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

Univerzita Hradec Králové
Ústav sociální práce

Statistika pro sociální práci

Martin Kořínek

Gaudeamus 2014

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

Publikace neprošla jazykovou úpravou.

Edice texty k sociální práci

Řada: Výzkumné metody v sociální práci - sv. 3

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

ISBN 978-80-7435-404-5

Obsah

1 Úvodem

Tato skripta jsou určena pro studenty, kteří mají svých ve studijních plánech jednosemestrovou výuku Statistiky pro sociální práci. Skripta by měla sloužit jako základní pomůcka (odrazový můstek), přičemž se předpokládá, že studenti sami využijí dalších pramenů pro získání detailnějších informací.

Skripta poskytuji vybrané kapitoly ze statistiky, které by sociální pracovním měl ovládat. Nedílnou součástí studia je řešení statistických příkladů a to s využitím jak jednoduššího, avšak rozhodně nikoli triviálního, programu, jakým je tabulkový procesor (například MS Excel), tak i s využitím profesionálnějšího programového vybavení jakým je například statistický paket NCSS či SPSS.

Pro zdárné studium těchto skript se předpokládá, že student již zná základy popisné statistiky.

Pro zájemce o následné studium statistiky je uveden seznam literatur, jenž zahrnuje fundamentální publikace (a i další informační zdroje), týkající se statistických metod používaných v praxi.

2 Nestačí jen průměry aneb opakování popisné statistiky

2.1 Příklady na úvod

Řekne-li někdo Schola ludus, ihned se před námi v našem podvědomí zjeví obraz fousatého učitele národů Jana Amose Komenského. Ti bystřejší si ze školních lavic pamatují, že tato dvě latinská slovíčka znamenají ono pověstné ŠKOLA HROU. Chceme-li totiž cokoli někoho naučit (či předložit něco nového nebo nepříliš známého), nejrychleji se nám to podaří, použijeme-li vhodných příkladů, přirovnání, jinak řečeno, vtáhneme-li plně čitatele do problému. A jelikož toto je i cílem těchto skript ze statistiky, připravili jsme si pro vás hned dva příklady.

Příklad 1 (SWOBODA, 1977). Každý snad zná tento vtip (který proto jen připomínáme). Setkají se dva statistikové a jeden si postěžuje, že mu žena doma neupeče kuře jak je rok dlouhý. Druhý se podiví a opáčí, vždyť jsi měl kuře v neděli. První se udiveně kouká (myslí si bůhví co o příteli), nedá mu to a podotkne něco o nepravdivém tvrzení. A druhý mu tedy to kuře hravě vypočítá: já měl celé kuře a ty prý žádné, takže 1 kuře a žádné kuře je jedno kuře - na něj jsme byli dva a tak jsme měli v neděli oba v průměru po půlce kuřete, tož si nestěžuj.

Příklad 2 (SWOBODA, 1977). V jednom městečku žije 1 010 obyvatel. Většina z nich (přesně 1 000), pracuje v blízké továrničce s platem 10 penízků měsíčně. Zbylých 10 obyvatel pracuje tamtéž, ovšem jako top manažeři s platem 10 000. Jaký je asi průměrný plat v tomto městečku?

Kupodivu uvedené příklady mají co dočinění se statistikou. Jak a proč, o tom si nyní povíme. Prvnímu příkladu jsme se zasmáli, což je v pořádku, ale dovedete jej vysvětlit a říci, kde je chyba? Pokusme se o to.

2.2 Co je a co není statistika

Především statistika se zabývá zpracováním a analýzou hromadných dat (HENDL, 2006), tedy nikoliv individuálních, o které se jedná v příkladu. Je to podobné, jako bychom z prospěchu dvou žáků ve třídě určily průměrný prospěch žáků na všech školách v celé české republice. Dva žáci nejsou nositeli hromadných dat! Shodných příkladů bychom mohli najít velké množství a dokonce i mezi těmi, které se považují za "skutečně správné".

První uvedený příklad můžeme vyvrátit i jiným způsobem. Existuje totiž charakteristika (míra, ukazatel), která nám určí, zda v daném případě je průměr vhodný či nikoli (CYHELSKÝ, 1981), popřípadě jak dosti přesně vystihuje realitu. "Užitečnost" či chcete-li "objektivnost" průměrů lze tedy měřit a to velice prostě. V uvažovaném souboru dat vypočteme průměr a směrodatnou odchylku. Prostým podílem z nich získáme relativní míru variability (měnlivost) zvanou variační koeficient. Podrobnější postup a význam výše načrtnutých pojmů si rozebereme o něco později.

Pokusme se nyní na chvilku pozastavit u statistiky jako vědecké disciplíny a trochu ji přiblížit. Domníváme se, že je v našem životě vesměs podceňována a je odklizena do ústraní. Podle našeho názoru je tato situace způsobena především díky mizivé a nepřesné (či dokonce mylné) informovanosti. Nesmíme se nechat plést vžitým neduhem, že statistika pouze sbírá více méně nepotřebná data a její metody končí u výpočtu průměrů či indexů. Dnešní statistika je vědou, která disponuje velice rozsáhlou a i poměrně složitou metodologií. Je schopna prezentované údaje zdařile analyzovat (SWOBODA, 1977).

Jiná otázka je ale věrohodnost prvotních údajů. Jsou prvním článkem statistického snažení, a proto na nich velice záleží. Jak jsou přesné prvotní podklady, tak bude věrohodná i jejich analýza. Z nepřesných, neúplných a zkreslených údajů nelze vyhodnotit přesný závěr (HENDL, 2005). Nechceme přece, aby se o statistice říkalo, že je to "přesná věda o nepřesných číslech". Můžeme však zde prozradit, že statistika má i několik testů, kterými je možno některé nehoráznosti vstupních čísel odhalit.

2.3 Statistika jako věda

Když už jsme nakousli otázku věrohodnosti statistiky, pozastavme se u jejího zařazení do systému věd. Jak jistě víme, statistiku využívá stále větší množství jiných oborů a nejen teoretických. Technické vědy ji používají již velmi dlouho, biologické se s ní přátelí také. Dokonce i vědy, kde bychom ji nehledali, ji účelně používají. Máme na mysli vědy humanitní či společenské, především sociologie a demografie. A výsledky z průzkumu veřejného mínění jsou podloženy z 30 % sociologií a ze 70 % statistikou.

Takže statistika je vlastně velkým pomocníkem podobně jako matematika. Někteří statistikové říkají, že statistika není nic jiného než vlastně určitý způsob filosofického myšlení, že dobrý statistik musí být především dobrým filosofem (KOZÁK, 1994).

Něco pravdy na tomto tvrzení je. Vraťme se k prvnímu příkladu. Víme co to je průměr a jak se počítá. Po jeho výpočtu jednoduše konstatujeme, že na každého připadá v průměru půlka kuřete. To je tedy přístup, kdy máme daná čísla a metodu a oprostíme se od všech možných vztahů. To je ale velký omyl a my se dopustíme chyby. Proč? Protože jsme zanedbali vztahy mezi čísly (ukazateli) a jejich nositeli (KOZÁK, 1994).

Nyní se trošku zamyslíme. Existuje jistý jev, který je nezávislý na našem vědomí. Abychom jej pochopili, musíme si jej pojmenovat. Tím se nám tento jev dostane do určitého systému a my jej můžeme popisovat[1]. Jestliže chceme jev kvantifikovat, musíme navíc určit jistý ukazatel, kterým budeme jev měřit. Jev a jeho pojmenování (zařazení) nám pomáhá poznat filosofie jako nejobecnější věda a dále daný vědní obor, který pojem blíže specifikuje. Teprve posléze nastupuje statistika, aby pomohla ve spojení s filosofií a konkrétním oborem určit správný ukazatel a tím i určit zařazení ukazatele do systému statistických měr.

Zdá se, že je to příliš složité? Nevadí, později sami zjistíte, že tento postup používáte bezděčně sami dost často, ani o tom nevíte. Proto je nezbytně nutné, znát vedle statistiky i obor, kde s ní pracujeme. Samozřejmě, je nutné dodat, že pouze perfektní znalost jistého oboru nám nebude stačit, chceme-li jej statisticky analyzovat. Musíme bezpečně znát i statistickou metodologii (ČERMÁK, 1968).

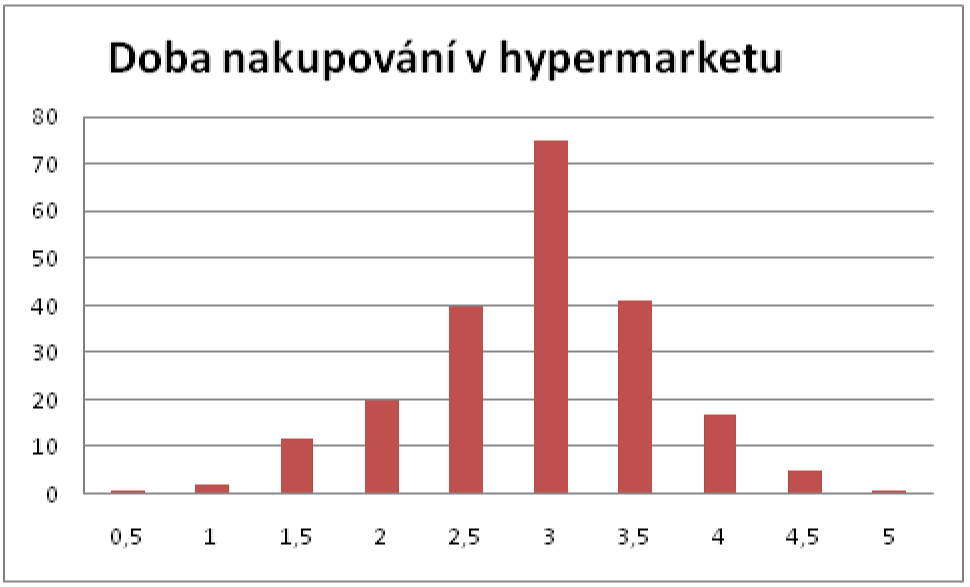
Z toho evidentně vyplívá, že dnes již existuje rozdělení statistiky dle použití. Máme tedy statistiku sociologickou, statistiku ekonomickou, lékaři používají taktéž některé své vlastní postupy, knihovnická statistika a mnoho dalších. Také ve fysice se mnoho zákonů dá odvodit statistickou metodou a s využitím pravděpodobnosti. Se statistikou se setkáváme na každém kroku (doslova při rozhodování před přecházení vozovky) a je jen na nás, jak ji budeme vědomě využívat.

A nyní se vraťme na začátek statistických metod. Pohleďme na průměry a řekneme si, co o nich známe. Následující řádky byste asi nejraději přeskočili, leč nečiňte tak, určitě se dozvíte alespoň jednu informaci, kterou jste dosud neslyšeli, nebo o které zatím nemáte ponětí. Zahleďme se na naše příklady.

[1] viz GLEICK, J.: Informace. Historie. Teorie. Záplava. Praha, Dokořán. 2013. ISBN 978-80-7363-415-5

2.4 Průměry trošku jinak

Proč počítáme průměry? Abychom jedním číslem vyjádřili určitou charakteristiku daného souboru (tedy toho, co měříme). Říkáme například, že průměrná doba nakupování v hypermarketu je *x* hodin. (Na následujícím grafu je průměrná doba nakupování právě 3 hodiny.)



Obrázek 1 - Graf rozdělení četností doby nakupování v hypermarketu

U grafu pak nám osa *x* charakterizuje dobu nakupování v hodinách a osa *y* pak počet (četnost) zákazníků, kteří nakupovali daný počet hodin. Podprůměrné a nadprůměrné doby vykazovalo méně zákazníků a průměrné (střední) doby vykazovalo zákazníků více. Pokud bychom provedli podobná šetření pro jiné typy nákupu (drogerie, automobil, pekařství atd.), zjistíme, že se grafy liší pouze ve stupnici osy *x*, tedy ve velikosti průměrné doby nakupování. Prostě jednou je průměr 3 hodiny a podruhé 30 minut (drogerie). Graf, křivka rozdělení četnosti se liší u obou případů pouze polohou, umístěním od počátku. Proto se průměrům říká "míry či charakteristiky polohy". Říkají, kde (v jakém místě) se nejpravděpodobněji setkáme s daným jevem, který analyzujeme.

Jak možná víte, průměrů máme několik. Především máme vážené a nevážené (prosté) tvary. Vážený tvar využíváme v případě, kdy máme k disposici rozdělení četností. Např. víme, kolik zákazníků má daný ukazatel roven třemi, u kolika zákazníků je roven čtyřmi, pěti atd. V rozdělení četností je soubor setříděn dle určitého znaku s uvedením počtu variant a s výskytem (četností) každé varianty.

Je již méně známé, že existuje průměr geometrický, harmonický a že dokonce si lze dle obecného vzorce vytvořit další jiné průměry (CYHELSKÝ, 1981). Ovšem vlastnosti takových průměrů jsou specielní a proto se jich využívá pouze k určitým výpočtům. Geometrického průměru se využívá k výpočtu průměrného koeficientu růstu a podobně. Platí různé převoditelné vzorce mezi průměry, ty však zde nebudeme rozebírat. Řeknete si, tolik průměrů, s nimi se dá spočítat skoro všechno. Spočítat ano, ale analyzovat pouze pomocí průměrů je naprosto nesprávné.

2.4.1 Další míry polohy

Opět poslouží oba příklady z počátku našeho povídání. Kdybychom trvali na tom, že průměrný plat v našem fiktivním městečku je 200 (přesně 198) penízků měsíčně, jistě se nám mnozí vysmějí. S průměry zde nevystačíme, a proto musíme sáhnout po jiné charakteristice. Míra polohy se totiž nevztahuje jen na průměry, ale na všechny ukazatele splňující podmínku, že musí být rovny či větší než nejmenší hodnota a rovny či menší než je největší hodnota v souboru. Nic jiného. Lze si jich vymyslet opět spoustu, my si zde uvedeme tyto: medián, modus.

Modus je nejčetnější varianta znaku (v případě městečka jasně vede 10 penízků, v případě doby nákupu v hypermarketu pak 3 hodiny). Medián je prostřední hodnota znaku, který je setříděn podle velikosti (SHAMRA, 2005). S pojmem medián souvisí i pojem kvantily, kvartily, percentily atd. Ty vyjadřují hodnotu znaku, který je v první respektive poslední čtvrtině (kvartily), či které jsou v příslušných desetinách řady (percentily, první, druhý, pátý atd.).

Jistě jste si všimli základního rozdílu mezi průměry a ostatními mírami polohy (mediánem, modem, kvantily). Průměry počítáme podle vzorců, kdežto ostatní míry zjišťujeme určeným postupem. Průměrem vypočítáme jistou hodnotu znaku, která je často fiktivní a proto v našem souboru nemusí vůbec figurovat (viz 200 penízků). Naproti tomu mediánem či modem označujeme konkrétní hodnotu znaku ze zkoumaného souboru.

Průměrem lze někdy vypočítat holý nesmysl, např. průměrný počet dětí v rodinách je 2,4. Medián a modus jasně hovoří 2. Podobně, budeme-li chtít změřit přesnější vzdálenost Praha - Hradec Králové. Pro jistotu změříme vzdálenost 10x a vypočteme průměr. A ouha. Přestože jsme měřily s přesností dejme tomu na desítky metrů, vyjde nám výsledek nejméně na milimetry. Taková přesnost je nejen zbytečná, ale dokonce troufalá a podezřelá. Použijeme-li na těchto 10 měření medián (o modu nelze hovořit, protože se z tak malého počtu měření těžko víckrát strefíme do stejného výsledku), problém se zbytečnou přesností a se zaokrouhlováním nás mine.

Přes veškeré tyto poznámky je však klasický (aritmetický) průměr velice vhodná míra polohy, bez které se nelze obejít zvláště v analýze, kde používáme složitější statistické metody. Většina metod s průměrem počítá, ale spíše jako s pomocnou hodnotou.

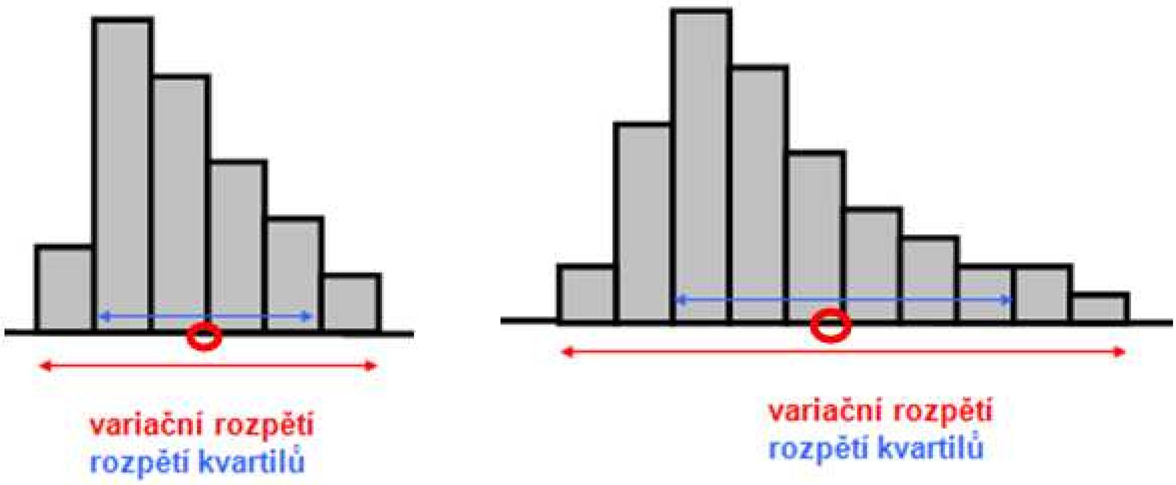
Jak tedy měřit polohu? Nejprve průměrem, který otestujeme, zda je výstižný (viz dále). V negativním případě požijeme modus - pokud je dáno rozdělení četností, nebo medián, pokud je soubor setříděn dle znaku vzestupně nebo obráceně.

2.4.2 Malé shrnutí

Pro praxi je nejdůležitější znát, že průměr máme vážený a prostý, vědět kdy který použít a ovládat pojmy medián, modus, horní a dolní kvartil. Zdatný statistik při prvním pohledu na řadu hodnot pozná, čím nejlépe ji charakterizovat. My nejsme ve statistice příliš sběhli a proto napřed vypočítáme průměr a pak jej otestujeme, zda-li je vhodný. A právě testováním průměrů se nyní budeme zabývat.

2.5 Měnlivost, variabilita, variace, diference, různorodost

Následující obrázky znázorňují grafy četností dvou souborů.



Obrázek 2 - Grafy četností u dvou souborů s různou variabilitou

První je zhuštěný a druhý je plochý. Je zřejmé, posuneme-li "kruh" (bod, značící průměr) o kousek vlevo, bude nám tento bod celkem dobře ještě charakterizovat soubor vpravo, leč pro soubor vlevo je podobný posun zcela nemyslitelný. Jinými slovy, rozpětí možných hodnot souboru vlevo je velmi malé, naopak rozpětí souboru vpravo je veliké. Jinak řečeno: hodnoty souboru vlevo se od sebe příliš neliší, naopak hodnoty souboru vpravo jsou hodně odlišné.

Tak tím jsme si vysvětlili pojmy z nadpisu tohoto odstavce. Soubor vpravo je více proměnlivý oproti souboru vlevo, který má menší variabilitu. Z obrázku je to zřejmé, ale z řady čísel, zvláště z velkého množství hodnot, vůbec ne. Nezbývá nám nic jiného, než tuto míru odlišnosti vypočítat.

Na to existuje opět vzorec. Ve skutečnosti je jich více, takže si můžeme vybrat. Jedny jsou založeny na průměrech, jiné na kvantilech atd. Je jich plno, ale nám postačí se seznámit s jednou, ale za to nejdůležitěji charakteristikou míry variability. Nazývá se rozptyl a vypočteme ji jako "průměrnou čtvercovou odchylku hodnot znaku od průměru" (RUMSEY, 2007).

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

Má jedinou malou nevýhodu - vychází v mocnině jednotek (v kvadrátu) a proto raději používáme její druhou odmocninu, která se nazývá směrodatná odchylka (v cizí literatuře - standardní odchylka) (SHARMA 2005).

$$\sigma = \sqrt{\sigma^2}$$

Je to tedy rozměrová míra s rozměrem shodným s rozměrem souboru (jednotkou souboru). Její další výhoda tkví v tom, že její hodnota nezáleží na průměru, ale pouze na vzájemných odchylkách všech hodnot znaku. Lze ji tedy počítat dvěma způsoby. Protože výpočet pomocí průměru je rychlejší a hlavně přehlednější dává se přednost tomuto způsobu výpočtu (CYHELSKÝ, 1981).

$$\sqrt{\overline{x^2} - \bar{x}^2}$$

Na okraj: dle vzorce je patrno, že se jedná o průměr čtverců, čili nezáporných čísel. Proto rozptyl a ani směrodatná odchylka nemůže mít nikdy zápornou hodnotu (ČERMÁK, 1968).

Občas nás napadne srovnávat dva či více souborů, které mají jinou jednotku. Budeme-li chtít porovnat variabilitu těchto souborů, pomocí směrodatné odchylky to nepůjde, protože ta, jak již bylo řečeno, je mírou rozměrnou. Nelze srovnávat jablka a hrušky (špendlíky a lokomotivy). Proto musíme určit míru variability bezrozměrnou veličinou. Její název zní - variační koeficient. Vypočte se pouhým poměrem směrodatné odchylky a průměru (CYHELSKÝ, 1981).

$$\boldsymbol{V} = \frac{\boldsymbol{\sigma}}{\overline{\boldsymbol{x}}}$$

A pozor, tato charakteristika je dosti důležitá. Říká nejen jak je soubor málo či hodně variabilní, ale pomůže nám odpovědět na otázku, zdali daný soubor hodnotíme průměrem správně nebo ne. Čím menší koeficient tím je soubor stejnorodější. Jestliže nám vyjde koeficient rovný 0,3 a vyšší (v procentech 30 % a více - procentní vyjádření je hojně používané) je nutno o adekvátnosti použití průměru pochybovat a raději pro míru polohy použít medián nebo modus (RUMSEY, 2007). Variační koeficient tedy skutečně měří (hodnotí) objektivnost průměru jako míry polohy.

2.6 Pravidlo šesti sigma

Je velmi praktické pravidlo, kterému, když porozumíme a budeme jej využívat, zdolali jsme první krok ke statistickému myšlení.

Mějme určitý soubor nezávislých měření. Vypočítáme průměr a směrodatnou odchylku (pro jistotu i variační koeficient). Určíme dva následující intervaly (HÁTLE, 1987):

1) k průměru přičteme a odečteme velikost odchylky a dostaneme spodní a horní mez prvního intervalu
2) od průměru přičteme a odečteme trojnásobek odchylky - tím dostaneme horní a dolní mez druhého intervalu.

Potom můžeme konstatovat, že námi měřená hodnota leží s pravděpodobností 66,6 % v prvním intervalu a s pravděpodobností téměř 100 % (přesněji 99,9 %) v intervalu druhém.

Proč pravidlo 6 sigma? Směrodatná odchylka se značí v některých případech malým řeckým písmenem sigma a my jsme pro druhý interval přidaly na obě strany 3 tyto odchylky, dohromady 6. Vysvětlení jsme zjednodušili, platnost je skoro universální, ale přesně se týká tzv. normálního rozdělení a dalších podmínek, které souvisí s teorií pravděpodobnosti a matematickou statistikou (HÁTLE, 1983).

3 Statistická analýza jednorozměrného souboru

3.1 Rozdělení četností

Základní numerický materiál statistického souboru obsahuje hodnoty sledovaného znaku (argumentu) v tom pořadí, v jakém byly průběžně zjišťovány. Takový zápis je nepřehledný a nelze z něj učinit žádný závěr. Zvlášť obtížné je to u souborů s velkými rozsahy. Z tohoto důvodu se volí jiný způsob sestavení základních dat.

Variační obor argumentu statistického souboru je interval omezený minimální a maximální hodnotou znaku souboru: x_{min} a x_{max} (MANĚNOVÁ, 2012). Tento variační obor rozkládáme na menší části (intervaly), nazývané třídy. Obsahuje-li soubor malý počet hodnot argumentu, volíme každou hodnotu tohoto argumentu za samostatnou třídu. Hodnota znaku je v tomto případě současně třídním znakem.

3.1.1 Intervalové rozdělení četností

Jestliže se u daného statistického souboru vyskytuje značně velký počet různých hodnot argumentu, potom do jedné třídy sdružujeme několik hodnot argumentu.

Pro volbu intervalů platí tyto zásady (CYHELSKÝ, 1981):

1. Intervaly se volí tak, aby bylo možno prvek zařadit do intervalu jednoznačně.
2. Šířka intervalu má být stejně velká. Doporučuje se, aby počet tříd byl 5 až 15 podle rozsahu souboru a účelu přehledu. Označíme-li R tzv. variační rozpětí,

   $$R = x_{max} - x_{min}$$

   pak pro šířku intervalu h lze použít přibližného vzorce

   $$h = 0{,}08\ R$$

   Této volbě odpovídá počet tříd $k = R / h$.
3. Hranice tříd a středy tříd mají být zaokrouhlená čísla. Střed třídy (intervalu) se vypočte podle vzorce

   $$x_j = 1/2\ (x\mathit{intervalu}_{max} - x\mathit{intervalu}_{min})$$

   Střed třídy je hodnotou třídního znaku.

3.1.2 Absolutní, relativní a kumulované četnosti

Počet prvků souboru patřící do určité třídy nazýváme absolutní četností. Výsledek roztřídění jednotek statistického souboru do tříd podle velikosti hodnot znaku se nazývá rozdělení četností (popřípadě rozložení četností). Skupinové rozdělení četností vzniká roztříděním jednotek statistického souboru do třídních intervalů. Součet všech absolutních četností se rovná rozsahu souboru (SHARMA, 2005):

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

kde k je počet tříd, n_i jsou četnost třídy i a n je rozsah souboru.

Pro názornější představu o struktuře sledovaného souboru z hlediska sledovaného znaku se používá relativní četnost. Relativní četnost se vypočítá jako poměr jednotlivých absolutních četností k rozsahu souboru (RUMSEY, 2007):

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

Relativní četnost vyjadřujeme jako čísla desetinná, nebo v procentech (tj. 100 f_j), jejich součet se rovná 1 (nebo 100 %).

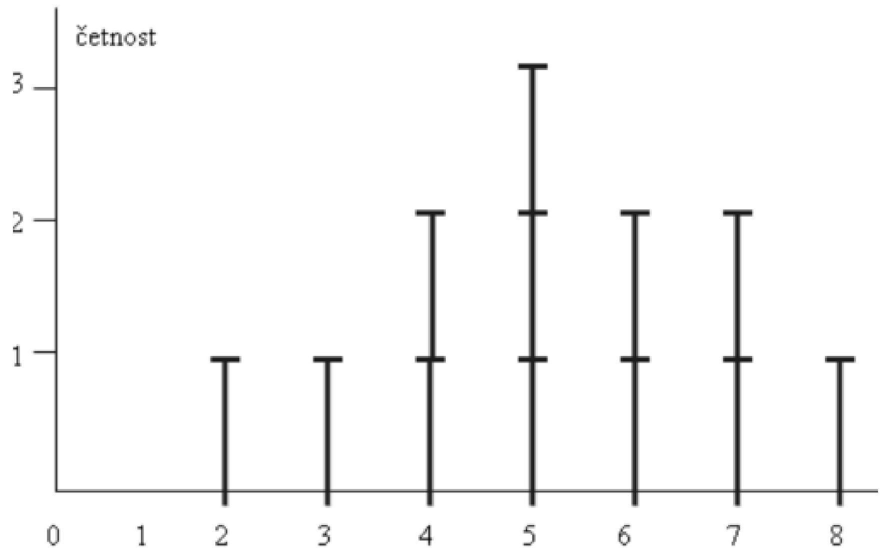
$$\sum_{i=1}^{k} f_i = 1$$

Rozdělení četností lze vhodně posoudit také podle kumulativních absolutních četností (*Fj*) a kumulativních relativních četností (*Rj*). Kumulativní četnost vyjadřuje součet všech četností od prvé do j-té třídy včetně, např. F5 = f1 + f2 + f3 + f4 + f5.

3.1.3 Grafické znázornění rozdělení četností

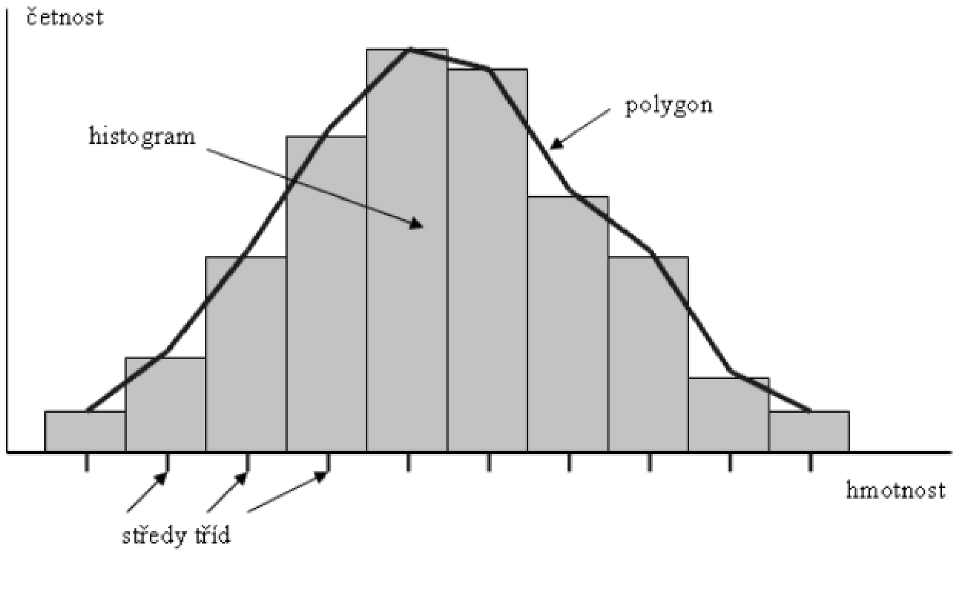
K znázornění rozdělení četností používáme nejčastěji spojnicových diagramů, histogramů četností a grafů kumulativních četností (ČERMÁK, 1968).

- Úsečkový diagram rozložení absolutních, popř. relativních četností dostaneme, jestliže na vodorovné ose zobrazíme středy jednotlivých tříd a v každém z nich sestrojíme ve směru svislé osy úsečku o délce rovné příslušné absolutní (popř. relativní) četnosti.



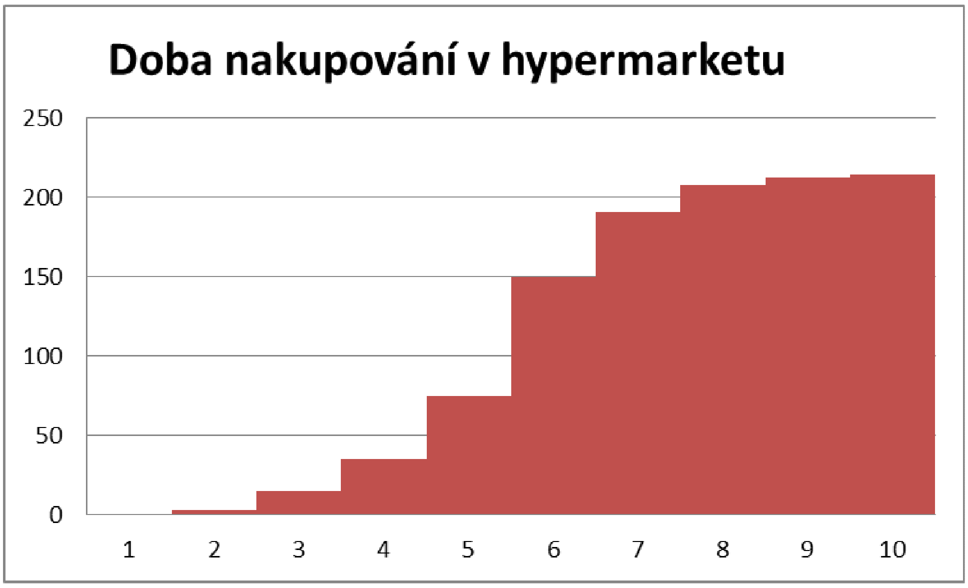
Obrázek 3 - Úsečkový graf četností

- Polygon rozložení četnosti (spojnicový diagram) dostaneme, jestliže koncové body úsečkového diagramu rozložení četností spojíme úsečkami a vytvoříme tak lomenou čáru, která pak představuje hledaný polygon.
- Histogram rozložení absolutních četností sestrojíme tak, že na vodorovné ose vyznačíme obrazy středů jednotlivých tříd a nad každou úsečkou zobrazující určitou třídu sestrojíme pravoúhelník s výškou rovnou příslušné absolutní četnosti.Horní obrys představuje histogram rozložení četností.



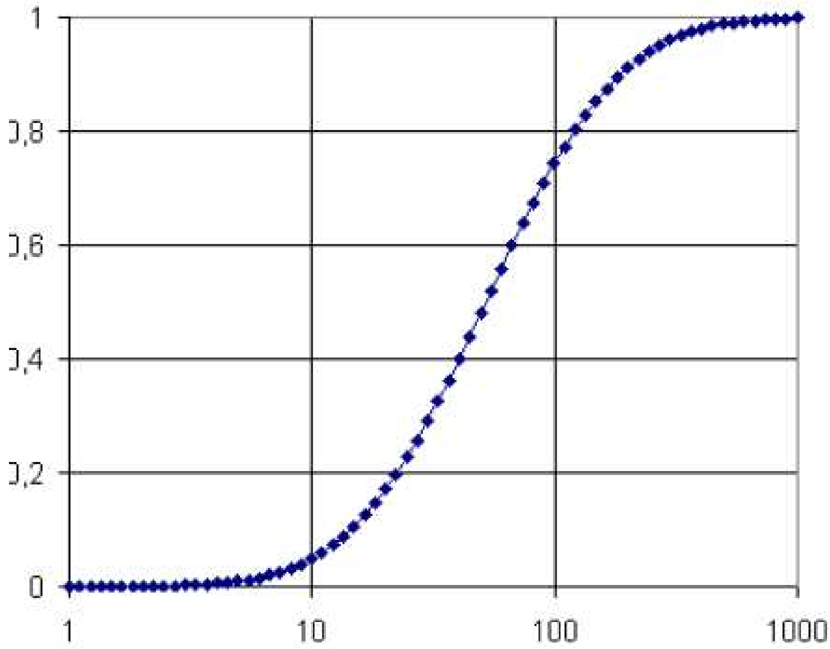
Obrázek 4 - Polygon a histogram rozdělení četností

- Graf, polygon nebo histogram kumulativních absolutních četností dostaneme analogicky jako histogram absolutních četností, přičemž každou kumulativní absolutní četnost přiřazujeme obrazu (na vodorovné ose) středu příslušné třídy.



Obrázek 5 - Histogram kumulované absolutní četnosti

- Ogivní křivku dostaneme, sestrojíme-li polygon kumulativních četností. Polygon kumulativních četností se někdy nazývá součtovou křivkou nebo S-křivkou.



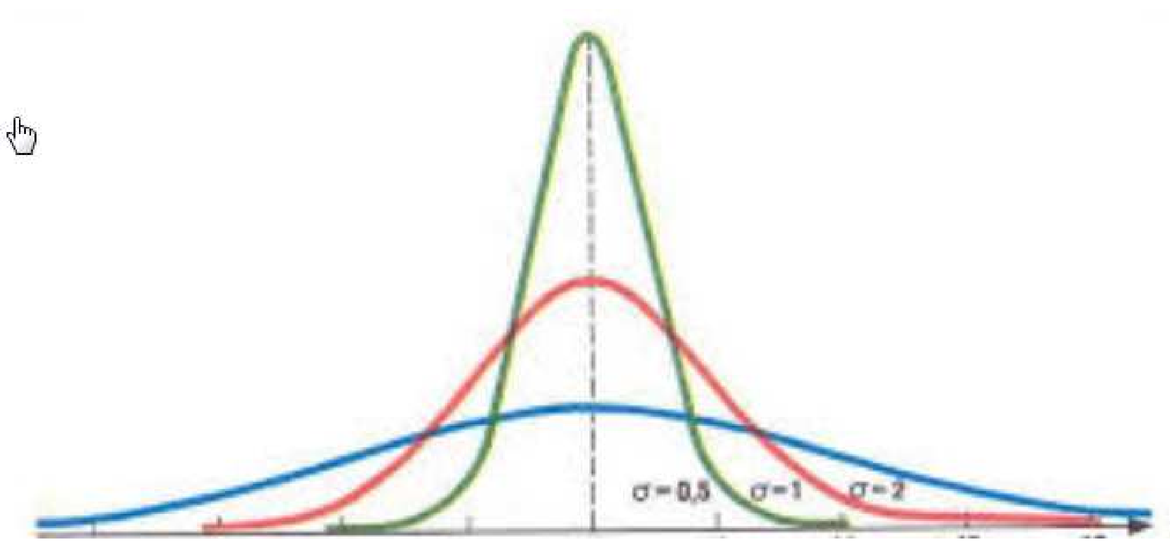
Obrázek 6 - Typická S křivka - spojnicový graf kumulované četnosti

V případě, že všechny třídy mají stejnou šířku, je vhodné připojit k svislé stupnici pro četnost také stupnici pro relativní četnost.

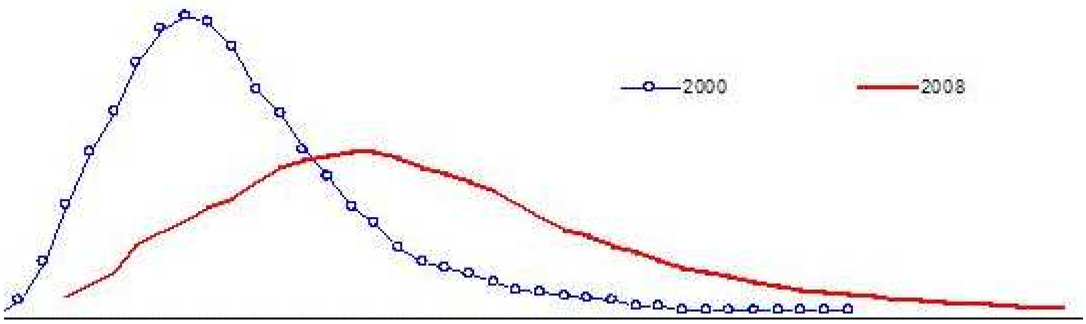
3.2 Statistické charakteristiky

Grafické znázornění rozdělení četností je sice úplné a názorné, ale pro řešení statistických úloh je potřebné shrnout informaci o rozdělení četností do vhodných číselných údajů. K tomuto účelu slouží statistické charakteristiky, což jsou veličiny, které dostatečně výstižně popisují hlavní vlastnosti rozdělení četností.

Základními druhy charakteristik jsou míry polohy, reprezentující střed daného rozdělení a míry variability (měnlivosti), ukazující na rozptýlení hodnot statistického souboru (SHARMA, 2005). Jeden z typických a častých tvarů rozdělení četností statistických dat jsou na následujících obrázcích.



Obrázek 7 - Různá rozdělení četností



Obrázek 8 - Další typy rozdělení četností

Pro jednoduchost byl polygon četností nahrazen plynulou křivkou. Schémata zobrazují, jak odlišný tvar křivky ukazuje na rozdílné vlastnosti zobrazovaného rozdělení. Z obrázků je zřejmé, že informaci o vlastnosti rozdělení nemůže podat jedna charakteristika, ale je žádoucí použít nejméně dvě, a sice míru polohy a míru variability.

K charakteristikám polohy, nebo-li středním hodnotám počítáme (CHRÁSKA, 2008):

- aritmetický průměr
- geometrický průměr
- harmonický průměr
- medián
- modus
- kvantily

Střední hodnota je číslo, které charakterizuje obecnou úroveň hodnot znaku. Zastupuje jednotlivé hodnoty statistického znaku a umožňuje jednoduché srovnávání.

K charakteristikám variability neboli rozptýlení patří (HENDL, 2006):

- rozptyl (disperse)
- směrodatná odchylka (standardní odchylka)
- variační koeficient
- průměrná odchylka
- variační rozpětí
- kvantilové rozpětí

Míry variability měří stupeň různosti určité vlastnosti (tj. statistického znaku) u statistického souboru.

3.3 Míry polohy

3.3.1 Aritmetický průměr

Aritmetický průměr je nejběžněji používaný průměr a tedy i nejčastěji používaný druh středních hodnot. Aritmetický průměr je definován jako úhrn hodnot statistického znaku dělený rozsahem souboru (CYHELSKÝ, 1981):

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

Tento průměr se nazývá prostým aritmetickým průměrem (na rozdíl od tzv. váženého aritmetického průměru, který je vysvětlen v jiném odstavci). Aritmetický průměr nahrazuje hodnoty znaku všech prvků souboru tak, že se nemění celkový úhrn hodnot znaku (ČERMÁK, 1968):

$$\sum_{i=1}^{n} x_i = n\,\bar{x}$$

Důležité vlastnosti aritmetického průměru (CYHELSKÝ, 1981):

- výpočet aritmetického průměru je založen na všech napozorovaných hodnotách a tudíž je ovlivňován krajními naměřenými hodnotami
- součet odchylek jednotlivých hodnot od aritmetického průměru je vždy roven nule

$$\sum(x_i - \bar{x}) = 0$$

- součet čtverců odchylek jednotlivých hodnot od aritmetického průměru je menší než součet čtverců odchylek od jakékoli jiné hodnoty, než je aritmetický průměr.

3.3.2 Geometrický průměr

Předpokládejme, že během deseti let vzrostl určitý druh produkce 4-krát. Označme objem produkce na začátku prvního roku Y_0 a na konci tohoto roku Y_1. Objem produkce na konci posledního roku bude

$$Y_{10}: \; Y_{10} = 4\, Y_0$$

Koeficient vzrůstu výroby bude (CYHELSKÝ, 1988)

$$K = Y_{10} / Y_0 = 4\, Y_0 / Y_0 = 4$$

Nás zaujímá průměrný roční koeficient růstu. Postupovat pomocí aritmetického průměru není možné, protože již v prvním roce je koeficient menší než 1:

$$k_1 = 4 / 10 = 0.4$$

a neznamená vzrůst objemu, ale jeho pokles.

Průměrný koeficient musíme vypočítat pomocí jiného průměru, přičemž je jasné, že tento koeficient musí být větší než 1. Objem výroby bude narůstat pomocí průměrného koeficientu růstu K_{pr} následovně:

Základní rok	Y_0
1. rok	$Y_1 = Y_0 \,.\, K_{pr}$
2. rok	$Y_2 = Y_1 \,.\, K_{pr} = Y_0 \,.\, K_{pr2}$

3. rok	$Y_3 = Y_2 . K_{pr} = Y_0 . K_{pr3}$
...	...
10: rok	$Y_{10} = Y_9 . K_{pr} = Y_0 . K_{pr10}$

Celkový koeficient tempa růstu

$$K = Y_{10} / Y_0 = Y_0 . K_{pr10} / Y_0$$
$$K = K_{pr10} = 4$$

Jednoroční průměrný koeficient

$$K_{pr} = \sqrt[10]{4}$$

Vypočítali jsme jej geometrickým průměrem. Geometrický průměr je definován vzorcem (CYHELSKÝ, 1981):

$$\bar{x}_G = \sqrt[n]{x_i \cdot x_2 \ldots\ldots x_n}$$

Geometrický průměr nahrazuje individuální hodnoty znaku tak, že se nemění jejich součin. S jeho užitím je možno se setkat zejména při určování průměrného koeficientu růstu časové řady, v teorii indexů a korelačním počtu. Geometrický průměr nelze použít v případě, že statistický znak nabývá nulové hodnoty, popř. záporné hodnoty. Při výpočtu geometrického průměru lze využít toho, že logaritmus geometrického průměru je aritmetickým průměrem logaritmů jednotlivých hodnot (CYHELSKÝ, 1981).

3.3.3 Harmonický průměr

Harmonický průměr představuje reciproký aritmetický průměr reciprokých hodnot argumentu *x*, (ČERMÁK, 1968), tj.

$$x_h = \frac{n}{\sum \frac{1}{x_i}}$$

Harmonický průměr nahrazuje individuální hodnoty znaku tak, že se tím nemění součet jejich převratných hodnot. Používá se tam, kde má smysl součet reciprokých hodnot. S jeho užitím je možno se setkat zejména v teorii indexů, tj. při stanovení průměrů z poměrných hodnot, jak ukazuje následující příklad:

K zhotovení určitého výrobku potřeboval každý z 5-ti pracovníků 8 min, kdežto šestý dobu 10 min. Vypočtěme průměrnou dobu k zhotovení výrobku uvedenou skupinou 6-ti pracovníků:

$$x_h = \frac{6}{\frac{5}{8} + \frac{1}{10}} = 8{,}28$$

Průměrná doba k zhotovení jednoho výrobku je 8,28 minut.

3.3.4 Vážené průměry

Vážené (vhodněji váhové) průměry je způsob výpočtu z rozdělení četností. Při tom je nutno mít na zřeteli, že některé hodnoty znaků se opakují. Počet opakování znaku udává jeho četnost. Označíme-li různé hodnoty znaku (nikoliv tedy hodnoty znaku každého prvku) jako $x_1, x_2, x_3, \ldots x_K$ kde *K* je počet variant prvku, $n_1, n_2, n_3 \ldots n_K$ jsou adekvátní četnosti jednotlivých znaků, pak vážený aritmetický průměr se vypočítá podle vzorce (RUMSEY? 2007)

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

Ze vzorce je patrno, že jednotlivým různým hodnotám znaku se přisuzuje různá závažnost tj. různá váha. Analogicky se z rozdělení četností vypočítává vážený geometrický průměr a vážený harmonický průměr

3.3.5 Medián

Medián statistického souboru je ta hodnota argumentu *x*, která rozděluje posloupnost hodnot tohoto argumentu, uspořádanou podle velikosti, na dvě stejně početné skupiny (CYHELSKÝ, 1981). Budeme jej značit Me(x) (nebo též $\tilde{x}$. Jsou-li hodnoty statistického souboru uspořádány podle velikosti v posloupnost a má-li soubor rozsah n = 2m + 1 (tj. lichý počet prvků), je medián definován

$$Me(x) = x_{m+1}$$

Medián má tedy hodnotu znaku, který má pořadí m+1, nebo-li (n+1)/2. Ze sudého počtu měření je hodnota mediánu rovna aritmetickému průměru dvou hodnot znaků pořadí m a m+1 (ČERMÁK, 1968):

$$Me(x) = 1/2\ (x_m + x_{m+1})$$

Je-li medián malý, informuje o tom, že polovina prvků souboru má hodnotu sledovaného znaku zcela jistě malou. Naopak, je-li medián velký, má polovina prvků hodnotu zcela jistě velkou.

Medián na rozdíl od průměrů není funkcí všech hodnot statistického souboru. To je výhodná vlastnost zejména tehdy, jestliže soubor obsahuje jednotky s extrémně odchýlenými hodnotami znaku nebo, jestliže nejsou všechny hodnoty znaku známy.

Vychází-li se při určení mediánu z intervalového rozdělení četností, lze přesně stanovit pouze mediánový interval. Hodnotu mediánu odhadujeme lineární interpolací.

3.3.6 Kvantily

Ve statistické řadě uspořádané podle velikosti jsme definovali medián. Můžeme určit i další podobné charakteristiky, které rozdělují daný statistický soubor na stejně početné skupiny. Takové charakteristiky nazýváme kvantily.

Kvantilem K_p nazýváme takovou hodnotu znaku x, která odděluje p% statistických jednotek s hodnotou znaku menší nebo stejnou jako K_p, od (100-p)% hodnot které jsou větší nebo stejné jako je K_p (LAMSER, 1970).

Tak např. pětadvacetiprocentní kvantil K_{25} rozděluje daný soubor na čtvrtinu prvků, které jsou menší, příp. stejné jako K_{25} a na zbylé tři čtvrtiny prvků s hodnotou znaku větší nebo stejnou. Medián je kvantil K_{50}. Z ostatních kvantilů se používají nejčastěji kvartily, decily a percentily.

Kvartily rozdělují uspořádanou řadu na čtyři stejně početné skupiny. Je zřejmé, že kvartily jsou celkem tři: K_{25} (dolní), K_{50} (medián), K_{75} (horní).

Obdobně decily, kterých je celkem devět, rozdělují soubor na 10 stejně početných částí, obsahujících vždy 10% prvků. U velmi rozsáhlých souborů lze použít i percentily, které rozdělují uspořádanou řadu na 100 stejně početných částí, obsahujících vždy 1% prvků (WONNACOTT, 1992).

3.3.7 Modus

Modus ($\hat{x}$) je hodnota znaku, které přísluší nejvyšší četnost (RUMSEY, 2007). Určení modu jako jedné ze středních hodnot, má značnou důležitost v případě, jde-li o vystižení typické hodnoty znaku v daném souboru. U rozdělení četností ve tvaru písmene U lze v protikladu k jednovrcholovým rozdělením, u nichž připadá v úvahu modus, určit antimodus, a to jednoduše jako hodnotu znaku s nejmenší četností.

Společně se u jednoho rozdělení mohou vyskytovat modus a antimodus jen tehdy, jde-li o rozdělení složené (CYHELSKÝ, 1981). Příkladem budiž součet krajně asymetrického rozdělení osob zemřelých v nízkém věku a přibližně symetrické rozdělení osob zemřelých ve středním a vysokém věku.

3.3.8 Vztahy mezi aritmetickým průměrem, mediánem a modem

K nejčastěji zastoupeným typem rozdělení je jednovrcholové rozdělení četností. Ubývá-li četností od vrcholu rozdělení na obě strany přibližně stejně, hovoříme o symetrickém rozdělení. U ekonomických a sociálních jevů se poměrně často setkáváme s rozdělením asymetrickým. Rozlišujeme pak rozdělení zešikmená kladně, jestliže variabilita nadprůměrných hodnot je větší než hodnot podprůměrných. V opačném případě, jestliže variabilita podprůměrných hodnot je větší, hovoříme o zešikmení záporném (WONNACOTT, 1992).

V symetrickém rozdělení spadá aritmetický průměr, medián a modus do jedné hodnoty. Jde-li o symetrické rozdělení tvaru U, má úlohu modu antimodus. Naproti tomu čím bude rozdělení četností asymetričtější, tím více se budou aritmetický průměr, medián a modus vzájemně lišit. V asymetrickém rozdělení četností zešikmeném kladně platí vztah

$$\hat{x} < \tilde{x} < x$$

V asymetrickém rozdělení četností zešikmeném záporně platí vztah

$$x < \tilde{x} < \hat{x}$$

Není-li asymetrické rozdělení příliš nesouměrné, je vzdálenost mediánu od aritmetického průměru většinou přibližně jednou třetinou vzdálenosti mezi modem a aritmetickým průměrem (CYHELSKÝ, 1981).

3.4 Míry variability

V odstavci "Statistické charakteristiky" bylo uvedeno, že porovnání sledovaného znaku u dvou nebo i více souborů pouze mírou polohy je neúplné. Je třeba zkoumat i to, jak se v souboru jednotlivé hodnoty znaku vzájemně liší, neboť zcela různá rozdělení četností znaku mohou mít stejné míry polohy. K posouzení měnlivosti čili kolísání hodnot slouží míry variability.

Míra variability ukazuje nejen na hodnotu rozptylu znaku, ale udává význam i hodnotě polohy. Hodnota polohy charakterizuje obecnou úroveň řady s malou variabilitou lépe, než úroveň řady s velkou variabilitou. Měření variability je tedy účelné také pro posouzení kvality měr polohy. Bez znalosti měr variability zkoumaného souboru údajů nemůžeme správně vyhodnotit výsledky pozorování a tak správně rozhodnout při řešení technických a ekonomických úloh (WONNACAOTT, 1992).

Míry variability mají vedle středních hodnot největší význam v analýze statistických dat. Rozvoj moderních statistických metod je právě svázán s hodnocením variability a často v pravděpodobnostním aspektu.

3.4.1 Rozptyl, směrodatná odchylka a variační koeficient

Z měr variabilnosti se daleko nejčastěji využívá rozptyl a z něj odvozené charakteristiky. Důvodem jsou jedinečné vlastnosti rozptylu, které jsou využívány při dalším kvantitativním zpracování a analýze.

V úvodní kapitole jsme si o rozptylu a směrodatné odchylce řekli to nejpodstatnější, přičemž jsme se zmínili i o variačním koeficientu. Proto zde nebudeme výše popsané míry opakovat. Nyní se zde zmíníme o důležitých vlastnostech rozptylu (SHARMA, 2005):

- rozptyl je míra variability nejen vzhledem k aritmetickému průměru, ale i mírou vzájemných odchylek znaku mezi sebou,
- rozptyl lze vyjádřit jako průměr ze čtverců hodnot zmenšený o čtverec aritmetického průměru (tzv. výpočtový vzorec,
- rozptyl kolem aritmetického průměru je minimální.

3.4.2 Kvantilové rozpětí

Kvantilové rozpětí je definováno jako rozdíl horního a dolního kvantilu. Nejčastějším případem kvantilového rozpětí je kvartilové rozpětí, které zachycuje oblast 50% případů ve střední části souboru

$$R_{50} = Q_{75} - Q_{25}$$

Místo kvartilového rozpětí se někdy používá kvartilová odchylka, která je polovinou kvartilového rozpětí (MANĚNOVÁ, 2012).

3.5 Rozdělení četností kvalitativního znaku

Při zpracování kvalitativního znaku jsou možnosti kvantifikace omezené. Kvalitativní znak se dobře zpřehlední sestavením rozdělení četností, při čemž pořadí obměn znaku nemusí byt jednoznačné (jak tomu je u uspořádané řady kvantitativních znaků). V těchto případech nemá žádný smysl hodnocení tvaru rozdělení četností ani výpočet kumulativních četností. Naopak výpočet relativních četností má význam - udávají procentní zastoupení každé obměny kvalitativního znaku v souboru (HENDL 2005).

Pořadí obměn znaku bývá dáno logickou posloupností (např. vzdělání od nejnižšího k nejvyššímu), od nejvyšší četnosti k nejnižší nebo zavedenou zvyklostí. V případě, že kvalitativní znak může nabývat velmi mnoho obměn, sdružují se obměny znaku do skupin.

Skupiny musí být přesně definované, aby nedocházelo k nejasnostem při zařazování jednotek do skupin.

Zvláštním případem je alternativní znak, který má pouze dvě varianty a lze jej snadno kvantifikovat. Jedné hodnotě znaku přisoudíme hodnotu 1 a druhé hodnotu 0 (nula). Pak lze vypočítat průměr i rozptyl (CYHELSKÝ, 1981).

$$\bar{x} = \frac{(1\,n_1 + 0\,n_2)}{(n_1 + n_2)} = \frac{n_1}{n}$$

Aritmetický průměr alternativního znaku se tedy rovná relativní četnosti výskytu alternativy s hodnotou znaku rovnou 1.

$$s_x^2 = \frac{n_1}{n}\frac{n_2}{n}$$

Rozptyl alternativního znaku se tedy rovná součinu relativních četností obou alternativ (HENDL, 2005).

Na alternativní znak se mohou převést i množné kvalitativní znaky tím, že se rozdělí na dvě skupiny. Vhodným převodem je volba jednoho znaku, který nás zaujímá za jednu alternativu a za druhou alternativu ostatní zbývající obměny znaku.

3.6 Popis složených statistických souborů

Při statistické analýze se často setkáváme se soubory, které se skládají z více dílčích souborů. Ty vznikají členěním celého souboru podle dalšího jiného znaku. Např. soubor studentů gymnázií v ČR se může členit na dílčí soubory podle regionů.

Ve všech těchto případech nás zaujímá, zda lze z charakteristik dílčích souborů usuzovat na charakteristiky celého souboru nebo jakým způsobem se podílejí dílčí charakteristiky na vlastnostech celku.

3.6.1 Průměr složeného souboru

Aritmetický průměr složeného souboru jako celku vyjadřujeme váženým aritmetickým průmětem z dílčích aritmetických průměrů. Jako váhy zde vystupují rozsahy dílčích souborů (ČERMÁK, 1968):

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

kde x_i je aritmetický.průměr i-tého dílčího souboru a n_i je rozsah dílčího i-tého dílčího souboru a n je součet všech četností.

3.6.2 Rozptyl složeného souboru

Podobně jako u průměru lze celkovou variabilitu vyjádřit pomocí dílčích variabilit. Ze všech měr variability má tuto vlastnost pouze rozptyl. Rozptyl složeného souboru možno vyjádřit pomocí dílčích rozptylů a pomocí rozptylu dílčích průměrů kolem celkového průměru (ČERMÁK, 1968). Tento vztah má důležitý význam v praktické statistické analýze. Ve zjednodušeném tvaru se zapisuje takto:

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

- První složka $\overline{s^2}$ představuje (vážený) průměr dílčích rozptylů
- Druhá složka $s_{\bar{x}}^2$ je rozptylem dílčích průměrů

Rozklad rozptylu na dvě složky umožňuje analyzovat, jak se na celkové variabilitě podílí tzv. vnitroskupinový rozptyl (první složka) a rozptyl mezi dílčími soubory, tzv. meziskupinový rozptyl (druhá složka). Jinými slovy, vnitroskupinovému rozptylu odpovídá průměr rozptylu dílčích souborů, meziskupinovému rozptylu odpovídá rozptyl dílčích průměrů.

4 Indexy

4.1 Úvod ke srovnávání statistických souborů

Porovnání, nebo-li srovnání dvou objektů je jeden ze základních článků úsudku a předpokladu možnosti zobecnění. Prostředkem srovnávání statistických souborů jsou srovnávací poměrná čísla, nebo-li indexy. Indexy jsou důležitou součástí statistiky. Index vzniká porovnáním veličin o stejné náplni, lišících se však časově, prostorově (místně) nebo věcně. Nejčastěji se indexů používá k časovému srovnání.

Při výpočtu indexů je nutné rozlišovat extenzitní veličiny a intenzitní veličiny (CYHELSKÝ, 1981). Extenzitní veličiny jsou takové, které charakterizují množství, rozsah, počet, objem. Extenzitní veličiny lze sčítat označují se písmenem Q nebo q (např. počet obyvatel je extenzitní veličina, lze ji shrnovat sčítáním).

Intenzitní veličiny jsou takové, které charakterizují úroveň, intenzitu, hladinu, hustotu (LAMSER, 1970). Intenzitní veličinou je např. počet obyvatel na 1 km^2, cena 1kg jablek, průměrná hodinová mzda. Shrnují se průměrováním, protože sčítání by dospělo k nesmyslným údajům. Intenzitní veličiny vzniknou poměrem dvou extenzitních veličin (např. počet obyvatel na 1 km^2 = počet obyvatel dělený plochou, na které byl proveden součet obyvatel). Intenzitní veličina je označována písmenem p.

Nejčastěji se porovnává novější veličina ke starší. Novější veličině se připisuje znak (index) 1 (q_1, p_1), starší veličině znak 0 (q_0, p_0).

Např. změnu prodeje jablek v říjnu vůči měsíci září zjistíme indexem:

p_0 = 250 tun v září , p_1 = 400 tun v říjnu

$I = p_1/p_0 = 400/250 = 1.6$

Indexy jsou bezrozměrná čísla a mohou nabývat hodnot větších než nula a nebo v procentech od 0 do 100 %.

4.2 Rozdělení indexů

Indexy se v prvé řadě rozlišují na individuální a souhrnné (CYHELSKÝ, 1988). Individuelní indexy posuzují stejnorodé veličiny, např. množství prodeje jablek, průměrnou absenci na jednoho žáka ve školách, odpracované hodiny, průměrná mzda. Souhrnné indexy se používají pro posouzení vývoje několika nesourodých veličin současně (např. vývoj celkové produkce koksu a současně i svítiplynu, produkce koksu se udává v tunách, svítiplynu v m^3). Tyto veličiny jsou nesouměřitelné a proto je nutno hledat něco, co mají společné, např. peněžní jednotky. To je jeden z důvodů, proč nelze použít individuálních indexů.

Individuální indexy se člení takto:

a. Jednoduché
    i. množství (např. porovnání množství výkupu jablek od určitého dodavatele v roce 1991 a 1992)
    ii. úrovně (např. porovnání ceny 1 kg jablek od určitého dodavatele v roce a1991 a 1992)
b. Složené
    i. množství (např. porovnání množství výkupu jablek od více dodavatelů v roce 1991 a 1992)
    ii. úrovně (např. porovnání průměrné ceny 1 kg jablek více dodavatelů v roce 1991 a 1992)

Podle způsobu výpočtu se složený index úrovně rozlišuje na (WONNACOTT, 1992)

- index proměnlivého složení
- index stálého složení
- index struktury

Metodou výpočtu lze u souhrnného indexu určit komponenty, nazývané:

- hodnotový index
- objemový index
- cenový index

4.3 Individuální složené indexy

Při sledování vývoje shrnované extenzitní veličiny se používají rovnou součty. V následující tabulce jsou údaje o prodeji jablek v září a říjnu ve třech krajích. Pomocí dosud známých indexů budeme charakterizovat změny objemu prodeje, tržeb a cen.

kraj	cena v Kč		prodáno v 100 kg		tržba v Kč	
	p_0 září	p_1 říjen	q_0 září	q_1 říjen	p_0q_0 září	p_1q_1 říjen
1. kraj	360	370	100	500	36 000	185 000

2. kraj	370	380	200	400	74 000	152 000
3. kraj	380	400	300	300	114 000	120 000
CELKEM			600	1200	224 000	457 000

Jednoduché individuální indexy

	cena	prodej	tržba
	p_1/p_0	q_1/q_0	p_1q_1/p_0q_0
1. kraj	1,028	5,000	5,139
2. kraj	1,027	2,000	2,054
3. kraj	1,053	1,000	1,053

Jednoduché individuální indexy ukazují, že ve všech jednotlivých krajích došlo v říjnu oproti září k zvýšení cen, objemu prodeje i tržeb až na prodej u 3. kraje, u něhož se objem prodeje nezměnil. Pomocí individuálních složených indexů extenzitních veličin můžeme vypočítat indexy objemu prodeje a tržby (WONNACOTT, 1992) za všechny 3 kraje.

- **index objemu prodeje** (= suma prodeje v říjnu / suma v září)

$$I_q = \frac{\sum q_1}{q_0}$$

= 1200 / 600 = 2,000

- **index tržby** (= suma tržby za říjen / suma za září)

$$I_{qp} = \frac{\sum q_1 p_1}{\sum q_0 p_0}$$

= 457000 / 224000 = 2,040

V rámci uvedených 3 krajů došlo k zvýšení objemu prodeje 2 krát a objemu tržeb 2,040 krát.

4.3.1 Individuální složené indexy intenzitních veličin

V předešlé tabulce jsou u jednotlivých krajů uvedeny ceny jablek. Ceny jsou intenzitní veličiny a proto je nelze za všechny kraje sčítat (což by vedlo k nesmyslům), ale musíme vypočítat jejich průměrnou hodnotu:

průměrná cena v říjnu (= součet tržeb v říjnu / součet objemu za všechny kraje prodeje v říjnu)

$$\overline{p_1} = \frac{\sum q_1 p_1}{\sum q_1}$$

$\overline{p_1}$ = 457 000 / 1200 = 380,83

Z odvozeného vzorce je patrné, že průměr z jednotlivých intenzitních veličin se vypočítává jako jejich vážený průměr (CYHELSKÝ, 1988). Důsledkem toho je, že průměrná cena je ovlivněna nejen dílčími cenami, ale i objemem prodeje. Např. jestliže se ceny nebudou měnit, ale objem prodeje se zvýší u kraje, který má nejvyšší ceny, průměrná cena se zvýší. Naopak průměrná cena se sníží, jestliže se sníží objem prodeje v kraji s nižší cenou, i když původní ceny nebyly změněné. Kvantitativní údaj o vývoji průměrných cen poskytuje individuální složený index (ČERMÁK, 1968), tj. poměr průměrné ceny v říjnu a v září.

$$I_p = \frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}$$

I_p = (457 000/1 200) : (224 000/600) = 380,83 : 373,33 = 1,020

Tento index se nazývá **indexem proměnlivého složení**, protože se v něm uplatňuje nejen změna intenzitní veličiny p (v našem příkladě jednotlivé ceny jablek), ale i změna struktury (neboli složení) extenzitní veličiny (v našem příkladě objem prodeje) (CYHELSKÝ, 1988).

Samotný vliv intenzitní veličiny můžeme posoudit, jestliže budeme isolovat vliv struktury. Toho dosáhneme tím, že ve vzorci pro index ponecháme extenzitní veličinu stálou a to buďto q_0 nebo q_1. Tento index nazýváme **index stálého složení**.

Při stálém q_1:

$$\frac{\sum p_1 q_1}{\sum q_1} : \frac{\sum p_0 q_1}{\sum q_1} = \frac{\sum p_1 q_1}{\sum p_0 q_1}$$

Index stálého složení při stálém q_0 (CYHELSKÝ, 1981):

$$\frac{\sum p_1 q_0}{\sum p_0 q_0}$$

Naší tabulku doplníme o následující propočty:

kraj	$p_0.q_1$	$p_1.q_0$
1.	180 000	37 000
2.	148 000	76 000
3.	114 000	120 000
CELKEM	442 000	233 000

Index cen stálého složení při stálém q_1 = 457 000/442 000 = 1,034

Index cen stálého složení při stálém q_0 = 233 000 / 224 000 = 1,040

Tak jako jsme v předcházejícím indexu stabilizovali extenzivní veličinu, abychom mohli sledovat vliv intenzitní veličiny, můžeme stabilizovat intenzitní veličinu.

Získaný index se nazývá **index struktury**. Ukazuje, jak se uplatňuje vliv změn ve struktuře extenzitní veličiny, která je nositelem dané intenzity, na změnu dané intenzitní veličiny (v našem případě vliv změn v objemu prodeje jablek). Index struktury může být vypočten buď za stálého p_0 nebo p_1.

Index struktury při stálém p_0 (CYHELSKÝ, 1981):

$$\frac{\sum p_0 q_1}{\sum q_1} : \frac{\sum p_0 q_0}{\sum q_0}$$

(442 000/ 1 200) / (224 000 / 600) = 0,987

Celkový závěr z analýzy změny objemu prodeje a tržeb, zejména změny cen jablek v říjnu oproti září za všechny kraje lze komentovat takto: Ukazuje se, že v říjnu oproti září se zvýšily průměrné ceny jablek asi o 2%, přestože došlo k mírnému přesunu ve struktuře prodeje ve prospěch krajů s levnějšími jablky.

Index proměnlivého složení je roven součinu indexu stálého složení a indexu struktury. Tento rozklad indexu proměnlivého složení umožňuje charakterizovat směr (tj. záporný nebo kladný vliv) a stupeň působení samotné intenzitní veličiny *p* a struktury *q* na celkovou změnu intenzitní veličiny (WONNACOTT, 1992).

Z vypočtených hodnot v příkladu zapíšeme rozklad indexu proměnlivého složení:

1,020 = 1,034 x 0,987

který je v souladu s naším závěrem o vlivu na vývoj průměrných cen.

4.3.2 Absolutní změna

Vhodným doplňkem indexů jsou absolutní změny porovnávaných veličin (značíme "delta" - Δ). Absolutní změna extenzitních veličin, které jsou vyjádřeny poměrem, se získá odečtením čitatele od jmenovatele:

$$\Delta_q = q_1 - q_0$$

U indexů intenzitních veličin se rozdíl zpravidla násobí odpovídající extenzitní veličinou z běžného období (LAMSER, 1970):

u indexu proměnlivého složení

$$\Delta_{pr.sl.} = \sum q_1 \, (\overline{p}_1 \;\; \overline{p}_0)$$

u indexu stálého složení

$$\Delta_{st.sl.} = \sum p_1 q_1 - \sum p_0 q_1$$

u indexu struktury

$$\Delta_{str.} = \sum q_1 \left(\frac{\sum p_0 q_1}{\sum q_1} - \sum q_1 \frac{\sum p_0 q_0}{\sum q_0} \right)$$

Součet absolutních změn vypočtených z indexu stálého složení a indexu struktury podle shora uvedených vzorců je roven absolutní změně vypočtené z indexu proměnlivého složení:

$$\Delta_{pr.sl} = \Delta_{st.sl.} + \Delta_{str.}$$

Pokračujme v našem příkladu. Absolutní efekt ze zvýšení cen, vázaný na objem běžného období, činí:

1 200 . (380,83 - 373,33) = 9 000 Kč

a lze jej rozložit na součet absolutního efektu z izolované změny cen, který se rovná

457 000 - 442 000 = 15 000 Kč

a absolutní efekt ze změny struktury prodeje:

1 200 . (368,83 - 373,33) = 6 000 Kč.

4.4 Souhrnné indexy

Souhrnné indexy (jak jsme již poznamenali) se používají při současném posouzení nesourodých veličin. Např. vývoj celé produkce koksárny, která vyrábí koks (měrná jednotka t) a svítiplyn (měrná jednotka m^3) lze posoudit tak, že ji vyjádříme v peněžních jednotkách (součet tuny a metry krychlové je nesmyslný).

Problematika souhrnných indexů bude vysvětlena na následujícím příkladě, uvádějící produkce a ceny koksárny v roce 1991 a 1992

Výrobek	Měrná jednotka	Produkce v tis.		cena za jednotku	
		1991	1992	1991	1992
koks	t	500	800	60	50
svítiplyn	m^3	1 200	1 300	30	25

Produkci označíme q_0, q_1, cenu označíme p_0, p_1, produkce v peněžních jednotkách v roce 1991 q_0p_0, v roce 1992 q_1p_1

Hodnota produkce celé koksárny je v roce 1991 $\sum q_0 p_0$, v roce 1992 $\sum q_1 p_1$

Tabulku se zadanými údaji rozšíříme o součiny a jejich součty, které potřebujeme v dalším výkladu.

	p_0q_0	p_1q_1	p_0q_1	p_1q_0
	30 000	44 000	48 000	27 500
	36 000	32 500	39 000	30 000
CELKEM	66 000	76 500	87 000	57 500

Pro porovnání hodnoty produkce r. 1992 k roku 1991 vypočítáme hodnotový index (index tržby):

$$I_h = \frac{\sum q_1 p_1}{\sum q_0 p_0}$$

76 000 / 66 000 = 1,159

Uvedeným vzorcem jsme vypočítali index hodnotový. Hodnotový index sleduje vývoj něčeho, co je vyjádřeno hodnotovými jednotkami. Měří změnu hodnoty shrnuté jak z extenzitní tak intenzitní veličiny. Měří tedy vliv obou těchto veličin současně.(touto vlastností je obdobou indexu proměnlivého složení). Vypočtený výsledek ukazuje, že hodnota produkce koksárny stoupla v roce 1992 1,159 krát (o 15,9 %).

Dalším souhrnným indexem je objemový index. Tento index měří vliv změny objemu, tedy extenzitní veličiny na hodnotu "něčeho" (standardu). Vylučuje tudíž změny intenzitní veličiny p a proto do výpočtu dosazujeme buďto p_0 a nebo p_1. Má tedy opět dva tvary (jako u individuelních indexů úrovně):

$$I_o = \frac{\sum q_1 p_0}{\sum q_0 p_0}$$

87 000 / 66 000 = 1,318

Tento objemový index udává zvýšení produkce o 31,8 % v případě ocenění produkce v obou letech cenami za základní období. Tento první tvar objemového indexu je v ekonomické praxi nejběžnější (WONNACOTT, 1992), protože pro měření vývoje se používá cen z určitého základního období.

Druhý tvar objemového indexu:

$$I_o = \frac{\sum q_1 p_1}{\sum q_0 p_1}$$

udává zvýšení produkce o 33 % při ocenění produkce v obou letech cenami běžného období.

Při vyloučení vlivu extenzitní veličiny, tedy pro zachycení vlivu intenzitní veličiny (kterou je velmi často cena) se používá tzv. cenový index. Má opět dva tvary, podle toho, zda dosazujeme q_0 anebo q_1:

$$I_c = \frac{\sum q_0 p_1}{\sum q_0 p_0}$$

57 500 / 66 000 = 0,871

$$I_c = \frac{\sum q_1 p_1}{\sum q_1 p_0}$$

76 500 / 87 000 = 0,879

Z indexů vyplývá: Kdyby došlo jen ke změně cen, zatímco objem produkce by byl stále stejný:

- jako v roce 1991, hodnota produkce by klesla na 87,1 %
- jako v roce 1992, hodnota produkce by klesla na 87,9 %

Podobně jako u individuálních indexů platí tento vztah (CYHELSKÝ, 1988):

$I_h = I_o$ se stálou p_0 x I_c se stálou q_1

což se používá nejčastěji, anebo:

$I_h = I_o$ se stálou p_1 x I_c se stálou q_0

Z výpočtů vyplývá tento obecný závěr: Výše uvedené indexy ukazují na vzestup hodnoty výroby koksárny v roce 1992 proti roku 1991 o 31,8 %, na vzestup fyzického objemu výroby a na pokles cen.

Poznali jsme, že dva typy indexů vedou k rozdílným výsledkům. K dosažení jednoznačnosti při měření fyzického objemu se používá Fisherova vzorce souhrnného indexu fyzického objemu (WONNACOTT, 1992):

$$I_q = \sqrt{\frac{\sum q_1 p_0}{\sum q_0 p_0} \, x \, \frac{\sum q_1 p_1}{\sum q_0 p_1}}$$

K dosažení jednoznačnosti při měření změny úrovně cen se používá Fisherova vzorce souhrnného cenového indexu (WONNACOTT, 1992):

$$I_p = \sqrt{\frac{\sum q_1 p_1}{\sum q_1 p_0} \, x \, \frac{\sum q_0 p_1}{\sum q_0 p_0}}$$

Pro náš příklad:

Fisherův objemový index $I_q = \sqrt{1{,}318 \, . \, 1{,}330} = 1{,}324$

Fisherův cenový index $I_p = \sqrt{0{,}879 \, . \, 0{,}871} = 0{,}875$

Kontrola výpočtu $I_h = I_q \, I_p = 1{,}324 \, . \, 0{,}875 = 1{,}158$

4.5 Indexy bazické a řetězové

Mějme časovou řadu těžby hnědého uhlí (která se skládá z těchto údajů: rok, těžba). Sledovanou veličinu (zde hodnoty těžby) lze srovnávat mezi sebou pomocí indexů. Zvolíme-li u všech indexů stejný časový základ, kterým může být obecně hodnota z libovolného období, dostaneme řadu tzv. **bazických indexů** (SHARMA, 2005).

$$I_b = \frac{q_i}{q_0}$$

kde $i = 1,2,3 \ldots k$

Absolutní přírůstky jsou zřejmě

$$\Delta_q = q_i - q_0$$

Řadu časových srovnání můžeme provést také tím, že srovnáváme hodnoty v jednotlivých obdobích s bezprostředně předcházejícím obdobím. Tyto indexy nazýváme **řetězové indexy** (SHARMA, 2005). Jde v podstatě o tempa růstu (CYHELSKÝ, 1981).

$$I_r = \frac{q_i}{q_{i-1}}$$

kde i = 2,3 ....k

Odpovídající absolutní přírůstky jsou:

$$\Delta_r = q_i - q_{i-1}$$

Mezi řetězovými a bazickými indexy existují vztahy, které umožňují bez znalosti původních dat přepočítat tyto indexy mezi sebou.

Výpočet řetězového indexu z bazických:

$$\frac{q_i}{q_0} = \frac{q_i}{q_0} \Big/ \frac{q_{i-1}}{q_0}$$

Výpočet bazického indexu z řetězových:

$$\frac{q_i}{q_0} = \frac{q_1}{q_0} \frac{q_2}{q_1} \dots \frac{q_i}{q_{i-1}}$$

5 Jednoduchá regrese

5.1 Úvod

Předcházející kapitoly byly věnovány popisu rozdělení četnosti pouze jednoho znaku. V praktických příkladech jsme poznali, že u statistických jednotek se sleduje současně více statistických znaků. Tyto znaky mohou mít mezi sebou různé souvislosti a vztahy, mohou být mezi sebou závislé (RUMSEY, 2007). Např. výše výdajů za potraviny v rodině bude jistě ve vztahu k počtu osob v domácnosti.

5.1.1 Typy závislostí

Z fyziky známe závislosti, kdy určité hodnotě jedné veličiny (nezávisle proměnné) odpovídá zcela přesně hodnota veličiny druhé (závisle proměnné). Např. pro volný pád platí[2]

$$s = 1/2\ g\ t^2\ (= 1/2 \times 9.81 \times t^2)$$

Každé zvolené hodnotě času t odpovídá jediná hodnota délky dráhy s. Závislost, kde určité hodnotě nezávisle proměnné veličiny odpovídá jediná hodnota veličiny závisle proměnné, se nazývá **závislost funkční**.

Mezi jevy a procesy tato jednoznačnost většinou neexistuje. Příkladem může být závislost mezi tělesnou výškou a věkem u mládeže. Takové **závislosti** se nazývají **statistické**.

5.1.2 Uspořádané dvojice

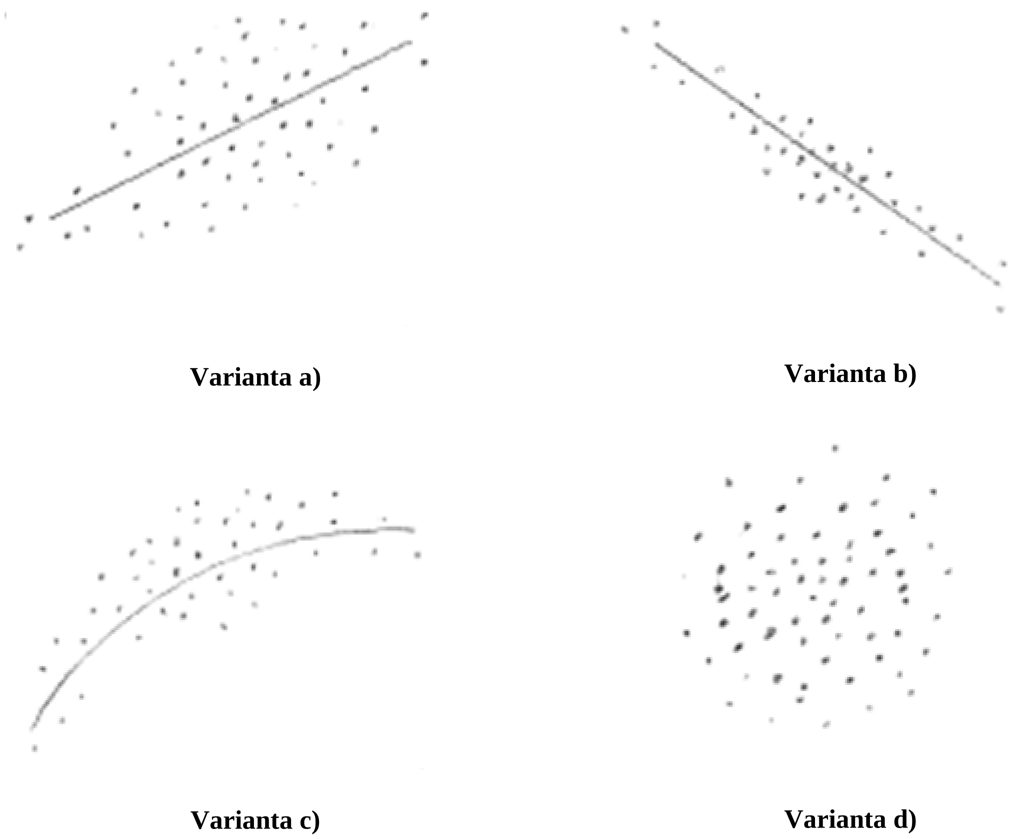
K úlohám statistiky patří zkoumání vztahů mezi danými jevy nebo objekty. K tomuto cíli se používá různých metod vícerozměrné analýzy, které se rozvíjely a stále se rozvíjejí v důsledku a to nejen zavádění výpočetní techniky. Ze všech metod je nejpropracovanější metoda regresní analýzy. Vzhledem na poslání těchto skript se omezíme pouze na popis vícerozměrných statistických souborů - statistickou indukci (o které byla zmínka při popisu jednorozměrných statistických souborů) zde pomineme.

Uvažujme soubor, na kterém měříme hodnoty dvou proměnných, takže získáváme dvojice výsledků měření: (x_1,y_1), (x_2,y_2), (x_3,y_3), ..., (x_n,y_n).

[2] viz FEYMAN, R., P., LEIGHTON, R., B., SANDS, M.: Feymanovy přednášky z fyziky - revidované vydání - I. Praha, Fragment. 2013. ISBN 978-80-253-1642-9díl.

5.1.3 Grafické znázornění vztahů

Výsledky lze znázornit tečkovým (bodovým) diagramem, který dobře vyhovuje pro první informaci o statistické závislosti. Na obrázku jsou schematicky zakresleny čtyři různé varianty statistických závislostí.



Obrázek 9 - Varianty bodového grafu (regresní analýza)

U obr. a) b) c) je zřetelně vidět určitou tendenci závislosti y na x. Takovou závislost nazýváme **korelací**. Na obr. d) jednotlivé body vyplňují zhruba plochu kruhu a proto o tendenci závislosti nelze uvažovat. Obě proměnné jsou v tomto případě korelačně nezávislé. Průběh závislosti na obr. a) b) c) je znázorněn čarou, kterou nejčastěji vystihujeme matematickou funkcí, a nazývá se **regresní funkce**. Pomocí ní můžeme ze známé hodnoty jednoho znaku odhadnout průměrnou hodnotu druhého znaku (LAMSER, 1970). Při tom máme na zřeteli, že odhadovaná hodnota se bude vlivem vedlejších a náhodných vlivů lišit od skutečnosti.

Přesnost odhadu je závislá na stupni rozptýlení hodnot závisle proměnné kolem příslušné regresní funkce. Odhady budou tím přesnější, čím méně jsou empirické hodnoty (tj. napozorované, zjištěné) rozptýleny kolem hodnoty regresní funkce (tj. vypočítané, teoretické) (KOZÁK, 1994). Na obr. b) budou odhady přesnější než na obr. a), protože jsou na něm body těsněji nakupeny kolem regresní přímky. Výpočet této těsnosti, neboli intenzity korelační závislosti je dalším hlavním úkolem **regresní analýzy.**

5.1.4 Kovariance

V následujících kapitolách se budeme setkávat s pojmem **kovariance**, proto si jej hned vysvětlíme. Kovariance je mírou rozptylu dvou proměnných veličin (CYHELSKÝ, 1981):

$$cov_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

kde n je počet dvojic (pozorování).

Kovariance je obdoba rozptylu jednorozměrné proměnné veličiny, což ukazuje i tento výpočet (odvození) (ČERMÁK, 1968):

$$cov_{xy} = \frac{\sum(x_i - \bar{x})(x_i - \bar{x})}{n} = \frac{\sum(x_i - \bar{x})^2}{n} = s_x^2$$

V literatuře se setkáváme s jednodušším (výpočetním) tvarem vzorce:

$$cov_{xy} = \overline{xy} - \bar{x}\,\bar{y}$$

5.1.5 Spojnice podmíněných průměrů jako regresní čára

Začněme příkladem závislosti čtvrtletních výdajů za potraviny na čtvrtletních příjmech domácnosti. Vstupní data jsou v následující, která je jen naznačena.

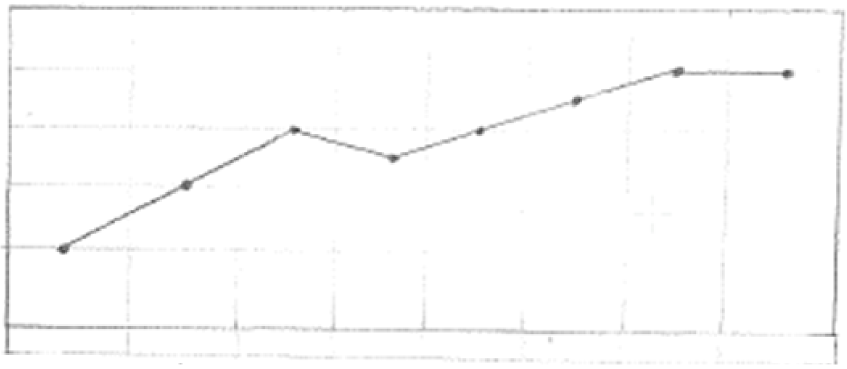
Čís. domácnosti	Příjmy (Kč)	Výdaje za potraviny (Kč)
1	6 277	1 818
2	4 860	1 289
3	13 808	1 473
4	3 590	1 068

atd...

Ze vstupních dat byla sestavena tabulka, v níž jsou skupiny nezávisle proměnné x (čtvrtletní příjem), skupinové průměry nezávisle proměnné x_i a podmíněné průměry závisle proměnné y_i (výdaje za potraviny). Podmíněný průměr závisle proměnné y_i je průměr hodnot y_i, které patří do skupiny, vytvořené z intervalu x (například pro interval 2 500 - 4 500 by výpočty mohly vypadat takto):

Skupiny příjmů	2 500-4 500
Skupinové průměry x_i	3 578
Podmíněné průměry y_i	1 514

Závislost mezi x_i a y_i, tj. mezi skupinovými průměry nezávisle proměnné a podmíněnými průměry závisle proměnné je znázorněna grafu:



Obrázek 10 - Závislost skupinových průměrů a podmíněných průměrů (závislost čtvrtletních výdajů za potraviny na čtvrtletních příjmech v domácnosti)

Na grafu je regresní čarou spojnice podmíněných průměrů závisle proměnné. Ta dává představu o tendenci výdajů za potraviny v závislosti na příjmech. Neznámou hodnotu závisle proměnné na základě zvolené hodnoty nezávisle proměnné (patřící do určité skupiny) odhadujeme jako podmíněný průměr závisle proměnné (odpovídající dané skupině) (SHARMA, 2005). Např. při příjmu 15 000 Kč je odhad výdajů 4 856 Kč.

5.1.6 Korelační poměr

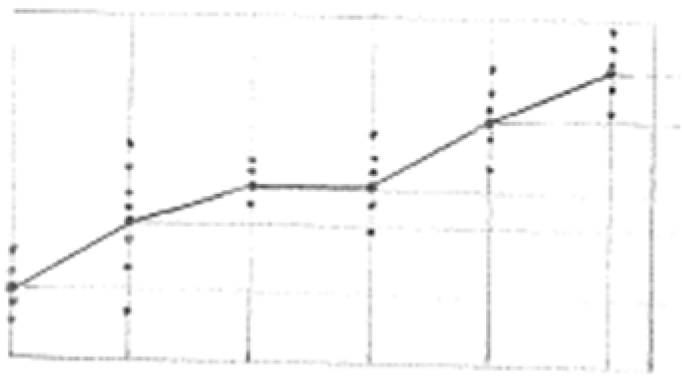
Odhady hodnot závisle proměnné budou tím přesnější, čím menší je variabilita napozorovaných (empirických) hodnot y_i ve všech skupinových četnostech. Čím je variabilita menší, tím je korelační závislost těsnější a naopak (HANOUSEK, 1992).

Měření stupně korelační závislosti je založeno na rozkladu celkového rozptylu závisle proměnné:

$$s_y^2 = \overline{s^2} + s_{\bar{y}}^2$$

- $\overline{s^2}$ je průměr podmíněných rozptylů
- $s_{\bar{y}}^2$ je rozptyl podmíněných průměrů

Geometrické znázornění rozptylů lze demonstrovat takto:



Obrázek 11 - Znázornění rozkladu celkového rozptylu

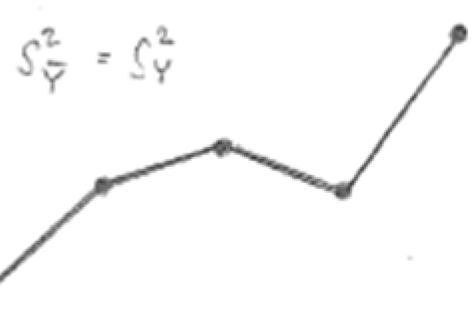
Korelační závislost y na x je tím těsnější, čím více se podílí rozptyl podmíněných průměrů $s_{\bar{y}}^2$ na celkovém rozptylu s_y^2. Číselně se korelační závislost vyjadřuje jejich poměrem, který se nazývá **poměr (index) determinace** (CYHELSKÝ, 1988):

$$I_{yx}^2 = \frac{s_{\bar{y}}^2}{s_y^2}$$

Podívejme se na dva extrémní případy tohoto poměru. Na následujícím obrázku neexistuje rozptyl hodnot y kolem podmíněných průměrů, čili průměr podmíněných rozptylů je nulový a rozptyl podmíněných průměrů se rovná celkovému rozptylu:

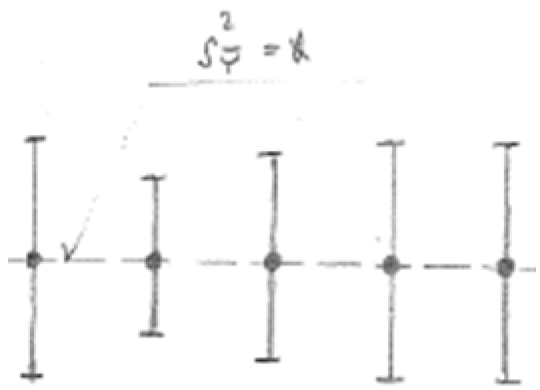
$$s_{\bar{y}}^2 = s_y^2$$

a tedy poměr determinace = 1



Obrázek 12 - Příklad neexistence rozptylu hodnot y

Další obrázek naopak ukazuje případ, kdy rozptyl podmíněných se průměrů rovná nule ($s^2_{\bar{y}} = 0$), takže průměr podmíněných rozptylů se rovná celkovému rozptylu ($\overline{s^2} = s^2_y$). Poměr determinace se rovná nule.



Obrázek 13 - Příklad neexistence rozptylu podmíněných průměrů

Poměr determinace se pohybuje v rozmezí od nuly do 1. Čím je blíže jedné, tím je korelační závislost silnější (LAMSER, 1970).

Často se používá druhá odmocnina poměru determinace, která se nazývá **poměr (index) korelace**:

$$I_{yx} = \sqrt{s^2_{\bar{y}}} = \frac{s_{\bar{y}}}{s_y}$$

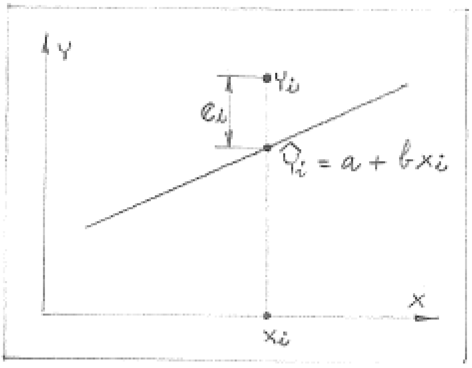
5.2 Lineární regrese

U odhadů závisle proměnné na základě podmíněných průměrů dostáváme jednu hodnotu y pro celý skupinový interval nepodmíněných proměnných x. To není výhodné. Individuelní hodnotu y odpovídající jedné hodnotě x můžeme získat, když pro regresní závislost zvolíme nějakou funkci (RUMSEY, 2007).

5.2.1 Rovnice přímky

Určíme-li za regresní funkci tvar přímky, hovoříme o jednoduché lineární regresi. Jednoduché proto, že se ve funkci vyskytuje jen jedna nezávisle proměnná x

$$\hat{Y}_i = a + bx_i$$



Obrázek 14 - Lineární regresní přímka

5.2.2 Rezidua

Z obrázku vyplývá, že individuelní hodnota závisle proměnné y_i se rovná součtu hodnoty regresní funkce $\hat{Y}_i$ a zbytku (reziduu) e_i:

$$y_i = \hat{Y}_i + e_i \text{, takže}$$

$$e_i = y_i - \hat{Y}_i$$

Důležitá je suma čtverců všech reziduálních odchylek (WONNACOTT, 1992):

$$Q_e = \sum e_i^2 = \sum (y_i - \hat{Y}_i)^2$$

Její minimální hodnota je podmínkou pro nejvhodnější proložení přímky mrakem v bodovém diagramu. Nejvhodnější poloha a sklon regresní přímky bude tedy odpovídat minimální hodnotě součtu čtverců odchylek od regresní přímky.

Poznámka: Pro zjednodušení symboliky budeme písmeno Q používat všeobecně pro sumu čtverců různých odchylek, např. Q_x je $\sum(x_i - \bar{x})^2$, Qy je $\sum(y_i - \bar{y})^2$. Analogicky použijeme písmeno Q i pro součet součinů dvojice odchylek x a y: $Q_{yx} = \sum(x_i - \bar{x})(y_i - \bar{y})$.

Rezidua tvoří jednorozměrný statistický soubor. Programy statistických paketů umožňují vytisknout hodnoty reziduálních odchylek, zobrazit je v bodovém diagramu nebo zobrazit jejich histogram.

Vhodné je posouzení číselných charakteristik reziduí, tj. rozptyl e (reziduální rozptyl) a směrodatnou odchylku e. Reziduální rozptyl je průměrná hodnota čtverce reziduální odchylky:

$$s_e^2 = \frac{Q_e}{n}$$

Při malém počtu *n* se do jmenovatele dosazuje (*n*-2) (KOZÁK, 1994). Druhá odmocnina tohoto rozptylu je směrodatná odchylka reziduální variability, tj. variability kolem regresní přímky. Poznamenejme, že reziduální rozptyl má v americké literatuře mnoho synonym.

5.2.3 Metoda nejmenších čtverců (MNČ) a odhad parametru *b*

Pro podmínku Qe=minimum se celkem jednoduchým řešením získají dvě lineární rovnice, z kterých se vypočítá neznámý parametr b pro rovnici přímky. Tomuto způsobu řešení se říká **metoda nejmenších čtverců** (HENDL, 2006).

$$b = Q_{xy} / Q_x$$

Známější alternativa vzorce je podíl kovariance a rozptylu proměnné x:

$$b = \frac{cov_{yx}}{s_x^2}$$

Oba vzorce jsou adekvátní, neboť

$$cov_{yx} = \frac{Q_{yx}}{n}$$

a

$$s_x^2 = \frac{Q_x}{n}$$

Nakonec, při známém b, vypočítáme z rovnice přímky

$$y_i = a + bx_y$$

(resp. $\bar{y} = a + b\bar{x}$)

parametr a:

$$a = \bar{x} - b\bar{y}$$

Příklad (CYHELSKÝ, 1988): Pro hodnoty nezávisle proměnné x a jim odpovídající hodnoty závisle proměnné y uvedené v následující tabulce (sloupce Y a X) je uvažován lineární vztah. Úkolem je vypočítat parametry regresní přímky. Pro sledování výpočtu jsou v tabulce uvedeny i další potřebné výpočty (slupce Y^2, X^2 a XY). U sum čtverců jsou kromě definičních vzorců uvedeny ještě výpočtové vzorce.

	Y	X	Y^2	X^2	XY
	36	9	1 296	81	324
	80	15	6 400	225	1 200
	44	10	1 936	100	440
	55	11	3 025	121	605
	35	10	1 225	100	350
CELKEM	250	55	13 882	627	2 919

Výpočty:

$$\bar{y} = \frac{\sum y_i}{n} = 250 / 5 = 50$$

$$\bar{x} = \frac{\sum x_i}{n} = 55 / 5 = 11$$

$$Q_{yx} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n} = 2\,919 - \frac{55.250}{5} = 169$$

$$Q_x = \sum (x_i - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} = 627 - \frac{552}{5} = 22$$

$$b = \frac{Q_{yx}}{Q_x} = \frac{169}{22} = 7{,}68$$

$$a = \bar{y} - b\,\bar{x} = 50 - 7{,}68\,.11 = -34{,}5$$

Regresní funkce má tedy tvar:

$$\hat{Y}_i = a + bx_i = -34{,}5 + 7{,}68.x_i$$

Například pro zvolenou hodnotu *x*=13 je

$$\hat{Y}_i = a + bx_i = -34{,}5 + 7{,}68\,.13 = 65{,}36$$

5.2.4 Index determinace a index korelace

Z metody nejmenších čtverců vyplývá tento rozklad součtu čtverců odchylek závisle proměnné:

$$Q_y = Q_{\hat{Y}} + Q_e$$

což znamená, že celkový součet čtverců se rovná teoretickému součtu čtverců plus reziduální součet čtverců (HÁTLE, 1983). Jak již víme, jednotlivé sumy čtverců jsou definovány takto:

$$Q_y = \sum (y_i - \bar{y})^2$$

$$Q_{\hat{y}} = \sum (\hat{y}_i - \bar{y})^2$$

$$Q_e = \sum (e_i - \bar{e})^2$$

Vlivem závislosti y na x dané regresní přímkou vzniká variabilita hodnot y_i, které odpovídá teoretický součet čtverců $Q_{\hat{y}}$. Předpokládáme, že vlivem náhodných činitelů dochází k variabilitě jednotlivých pozorovaných hodnot kolem regresní přímky (reziduální odchylky) a té odpovídá reziduální součet čtverců. Poměr teoretického a celkového součtu čtverců odchylek se interpretuje jako podíl variability y vysvětlený regresní funkcí a je tedy vhodnou mírou těsnosti korelace (CYHELSKÝ, 1981):

$$I^2 = \frac{Q_{\hat{y}}}{Q_y}$$

Nazývá se index determinace a nabývá hodnot od 0 do 1. Kladná hodnota odmocniny indexu korelace se nazývá index korelace:

Příklad (pokračování):

$$Q_y = 1\ 382$$

$$Q_{\hat{y}} = 1\ 298{,}2$$

$$Q_\varepsilon = 83{,}8$$

$$I^2 = \frac{1\ 298{,}2}{1382} = 0{,}939$$

$$I = \sqrt{0{,}939} = 0{,}97$$

Závěr: Korelační závislost je velmi těsná, 94% rozptylu pozorovaných hodnot y je způsobeno regresní funkcí, pouze 6% rozptylu je způsobeno náhodnými, blíže neurčenými vlivy.

5.2.5 Koeficient korelace

Poznamenejme, že index determinace a index korelace se používá i pro jiné regrese než je přímková. Specielně pro přímkovou regresi se používá převážně tzv. **koeficient korelace**. Jeho absolutní hodnota je stejná jako u indexu korelace, nabývá však hodnot od -1 do +1. Z toho plyne, že koeficient korelace je i ukazatelem smyslu závislosti: při záporné hodnotě ukazuje na nepřímou korelační závislost (viz předcházející obrázky), při kladné hodnotě na přímou korelační závislost (HENDL, 2006).

V literatuře se můžeme setkat s různými definičními a výpočtovými vzorci. Jeden z nejjednodušších:

$$r_{yx} = \frac{cov_{yx}}{s_x^2}$$

Všimněme si, že u lineární regrese se koeficient korelace rovná parametru *b*. Parametru *b* se u rovnice přímky říká regresí koeficient a udává sklon regresní přímky (SHARMA, 2005).

5.3 Nelineární modely regrese

V praxi se stává, že v některých případech přímková regrese nevystihuje dobře průběh závislosti. Potom za regresní funkci volíme křivku, např. parabolu:

$$\hat{Y} = b_0 + b_1 + b_2 X^2$$

nebo polynom vyššího stupně. Příklady dalších funkcí mohou být tyto (MELOUN, 2006):

$$\hat{Y} = b_0 + \frac{b_1}{X} + \frac{b_2}{X^2}$$

$$\hat{Y} = b_0 + b_1 \ln X$$

Poznamenejme si, že všechny uvedené funkce jsou sice nelineární v nezávisle proměnné x, ale lineární v parametrech b0, b1, b2. U funkcí, které jsou lineární v parametrech, vypočítávají se tyto neznámé parametry ze soustav lineárních rovnic získaných opět metodou nejmenších čtverců.

5.3.1 Transformace funkcí nelineárních v parametrech.

Regresní funkce nelineární v parametrech mají složitější tvary např.:

$$\hat{Y} = b_1 . b_2 . X$$

$$\hat{Y} = {b_1}^{b_2} . X$$

Funkce nelineární v parametrech lze často upravit jednoduchou transformací na lineární tvar. Po transformaci je nová funkce již v parametrech lineární a proto se řeší opět metodou nejmenších čtverců a soustavou lineárních rovnic. Nejčastěji se používá logaritmická transformace (MELOUN, 2006).

Např. exponenciální funkce

$$\hat{Y} = ab^x$$

se transformuje na funkci

$$\ln \hat{Y} = \ln a + x . \ln b$$

Po substituci

$$Y := \ln \hat{Y}, \alpha := \ln a, \beta := \ln b$$

se řeší funkce

$$Y = \alpha + \beta . x$$

Místo vstupních dat y_i zavádíme ln y_i. Počítáme-li konkrétní hodnoty $\hat{y}_i$, výsledek odlogaritmujeme.

Příklad: Následující data obsahují závislost úrody stromů určité odrůdy na věku stromů po přesazení. Předpokládá se regresní funkce tvaru exponenciální křivky, tj. typu y = a b^x

Nejprve budeme transformovat vstupy pomocí transformace (substituce): log y = w. Vstupní data jsou následující:

věk (x): 2, 2, 3, 5, 4, 3, 2, 3, 4, 4, 3, 5

úroda (y): 2, 3, 2, 10, 5, 3, 1, 4, 7, 6, 3, 12

Výsledkem řešení je regresní funkce:

w = -0.296962 + 0.265011 x

Po odlogaritmování dostáváme finální verzi exponenciálního modelu:

$$\hat{Y} = 0{,}505 . 1{,}841^x$$

5.3.2 Index determinace a index korelace

Pro index determinace u nelineárních regresních funkcí platí stejný vzorec jako pro přímkovou závislost:

$$I^2 = \frac{Q_{\hat{y}}}{Q_y}$$

Zde však znamená sumu čtverců odchylek způsobenou regresní křivkou. Opět platí

$$Q_y = Q_{\hat{Y}} + Q_e$$

kde Q_e je suma čtverců odchylek nezávisle proměnné od regresní křivky. Vlivem závislosti y na x dané regresní křivkou vzniká variabilita hodnot y_i, které odpovídá teoretický součet čtverců $Q_{\hat{Y}}$.

Podíl teoretického součtu čtverců $Q_{\hat{Y}}$z celkového součtu čtverců Q_y se vysvětluje vlivem regresní funkce. Působením vedlejších a nahodilých činitelů kolísají pozorované hodnoty y_i kolem regresní křivky a tomu odpovídá reziduální součet čtverců Q_e.:

Index korelace je kladná hodnota z druhé odmocniny indexu determinace:

Při interpretaci indexu korelace je třeba pamatovat na to, že index korelace měří těsnost závislosti závisle na zvolené regresní funkci. Nízká hodnota indexu korelace nemusí ukazovat na malou intenzitu korelační závislosti, ale na zvolenou nevhodnou regresní funkci. Z tohoto důvodu míru těsnosti statistické závislosti používáme jako kriterium posuzování vhodnosti regresní funkce (MELOUN, 2006).

U transformovaných funkcí se ovšem index determinace a korelace nevztahuje na původní proměnnou, ale na transformovanou proměnnou.

5.4 Volba regresní funkce

V předcházejících kapitolách jsme si vysvětlili, že regresní funkce může být lineární nebo nelineární. V některých případech není dopředu zřejmé, o jaký konkrétní typ funkce jde. Prvním krokem při řešení regrese je v tomto případě její výběr. Velmi dobře při tom může posloužit bodový diagram tj. znázornění dvojic hodnot y_i, x_i. Zpravidla je výhodné zvolit několik typů funkcí a podle určitých kriterií posoudit vhodnost výběru funkce (CYHELSKÝ, 1981).

Jak jsme již uvedli v předcházející kapitole, jedním z kriterií vhodnosti volby typu regresní funkce je index determinace, popřípadě index korelace. U přímkové regrese použijeme koeficient determinace nebo korelace. (Koeficient determinace je vlastně zvláštním případem indexu determinace.) V praxi zpravidla vystačíme s některým z typů lineární regresní funkce nebo s nelineární regresní funkce transformované na lineární funkci v parametrech.

Jiným vhodným kriteriem pro určení typu regrese je posouzení sumy čtverců reziduálních odchylek Q_e, která má mít menší hodnoty.

6 Závěr

Dostali jsme se na samý konec těchto skript, která si vzala za úkol seznámit nás s některými vybranými statistickými kapitolami, metodami. Jednalo se však o pouhé naznačení a to tak, aby čtenář byl schopen do sebe vstřebat další informace a to z fundovanějších, především matematicky zaměřených, knih, článků a skript.

Jak jsme si již několikráte naznačila, statistika se za minimálně posledních sto let rozrostla do velké šíře a obohatila mnohé vědní obory včetně oblasti sociální analýzy. Z tohoto pohledu je poměrně důležité, abychom měli o statistice nikoli pouze povrchní znalosti, ale abychom ji dokázali správně používat a hlavně využívat.

Tato skripta snad trošku pomohla odkrýt některé taje statistiky a také navnadila dychtivé čtenáře k dalšímu studiu této stále se rozvíjející se vědy.

7 Literatura

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

LAMSER, V., RŮŽIČKA, L. *Základy statistiky pro sociology*. Praha: Nakladatelství Svoboda, 1970. ISBN 25-612-70.

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

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

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

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

8 Rejstřík

R

Ř

S

Š

T

U

V

Z

Řada: Výzkumné metody v sociální práci – sv. 3

Název: **Statistika pro sociální práci**
Rok a místo vydání: 2014, Hradec Králové
Vydání: první
Náklad: 200
Vydalo nakladatelství Gaudeamus při Univerzitě Hradec Králové jako svou 1344. publikaci.

ISBN 978-80-7435-404-5