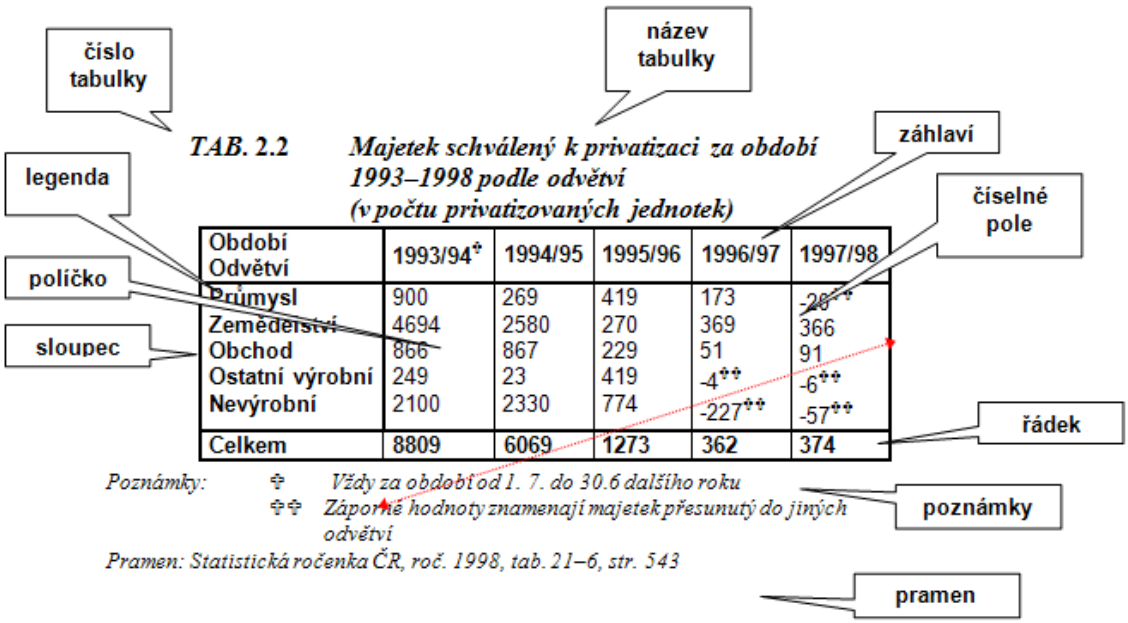
Naproti tomu graf dokáže vystihnout a vyzdvihnout základní charakteristiky – na první pohled je vidět nejmenší a největší hodnota, správně zkonstruovaný graf dokonce dokáže podat informace o dynamice zkoumaného jevu, o rozdělení a také o vzájemném vztahu dvou či více jevů (proměnných). Toto v tabulce lze objevit velice obtížně či to je dokonce nemožné. Na druhou stranu z grafu stěží vyčteme konkrétní a přesné naměřené (pozorované) hodnoty. Poznamenejme, že z jedné tabulky lze zkonstruovat několik grafů, či naopak konstruujeme graf z dat z více tabulek (které jsou ovšem v jistém vztahu).

Z výše popsaného nám vyplývá, že statistická analýza se neobejde bez obou vyjadřovacích prostředků – tabulky a grafu.

4.2 Tabulka

S tabulkami se setkáváme dennodenně a tak je pravděpodobné (resp. zcela jasné), že o tabulkách se nedozvíme nic nového. Podívejme se na standardní příklad statistické tabulky:

4.2.1 Popis statistické tabulky



TAB. 2.2 Majetek schválený k privatizaci za období 1993–1998 podle odvětví (v počtu privatizovaných jednotek)

Období Odvětví	1993/94*	1994/95	1995/96	1996/97	1997/98
Průmysl	900	269	419	173	-26**
Zemědělství	4694	2580	270	369	366
Obchod	866	867	229	51	91
Ostatní výrobní	249	23	419	-4**	-6**
Nevýrobní	2100	2330	774	-227**	-57**
Celkem	8809	6069	1273	362	374

Poznámky: * *Vždy za období od 1. 7. do 30.6 dalšího roku*
** *Záporné hodnoty znamenají majetek přesunutý do jiných odvětví*

Pramen: Statistická ročenka ČR, roč. 1998, tab. 21–6, str. 543

Obrázek 2 - Statistická tabulka

Každá statistická tabulka musí obsahovat název, ze kterého je zřejmá náplň (obsah dat). Nevhodným názvem je příliš obecný „Tržby“, adekvátním názvem by byl „Měsíční tržby provozovny FINGON za zimní sportovní oblečení v roce 2010“. Vedle názvu tabulky bychom měli uvést i její číslo, protože v textu se můžeme na číslo tabulky snadněji odvolávat než na celý název (název můžeme při korektuře a finální úpravě měnit, číslo spíše nikoli).

Tabulka musí obsahovat hlavičku a legendu – popis sloupců a řádků. Poznamenejme, že pouze hlavička (popis sloupců) může obsahovat i symboly pro jednotky (ks, Kč, kg atd.). Nikde jinde se v tabulce (a už vůbec ne v jednotlivých polích) nesmí znak jednotky vyskytovat (Kolektiv, 1967)!

Pod statistickou tabulkou bychom měli uvést zdroj či pramen, odkud jsme data získali. V úvahu přichází odkaz na literaturu (knihu, časopis), ověřený zdroj na internetu (http adresa), případně vlastní výzkum (pozorování).

Poznámky pod tabulkou mohou odkazovat na sloupec či jednotlivá data a slouží k vysvětlení některých nekonzistentností (jinak měřená data a podobně).

Statistická tabulka by měla obsahovat agregační (součtový či průměrový) řádek (za jednolité sloupce) a mnohdy také součtový sloupec (součty za jednotlivé řádky).

Data ve statistické tabulce (jednotlivé buňky) by měly být správně formátované – nejlépe čísla zarovnána doprava se stejným počtem desetinných míst, aby jednotlivé řády byly pod sebou. Zcela nevhodné je zarovnání sloupců na střed. (I náš ukázkový příklad není ideální, zarovnání čísel doleva není vhodné – jednotlivé řády nejsou pod sebou a chybí tudíž pseudografická představa o řádu jednotlivých čísel).

Rovněž bychom se měli vyvarovat použití barev ve statistických tabulkách.

4.2.2 Statistické symboly v tabulkách

Statistická tabulka používá i některé speciální znaky, které pro nezasvěceného mohou být matoucí. Jedná se o tyto čtyři symboly (Kolektiv, 1967):

×	(ležatý křížek)
–	(ležatá čárka)
.	(tečka)
0	(nula)

- ležatý křížek se používá v případě, kdy vyplnění políčka by bylo nelogické. Standardním příkladem je součtový řádek – tedy součet za sloupec, který nelze sčítat (nelze sčítat hrušky a jablka v kusech, nelze sčítat kvalitativní údaje – barva vlasů atd.)
- ležatá čárka představuje nulový údaj, žádný případ, žádný výskyt (pokud měříme skok do dálky, pak nule ležatá čárka znamená, že sportovec tuto disciplínu neabsolvoval)
- tečka identifikuje neznámý, nespolehlivý údaj (například při měření výšky studenta bylo naměřeno 1024 cm, což je zcela jistě chybné měření a nelze s ním dále počítat)

- nula neznamená nulový údaj (viz znak ležatá čárka), ale znamená méně než polovinu zvolené měrné jednotky. Pokud zjišťujeme hrubý měsíční příjem a jednotkovou je tisíc Kč (zapisujeme tedy celé tisíce) a jeden dotazovaný odpoví, že jeho čistý měsíční příjem je 430 Kč, pak do tabulky zaznamenáme 0, protože 430 je méně než 500, což je polovina měrné jednotky (polovina 1 000).

4.2.3 Druhy tabulek

Na konci výkladu o statistických tabulkách je nutno poznamenat, že vedle statistických tabulek, pro které platí výše uvedené poznámky, se lze setkat s dalšími typy tabulek (Cyhelský, 1981).

Tabulky **prezentační** slouží pro prezentaci (pro prezentaci, dat, výsledků zpracování nebo výsledků analýzy) a to buď pro úzký okruh konkrétních příjemců (management, a podobně) či naopak pro širokou veřejnost.

Mezi tabulky **pracovní** (např. na listu v MS Excelu) můžeme zařadit pomocné či rozpracované tabulky, které slouží pouze tomu, kdo s daty pracuje, slouží jako podklady pro následné finální statistické tabulky.

Tabulky **důležitých konstant,** které nalezneme například v přílohách statistických učebnic, obsahují hodnoty testových kritérií (jejich kritické hodnoty). Obdobou jsou tabulky matematické či fyzikální.

4.3 Grafy

Grafické znázornění má zcela zvláštní postavení. Příčinou toho je skutečnost, že na rozdíl od „holých" čísel vyjadřuje graf {či grafický obrazec) informaci v přehledné a srozumitelné formě, a to pro nejširší okruh čtenářů. Každý z nás se raději zahledí na graf tempa růstu, než by hledal shodné údaje v tabulkách. Na druhé straně musíme konstatovat, že grafy mají i své nevýhody. Informace v nich obsažená je sice srozumitelná, ale tato srozumitelnost je na úkor přesnosti (Kořínek, 1989).

V současné době se stále hovoří o „informační inflaci". Množství informací, které se nám dostává do rukou, ať již ve sféře pracovní či mimopracovní, se neustále zvětšuje.

Abychom mohli tyto informace plně využít, musíme zefektivnit jejich „konzumaci". Jednou cestou, která se jeví poměrně schůdnou, je cesta zhuštění informací pomocí grafů.

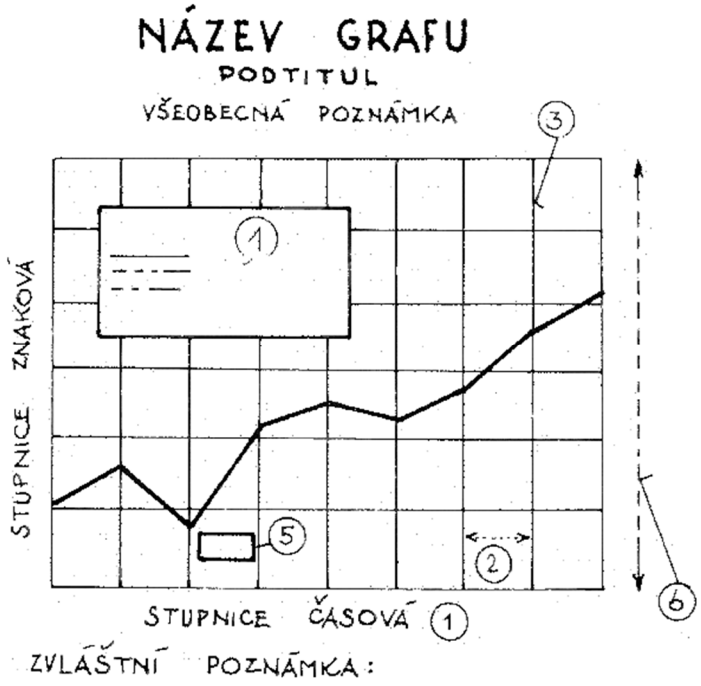
4.3.1 Co to je graf

Řekli jsme si, že proti přesnosti tabelovaných údajů stojí názornost grafů. Vhodně sestavený graf umožňuje v podstatě „jediným pohledem" zhodnotit charakter (tj. úroveň, dynamiku, strukturu, variabilitu) i vzájemný vztah a závislost zobrazených údajů. Větší pracnost grafických metod se vyrovnává větší srozumitelností i rozšířením výkladu.

Co to vlastně graf je? Jedna z mnohých definic praví, že „graf je kresba provedená podle určitých dohodnutých a uznávaných pravidel a zobrazující určitou kvantitativní či kvalitativní informaci" (Kolektiv, 1967). V širším smyslu lze grafem chápat i různé informační a výstražné značky, plány a schémata, využívající různé kresebné techniky a prostředky. My se budeme zabývat grafy v tom užším pojetí.

4.3.2 Části grafu

Základní prvky grafu vidíme na následujícím obrázku (Kořínek, 1989).



Obrázek 3 - Prvky grafu

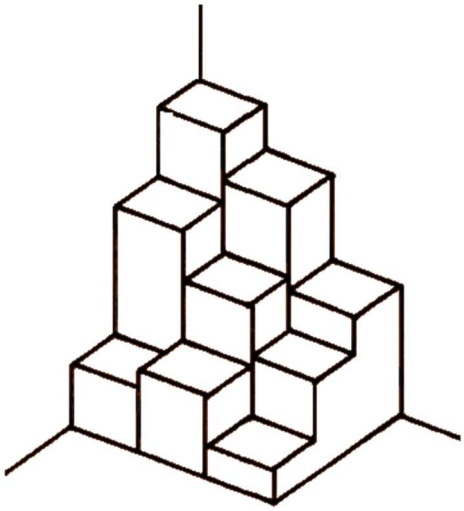
- stupnice a souřadnicové osy včetně adekvátních názvů,
- grafický interval (měřítko),
- klíč,
- vysvětlivky,
- délku stupnic.

V některých případech (zvláště jde-li o ustálený typ grafu, který se opakuje ve stejné podobě), mohou některé části chybět.

Stupnice mohou být přímočaré, křivočaré (podle nositelky stupnice, tj. čáry), nebo rovnoměrné a nerovnoměrné (podle úměrnosti grafických intervalů).

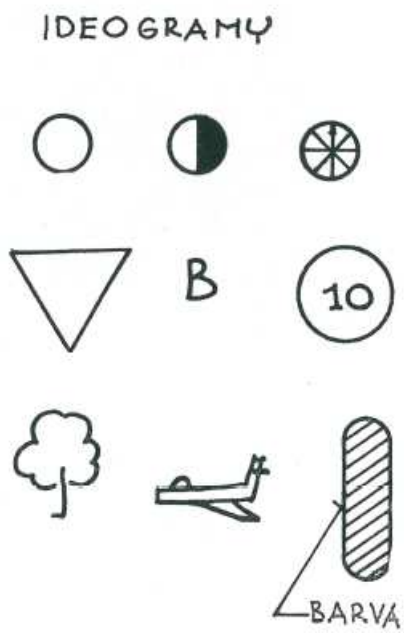
U systémů souřadnic (souřadnicových os) se používají především tyto tři druhy (Kolektiv, 1967):

1. **Soustava pravoúhlých souřadnic**, nazývaná někdy kartézská či Descartova, je nejznámějším souřadným systémem. Horizontální osa se nazývá osou úseček (*x*) a vertikální osa (*y*) osou pořadnic. Celé polesouřadnic se dělí na čtyři kvadranty, přičemž se nejčastěji používá první kvadrant. Polohu libovolného bodu, kterou vyjadřujeme vždy dvojicí čísel, určíme délkou kolmic k ose úseček a k ose pořadnic.
2. **Soustava polárních souřadnic** (též radiální či paprskovitá) je vhodná především pro grafické znázorňování periodicky se opakujících jevů a pro některé úkoly zkoumání struktury. Poloha bodu, která je opět vyjádřena dvojicí čísel, je dána nejkratší přímou vzdáleností od daného pólu a nejkratší vzdáleností po oblouku kružnice (se středem v pólu) od hlavního polopaprsku ve směru hodinových ručiček. Vzdálenosti od pólu se říká průvodič (poloměr, radiusvektor) a obloukové vzdálenosti od hlavního paprsku odchylka (úhel, amplituda).
3. **Soustava prostorových souřadnic**umožňuje vzhledem k trojrozměrnosti prostoru vyjádřit jediným bodem tři veličiny vztahující se k určité jednotce. K dvěma souřadným osám *x* a *y* přibude třetí (*z*), která s oběma předchozími svírá pravý úhel. Grafy v takovém systému nazýváme stereogramy (viz následující obrázek). Protože pracnost sestrojení stereogramů je odstraněna počítačovou technikou, používají se v literatuře čím dál tím častěji.



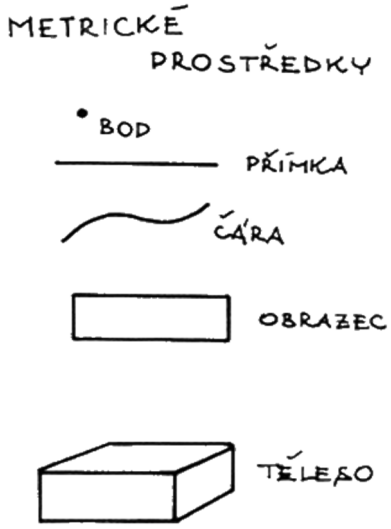
Obrázek 4 - Trojrozměrný graf (stereogram)

Souřadné osy a stupnice, jíž samy o sobě určují grafický Interval. Klíčem se rozumí výklad, který nám umožní, abychom se mohli v grafu orientovat a nebyl pro nás „španělskou vesnici". Součástí klíče je přehled použitých ideogramů a grafických prvků s metrickým významem. Ideogramy jsou grafické prostředky mající kvalitativní symbolický význam a umožňující tedy rozlišovat nebo naopak agregovat (shrnovat) jednotlivé jednotky.



Obrázek 5 - Ukázka ideogramů

Grafické prvky s metrickým významem mají, jak již jejich název proklamuje, kvantitativní význam. Ten se prokazuje např. počtem, rozměrem či polohou.



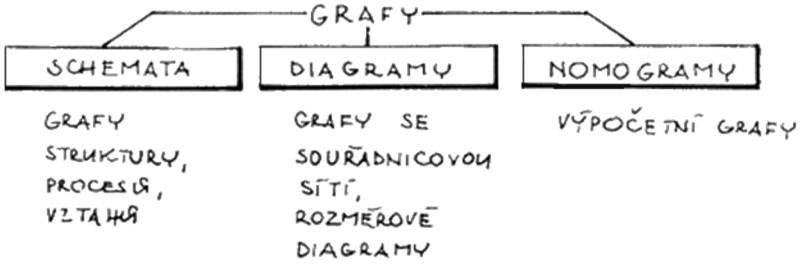
Obrázek 6 - Ukázka metrických prvků

Podíváme-li se na oba předešlé obrázky (a máme-li jistou dávku obrazotvorností), napadne nás velice správná myšlenka, že rozdíl mezi ideogramy a metrickými prostředky je pouze podmíněný: v jednom grafu může být ídeogramem to, co v druhém metrickým prostředkem.

U složitějších grafů má být obsahem klíče i způsob jejích čtení.

4.3.3 Klasifikace grafů

Grafy (grafická znázornění) dělíme do tří skupin. Jsou to schémata, diagramy a nomogramy (Kolektiv, 1967).



Obrázek 7 - Klasifikace grafů

Schémata vyjadřují různé struktury a vztahy znázorňovaného jevu (povahy spíše kvalitativní). Znázorňují tedy umístění daného jevu (čí procesu) v určité věcné, prostorové číičasové soustavě. V širším slova smyslu to mohou být i grafy znázorňující vztahy pojmů. Mezí schémata patří:

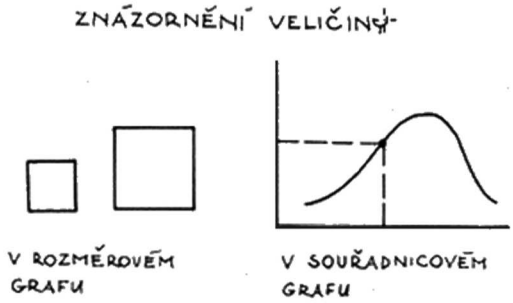
- schémata věcných soustav (klasifikační, organizační),
- schémata prostorových soustav (topogramy),
- schémata časových soustav (chronogramy a harmonogramy).

Namísto soustavy souřadnic (jelikož se nezobrazuje kvantita, ale kvalita) užívají schémata zonální soustavu. Ta člení graf na jednotlivé části podle struktury celku a pomáhá identifikovat jednotlivé grafické prvky.

Nomogram je výpočetní graf nebo soustava výpočetních grafů, které slouží k přibližnému určení hodnoty určitého algebraického výrazu pro danou kombinací proměnných, Nomogram se skládá ze soustavy čar a stupnic, které umožňují grafický výpočet. Způsob čtení bývá určen klíčem. Rozlišujeme nomogramy průsečíkové a spojnicové.

Diagram je druh grafu, který využívá pro orientací grafického prvku soustavy souřadnic, nebo grafického měřítka. Podle tohoto kritéria se diagramy člení na rozměrové a souřadnicové.

O rozměrovém diagramu (viz následující obrázek) hovoříme tehdy, je-li kvantitativní charakteristika znázorněna poměrnou velikostí grafického obrazu, tj. jeho rozměry.



Obrázek 8 - Rozměrový diagram

Velikost zobrazované veličiny se určí podle velikosti grafického prvku. Pro znázornění jednotlivých jevů se nejčastěji používá plošných útvarů (čtverec, obdélník, kruh, trojúhelník). V tomto případě jim říkáme plošné grafy, jejích výhody (Kořínek, 1989):

- použití různých obrazců zpestřuje výklad,

- plochy je možno zřetelně členit a z toho vyplývá možnost zobrazovat í strukturu daného jevu a její změny.

K nevýhodám pak můžeme zařadit:

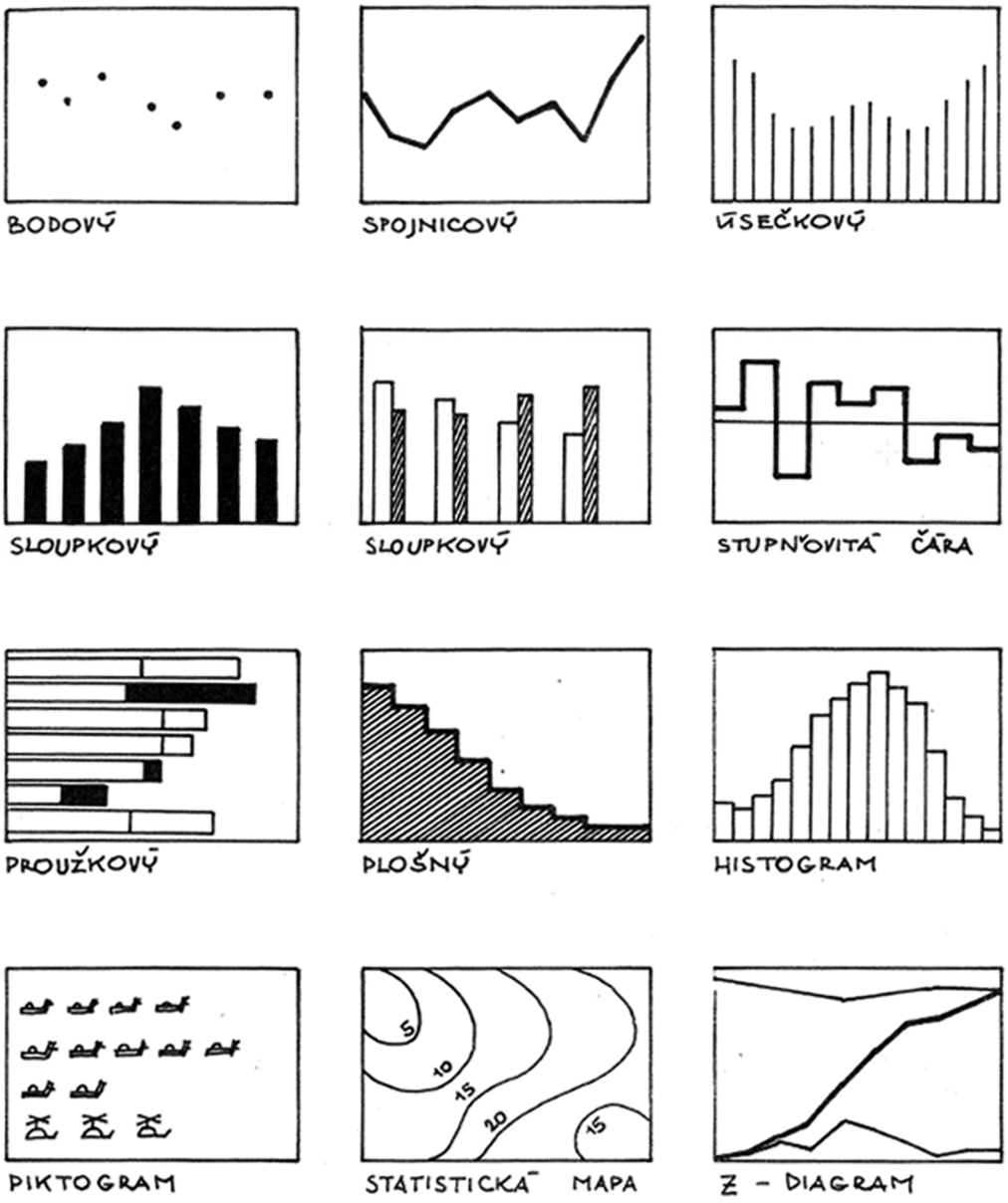
- větší prostorová náročnost ve srovnání s diagramy souřadnicovými, které jsou „hutnější",
- obvykle pracnější konstrukce, proto nejsou vhodné pro pracovní či analytické grafy,

Obecně lze velikost každého plošného obrazce (pravidelného) charakterizovat bud jeho základními rozměry (strany čtverce, poloměr kruhu, výška trojúhelníka), nebo jeho plochou. Podle užitého grafického prvku se určuje i název grafu (Cyhelský, 1981):

- kruhový graf
- čtvercový graf,
- trojúhelníkový graf

Mezi neznámější strukturální rozměrové grafy patři výsečový diagram a procentní čtverec.

V souřadnicovém diagramu je kvantitativní charakteristika určena polohou v soustavě souřadnic. Základním typem souřadnicového diagramu je bodový diagram, který lze vhodnou volbou grafických prvků rozvinout v další známé typy diagramů, jako jsou: spojnicový, čárový, úsečkový, sloupcový U diagramu rozdělení četností se spojnicový diagram nazývá polygonem a sloupcový diagram histogramem. Na další obrázku si některé výše jmenované diagramy můžeme prohlédnout (Kořínek, 1989).



Obrázek 9 - Některé typy diagramů

Některé diagramy mají zvláštní názvy, např. regulační diagram, křivka koncentrace, Z diagram (Cipra, 1986).

5 Úvod do deskriptivní statistiky

Při statistickém zpracování dat je nutné většinou zpracovat velký (či alespoň větší) objem vstupních údajů, vstupních dat, přičemž se může jednat jak o data kvantitativní ale i o data kvalitativní (či pochopitelně jejich kombinace).

Cílem statistické analýzy dat získat souhrnné údaje (statistiky) sledovaném jevu a to tak, abychom na základě těchto výsledků lépe porozuměli řešenému problému. „*Porozumění vzniká z kombinace znalostí o kontextu, jak data vznikla, a schopnosti využít statistické grafy a numerické výpočty*.“ (Hendl 2006, str. 85).

V předešlé kapitole jsme si naznačili, že pro statistický popis dat se používají dva základní vyjadřovací prostředky:

- statistická tabulka a
- statistický graf.

Obecně lze konstatovat, že statistikou analýzu lze rozvrhnout do následujících etap (Maněnová, 2012):

1. Nejprve zobrazíme data pomocí tabulek a grafů.
2. Vypočítáme základní charakteristiky měr polohy a variability.
3. Zkusímevystihnout základní tendence v datech, případně odchylky od těchto tendencí.
4. Můžeme použít pravděpodobnostní model, který vystihne stručným způsobem základní konfiguraci dat.

Tato kapitola je věnovaná úvodu do deskriptivní statistiky, a proto se budeme zabývat prvními dvěma body. Znamená to, že se seznámíme s následujícími statistickými procedurami:

- uspořádání dat a sestavení tabulek četností,
- grafické znázornění naměřených dat,
- výpočet charakteristik polohy,
- výpočet charakteristik rozptýlení (měr variability).

5.1 Uspořádání dat a sestavování tabulek četností

5.1.1 Čárkovací metoda

Po získání dat je naším prvním úkolem je (statisticky) utřídit. Jedním ze standardních postupů je čárkovací metoda, při které nejdříve zapíšeme do prvního sloupce zleva všechny různé obměny znaku, jichž bylo při měření dosaženo. Jednotlivé hodnoty přitom uvádíme nejčastěji seřazené podle velikosti (od nejmenší po největší). Poté procházíme jednotlivá pozorování (konkrétní hodnoty) a pomocí čárek zaznamenáváme jejich výskyt do druhého sloupce (Sharma, 2005).

Názorněji si to předvedeme na jednoduchém příkladu[16]. Skupině 25 žáků byl předložen didaktický test. Jednotliví žáci získali následující počty bodů: 16, 16, 15, 17, 16, 17, 18, 15, 16, 17, 17, 18, 16, 20, 16, 18, 20, 16, 17, 20, 19, 19, 18, 16, 17.

Nyní pomocí čárkovací metody sestavíme tuto tabulku:

Tabulka 1 - Čárkovací metoda

hodnota	
15	||
16	|||||||
17	|||||
18	||||
19	||
20	|||

V našem pozorování se vyskytovaly pouze hodnoty od 15 do 20 (sloupec vlevo). Na první pohled je zřejmé, že hodnota 16 se vyskytovala nejčastěji (nejvíce čárek je právě u této hodnoty). Sloupec vpravo vyjadřuje počet výskytů dané konkrétní obměny znaku a počet čárek pak charakterizuje **absolutní četnost**.

Poznamenejme si, že čárkovací metodu použijeme pouze v případě, že data zpracováváme ručně. Pokud použijeme odpovídající software, pak program nám automaticky tabulku s četnostmi vygeneruje. Na druhou stranu jsme si tento postup

[16]Příklad převzat z (Maněnová 2012)

z metodologického hlediska uvedli, protože demonstruje, jak se z prvotních dat k tabulce četnosti dospěje.

5.1.2 Tabulka rozdělení četností

Z tabulky výše uvedené tabulky nejprve vytvoříme tabulku rozdělení četností, což znamená, že čárky nahradíme číslem, počtem čárek.

Tabulka 2 - Absolutní četnosti

hodnota	absolutní četnost
15	2
16	8
17	6
18	4
19	2
20	3
Celkem	25

Tabulku jsme také doplnili o součtový řádek, který mimo jiné slouží i jako kontrola – je vidět, že jsme celkem pozorovali 25 výskytů (počet zkoumaných žáků bylo 25). Poznamenejme, že celkovou četnost nalezneme v řádku Celkem a označujeme ji *n*. Jednotlivé absolutní četnosti pak značíme n_i. Platí následující vztah – součet všech četností je roven celkové četnosti (součet všech četností v jednotlivých řádcích je roven celkovému počtu pozorování) (Sharma, 2005):

$$\sum_{i=1}^{k} n_i = n \quad (1)$$

Obvykle jsou absolutní četnosti doplněny o **četnosti relativní**, které charakterizují podíl dané skupiny na celku (jak velká část z celkového počtu hodnot připadá na danou hodnotu, kategorii).

Relativní četnost f_i vypočítáme dle následujícího vzorce (Cyhelský, 1981):

$$f_i = \frac{n_i}{n} \quad (2)$$

Naší tabulku doplníme o sloupec relativní četnosti.

Tabulka 3 - Relativní četnosti

hodnota	absolutní četnost	relativní četnost
15	2	0,08
16	8	0,32
17	6	0,24
18	4	0,16
19	2	0,08
20	3	0,12
Celkem	25	1,00

Pro objasnění si odpovíme dvě otázky: Jak se vypočítá hodnota 0,24 a co toto číslo znamená (relativní četnost u řádku 17). Relativní četnost vypočítáme jako podlá absolutní četnosti dané skupiny a celkové četnosti, celkového počtu pozorování. V tomto případ se tedy jedná o podíl čísel 6 a 25 (6/25 = 0,24). Relativní četnosti je možno také vyjádřit v procentech, vypočítaná hodnota f se tedy vynásobí 100%. Číslo 0,24 znamená, že podíl žáků, kteří při testu získali 17 bodů je 24%.

Tabulku absolutních četností a relativních četností vytváříme pro téměř všechny případy – jak číselné hodnoty (viz náš příklad), tak i pro hodnoty kvalitativního znaku (například barva vlasů, členství v politické straně atd.).

Z formálního hlediska je důležité upozornit, že čísla ve sloupci relativní četnost by měla mít stejný počet desetinných míst, aby řády byly zarovnány pod sebou. Stejným porušením je situace, kdy relativní četnost je uváděna v procentech a znak procento je uveden u každého čísla. Zopakujme si, že jakýkoli symbol (a tudíž i znak procento) se v jednotlivých buňkách neuvádí, lze jej uvést pouze v hlavičce.

Nakonec upozorňujeme, že součet relativních četností musí být vždy roven 1 (či v případě procentního vyjádření roven 100). Jakákoli jiná hodnota je nepřípustná (nelze proto konstatovat, že součet relativních četností je 99.9 vzhledem k zaokrouhlování). Konec konců i toto můžeme zařadit mezi kontrolní mechanismy tabulky rozdělení četností.

Kumulované četnosti

V mnohých případech je vhodné doplnit naší tabulku i o kumulované četnosti. Upozorněme, že kumulativní četnosti lze uvádět pouze v případech, kdy jednotlivé obměny sledovaného znaku můžeme jednoznačně seřadit. V našem případě lze znak počet

získaných bodů seřadit podle velikosti. Znak počet dětí v rodině lze také seřadit, naopak kvalitativní znak barva očí jednoznačně seřadit nelze, a proto kumulativní četnosti ztrácejí logiku.

Kumulativní četnost je součet četnosti v určitém řádku tabulky a četnosti ve všech předchozích řádcích dohromady (Rumsey, 2007).

Absolutní kumulovaná četnost (kn_i)

n_1

n_1+n_2

$n_1+ n_2+n_3$

…..

$n_1+n_2+ n_3+\ldots+n_n$

Relativní kumulovaná četnost (kf_i):

f_1

f_1+f_2

$f_1+f_2+f_3$

…..

$f_1+f_2+ f_3+\ldots+f_n$

Naší tabulku doplníme o obě kumulované četnosti:

Tabulka 4 - Kumulované četnosti

hodnota	absolutní četnost	relativní četnost	kn_i	kf_i
15	2	0,08	2	0,08
16	8	0,32	10	0,40
17	6	0,24	16	0,64
18	4	0,16	20	0,80
19	2	0,08	22	0,88
20	3	0,12	25	1,00
Celkem	25	1,00	X	X

Opět si pro objasnění sloupce kumulovaná absolutní četnost (kn_i) odpovězme na tyto dvě otázky: Jak se vypočítá číslo 16 a co nám říká? Kumulativní četnost je součtem absolutních četností do daného řádku včetně – tedy 16 = 6 + 8 + 2. A tato hodnota nám říká, že 16 žáků z celkového počtu 25 získalo nejvýše 17 bodů (tedy 15, 16 či maximálně 17).

U kumulované relativní četnosti (*kf)* je to zcela podobné. Například číslo 0,8 nám říká, že 80% všech sledovaných žáků získalo maximálně 18 bodů a vypočítá se jako součet relativních četností (0,15 + 0,24 + 0,32 + 0,08).

Intervalové rozdělení četností

Pokud byl při měření získán velký počet různých hodnot (například výška postavy, váha postavy či při měření spojitých veličin – čas a podobně), pak tabulka četností dle předchozího příkladu bude obsahovat příliš velký počet řádků a bude nepřehlednou. Navíc z takové tabulky se statistik žádné důležité informace nedozví.

V takovýchto případech se získaná data seskupují do **intervalů**, do uměle vytvořených skupin (Chráska, 2008). Při tvorbě těchto intervalů musíme rozhodnout o jejich počtu, o mezích těchto intervalů a o jejich velikosti (šířce).

Pro tvorbu intervalů platí několik zásadních pravidel (Cyhelský, 1981):

- Preferujeme intervaly (třídy) s konstantní šířkou,
- počet tříd (intervalů) koresponduje s rozsahem souboru a je v rozmezí 6 až 15,
- šířku, hranice a středy tříd volíme s ohledem na maximální přehlednost,
- vždy je bezpodmínečně nutné nesporné vymezení hranic tříd (mezí intervalů),
- první a poslední třída (interval) mohou být otevřené.

Počet intervalů je závislý na celkovém počtu pozorování a lze použít tyto odhady:

Tabulka 5 - Odhad počtu intervalů s ohledem na velkost souboru

n	**počet intervalů**
< 50	5-6
50-100	6-8
> 100	8-10

Poznamenejme, že počet intervalů lze přibližně stanovit podle celé řady empirických vzorců, např. Sturgesovým pravidlem (Chráska, 2008):

$$k = 1 + 3,3 \cdot \log n \quad (3)$$

čí přibližným odhadem (Cyhelský, 1981)

$$k \approx \sqrt{n} \quad (4)$$

kde *k* je počet intervalů a *n* je počet zjištěných údajů (počet pozorování). Šířku intervalu pak získáme ze vztahu (Chráska, 2008):

$$h \approx 0,08 \cdot R \quad (5)$$

kde *h* je šířka intervalu a *R* je variační šíře vypočtená jako $x_{max} - x_{min}$ (tedy rozdíl mezi největší a nejmenší naměřenou hodnotou).

Všechny výše uvedené výpočty je nutno zaokrouhlit na vhodné celé číslo, desetiny či setiny dle povahy zpracovávaných dat.

Pokud použijeme statistický software, tak nutnost definovat jednotlivé skupiny (intervaly) odpadá, neboť program vygeneruje dle pravidel adekvátní skupiny automaticky, případně můžeme návrh upravit. Výše uvedený postup jsme si ukázali, abychom lépe pochopily, proč a jak data správně statisticky zpracovat.

Je zřejmé, že takto uměle tvoříme skupiny pouze u číselných dat, u kvalitativních znaků se umělé skupiny tvořit nedají (pouze slučováním „podobných“ hodnot znaků – například při definování hodnoty barva očí bychom mohli odpověď modrá, tmavě modrá a šedomodrá spojit do skupiny modrá).

Pro konstrukci intervalového rozdělení četností si ukážeme následující příklad. Ve vybrané třídě jedné základní školy bylo provedeno měření výšky žáků (v cm). Obdrželi jsme následující hodnoty:144, 149, 145, 142, 146, 147, 141, 150, 143, 146, 150, 141, 148, 148, 144, 141, 145, 148, 144, 143, 155, 133, 158, 154, 151, 140, 136, 137, 153, 139, 138[17].

[17]Příklad převzat z CYHELSKÝ, L., HUSTOPECKÝ, J., ZÁVODSKÝ, P.: *Příklady k základům statistiky*. Praha,

Naším úkolem je naměřená data uspořádat do tabulky.

Na první pohled je zřejmé, že existuje mnoho variant obměn znaku a budeme muset proto vytvořit umělé skupiny, vytvořit intervalové rozdělení četností.

Dle výše uvedených vzorců bychom vypočetli následující charakteristiky:

n = 31 (počet pozorování)

max = 158 (nejvyšší hodnota)

min = 133 (nejnižší hodnota)

R = 25 (variační rozpětí)

K = 5,92149359 (počet intervalů dle $\boldsymbol{k = 1 + 3{,}3 \cdot \log n}$(3))

l = 2 (šířka intervalu dle $\boldsymbol{h \approx 0{,}08 \cdot R}$(5))

K vytvoření jednotlivých intervalů nám však postačí i rychlý odhad o počtu intervalů (viz Tabulka **5** - Odhad počtu intervalů s ohledem na velkost souboru) a dále nad tabulkou uvedená pravidla.

Po úvaze se rozhodneme pro šířku intervalu 5cm a meze tak, aby se jednalo o celá čísla. První interval bude polootevřený s horní mezí 135, druhý interval bude rovněž polootevřený s mezemi 135 a 140, tedy zápis (135;140> atd. Tabulka intervalového rozdělení četností bude vypadat takto:

Tabulka 6 - Intervalové rozdělení četností

skupiny	n_i	f_i	kn_i	kf_i
135	1	0,03	1	0,03
140	5	0,16	6	0,19
145	11	0,35	17	0,55
150	8	0,26	25	0,81
155	5	0,16	30	0,97
160	1	0,03	31	1,00
součet	31	1,00	x	x

Ve sloupci skupiny jsou uvedeny pouze horní meze, v dalších sloupcích jsou pak postupně absolutní četnost, relativní četnost, kumulovaná absolutní četnost a kumulovaná relativní četnost. Výpočty a komentování je shodné s příkladem uvedeným výše (viz

SNTL/ALFA,1988. 04-317-88.

Tabulka **4** - Kumulované četnosti). Tedy například číslo 0,55 v posledním sloupci nám říká, že přes 50 % žáků měří do 145 cm.

5.2 Grafické znázornění naměřených dat

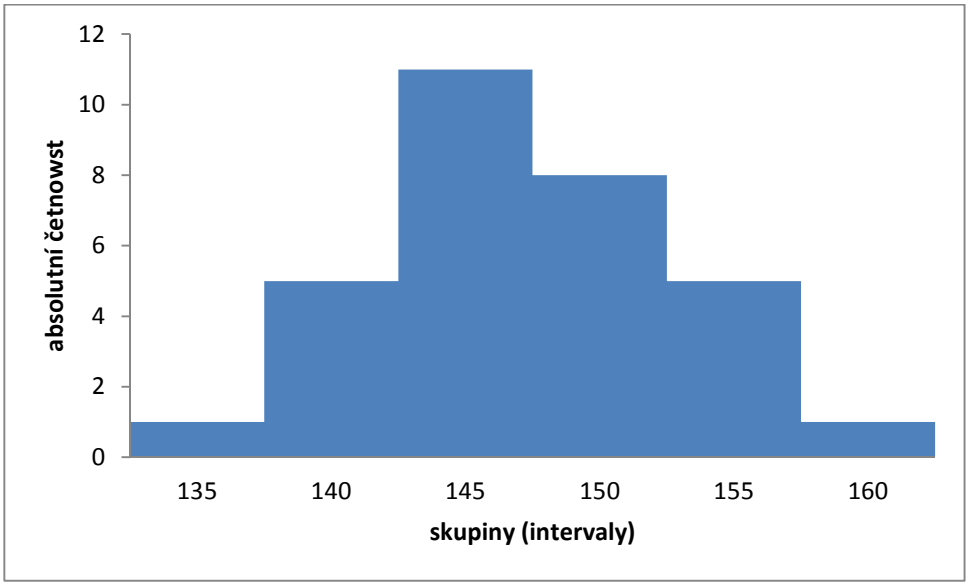
Data a výpočty z tabulky rozdělení četností je vhodné prezentovat také v názorné grafické podobě. K tomuto účelu se používají histogramy, polygony četnosti, kumulativní křivky a výsečové (koláčové) grafy.

5.2.1 Histogram

Histogram je sloupcový graf, kdy na (vodorovné) ose *x* jsou uvedeny jednotlivé obměny znaku a na ose (svislé) *y* pak četnosti, případně relativní četnosti (Kolektiv, 1967).

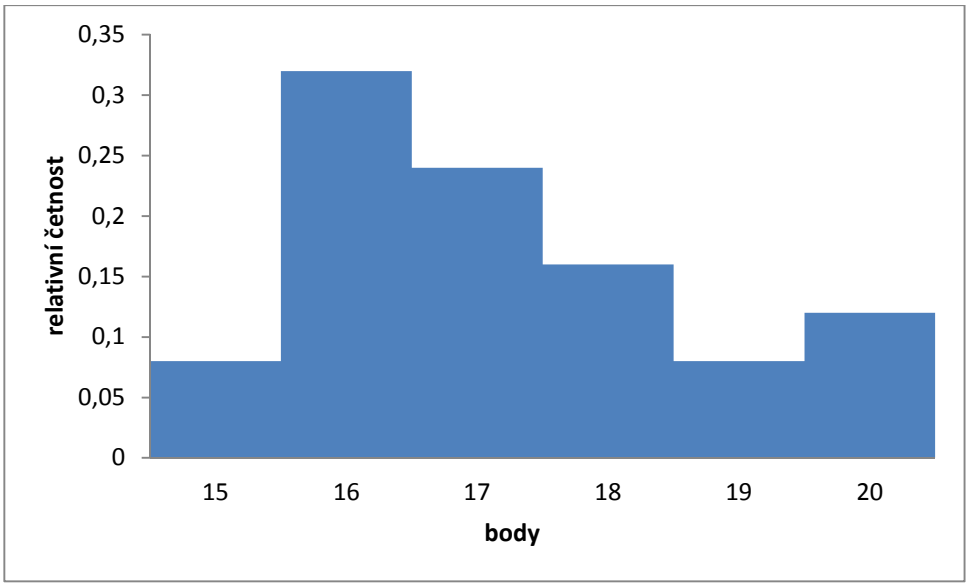
Histogram lze použít pochopitelně i pro zobrazení intervalového rozdělení četností. Histogramem lze zobrazit i kumulované četnosti.

První příklad ukazuje histogram absolutních četností našeho druhého příkladu



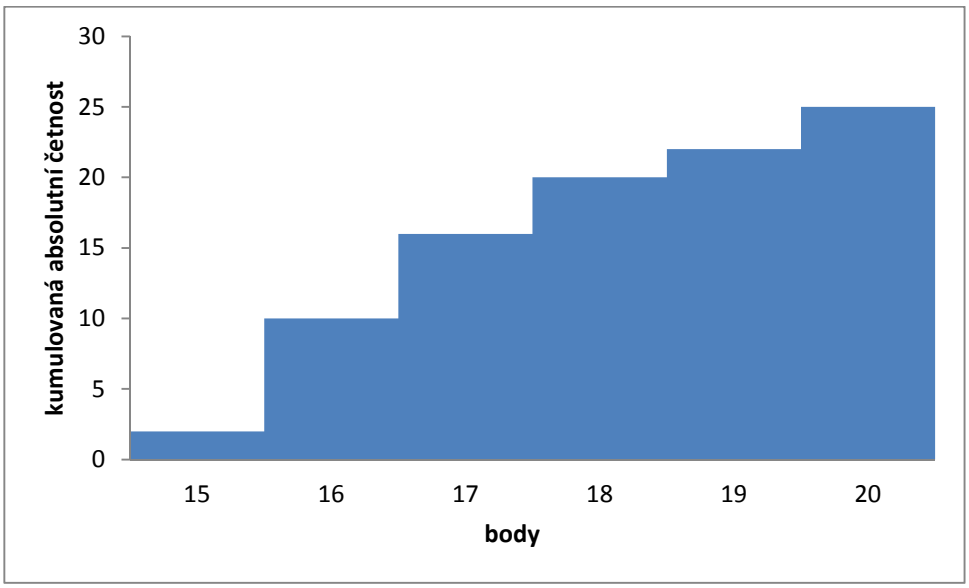
Obrázek 10 - Histogram intervalových absolutních četností

Druhý obrázek pak představuje histogram relativních četností pro první příklad.



Obrázek 11 - Histogram relativních četností

Jak bylo již řečeno, pomocí histogramu lze zobrazit i kumulované četnosti – náš příklad ukazuje kumulované absolutní četnosti našeho prvního příkladu.



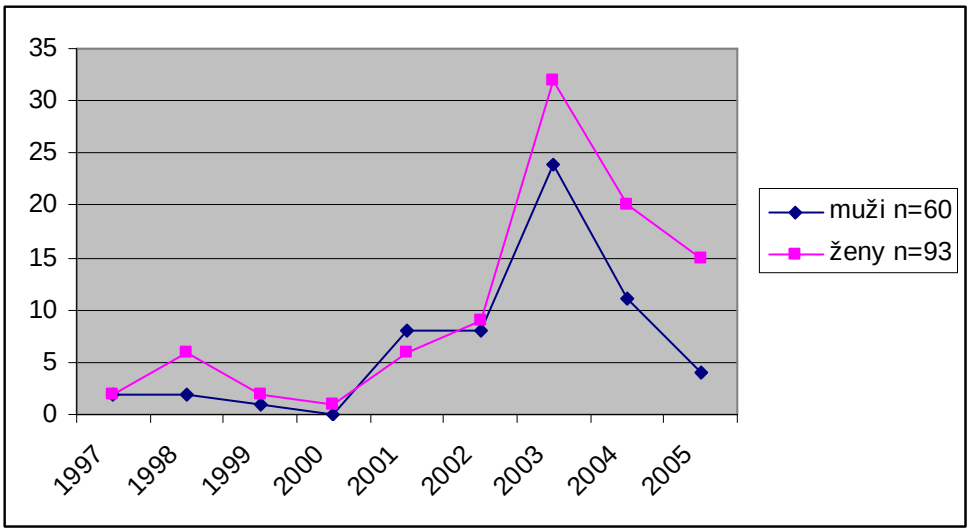
Obrázek 12 - Histogram kumulované absolutní četnosti

Uvědomme si, že histogram je nejobecnějším grafem a lze ho použít téměř vždy, pro každé rozdělení četností, pro kvantitativní ale i kvalitativní data (na ose *x* budou jednotlivé kvalitativní obměny – například barva očí modrá, zelená, hnědá atd.).

Poslední poznámka je k zobrazování absolutních a relativních četností a nevztahuje se k histogramu, ale obecně ke všem grafům. Protože relativní četnost je v podstatě pouze normovaná četnost absolutní, průběh obou grafů (s absolutní a relativní četností) je zcela shodný, rozdíl je pouze v popisu osy *y* – jednou jsou uvedena hodnoty absolutní a podruhé hodnoty relativní. Z toho důvodu si pro grafické zobrazení vybereme pouze jednu variantu (absolutní či relativní) a nikdy nekonstruujeme oba grafy. A tato poznámka pochopitelně platí i pro alternativy graf kumulovaných absolutních versus kumulovaných relativních četností.

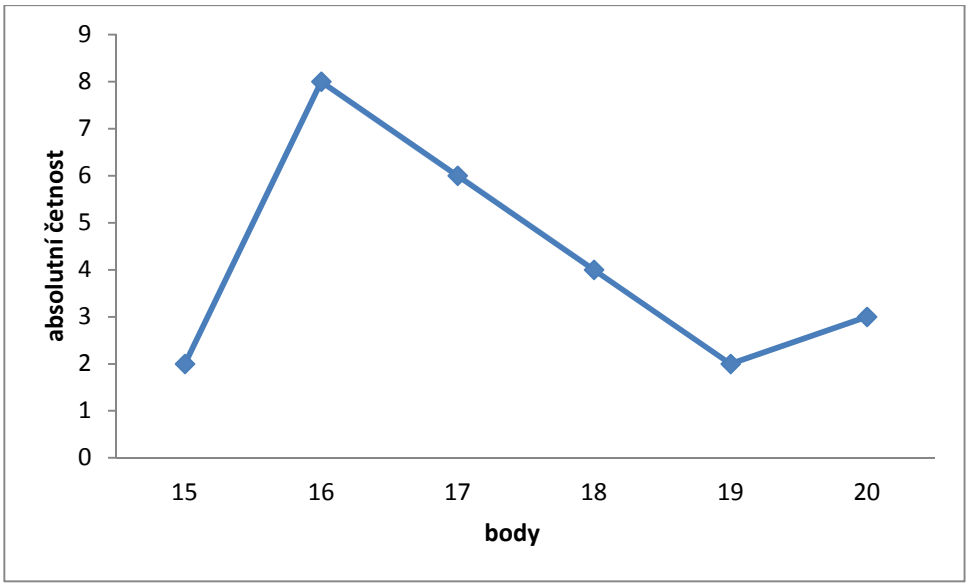
5.2.2 Polygon četnosti

„Polygon četností se liší od histogramu tím, že se jedná o diagram spojnicový. (např. výskyt onemocnění v jednotlivých letech, měsících).“ (Maněnová, 2012)



Obrázek 13 - Ukázka polygonu četností

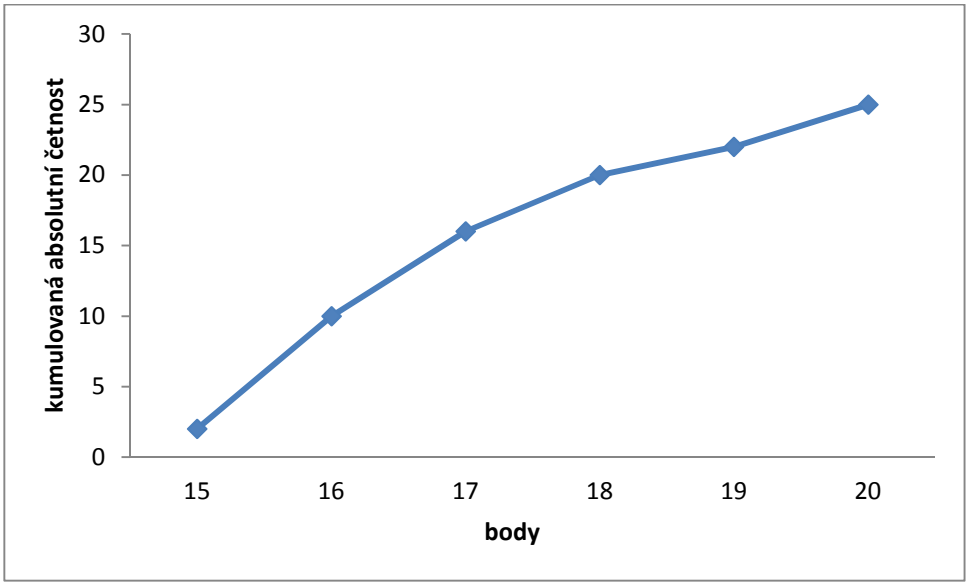
Musíme si však uvědomit, že spojnicový graf evokuje vývoj, dynamiku. Proto lze polygon četnosti použít pouze v případě, kdy na ose *x* je čas – hodnoty roky, měsíce, dny, hodiny a podobně. Případně lze polygon četnosti ještě použít v situaci, kdy lze obměny znaku na ose x jednoznačně seřadit – například naše úloha s bodovým ohodnocením žáků.



Obrázek 14 - Polygon četnosti - spojnicový graf absolutních četností

Nicméně doporučuje se, aby polygon četnosti – spojnicový graf – byl použit výhradně s „časovou“ osou *x* (Vonnacott, 1992). Pro příklad bodového ohodnocení by histogram byl vhodnější.

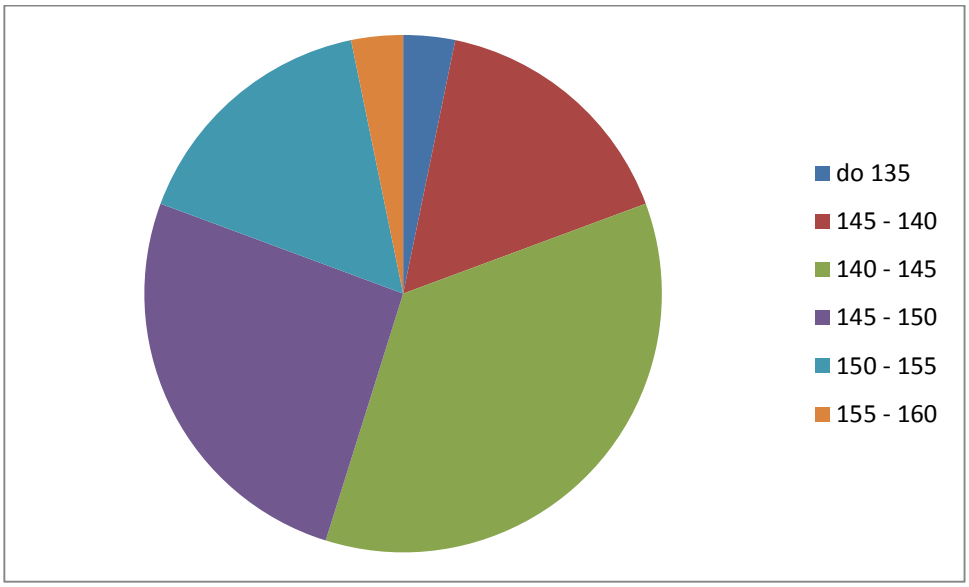
Polygon četnosti v případě, kdy požijeme kumulované četností, bývá označován jako **Galtonova ogiva** (Cyhelský, 1981).



Obrázek 15 - Galtonovaogiva – spojnicový graf kumulované četnosti

5.2.3 Koláčový graf

Koláčový (též výsečový, kruhový případně prstencový) graf je velmi oblíbený a vhodný v případě, že se zobrazuje struktura zkoumaného souboru. Protože celý kruh (koláč) chápeme jako celek (tedy 100%), jednotlivé části znázorňují procentní podíl (Kolektiv, 1967). Z tohoto důvodu je koláčový graf nejvhodnější pro zobrazení relativní četnosti.



Obrázek 16 Koláčový graf - intervalové rozdělení

Náš obrázek znázorňuje graf relativní četnosti k příkladu druhému, k intervalovému rozdělení četností. (viz legenda v pravé části grafu).

Poznamenejme, že se objevují dvě nejčastější chyby v použití koláčového grafu. První spočívá v tom, že k zobrazení se použijí data z kumulativních četností. Toto odporuje logice konstrukce koláčového grafu, neboť některé hodnoty se v koláči (resp. jednotlivých výsečích) objevují vícekrát, což je neodůvodnitelné. Druhá chyba nastane, pokud k zobrazení do koláčového grafu vybereme více jak jednu proměnnou (více sloupců), jeden jev, který zkoumáme. Příkladem budiž tabulka s počty kuřáků v jednotlivých věkových skupinách a to zvlášť pro ženy (sloupec A) a muže (sloupec B). Pokud do koláčového grafu vložíme oba sloupce, pak jedna výseč nám neposkytuje odpovídající správné informace (správně bychom měli vytvořit dva koláčové grafy, první by vyjadřoval věkovou strukturu u kuřáků a druhý věkovou strukturu u kuřaček).

5.3 Charakteristiky polohy

Jak jsme si již řekli, při statistickém zpracování dat potřebujeme všechna naměřená data nějakým způsobem výstižně a stručně charakterizovat (jinak řečeno, potřebujeme vypočítat „statistiku“, jedno číslo, které daný soubor charakterizuje, popisuje).

K popisu polohy neboli středu či těžiště zkoumaného souboru, můžeme použít celou řadu statistik. Mezi nejznámější se řadí **aritmetický průměr**, **medián** nebo **modus** (Cyhelský, 1981).

5.3.1 Průměr

Aritmetický průměr $\bar{x}$ z číselných hodnot x_1 , x_2 , x_3, … x_n lze vypočítat podle vzorce

$$\overline{\boldsymbol{x}} = \frac{\boldsymbol{x_1+x_2+x_3+..+x_n}}{\boldsymbol{n}} \quad \textbf{(6)}$$

kde *n* je celková četnost všech hodnot (rozsah souboru). Velmi často se používá i zkrácený tvar této rovnice s použitím symbolu „suma“:

$$\overline{\boldsymbol{x}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \boldsymbol{x_i} \quad \textbf{(7)}$$

Výše uvedeným vzorcem budeme počítat aritmetický průměr ze získaných dat, z primárních dat. Pokud všechna vstupní data nemáme k dispozici a známe tedy jen tabulku rozdělení četností, pak musíme použít vážený tvar aritmetického průměru (Čermák, 1968):

$$\overline{\boldsymbol{x}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{k}} \mathbf{n_i} \cdot \boldsymbol{x_i} \quad \textbf{(8)}$$

kde *n* je celková četnost všech hodnot, x_i je určitá hodnota, n_i četnost hodnoty x_i a *k* je počet řádků (skupin, tříd) v tabulce četností.

Pokud jsou v tabulce četností data seskupena do intervalů, potom nejdříve určíme střed každého intervalu a ten potom dosadíme jako hodnotu x_i do příslušného vzorce (v tomto případě se však jedná o „odhad“ aritmetického průměru).

Mezi výhody aritmetického průměru se řadí ta vlastnost, že jeho matematické vyjádření je jednoduché a také, že je velmi důležitý pro odvozování dalších důležitých vztahů (rozptyl, korelace, kovariance, regresní koeficient, metoda nejmenších čtverců atd.).

Je důležité si uvědomit, že, jeho hodnota závisí na všech prvcích souboru dat. Z toho vyplývá i nevýhoda průměru – jeho citlivost k extrémním hodnotám, tj. hodnotám, které se od ostatních značně odchylují. Extrémní hodnoty mohou značně zkreslit hodnotu průměru (obvyklým a zcela správným příkladem je průměrná mzda celé ekonomicky aktivní populace).

Nyní vypočítáme aritmetický průměr u dat z výše uvedeného příkladu, skupina 25 žáků s výsledky testu.Naším úkolem je určit průměrný počet bodů v testu.

Pokud budeme mít k dispozici všechna data, pak stačí jednotlivé výsledky, body, celkem 25 čísel sečíst a poté vydělit celkovým počtem, tedy 25. Aritmetický průměr je pak 17,2 bodů.

V případě, že máme k dispozici pouze tabulku rozdělení četností, k výpočtu musíme použít vážený tvar aritmetického průměru a tabulku doplníme o výpočtový sloupec:

hodnota	absolutní četnost	
x_i	n_i	$x_i * n_i$
15	2	30
16	8	128
17	6	102
18	4	72
19	2	38
20	3	60
Celkem	25	430

Tabulka 7 - Výpočet váženého aritmetického průměru

Aritmetický průměr pak vypočítáme 430 / 25 = 17,2 (výsledek je pochopitelně stejný jako při výpočtu ze vstupních dat).

Vedle aritmetického průměru se můžeme setkat ještě s průměrem geometrickým[18] ($\bar{x}_G$), který je vhodný pokud potřebujeme postihnout průměrné tempo růstu:

[18]KOŘÍNEK, M.:*Rozptyl a některé další míry variability.* In Statistika 8/1988, ss. 374-376.

$$\overline{x}_G = \sqrt[n]{x_i \cdot x_2 \ldots . x_n} \quad (9)$$

Následující příklad demonstruje výpočet geometrického průměru z indexů růstu z jednotlivých údajů z let 2000 – 2005:

Tabulka 8 - Geometrický průměr

Geometrický průměr

roky	Index růstu
2000	
2001	1,110
2002	1,113
2003	0,990
2004	1,115
2005	1,120
průměr	**1,088**

Geometrický průměr byl vypočítán jako 5odmocnina ze součinu čísel ze sloupce index růstu.

Poznamenejme, že průměry lze vypočítávat pouze z kvantitativních dat. Je zřejmě nelogické pokusit se vypočítat průměrnou barvu vlasů (u kvalitativních dat používáme pro míru polohy raději modus – nejčastější variantu znaku).

5.3.2 Modus

„Modus ($\hat{x}$) je hodnota, která se v daném souboru dat vyskytuje nejčastěji (která má největší četnost). Modus slouží jako typická hodnota charakteristika polohy.“ (Maněnová, 2012)

Například modus v našem příkladu viz bodovým hodnocením žáků (viz Tabulka **2** - Absolutní četnosti) je hodnota 16, protože ve sloupci absolutní četnost je u této hodnoty největší číslo.

Poznamenejme si, že modus je vždy jedna z naměřených reálných hodnot, nejedná se o vypočítané teoretické číslo, jako například průměr.

V případě tabulky četností s intervaly lze modus odhadnout jako střed intervalu s největší četností (Chráska, 2008).

Modus je nezávislý na extrémních hodnotách měřené veličiny, proto jej řadíme mezi robustní charakteristiky (na rozdíl od průměru). Modus lze určovat jak u kvantitativních veličin, tak i u kvalitativních (například – nejčastější barva očí je hnědá).

5.3.3 Medián

„Medián ($\tilde{x}$) je prostřední hodnota z řady hodnot seřazených podle velikosti. Je to ta hodnota, která rozděluje soubor dat na dvě stejné části (počet hodnot menších nebo stejně velkých jako medián je stejný jako počet hodnot větších nebo stejně velkých jako medián.)“ (Maněnová, 2012)

Medián sice patří mezi kvantilové charakteristiky, viz dále, ale protože jeho význam je značný, zmiňujeme se o něm v tomto samostatném textu.

V případě, že celkový počet pozorování je číslo sudé, určí se medián jako aritmetický průměr ze dvou prostředních hodnot (u deseti pozorování, tedy u sudého počtu pozorování, leží medián „mezi“ 5. a 6. pozorováním – pochopitelně pracujeme již se seřazenými hodnotami).

Zkusme nyní určit medián z našeho příkladu – bodové hodnocení žáků. Na pomoc si vezmeme tabulku rozdělení četností:

Tabulka 9 - Četnosti - výpočet mediánu (a kvartilů)

hodnota	n_i	f_i	kn_i	kf_i
15	2	0,08	2	0,08
16	8	0,32	10	0,40
17	6	0,24	16	0,64
18	4	0,16	20	0,80
19	2	0,08	22	0,88
20	3	0,12	25	1,00
Celkem	25	1,00	X	X

Medián je prostřední hodnota ze seřazených hodnot. Pro medián platí, že polovina (tedy 50%) hodnot je menších či rovna této hodnotě (mediánu) a polovina hodnot je větších či rovna této hodnotě. No a právě poslední sloupec tabulky – kumulované relativní četnosti – nám poskytují informací o procentním rozložení souboru.

Číslo 0,40 nám říká, že 40% žáků má bodové ohodnocení 15 či 16. Číslo 0,64 nám říká, že 64% žáků má hodnocení 15, 16 či 17. Ale z toho plyne, že tedy přesně uprostřed leží hodnota 17, je to tedy hodnota mediánu. (Protože hodnotu 17 mají žáci mezi 40 až 64 procenty, tedy polovina – 50%– leží v této skupině.)

V případě intervalového rozdělení je možno medián odhadnout středem tohoto intervalu. Existují sice přesnější teoretické odhady (viz Cyhelský, 1981), ale v praxi se příliš nepoužívají.

Medián není citlivý k extrémním hodnotám, a proto se jedná o robustní charakteristiku. Jak bylo zřejmé z výkladu, medián lze použít pro kvantitativní data, ale i pro kvalitativní data, pokud je lze jednoznačně seřadit. Naopak pro určení mediánu u znaku barva očí je zcela nelogické.

5.3.4 Kvantily

U výkladu mediánu jsme naznačili, že medián patří mezi kvantilové charakteristiky. Obecná definice kvantilu zní: „*P* procentní kvantil je hodnota, která rozděluje seřazení hodnoty tak, že *p* procent hodnot je menších či rovno tomuto kvantilu a *1-p* procent hodnot je větších či rovno tomuto kvantilu.“ (Cyhelský, 1981)

Definice je na první pohled nepříliš jasná, ale zkusme si za P dosadit do této definice 50 a zjistíme, že se jedná o definici mediánu, který jsme si popsali výše.

Vedle mediánu, tedy prostřední hodnoty, deskriptivní statistika používá velice často i 25procentní a 75procentní kvantil.

Dle definice 25procentní kvantil $x_{0,25}$odděluje ze seřazených hodnot první čtvrtinu (25%) nejmenších hodnot. A naopak 75procentní kvantil $x_{0,75}$odděluje horní čtvrtinu nejvyšších hodnot (Hendl, 2005).

Všechny tři hodnoty – 25, 50 a 75 procentní kvantily označujeme jako kvartily. Kvartily (máme tří kvartily) rozdělují soubor seřazených hodnot na jednotlivé čtvrtiny. Alternativní označení je dolní kvartil (25), prostředníkvartil (medián, 50) a horní kvartil (75), nebo také pak Q_1 pro dolní, Q_2 promedián *a* Q_3 pro horní kvartil.

Naším úkolem bude zjistit hodnoty dolního a horního kvartilu u příkladu bodového ohodnocení žáků. Pokud jsme pochopili zjištění mediánu, pak (viz Tabulka **9** - Četnosti - výpočet mediánu (a kvartilů)) jistě rychle zjistíme i hodnoty obou krajních kvartilů.

Dolní kvartil má hodnotu 16, horní kvatil pak 18. Tyto hodnoty jsme nalezli pomocí sloupce kumulovaná relativní četnost, přesně pro dolní kvartil pak 0,40 a 0,80 pro kvatil horní.

Pro vhodnost použití kvartilů platí stejné poznámky jako pro medián (viz výše).

5.4 Charakteristiky měnlivosti (míry variability)

Pomocí charakteristik polohy (medián, modus, průměr) získáme základní představu o datech, která zpracováváme.Avšak tyto charakteristiky neříkají nic o skladbě (rozložení)pozorovaných hodnot. Informaci o tom, jak dalece jsou jednotlivé hodnoty kolem střední hodnoty nakupeny či naopak rozptýleny, vyjadřují míry variability (Sharma, 2005)

V této kapitole si některé nejpoužívanější míry variability představíme.

5.4.1 Variační rozpětí

Jako velmi jednoduchá míra, která slouží k posouzení rozptýlení hodnot (posouzení variability), se používá variační rozpětí, které je definované jako (Rumsey, 2007)

$$R = x_{max} - x_{min} \quad (10)$$

tedy rozdíl mezi největší a nejmenší naměřenou hodnotou.

Je zřejmé, že tato míra je velmi citlivá na extrémní hodnoty a tak se používá pouze pro první odhad a dále pro doplnění dalších měr variability

5.4.2 Konstrukce míry variability – průměrná absolutní odchylka

Zkusme se nyní zamyslet nad rozložením pozorovaných hodnot a vymyslet míru, která by dobře charakterizovala toto rozptýlení. Pro jednoduchost budeme uvažovat o dvou souborech, které se skládají z jednoduchých čísel – soubor 1 (1, 2, 3) a soubor 2 (100, 200, 300). Na první pohled je zřejmé, že pozorování (čísla) u prvního souboru se od sebe příliš neliší, u druhého souboru je rozptýlení mezi daty velké (dokazuje to i variační rozpětí, u souboru 1 je to hodnota 2 a u souboru se jedná o hodnotu řadově vyšší, 200).

Pokud chceme zjistit, jak se jednotlivé napozorované hodnoty od sebe liší, asi nás bude zajímat, jak jsou rozptýleny (napravo a nalevo) kolem teoretického těžiště. Víme, že pro určení těžiště lze použít průměr (viz 5.3). Základem naší míry bude tedy porovnání dané naměřené hodnoty s tímto těžištěm. Pro porovnání můžeme použít podíl (index), ale logičtější bude použití rozdílu. Jádro míry variability bude tedy rozdíl

$$x, - \bar{x} \qquad (11)$$

Zřejmě požadujeme, aby míra variability daného souboru bylo jedno číslo. Proto výše uvedené rozdíly (každé hodnoty od těžiště, od středu) sečteme, čímž dostaneme jednu hodnotu. Vzorec (11) doplníme o sumu:

$$\sum(x, - \bar{x}) \qquad (12)$$

Zkusíme si tento postup u našich obou příkladů. U souboru je průměr roven 2 a jednotlivé rozdíly se počítají takto::

1 – 2 = -1, 2 – 2 = 0, 3 – 2 = 1, součet je tudíž 0

A soubor 2, u kterého je průměr roven 200:

100 – 200 = -100, 200 – 200 = 0, 300 – 200 = 100, součet je 0

U obou případů je součet roven 0, tedy námi navržená charakteristika neodráží variabilitu, rozptýlenost naměřených hodnot. Je to zřejmé, vždyť kladné rozdíly jsou v součtu vyrušeny rozdíly zápornými.

Pro míru variability by bylo vhodné, aby se sčítaly všechny výkyvy, napravo i nalevo od těžiště (průměru). Naším úkolem je tedy vymyslet matematickou transformaci, která by z kladného čísla udělala kladné a ze záporného čísla kladné.

Ze základní školy víme, že jednou z možností je použít absolutní hodnotu. Náš základ pro výpočet m míry variability by pak měl tuto podobu:

$$\sum[x, - \bar{x}] \qquad (13)$$

Například u souboru 1 bychom dostali tyto mezivýsledky:

1 – 2 = 1, 2 – 2 = 0, 3 – 2 = 1, součet je tudíž 2

Nyní naší mírou můžeme porovnávat variabilitu u více souborů. Ale po otestování na více různých pozorování nás asi napadne, že pouhý součet všech odchylek od středu nebude asi vhodné v případě, kdy budeme porovnávat variabilitu u souborů, které mají rozdílný počet pozorování. Jinými slovy, dovedeme si představit situaci, kdy u konzistentního souboru (hodnoty kterého jsou velmi blízko středu, těžiště), ale který se skládá z mnoha pozorování, bude námi navržená míra variability (13) větší než u souboru, který se skládá z hodnosti, jež jsou velmi vzdálené od středu, ale počet měření je malý (součet mnoha malých čísel může být větší než součet několika čísel velkých).

Náprava je triviální, postačí vypočítat nikoli součet všech odchylek, rozdílností, ale vypočítat průměrnou odchylku. Celkový součet jednotlivých rozdílností vztáhnout na celkový počet pozorování. A výsledný vzorec je vymyšlen:

$$\frac{\sum[x_i - \bar{x}]}{n} \quad \textbf{(14)}$$

Této charakteristice se říká průměrná absolutní odchylka – z názvu (a našeho postupu) je patrné, jaká je její konstrukce. Velice často se označuje jako $\bar{d}_x$.

5.4.3 Rozptyl a směrodatná odchylka

Vraťme se ještě na chvilku k předešlému textu a problému s vyhledáním vhodné matematické operace, která kladné číslo transformuje na kladné a záporné na kladné. Vedle absolutní hodnoty je další možností druhá mocnina – pro náš soubor 1 by jednotlivé členy byly následující:

$$(1 - 2)^2 = 1,\ (2 - 2)^2 = 0,\ (3 - 2)^2 = 1$$

Námi navržená míra variability (14) by při použití druhé mocniny místo absolutní hodnoty měla tvar

$$\boldsymbol{\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \textbf{(15)}$$

Uvedená míra se nazývá rozptyla dle definice (konstrukce) to je aritmetický průměr čtverců odchylek jednotlivých hodnot od aritmetického průměru. Rozptyl označujeme $\boldsymbol{\sigma^2}$ (Chráska, 2008)

Poznamenejme, že tato míra variability se používá nejčastěji a že tvoří základ mnoha dalších charakteristiky (korelace, kovariance, lineární regrese atd.). Proto má ve statistice dominantní postavení.

Na tomto místě se zmíníme o pojmu výpočetní vzorec. Všechny výše uvedené vzorce jsou definiční – na první pohled je z nich zřetelné, jak je míra konstruován, jaká je logika, filozofie dané míry. Mnohdy jsou však výpočty dle definičních vzorců náročnější a tak statistika vedle definičníhovzorce nabízí mnohdy i vzorec výpočetní. Výpočet dle těchto vzorců je snadnější, na druhou stranu z těchto výpočetních vzorců se statistik nedoví, jaká je podstata daného vzorce. Proto se vždy musíme nejprve seznámit se vzorci definičními a teprve poté můžeme používat vzorce výpočetní. Nicméně je logické, že s nástupem statistického software není potřeba se s výpočetními vzorci již seznamovat, pomalu zapadají v zapomnění (přestože se i ve statistickém programu těchto vzorců pro zrychlení používá).

Například pro rozptyl vypadá výpočetní vzorec následovně:

$$\overline{\boldsymbol{x^2}} - \overline{\boldsymbol{x}}^{\boldsymbol{2}} \quad (16)$$

případně rozepsaný tvar (Cyhelský, 1981):

$$\frac{\boldsymbol{1}}{\boldsymbol{n}}\sum \boldsymbol{x_i^2} - \left(\frac{\boldsymbol{1}}{\boldsymbol{n}}\sum \boldsymbol{x_i}\right)^{\boldsymbol{2}} \quad (17)$$

Protože rozptyl se vypočítává ve čtvercích tak i daná jednotka je umocněna – pokud sledujeme znak výška, který měříme v cm, pak rozptyl by mohl mít hodnotu 25, ale pozor cm^2. Z tohoto důvodu se ve statistice používá druhá odmocnina rozptylu

$$\boldsymbol{\sigma} = \sqrt{\boldsymbol{\sigma^2}} \quad (18)$$

kterou nazýváme směrodatná odchylka. Její hodnota by v předchozím příkladu (výška) byla tedy 5 cm.

Rozptyl a směrodatná odchylka charakterizují kolísání jednotlivých hodnot kolem aritmetického průměru. Čím více a čím častěji se jednotlivé hodnoty odchylují od aritmetického průměru, tím je rozptyl i směrodatná odchylka větší. Výpočet rozptylu a směrodatné odchylky je oprávněný v těch případech, kdy zpracováváme metrická data (intervalová nebo poměrová data). (Maněnová, 2012)

5.4.4 Variační koeficient

V případě, že bychom chtěli porovnat variabilitu u souborů, která mají značně (statisticky významně) odlišné průměry, pak pro objektivní porovnání variabilit je vhodné použít variační koeficient, který poměří vypočítanou variabilitu s průměrem.

$$\boldsymbol{V = \frac{\sigma}{\bar{x}}} \quad \textbf{(19)}$$

Variační koeficient může bát vyjádřen i v procentech (V * 100%). Je důležité upozornit, že variační koeficient se také používá k „ověření“ průměru jako vhodné míry polohy. Pokud je V větší než 0,3, pak je variabilita natolik velká, že průměr nemusí dobře vystihovat míru polohy a raději použijeme medián či modus (Rumsey, 2007)

5.4.5 Kvartilové rozpětí a kvartilová odchylka

V kapitole o mírách polohy jsme si vysvětlily pojem kvartily (viz kapitola 5.3.4). Kvartily je možno použít také k vyjádření míry variability, stačí si uvědomit, že mezi dolním kvartilem a horním kvartilem se pohybuje přesně 50 procent pozorování.

Kvartilové rozpětí je analogické rozpětí variačnímu (viz 5.4.1) a vypočítá se jako (Čermák, 1968)

$$\boldsymbol{Q_3 - Q_1} \quad \textbf{(20)}$$

Kvartilová odchylka $\boldsymbol{Q}$ je definovaná jako:

$$\boldsymbol{Q = \frac{Q_3 - Q_1}{2}} \quad \textbf{(21)}$$

a je mírou rozptýlení hodnot kolem mediánu.

V případě přibližně symetrického rozdělení četností (a při dostatečně velkém souboru dat) platí, že $Q_1 = Q_2 - Q$ a $Q_3 = Q_2 + Q$.

6 Závěr

Právě dočtená skripta si kladla za úkol seznámit čtenáře s úvodem do statistiky. Značná část textu byla zaměřena na vysvětlení pojmu statistika, na začlenění statistiky do vědních oborů a pochopitelně výkladu o historii statistiky jako zprvu nevyčleněného avšak později samostatného oboru (jenž je však stále v podvědomí jako součást matematiky).

Po tomto, jak jsme naznačili poměrně rozsáhlém, úvodu nás učební text zavedl do statistických zákoutí - čtenář se seznámil s pojmy proměnná, statistický soubor, statistické zkoumání, výběrový soubor, validita. Vlastní "statistiku" jsem začali zkoumat při popisu tabulek a grafů, při sestrojování tabulek rozdělení četností, při generování intervalového rozdělení četností.

V poslední třetině jsme si vysvětlili popisnou, deskriptivní, statistiku. Míry polohy (průměr, medián) a míry variability (rozptyl, směrodatná odchylka).

Autor měl za úkol poodhalit tajemnou roušku, která statistiku (a matematiku) rádoby zahaluje. Dokladem úspěšného pokusu pak bude čtenář, který se statistiky přestane bát a který sáhne po další statistické literatuře a začne studovat. A začne statistiku prakticky používat. Finále je (možná) v nedohlednu - čtenář, který začne dokonce statistiky uvažovat.

Pro čtenáře, kteří chtějí statistiku používat (a v běžném životě se podstatně lépe orientovat) je uveden seznam literatury, který pomůže překonat před statistikou ostych a hlavně dokáže rozšířit statistické obzory. Autor přeje takovýmto čtenářům mnoho objevných zážitků.

7 Literatura

CYHELSKÝ, L.: *Úvod do teorie statistiky*. Praha, SNTL/ALFA, 1981. 04-318-81

CYHELSKÝ, L., HUSTOPECKÝ, J., ZÁVODSKÝ P.*: Příklady k základům statistiky*. Praha, SNTL/Alfa, 1988. 04-317-88

CIPRA, T: *Analýza časových řad s aplikacemi v ekonomii*. Praha, SNTL 1986. ISBN 04-012-86

ČERMÁK, V.: *Výběrové statistické zjišťování*. Praha, SNTL/ALFA, 1980. 04-326-80

ČERMÁK, V.*: Statistika. II. Díl*. Praha, SNTL/ALFA, 1968. 04-302-68

HANOUSEK, J., CHARAMZA, P.: *Moderní metody zpracování dat – matematická statistika pro každého*. Praha, Grada, 1992. ISBN 80-55623-31-5

HÁTLE, J., KAHOUNOVÁ, J.: Úvod do teorie pravděpodobnosti. Praha, SNTL, 1987.

HÁTLE, J., KAHOUNOVÁ, J.:Teorie pravděpodobnosti s příklady. Praha, SPN, 1983.

BEDNÁR, J.:*Testování statistických hypotéz*. Brno, ÚM FSI, 2006.

HENDL, J.: *Kvalitativní výzkum. Základní metody a aplikace*. Praha, Portál, 2005. ISBN 80-7367-040-8

HENDL, J.: *Přehled statistických metod zpracování dat. Analýza a metaanlýza dat*. Praha, Portál, 2006. ISBN 80-7367-123-9

CHRÁSKA, M.: *Metody pedagogického výzkumu. Základy kvantitativního výzkumu*. Praha, Grada, 2008. ISBN 978-80-247-1369-4

MANĚNOVÁ, M., ČIHÁK, M., KOŘÍNEK, M., SKUTIL, M.: *Statistické zpracování dat*. Hradec Králové: Gaudeamus. 2012. ISBN 978-80-7435-192-1

MELOUN, M., MILITKÝ, J.:Kompendium statistického zpracování dat. Praha, Academia. 2006. ISBN 80-200-1396-2

KOŘÍNEK, M.: *Seznamujeme se s grafy. Více grafů v jednom*. In Letectví a PVO 2/1989, ss. 26-29

KOZÁK, J, HINDLS, R, ARLT, J.: *Úvod do analýzy ekonomických časových řad*. Praha, VŠE. 1994. ISBN 80-7079-760-6

LAMSER, V., RŮŽIČKA, L.: *Základy statistiky pro sociology*. Praha, Nakladatelství Svoboda, 1970.25-612-70

RUMSEY, D.: *Intermediate statistics for dummies*. Hoboken, Wiley Publishing, Inc, 2007. ISBN 978-0-470-04520-6

SHARMA, A., K.: *Text book ofelementary statistics*. New Delhi, Discovery Publishing House, 2005. ISBN 81-7141-953-4

SWOBODA, H.: *Moderní statistika*. Praha, Svoboda, 1977. 25-004-77

WONNACOTT, R., J., WONNACOTT, T., H.: *Statistika pro obchod a hospodářství. Úvod do statistiky pro ekonomiku a podnikání*. Praha, Victoria Publishing, 1992. ISBN 80-85605-09-0

ŽVÁČEK J.: *Statistické výpočetní prostředí 2007*. In Informační Bulletin České Statistické Společnosti. Listopad 2007, roč. 18, č. 3, s. 1 -15. ISSN 1210-8022.

KLETEČKOVÁ, M.: *Statistický systém STATISTICA.* In Informační Bulletin České Statistické Společnosti. Prosinec 1998, roč. 9, č. 3, s. 9 -12. ISSN 1210-8022.

KÁRNÍK, I., SVOBODA, L.: *STATGRAPHICS – studnice poznání*. In Informační Bulletin České Statistické Společnosti. Červen 1997, roč. 8, č. 1, s. 20 -22. ISSN 1210-8022.

KOŘÍNEK, M: *Možnosti Zkušenosti se softwarem používaným při výuce statistiky v humanitních oborech na Pedagogické fakultě Univerzity Hradec Králové*. In Nové technologie ve vzdělávání (vzdělávací software a interaktivní tabule). On-line mezinárodní vědecko-odborná konference Olomouc 2010. URL: <http://www.kteiv.upol.cz/ntvv/?konf=konference2&detail-prispevku=57> [citováno 10. října 2010].

TVRDÍK, J.: *STAT a NCSS z pohledu uživatele*. In Informační Bulletin České Statistické Společnosti. Prosinec 1997, roč. 8, č. 3, s. 5 -16. ISSN 1210-8022.

KOLEKTIV [online]. Olomouc: 2011: [cit. 2012-04-15]. *Historie matematické statistiky*. Dostupné z WWW: <http://mant.upol.cz/soubory/MC/ps01-uvod.pdf>

KOLEKTIV: *Stručný statistický slovník*. Praha, Nakladatelství Svoboda, 1967. 25-105-67

KOŘÍNEK, M: *Statistický programový paket SPSS a Boxova-Jenkinsova analýza časových řad*. In MAA. Číslo 3, 1992. s. 74-77.

KOŘÍNEK, M: *Výuka statistiky a Excel 97*. In Pedagogický software ’97 (sborník přednášek). České Budějovice. 1997. s. 87-89. ISBN 80-85645-26-2.

ČESKÝ STATISTICKÝ ÚŘAD [online]. Praha: 2011, [cit. 2011-03-26]. *Historie statistiky v Čechách.* Dostupné z WWW: <http://www.czso.cz/csu/redakce.nsf/i/historie_statistiky_v_cechach>.

DISMAN, M.: *Jak se vyrábí sociologická znalost. Příručka pro uživatele*. Praha, Karolinum, 2006. ISBN 80-246-0139-7

ŽÁK, L. [online]. Brno: 2006 [cit. 2012-09-6]. *Historie statistiky a pravděpodobnosti*. Dostupné z WWW: <http://mathonline.fme.vutbr.cz/download.aspx?id_file=471

8 Rejstřík

M

N

O

P

Q

R

Ř

S

T

V

Z

Edice texty k sociální práci

Řada: Výzkumné metody v sociální práci - sv. 2

Název: **Statistické zpracování dat**
Rok a místo vydání: 2014, Hradec Králové
Vydání: první
Náklad: 200
Vydalo nakladatelství Gaudeamus při Univerzitě Hradec Králové jako svou 1339. publikaci.

ISBN 978-80-7435-399-4