distgeo.upol.cz

Základy statistiky

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

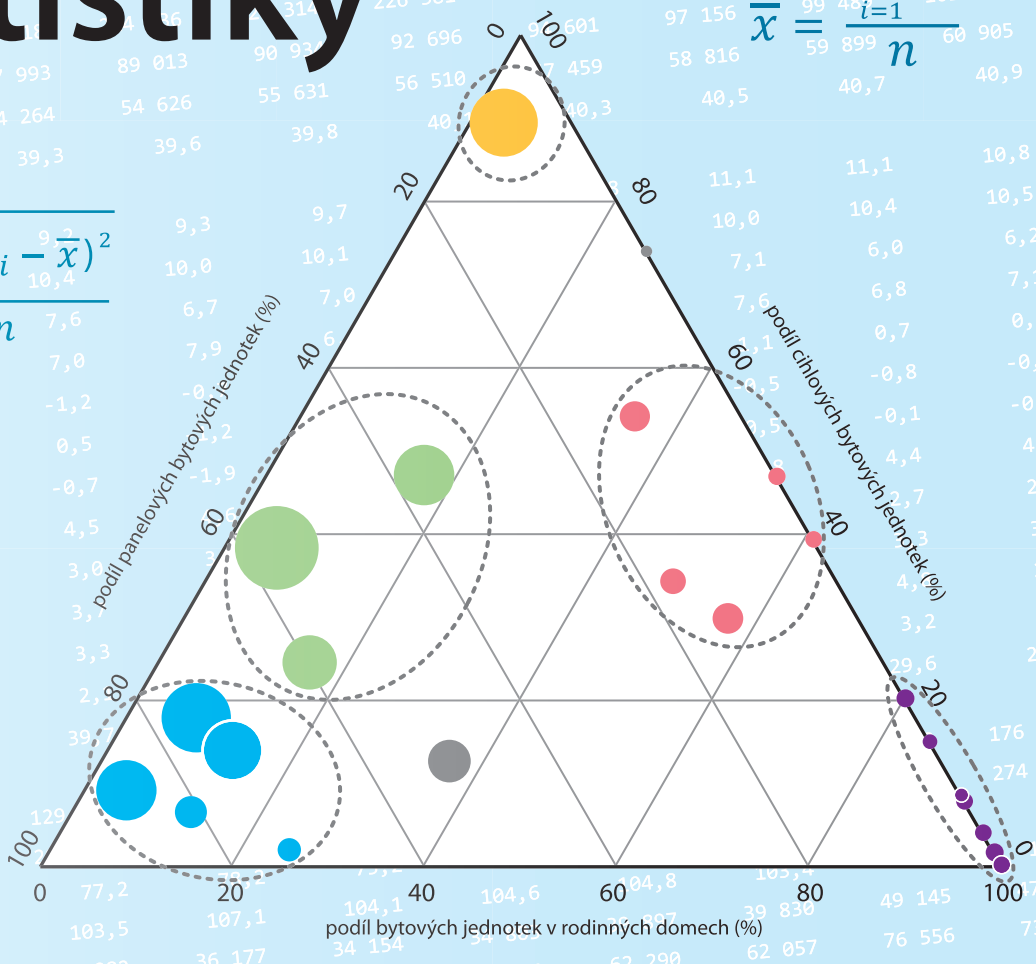
$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

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

$$v = \frac{s}{\overline{x}}$$

$$\mu \pm 2\sigma$$

$$\chi^2$$



Petr Kladivo

Univerzita Palackého v Olomouci
Přírodovědecká fakulta

Základy statistiky

Petr Kladivo

Olomouc 2013

Oponenti: RNDr. Šárka Brychtová, Ph.D.
RNDr. Miloš Fňukal, Ph.D.
Mgr. Petr Zemánek, Ph.D.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Název projektu: Rozšíření akreditace studia učitelské geografie na PřF UP v Olomouci o kombinovanou formu
Reg. číslo: CZ.1.07/2.2.00/18.0014

1. vydání

ISBN 978-80-244-3841-2 (tištěná verze)
ISBN 978-80-244-3842-9 (online verze)

Obsah

Úvod

Hlavním cílem učebního textu je poskytnout čtenáři přehledný materiál, který mu poskytne možnosti se seznámit se základními statistickými metodami uplatnitelnými v geografii. Text je systematicky rozčleněn tak, aby v přehledné formě poskytnul jednak nezbytný teoretický a metodologický rámec, ale současně i vysvětlil možnosti aplikace probíraných metod na konkrétních geografických úlohách. Jednotlivé kapitoly jsou proto doplněny řešenými cvičeními a dále příklady, na kterých si student může aplikace vyzkoušet samostatně. Každý vysvětlovaný příklad aplikace má jasně a podrobně popsaný postup a studentům objasňuje důležitá metodologická rozhodnutí. Částečnou předlohou a inspirací učebního textu jsou publikace Brázdil R. a kol. (1995): Statistické metody v geografii – cvičení a Hendl (2009): Přehled statistických metod, obzvlášť pro kapitoly 5 a 6, jejichž část byla převzata a doplněna aplikacemi a příklady pro potřeby tohoto textu.

Čtenář bude postupně seznámen se základními statistickými pojmy, základními charakteristikami statistických souborů, aby je mohl vzájemně srovnávat (zejména ukazateli polohy a variability dat). Následovat bude teorie nespojitých i spojitých rozdělení náhodných veličin se zaměřením na jejich geografické aplikace, dále se naučí posuzovat statistickou významnost dosažených výsledků prostřednictvím testování hypotéz. Velký důraz bude kladen na získání znalostí v oblasti korelačního počtu a regresní analýzy, protože schopnost posouzení těsnosti vztahu mezi dvěma proměnnými a jeho matematické vyjádření patří k základním dovednostem geografů.

Kromě konkrétních statistických metod se seznámíme i s jejich principy, základními předpoklady pro jejich použití a také omezeními plynoucími z jejich specifik. Z důvodu zachování přehlednosti a celkové vypovídací hodnoty učebního textu obsahuje teoretická část jen nezbytně nutné vědomosti. Proto je na konci zařazena kapitola věnující se jednoduché rešerši a přehledu zdrojů, literatury a odkazů na místa, kde je možné dohledat podrobnější statisticko-matematické souvislosti.

Závěr skripta je doplněn o kapitolu věnující se základním analýzám časových řad, které přestavují podstatnou část metod geografických výzkumů, text je navíc obohacen o ukázky grafické interpretce dosažených výsledků, na kterou se v dnešní době klade nemalý důraz. Smyslem je, aby student prohloubil své geografické znalosti a vnímání, aby postřehl souvislosti mezi geografickými jevy, porozuměl jednotlivým úlohám, aby pro jejich vyřešení vybral vhodnou metodu, porozuměl způsobu jejího použití, byl schopen ji aplikovat i pro odlišný případ a především interpretovat korektně její výsledky.

Předpokládané vstupní znalosti

Text je „šit na míru“ studentům, kteří absolvovali gymnázium, předpokládá proto zvládnutí středoškolského učiva. Autoři dále u svých čtenářů očekávají předchozí absolvování vysokoškolského studijního předmětu Úvod do studia geografie. Písemné úkoly, které budete v předmětu vypracovávat, by proto měly mít formální i obsahovou úroveň odborných textů (používání odborné terminologie, citační aparát, uvážlivý výběr odborných zdrojů, atd.).

Písemné kontrolní úkoly, komunikace s tutorem

Během semestru studenti zpracují v písemné formě 6 kontrolních úkolů (dva dlouhé a čtyři krátké). S těmito úkoly budou seznámeni tutorem na prvním setkání. K odevzdávání úkolů, k diskuzím s kolegy a k dotazům tutorovi budou studenti využívat e-learningový portál katedry geografie Přírodovědecké fakulty UP v Olomouci, který je dostupný na internetové adrese http://geomoodle.upol.cz/. Výukový portál katedry tak tvoří organický doplněk této studijní opory.

Vysvětlivky k ikonám

Průvodce studiem
Prostřednictvím průvodce studiem k vám promlouvá autor textu. V průběhu četby vás upozorňuje na důležité pasáže, nabízí vám metodickou pomoc a nebo předává důležitou vstupní informaci ke studiu kapitoly.

Příklad
Příklad objasňuje probírané učivo, případně propojuje získané znalosti s ukázkou jejich praktické aplikace.

Úkoly
Pod ikonou úkoly najdete dva druhy úkolů. Buď vás autor vybídne k tomu, abyste se nad nějakým problémem zamysleli a uvedli svůj vlastní názor na položenou otázku, nebo vám zadá úkol, kterým prověřuje získané znalosti. Správné řešení zpravidla najdete přímo v textu.

Pro zájemce
Část pro zájemce je určena těm z vás, kteří máte zájem o hlubší studium dané problematiky. Najdete zde i odkazy na doplňující literaturu. Pasáže i úkoly jsou zcela dobrovolné.

Řešení
V řešení můžete zkontrolovat správnost své odpovědi na konkrétní úkol nebo v něm najdete řešení konkrétního testu. Váže se na konkrétní úkoly, testy! Nenajdete zde databázi správných odpovědí na všechny úkoly a testy v textu!

Shrnutí
Ve shrnutí si zopakujete klíčové body probírané látky. Zjistíte, co je pokládáno za důležité. Pokud shledáte, že některému úseku nerozumíte, nebo jste učivo špatně pochopili, vraťte se na příslušnou pasáž v textu. Shrnutí vám poskytne rychlou korekci!

Kontrolní otázky a úkoly
Prověřují, do jaké míry jste pochopili text, zapamatovali si podstatné informace a zda je dokážete aplikovat při řešení problémů. Najdete je na konci každé kapitoly. Pečlivě si je promyslete. Odpovědi můžete najít ve více či méně skryté formě přímo v textu. Někdy jsou tyto otázky řešeny na tutoriálech. V případě nejasností se obraťte na svého tutora.

Pojmy k zapamatování
Najdete je na konci kapitoly. Jde o klíčová slova kapitoly, která byste měli být schopni vysvětlit. Po prvním prostudování kapitoly si je zkuste nejprve vyplnit bez nahlédnutí do textu! Teprve pak srovnejte s příslušnými formulacemi autora. Pojmy slouží nejen k vaší kontrole toho, co jste se naučili, ale můžete je velmi efektivně využít při závěrečném opakování před testem.

1 Základní statistické pojmy

Cíl
Po prostudování této kapitoly budete umět:

- posoudit pozici statistiky ke geografickým disciplínám,
- vysvětlit objekty studia statistiky v geografii,
- rozlišovat věcné, prostorové a časové atributy statistických znaků, jednotek.

Doba potřebná k prostudování kapitoly: **45 minut**.

Průvodce studiem
V první kapitole si řekneme o nezbytnosti statistiky pro studium geografických jevů, seznámíme se s předmětem jejího studia a terminologicky si zakotvíme základní statistické pojmy tak, abychom s nimi mohli v průběhu dalšího studia pracovat.

1.1 Statistika, popisná statistika, statistika v geografii

Statistika je vědním oborem, který se zabývá zkoumáním jevů, které mají hromadný charakter. Zkoumaný jev tedy musí příslušet určité části velkého množství prvků (předmětů, osob, událostí apod.), nebo musí být dána možnost opakovaně získat požadované informace o zkoumaném jevu za podmínek, za nichž jev může nastat. Statistika se pak zabývá zjišťováním, zpracováním, rozborem, hodnocením a výkladem údajů o tomto jevu. Tyto údaje shromažďujeme za účelem popisu rozsáhlých souborů, nebo k redukci rušivých odchylek způsobovaných jevy jinými než je sledovaný jev.

Statistiku rozdělujeme na deskriptivní, jejímž cílem je hlavně popis a matematickou, která čerpá z teorie pravděpodobnosti.

Popisná statistika

Popisná (deskriptivní) statistika se zabývá popisem stavu nebo vývoje hromadných jevů. Nejprve se vymezí soubor prvků, na nichž se bude uvažovaný jev zkoumat. Následně se všechny prvky vyšetří z hlediska studovaného jevu. Výsledky šetření – kvalitativní i kvantitativní, vyjádřeny především číselným popisem – tvoří obraz studovaného hromadného jevu vzhledem k vyšetřovanému souboru. Z popisné statistiky se postupem času vyčlenily dílčí statistické disciplíny, z nichž zřejmě nejvýznamnější je **matematická statistika**, která je založena především na teorii pravděpodobnosti.

Matematická statistika

Statistika se prolíná prakticky všemi dílčími geografickými disciplínami, které z jejích výsledků čerpají. **Statistika v geografii** pak je dílčí geografickou disciplínou, která na geografické jevy s hromadným charakterem (fyzicko-geografické, sociální, ekonomické, demografické aj.) aplikuje poznatky popisné a matematické statistiky. Ve své podstatě tak tvoří nosnou platformu pro prakticky všechny geografické subdisciplíny, které z jejích metod čerpají.

1.2 Základní pojmy

Mezi základní statistické pojmy, se kterými budeme v celém učebním textu pracovat, jsou: statistická jednotka, statistický znak a statistický soubor.

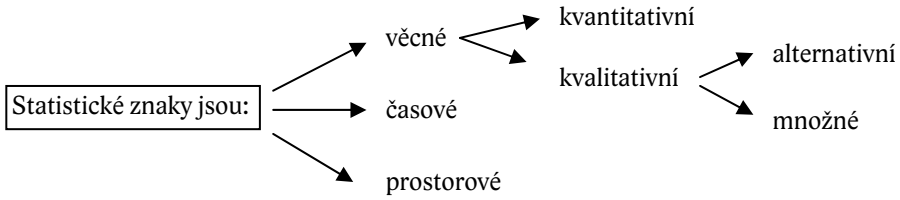
1.2.1 Statistická jednotka

Statistickou jednotkou rozumíme základní, přesně vymezený objekt, prvek, nebo jev, který je předmětem pozorování, resp. statistického šetření. Přesné vymezení statistické jednotky spočívá v jejím určení ve smyslu věcném, časovém a prostorovém.

1.2.2 Statistický znak

Každý statistický znak je věcně, časově a prostorově vymezen a je nezbytné toto v praktických úlohách zohledňovat a respektovat.

Statistickým znakem je charakteristika některé z vlastností statistické jednotky. Tyto charakteristiky, kterými můžeme rovněž rozumět měřitelné projevy jejich vlastností, rozlišit do tří kategorií (viz obr. 1) na znaky prostorové, věcné a časové.



Obr. 1 Statistické znaky (Pramen: autor).

Atribut prostoru v podstatě znamená lokaci studované vlastnosti statistické jednotky, znaky časové její časové zařazení a znaky věcné vyjadřují kvantitativní či kvalitativní ukazatel. Přitom kvantitativní znamenají měřitelné údaje dané jednotky (ptáme se „kolik“ nebo „jak velká, vysoká,...)“ – např. výška, hmotnost, zisk, objem výroby, počet zaměstnanců), zatímco kvalitativními znaky rozlišujeme vlastnosti, které nejsou měřitelné. Alternativní znaky mohou nabývat pouze dvou hodnot (ptáme-li se např. na pohlaví), množné pak více hodnot (např. zjišťujeme-li národnost, náboženství atd.).

Pomocí shodných (společných) znaků statistických jednotek vymezujeme jejich příslušnost ke statistickým souborům.

Příklad / Příklad z praxe

Zaměřme se na Sčítání lidu, domů a bytů. Statistickými jednotkami jsou osoby, domácnosti nebo např. byty a domy. Statistickým znakem pak může být věk, národnost, bydliště, pohlaví, počet členů domácnosti, stáří domu, vybavenost bytu apod.
Dalším příkladem statistické jednotky může být průmyslový podnik se statistickými znaky např. roční obrat, počet zaměstnanců, odvětví podnikání, průměrná mzda zaměstnanců apod.
Jako příklady statistických jednotek z fyzické geografie uveďme teplota vzduchu v určitý čas na určitém místě, podobně průtok, stav vodní hladiny, tlak vzduchu, srážkový úhrn apod.

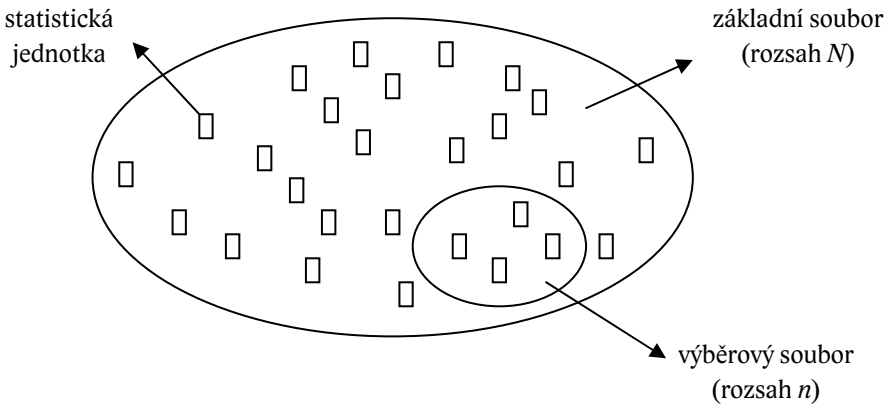
Úkol / Úkol k zamyšlení

Uveď příklady statistických jednotek a znaků z různých geografických disciplín.

1.2.3 Statistický soubor

Statistickým souborem rozumíme souhrn statistických jednotek stejného druhu. Soubory jsou rovněž jednoznačně časově, věcně a prostorově vymezeny.

Rozsah výběrového souboru označujeme n, základního N.



Obr. 2 Základní a výběrový statistický soubor (Pramen: autor).

Rozsahem souboru (označujeme *n*, jedná-li se o základní soubor *N*) rozumíme počet jednotek, které obsahuje. Za základní soubor považujeme takový, který obsahuje všechny statistické jednotky, na které se vztahuje statistické šetření (jeho rozsah může být konečný, nebo nekonečný). Výběrový soubor je pak část (výběr) ze základního souboru. Jednotky výběrového souboru vybíráme ze základního souboru buď náhodně (náhodný výběr), nebo podle určitých pravidel.

Příklad / Příklad z praxe

Za základní statistický soubor můžeme prohlásit například počet obyvatel v České republice k 1. 3. 2001 (SLDB 2001). Jeho rozsah je $N = 10\,230\,060$ (ČSÚ). Výběrových souborů z tohoto základního souboru je celá řada. Uveďme např. počet obyvatel v Pardubickém kraji ($n = 508\,281$), počet žen v ČR ($n = 5\,247\,989$), počet obyvatel ve věku 20 až 29 let ($n = 1\,708\,699$), počet rozvedených mužů ($n = 352\,079$), počet obyvatel romské národnosti ($n = 11\,746$), počet věřících obyvatel ($n = 3\,288\,088$) atd. Z těchto výběrových souborů se dají vybírat další dílčí soubory s ještě menším rozsahem.

Pro zájemce

tatistické znaky lze kategorizovat i do jiných kategorií založených ale na podobných principech. Příklad takového třídění je např. následující (podle stupně kvantifikace): **1) znaky nominální**, u kterých lze interpretovat pouze rovnost (pohlaví, barva pleti, národnost aj.); **2) znaky ordinální**, tzv. pořadové znaky (školní klasifikace, pořadí určené na základě hodnocení - počtu bodů); **3) znaky metrické** (též kardinální), charakteristické přesně např. naměřenou hodnotou, lze u nich přesně posoudit rozdíl mezi hodnotami (o kolik se liší), patří sem například teplota, tlak, ale i rozloha, plocha povodí, počet obyvatel, HDP/obyv. apod.

SHRNUTÍ

Úkolem statistiky v geografii je studium hromadných geografických jevů prostřednictvím statistických souborů, ve kterých jsou seskupeny statistické jednotky stejného druhu. Popisná část statistiky tyto soubory vyhodnocuje především pomocí jejich číselných charakteristik, matematická nebo pravděpodobnostní statistika pak posuzuje vztahy, rozdíly a závislosti mezi statistickými soubory resp. mezi hromadnými jevy a snaží se je zobecnit.

Kontrolní otázky a úkoly

1. Uveď konkrétní příklady věcného, kvalitativního, alternativního statistického znaku.
2. Jaký je vztah mezi základním a výběrovým statistickým souborem?
3. Statistickou jednotkou je měsíční úhrn srážek v Olomouci. Měření provádím v letech 2001 a 2010. Jaký bude rozsah souboru získaných hodnot?

Pojmy k zapamatování

Pojem 1: statistická jednotka, statistický znak a jejich určení a typy

Pojem 2: základní a výběrový statistický soubor

Pojem 3: rozsah souboru, náhodný výběr

2 Třídění dat a rozdělení četností

Cíl
Po prostudování této kapitoly budete umět:

- rozlišovat pojmy absolutní, relativní, kumulativní četnost,
- roztřídit data do optimálního počtu intervalů,
- tabulkově a graficky prezentovat rozložení četností ve statistickém souboru.

Doba potřebná k prostudování kapitoly: **60 minut**.

Průvodce studiem
Představme si rozsáhlý statistický soubor, např. obce České republiky s jejich počtem obyvatel. Rozsah takového souboru je n = 6 251. Pro jeho přehlednou grafickou prezentaci je třeba taková data kategorizovat, roztřídit obce do intervalů podle počtu obyvatel. Podle jakých kritérií třídíme data do intervalů, jak výsledky prezentujeme a jakých pravidel se máme držet, si řekneme v následující kapitole.

2.1 Četnosti

Četností rozumíme počet prvků se stejnou hodnotou statistického znaku (každý statistický soubor tak generuje své rozdělení četností) nebo četností myslíme počet prvků s hodnotami znaku patřícími do určitého intervalu (nebo třídy) – pak se bavíme o tzv. skupinovém (intervalovém) rozdělení četností.

2.1.1 Absolutní, relativní a kumulativní četnost

Absolutní četnosti budeme označovat n_i, relativní f_i.

Absolutní četnost (označujeme n_i) vyjadřuje absolutní hodnotou četnost zastoupených hodnot ve statistickém souboru, resp. v daném intervalu.

Relativní četnost f_i vyjadřuje četnost pomocí relativních hodnot, výpočet je dán vztahem:

$$f_i = \frac{n_i}{n},$$

tj. je dána podílem jednotlivých absolutních četností n_i k rozsahu souboru n. Může být uvedena desetinným číslem, nebo procentuálně.

Kumulativní četnosti absolutní N_i, resp. relativní F_i udávají úhrnnou četnost statistických jednotek s hodnotami znaku menšími, nebo rovnými hodnotě znaku nebo horní hranici intervalu, při seřazení hodnot nebo intervalů podle pořadí neklesajících hodnot znaku. Kumulovanou četnost lze vyjadřovat a počítat z absolutních četností i z relativních četností.

2.2 Třídění dat do intervalů

2.2.1 Intervaly a jejich parametry, terminologie

Hranice intervalu – neboli mez intervalu –, ať už horní nebo dolní, určuje, které hodnoty do intervalu patří.

Délka intervalu – nebo též rozpětí či šířka – je rozdíl (kladný) dvou po sobě následujících dolních (nebo horních) hranic intervalů.

Střed intervalu – označujeme x_s – je důležitou hodnotu, která při výpočtech z intervalového rozdělení četností zastupuje příslušný interval. Střed intervalu spočítáme jako aritmetický průměr horní a dolní hranice intervalu, neboli: (a+b) : 2, kde a (b) je dolní (horní) hranice intervalu.

Typologie intervalů

<a; b> – uzavřený interval, množina všech x, pro která platí **a ≤ x ≤ b**

(a; b) – otevřený interval, množina všech x, pro která platí **a < x < b**

<a; b) – uzavřený interval zleva, množina všech x, pro která platí **a ≤ x < b**

(a; b> – uzavřený interval zprava, množina všech x, pro která platí **a < x ≤ b**

2.2.2 Princip třídění dat

Jak třídíme data?

Jednotlivé intervaly, do kterých sledovaný statistický soubor rozdělíme, vzniknou roztříděním jeho hodnot podle určitých kritérií:

- každý interval je přesně vymezen svojí horní a dolní hranicí,
- jsou vymezeny tak, aby šel každý prvek jednoznačně zařadit,
- intervaly se nesmí překrývat,
- šířka intervalů by měla být stejná (pro snadnější výpočty),
- počet intervalů volit optimálně („ani málo, ani příliš“).

Jak rozdělit data do intervalů?

Exaktní pravidla pro určení optimálního počtu intervalů neexistují, celý algoritmus bude vždy obsahovat subjektivní prvek. Přesto se setkáme s doporučeními, jak postupovat, následující algoritmus představuje jedno z nich:

- určíme R jako rozdíl mezi maximální a minimální hodnotou (jedná se o variační rozpětí) sledovaného souboru, tzn. $R = x_{\min} - x_{\max}$,
- výpočet počtu intervalů (tříd) označme k – rozdělíme na tři případy:
  - je-li rozsah souboru $n > 100$, pak $k = 10 \cdot \log n$ ... (i)
  - je-li rozsah souboru $40 < n \leq 100$, pak $k = 2 \cdot \sqrt{n}$ ... (ii)
  - je-li rozsah souboru $n \leq 40$, pak $k = 1 + 1{,}4426 \cdot \ln \mathrm{n}$ ... (iii)

- výpočet šířky intervalu *h* je pak dán vztahem:

$$h = \frac{R}{k}.$$

Jak už bylo uvedeno výše, volba počtu intervalů se těmito pravidly nemusí řídit, může být intuitivní, provedená na základě analýzy struktury studovaných dat nebo na základě zkušeností.

2.3 Grafické vyjádření rozdělení četností

Zjištěné četnosti nejpřehledněji uvádíme v „tabulkách intervalového (skupinového) rozdělení četností“ – viz tab. 2. Pro přehlednost, nadhled, nebo lepší orientaci prezentujeme data z těchto tabulek graficky, nejčastěji pomocí histogramu, polygonu a součtové čáry.

2.3.1 Histogram

Histogram je graf vyjadřující rozložení četností ve statistickém souboru. Jedná se o graf sloupcový, při jeho konstrukci nezáleží na tom, zda jako zdrojová data uvažujeme absolutní, nebo relativní četnosti (pro oba způsoby vypadá diagram stejně). Na vodorovnou osu x nanášíme intervaly v příslušných jednotkách, na ose svislé y se vynáší absolutní (relativní) četnosti. Jak již bylo řečeno, jedné se o sloupcový graf, kde šířce sloupce odpovídá délka (šířka) intervalu a výšce pak četnost v daném intervalu. Z vhodně sestrojeného histogramu lze vypozorovat rozložení hodnot ve statistickém souboru, jejich rozmístění okolo střední hodnoty, rovněž jejich rozptyl v souboru a dají se určit další charakteristiky, jako například modální interval aj.

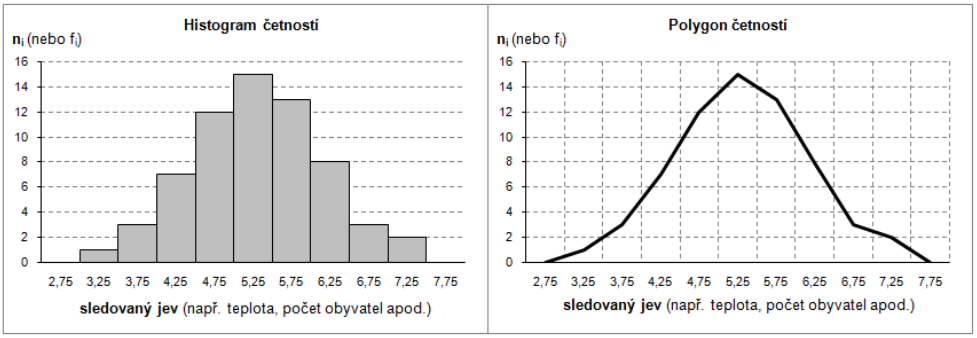
Histogramem rozumíme sloupcový graf prezentující četnosti. Můžeme ho sestrojit z četností absolutních i relativních, tvar bude mít stejný.

2.3.2 Polygon

Polygon je obdobou histogramu, také vyjadřuje rozložení četností ve statistickém souboru, liší se pouze typem grafu. Zatímco v případě histogramu se jedná o sloupcový graf, u polygonu jde o spojnicový typ grafu. Z vlastností polygonu vyplývá, že v případě jeho sestrojení z relativních četností ohraničuje křivka polygonu plochu o velikosti 1 (v případě vyjádření v procentech pak 100 %).

Rozdíl mezi polygonem a histogramem je pouze v typu grafu.

Polygon i histogram tak znázorňují stejné údaje poněkud odlišným způsobem, nezáleží na výběru z absolutních nebo relativních četností.



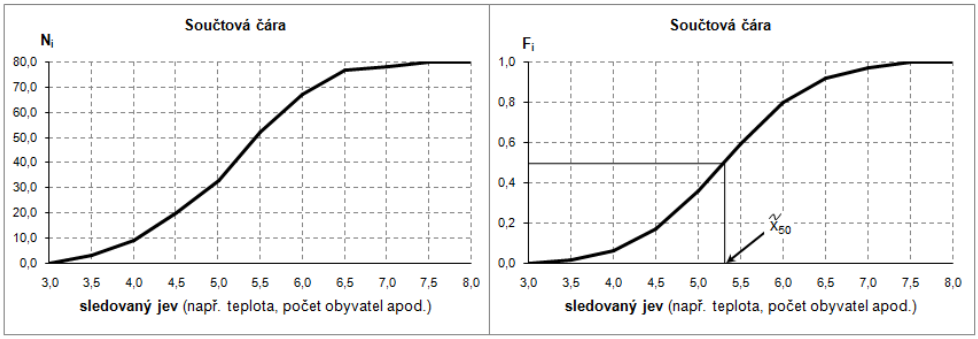
Obr. 3 Ukázka histogramu a polygonu četností (Pramen: autor).

2.3.3 Součtová čára

Součtovou čarou prezentujeme kumulativní četnosti, lze využít i histogram kumulativních četností.

Součtová čára slouží pro znázornění kumulovaných četností. Při konstrukci se vynáší hodnoty kumulovaných četností (nezáleží na tom, zda absolutní, či relativní, ale častěji se používají relativní) k horním hranicím intervalů, body se spojí lomenou čarou.

Z grafu součtové čáry lze vyčíst řadu charakteristik, mj. hodnoty kvartilů, medián atd.



Obr. 4 Ukázky součtové čáry, vpravo s vyznačením mediánu (Pramen: autor).

Pro zájemce

Určit optimální počet intervalů a jejích šířku závisí na povaze problematiky, resp. jevu, který analyzujeme. Zpravidla se řídíme tím, aby měly všechny intervaly, do kterých data třídíme konstantní šířku, tj. aby byly všechny stejně velké, což praktické vzhledem k výpočtům i prezentaci datového souboru. Setkáme se ale s řadou případů, kdy toto pravidlo dodržet nelze. Ukázkovým příkladem je třídění obcí České republiky do intervalů podle počtu obyvatel. Je zřejmé, že při řešení takovéto úlohy musíme počet intervalů a jejich velikost, resp. hranice, určit uměle, dodržet konstantní šířku intervalů je nevhodné.

Příklad / Příklad z praxe

Máme k dispozici fiktivní data – mzdy (v tis. Kč) 30 zaměstnanců firmy (viz tab. 1). Data vhodně roztřiďte do tříd, graficky a tabulkově prezentujte.

Tab. 1 Data pro příklad.

22	25	19	17	25	31	8	17	18	15
27	17	16	22	24	21	18	23	18	13
21	12	22	10	16	13	29	19	21	22

Pramen: Autor.

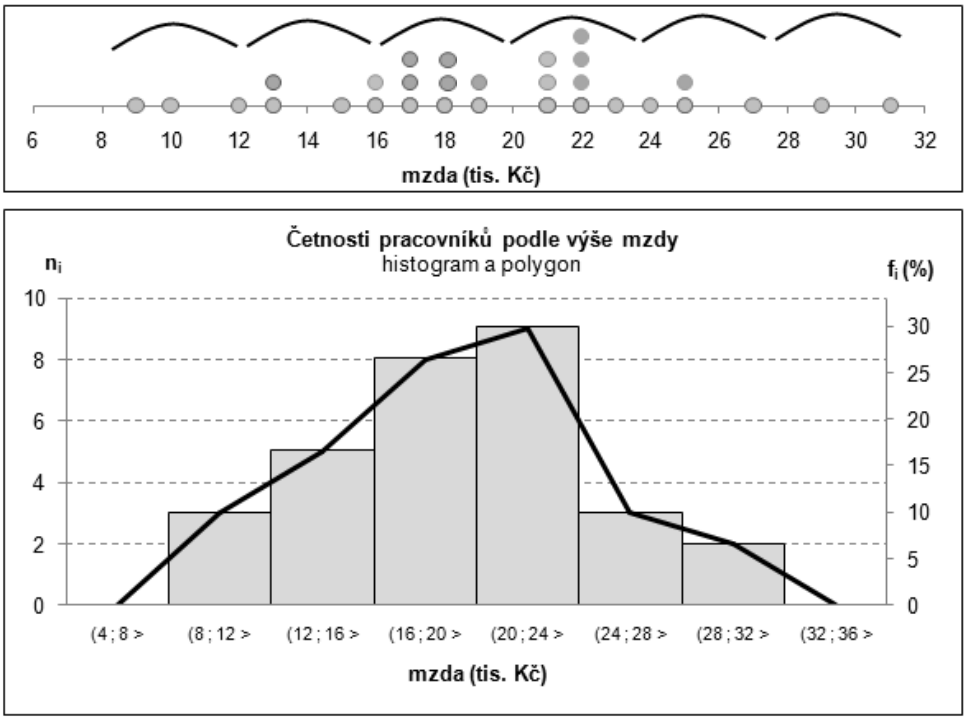
Řešení:

Pro určení optimálního počtu intervalů k použijeme vztah (iii), protože rozsah souboru $n = 30$. Tedy $k = 1+1{,}4426 \cdot \ln 30 = 5{,}9$. Data tedy roztřídíme do šesti tříd, variační rozpětí $R = x_{min} - x_{max} = 31 - 8 = 23$. Šířka intervalů $h = 23 : 6 = 3{,}8$. Vzhledem k povaze dat může pracovat s šířkou intervalů 4. Začneme-li minimální hodnotou 8, dostaneme první interval (8; 12>. Zkonstruujeme zbývající intervaly, roztřídíme do nich původní hodnoty a spočítáme četnosti příslušné jednotlivým intervalům (viz tab. 2).

Tab. 2 Řešení příkladu.

Mzda (tis. Kč)	x_s	n_i	N_i	f_i (%)	F_i (%)
(8 ; 12 >	10	3	3	10,0	10,0
(12 ; 16 >	14	5	8	16,7	26,7
(16 ; 20 >	18	8	16	26,7	53,3
(20 ; 24 >	22	9	25	30,0	83,3
(24 ; 28 >	26	3	28	10,0	93,3
(28 ; 32 >	30	2	30	6,7	100,0
	-	30	-	100,0	-

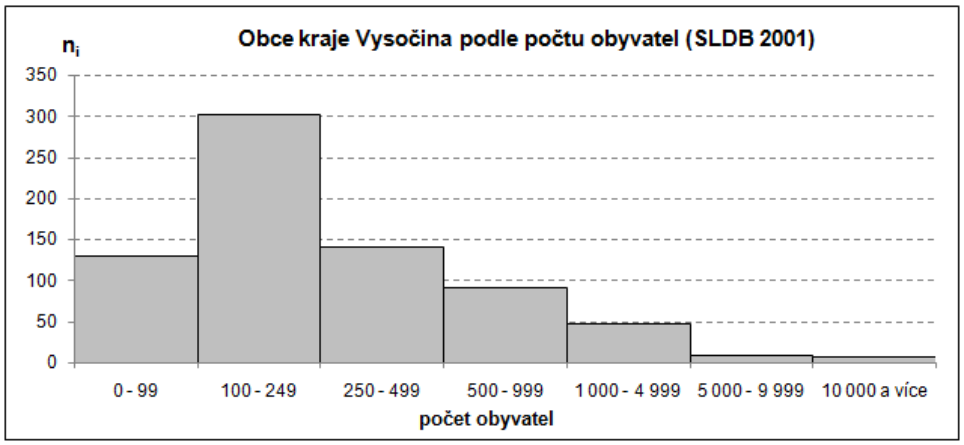
Pramen: Autor.



Obr. 5 Prezentace neroztříděného (nahoře) a roztříděného souboru. (Pramen: autor).

Příklad / Příklad z praxe

Minulý příklad byl ukázkou třídění dat do stejně velkých intervalů, setkáme se ale také s tříděním do intervalů s různou šířkou, typická ukázka viz obr. 6.



Obr. 6 Ukázka třídění dat do nestejně velkých intervalů. (Pramen: autor, data ČSÚ).

Úkol / Úkol k zamyšlení

Máte k dispozici statistický soubor (viz níže) – fiktivní data o průměrné roční teplotě na meteorologické stanici. Data roztřiďte do intervalů a tabulkově i graficky je prezentujte (histogram, součtová čára).

7,4	8,3	8,5	10,9	7,9	10,8	9,9	9,4	9,3	8,5
9,6	9,4	8,2	9,7	8,4	9,4	10,7	8,8	9,5	9,0
8,1	10,3	7,7	8,8	8,6	9,8	9,4	8,9	9,6	9,2
9,1	9,9	10,0	8,9	10,2	9,3	9,6	8,7	9,9	9,4
7,9	10,1	11,1	9,3	10,5	8,5	9,1	9,1	8,8	9,6

Doporučení: *Zvolte šířku intervalu h = 0,5 ºC, jako dolní hranici prvního intervalu zvolte teplotu 7,0 ºC.*

SHRNUTÍ

Umět rozdělit údaje ze statistického souboru do tříd patří k elementárním dovednostem práce s daty. K určení optimálního počtu intervalů, do kterých třídíme, lze využít některé z existujících algoritmů, často se však jedná o záležitost subjektivní, která vychází buď z doporučení, nebo ze zkušeností. Nedílnou součástí celého procesu je korektní tabulková a grafická prezentace ať už neroztříděných nebo roztříděných statistických souborů.

Kontrolní otázky a úkoly

1. Čemu je roven součet všech absolutních (n_i) a relativních (f_i) četností ve statistickém souboru?
2. Jaký je rozdíl mezi polygonem a součtovou čárou?
3. Uveď příklady z geografie, kde se nehodí třídit data do intervalů se stejnou šířkou.

Pojmy k zapamatování

Pojem 1: četnost, absolutní, relativní, kumulativní četnosti

Pojem 2: histogram, polygon, součtová čára

Pojem 3: variační rozpětí, horní, dolní hranice a střed intervalu

3 Základní statistické charakteristiky

Cíl
Po prostudování této kapitoly budete umět:

- vypočítat a okomentovat číselné charakteristiky statistických souborů,
- na základě vypočítaných hodnot mezi sebou statistické soubory porovnat,
- vybrat reprezentativní číselné charakteristiky pro statistický soubor.

Doba potřebná k prostudování kapitoly: **120 minut**.

Průvodce studiem
Jedním ze základních úkolů statistiky je schopnost porovnávání statistických souborů mezi sebou. Jednou z variant je prezentace rozložení četností v těchto souborech, kterou jsme si uvedli v minulé kapitole. Další možností je srovnávání pomocí číselných charakteristik, o kterém si řekneme nyní. Číselnou charakteristikou rozumíme hodnoty (průměry, odchylky apod.), které nám budou statistické soubory reprezentovat, a na jejich základě budeme schopni soubory porovnávat.
Probereme si čtyři základní skupiny statistických charakteristik, budou to charakteristiky úrovně (též polohy), charakteristiky variability, charakteristiky šikmosti a konečně charakteristiky špičatosti.
Součástí této kapitoly budou vzorce, podle kterých se jednotlivé číselné charakteristiky počítají.

Terminologie a symbolika, kterou budeme dále používat:
Neroztříděný statistický soubor:
x_i *– prvek statistického souboru (statistická jednotka),*
n *– rozsah souboru; soubor se tedy skládá z prvků* $x_1, x_2, \ldots, x_n$.

Roztříděný statistický soubor:
n_i *– četnost příslušného intervalu (např.* n_1 *– četnost prvního intervalu),*
x_{si} *– střed příslušného intervalu (např.* x_{s1} *– střed prvního intervalu),*
k *– počet intervalů, do kterých jsou data roztříděna,*
n *– rozsah souboru.*

3.1 Charakteristiky úrovně, polohy

Statistickými charakteristikami (ukazateli) úrovně, resp. polohy statistického souboru, rozumíme hodnoty zkoumaného znaku, které udávají velikost jevu v daném souboru a udávají polohu četností. Slouží k porovnávání dvou i více souborů, charakteristiky úrovně vlastně zastupují všechny hodnoty statistického souboru (typicky např. aritmetický průměr). Nejčastěji používanými charakteristikami úrovně jsou střední hodnoty (průměry, modus, medián apod.), dále sem řadíme např. kvantily (kvartily, decily, percentily).

3.1.1 Střední hodnoty

Střední hodnoty patří k nejdůležitějším a nejpoužívanějším charakteristikám statistických souborů vůbec. Obzvlášť průměr, modus a medián.

O středních hodnotách se bavíme v případě různých druhů průměrů (aritmetický, harmonický, geometrický, vážený), řadíme sem také modus, medián a aritmetický střed.

Aritmetický průměr

Je patrně nejpoužívanější statistickou charakteristikou, jejíž výpočet je velmi jednoduchý – jde o úhrn hodnot statistického znaku, dělený rozsahem souboru:

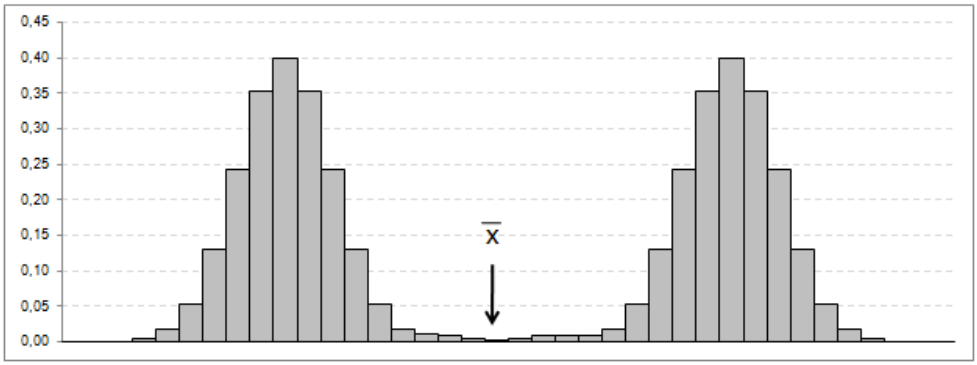
$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Mezi základní vlastnosti aritmetického průměru patří:

- algebraický součet všech odchylek jednotlivých hodnot znaku od aritmetického průměru je roven nule,
- je-li znak konstantní, průměr je roven této konstantě,
- přičteme-li ke všem hodnotám znaku konstantu k, zvětší se i průměr o tuto konstantu,
- vynásobíme-li všechny hodnoty znaku konstantou k, je i průměr k-krát větší.

Typický průměr alespoň přibližně vystihuje nejčastější hodnotu v souboru, netypický nikoliv.

Kromě té výhody, že výpočet aritmetického průměru je velmi jednoduchý, má tato charakteristika i některé nevýhody, a to zejména tu, že nemusí vždy podávat správnou informaci. Může být zkreslen extrémní (výraznou maximální nebo minimální) hodnotou v případě, že vycházíme ze souboru s nižším rozsahem, rovněž rozdělení hodnot v souboru může mít dva nebo více vrcholů, a ty jedním ukazatelem nelze popsat. Pak mluvíme o „typickém“ průměru – kdy je většina hodnot souboru „blízká“ průměru – a naopak o „netypickém“ průměru.



Obr. 7 Statistický soubor (bimodální) s tzv. netypickým průměrem. (Pramen: autor).

Vážený aritmetický průměr

Vážený průměr se využívá v případě, kdy prvky statistického souboru mají různou důležitost, tj. že každému prvku statistického souboru x_i je přiřazena jeho váha n_i. Typickým, i když negeografickým příkladem jsou získané známky ze zkoušek absolvovaných předmětů, váhami pak jsou kredity příslušné těmto předmětům. Vážený průměr dostaneme jako součet součinů prvků a jejich vah dělený celkovým součtem vah, tj. ze vztahu:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{\sum_{i=1}^{k} n_i}$$

Pro výpočet aritmetického průměru roztříděného statistického souboru (kdy neznáme vstupní data), se používá právě váženého průměru, ve vzorci stačí nahradit x_i za x_{si} – středy intervalů a jednotlivé váhy (n_i) jsou vlastně četnosti příslušné jednotlivým intervalům.

Příklad / Příklad z praxe

Máte k dispozici údaje o počtu zaměstnanců podniku v jednotlivých mzdových tarifních třídách. Spočítejte průměrnou tarifní třídu s využitím váženého průměru.

Tarifní třída	1	2	3	4	5	6	7	8
Počet zaměst.	8	12	18	36	63	46	23	14

Geometrický průměr

Geometrický průměr využijeme při analýze časové řady, konkrétně při výpočtu průměrného tempa růstu.

Používá se v případech, kdy hodnoty tvoří alespoň přibližně geometrickou řadu. Tehdy má smysl uvažovat o použití geometrického průměru. V geografii se pomocí geometrického průměru analyzují zpravidla časové řady, typickou úlohou je výpočet průměrného tempa růstu. Geometrický průměr se počítá jako *n*-tá odmocnina ze součinu všech hodnot souboru:

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

Aritmetický střed

Aritmetický střed sice podává okamžitou informaci kde je střed souboru, ale může být výrazně zkreslen odlehlou hodnotou.

Jde spíše doplňkový ukazatel, popřípadě podává prvotní informaci o rozložení hodnot ve statistickém souboru. V případě, že jsou hodnoty v něm rozloženy rovnoměrně, podává poměrně kvalitní informaci v tom smyslu, že se aritmetický střed v takovém případě blíží aritmetickému průměru. Z jeho vlastní definice (jedná se o aritmetický průměr maximální a minimální hodnoty v souboru) pak plynou i případné nevýhody. Je-li maximální, nebo minimální hodnota výrazně „vychýlena“ či „vzdálena“ od ostatních hodnot, není jeho použití vhodné a nemá příliš velkou vypovídající hodnotu.

$$x_{st} = \frac{x_{\max} + x_{\min}}{2}$$

Modus

Modem nazýváme nejčetnější (nejčastější) hodnotu kvantitativního znaku studovaného souboru, to v případě, že vycházíme z neroztříděného souboru, tedy ze všech jeho hodnot. Na první pohled je tak zřejmé, že pro snadné nalezení modu je vhodné seřadit hodnoty znaku vzestupně nebo sestupně. V případě souboru roztříděného do intervalů hovoříme o intervalu s největší četností jako o „modálním intervalu“ a hodnotu modu (přibližnou) jsme schopni spočítat pomocí následujícího vzorce:

$$\hat{x} = L + h\frac{n_2}{n_1 + n_2},$$

kde L je dolní hranice modálního intervalu, h je šířka modálního intervalu, n_1 je četnost intervalu, který předchází modálnímu a n_2 je četnost intervalu, který následuje po modálním.

Důležitost modu se projeví při vystižení typické hodnoty znaku v daném souboru a následně při porovnávání typických hodnot souborů.

Medián

Medián

Medián je prvek řady (hodnot sledovaného znaku), uspořádané v neklesajícím (rostoucím) pořadí, který ji rozděluje na dvě části v tom smyslu, že polovina prvků této řady má menší hodnotu znaku a polovina má větší hodnotu znaku, než je hodnota mediánu. Jinými slovy lze prohlásit, že za medián považujeme hodnotu, která nám dělí vzestupně seřazené hodnoty statistického souboru na dvě stejné poloviny. Označujeme ho $\tilde{x}_{50}$.

Má-li soubor rozsah *n* a jeho hodnoty jsou vzestupně uspořádané, pak je, v případě, že n je liché, medián hodnota, která má pořadové číslo

$$\frac{n+1}{2}.$$

Pro *n* sudé za medián považujeme aritmetický průměr hodnot, které se nachází na pozicích

$$\frac{n}{2} \text{ a } \frac{n}{2}+1.$$

Výhodou mediánu je, že zachycuje úroveň (polohu) hodnot lépe než průměr.

Příklad / Příklad z praxe

Vypočítejte aritmetický průměr, aritmetický střed a určete modus a medián ze vstupních dat z příkladu na str. 14.

3.1.2 Kvantily

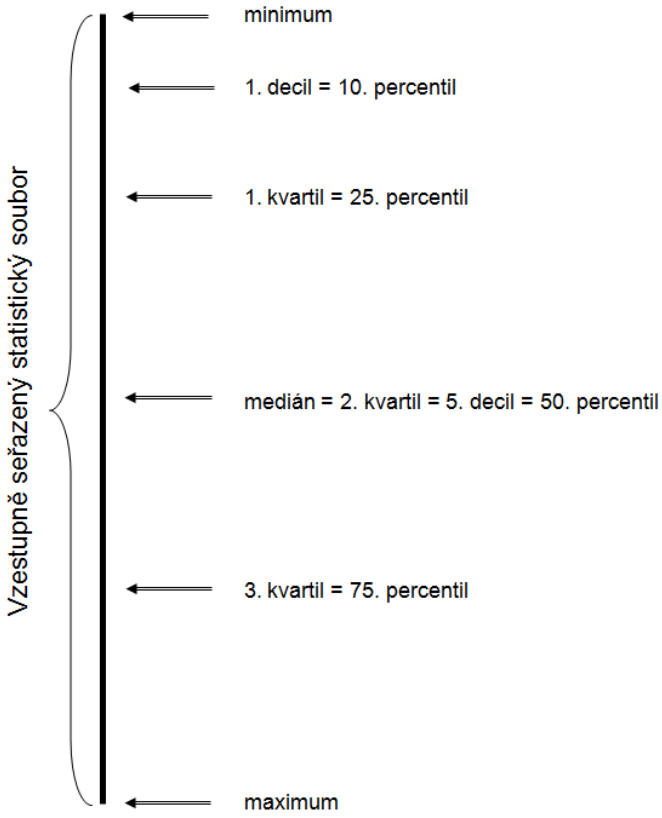
Hodnoty kvantilů informují o rozložení dat ve vzestupně seřazeném statistickém souboru.

Kvantily se řadí mezi charakteristiky úrovně, střední hodnotou je však pouze jeden z nich, a to medián. Kvantily obecně fungují na stejném principu jako medián. Jak již bylo uvedeno, za medián považujeme hodnotu, která dělí vzestupně seřazené hodnoty statistického souboru na dvě stejné poloviny.

Kvartily jsou takové hodnoty, které od sebe oddělují čtvrtiny vzestupně seřazených hodnot souboru. Jsou tedy celkem tři. První (dolní) kvartil odděluje první čtvrtinu hodnot od zbylých tří čtvrtin, druhý (prostřední) kvartil odděluje první dvě čtvrtiny od zbylých dvou a je tedy totožný s mediánem, třetí (horní) kvartil odděluje první tři čtvrtiny hodnot od poslední čtvrtiny.

Obdobně v souboru identifikujeme **decily**, kterých je v každém statistickém souboru celkem devět a dělí ho na jednotlivé desetiny, a konečně **percentily**, které ho dělí na setiny. Percentilů je v souboru 99.

Označení:	$\tilde{x}_{25}$, $\tilde{x}_{50}$ a $\tilde{x}_{75}$	1., 2. a 3. kvartil
	$\tilde{x}_{10}$, $\tilde{x}_{20}$, ..., $\tilde{x}_{90}$	1., 2., ..., 9. decil
	$\tilde{x}_{1}$, $\tilde{x}_{2}$, ..., $\tilde{x}_{99}$	1., 2., ..., 99. percentil



Obr. 8 Rozložení kvantilů ve statistickém souboru (Pramen: autor).

Příklad / Příklad z praxe

Níže uvedená data (zdroj: ČSÚ) prezentují počty nevěst podle věku v České republice za rok 2006. Určete medián věku nevěst a 1. a 3. kvartil.

věk	počet	f_i	F_i	věk	počet	f_i	F_i
16	17	0,0003	0,0003	**29**	3 633	0,0687	0,6384
17	22	0,0004	0,0007	**30**	3 050	0,0577	0,6961
18	388	0,0073	0,0081	**31**	2 297	0,0435	0,7396
19	644	0,0122	0,0203	**32**	1 782	0,0337	0,7733
20	1 054	0,0199	0,0402	**33**	1 432	0,0271	0,8004
21	1 592	0,0301	0,0703	**34**	1 071	0,0203	0,8206
22	2 139	0,0405	0,1108	**35–39**	3 287	0,0622	0,8828
23	2 795	0,0529	0,1637	**40–44**	2 144	0,0406	0,9234
24	3 624	0,0686	0,2322	**45–49**	1 496	0,0283	0,9517
25	4 116	0,0779	0,3101	**50–54**	1 211	0,0229	0,9746
26	4 684	0,0886	0,3987	**55–59**	779	0,0147	0,9893
27	4 727	0,0894	0,4881	**60+**	564	0,0107	1,0000
28	4 312	0,0816	0,5697	**celkem**	**52 860**	**1,0000**	

3.2 Charakteristiky variability

Jedná se o hodnoty, které charakterizují stupeň proměnlivosti statistického znaku (resp. hodnot sledovaného jevu) v daném statistickém souboru. Měříme proměnlivost vzhledem

Platí pro něj totéž, co pro aritmetický střed.

k typické hodnotě souboru, zpravidla vzhledem k průměru nebo mediánu. Charakteristiky variability jsou důležitým doplňkem informací, které poskytují střední hodnoty. Jak najít střední odchylku s nejlepší vypovídající schopností si ukážeme na následujícím příkladu:

Máme k dispozici statistický soubor o rozsahu pěti hodnot: 20; 30; 40; 60; 100. Snadno nalezneme aritmetický průměr:

$$\frac{20+30+40+60+100}{5}=50\,.$$

První možností, jak hledat průměrnou odchylku je konstrukce absolutních odchylek. Jejich nevýhodou je (vzhledem k vlastnostem aritmetického průměru), že dávají součet 0, tedy jejich průměr je také nulový.

Druhou možností je uvažovat nezáporné hodnoty absolutních odchylek (viz obr. 9). Jejich součet je 120 a průměr 24 (120 : 5). Dostáváme tzv. „průměrnou odchylku“. Třetí a z matematického pohledu nejlepší metodou je výpočet kvadratických odchylek (absolutní odchylky umocněné na druhou). Jejich průměr 800 (4 000 : 5) nazýváme rozptyl statistického souboru. Pokud tento průměr (800) odmocníme, čímž se vrátíme do původního rozměru dat, dostaneme hodnotu 28 a nazveme ji „směrodatnou odchylkou“. Jedná se o nejčastěji používanou charakteristiku variability a současně tu nejvhodnější. Přehled vybraných charakteristik variability je uveden dále v textu.

x_i	absolutní odchylky $(x_i - \bar{x})$	absolutní odchylky $\lvert x_i - \bar{x}\rvert$	kvadratické odchylky $(x_i - \bar{x})^2$
20	-30	30	900
30	-20	20	400
40	-10	10	100
60	10	10	100
100	50	50	2 500
Σ	**0**	**120**	**4 000**
průměr	**0**	**24**	**800**

Obr. 9 Konstrukce vybraných odchylek od aritmetického průměru. (Pramen: autor).

Variační rozpětí

Jde o nejjednodušší ukazatel variability souboru, určí se jako rozdíl minimální a maximální hodnoty ve sledovaném souboru, tedy

$$R = x_{\min} - x_{\max}\,.$$

Jedná se o ukazatel jednoduchý, ale protože závisí pouze na dvou extrémních hodnotách, nemusí být dostatečně výstižný, maximální a minimální hodnota může být „nahodilá“. Tato ne příliš dokonalá míra variability slouží především k první informaci o variabilitě souboru.

Průměrná odchylka

Průměrné odchylky vyjadřují míru odlišnosti (variace) od střední hodnoty (průměru, mediánu). Jsou doplňkovou informací ke střední hodnotě a spočítají se jako aritmetický průměr absolutních hodnot odchylek (rozdílů) všech hodnot znaku od střední hodnoty (aritmetického průměru, mediánu...). Pokud vydělíme průměrnou odchylku střední hodnotou (průměrem, nebo mediánem), dostaneme relativní bezrozměrnou míru.

Výpočty průměrné odchylky (od průměru, mediánu):

$$\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \text{ resp. } \bar{d}_{\tilde{x}} = \frac{\sum_{i=1}^{n} |x_i - \tilde{x}|}{n}.$$

Výpočet průměrné odchylky z intervalového rozdělení četností:

$$\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^{k} |x_s - \bar{x}| \cdot n_i}{\sum_{i=1}^{k} n_i}.$$

Střední diference

Je definována jako aritmetický průměr absolutních hodnot všech možných vzájemných rozdílů n jednotlivých hodnot sledovaného znaku x, Vhodná míra variability pro soubory s malým rozsahem, v ostatních případech je její výpočet zbytečně pracný:

$$\Delta = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|}{n \cdot (n-1)}.$$

Rozptyl

Rozptyl vypočítáme jako průměr ze čtverců odchylek jednotlivých hodnot znaku od jejich aritmetického průměru. Použít můžeme vzorec pro výpočet rozptylu neroztříděného souboru (vzorec vlevo), nebo uvažovat soubor roztříděný do intervalů (vzorec vpravo).

Rozptyl je nejdůležitější charakteristikou variace hodnot znaků ve statistickém souboru.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \qquad s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i}{\sum_{i=1}^{k} n_i}$$

Vybrané vlastnosti rozptylu:

- pokud odečteme od všech hodnot statistického souboru stejnou konstantu *k*, rozptyl souboru zůstane nezměněn,
- po vynásobení všech hodnot statistického souboru stejnou konstantu *k*, rozptyl musíme vynásobit druhou mocninou této konstanty.

Směrodatná odchylka představuje nejčastější a nejvhodnější charakteristiku variability.

Směrodatná odchylka

V praxi se směrodatnou odchylkou setkáváme častěji než s rozptylem, je definována jako druhá odmocnina z rozptylu a vlastně se jedná o míru rozptylu hodnot sledovaného znaku x_i kolem průměru.

Vzorce pro výpočet směrodatné odchylky z neroztříděného souboru, nebo-li ze všech hodnot (vlevo) a z intervalového rozdělení četností (vpravo).

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

Vybrané vlastnosti směrodatné odchylky:

- pokud odečteme od všech hodnot statistického souboru stejnou konstantu *k*, směrodatná odchylka zkoumaného souboru zůstane nezměněna
- po vynásobení všech hodnot statistického souboru konstantou *k,* se směrodatná odchylka musí vynásobit také touto konstantou

Variační koeficient

Variační koeficient je nejpoužívanější relativní mírou variability.

Variační koeficient je dán poměrem směrodatné odchylky a aritmetického průměru a z definice tohoto poměru plyne, že jde o ukazatel (míru) bezrozměrný.

$$v = \frac{s}{\bar{x}} \qquad v = \frac{s}{\bar{x}} \cdot 100[\%]$$

Variační koeficient se uvádí desetinným číslem (aplikací vzorce vlevo), nebo po vynásobení stem v procentech (vzorec vpravo).

3.3 Charakteristiky šikmosti

Charakteristikami šikmosti (symetrie, asymetrie) myslíme míry (čísla), která charakterizují nerovnoměrné (nesouměrné) rozložení četností ve statistickém souboru. Pomocí nich jsme schopni odhadnout tvar rozdělení četností (resp. jeho souměrnost, nebo nesouměrnost), souměrné rozdělení četností má míry šikmosti nulové.

Míra šikmosti (založená na variačním rozpětí)

Jde o jednoduchou charakteristiku šikmosti co do výpočtu, ale jinak je to míra poměrně nedokonalá, ovlivněná maximální a minimální hodnotou souboru, které mohou být „nahodilé“. Hodnoty míry šikmosti se pohybují v intervalu (-1;1):

$$s = \frac{x_{\max} + x_{\min} - 2\tilde{x}}{x_{\max} - x_{\min}}.$$

Obdobným ukazatelem je i míra šikmosti založená na rozpětí kvantilů, jejich společným nedostatkem je to, že při výpočtu neuvažují hodnoty znaku, pouze vybrané extrémní nebo co do polohy významné hodnoty.

Koeficient šikmosti

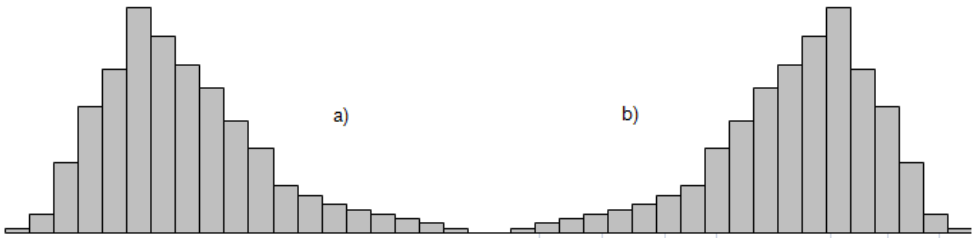
Tato míra šikmosti je, na rozdíl od míry šikmosti založené na variačním rozpětí, nebo založené na rozpětí kvantilů, dokonalejším ukazatelem, je definována jako aritmetický průměr z třetích mocnin odchylek jednotlivých hodnot znaku od aritmetického průměru vydělený třetí mocninou směrodatné odchylky (viz vzorec pro jeho výpočet ze skupinového rozdělení četností):

$$\alpha = \frac{\sum_{i=1}^{k} \left(x_s - \bar{x}\right)^3 \cdot n_i}{n \cdot s^3},$$

přičemž je-li: $\alpha > 0$, pak je rozdělení četností zešikmeno doleva (kladná šikmost)

$\alpha = 0$, pak je rozdělení četností souměrné (nulová šikmost)

$\alpha < 0$, pak je rozdělení četností zešikmeno doprava (záporná šikmost)



Obr. 10 Rozložení četností ve statistickém souboru – ukázky šikmosti:

a) $\alpha > 0$ (konkrétně 0,83), **b)** $\alpha < 0$, (-0,83). (Pramen: autor).

3.4 Charakteristiky špičatosti

Jedná se o čísla, která charakterizují koncentraci prvků souboru v blízkosti určité hodnoty znaku, jejich úkolem je poskytnout představu o tvaru rozdělení četností co do špičatosti nebo plochosti.

Míra koncentrace kolem mediánu

Tato míra špičatosti je, podobně jako míra šikmosti založené na variačním rozpětí a míra šikmosti založené na rozpětí kvantilů, nedokonalý ukazatel, který může být ovlivněn „nahodilými“ extrémními hodnotami:

$$K = \frac{x_{\max} - x_{\min}}{\tilde{x}_{75} - \tilde{x}_{25}}.$$

S rostoucím *K* je rozdělení četností „špičatější“ (dochází k větší koncentrovanosti hodnot v okolí mediánu), naopak s klesající hodnotou K se rozložení četností „zplošťuje“.

Koeficient špičatosti

Dokonalejším ukazatelem než míra koncentrace kolem mediánu je koeficient špičatosti. Je definován jako průměrná hodnota součtu čtvrtých mocnin odchylek hodnot znaku od arit-

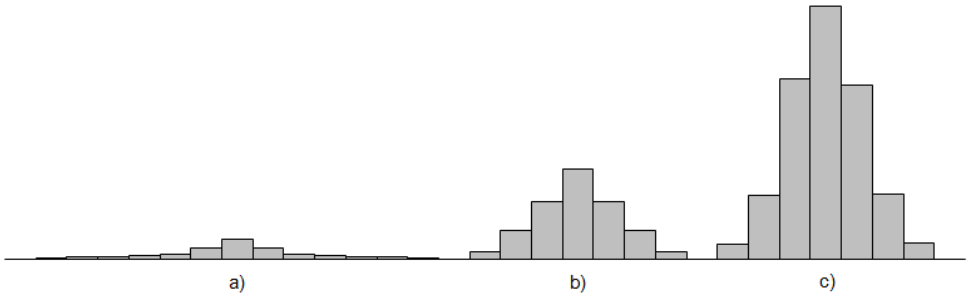
metického průměru dělených čtvrtou mocninou směrodatné odchylky (viz vzorec pro jeho výpočet ze skupinového rozdělení četností).

$$\varepsilon = \frac{\sum_{i=1}^{k} (x_s - \bar{x})^4 \cdot n_i}{n \cdot s^4}$$

Pokud: $\varepsilon > 0$, pak je rozdělení četností kladně zašpičatělé (špičaté)

$\varepsilon = 0$, pak je rozdělení četností normálně zašpičatělé

$\varepsilon < 0$, pak je rozdělení četností záporně zašpičatělé (ploché)



Obr. 11 Rozložení četností ve statistickém souboru - ukázky špičatosti:

a) $\varepsilon < 0$, **b)** $\varepsilon = 0$, **c)** $\varepsilon > 0$ (Pramen: autor).

Pro zájemce

Pokusíme se vysvětlit si termín ***„stupně volnosti“***, což je termín velmi často používaný v případě, že přecházíme v úvahách od výběrového souboru na soubor základní, většinou bezrozměrný. Je-li k dispozici pouze jedna naměřená nebo jinak zjištěná hodnota, tedy výběrový soubor o rozsahu $n = 1$, i takovýto výběr nám poskytuje informaci o průměru základního souboru. Ale nemáme žádnou možnost dozvědět se cokoliv o charakteristice variability výběru (odkud kam jsou hodnoty uspořádány, jak jsou rozmístěny...), o variabilitě zkrátka nemůžeme usuzovat z jedné jediné hodnoty. Uvažovat něco o rozptýlenosti dat můžeme od n většího než 1. Pro výpočet variability výběru a následně i její odhad pro základní soubor tak musíme nutně mít k dispozici n-1 jednotek. Člen n–1 tedy považujeme za správného dělitele pro výpočet rozptylu a směrodatné odchylky, který slouží k odhadům parametrů základního souboru. Příklady si ukážeme v následující kapitole.

Příklad / Příklad z praxe

Máte k dispozici intervalové rozdělení četností (viz níže). Spočítejte charakteristiky polohy – průměr, modus, určete interval, kde leží medián; variability – rozptyl, směrodatnou odchylku, variační koeficient; koeficient šikmosti a špičatosti.

interval č.	x_s	n_i	**interval č.**	x_s	n_i
1	0,5	0	**11**	10,5	24
2	1,5	1	**12**	11,5	31
3	2,5	2	**13**	12,5	36
4	3,5	3	**14**	13,5	42
5	4,5	4	**15**	14,5	48
6	5,5	5	**16**	15,5	35
7	6,5	6	**17**	16,5	27
8	7,5	8	**18**	17,5	15
9	8,5	10	**19**	18,5	4
10	9,5	18	**20**	19,5	1

Úkol / Úkol k zamyšlení

Jak se změní vzorce pro výpočty charakteristik polohy, variability, šikmosti a špičatosti z intervalového rozdělení četností, nebude-li ve výpočtu figurovat absolutní četnost, ale relativní četnost?

SHRNUTÍ

Výpočty číselných charakteristik příslušných statistickým souborům také patří k elementárním dovednostem nezbytným pro schopnost porovnávání souborů mezi sebou a vyvozování primárních informací o sledovaném geografickém jevu. Jsme díky nim schopni najít typickou hodnotu jevu, kterou nejčastěji ztotožňujeme s průměrem, modem nebo mediánem, umíme posoudit, jak jsou data v souborech rozprostřena, jestli více či méně oscilují od střední hodnoty, je-li rozdělení četností symetrické, ploché nebo špičaté. Tyto prvky popisné statistiky představují základ pro pravděpodobnostní statistiku, která na tu popisnou bezprostředně navazuje.

Kontrolní otázky a úkoly

1. Vysvětli rozdíl mezi typickým a netypickým aritmetickým průměrem.
2. Která z charakteristik variability je nejpoužívanější a proč? Přibliž její výpočet.
3. Vysvětli, jaký je rozdíl mezi aritmetickým průměrem vypočítaným z netříděných dat a mezi průměrem vypočítaným z intervalového rozdělení četností (tj. neznáme všechny vstupní hodnoty, pouze intervaly a jim příslušné četnosti).

Pojmy k zapamatování

Pojem 1: průměr, modus, medián, kvantily

Pojem 2: směrodatná odchylka, průměrná odchylka, rozptyl

Pojem 3: variační koeficient, šikmost, špičatost

4 Teorie rozdělení

Cíl
Po prostudování této kapitoly budete umět:

- vysvětlit princip přechodu studia od výběrového souboru k základnímu,
- objasnit rozdíly mezi spojitou, nespojitou náhodnou veličinou a jejími rozděleními,
- posoudit extremitu geografických jevů.

Doba potřebná k prostudování kapitoly: **120 minut**.

Průvodce studiem
Z předcházející kapitoly umíme statistickým souborům vypočítat a přiřadit jejich číselné charakteristiky, na jejichž základě je i můžeme vzájemně srovnávat. To jsou nezbytné dovednost pro úspěšné zvládnutí kapitoly následující. Jejím cílem bude čtenáři přiblížit a objasnit přechod od studia výběrových souborů, resp. souborů s konečným rozsahem směrem k zobecnění, k úvahám, jak se chová soubor základní, bezrozměrný.
Dostáváme se tak od popisné statistiky ke statistice pravděpodobnostní, která má také za úkol přiřazovat hodnotám geografických jevů pravděpodobnosti, se kterými mohou nastat, nebo je klasifikovat z hlediska extremity, tzn. identifikovat, je-li údaj normální, podnormální apod.
Studium si rozdělíme na dvě části, zvlášť se podíváme na náhodné veličiny nespojité, z jiného pohledu na ty spojité. Bude nás zajímat tvar tzv. pravděpodobnostní křivky, která nám napoví o četnostech a jejich rozložení (rozdělení) v bezrozměrném statistickém souboru. Taková rozdělení nazveme „teoretická rozdělení" náhodné veličiny a uvedeme si jejich základní příklady v závislosti na jejich tvaru a dalších parametrech.

4.1 Náhodná veličina

Za náhodnou veličinu (v obecné rovině, nikoliv jenom v geografii) považujeme proměnnou, pro kterou nelze na základě určité zákonitosti předem stanovit její konkrétní hodnotu. Pokud tato proměnná může nabývat jakékoliv hodnoty (v určitém intervalu), nazveme ji *spojitou* náhodnou veličinou, v opačném případě hovoříme o veličině nespojité neboli *diskrétní*.

Příklady náhodných veličin geografii:

Spojité:

teplota, vlhkost a tlak vzduchu; srážkové úhrny; průtoky; hrubé míry – porodnosti, úmrtnosti apod.; index stáří; průměrný věk; míra nezaměstnanosti; dokončené byty na 1 000 obyvatel; atd.

Nespojité:

nejrůznější četnosti, např. četnosti srážkových období; četnosti věkových kategorií; počet suchých měsíců v roce; narození chlapce nebo dívky; apod.

4.2 Teoretické rozdělení náhodné veličiny

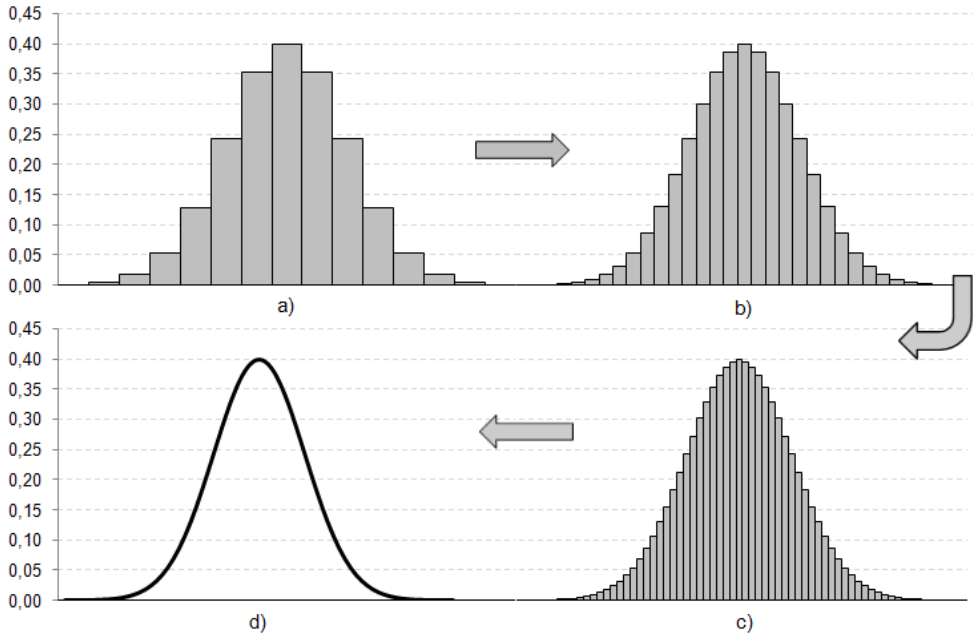
Ve statistice pracujeme často s výběrovými soubory o rozsahu n, jejichž grafickým znázorněním je histogram. Budeme-li zvětšovat rozsah souboru (při předpokladu, že náhodná

veličina je spojitá) a hodnoty třídit do stále menších intervalů, dostaneme histogramy, které se budou stále více blížit hladké křivce (viz obr. 12).

Této hladké křivky dosáhneme v teoretickém limitním případě, kdy soubor o nekonečně velkém rozsahu třídíme do nekonečně mnoha nekonečně úzkých intervalů. Dostaneme tak frekvenční (též pravděpodobnostní) funkci f(x) – neboli hustotu pravděpodobnosti. Analogicky bychom mohli přejít od součtové čáry ke spojité křivce F(x) – k tzv. distribuční nebo součtové funkci. Frekvenční funkce tak představuje teoretické rozdělení četností základního souboru o parametrech:

..... střední hodnota ..... směrodatná odchylka

Limitní přechod spočívá v neustálém třídění stále většího počtu hodnot do zvětšujícího se počtu zužujících se intervalů.



Obr. 12 Konstrukce frekvenční funkce tzv. limitním přechodem. (Pramen: autor).

4.2.1 Normální (Gaussovo) rozdělení

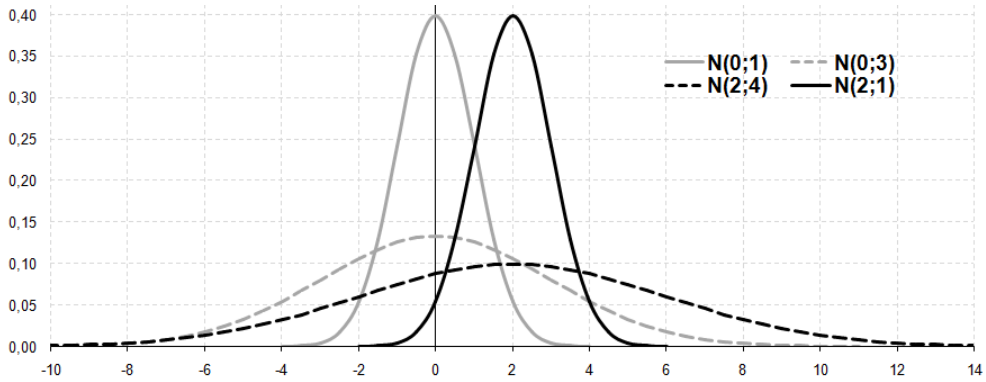
Patří mezi nejčastěji používaná rozdělení spojité náhodné veličiny. Bylo pozorováno při opakovaném měření téže veličiny za stálých podmínek, kdy se jednotlivé hodnoty více či méně odlišovaly od průměrné hodnoty. Normální rozdělení příslušné střední hodnotě μ a směrodatné odchylce σ je zpravidla označováno N(μ, σ^2).

Frekvenční funkce normálního rozdělení má tvar

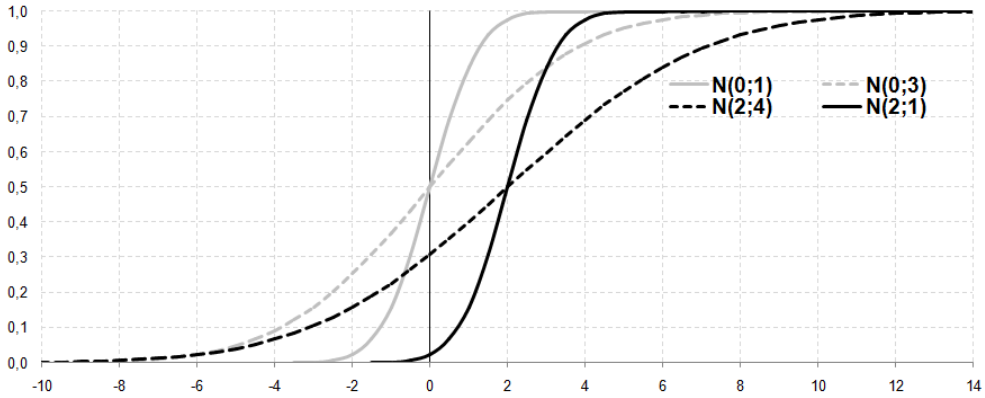
$$f(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

funkce distribuční pak

$$F(x) = \frac{1}{\sigma \cdot \sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$



Obr. 13 Ukázky Gaussových křivek příslušných normálním rozdělením. (Pramen: autor).



Obr. 14 Ukázky distribučních funkcí normálních rozdělení. (Pramen: autor).

Normované normální rozdělení

Jistou nevýhodou normálního rozdělení je jeho závislost na dvou parametrech (μ, σ^2), proto ho v praxi často normujeme pomocí „substitučního“ výrazu

$$z = \frac{x - \mu}{\sigma}.$$

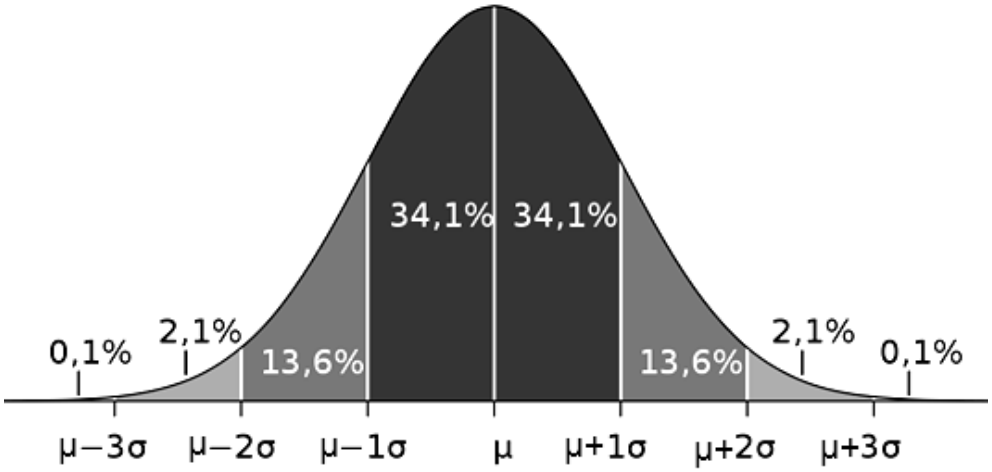
Po jeho aplikaci dostaneme frekvenční a distribuční funkci ve tvaru:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2}} \qquad F(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{-\frac{z^2}{2}} dz$$

Takto upravené„normované“ normální rozdělení již nezávisí na parametrech, a má následující vlastnosti:

- zvonovitý tvar, asymptoticky se přibližuje ose x,
- souměrná podle osy, která prochází vrcholem,
- x-ová souřadnice vrcholu je aritmetickým průměrem normálního rozdělení,
- aritmetický průměr se rovná modu a mediánu,

- normální křivka omezuje plochu 100 % (nebo 1),
- lze tak určit pravděpodobnosti, s nimiž leží hodnoty v určitém intervalu (viz obr. 15):
    - v intervalu $\mu \pm \sigma$ ... leží 68,28 % všech hodnot
    - v intervalu $\mu \pm 2\sigma$ ... leží 95,45 % všech hodnot
    - v intervalu $\mu \pm 3\sigma$ ... leží 99,73 % všech hodnot
- nebo z opačného pohledu
    - 95 % hodnot odpovídá intervalu $\mu \pm 1{,}65\,\sigma$
    - 99 % hodnot odpovídá intervalu $\mu \pm 2{,}58\,\sigma$



Obr. 15 rozložení hodnot pod křivkou normálních rozdělení. (Pramen: autor).

V geografii se často setkáváme s rozdělením jevů podle extremity. Tato typologie, která vychází z aplikace normálního rozdělení je uvedena v následující tabulce 3:

Tab. 3 Normální rozdělení a extremita jevů.

slovní označení extremity	**meze**	**pravděpodobnost výskytu jevu (%)**
extrémně podnormální	do $\mu - 3\sigma$	0,135
silně podnormální	μ –3 až –2 σ	2,190
podnormální	μ –2 až –1 σ	13,590
normální	μ –1 až +1 σ	68,270
nadnormální	μ +1 až +2 σ	13,590
silně nadnormální	μ +2 až +3 σ	2,190
extrémně nadnormální	μ +3 σ a více	0,135

Pramen: Autor.

Pro zájemce

Sestrojit grafy frekvenčních a distribučních funkcí normálního rozdělení v softwarovém rozhraní tabulkového procesoru Excel je poměrně snadné. Stačí využít funkce „normdist" a vhodně zadat parametry – pro jaká *x* hledáme hodnotu frekvenční resp. distribuční funkce; střední hodnotu; směrodatnou odchylku a požadavku na frekvenční („nepravda") nebo distribuční funkci („pravda"). Hodnoty, které dostaneme, pak snadno vyneseme do bodového grafu.

Ne všechny geografické jevy se ale řídí normálním rozdělením. Data, která máme k dispozici, musíme buď transformovat (vhodnou transformací, např. logaritmickou) nebo využít některá z dalších teoretických rozdělení spojité náhodné veličiny. Nejčastějšími příklady jsou rozdělení: Fisherovo (též F-rozdělení), Studentovo (t-rozdělení), nebo rozdělení χ^2 („chí kvadrát"). Jejich konstrukce a vlastnosti vychází ze stejných principů, které jsme ukázali u normálního rozdělení.

Příklad / Příklad z praxe

Víme-li, že se studovaná veličina řídí určitým rozdělením, máme v ruce silný nástroj k tomu, abychom mohli určit s jakou pravděpodobností bude její určitá mez překročena, kolik hodnot z uskutečněných měření padne do určitého intervalu atd.
Jestliže má veličina N normované normální rozdělení – tj. N(0,1) – určete:
a) pravděpodobnost, že $N > 1{,}64$
b) pravděpodobnost, že $N < -1{,}64$
c) pravděpodobnost, že $1{,}0 < N < 1{,}5$
d) pravděpodobnost, že $-2 < N < 2$
Doporučení: Pracujte s distribuční funkcí normálního rozdělení, nebo v Excelu vhodně využijte funkci NORMDIST.
Výsledky: a) 5,1 %; b) 5,1 %; c) 9,2 %; d) 95,4 %.

Úkol / Úkol k zamyšlení

Čas potřebný na vypracování testu na VŠ má normální rozdělení s průměrnou dobou 105 minut a směrodatnou odchylkou 20 minut.

a) kolik procent studentů dokončí test do dvou hodin?
b) kolik času by mělo být dáno, aby test mohlo dokončit 95 % studentů?

4.2.2 Binomické rozdělení

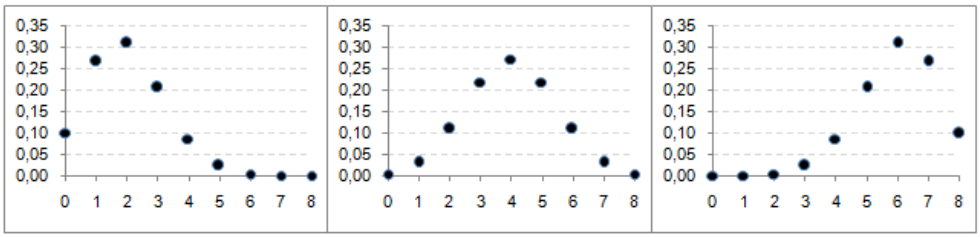
Průvodce studiem
Podívejme se na problematiku nespojité náhodné veličiny nejdříve „negeografickým" způsobem.
Basketbalista v tréninku pravidelně promění z 10 sedmimetrových hodů 7. Zajímá nás, kolik jich promění s největší pravděpodobností v zápase, hází-li sedmimetrových hodů 20. S jakou pravděpodobností promění více než 15 hodů? S jakou pravděpodobností promění přesně 15 hodů z 20?
To je typický příklad na binomické rozdělení. Proč? Existují pouze dvě varianty výsledku pokusu, který je v tomto případě hod na koš. Buď hráč promění, nebo nikoliv. V našem případě proměňuje v koš 7 hodů z 10, tzn. pravděpodobnost úspěchu (p) je 7/10, tj. 0,7 (nebo též 70 %). Pravděpodobnost neúspěchu (q) je logicky 1 – 0,7 = 0,3. Jak vypadá frekvenční a distribuční funkce tohoto rozdělení? Jaké jsou odpovědi na naše otázky? Dozvíme se v následující podkapitole.

Na rozdíl od normálního rozdělení je binomické rozdělení nejtypičtějším rozdělením diskrétní náhodné veličiny. Udává rozdělení výsledků při opakování jednoho a téhož pokusu za stejných podmínek, přičemž výsledkem pokusu mohou být pouze 2 alternativy: A, nebo B. Pravděpodobnost, že nastane alternativa A označíme jako *p*, pravděpodobnost, že nastane alternativa B, označíme jako *q*, přitom musí platit, že $p + q = 1$.

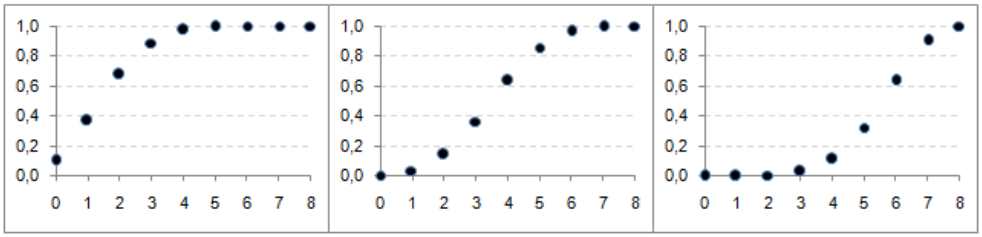
Za předpokladu, že provedeme uvažovaný pokus *n*-krát, hledáme pravděpodobnost, že alternativa *A* (s pravděpodobností p) nastane právě *x*-krát.

Výpočet pravděpodobnosti provádíme pomocí následující rovnice, která vlastně udává obecný člen binomického rozvoje a vyjadřuje rozdělení pravděpodobností binomického rozdělení:

$$f(x) = \binom{n}{x} \cdot p^x \cdot q^{n-x} = \frac{n!}{x!\cdot(n-x)!} \cdot p^x \cdot q^{n-x}.$$



Obr. 16 Ukázky frekvenčních funkcí binomického rozdělení pro n = 8 a postupně p = 0,25; 0,5 a 0,75. (Pramen: autor).



Obr. 17 Ukázky distribučních funkcí binomického rozdělení pro n = 8 a postupně p = 0,25; 0,5 a 0,75. (Pramen: autor).

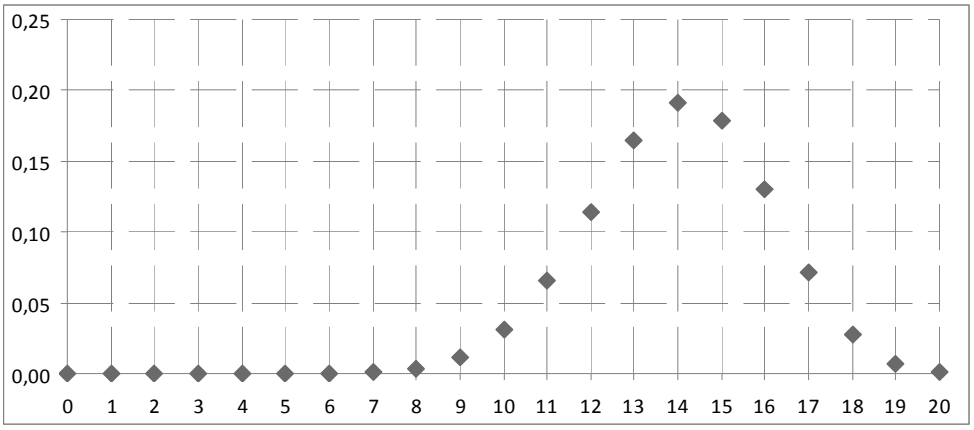
Pro zájemce

Modelovat grafy frekvenčních a distribučních funkcí binomického rozdělení lze velmi jednoduše v softwarovém rozhraní tabulkového procesoru Excel, a to s využitím funkce „BINOMDIST“ (binomial distribution) a vhodně zadaných parametrů – pravděpodobnost úspěchu (*p*), počet pokusů a požadavku na frekvenční („nepravda“) nebo distribuční funkci („pravda“).

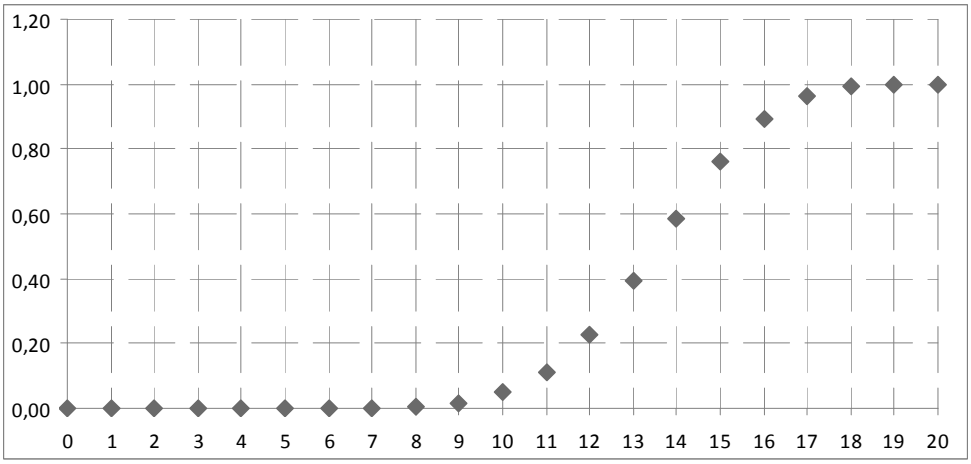
Příklad / Příklad z praxe

Zodpovězení otázek z „průvodce studiem“ na str. 32: Pravděpodobnost $p = 0,7$, počet pokusů $n = 20$. Frekvenční a distribuční funkce tohoto binomického rozdělení – viz obr. 18 a 19. Pravděpodobnost proměnění právě 15 hodů je hodnota frekvenční funkce pro $x = 15$, tj. 0,179 (17,9 %). Pravděpodobnost, se kterou hráč promění méně než 15 hodů je hodnotou distribuční funkce pro $x = 14$, tj. 0,584 (58,4 %). Pravděpodobnost, se kterou promění

více než 15 hodů je 1 – (minus) hodnota distribuční funkce pro $x = 15$ (proměnění 15 nebo méně hodů), tj. 1 – 0,762 = 0,238 (23,8 %).



Obr. 18 Frekvenční funkce binomického rozdělení pro n = 20 a p = 0,7. (Pramen: autor).



Obr. 19 Distribuční funkce binomického rozdělení pro n = 20 a p = 0,7. (Pramen: autor).

Teorie binomického rozdělení se v geografii často využívá, např. při stanovování pravděpodobností roků s určitým počtem suchých měsíců apod. Dalším z příkladů teoretického rozdělení nespojité náhodné veličiny je například rozdělení Poissonovo.

Úkol / Úkol k zamyšlení

Pokuste se vymyslet vhodné uplatnění binomického rozdělení na geografické jevy.

SHRNUTÍ

Kapitola „teoretická rozdělení“ představuje stručný vstup do problematiky pravděpodobnostní statistiky. Nejdůležitějším poznatkem je tzv. limitní přechod, kdy se snažíme sestrojit hladkou křivku teoretického rozdělení, tzv. hustotu pravděpodobnosti (resp. frekvenční funkci). Podstatou je snaha objasnit chování základního souboru, vycházíme přitom ze

souboru výběrového, jehož rozložení četností a číselné charakteristiky jsme schopni spočítat a graficky prezentovat.

Kontrolní otázky a úkoly

1. Co je Gaussova křivka, jak ji lze sestrojit, co znamená tzv. limitní přechod od histogramu k hladké křivce?
2. Uveď příklady spojitých a nespojitých geografických veličin.
3. Sestroj s využitím funkcí MS Excel frekvenční funkce normálních rozdělení N(0,3) a N(2,6) a binomického rozdělení pro p = 0,4 a n = 10.

Pojmy k zapamatování

Pojem 1: spojitá, nespojitá náhodná veličina

Pojem 2: teoretické rozdělení a jeho příklady

Pojem 3: distribuční, frekvenční funkce

5 Odhady parametrů

Cíl
Po prostudování této kapitoly budete umět:

- bodově odhadnout střední hodnotu a směrodatnou odchylku základního souboru,
- intervalově a s určitou pravděpodobností odhadnout střední hodnotu a směrodatnou odchylku základního souboru,
- vlastními slovy popsat princip a význam odhadování parametrů.

Doba potřebná k prostudování kapitoly: **60 minut**.

Průvodce studiem
Již v minulé kapitole jsme uvedli, že při zpracovávání dat, analýzách i při vytváření teorií pracujeme častěji se soubory výběrovými než základními. Děje se tomu tak hned z několika důvodů. Práce se základními soubory může být velice komplikovaná pro jejich velký rozsah (v některých případech i nekonečnost), v řadě případů se musíme na výběrový soubor spolehnout z důvodu náročnosti měření, nebo jiného šetření. Pokud je výběr ze základního souboru proveden náhodně (tzn., že každý člen základního souboru má stejnou pravděpodobnost dostat se do základního výběru), hovoříme o tzv. náhodném výběrovém souboru.
Cílem této kapitoly je naučit se odhadovat charakteristiky (střední hodnotu, rozptyl, směrodatnou odchylku) základního souboru pomocí charakteristik souboru výběrového. V praxi to znamená, že usuzujeme, postupujeme či přecházíme z části na celek a zevšeobecňujeme závěry, používáme tedy statistickou indukci.

5.1 Princip odhadů

Rozeznáváme dva základní typy odhadů – bodový a intervalový. Bodový je jednodušší, ale vhodnější je použití intervalového.

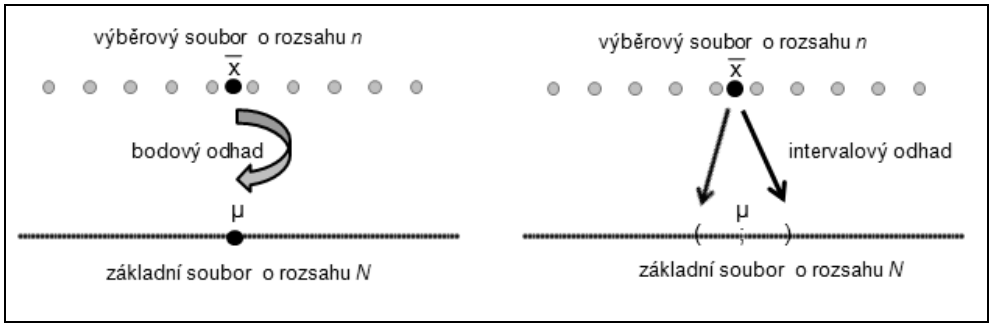
Neznámé charakteristiky základního souboru odhadujeme pomocí příslušných výběrových charakteristik s určitou přesností a spolehlivostí. Přesnost odhadu dané charakteristiky je určena násobkem střední výběrové chyby, kterou je směrodatná odchylka příslušné charakteristiky ze všech teoreticky možných výběrů. Spolehlivost odhadu je dána pravděpodobností, se kterou je možné určitý odhad považovat za správný. Určení přesnosti a spolehlivosti odhadu předpokládá znalost rozdělení výběrových charakteristik. U velkých výběrů (zpravidla při n > 30) se výběrové rozdělení aproximuje většinou rozdělením normálním, zatímco u souborů menších (n < 30) uvažujeme jiná rozdělení. Kvalita výběru je podmíněna tím, jakou metodou je proveden, správné reprezentativnosti dosahujeme zpravidla náhodným výběrem.

Rozeznáváme dva základní typy odhadů – bodový a intervalový. Abychom si mohli vysvětlit princip, na jakém jsou založeny, uveďme si nyní terminologicky vztahy mezi výběrovým a základním statistickým souborem:

- základní soubor: N je rozsah, a_i je i-tý prvek základního souboru, μ je aritmetický průměr základního souboru, σ je pak směrodatná odchylka základního souboru,
- výběrový soubor: n je rozsah, x_i označuje i-tý prvek výběrového souboru, $\overline{x}$ aritmetický průměr výběrového souboru (výběrový průměr) a s směrodatnou odchylku výběrového souboru (výběrová směrodatná odchylka).

Jak už je zřejmé z názvu metody, bodový odhad charakteristik základního souboru provedeme pomocí jedné hodnoty, zatímco při odhadu intervalovém konstruujeme interval,

ve kterém bude střední hodnota základního souboru s určitou pravděpodobností ležet (schematicky viz obr. 20).



Obr. 20 Princip odhadování parametrů (Pramen: autor).

5.1.1 Bodové odhady

Bodový odhad střední hodnoty základního souboru je dán následujícím vztahem

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

Ze vzorce je zřejmé, že bodový odhad střední hodnoty základního souboru stanovíme jako aritmetický průměr souboru výběrového.

U bodového odhadu směrodatné odchylky je situace poněkud složitější. Používá se tu princip statistické (matematické) indukce, pro naše potřeby postačí, uvedeme-li si bodový odhad směrodatné odchylky základního souboru přímo bez odvozování

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

Vztah vychází z definice směrodatné odchylky statistického souboru, při aplikování postupu statistické indukce je ale ve jmenovateli vzorce „$n - 1$“ namísto pouhého „n“, ve výše uvedeném vzorci „s“ znamená směrodatnou odchylku výběrového souboru. Výraz „$n - 1$“ ve jmenovateli vyjadřuje tzv. „stupně volnosti“, tento termín jsme si objasnili v sekci „pro zájemce“ v kapitole 3.

Následujícího vztahu se využívá při odhadu směrodatné odchylky výběrových průměrů, jež je důležitá při určování spolehlivosti či přesnosti odhadu hledané charakteristiky, v tomto případě aritmetického průměru:

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n-1}}$$

Při bodových odhadech dochází volbou různých výběrů k nepřesnostem, hodnoty výběrových charakteristik se s různými výběry liší a bodové odhady jsou tak zatíženy chybou. Je tedy nezbytné určit odchylky od skutečných charakteristik základního souboru, jinými slovy určit přesnost a těsnost odhadu. K tomu využíváme intervalové odhady neboli intervaly spolehlivosti.

5.1.2 Intervalové odhady

V kapitole věnované normálnímu rozdělení jsme pomocí aritmetického průměru μ a násobků směrodatné odchylky σ základního souboru stanovili pravděpodobnosti (resp. meze pravděpodobností), s nimiž hodnoty sledovaného jevu leží v určitých intervalech.

Například pokud zvolíme za tyto meze hodnoty $\mu \pm 3\sigma$, znamená to, že všechny odchylky od střední hodnoty, které neleží v těchto mezích, tzn. za nepřípustné budeme považovat odchylky $(x_i - \mu) > 3\sigma$ a $(x_i - \mu) < -3\sigma$, přičemž vnitřní interval omezený trojnásobkem σ považujeme za interval spolehlivosti a hodnoty $\mu \pm 3\sigma$ nazveme meze spolehlivosti.

Kritický obor je tvořen intervaly, které navazují na interval spolehlivosti (z obou stran). Plocha omezená částí normální křivky a pořadnicemi v bodech mezí spolehlivosti se nazývá oblast přijetí, ostatní část plochy je tzv. oblast zamítnutí.

Z teorie normálního rozdělení víme, že:

v intervalu	...	$\mu \pm \sigma$	...	leží 68,28 % všech hodnot
		$\mu \pm 2\sigma$	...	leží 95,45 % všech hodnot
		$\mu \pm 3\sigma$	...	leží 99,73 % všech hodnot

a naopak	...	95,0 % hodnot odpovídá intervalu	$\mu \pm 1{,}960\sigma$
		99,0 % hodnot odpovídá intervalu	$\mu \pm 2{,}576\sigma$
		99,9 % hodnot odpovídá intervalu	$\mu \pm 3{,}291\sigma$

Z předešlého odstavce a vlastností normálního rozdělení vyplývá i následující tabulka nejčastěji používaných intervalů spolehlivosti:

Tab. 4 Kritické obory pro intervaly spolehlivosti.

násobky směrodatné odchylky	oblast	
	přijetí	zamítnutí
± 1,960	95,0 %	5,0 %
± 2,576	99,0 %	1,0 %
± 3,291	99,9 %	0,1 %

Pramen: Autor.

Šířka intervalu spolehlivosti záleží na rozsahu náhodného výběru – čím je rozsah větší, tím je přesnější odhad skutečné hodnoty odhadovaného parametru. Intervaly spolehlivosti podle jednotlivých výběrů se od sebe liší, neboť jsou rozdílné charakteristiky jednotlivých výběrů (podobnou situaci jsme již řešili v případě bodových odhadů). Nicméně stanovíme-li 95% interval spolehlivosti na základě jednoho náhodného výběru, zahrne s pravděpodobností 95 % skutečnou hodnotu odhadovaného parametru.

Konstrukce intervalového odhadu střední hodnoty základního souboru μ pro výběrové soubory s rozsahem n > 30:

Z kapitoly o principech odhadů parametrů víme, že

$$\mu_{\bar{x}} = \mu \qquad \text{a} \qquad \sigma_{\bar{x}} = \frac{\sigma}{n},$$

takže v souladu s předchozí teorií např. interval $\mu \pm 2{,}576\ \sigma$ zahrne 99 % všech výběrových průměrů. Výběrový průměr je téměř s jistotou součástí daného intervalu, tedy můžeme psát, že:

$$\mu_{\bar{x}} - 2{,}576 \cdot \sigma_{\bar{x}} \le \bar{x} \le \mu_{\bar{x}} + 2{,}576 \cdot \sigma_{\bar{x}}\,.$$

Násobek směrodatné odchylky nahradím výrazem u_p, kde index p značí pravděpodobnost (vyjádřenou desetinným číslem), se kterou náhodná veličina překročí kritickou hodnotu. Pro $p = 0{,}01$ je $u_p = \pm 2{,}576$.

$$\mu_{\bar{x}} - u_p \cdot \sigma_{\bar{x}} \le \bar{x} \le \mu_{\bar{x}} + u_p \cdot \sigma_{\bar{x}}$$

Dalšími úpravami dostaneme:

$$\bar{x} - u_p \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + u_p \cdot \frac{\sigma}{\sqrt{n}}\,.$$

Směrodatnou odchylku σ většinou neznáme, proto ji nahradíme jejím bodovým odhadem a dostaneme pro intervalový odhad střední hodnoty následující vztah:

$$\bar{x} - u_p \cdot \frac{s}{\sqrt{n-1}} \le \mu \le \bar{x} + u_p \cdot \frac{s}{\sqrt{n-1}}\,.$$

Předchozí vzorec je tzv. ***intervalem spolehlivosti***.

Při praktických analýzách, výpočtech a šetřeních často potřebujeme určit rozsah n náhodného výběru, aby spolehlivě (s určitou pravděpodobností) reprezentoval základní soubor, jinými slovy řečeno, aby se z dat výběru podařilo odhadnout neznámou charakteristiku (v tomto případě průměr) s předem zvolenou přesností. Rozsah tohoto výběru je dán následujícím vztahem:

$$n = u_p{}^2 \cdot \frac{s^2}{\delta^2},$$

kde δ je polovina požadované šířky intervalu spolehlivosti (neboli dané přesnosti).

Směrodatná chyba aritmetického průměru je dána vztahem:

$$c_{\bar{x}} = \frac{s}{\sqrt{n}}\,.$$

Pomocí ní jsme schopni určit pravděpodobnou chybu výběrového průměru, např. ze vztahu:

$$pc_{\bar{x}} = 0{,}675\frac{s}{\sqrt{n}}.$$

Přičemž tuto rovnici můžeme použít ke zjištění rozsahu výběru nutného k odhadu průměru tak, aby jeho chyba měla předem zvolenou velikost. Musíme vyjít ze vztahu:

$$\frac{N'}{n} = \frac{pc_{\bar{x}}^{\ 2}}{PC_{\bar{x}}^{\ 2}},$$

kde n je rozsah výběru, z něhož byl průměr vypočten, N' je hledaný rozsah, čitatel vpravo je vypočtená pravděpodobná chyba a jmenovatel je zvolená pravděpodobná chyba.

Konstrukce intervalového odhadu střední hodnoty základního souboru μ pro výběrové soubory s rozsahem n < 30:

Problematika této úlohy je založena na stejném principu jako intervalový odhad pro soubory o velkém rozsahu, s jednou jedinou změnou – a to že hodnoty u_p zaměníme hodnotami t_p, tedy kritickými hodnotami t-rozdělení pro n – 1 stupňů volnosti (tyto hodnoty jsou uvedeny v tabulkách). Výsledná podoba ***intervalu spolehlivosti***:

$$\bar{x} - t_p \cdot \frac{s}{\sqrt{n-1}} \le \mu \le \bar{x} + t_p \cdot \frac{s}{\sqrt{n-1}}.$$

Rozdíl mezi kritickými hodnotami t_p a u_p se s rostoucím rozsahem výběrového souboru zmenšuje, pokud je $n > 30$, můžeme místo kritických hodnot t_p t-rozdělení použít kritické hodnoty u_p normálního rozdělení.

Konstrukce intervalového odhadu směrodatné odchylky σ základního souboru:

Interval spolehlivosti pro směrodatnou odchylku základního souboru dostaneme aplikací následujícího vzorce:

$$\sqrt{\frac{ns^2}{\chi^2_{0,5p}}} \le \sigma \le \sqrt{\frac{ns^2}{\chi^2_{1-0,5p}}},$$

kde χ^2_p jsou kritické hodnoty teoretického rozdělení χ^2 s n – 1 stupni volnosti, které najdeme v tabulkách.

Pozn.: Index p opět značí pravděpodobnost (vyjádřenou desetinným číslem), se kterou náhodná veličina překročí kritickou hodnotu. Tzn., že hledáme-li 99% interval spolehlivosti, je p = 0,01.

Pro zájemce

Vypočítat pravděpodobnostní intervaly spolehlivosti pro střední hodnoty základního souboru lze opět velmi jednoduše v Excelu, a to s využitím funkce „CONFIDENCE“ a vhodně zadaných parametrů. Výsledkem výpočtu je polovina šířky hledaného intervalu spolehlivosti, jeho dolní (horní) hranici dostaneme odečtením (přičtením) této hodnoty od aritmetického průměru výběrového souboru.

Příklad / Příklad z praxe

Náhodný výběr 5 států USA má následující rozlohy (tis. mil čtverečních):

147 84 24 85 159

a) vypočtěte 95% interval spolehlivosti pro střední rozlohu všech 50 států USA
b) vypočtěte 95% interval spolehlivosti pro celkovou rozlohu USA
c) je její skutečná hodnota (3 620 000) zahrnuta v tomto intervalu?

Doporučení: Využijte vzorce uvedené v této kapitole, nebo v rozhraní Excel statistickou funkci „CONFIDENCE“.

Řešení: a) (31 942; 167 658); b) (1 597 109; 8 382 891); c) ANO

Úkol / Úkol k zamyšlení

Máte k dispozici výběrový soubor, se kterým jsme již pracovali. Na jeho základě sestrojte 95% a 99% intervaly spolehlivosti pro střední hodnotu souboru základního.

7,4	8,3	8,5	10,9	7,9	10,8	9,9	9,4	9,3	8,5
9,6	9,4	8,2	9,7	8,4	9,4	10,7	8,8	9,5	9,0
8,1	10,3	7,7	8,8	8,6	9,8	9,4	8,9	9,6	9,2
9,1	9,9	10,0	8,9	10,2	9,3	9,6	8,7	9,9	9,4
7,9	10,1	11,1	9,3	10,5	8,5	9,1	9,1	8,8	9,6

SHRNUTÍ

Relativní jednoduchost výpočtu bodového odhadu základního statistického souboru má svá úskalí v tom, že může být zkreslený. Proto považujeme za efektnější a nakonec i efektivnější metodu intervalového odhadu. S využitím pokročilého statistického softwaru, ale i běžně dostupného Excelu, nejde o nikterak náročnou proceduru. Pro korektní interpretaci sestrojených intervalů spolehlivosti a pochopení jejich konstrukce je nezbytné zvládnutí kapitoly o teoretických rozděleních náhodných veličin.

Kontrolní otázky a úkoly

1. Vysvětli rozdíl mezi bodovým a intervalovým odhadem.
2. Který z intervalů spolehlivosti je širší: 95% nebo 99%?
3. Popiš vlastními slovy princip odhadů, jaké může být jejich uplatnění v geografii?

Pojmy k zapamatování

Pojem 1: výběrový průměr, výběrová směrodatná odchylka, bodový odhad

Pojem 2: intervalový odhad, interval spolehlivosti

Pojem 3: stupně volnosti

6 Testování statistických hypotéz

Cíl
Po prostudování této kapitoly budete umět:

- posoudit statistickou významnost rozdílu mezi středními hodnotami souborů,
- posoudit, zda soubor pochází z určitého teoretického rozdělení,
- korektně formulovat pracovní a nulovou hypotézu.

Doba potřebná k prostudování kapitoly: **60 minut**.

Průvodce studiem
Cílem celé problematiky je ověření určitého předpokladu. Nejčastěji zjišťujeme, zda zkoumaný výběr pochází ze základního souboru, který má určité rozdělení, nebo můžeme ověřovat, zda dva výběry pocházejí z téhož základního souboru (zda jsou rozdíly mezi jejich charakteristikami statisticky významné, či nikoliv).

6.1 Princip testování

Obecný postup testování, prakticky využitelný pro naprostou většinu statistických testů se řídí souborem pravidel uvedených v následujících šesti krocích:

- Zvolíme hladinu významnosti (označujeme ji p, hladina významnosti je vlastně pravděpodobnost, že náhodná odchylka překročí danou hodnotu – tzv. kritickou hodnotu. Snažíme se ji tedy volit co nejnižší, zpravidla $p = 0{,}05$ (5 %), nebo $p = 0{,}01$ (1 %), přičemž odchylky, které se vyskytují s pravděpodobností menší, než je hladina významnosti, označujeme za statisticky významné na zvolené hladině významnosti).

- Formulujeme nulovou hypotézu. Statistickou hypotézou rozumíme každý předpoklad o neznámé vlastnosti základního souboru, zatímco nulová hypotéza (H_0), neboli prověřovaná hypotéza, je „speciální hypotézou“ o charakteristikách základního souboru. Nulová hypotéza je zpravidla negací pracovní hypotézy, pro jejíž ověření byl daný pokus (nebo pozorování) uspořádán.

- Zvolíme vhodné testovací kritérium (závisí na povaze řešeného problému). Každé testovací kritérium má své určité rozdělení – např. t-rozdělení, χ^2 („chí kvadrát“) rozdělení, F-rozdělení...).

- Vypočteme velikost testovacího kritéria.

- Porovnáme tuto hodnotu s kritickou hodnotou. Ve statistických tabulkách jsou uvedeny kritické hodnoty rozdělení příslušných testovacím kritériím pro nejčastěji používané hladiny významnosti a pro různé rozsahy výběru (tzv. stupně volnosti).

- Vyslovíme závěr. O platnosti testované hypotézy rozhodneme po porovnání vypočtené hodnoty testovacího kritéria s kritickou hodnotou z tabulek, tzn. je-li vypočtené kritérium větší než kritická hodnota, obecně nastává případ, který jsme očekávali s nepatrnou pravděpodobností (tzn. 5 nebo 1 %). Usuzujeme, že takový případ je téměř nemožný a že testovaná odchylka nemá charakter náhodný. Zamítáme nulo-

vou hypotézu a vyslovujeme závěr, že na zvolené hladině významnosti je rozdíl mezi testovanými charakteristikami statisticky významný. Je-li vypočtené testovací kritérium menší než tabulková kritická hodnota, nastal případ, který očekáváme s pravděpodobností 1 – p (tedy s pravděpodobností 95 nebo 99 %), tedy s takovou pravděpodobností, že jeho výskyt můžeme považovat za téměř jistý. Usuzujeme, že rozdíl mezi testovanými charakteristikami není a nezamítáme nulovou hypotézu. Na zvolené hladině významnosti není rozdíl statisticky významný.

6.1.1 χ^2 – test

Tento test je nejčastěji využívaným testem shody.

Tento test se nazývá testem shody, jeho princip spočívá v tom, že posuzujeme, jak se rozložení četností pozorovaného (výběrového) souboru liší od základního souboru. Při jeho použití dáváme do souvislosti empirické hodnoty zjištěné ze statistického šetření a teoretické (očekávané) hodnoty. Hodnotíme rozdíly mezi četnostmi pozorovanými a teoretickými.

Vzorec pro výpočet testového kritéria má tvar:

$$\chi^2 = \sum_{j=1}^{k} \frac{(n_{e,j} - n_{t,j})^2}{n_{t,j}},$$

kde $n_{e,j}$ jsou empirické četnosti a $n_{t,j}$ teoretické četnosti.

Takto definované testové kritérium má χ^2 rozdělení s $k - 1$ stupni volnosti (k je počet intervalů). Kritické hodnoty tohoto rozdělení najdeme v tabulkách. Kromě χ^2 – testu, který se nedá vždy použít, můžeme zvolit jiný test shody, a to Kolmogor-Smirnovův test. Ten nevychází z pravděpodobnostní funkce rozdělení, ale z funkce distribuční.

Příklad / Příklad z praxe

Máte k dispozici intervalové rozdělení četností (empiricky zjištěné četnosti). Pomocí testu shody (test χ^2) ověřte na hladině významnosti $p = 0{,}05$, zda tento výběr pochází ze souboru základního, který má normální rozdělení.

třída	středy tříd	$n_{e,j}$
1	7,5	5
2	8,0	9
3	8,5	20
4	9,0	32
5	9,5	34
6	10,0	44
7	10,5	39
8	11,0	8
9	11,5	8
10	12,0	1

Doporučený postup: Využijte Excel, vypočítejte charakteristiky výběru – průměr a směrodatnou odchylku. Nulová hypotéza: „odlišnost mezi n_{ei} a n_{ti} je náhodná“. K jednotlivým intervalům pomocí funkce „NORMDIST“ spočítejte teoretické četnosti, jako parametry normálního rozdělení využijte charakteristiky výběru. Pomocí testového kritéria (χ^2-testu) otestujte shodu mezi empirickými a teoretickými četnostmi.

6.1.2 F-test

F-test, neboli test rozptylů, vždy předchází t-testu.

Pomocí tohoto testu zjišťujeme významnost rozdílu mezi dvěma rozptyly. Za testové kritérium uvažujeme poměr odhadů dvou rozptylů základního souboru:

$$F = \frac{\hat{\sigma}_1^{\,2}}{\hat{\sigma}_2^{\,2}}.$$

Takto definované testové kritérium má Fisherovo (F) rozdělení a jeho kritické hodnoty najdeme opět v tabulkách.

6.1.3 t-test

Posledním typem testu, který si ukážeme, je t-test. Je založen na podobném principu jako předchozí F-test a používáme ho k testování rozdílu výběrového průměru a známého průměru základního souboru, nebo k testování významnosti rozdílu dvou výběrových průměrů, a to v případě, že F-testem jsme ověřili rovnost rozptylů, a t-test můžeme použít i k testování rozdílu dvou výběrových průměrů, jestliže jsme F-testem ověřili nerovnost rozptylů.

Testové kritérium k testování rozdílu mezi průměrem výběrového souboru a známým průměrem základního souboru (počet stupňů volnosti je n–1):

$$t = \frac{\left|\bar{x} - \mu\right| \cdot \sqrt{n-1}}{s}.$$

Testové kritérium k testování významnosti rozdílu dvou výběrových průměrů ze předpokladu rovnosti rozptylů výběrových souborů je dáno vztahem (počet stupňů volnosti je v tomto případě $n_1 + n_2 - 2$):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{n_1 s_1^{\,2} + n_2 s_2^{\,2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}.$$

Postup při použití tohoto testu je opět podobný s obecným postupem testování statistických hypotéz, nejdříve zvolíme hladinu významnosti p, poté vypočítáme aritmetické průměry a směrodatné odchylky obou souborů, ověříme nulovou hypotézu F-testem, vypočítáme hodnotu testového kritéria, určíme počet stupňů volnosti a najdeme pro ně příslušnou kritickou hodnotu t_p, porovnáme ji s hodnotou t a vyslovíme závěr, tzn. je-li $t > t_p$ zamítáme nulovou hypotézu a tvrdíme, že rozdíl průměrů je statisticky významný na zvolené hladině významnosti (popř. že se výběrový průměr na zvolené hladině významnosti významně liší od známé hodnoty aritmetického průměru základního souboru), v opačném případě nulovou hypotézu nezamítáme a považujeme rozdíl průměrů za nevýznamný.

V případě, že F-testem zjistíme, že mezi rozptyly je statisticky významný rozdíl, testové kritérium k testování významnosti rozdílu dvou průměrů bude mít tuto podobu

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{\sqrt{\frac{s_1^{\,2}}{n_1 - 1} + \frac{s_2^{\,2}}{n_2 - 1}}}.$$

Hodnotu testového kritéria v tomto případě nebudeme porovnávat s kritickou hodnotou z tabulek, ale s hodnotou t_p^+, kterou vypočítáme podle vzorce

$$t_p^+ = \frac{\frac{s_1^2}{n_1 - 1} \cdot t'_p + \frac{s_2^2}{n_2 - 1} \cdot t''_p}{\frac{s_1^2}{n_1 - 1} + \frac{s_2^2}{n_2 - 1}},$$

kde t'_p a t''_p jsou tabulkové hodnoty pro 1 a 2 stupňů volnosti.

Pomocí t-testu lze také testovat soubory, které vzniknou měřením ukazatelů dvakrát, pokaždé za jiných podmínek (pak se jedná o tzv. t-test pro párové hodnoty), který má své vlastní testové kritérium, založené na rozdílech jednotlivých párových hodnot.

Pro zájemce

Při testování se můžeme dopustit chyb. Například té, že nulová hypotéza platí a my jsme ji zamítli (tzv. chyba 1. druhu) anebo nulová hypotéza neplatí a my jsme ji testem nezamítli (tzv. chyba 2. druhu). Čtyři možné případy, které mohou při testech nastat, uvádí tab. 5.

Tab. 5 Možné výsledky testování statistických hypotéz.

	výsledek testu	
realita	**H_0 nezamítáme**	**H_0 zamítáme**
H_0 platí	rozhodli jsme správně	chyba I. druhu
H_0 neplatí	chyba II. druhu	rozhodli jsme správně

Pramen: Autor

Příklad / Příklad z praxe

Máme k dispozici dva výběrové soubory o rozsahu n = 31 hodnot s následujícími charakteristikami:
1. soubor: průměr 8,65; rozptyl 8,53
2. soubor: průměr 9,44; rozptyl 9,78
Otestujte na hladině významnosti p = 0,05 statistickou významnost rozdílů mezi průměry a rozptyly.

Doporučení: Použijte nejprve F-test, nulové hypotézy: průměry (rozptyly) jsou stejné, resp. není mezi nimi statisticky významný rozdíl.
Řešení: hodnota F-kritéria 9,78 : 8,83 = 1,15; kritická hodnota F-rozdělení pro 30 stupňů volnosti je 2,07. Platí, že 1,15 < 2,07, tj. nemůžeme zamítnout nulovou hypotézu. Závěr: mezi rozdíly v rozptylech není statisticky významný rozdíl.
Obdobně t-testem ověříme, že nejsou statisticky významné rozdíly mezi průměry. Závěrečná interpretace: oba výběrové soubory mohou pocházet z jednoho základního souboru.

Úkol / Úkol k zamyšlení

Ověřte vztah mezi funkcemi „TTEST“ a „TINV“ v Excelu. Využijte nápovědy k těmto funkcím.

SHRNUTÍ

Testování statistických hypotéz se řídí přesnými pravidly, celý algoritmus je logický a obecný pro většinu testů. Po počátečním studiu problému a formulování nulové hypotézy volíme vhodné testové kritérium a hladinu významnosti. Přitom každé testové kritérium má své popsané rozdělení s kritickými mezemi uvedenými ve statistických tabulkách. Hodnotu kritéria vypočítáme a porovnáme s kritickou tabulkovou hodnotou, což nám umožní vynést verdikt o testované hypotéze.

Provádět celý algoritmus testování nemusíme ručně, stejných výsledků dosáhneme i s využitím statistických softwarů, včetně dostupného Excelu.

Kontrolní otázky a úkoly

1. V čem spočívá test shody?
2. Popiš vlastními slovy princip testování statistických hypotéz.
3. Pokud nemáš k dispozici tabulky, pokus se vygenerovat kritické hodnoty F-rozdělení nebo t-rozdělení prostřednictvím Excelu (funkce „TTEST“ nebo „TDIST“, „FTEST“ nebo „FDIST“).

Pojmy k zapamatování

Pojem 1: nulová hypotéza, testové kritérium

Pojem 2: test shody

Pojem 3: hladina významnosti

7 Závislosti mezi náhodnými veličinami

Cíl
Po prostudování této kapitoly budete umět:

- změřit těsnost korelační závislosti mezi dvěma jevy,
- posoudit statistickou významnost korelační závislosti,
- vysvětlit závislost dvou proměnných matematickým modelem.

Doba potřebná k prostudování kapitoly: **120 minut**.

Průvodce studiem
Cílem této kapitoly je analyzovat a charakterizovat vztah dvou jevů (resp. dvou náhodných veličin), tento vztah (případně závislost) změřit, a pokud existuje, tak ho vyjádřit matematicky (nejlépe pomocí funkce).
Až do této kapitoly jsme se věnovali jednomu statistickému souboru, který jsme zkoumali pomocí jeho charakteristik, nebo jsme pomocí těchto charakteristik porovnávali statistické soubory mezi sebou. Pokaždé se ale jednalo o tzv. jednorozměrné soubory (tzn., sledovali jsme pouze jeden jev). Nyní se ale dostáváme do situace, kdy budeme zkoumat, jak souvisí změna statistického znaku jednoho výběru se změnou statistického znaku druhého výběru, nebo zdali změna jednoho není podmíněna změnou druhého. Budeme také studovat, jestli na sobě závisí znaky ve vícerozměrném souboru.
Touto problematikou se zabývají dva dílčí obory statistiky, a to korelační a regresní analýza (v některé literatuře najdeme označení korelační a regresní počet).
Korelace si klade za cíl vyjádřit vzájemný vztah mezi dvěma procesy nebo veličinami. Pokud se jedna z nich mění, mění se i druhá a naopak. Pokud se mezi dvěma procesy ukáže korelace, je pravděpodobné, že na sobě závisejí, nelze z toho však ještě usoudit, že by se podmiňovaly, že by jeden z nich byl příčinou a druhý následkem. To samotná korelace nedovoluje rozhodnout. K tomu nelze použít pouze matematický aparát, ale musíme tuto závislost (stejně tak jako určení nezávislé a závislé veličiny) logicky zdůvodnit.
Zatímco pod pojmem regresní analýza rozumíme statistické metody, jež slouží k odhadování hodnot tzv. závislé veličiny (někdy též tzv. vysvětlované proměnné) na základě znalosti veličiny nezávislé (resp. vysvětlující proměnné).
Zjednodušeně řečeno: korelace slouží k analyzování těsnosti (síly) vztahu dvou náhodných veličin (ale ne k předpovědi), zatímco regrese hledá způsob této závislosti a umožňuje předpovědi.

7.1 Korelační počet

Úkolem korelačního počtu je změřit těsnost vztahu mezi dvěma proměnnými, nebo těsnost změny hodnoty znaku závisle proměnné při změně hodnoty znaku nezávisle proměnné. Stanovení této těsnosti (těsnosti korelační závislosti) je nutným krokem, jež předchází regresní analýze a vyjádření této závislosti matematickou funkcí.

Korelační koeficient se řadí k nejdůležitějším charakteristikám hodnocení korelační závislosti. Předpokládá linearitu studovaných proměnných.

Zmíněnou těsnost závislosti dvou jevů (dvou náhodných veličin) X a Y změříme pomocí charakteristiky „koeficient korelace“ (též korelační koeficient, zpravidla označovaný r_{xy}, viz vzorec)

$$r_{xy} = r_{yx} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Tento vzorec, který je založen na tzv. kovarianci (označujeme s_{xy}, viz vzorec níže), což je obdoba rozptylu

$$s_{yx} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}),$$

lze zjednodušit na následující tvar:

$$r_{xy} = r_{yx} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \cdot \sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}},$$

který závisí přímo na jednotlivých hodnotách proměnných X a Y. Použití korelačního koeficientu předpokládá normální rozdělení obou výběrů (pokud tomu tak není, je třeba oba výběry na toto rozdělení převést), další podmínkou je linearita vztahu x_i a y_i, tzn., že regresní funkcí musí být přímka. Výše zmiňovaný koeficient se nazývá v odborné literatuře často též „Pearsonův korelační koeficient“. V praxi se též můžeme setkat ještě s tzv. „Spearmanovým koeficientem“, který nebere v potaz jednotlivé hodnoty sledovaných jevů, ale jejich pořadí.

Důležitým prvkem korelační a regresní analýzy, který nám může okamžitě napovědět o vztahu mezi dvěma veličinami je tzv. „korelační pole (diagram)“, což je bodový graf zobrazující obě náhodné veličiny X a Y.

Vlastnosti korelačního koeficientu:

- hodnoty se pohybují v intervalu <–1; 1>,
- v případě, že $r_{xy} = 1$, hovoříme o tzv. přímé korelační závislosti, kdy přírůstek nezávisle proměnné znamená přírůstek závisle proměnné,
- v případě, že $r_{xy} = –1$, hovoříme o tzv. nepřímé korelační závislosti, kdy přírůstek nezávisle proměnné znamená úbytek závisle proměnné,
- hodnotu $(r_{xy})^2$ nazýváme koeficientem determinace, jeho hodnoty se pohybují v intervalu <-0; 1> a jde o doplňkový údaj ke korelačnímu koeficientu,
- statistická závislost (resp. její významnost) se posuzuje pomocí t-testu, testujeme korelační koeficient, testové kritérium t je dáno vztahem (t-rozdělení s n–2 stupni volnosti)

$$t = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \cdot \sqrt{n-2},$$

- o statistické významnosti vypočtené hodnoty korelačního koeficientu se můžeme dozvědět i ze statistických tabulek.

7.2 Regresní analýza

V této podkapitole budeme řešit takovou statistickou úlohou, jejíž náplní bude hledání a zkoumání závislostí proměnných, jejichž hodnoty jsme získali při realizaci šetření, experimentů, nebo uvažujeme soubory statistických dat, přičemž tyto proměnné (jevy, veličiny) považujeme za náhodné. Dvojice náhodných proměnných (závislých, jejichž závislost jsme ověřili korelační analýzou) je reprezentována nezávisle proměnnou X (x_1, ..., x_n) a závisle proměnnou Y (y_1, ..., y_n).

Jak již bylo uvedeno v teoretickém úvodu ke kapitole 7, k popisu a vyšetřování závislosti Y na X užíváme regresní analýzu, přičemž tuto závislost vyjadřujeme regresní funkcí.

Cílem regresní analýzy je nalézt tvar (předpis) regresní funkce. Obvykle jej volíme tak, aby co nejvíce odpovídal vyšetřované nebo uvažované závislosti. Bývá zvykem volit regresní funkci s co nejmenším počtem regresních koeficientů, avšak dostatečně flexibilní a s požadovanými vlastnostmi (např. monotonie, předepsané hodnoty, asymptoty aj.). Vychází se přitom povětšinou ze zkušenosti, avšak v současné době se při realizaci regresní analýzy s využitím statistických softwarů dají často úspěšně použít vhodné databáze regresních funkcí.

Cílem regresní analýzy je nalézt tvar (předpis) regresní funkce, pomocí které budeme schopni predikovat hodnotu závislé proměnné pro jakékoliv hodnoty nezávisle proměnné.

Regresní funkce rozdělujeme na lineární a nelineární (vzhledem k regresním koeficientům). Některé nelineární regresní funkce (např. kvadratickou, logaritmickou, exponenciální regresi) můžeme vhodnou transformací převést na lineární (např. mocninnou nebo exponenciální funkci logaritmujeme). Jde sice o běžně používaný postup, kdy však nakonec řešíme jiný regresní model nežli původně uvažovaný.

Lineární regrese

Lineární regrese je nejjednodušší případ regresní funkce, kdy regresní čárou je přímka. Tato přímka je dána vztahem $y = a + bx$, což je „analytický výraz“, který vyjadřuje výskyt hodnot y (závisle proměnná), očekávaných s největší pravděpodobností a podmíněných změnami x (nezávisle proměnná).

Průběh regresní přímky vyjádření koeficientů a, b je výsledkem metody nejmenších čtverců (jedná se o nejčastěji uváděný způsob určení regresní čáry). Metoda spočívá v podmínce, aby se hledaná přímka co nejvíce přimykala bodům korelačního pole tak, že součet druhých mocnin (čtverců) vzdáleností bodů pole od přímky musí být minimální.

K výpočtům koeficientů a, b regresní přímky se používá celá řada softwarových programů, pro „ruční“ výpočet můžeme použít vzorce

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^{\,2} - n\bar{x}_2} \qquad \text{a} \qquad a = \bar{y} - b\bar{x}\,.$$

Pro zájemce

Máme k dispozici čtyři dvojice náhodných veličin – viz následující tabulka:

Tab. 6 Dvojice náhodných proměnných.

x1	y1	x2	y2	x3	y3	x4	y4
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Pramen: Hendl 2009.

Ke každé proměnné vypočítejte její aritmetický průměr, směrodatnou odchylku a u každého páru proměnných ověřte korelačním koeficientem těsnost závislosti (vše spočítejte s přesností na dvě desetinná místa). K čemu jste dospěli? Následně sestrojte pro každou dvojici proměnných korelační diagram a vše okomentujte.

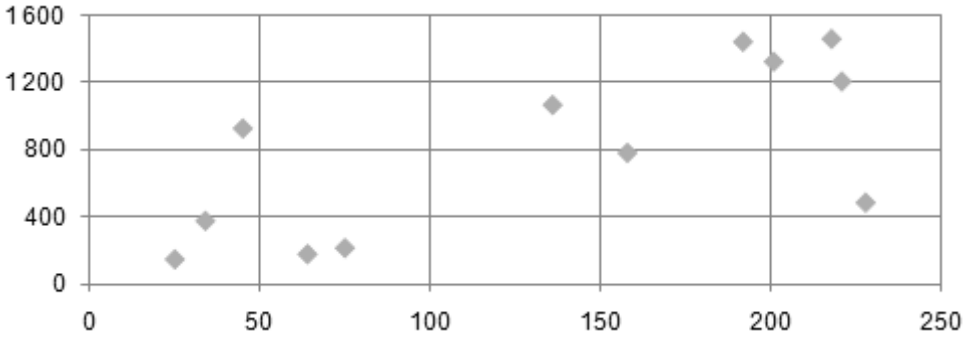
Příklad / Příklad z praxe

Máme k dispozici párové hodnoty x_i a y_i. Prověřte těsnost korelační závislosti mezi proměnnými X a Y, je-li statisticky významná, sestrojte pomocí nástrojů regresní analýzy matematický model.

Tab. 7 Zdrojová tabulka k příkladu.

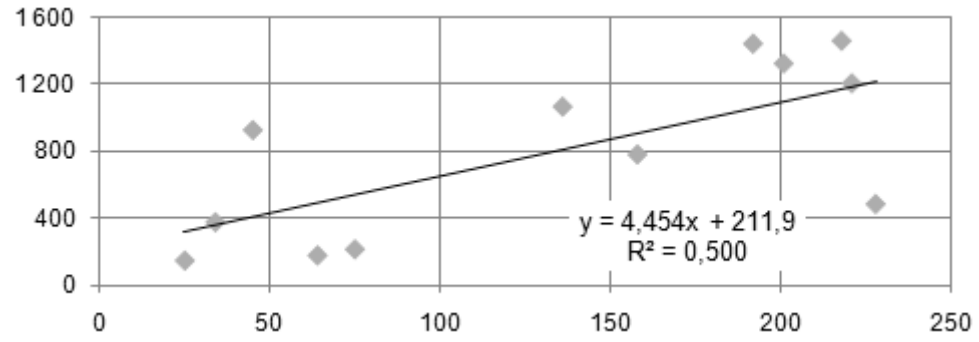
x_i	25	45	34	192	136	218	221	201	228	158	64	75
y_i	155	930	383	1443	1069	1460	1208	1325	491	785	186	222

Řešení: Nejdříve sestrojíme korelační diagram (viz obr. 21)



Obr. 22 Korelační diagram (Pramen: autor).

Z něj je zřejmé, že by korelační vztah (lineární) mohl existovat. Hodnota korelačního koeficientu r = 0,71, což je pro n = 12 párových hodnot statisticky významná hodnota. Další obrázek 22 vyjadřuje regresní lineární model.



Obr. 22 Korelační diagram s regresním modelem. (Pramen: autor).

Úkol / Úkol k zamyšlení

Vraťme se ještě k předchozímu řešenému příkladu. Regresní model slouží rovněž k predikci. Jaká bude nejpravděpodobnější hodnota závisle proměnné veličiny pro x = 250?

SHRNUTÍ

Korelační a regresní analýza představuje jednu ze zásadních kapitol statistiky. Umožňuje nám nejen posoudit vztahy mezi geografickými jevy, náhodnými veličinami, ale v případě, že se nám podaří sestrojit vhodný regresní model, umožňuje nám i předpovídat hodnoty závisle proměnné podle toho, jak se změní proměnná nezávislá.

Těsnost korelační závislosti měříme korelačním koeficientem (nejčastěji Pearsonovým). Ten lze spočítat i v excelovském rozhraní pomocí funkce „CORREL“. Ani vysoká hodnota korelačního koeficientu nemusí znamenat kauzalitu mezi proměnnými, tu musíme nějak logicky zdůvodnit. Pokud se nám to podaří, dostává se na řadu modelování pomocí nástrojů regresní analýzy. Korelační koeficient se využívá v případě linearity mezi proměnnými, u regresí nelineárních (logaritmické, exponenciální apod.) se využívá tzv. koeficient determinace, resp. koeficient spolehlivosti (R^2), který nám říká, jak úspěšný je námi sestrojený regresní model, resp. jaký podíl rozptylu původních dat nám objasňuje.

Kontrolní otázky a úkoly

1. Jaký je rozdíl mezi korelační a regresní analýzou?
2. V jakém intervalu se pohybují hodnoty korelačního koeficientu?
3. Vytipujte příklady dvojic nezávisle a závisle proměnné z oblasti fyzické i ekonomické geografie.

Pojmy k zapamatování

Pojem 1: korelační koeficient, korelační diagram

Pojem 2: koeficient determinace, index spolehlivosti

Pojem 3: přímá a nepřímá závislost, lineární regresní analýza

8 Vybrané statistické metody

Cíl
Po prostudování této kapitoly budete umět:

- analyzovat a graficky prezentovat průběh časové řady,
- vyjádřit číselně a graficky koncentraci vybraného jevu v prostoru,
- sestrojit a používat trojúhelníkový graf.

Doba potřebná k prostudování kapitoly: **60 minut**.

Průvodce studiem
V této kapitole se seznámíme s některými statistickými metodami, které buď nebylo možné zařadit do žádné z předchozích kapitol, nebo je vhodnější zmínit je samostatně. Patří sem například nástroje analýzy časových řad, dále si řekneme, jak vhodně změřit koncentraci jevu v prostoru a graficky ji vyjádřit a závěrem představíme jednu efektní grafickou metodu v podobě trojúhelníkového grafu.

8.1 Časové řady

Statistická řada je posloupnost hodnot znaku, které jsou určitým způsobem uspořádány. Je-li toto uspořádání realizováno na základě časového sledu hodnot znaku, nazýváme takovou řadu *časovou řadou*. Při analýze časových řad je nutné dodržovat zásady statistického šetření – používat stejně velká časová období, stejně velká území, stejné měrné jednotky apod.

Bazický index

Bazický index je index se stálým základem, lze jej tedy spočítat podle vztahu

$$k_i' = \frac{x_i}{x_z} \quad \text{event.} \quad k_i' = \frac{x_i}{x_z} \cdot 100[\%],$$

kde hodnota x_z je první hodnotou časové řady, tzv. základ, s níž srovnáváme všechny ostatní hodnoty řady. Při výpočtu bazického indexu je tedy vždy hodnota prvního časového momentu brána jako 100 %.

Řetězový index

Řetězový index neboli koeficient růstu je indexem s pohyblivým základem. Koeficienty růstu spočítáme podle vztahu

$$k_i = \frac{x_i}{x_{i-1}} \quad \text{event.} \quad k_i = \frac{x_i}{x_{i-1}} \cdot 100[\%],$$

řetězový index tak vyjadřuje, o kolik procent vzrostla hodnota v okamžiku t_i ve srovnání s hodnotou předchozí, tj. v čase t_{i-1}. Při výpočtu řetězového indexu považujeme za základ (100 %) hodnotu předchozího časového momentu.

Příklad / Příklad z praxe

V tabulce 8 jsou uvedeny počty obyvatel okresů Přerov a Bruntál ze sčítání mezi lety 1869 a 2001. Doplňte tabulku o bazické a řetězové indexy a následně je prostřednictvím spojnicového grafu prezentujte a okomentujte.

Tab. 8 Vývoj počtu obyvatel okresů Přerov a Bruntál v letech 1869–2001

Rok	Počet obyvatel		Bazický index (%)		Řetězový index (%)	
	Přerov	Bruntál	Přerov	Bruntál	Přerov	Bruntál
1869	86 128	143 985	100,0	100,0	100,0	100,0
1880	95 695	148 047	111,1	102,8	111,1	102,8
1890	101 648	147 424	118,0	102,4	106,2	99,6
1900	108 581	141 337	126,1	98,2	106,8	95,9
1910	119 383	140 940				
1921	120 794	133 195				
1930	127 479	140 874				
1950	117 963	82 837				
1961	127 683	90 283				
1970	133 823	91 894				
1980	139 516	99 836				
1991	138 379	108 965				
2001	135 886	105 139				

Pramen: ČSÚ.

Průměrné tempo růstu

Průměrné tempo růstu časové řady se spočítá jako geometrický průměr:

$$\overline{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}$$

Příklad / Příklad z praxe

Níže uvedená data vyjadřují roční tempo růstu (v %) rozvojové země. Vypočítej průměrné roční tempo růstu za celé období.

2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
3,5	4,7	7,6	5,8	12,5	16,7	15,3	5,8	10,6	10,8

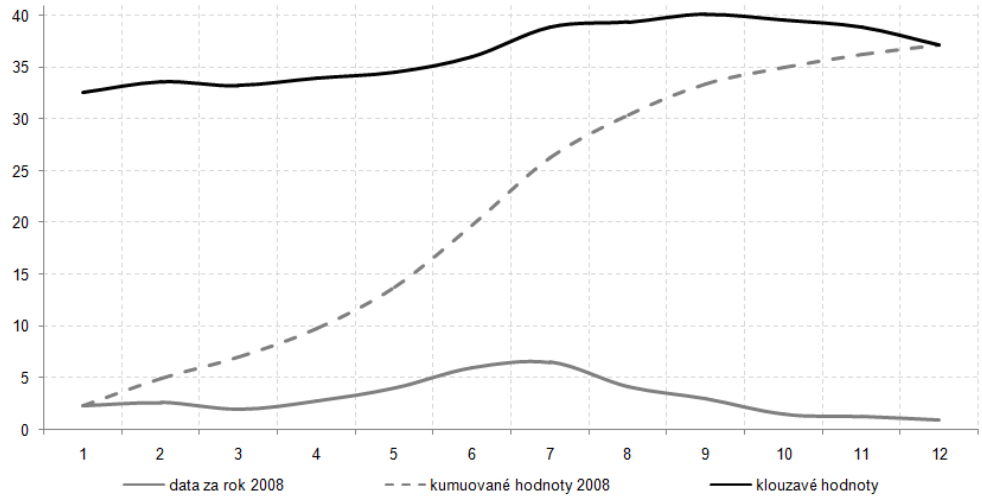
Doporučení: Do vzorce zadejte hodnoty ve tvaru 1+3,5/100=1,035 pro rok 2001.

Metoda klouzavých úhrnů a Z-diagram

Klouzavé úhrny jsou vhodnou metodou pro porovnávání hodnot v odpovídajících si časových intervalech, tj. řečeno v obecné rovině – porovnáváme úroveň statistické řady s úrovní statistické řady v předešlém období. Rostou-li hodnoty klouzavých úhrnů, znamená to, že velikost ukazatelů ve druhém období je vyšší než v prvním. Řadu klouzavých úhrnů sestrojíme tak, že tvoříme vždy součty hodnot sledovaného jevu za posledních 12 měsíců (pokud tedy porovnáváme dvě roční řady s údaji za jednotlivé měsíce) a tyto součty posouváme vždy

o jeden měsíc. Vyjdeme ze součtu měsíčních hodnot za první rok, od něj odečteme lednovou hodnotu z prvního roku a přičteme lednovou hodnotu roku druhého. Tak dostaneme první klouzavý úhrn, další vypočítáme analogickým postupem (tzn., že odečteme a přičteme příslušné únorové hodnoty, pak březnové atd.). Poslední klouzavý úhrn je roven součtu všech měsíčních hodnot ve druhém sledovaném roce.

Tato metoda má své uplatnění ve fyzické geografii (např. při prezentaci srovnání srážkových úhrnů ve dvou časových obdobích), ale aplikovat ji lze i v ekonomické geografii, např. při hodnocení intenzity bytové výstavby apod. Nejpoužívanějším grafickým znázorněním klouzavých úhrnů je speciální spojnicový graf – tzv. Z-diagram (viz obr. 23).

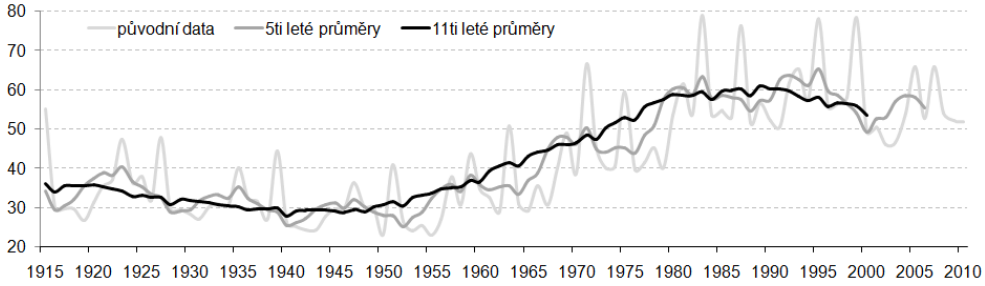


Obr. 23 Ukázka Z-diagramu, intenzita bytové výstavby, smyšlená data. (Pramen: autor).

Z-diagram zobrazuje klouzavé úhrny, kumulované četnosti a hodnoty časové řady, kterou analyzujeme. Pro jeho sestrojení musíme tedy umět spočítat klouzavé úhrny a kumulované četnosti. Všechny tři řady zobrazíme do spojnicového grafu (každou datovou řadu zvlášť), kde osa x nese jednotlivé měsíce, osa y pak sledovaný jev.

Metoda klouzavých průměrů

Jde o metodu sloužící ke shlazování dlouhodobých časových řad. Původní data mohou být značně rozkolísaná a pak je velmi obtížné nalézt v časové řadě trend. Proto se používá metody klouzavých průměrů, a to n-letých, kde n je liché číslo (typické jsou průměry 5tileté, 7mileté, ale i 11tileté). Shlazení dat spočívá v tom, že hodnotu časové řady nahradíme průměrem okolních hodnot, v případě 5tiletých průměrů tedy každou hodnotu nahradíme průměrem vypočítaným z dané hodnoty, dvou předešlých a dvou následujících. V takto shlazené řadě už je analýza trendu podstatně snadnější, viz obr. 24.



Obr. 24 Proces shlazování časové řady, smyšlená data. (Pramen: autor).

8.2 Koncentrace jevu v prostoru

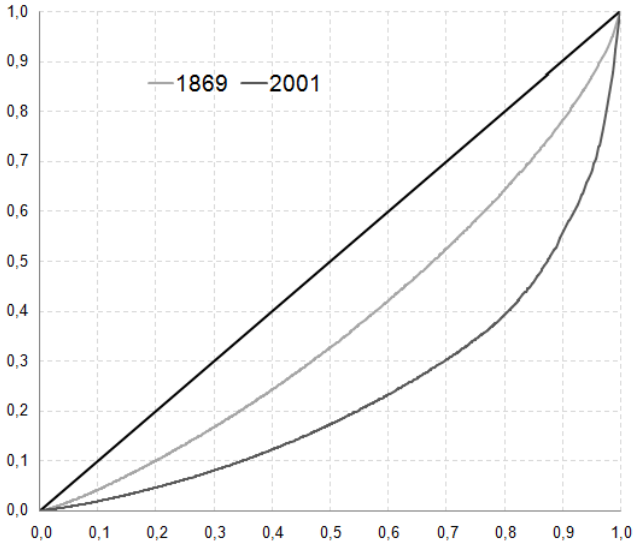
Lorenzův oblouk

Jak hodnotíme koncentraci jevu v prostoru?

Nejčastěji používaným grafickým vyjádřením koncentrace jevu v prostoru (např. koncentrace bohatství ve společnosti, koncentrace průmyslové, zemědělské výroby nebo obyvatelstva v území) je Lorenzův oblouk (Lorenzova křivka; pojmenována podle amerického ekonoma Maxe Otto Lorenze).

Vlastní koncentraci analyzujeme na základě toho, jak je křivka vzdálena od diagonály v grafu. Čím více se křivka přibližuje diagonále, tím víc je jev prostoru rovnoměrněji rozmístěn (samotná diagonála vlastně představuje naprosto rovnoměrné rozmístění). Čím víc se od diagonály vzdalujeme, tím je jev v prostoru koncentrovanější (v určitých oblastech).

Na obr. 25 je znázorněna změna územní koncentrace obyvatelstva v kraji Vysočina mezi lety 1869 a 2001. Zatímco v roce 1869 bylo obyvatelstvo na ploše kraje rozmístěno ještě poměrně rovnoměrně (křivka blízko diagonály), v roce 2001 je obyvatelstvo podstatně koncentrovanější (koncentrovanější polovina populace žila na 20 % rozlohy kraje).



Obr. 25 Koncentrace obyvatelstva v kraji Vysočina v letech 1869 a 2001.

(Pramen: autor na základě dat ČSÚ).

Popis konstrukce Lorenzovy křivky (vždy musíme mít k dispozici data o analyzovaném jevu – počet obyvatel, objem průmyslové výroby apod. a jejich rozlohu – za územní jednotky – obce, okresy apod.):

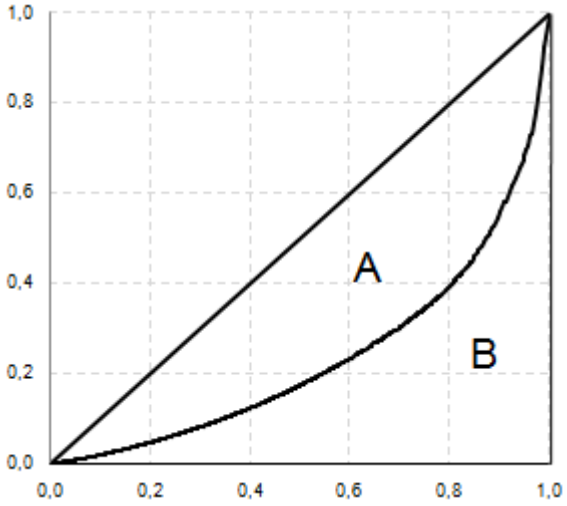
- určení podílu (v případě koncentrace obyvatelstva v území je to hustota v ob./km^2),
- seřazení dat podle daného poměru od největšího po nejmenší,
- výpočet relativních a kumulovaných hodnot pro dané prostorové jednotky,
- vynesení kumulovaných hodnot do bodového grafu.

Giniho koeficient

V případě Giniho (výslovnost „džiniho“) koeficientu jde o vyjádření téhož, jako v případě Lorenzovy křivky, jen nikoliv metodou grafickou, ale číselnou. Je tedy číselnou charakteristikou diverzifikace a má uplatnění v ekonomii, sociologii, kde se jím poměřuje například rozložení bohatství v jednotlivých územních celcích, nejčastěji státech.

Označíme-li obsah plochy mezi diagonálou a Lorenzovým obloukem jako A, plochu pod Lorenzovým obloukem jako B (viz obr. 26), pak je Giniho koeficient dán vztahem

$$G = \frac{A}{A + B}.$$



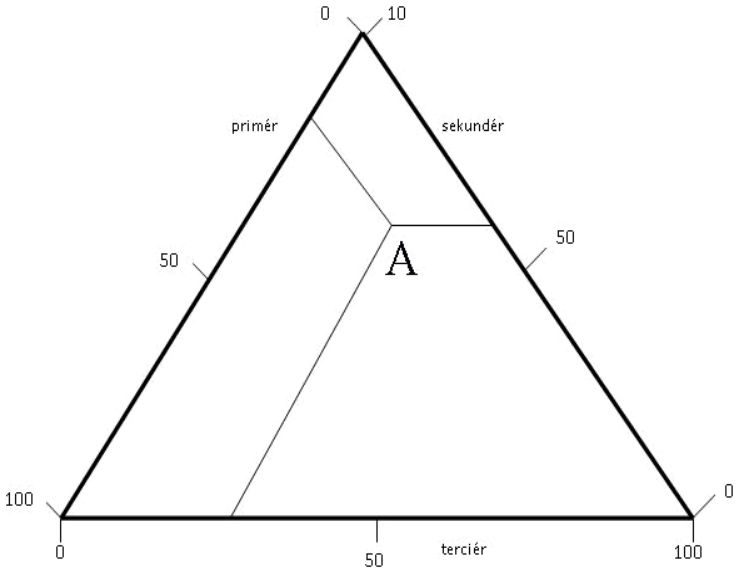
Obr. 26 Giniho koeficient (Pramen: autor).

Vrátíme-li se k obrázku 25 – tedy ke koncentraci obyvatelstva na ploše – je hodnota Giniho koeficientu pro rok 1869 G = 0,255 a pro rok 2001 G = 0,532.

Metoda je pojmenována podle italského statistika, demografa a sociologa Corrada Giniho, který se ve svých pracích věnoval měření nerovnoměrností ve společnosti.

8.3 Trojúhelníkový graf (Ossanův trojúhleník)

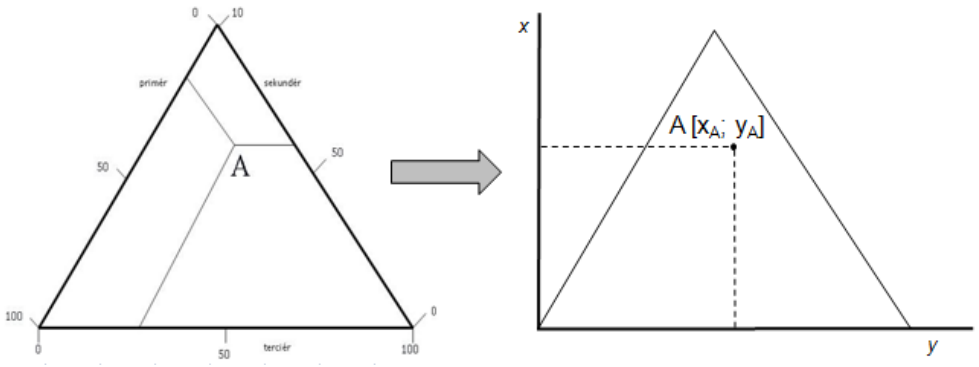
Poslední metodou, o které si řekneme je metoda grafická, která slouží k prezentaci jednotek, u kterých sledujeme jev mající tři souřadnice, jejichž součtem dostáváme 1 (nebo 100 %). Například sledujeme v územních jednotkách zaměstnanost v sektorech hospodářství, zmíněné tři souřadnice představují zaměstnanost v primérním sektoru, sekundárním a terciérním (v součtu 100 % zaměstnaných). Takováto data lze jednoduše graficky prezentovat, konkrétně prostřednictvím tzv. trojúhelníkového grafu – viz obr. 27. Každá jednotka je zobrazena prostřednictvím jednoho bodu v rovnostranném trojúhelníku, jehož strany jsou nositelkami stupnic. Zobrazený bod A [20; 55; 25] tak představuje územní jednotku se zaměstnaností I. – 20 %; II. – 55 %; III. – 25 %.



Obr. 27 Trojúhelníkový graf. (Pramen: autor).

Abychom takovýto graf nemuseli sestrojovat ručně, lze ho sestrojit jako bodový graf v Excelu, je ale nezbytná transformace tří souřadnic do pravoúhlé sítě XY. Pokud osy x a y proložíme trojúhelníkem tak, jak je uvedeno na obrázku 28, můžeme vyjádřit souřadnice bodu A pomocí následujících transformačních rovnic, které vycházejí z prosté Pythagorovy věty ($Z_{I.}$ a $Z_{II.}$ představují zaměstnanost v prvním a druhém sektoru, obecně 1. a 2. souřadnici bodu z trojúhelníkového grafu):

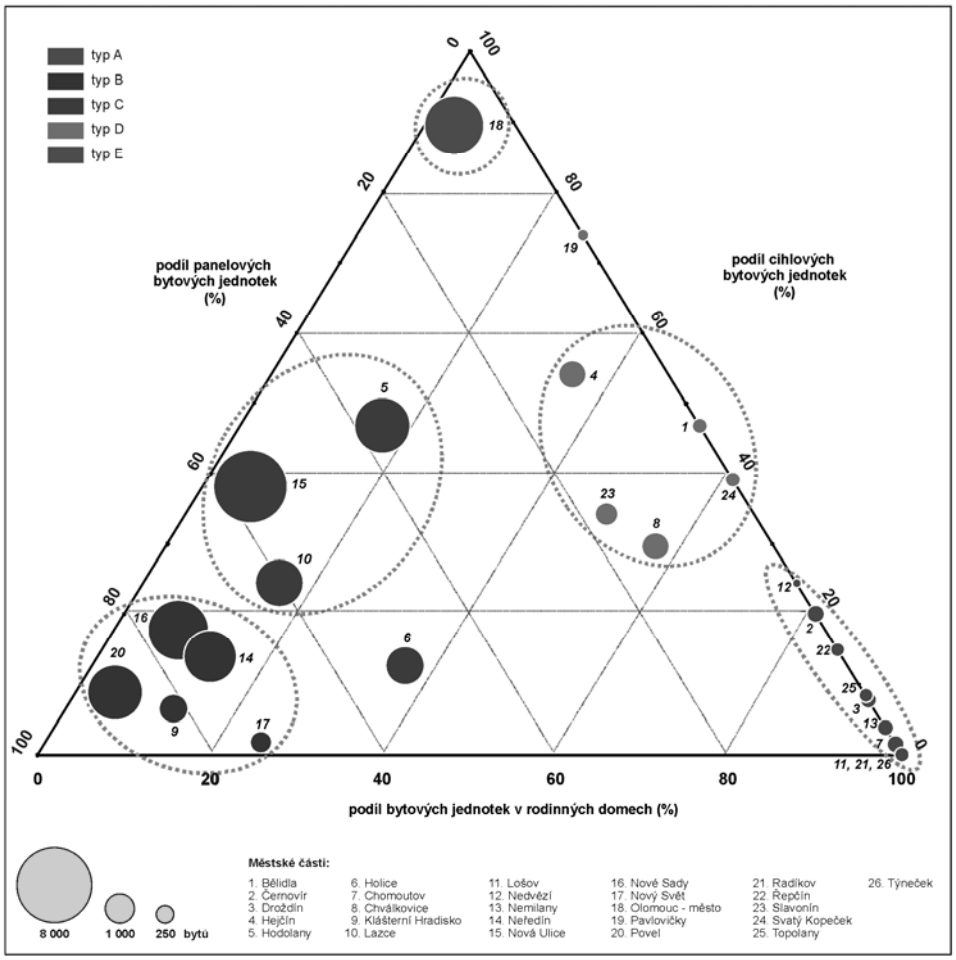
$$x_A = 100 - \frac{Z_{II.}}{2} - Z_{I.}, \qquad y_A = \frac{\sqrt{3}}{2} \cdot Z_{II.}$$



Obr. 28 Transformace souřadnic trojúhelníkového grafu do pravoúhlé soustavy souřadnic. (Pramen: autor).

Máme-li souřadnice transformovány do pravoúhlé sítě, nic nám nebrání k vynesení studovaných jednotek do bodového grafu. Snadno dokreslíme strany trojúhelníka jako spojnice vrcholů, popř. další význačné úsečky, např. těžnice.

O metodě trojúhelníkového grafu se zmiňujeme, protože se jedná o dobrý nástroj k provádění jednoduchých klasifikací nebo typologií. To je ukázáno na obr. 29, který zobrazuje městské části Olomouce z pohledu struktury bytového fondu. Podle toho, jak jsou body v grafu uskupeny, můžeme identifikovat jednotlivé typy městských částí.



Obr. 29 Ukázka trojúhelníkového grafu. (Pramen: autor na základě dat ČSÚ)

Pro zájemce

Pokuste se nalézt další vhodné uplatnění trojúhelníkového grafu.

SHRNUTÍ

V poslední, osmé, kapitole jsme si uvedli jednoduché metody sloužící k analýze časových řad – bazické, řetězové indexy, průměrné tempo růstu, metodu klouzavých průměrů apod.

Kontrolní otázky a úkoly

1. Vysvětli rozdíl mezi bazickým a řetězovým indexem.
2. Jak lze vyjádřit koncentraci jevu v prostoru?
3. Popiš konstrukci Lorenzova oblouku.

Pojmy k zapamatování

Pojem 1: tempo růstu, bazický, řetězový index

Pojem 2: Lorenzův oblouk

Pojem 3: Giniho koeficient

Závěr

Cílem publikace bylo představit vybrané základní statistické metody a jejich aplikaci v geografických úlohách. Postupně jsme přešli od metod popisné statistiky, které sloužily k prvotní analýze, popisu a komparaci statistických souborů k pravděpodobnostní statistice, kde jsme si kladli za cíl zobecnit zjištěné výsledky zkoumání výběrových souborů a přejít k souborů základním. Seznámili jsme se s vybranými teoretickými rozděleními a jejich vlastnostmi, odhadovali jsme jejich parametry a vše si vysvětlili na příkladech.

Finální část jsme zaměřili na analýzu závislostí náhodných veličin, probrali jsme vstup do jinak složitých konstrukcí korelační a regresní analýzy. Učební text by měl představovat pouze první vstup do problematiky statistiky v geografii, předpokládá se, že čtenář si rozšíří spektrum zde uvedených metod i z dalších zdrojů věnujících se podobné problematice. Pokud učební text čtenáři pomohl, nebo ho dokonce zaujal, pak splnil svůj účel.

Použité zdroje

- Barber, G. M. (1996): *Elementary Statistics for Geographers*. New York: Guilford.
- Brázdil R. a kol. (1995): *Statistické metody v geografii – cvičení*. Brno: Masarykova univerzita.
- Flowerdew, R., Martin, D. (2005): *Methods in Human Geography: A guide for students doing a research project*. Prentice Hall.
- Fotheringham, A. S., Brunsdon, Ch., Charlton, M. (2000): *Quantitative Geography: Perspectives on Spatial Data Analysis*. Sage Publications.
- Haining, R. (1990): *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.
- Hammond, R., McCullagh, P. (2009): *Quantitative Techniques in Geography: An Introduction*. Oxford: University Press.
- Hebák, P. a kol. (2007): *Vícerozměrné statistické metody 1, 2, 3*. Praha: Informatorium.
- Hendl, J. (2009): *Přehled statistických metod zpracování dat*. Praha: Portál.
- Kladivo, P., Toušek, V., Janota, M. (2010) *Aplikace v regionální a sociální geografii* (on-line). Cit. 2013-01-20. Dostupné z http://aplikacergsg.csi.muni.cz.
- McGrew, Ch. J., Monroe, Ch. (1999): *An Introduction to Statistical Problem Solving in Geography*. McGraw-Hill Higher Education.
- Rogerson, P. A. (2006): *Statistical Methods for Geography : A Student Guide*. London: SAGE Publications.
- Silk, J. (1979): *Statistical Concepts in Geography*. Allen and Unwin.
- Shaw, G. O. P. (1985): *Statistical Techniques in Geographical Analysis*. John Wiley & Sons.

V textu a příkladech dále použita data ČSÚ volně dostupná na www.czso.cz.

Profil autora

Mgr. Petr Kladivo Ph.D.

Narodil se 3. 10. 1981 v Poličce. V letech 2000–2005 absolvoval magisterské studium na Přírodovědecké fakultě UP v Olomouci – učitelství pro střední školy s aprobací matematika a zeměpis. V letech 2006–2012 absolvoval pod vedením doc. RNDr. Václava Touška, CSc. Doktorské studium na Geografickém Ústavu Přírodovědecké fakulty MU v Brně.

Během své odborné činnosti se podílel jako řešitel nebo spoluřešitel na řadě výzkumných projektů (GAČR, GAAV, FRVŠ...). Během své akademické činnosti pedagogicky působil na MU v Brně, UJEP v Ústí nad Labem a UP v Olomouci (výuka předmětů Kvantitativní metody, Statistika pro geografy, Metody RG výzkumu, Matematika, Teorie regionů a osídlení aj.), spoluautorsky se podílel na učebním textu Aplikace v regionální a sociální geografii. Pod jeho vedením byla vedena a úspěšně obhájena řada kvalifikačních prací.

Profesní specializace - statistické, kvantitativní metody a jejich aplikace v geografii, urbánní geografie, geografie příhraničních regionů, geografické modelování.

Mgr. Petr Kladivo, Ph.D.

Základy statistiky

Výkonný redaktor prof. RNDr. Tomáš Opatrný, Dr.
Odpovědný redaktor Bc. Otakar Loutocký
Technická redakce autor
Návrh obálky Martin Jurek
Technické zpracování obálky Jiří Jurečka

Vydala a vytiskla Univerzita Palackého v Olomouci
Křížkovského 8, 771 47 Olomouc
www.vydavatelstvi.upol.cz
www.e-shop.upol.cz
vup@upol.cz

Publikace neprošla ve vydavatelství redakční jazykovou úpravou

1. vydání

Olomouc 2013

Ediční řada – Studijní opory

ISBN 978-80-244-3841-2 (tištěná verze)
ISBN 978-80-244-3842-9 (online verze)

Neprodejná publikace

Online verze publikace dostupná na
http://geography.upol.cz/soubory/studium/e-ucebnice/978-80-244-3842-9.pdf

VUP 2013/749 (tištěná verze)
VUP 2013/750 (online verze)

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

9 788024 438412