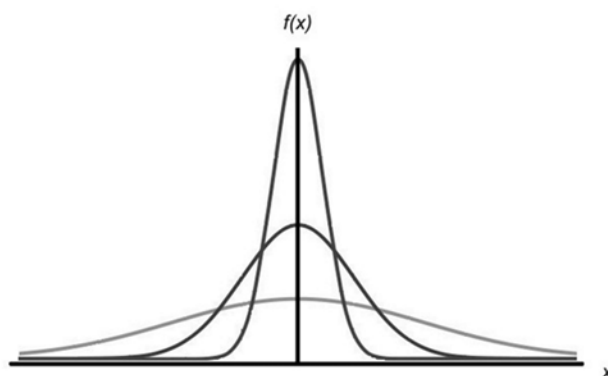


VYSOKOŠKOLSKÉ UČEBNÉ TEXTY
PREŠOVSKÁ UNIVERZITA
Fakulta humanitných a prírodných vied



Monika IVANOVÁ, Jaroslav HOFIERKA

ZÁKLADY ŠTATISTICKÝCH METÓD V GEOGRAFII



Prešov 2009

Autori: RNDr. Monika Ivanová, PhD., Doc. Mgr. Jaroslav Hofierka, PhD.

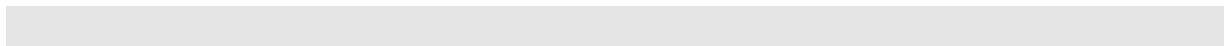
Recenzovali: Doc. PhDr. RNDr. Martin Boltížiar, PhD., Fakulta prírodných vied, Univerzita
Konštantína Filozofa v Nitre
Mgr. Mária Majherová, PhD., Fakulta humanitných a prírodných vied,
Prešovská univerzita v Prešove

Za odbornú a jazykovú stránku tohto vysokoškolského učebného textu zodpovedajú autori.

Vydala Fakulta humanitných a prírodných vied Prešovskej univerzity v Prešove ako
vysokoškolský učebný text pre študentov geografických študijných programov Prešovskej
univerzity v Prešove.

© RNDr. Monika Ivanová, PhD., Doc. Mgr. Jaroslav Hofierka, PhD., 2009

ISBN 978 – 80 – 555 - 0091-1



OBSAH

Predslov.....	4
1 Základné štatistické pojmy.....	5
2 Štatistický výskum	10
3 Štatistické rady a početnosti.....	16
4 Stredné hodnoty	29
5 Kvantily.....	51
6 Miery variability	55
7 Momenty polohy, miery šikmosti, špicatosti a koncentrácie, teoretické rozdelenia náhodných premenných.....	64
8 Časové rady a miery rastu.....	90
9 Výberové skúmanie a štatistické hypotézy.....	101
10 Závislosti	114
11 Regresná analýza.....	124
12 Analýza viacrozmerných štatistických údajov	129
Prílohy	136
Zoznam použitej literatúry.....	142
Zoznam grafov, tabuliek a obrázkov.....	144

Pri riešení geografických problémov je vždy nutné spracovať množstvo údajov, ktoré je potrebné správne analyzovať a vyhodnotiť. Rýchle tempo vývoja spoločnosti, jej informatizácia a dostupnosť veľkého množstva digitálnych dát si vyžadujú poznanie základných štatistických metód, ktoré umožňujú pochopiť stav analyzovaných javov, prebiehajúce súvislosti, vzťahy a zákonitosti. Štatistické metódy preto patria medzi základné metodické nástroje geografie a ich význam v geografii neustále nabera na význame. Zvyšuje sa množstvo a rozmanitosť dostupných štatistických údajov a meraní, ktoré je nutné spracovať, analyzovať a interpretovať. Metodický aparát modernej štatistiky poskytuje množstvo metód, ktoré umožňujú lepšie pochopiť zákonitosti fungovania javov v krajinskej sfére.

V predložených učebných textoch sa venujeme základným metódam, ktoré využívame v geografickej praxi. Pri skúmaní zákonitostí a väzieb v krajinskej sfére sa nevyhneme ani analýze viacrozmerných štatistických súborov. Z tohto dôvodu bola do učebných textov zaradená aj kapitola venujúca sa základným princípom analýzy viacrozmerných štatistických súborov. Metódy viacrozmernej analýzy nie je možné realizovať bez využitia špecializovaného štatistického softvéru. V súčasnosti k najznámejším softvérovým balíkom patria SPSS, Statistica, R a Statgraphics.

Učebné texty vznikli na Katedre geografie a regionálneho rozvoja FHPV PU v Prešove ako základný študijný materiál najmä pre študentov geografických študijných programov. Využiť ho však môžu aj študenti iných odborov, ktorých štúdium si vyžaduje prácu so štatistickými dátami. Štruktúra učebných textov je koncipovaná tak, aby poskytla základné informácie z oblasti opisnej štatistiky, ktoré prostredníctvom konkrétnych príkladov pomáhajú hlbšie vniknúť do skúmanej problematiky a vytvárajú tak priestor pre praktické uplatnenie získaných informácií.

Úvodné kapitoly oboznamujú študentov so základnými štatistickými pojmami, zaoberajú sa predmetom a významom štatistiky, ako aj etapami štatistického výskumu. Druhá časť učebných textov je venovaná frekvenčnej analýze, tvorbe štatistických výstupov, práci so štatistickými softvérmi, stredným hodnotám a vybraným mieram variability. Obsahom ďalších kapitol je analýza šikmosti, špicatosti, koncentrácie a teoretických rozdelení. Záverečné kapitoly poskytujú pohľad na časové rady, štatistické hypotézy, skúmajú závislosť medzi kvalitatívnymi a kvantitatívnymi štatistickými znakmi, charakterizujú priebeh závislosti. Riešenú problematiku uzatvára analýza viacrozmerných údajov, ktorá predstavuje prechod k uplatneniu pokročilejších štatistických metód. Učebné texty obsahujú viacero riešených príkladov, štatistických tabuliek, grafov a obrázkov, čím predstavujú relatívne samostatnú učebnú pomôcku, ktorá vhodne dopĺňa obsah prednášok a seminárov.

Tvorba učebných textov vyplynula v neposlednom rade i z potreby aplikácie rôznych štatistických metód tak v oblasti humánnej, ako aj fyzickej geografie. Pri ich tvorbe sme sa snažili v maximálnej miere použiť zrozumiteľnú a praktickú formu obohatenú o viacero príkladov najmä z oblasti geografie a jej príbuzných disciplín. K dosiahnutiu tohto cieľa prispeli aj recenzenti Doc. PhDr. RNDr. Martin Boltižiar, PhD. a Mgr. Mária Majherová, PhD., ktorým sa aj touto cestou chceme poďakovať za ich pripomienky a návrhy pri skvalitnení tohto diela.

1 Základné štatistické pojmy

Pôvod slova **štatistika** pochádza z latinského slova **status**, čo znamená štát, stav. V prvopočiatoch svojho vývoja plnila štatistika takmer výlučne popisnú funkciu a slúžila na zber dát, ktoré poskytovali informácie o stave štátu.

Prvé záznamy o zisťovaní počtu obyvateľov pochádzajú ešte z obdobia pred naším letopočtom, keď vznikali prvé historické zdroje štatistiky. Ich úlohou bolo úradné zisťovanie a evidencia. Dôkazom sú štatistické prieskumy v Babylónii, Egypte, Číne a Izraeli. Obdobné súpisy sa vykonávali i počas helénskej kultúry v starovekom Grécku a Ríme (Bačík 2007).

Stredoveká svetská moc nebola veľmi naklonená myšlienke štatistických prieskumov, a tak takmer došlo k úpadku úradného zisťovania. K jeho obnove došlo až v 16. -17. storočí, a to prostredníctvom nejednotných cirkevných matrik cez systematické matričné zápisy až po jednotnú evidenciu a výkazníctvo pre potreby štátu (Cyhelský 1990 in Bačík 2007)

Samotný pojem sa začal systematickejšie používať až koncom 18. storočia. Najviac sa o jeho rozšírenie zaslúžil nemecký profesor Gottfried Achenwall (1719-1772), ktorý ho používal pre označenie vedy o tzv. štátnych pozoruhodnostiach. Podobu dnešnej štatistiky ovplyvnila do značnej miery **teória pravdepodobnosti**. Jej základy položili významní matematici 17. - 19. storočia (Daniel a Nikolaus Bernoulli, Lagrange, Euler, Laplace, de Moivre, Gauss). Spočiatku sa teória pravdepodobnosti a matematická štatistika vyvíjali izolovane, no v neskoršom období došlo k ich prelínaniu, čo malo za následok rozvoj novej disciplíny – **matematickej štatistiky**. Pre štatistiku 19. a začiatku 20. storočia bolo príznačné úsilie získať čo najviac dát o skúmanom jave. Osobitný impulz pre rozvoj a využívanie štatistických metód v geografii priniesol nástup počítačových technológií v 50-tych a 60-tych rokoch 20. storočia, ktoré umožňovali rýchle spracovanie veľkého množstva údajov. V tejto súvislosti hovoríme aj o tzv. kvantitatívnej revolúcii, ktorá priniesla aj viacero nových geografických smerov bádania. Kvantitatívne metódy sa tak stali nevyhnutnou súčasťou metodického aparátu súčasnej geografie. Štatistické metódy sú však využívané aj v mnohých iných oblastiach (napr. ekonómia, zdravotníctvo, školstvo, atď.), jednoducho všade tam, kde vzniká množstvo dát, ktoré je potrebné spracovať a vyhodnotiť.

V súčasnosti sa pojem štatistika používa najčastejšie v nasledovných významoch:

- **praktická činnosť** – zameraná na získavanie, spracovanie a vyhodnocovanie štatistických dát, t.j. údajov o hromadných javoch,
- **štatistické údaje** – údaje o hromadných javoch bezprostredne získané pozorovaním alebo z nich vypočítané charakteristiky,
- **vedná disciplína** – zaoberajúca sa metódami skúmania a vyhodnocovania štatistických údajov, t.j. údajov o hromadných javoch.

Základné črty štatistiky ako vednej disciplíny:

- pracuje s **hromadnými javmi**,
- hľadá ich zákonitosti a k tomu využíva **hromadné pozorovanie**,
- zaoberá sa **kvantitatívnymi metódami hodnotenia** a k tomu používa matematický aparát.

Predmetom výskumu **štatistických metód** sú hromadné javy. **Hromadným javom** nazývame každý prírodný alebo spoločenský jav, ktorý skúmame na veľkom počte prípadov (hromadne). Ide o také javy (predmety, udalosti, procesy), ktoré sú výsledkom pôsobenia veľkého množstva príčin a ich vlastnosti sa neprejavujú v jednotlivých javoch, ale v ich súbore (Brázdil, et. al. 1995). Každý hromadný jav teda pozostáva z mnohých **individuálnych javov**.

Rôzne prírodné a spoločenské javy sa skúmajú ako hromadné javy preto, lebo takýto spôsob umožňuje poznať ich podstatu a vlastnosti, povahu činiteľov, ktoré ich podmieňujú a ovplyvňujú v ich vzájomných vzťahoch a súvislostiach. Bez hromadného pozorovania by sme nemohli o príslušnom jave robiť zovšeobecňujúce závery.

Hromadné pozorovanie sa teda zakladá na skúmaní dostatočne veľkého počtu prípadov. Rozlišujeme:

- **jednoduché pozorovanie** – do priebehu pozorovaných javov nijakým spôsobom nezasahujeme a neovplyvňujeme ich,
- **experiment** – vytvorí sa súbor kontrolovaných podmienok, v ktorých sa pozorovaný jav opakuje (Gregorová, Fillová 2004).

Štatistika pracuje predovšetkým s číselnými (kvantitatívnymi) hodnotami. Nakoľko i kvalitatívne hodnoty často transformuje do čísel a potom ich vyhodnocuje, teda aj metódy, ktoré pri tom používa, sa označujú ako kvantitatívne. Pojmom štatistické údaje možno preto označiť číselné (kvantitatívne) údaje o hromadných javoch.

Základným objektom pozorovania, na ktorom možno skúmať konkrétny prejav hromadného javu, je **štatistická jednotka**. Ide o takú jednotku (prvok, predmet, jav), ktorá je už pred začiatkom vlastného štatistického výskumu vecne, časovo a priestorovo vymedzená.

Vecné vymedzenie vo všeobecnosti určuje, čo pre daný prípad je alebo nie je štatistickou jednotkou. Je to najvýznamnejšie a občas aj najnáročnejšie vymedzenie. Niekedy je potrebné k definícii štatistickej jednotky pripojiť aj podrobné vysvetlenie ako postupovať v jednotlivých prípadoch (napr. študent 1. ročníka denného štúdia študijného programu geografia v regionálnom rozvoji na FHPV PU).

Časové vymedzenie určuje časové obdobie alebo časový moment, pre ktorý bola štatistická jednotka zaradená do štatistického výskumu (napr. sčítanie k 1.1. 1991).

Priestorové vymedzenie vymedzuje územie, na ktoré sa bude vzťahovať štatistický výskum (napr. štát, vyšší územný celok).

Štatistické jednotky rovnakého druhu vytvárajú štatistický súbor. **Štatistický súbor** je množina štatistických jednotiek (prvkov, osôb, vecí, javov), ktoré sú zhromaždené na základe istých spoločných vlastností a sú predmetom štatistického výskumu.

Z hľadiska obsahu (to, čo daného súboru patrí) možno rozlíšiť:

- **Základný štatistický súbor** - vyčerpávajúci (úplný) súbor štatistických jednotiek, ktoré svojím opisom vymedzujú daný štatistický súbor. Napríklad pri sčítaní ľudu je to množina všetkých obyvateľov štátu alebo všetkých domácností, pri skúmaní produkcie priemyselného odvetvia tvorí základný štatistický súbor množina všetkých podnikov konkrétneho rezortu.
- **Výberový štatistický súbor (výber)** – je podmnožinou základného súboru, ktorého štatistické jednotky boli vybraté podľa určitej pravdepodobnostnej schémy. Pristupujeme k nemu v prípade, že základný súbor je veľmi rozsiahly a skúmanie všetkých štatistických

jednotiek by mohlo byť príliš zdĺhavé, nákladné alebo niekedy aj nerealizovateľné. Výberový súbor musí byť **reprezentatívny** pre základný súbor, ktorého je podmnožinou. Reprezentatívnosť znamená, že určité štatistické vlastnosti (charakteristiky) sú takmer zhodné s vlastnosťami základného súboru (napr. stredné hodnoty, variabilita a pod.). Ich reprezentatívnosť závisí od rozsahu a postupu výberu, ako aj od rozsahu základného súboru.

Výber môže byť:

- **náhodný (pravdepodobnostný)** – používa sa pri homogénnych štatistických súboroch, kde každá štatistická jednotka má pravdepodobnosť dostať sa do výberového súboru a skutočnosť, či sa tam dostane, závisí na náhode (napr. žrebovanie pri športke),
- **zámerný (stratifikovaný)** – uplatniteľný pri veľmi heterogénnych vlastnostiach základného súboru, kde je potrebné poznať rozloženie a relatívnu početnosť štatistických jednotiek s danou vlastnosťou v základnom súbore,
- **samovýber** – štatistická jednotka sama rozhoduje, či bude zaradená do výberového skúmania (napr. osoby, ktoré sa prihlásia na výzvu).

Pod pojmom **štatistický znak (premenná, atribút)** chápeme vlastnosť štatistickej jednotky, ktorú skúmame.

Štatistické znaky možno rozdeliť do nasledovných skupín:

1. Podľa toho, či sa vyskytujú alebo nevyskytujú u všetkých jednotiek štatistického súboru:

- **spoločné** – majú ich všetky štatistické jednotky súboru (Ak základným súborom sú zapísaní študenti 1. ročníka GRR pre akademický rok 2007/2008, tak spoločným znakom je, že ide o študentov 1. ročníka daného študijného programu zapísaných v uvedenom akademickom roku),
- **variabilné** – sú predmetom štatistického skúmania, ich význam nie je dôležitý z hľadiska sledovaného cieľa (Napri. pri prijímacích pohovoroch na vybranú vysokú školu je jedno, akú strednú školu majú uchádzači, podstatné je, aby mali maturitu. Variabilným znakom je v tomto prípade druh absolvovanej strednej školy).

2. Podľa toho, či sú štatistické znaky priamym výrazom vlastnosti štatistickej jednotky alebo ju charakterizujú len sprostredkovane:

- **priame** – dajú sa spravidla priamo odmerať, resp. určiť (výška v cm, váha v kg...)
- **nepriame (odvodené)** – sú merateľné prostredníctvom nejakej ďalšej vlastnosti (napr. výbušná sila dolných končatín je sila, ktorá by mala byť meraná v N, ale merá sa v cm – ako ďaleko skočíš).

3. Podľa počtu variantov, ktoré štatistické znaky nadobúdajú¹:

- **alternatívne** – výber z dvoch 2 možností (pohlavie; zdravotný stav vo význame vyhovujúci- nevyhovujúci),
- **množné** – viac ako 2 možnosti (náboženské vierovyznanie, národnosť, druh absolvovanej strednej školy...).

¹ Delenie na alternatívne a množné znaky sa uplatňuje len pri kvalitatívnych znakoch.

Podľa toho, z akého hľadiska charakterizujú štatistickú jednotku:

- **časové** – charakterizujú štatistickú jednotku z hľadiska času (údaje k príslušnému roku, polroku, štvrťroku...),
- **priestorové** – charakterizujú štatistickú jednotku z hľadiska priestoru (trvalé bydlisko, región...),
- **vecné** – vyjadrujú vecnú stránku štatistickej jednotky (kúpna sila obyvateľstva...) .

U vecných štatistických znakov uvažujeme aj o tom, či ich možno vyjadriť číslom, teda znázorňujú kvantitu sledovaného znaku (**kvantitatívne štatistické znaky**), alebo ich vyjadrujeme slovné, teda máme na mysli kvalitu sledovaného javu (**kvalitatívne štatistické znaky**). V prípade kvantitatívnych štatistických znakov má význam uvažovať aj o ich spojitosti. **Nespojité (diskrétné) štatistické znaky** sú také, ktoré môžu nadobúdať celočíselné hodnoty (počet členov domácnosti), **spojité** akúkoľvek reálnu hodnotu (priemerná známka z maturity).

Riešené úlohy:

Štatistický výskum má priniesť odpoveď na otázku, aký je vzťah medzi počtom členov slovenských domácností a peňažnými príjmami na jednej strane a peňažnými výdavkami za potraviny na druhej strane. Čo bude štatistickým súborom a čo štatistickou jednotkou? Ktoré variabilné znaky sa budú skúmať? Klasifikujte tieto znaky.

Štatistický súbor: Slovenské domácnosti s trvalým pobytom na území Slovenskej republiky skúmané ku dňu sčítania obyvateľstva (k 1.1. 1991).

slovenské domácnosti – vecné vymedzenie

s trvalým pobytom na území Slovenskej republiky – priestorové vymedzenie

ku dňu sčítania obyvateľstva (k 1.1. 1991) – časové vymedzenie

Štatistická jednotka vyššie vyšpecifikovaného štatistického súboru: slovenská domácnosť

Štatistické znaky:

počet členov slovenských domácností – kvantitatívny, spojitý

peňažné príjmy za potraviny – kvantitatívny, spojitý

peňažné výdavky za potraviny – kvantitatívny, spojitý

Rozhodnite, ktoré z uvedených štatistických znakov sú kvantitatívne, resp. kvalitatívne. V prípade kvantitatívnych štatistických znakov uveďte, či ide o znaky spojité alebo nespojité.

a) priemerná výška v cm – kvantitatívny, spojitý,

b) počet áut – kvantitatívny, nespojitý,

c) farba vlasov – kvalitatívny,

d) náboženské vierovyznanie – kvalitatívny,

e) pohlavie – kvalitatívny,

Zistite, či ide o znaky alternatívne alebo množné:

- a) *druh absolvovanej strednej školy* – množný,
- b) *pohlavie*- alternatívny,
- c) *náboženské vierovyznanie* – množný.

Rozhodnite, či v daných prípadoch pôjde o základný alebo výberový štatistický súbor.

- a) *skúmame výskyt malárie na vzorke 1 000 respondentov (výberový – pracujem s určitou vzorkou respondentov),*
- b) *skúmame návštevnosť krúžku v 2. A triede istej základnej školy v Prešove za prítomnosti všetkých žiakov triedy (základný – predmetom záujmu sú všetci respondenti).*

Neriešené úlohy:

1. Štatistický výskum má priniesť odpoveď na otázku, aký je vzťah medzi priemerným prospechom vysokoškolských študentov v prvom semestri na vysokej škole na jednej strane, ich priemerným prospechom z maturity a druhom absolvovanej školy na druhej strane. Čo bude štatistickým súborom a čo štatistickou jednotkou? Ktoré variabilné znaky sa budú skúmať? Klasifikujte tieto znaky.

2. Rozhodnite, ktoré z uvedených štatistických znakov sú kvantitatívne, resp. kvalitatívne, alternatívne, resp. množné, spojité, resp. nespojité:

- a) *priemerná váha v kg,*
- b) *návštevnosť krúžku,*
- c) *druh ovocia,*
- d) *priemerný vek študenta,*
- e) *vlastníctvo auta,*
- f) *typ auta.*

3. Zistite, či ide o znaky alternatívne alebo množné:

- a) *rok výroby auta,*
- b) *zdravotný stav (vo význame uspokojivý - neuspokojivý),*
- c) *farba vlasov.*

4. Rozhodnite, či v daných prípadoch pôjde o základný alebo výberový štatistický súbor:

- a) *skúmame množstvo vynaloženej práce na vypracovanie určitého výrobku u pracovníkov bližšie špecifikovaného závodu,*
- b) *skúmame vzdelanostnú úroveň obyvateľov kraja SR podľa získaných údajov o 20 % respondentov v každom okrese uvedeného kraja.*

2 Štatistický výskum

Štatistický výskum pozostáva z troch základných etáp:

- **štatistické zisťovanie** - ide o proces získavania a zhromažďovania údajov o skúmaných javoch,
- **štatistické spracovanie** - usporiadanie štatistických jednotiek do skupín podľa triediacich znakov,
 - zameranie sa na opis štatistického súboru, testovanie, ako aj na výstup výsledkov po ich formálno-technickej a metodologickej stránke,
- **štatistický rozbor** – konečná fáza štatistického výskumu, ktorú tvorí analýza výsledkov štatistického spracovania a formulovanie záverov, ktoré z nich vyplývajú.

ŠTATISTICKÉ ZISŤOVANIE

Štatistické zisťovanie predstavuje úvodnú etapu štatistického výskumu. Jeho cieľom je získanie potrebných štatistických údajov potrebných pre analýzu hromadných javov.

Podľa spôsobu zisťovania údajov o hromadných javoch rozoznávame:

- Výkazníctvo** – pravidelné predkladanie vopred navrhnutých a schválených formulárov (výkazov).
- Súpis (cenzus)** – zisťuje javy, o ktorých sa nevedie bežná evidencia alebo nie je dostatočná. Viaz sa k rozhodujúcemu okamihu alebo je vykonaný za určité obdobie.
 - Makrocenzus** – súpis, pri ktorom sa zisťuje mnoho rôznych znakov o štatistických jednotkách na rozsiahlom území, spravidla na celom území štátu (napr. súpis obyvateľstva).
 - Mikrocenzus** - týka sa menších okruhov štatistických jednotiek, má užší obsah (napr. súpis odborníkov v podnikoch určitého odvetvia).
- Štatistický (znalecký) odhad** – spočíva v subjektívnom ohodnotení určitého javu expertom.
- Anketa** – ústne alebo písomné zisťovanie názoru na určitý jav u osôb alebo inštitúcií, ktoré sú spravidla náhodne vybraté. Aby anketa mala výpovednú hodnotu, je potrebné, aby priniesla odpovede od dostatočného počtu respondentov. Veľký význam má správna formulácia otázok. Napr. otázky by mali byť zrozumiteľné a jednoznačné, no zároveň by nemali navádzať na určitú odpoveď. Výsledky ankety však nemožno zovšeobecňovať na celý základný súbor, pretože súbor respondentov zvyčajne nie je dostatočne reprezentatívny. Špecializované agentúry vykonávajú tzv. prieskumy verejnej mienky, kde výber respondentov (okolo 1000 respondentov) zvyčajne zodpovedá vlastnostiam základného súboru.
- Monografia** – podrobný opis určitého javu z najrôznejších stránok odborníkmi, ktorí sú schopní zodpovedne zhrnúť svoje poznatky a názory o danom jave a všestranne ho posúdiť.

ŠTATISTICKÉ SPRACOVANIE

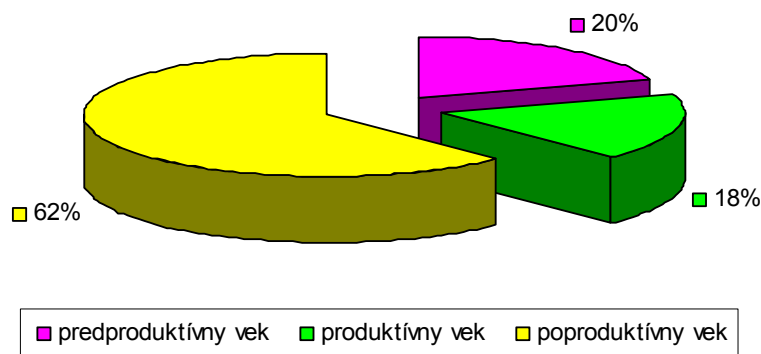
Druhy štatistického triedenia

Prvým krokom spracovania štatistických údajov je ich **triedenie**. Chápeme pod ním **usporiadanie** štatistického súboru (rozdelenie jeho jednotiek) do **skupín** (tried) podľa určitého štatistického znaku alebo znakov (**triediaceho atribútu**).

Ak sa pri triedení použije **len jeden triediaci znak**, ide o **jednostupňové triedenie** (graf 1).

Ak sa štatistický súbor triedi podľa **dvoch alebo viacerých znakov**, ide o **dvojstupňové**, resp. **viacstupňové triedenie** (tab.1)

Graf 1 Rozdelenie obyvateľov obce Vinné z hľadiska produktivity (rok 2001)



Spracované podľa: ŠÚ Prešov

Tab. 1 Rozdelenie obyvateľov obce Vinné podľa pohlavia a národnosti (rok 2001)

Národnosť	Muži	Ženy	Spolu
Slovenská	744	843	1587
Rusínska	4	1	5
Česká	2	1	3
Maďarská	2	0	2
Ukrajinská	0	1	1
Poľská	0	1	1
Ostatné, nezistené	3	6	9
Spolu	755	853	1608

Spracované podľa: ŠÚ Prešov

V širšom slova zmysle pod triedením chápeme **usporiadanie** štatistického súboru (rozdelenie jeho jednotiek) z **časového** (tab. 2), **priestorového** (tab. 3) alebo **vecného hľadiska** (tab. 4).

Tab. 2 Vývoj počtu obyvateľov obce Vinné od roku 1828 do roku 1948

Rok	1828	1869	1880	1890	1900	1910	1921	1930	1940	1948
Počet obyvateľov	1061	1524	1325	1305	1413	1472	1438	1428	1500	1407

Spracované podľa: ŠÚ Košice

Tab. 3 Miera zamestnanosti vo vybraných štátoch Európy (rok 1999)

Krajina	Miera zamestnanosti v percentách
Slovensko	58,1
Belgicko	59,3
Francúzsko	60,9
Holandsko	71,7
Taliansko	52,7

Spracované podľa: Štatistická ročenka SR 2005

Tab. 4 Rozvody podľa veku manželov v roku 2004

Vek muža	Počet rozvodov
0-19	8
20-24	297
25-29	1589
30-34	2257
35-39	1980
40-44	1780
45-49	1435
50-54	914
55-59	382
60 a viac	247
Spolu	10889

Spracované podľa: Štatistická ročenka SR 2005

Každé štatistické triedenie sa musí vykonávať podľa určitých pravidiel:

→ **Pravidlo úplnosti** – skupiny vytvárame tak, aby každá štatistická jednotka zo súboru bola zaradená do niektorej z nich. Ak určitý prvok nemôžeme zaradiť do nijakej z vytvorených skupín, vytvoríme ďalšiu skupinu „Ostatné“, „Iné“ a pod.

→ **Pravidlo jednoznačnosti triedenia** - skupiny sa nesmú navzájom prekrývať (jeden prvok môže patriť len do jednej skupiny). Častou chybou je vytvorenie intervalových skupín, ktorých hranice tvorí to isté číslo. Príklad nesprávne vytvorených intervalov: do

15, 15-60, 60 a viac, príklad správne vytvorených intervalov: do 15, 16-60, 61 a viac. Pri stanovení hraníc intervalov je potrebné vychádzať aj z hodnôt štatistických jednotiek (celé alebo reálne čísla).

Pri triedení podľa **kvalitatívnych štatistických znakov** sa štatistický súbor rozdeľuje podľa variantov slovného znaku, pričom samotný znak, ako aj jeho obmeny musia byť vyjadrené slovne. V štatistike využívame dva základné druhy triedenia podľa slovných znakov:

→ dichotomické (binárne) - štatistický súbor sa rozdelí na dve triedy, pretože triediaci kvalitatívny znak má iba dva varianty (napr. podľa pohlavia, návštevnosť krúžku vo význame navštevujem – nenavštevujem),

→ polytomické (množné) - štatistický súbor sa rozdelí na viac tried, pretože triediaci znak má viac ako dva varianty (napr. náboženské vierovyznanie, druh absolvovanej strednej školy a pod.) .

Ak delíme štatistický súbor podľa **kvantitatívnych štatistických znakov**, celkový počet štatistických jednotiek v súbore sa rozdelí do jednotlivých skupín nasledovne:

→ individuálne triedenie - každej hodnote atribútu je priradená jedna trieda (tab. 5),

→ intervalové (skupinovú) triedenie - číselné hodnoty sú rozdelené do rovnakých alebo nerovnakých intervalov (tab. 6).

Počet štatistických jednotiek, ktoré patria do určitej triedy, nazývame triednou početnosťou.

Označuje sa symbolom n_i pre $i=1,2,...,k$. Platí: $\sum_{i=1}^k n_i = n_1 + n_2 + ... + n_k$. Pre stanovenie počtu

tried a triednych intervalov neexistuje žiadne všeobecne platné pravidlo. Niektorí autori odporúčajú stanoviť počet intervalov od 5 do 20, iní vychádzajú z rozsahu súboru a počítajú počet tried podľa nasledovných vzťahov:

$k \leq 5 \log n$, $k \approx \sqrt{n}$, $k \approx 1 + 3,3 \log n$, kde k je počet tried a n počet štatistických jednotiek.

Rozpätie (šírka) intervalu (h) závisí od variačného rozpätia, teda rozdielu medzi najväčšou a najmenšou hodnotou znaku a od počtu tried. Jeho približnú hodnotu možno určiť podľa

vzt'ahu: $h = \frac{x_{\max} - x_{\min}}{k}$.

Tab. 5 Výsledky písomnej previerky zo slovenského jazyka

Známka	Výskyt známky v skupine (n_i)
1	7
2	10
3	6
4	2
5	1

Spracované podľa: ilustratívne údaje autorov

Tab. 6 Rozdelenie študentov 1. ročníka externého štúdia odboru GRR podľa ich veku

<i>Vek študenta</i>	<i>Počet študentov(n_i)</i>
<18-21)	10
<21-24)	15
<25-28)	10
<28-31)	14
<31-34)	3

Spracované podľa: ilustratívne údaje autorov

Techniky štatistického spracovania údajov

Základnými technikami štatistického spracovania sú ručné a automatické spracovanie štatistických údajov.

- **Ručné spracovanie** sa vykonáva pomocou bežnej kancelárskej techniky a za bezprostrednej osobnej účasti spracovateľa. Patrí tu:

- **Čiarková metóda** – najjednoduchšia metóda ručného spracovania. Čiarky sa zapisujú tak, že každú piatu jednotku v skupine označíme preškrtnutím predchádzajúcich štyroch (|||| – |||| – |||| ||). Zápis pri viac ako 4-6 čiarkach je neprehľadný (||||| ||||| |||||).

- **Metóda ručných štítkov** – veľmi pracná, dnes už neefektívna metóda (napr. evidencia údajov o zamestnancoch). Pre každú štatistickú jednotku sa vytvorí samostatný štítok (kartička), na okraj ktorého sa postupne uvedú údaje o jednotlivých štatistických znakoch. Podľa nich sa rozčleňujú štítky na potrebné skupiny a spočítavajú sa.

- **Metóda sústredovacích (koncentračných) tabuliek** – štatistický súbor sa rozdelil na niekoľko skupín podľa určitého znaku či znakov a pre každú takto definovanú skupinu sa vytvorila samostatná tabuľka, do ktorej sa zapísali ďalšie údaje o štatistických jednotkách patriacich do danej skupiny.

- **Automatizované spracovanie** je realizované prostredníctvom výpočtovej techniky. K najpoužívanejším počítačovým programom využívaných pre spracovanie štatistických údajov radíme: MS Excel (z balíka MS Office), Statistica, StatGraphics, R, SPSS, NCSS a pod.

Obr. 1 Databázová tabuľka štatistického softvéru SPSS

Cars [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : mpg 18

	engine	horse	weight	accel	year	origin	cylinder
1	307	130	Vehicle Weight (lbs.)	12	70	American	8 Cylinders
2	350	165	3693	12	70	American	8 Cylinders
3	318	150	3436	11	70	American	8 Cylinders
4	304	150	3433	12	70	American	8 Cylinders
5	302	140	3449	11	70	American	8 Cylinders
6	429	198	4341	10	70	American	8 Cylinders
7	454	220	4354	9	70	American	8 Cylinders
8	440	215	4312	9	70	American	8 Cylinders
9	455	225	4425	10	70	American	8 Cylinders
10	390	190	3850	9	70	American	8 Cylinders
11	133	115	3090	18	70	European	4 Cylinders
12	350	165	4142	12	70	American	8 Cylinders
13	351	153	4034	11	70	American	8 Cylinders
14	383	175	4166	11	70	American	8 Cylinders
15	360	175	3850	11	70	American	8 Cylinders
16	383	170	3563	10	70	American	8 Cylinders
17	340	160	3609	8	70	American	8 Cylinders
18	302	140	3353	8	70	American	8 Cylinders
19	400	150	3761	10	70	American	8 Cylinders

Data View Variable View

SPSS Processor is ready

3 Štatistické rady a početnosti

Výsledkom základného spracovania štatistických súborov vznikajú tzv. štatistické rady. **Štatistický rad** je množina hodnôt **jediného atribútu** všetkých štatistických jednotiek daného súboru. **Na základe všeobecného usporiadania štatistického súboru sa rady delia na:**

- vecné - vznikajú na základe vecného atribútu,
- časové - štatistické údaje majú časovú postupnosť (napr. počet obyvateľov SR za jednotlivé roky),
- geografické (priestorové) rady - štatistické údaje sú zhromaždené podľa územných jednotiek, regiónov (napr. počet obyvateľov za jednotlivé štáty Európskej únie).

Empirický rad – štatistický rad získaný štatistickým zisťovaním. Hodnoty sú uvedené tak, ako boli zaznamenané.

Príklad: 8,2,7,1,15,4.

Usporiadaný (variačný) rad – rad usporiadaných číselných hodnôt (vzostupné - od minimálnej po maximálnu hodnotu, alebo zostupné - od maximálnej po minimálnu hodnotu).

Príklad: 1,2,4,7,8,15.

POČETNOST'

V prípade štatistických radov, ktorých hodnoty boli rozdelené do tried, hovoríme aj o sekundárnych radoch, resp. radoch odvodených triedením. Počet jednotiek v jednotlivých triedach (intervaloch) vyjadruje **početnosť**. Počet jednotiek v jednotlivých triedach (rozdelenie početností) podáva dôležitú informáciu o **vnútornej štruktúre štatistického súboru**. Proces zisťovania a charakterizovania rozloženia početností sa nazýva **frekvenčnou analýzou** a je základnou metódou kvantitatívnej analýzy v geografii.

Rady rozdelenia početností (rozdelenie početností) – sú rady, v ktorých je každej hodnote atribútu priradená početnosť vyjadrujúca počet jednotiek, ktoré nadobúdajú túto hodnotu.

Rady skupinového rozdelenia početností – sú rady, v ktorých sú hodnoty atribútu rozdelené do intervalov. Pre každý interval sa určí kvantitatívna hodnota (početnosť), ktorá vyjadruje koľko jednotiek súboru nadobúda hodnoty z daného intervalu.

Početnosť:

Absolútna početnosť (n_i) – počet štatistických jednotiek súboru, ktoré nadobúdajú danú hodnotu.

Relatívna početnosť (f_i) – podiel jednotiek zo všetkých jednotiek, ktoré nadobúdajú hodnotu daného atribútu (znaku) x_i .

$$f_i = \frac{n_i}{n} \text{ pre } i=1, 2, \dots, k$$

Kumulatívna početnosť – získa sa sčítaním (kumulovaním) všetkých početností pre danú hodnotu atribútu (intervalu). Udáva, koľko prvkov štatistického súboru má hodnotu znaku

nižšiu alebo rovnú danej hodnote. Pri zisťovaní kumulatívnej početnosti musia mať jednotlivé rady rozdelenia početností charakter **variačného (usporiadaného) radu**.

Príklad: Hráte hru a v každom kole získavate body. Ak v 1. kole získate 3 body, v 2. kole 8 bodov, v 3. kole 5 bodov a v 4. kole 10 bodov, zápis je kumulatívny - vždy ide o sumárny (kumulovaný) zisk, nie o zisk v určitom kole.

Absolútna kumulatívna početnosť (N_i) – súčet absolútnych početností všetkých hodnôt znaku. Súčet všetkých absolútnych početností sa rovná **rozsahu súboru** (n) – počtu všetkých štatistických jednotiek v súbore.

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

Relatívna kumulatívna početnosť (F_i) – súčet relatívnych početností všetkých hodnôt znaku. Súčet všetkých relatívnych početností sa rovná 1, resp. 100% .

$$F_i = \sum_{i=1}^k f_i \qquad f_i = \frac{n_i}{n} \text{ pre } i=1, 2, \dots, k.$$

VYJADRENIE RADOV ROZDELENIA POČETNOSTÍ

Aby sme mohli posúdiť rozdelenie početností v jednotlivých triedach, je vhodné prehľadne vyjadriť početnosť graficky alebo pomocou tabuľky. Tabuľku, ktorá vyjadruje rozloženie početností, nazývame tiež **frekvenčnou tabuľkou** (tab. 7).

Príklad: Skúmame ročný úhrn zrážok v ombrometrických staniciach okresu Prešov v roku 2001.

Tab. 7 Frekvenčná tabuľka meraní ročných úhrnov zrážok

i	x_i	n_i	f_i (%)	N_i	F_i (%)
1	220	2	12	2	12
2	250	1	9	3	21
3	270	4	26	7	47
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

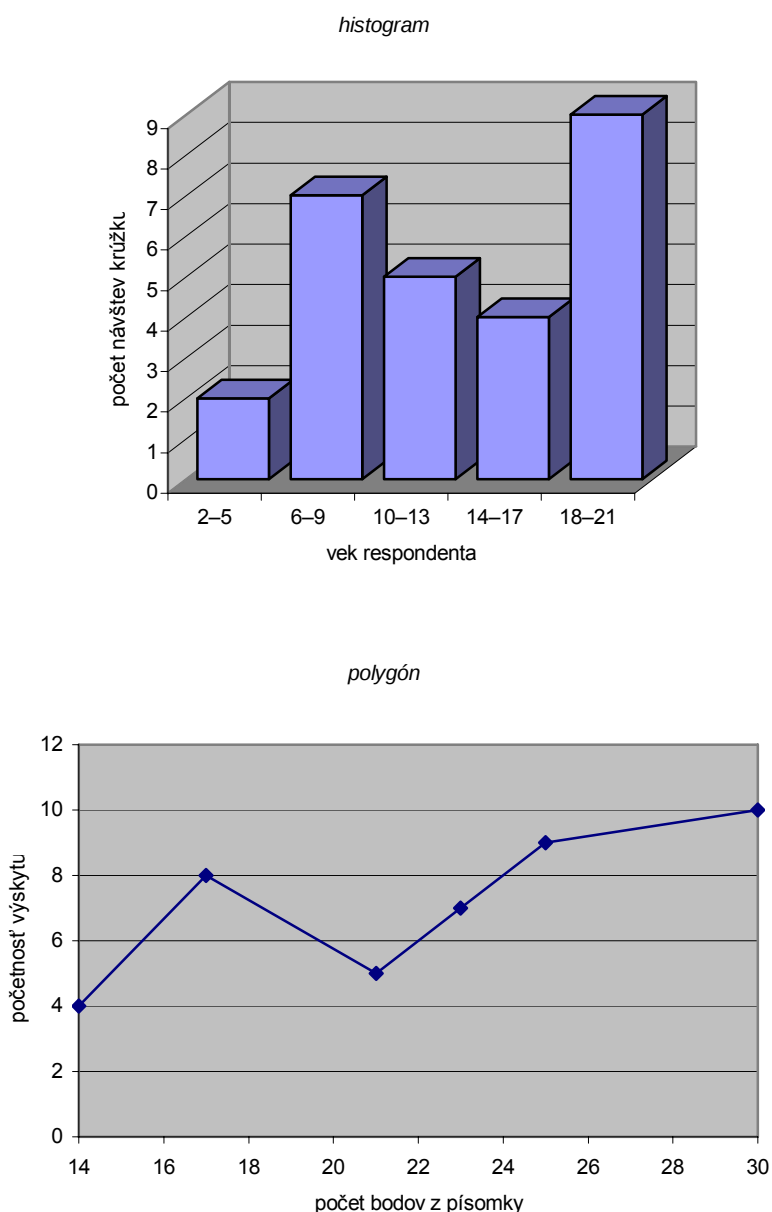
Grafické vyjadrenie rozdelenia početností je zvyčajne názornejšie. Pri **grafe rozdelenia početností** sa využíva kartézská súradnicová sústava. Na os **x** sa nanášajú **hodnoty znakov** jednotlivých prvkov (x_i), resp. ich tried a na os **y** **početností** (n_i, f_i, N_i, F_i).

Veľmi časté grafické vyjadrenie rozdelenia početností je v podobe **histogramu**, čo je stĺpcový graf. Vhodný je na vyjadrenie intervalového rozdelenia početností (graf 2). Je zložený zo stĺpcov, ktoré reprezentujú jednotlivé intervaly. **Intervalové hodnoty** sa zvyčajne nanášajú na os **x**. Výška stĺpca zodpovedá hodnote početnosti intervalu (početnosť môže byť absolútna alebo relatívna). **Početnosť** sa označuje na os **y**. Histogramové vyjadrenie je veľmi časté pri

vekových pyramídach (graf 3), kde sa však pre názornosť tvaru vekovej pyramídy hodnoty znaku (vekové triedy) nanášajú na os y a početnosti na os x.

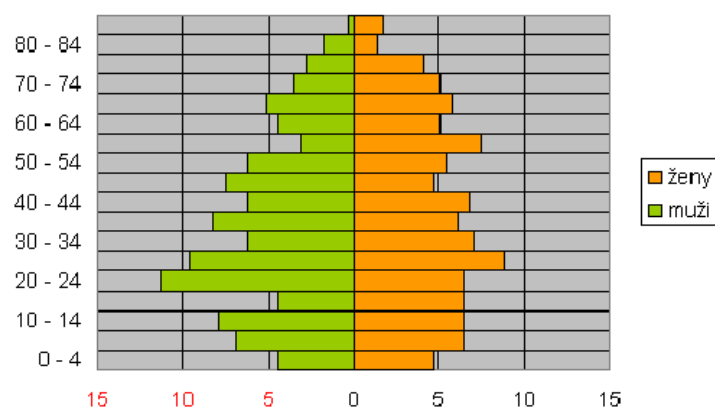
Ďalšou formou grafického vyjadrenia rozdelenia početností je **polygón**, t.j. spojnicový líniový (čiarový) graf, ktorý sa vytvára spojením bodov (graf 2). Na os x sa nanáša **hodnota znaku, resp. stred intervalu** a na os y sa nanáša **hodnota početnosti** (absolútna alebo relatívna). Na vyjadrenie kumulatívnych početností sa používa tzv. **ogivná krivka (súčtová krivka, S-krivka, ogiva (Galtonova ogiva – Galton zaviedol pojem ogiva))** - neklesajúca krivka, ktorá vznikne pospájaním bodov, súradnice ktorých predstavuje **hodnota znaku, resp. stred intervalu** na osi x a **kumulatívna početnosť** na osi y.

Graf 2 Histogram a polygón



Spracované podľa: ilustratívne údaje autorov

Graf 3 Veková pyramída



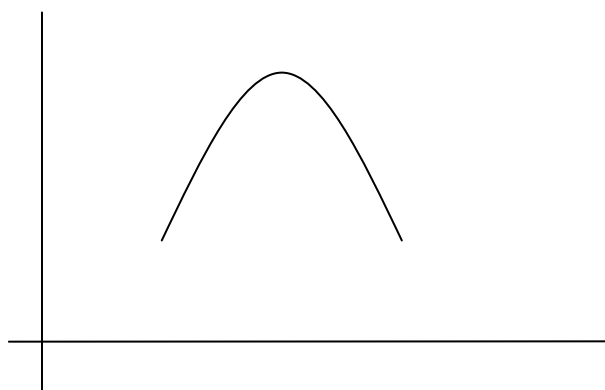
Spracované podľa:

http://www.rajcany.sk/geograficke_udaje_o_obci/geograficke_udaje_o_obci_soubory/image010.gif

Typy rozdelenia početností

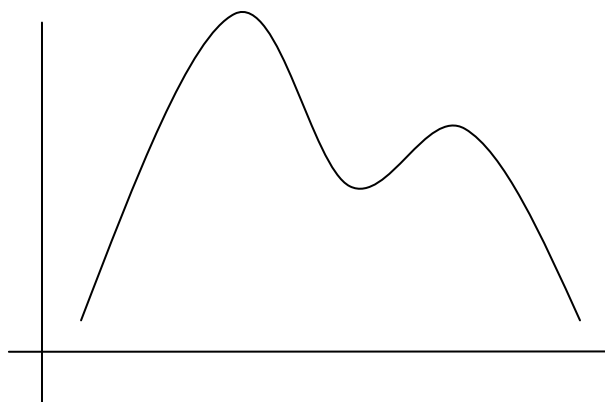
Početnosti môžu byť rozdelené rôzne, preto aj grafy rozdelenia početností majú rôzne tvary. Z tvaru rozdelenia početností môžeme usudzovať určité vlastnosti skúmaného javu alebo očakávaný vývoj v prípade časových radov. Napríklad z tvaru vekovej pyramídy vieme odhadnúť očakávaný vývoj počtu obyvateľov v určitých vekových kategóriách. Z hľadiska tvaru nás zaujíma najmä sústredenie početností do určitej triedy a v jej okolí (tvorba vrcholov, lokálnych extrémov v grafe) a potom symetria rozdelenia. Na základe toho rozoznávame niekoľko charakteristických typov rozdelení početností:

Jednovrcholové (unimodálne) rozdelenia - početnosti sa sústreďujú do jedného lokálneho maxima (vrcholu).

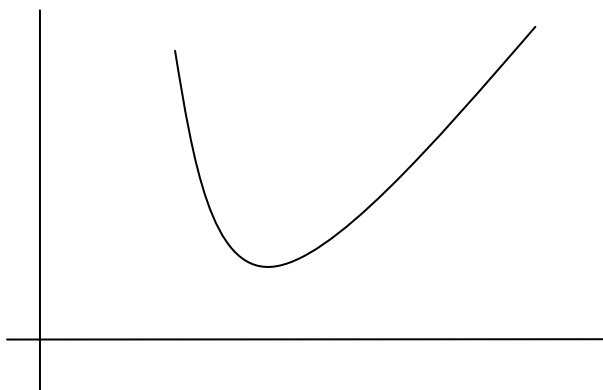


Viacvrcholové (multimodálne) rozdelenia - početnosti sa sústreďujú vo viacerých častiach rozdelenia. Ak má rozdelenie početnosti viacero vrcholov, svedčí to o určitej heterogenite štatistického súboru. V takomto prípade je potrebné preskúmať príčiny tejto heterogenity a zvážiť, či by nebolo vhodnejšie rozdeliť štatistický súbor na viacero homogénnejších častí a výskum ďalej realizovať v rámci nich.

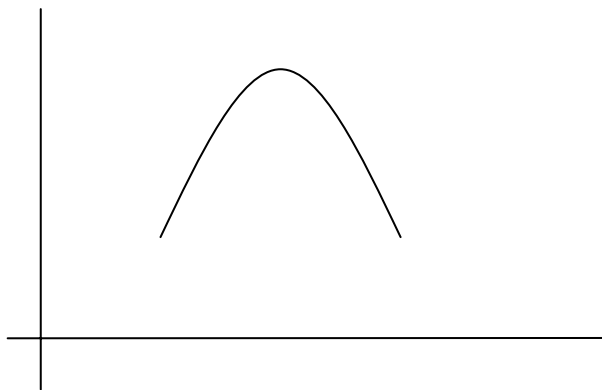
Dvojvrcholové (bimodálne) rozdelenia – početnosti sa sústreďujú v dvoch lokálnych extrémoch.



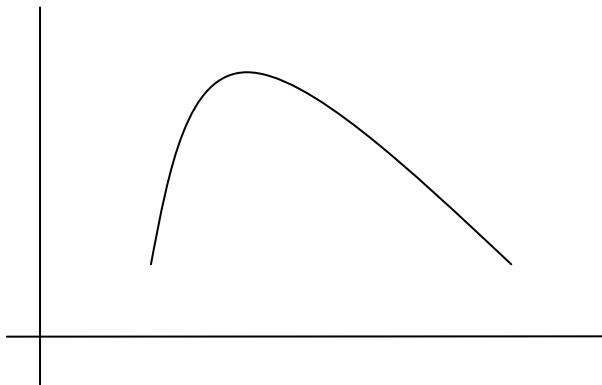
U – rozdelenia (antimodálne) – protiklad k unimodálnemu rozdeleniu, početnosti sú koncentrované v najvyšších a najnižších triedach a smerom k stredu rozdelenia sa znižujú.



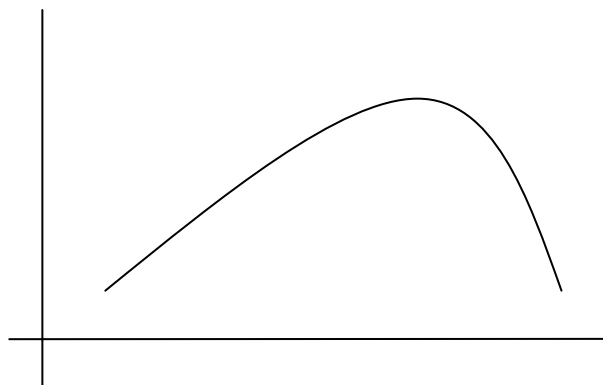
Symetrické (súmerné) rozdelenie – početnosti sa od stredu rozdelenia s najvyššou početnosťou sa rovnakým spôsobom znižujú a zväčšujú smerom k nižším a vyšším hodnotám štatistického znaku. Dokonalé symetrické rozdelenia v praxi zväčša neexistujú, častejšie sú **približne symetrické rozdelenia**.



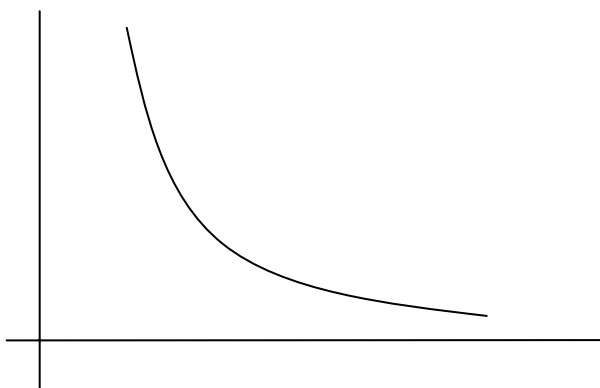
Asymetrické (nesúmerné) rozdelenia. Ak je graf početností vzhľadom k svojmu vrcholu zošikmený doľava, ide o **ľavostranne asymetrické rozdelenie**, resp. rozdelenie s **kladnou šikmost'ou (asymetriou) - pozitívne rozdelenie**. Početnosti sa sústreďujú viac pri nižších hodnotách znaku a vyššie hodnoty sú menej početné:



Ak je graf početností vzhľadom k svojmu vrcholu zošikmený doprava, ide o **pravostranne asymetrické rozdelenie**, resp. rozdelenie so **zápornou šikmost'ou (asymetriou) - negatívne rozdelenie**. Početnosti sa viac sústreďujú pri vyšších hodnotách znaku.



Extrémne zošikmené (asymetrické) jednovrcholové (unimodálne) rozdelenie - najvyššia, resp. najnižšia hodnota znaku má výrazne najväčšiu početnosť a početnosti ostatných hodnôt znaku sústavne klesajú.



Rektangulárne (konštantné) rozdelenie - početnosti sú rovnomerne rozložené vo všetkých triedach.



Skúmaním veľkého množstva empirických radov sa zistilo, že niektoré javy majú charakteristický tvar rozdelenia početností. Napríklad mnohé prírodné javy majú tzv. normálne rozdelenie početností, ktoré je jednovrcholové a symetrické. Poznanie charakteristických tvarov rozdelenia početností umožňuje empirický rad porovnať s tzv. teoretickým rozdelením, ktoré je definované matematickou funkciou. Príklady **teoretických rozdelení početností**: binomické, hypergeometrické, normálne (Gaussovo), Poissonovo, exponenciálne, Studentovo, χ^2 (chi-kvadrát), Fisherovo-Snedecrovo F-rozdelenie (pozri kapitolu 7).

Riešené úlohy:

1. Daný je nasledovný empirický rad 200 hodnôt.

12 15 18 21 24 27 30 33 36 39 42 45 49 51 54 57 60 63 66 52
 13 13 14 12 22 22 28 28 29 35 35 40 41 41 42 42 60 63 65 48
 25 28 32 33 46 49 58 61 65 34 37 58 45 45 58 56 62 67 60 32
 25 28 32 33 46 49 58 61 65 34 37 58 45 45 56 56 62 67 64 15
 66 45 60 63 66 12 13 18 21 24 48 51 15 18 21 39 42 45 21 24
 65 40 60 63 65 13 13 14 12 22 41 41 13 14 12 35 35 40 12 22
 60 53 52 67 60 25 28 32 33 46 45 45 26 32 33 34 37 58 33 46
 64 58 62 67 64 25 25 32 33 46 45 45 28 32 33 34 37 58 33 46
 5 6 67 41 44 47 49 51 32 35 66 37 66 38 61 38 59 41 27 66
 25 28 32 33 46 49 56 61 65 34 22 22 28 28 29 35 35 40 41 41

Minimum je 5, maximum 67. Atribút (výsledok testu z prijímacích pohovorov) priradený študentovi mohol nadobúdať hodnotu od 0 do 70. Štatistickou jednotkou je študent, ktorý sa zúčastnil na prijímacích skúškach na škole X v meste N.

Empirický rad vytried'te na:

- rad početností a vytvorte kompletnú frekvenčnú tabuľku,
- na rady skupinového rozdelenia početností s rozsahom intervalu 15 a s optimálnym počtom intervalov a vytvorte kompletnú frekvenčnú tabuľku,
- pre všetky prípady skupinového rozdelenia početností zostrojte grafy rozdelenia početností a z grafov určite, k akému typu rozdelenia početností patria.

Riešenie:

- Rad početností doplnený na frekvenčnú tabuľku (pozri str. 24).

Hodnoty N_i, f_i, F_i získame pomocou nasledovných vzorcov pre $i=1,2,\dots,k$:

$$N_i = \sum_{j=1}^i n_j, f_i = \frac{n_i}{n}, F_i = \sum_{j=1}^i f_j.$$

- Rady skupinového rozdelenia početností doplnené na frekvenčnú tabuľku:

- šírka intervalu $h=15$.

Hodnotu x_i pre stanovenie pevnej šírky intervalu vypočítame tak, že k x_{min} hodnote pripočítame $h-1$ štatistických jednotiek súboru, čím dostaneme maximálnu hodnotu rozpätia prvého intervalu. Pri stanovení druhého intervalu v poradí začíname najbližším číslom z oboru celých čísel vyskytujúcom sa po maximálnej hodnote z prvého intervalu, pričom opäť k nej pripočítame hodnotu $h-1$ a postup opakujeme. Absolútne početnosti n_i pre jednotlivé intervaly si následne určíme z tabuľky a). Rad skupinového rozdelenia početností doplníme na kompletnú frekvenčnú tabuľku obdobne ako v úlohe a).

x_i	n_i	N_i	f_i	F_i
5-19	23	23	11,5	11,5
20-34	58	81	29	40,5
35-49	60	141	30	70,5
50-64	42	183	21	91,5
65-79	17	200	8,5	100

x_i	n_i	N_i	f_i	F_i
5	1	1	0,5	0,5
6	1	2	0,5	1
12	6	8	3	4
13	6	14	3	7
14	3	17	1,5	8,5
15	3	20	1,5	10
18	3	23	1,5	11,5
21	4	27	2	13,5
22	6	33	3	16,5
24	3	36	1,5	18
25	5	41	2,5	20,5
26	1	42	0,5	21
27	2	44	1	22
28	10	54	5	27
29	2	56	1	28
30	1	57	0,5	28,5
32	9	66	4,5	33
33	10	76	5	38
34	5	81	2,5	40,5
35	7	88	3,5	44
36	1	89	0,5	44,5
37	5	94	2,5	47
38	1	95	0,5	47,5
39	3	98	1,5	49
40	4	102	2	51
41	8	110	4	55
42	4	114	2	57
44	1	115	0,5	57,5
45	11	126	5,5	63
46	7	133	3,5	66,5
47	1	134	0,5	67
48	2	136	1	68
49	5	141	2,5	70,5
51	3	144	1,5	72
52	1	145	0,5	72,5
54	1	146	0,5	73
56	3	149	1,5	74,5
57	1	150	0,5	75
58	10	160	5	80
59	1	161	0,5	80,5
60	7	168	3,5	84
61	4	172	2	86
62	4	176	2	88
63	4	180	2	90
64	3	183	1,5	91,5
65	7	190	3,5	95
66	5	195	2,5	97,5
67	5	200	2,5	100

- optimálny počet intervalov

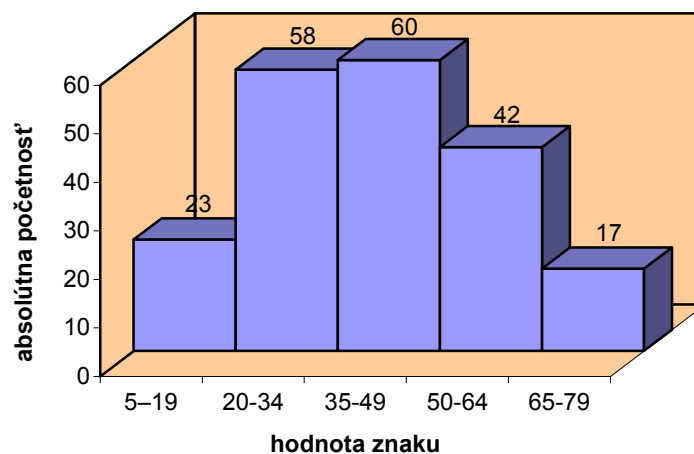
Stanovenie optimálnej šírky intervalu h vychádza z poznania k - počtu tried štatistického súboru a variačného rozpätia (rozdielu medzi najväčšou a najmenšou hodnotou štatistického znaku v súbore). Počet tried si vypočítame podľa vzorca $k \approx 1 + 3,3 \log n = 1 + 3,3 \log 200 \approx 9$, kde n predstavuje počet štatistických jednotiek súboru (v našom prípade 200). Následne určíme šírku intervalu:

$$h = \frac{x_{\max} - x_{\min}}{k} = \frac{67 - 5}{9} \approx 7.$$

x_i	n_i	N_i	f_i	F_i
5–11	2	2	1	1
12–18	21	23	10,5	11,5
19–25	18	41	9	20,5
26–32	25	66	12,5	33
33–39	32	98	16	49
40–46	35	133	17,5	66,5
47–53	12	145	6	72,5
54–60	23	168	11,5	84
61–67	32	200	16	100

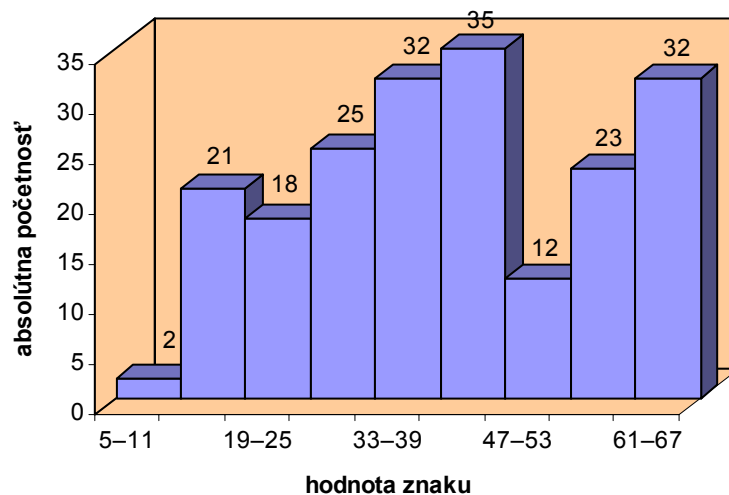
c) Grafy rozdelenia početností:

- pre šírku intervalu 15



Odpoveď: Unimodálne asymetrické rozdelenie početností.

- pre optimálny počet intervalov



Odpoveď: Unimodálne asymetrické rozdelenie početností.

2. V tabuľke sú uvedené informácie o priemernom úhrne zrážok (v mm) podľa zrážkomernej stanice Jovsa za obdobie 1931-60. Z uvedených hodnôt vytvorte rad početností a ten doplňte na kompletnú frekvenčnú tabuľku.

Priemerný úhrn zrážok (v mm) za obdobie 1931-60

Obdobie	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.
Jovsa	60	53	41	45	66	89	85	85	55	66	72	71

Spracované podľa: SHMÚ Košice

Riešenie:

a) Rad početností doplnený na frekvenčnú tabuľku:

Rad početností získame využitím čiarovej metódy. Následne ho doplníme na kompletnú frekvenčnú tabuľku, kde hodnoty N_i , f_i a F_i vypočítame pomocou nasledovných vzorcov pre $i=1, 2, \dots, k$:

$$N_i = \sum_{i=1}^k n_i, f_i = \frac{n_i}{n}, F_i = \sum_{i=1}^k f_i.$$

x_i	n_i	N_i	f_i^*	F_i^*	f_i	F_i
41	1	1	8,3	8,3	8,325	8,325
45	1	2	8,3	16,6	8,325	16,65
53	1	3	8,3	24,9	8,325	24,975
55	1	4	8,3	33,2	8,325	33,3
60	1	5	8,3	41,5	8,325	41,625
66	2	7	16,7	58,2	16,7	58,325
71	1	8	8,3	66,5	8,325	66,65
72	1	9	8,3	74,8	8,325	74,975
85	2	11	16,7	91,5	16,7	91,675
89	1	12	8,3	99,8	8,325	100

Poznámka: Hodnoty f_i^* sme boli nútení upraviť, nakoľko posledná hodnota v stĺpci F_i^* bola menšia ako 100 %. Pri úprave hodnôt f_i^* sme počítali veľkosť absolútnej chyby. V uvedenom prípade sme mali dve možnosti: zaokrúhliť smerom nahor čísla 16,7 o 1 desatinné miesto (absolútna chyba mala veľkosť: $16,8-16,7=0,1$ alebo zaokrúhliť smerom nahor číslo 8,3. Nakoľko máme až 8 čísel s tou istou číselnou hodnotou a všetky čísla je nutné zaokrúhliť o rovnakú veľkosť absolútnej chyby tak, aby sme navýšili hodnotu posledného čísla v stĺpci F_i^* o 2 desatinné miesta, počítali sme podiel $0,2/8=0,025$. Teda každé číslo s hodnotou 8,3 by sme mali zvýšiť na 8,325. Veľkosť absolútnej chyby by v tomto prípade bola: $8,325-8,3=0,025$, čo je menej ako veľkosť absolútnej chyby u čísla 16,7. Z tohoto dôvodu sme navyšovali čísla 8,3. Upravené hodnoty po navýšení sú uvedené v stĺpci f_i , z ktorého sme následne vypočítali hodnoty F_i .

Neriešené úlohy:

1. V tabuľke sú uvedené informácie o priemernom úhrne zrážok (v mm) podľa zrážkomernej stanice Vinné za obdobie 1931-60.

Priemerný úhrn zrážok (v mm) za obdobie 1931-60

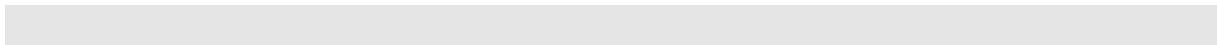
Obdobie	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.
Vinné	45	44	34	42	59	81	84	73	47	56	62	57

Spracované podľa: SHMÚ Košice

Z uvedených hodnôt vytvorte:

- a) rad početností a ten doplňte na kompletnú frekvenčnú tabuľku,
- b) rady skupinového rozdelenia početností s rozsahom intervalu 15 a s optimálnym počtom intervalov a vytvorte kompletnú frekvenčnú tabuľku,

- c) *pre skupinové rozdelenie početností zostrojte graf rozdelenia početností a z grafu určite, k akému typu rozdelenia početností patrí.*



4 Stredné hodnoty

Stredné hodnoty sú súhrnné, zovšeobecňujúce, reprezentatívne číselné hodnoty, ktoré vyjadrujú úroveň hodnôt znaku štatistického súboru. Môžu to byť hodnoty, ktoré sú pre štatistický súbor typické, očakávané, najčastejšie alebo hodnoty, okolo ktorých sú rozložené všetky ostatné hodnoty.

Umožňujú jednoduché a prehľadné porovnávanie skúmaného javu v dvoch aj viacerých štatistických súboroch na všeobecnej úrovni.

Označujú sa aj ako momenty polohy, pretože udávajú jedným číslom polohu skúmaného súboru na osi x .

Môžeme ich rozdeliť do dvoch skupín:

- stredné hodnoty, ktorých veľkosť závisí od hodnôt znaku všetkých jednotiek súboru,
- stredné hodnoty, ktorých veľkosť nie je závislá od hodnôt znaku všetkých jednotiek súboru.

STREDNÉ HODNOTY ZÍSKANÉ VÝPOČTOM ZO VŠETKÝCH HODNÔT SÚBORU

Veľkosť takýchto stredných hodnôt závisí od veľkosti hodnôt znaku všetkých jednotiek súboru. Túto skupinu reprezentujú priemery.

Aritmetický priemer $\bar{x}$

Aritmetický priemer je najčastejšie používanou strednou hodnotou. Vypočíta sa ako súčet hodnôt znaku delený počtom jednotiek. Výhodou je jednoduchosť výpočtu, nevýhodou citlivosť na existenciu extrémnych (vybočujúcich) hodnôt, ktoré priemer zmenia tak, že nepredstavuje reprezentatívnu hodnotu. Príklad: Priemerná mesačná mzda v spoločnosti, kde štyria zamestnanci majú mzdu 700 EUR a jej riaditeľ 2 200 EUR je 1 000 EUR. Táto suma nereprezentuje adekvátne mzdu zamestnancov a ani riaditeľa.

Uvedený problém citlivosti aritmetického priemeru na extrémne hodnoty sa rieši použitím inej strednej hodnoty (napr. mediánu) alebo vylúčením extrémnej hodnoty zo štatistického súboru, najmä ak máme dostatočne veľký štatistický súbor a existuje dôvod sa domnievať, že extrémna hodnota mohla byť spôsobená nejakou chybou (napr. preklepy, chyby merania).

Výpočet aritmetického priemeru:

a) jednoduchá forma

- súčet všetkých hodnôt kvantitatívneho znaku x_i , kde $i=1,2,\dots,n$ delíme počtom týchto hodnôt, teda rozsahom súboru n

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

b) vážená forma

- ak sú hodnoty usporiadané do radov početností a radov skupinového rozdelenia početností, využíva sa viacnásobný výskyt rovnakej hodnoty, a teda rovnaké hodnoty nesčítavame, ale násobíme ich počtom (n_i):

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \dots + x_k \cdot n_k}{n} = \frac{\sum_{i=1}^k x_i \cdot n_i}{n}$$

Aritmetický stred intervalu:

Pri skupinovom rozdelení početností sa hodnota x_i získa ako aritmetický stred intervalu.

$$s_i = \frac{x_{\max} + x_{\min}}{2}, \text{ kde}$$

s_i predstavuje stred i -teho intervalu,

$x_{\max}$ je jeho horná hranica,

$x_{\min}$ jeho dolná hranica.

Aritmetický stred intervalu je teoretická hodnota, ktorá nahrádza všetky hodnoty z daného intervalu (platí to len v prípade, ak je rozdelenie početností v intervale relatívne spojité, nie extrémne skokovité).

c) metóda vhodne zvoleného počiatku

Metóda vhodne zvoleného počiatku umožňuje zjednodušený výpočet aritmetického priemeru. Vychádza z vlastnosti aritmetického priemeru, že ak pripočítame, odpočítame, vydělíme, alebo vynásobíme každú hodnotu štatistického súboru konštantou, aj hodnota aritmetického priemeru sa zväčší, zmenší, vydělí, vynásobí o túto konštantu.

Pri aplikácii metódy postupujeme tak, že zmenšíme všetky hodnoty znaku o konštantu a (zväčša stred toho intervalu, ktorý je blízky hodnote hľadaného priemeru) a takto zmenšené hodnoty delíme konštantou h (zväčša je to šírka intervalu). Konštanty volíme tak, aby nová premenná mala priemer blízky nule a aby rozdiely medzi stredmi intervalov boli jednotkové.

Vytvoríme tak novú premennú, ktorú vyjadríme vzťahom: $z_i = \frac{x_i - a}{h}$. Z pomocnej premennej

z_i následne vypočítame aritmetický priemer pomocou vzorca $\bar{z} = \frac{\sum_{i=1}^k z_i \cdot n_i}{n}$. Vážený priemer pôvodných hodnôt dostaneme zo vzťahu: $\bar{x} = \bar{z} \cdot h + a$.

Šírka intervalu

Šírku intervalu h možno vypočítať podľa nasledovného vzťahu:

- a) v prípade, že hodnoty x_i nadobúdajú len celočíselné hodnoty, $h = x_{\max} - x_{\min} + 1$,
- b) v prípade, že hodnotou x_i môže byť akékoľvek reálne číslo, $h = x_{\max} - x_{\min}$.

Vlastnosti aritmetického priemeru:

1. Stálosť súčtu hodnôt znaku - ak každú hodnotu znaku nahradíme aritmetickým priemerom, ich súčet sa nezmení:

$x_1 + x_2 + \dots + x_k = \bar{x} + \bar{x} + \dots + \bar{x}$, teda $\sum_{i=1}^k x_i n_i = \bar{x} n$ pre jednoduchú formu

$x_1 n_1 + x_2 n_2 + \dots + x_k n_k = \bar{x} n_1 + \bar{x} n_2 + \dots + \bar{x} n_k$, teda $\sum_{i=1}^k x_i n_i = \bar{x} n$ pre váženú formu.

2. Súčet odchýlok jednotlivých hodnôt znaku od ich aritmetického priemeru sa rovná 0:

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

3. Súčet štvorcov odchýlok hodnôt znaku od aritmetického priemeru je minimálny, t.j. menší ako súčet štvorcov odchýlok od ľubovoľne zvoleného čísla.

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

4. Aritmetický priemer radu pozostávajúceho z konštanty c sa rovná tejto konštante:

$$x_1 = x_2 = \dots = x_k = c \rightarrow \sum_{i=1}^k x_i = nc \rightarrow c = \bar{x}$$

5. Aritmetický priemer zo súčtu hodnôt znaku x_i a y_i sa rovná súčtu ich aritmetických priemerov:

$$\overline{x+y} = \frac{1}{n} \sum_{i=1}^n (x_i + y_i) = \frac{1}{n} \sum_{i=1}^n x_i + \frac{1}{n} \sum_{i=1}^n y_i = \bar{x} + \bar{y}$$

6. Aritmetický priemer z rozdielu hodnôt znaku x_i a y_i sa rovná rozdielu ich aritmetických priemerov:

$$\overline{x-y} = \frac{1}{n} \sum_{i=1}^n (x_i - y_i) = \frac{1}{n} \sum_{i=1}^n x_i - \frac{1}{n} \sum_{i=1}^n y_i = \bar{x} - \bar{y}$$

Harmonický priemer $\bar{x}_h$

Harmonický priemer sa používa vtedy, keď je medzi hodnotami znaku a výsledným javom nepriamy vzťah, t.j. keď sú váhy hodnôt znaku dané nepriamo alebo keď sa priemer počíta z pomerných čísel.

Výpočet harmonického priemeru:

a) jednoduchá forma

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

b) vážená forma

- ak máme vyjadrené aj početnosti hodnôt znaku

$$\bar{x}_h = \frac{n_1 + n_2 + \dots + n_k}{\frac{1}{x_1} \cdot n_1 + \frac{1}{x_2} \cdot n_2 + \dots + \frac{1}{x_k} \cdot n_k} = \frac{\sum_{i=1}^k n_i}{\sum_{i=1}^k \frac{n_i}{x_i}}$$

c) z pomerných čísel

Pre pomerené čísla označené symbolom w platí $w = \frac{y}{x}$, pričom y predstavuje číselnú hodnotu, ktorej veľkosť na jednu jednotku x sa vyjadruje číslom w , a x je základ pomerného čísla. Ak poznáme n pomerných čísel w_i pre $i=1, 2, \dots, n$ a máme vypočítať ich priemer, musíme brať do úvahy skutočnosť, že každé pomerné číslo v štatistickom súbore môže mať rôznu váhu, ktorá je daná základom pomerného čísla (Bakytová 1975).

Priemer z pomerných čísel sa určí:

→ ak sú známe priame váhy (základ pomerného čísla), použijeme tvar váženého aritmetického priemeru

$$\bar{w} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n x_i},$$

→ ak poznáme len nepriame váhy (čitateľa pomerných čísel), použijeme tvar váženého harmonického priemeru

$$\overline{w} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n \frac{y_i}{w_i}},$$

→ v prípade, že poznáme čitateľa i menovateľa pomerných čísel, môžeme ich priemer určiť priamo

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

Vlastnosti harmonického priemeru:

1. Stálosť súčtu obrátených hodnôt znaku – ak každú hodnotu znaku nahradíme harmonickým priemerom, ich súčet sa nezmení.

$$\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} = \frac{1}{x_h} + \frac{1}{x_h} + \dots + \frac{1}{x_h}$$

$$\sum_{i=1}^n \frac{1}{x_i} = \frac{n}{x_h} \rightarrow \overline{x_h} = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} \rightarrow \text{pre jednoduchú formu}$$

$$\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_k}{x_k} = \frac{n_1}{x_h} + \frac{n_2}{x_h} + \dots + \frac{n_k}{x_h}$$

$$\sum_{i=1}^k \frac{n_i}{x_i} = \frac{\sum_{i=1}^k n_i}{x_h} \rightarrow \overline{x_h} = \frac{\sum_{i=1}^k n_i}{\sum_{i=1}^k \frac{n_i}{x_i}} \rightarrow \text{pre váženú formu}$$

2. Harmonický priemer konštanty sa rovná danej konštante.

$$x_1 = x_2 = x_3 = \dots = x_n = c$$

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

3. Ak násobíme (delíme) váhy v štatistickom súbore ľubovoľnou konštantou c , veľkosť vypočítaného harmonického priemeru sa nezmení.

$$\bar{x}_h = \frac{\sum_{i=1}^k cn_i}{\sum_{i=1}^k \frac{cn_i}{x_i}} = \frac{c \sum_{i=1}^k n_i}{c \sum_{i=1}^k \frac{n_i}{x_i}} = \frac{\sum_{i=1}^k n_i}{\sum_{i=1}^k \frac{n_i}{x_i}}$$

Geometrický priemer $\bar{x}_g$

Geometrický priemer sa používa v prípadoch, kedy má reálny zmysel súčin hodnôt skúmaného znaku (multiplikatívny vzťah medzi veličinami). Definujeme ho ako n -tú odmocninu zo súčinu n_i -tych mocnín jednotlivých hodnôt znaku (n je rozsah súboru). Využíva sa pri výpočte priemerného, resp. plánovaného tempa rastu (napr. počtu obyvateľov miest). Geometrický priemer je možné počítat iba z hodnôt, ktoré sú kladné. Ak je $\bar{x}_g > 1$ dochádza k priemernému nárastu skúmaného javu, ak je $\bar{x}_g < 1$, dochádza k priemernému poklesu.

jednoduchá forma:

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n} = \sqrt[n]{\prod_{i=1}^n x_i}$$

vážená forma:

$$\bar{x}_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^k x_i^{n_i}}$$

Geometrický priemer sa niekedy nazýva aj logaritmický priemer. Odmocniny vyšších radov sa počítajú komplikovane, preto sa pre výpočet geometrického priemeru využíva logaritmovanie. Zlogaritmovaním pôvodného výrazu dostávame:

jednoduchá forma

$$\log \bar{x}_g = \frac{1}{n} \cdot (\log x_1 + \log x_2 + \dots + \log x_n) = \frac{1}{n} \cdot \sum_{i=1}^n \log x_i$$

vážená forma

$$\log \bar{x}_g = \frac{n_1 \cdot \log x_1 + n_2 \cdot \log x_2 + \dots + n_k \cdot \log x_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i \cdot \log x_i}{\sum_{i=1}^k n_i}$$

STREDNÉ HODNOTY ZÍSKANÉ VÝPOČTOM IBA Z NIEKTORÝCH HODNÔT SÚBORU

Niekedy pri štatistickom rozbere nastane situácia, kedy priemery dostatočne necharakterizujú štatistický súbor:

- ich hodnoty sú výrazne ovplyvnené extrémnymi hodnotami súboru,
- nepoznáme hodnoty všetkých prvkov súboru.

Vtedy je potrebné použiť iné stredné hodnoty, ktoré nie sú odvodené zo všetkých údajov súboru, ale iba z niektorých hodnôt.

Medián

- stredná hodnota, ktorá rozdeľuje usporiadaný štatistický súbor na dve rovnako početné časti,
- je necitlivý voči extrémnym hodnotám a jeho určenie je veľmi jednoduché, nakoľko nepotrebujeme poznať všetky hodnoty súboru,
- označenie: Q_1^2 , resp. $\tilde{x}$.

Výpočet mediánu v štatistickom súbore:

a) s nepárnym počtom jednotiek:

- medián je reálna hodnota,
- je to hodnota r-tej štatistickej jednotky

$$\tilde{x} = x_r,$$

ktorej poradie vypočítame podľa vzťahu:

$$r = \frac{n+1}{2},$$

Pr.: 2, 5, 7, 8, 11 $\rightarrow \tilde{x} = 7$

b) s párnym počtom jednotiek:

- medián je teoretická hodnota,
- je to priemer hodnoty znaku r-tej a (r+1)-ej štatistickej jednotky,

$$\tilde{x} = \frac{x_r + x_{r+1}}{2},$$

pričom $r = \frac{n}{2}.$

$$\text{Pr.: } 1, 3, \underline{5}, 7, 9, 11 \quad \rightarrow \tilde{x} = 6 \quad (7+5)/2$$

c) pre intervalové rozdelenie početností:

- bezprostredne je možné určiť iba mediánový interval, t.j. interval, do ktorého patrí x_r a x_{r+1} ,
- konkrétne hodnoty mediánového intervalu nie sú známe, a tak sa medián určí lineárnou interpoláciou:

1. pre relatívne početnosti:

$$\tilde{x} = a + h \cdot \frac{0,5 - \sum_{i=1}^{r-1} f_i}{f_{\tilde{x}}},$$

a - dolná hranica mediánového intervalu,

h - šírka mediánového intervalu,

$\sum_{i=1}^{r-1} f_i$ - relatívna kumulatívna početnosť po mediánový interval,

$f_{\tilde{x}}$ - relatívna početnosť mediánového intervalu,

2. pre absolútne početnosti:

$$\tilde{x} = a + h \cdot \frac{\frac{n+1}{2} - \sum_{i=1}^{r-1} n_i}{n_{\tilde{x}}},$$

$\sum_{i=1}^{r-1} n_i$ - absolútna kumulatívna početnosť po mediánový interval,

$n_{\tilde{x}}$ - absolútna početnosť mediánového intervalu.

Modus $\hat{x}$ (z lat. modus – spôsob, miera)

- najčastejšie sa vyskytujúca hodnota roztriedeného štatistického súboru, teda hodnota s najvyššou (maximálnou) početnosťou,

- na grafe rozdelenia početností pri jednovrcholovom rozdelení - je modus x-ovou súradnicou jeho vrcholu,
- pri dvojvrcholových rozdeleniach možno obidva vrcholy krivky pokladať za modus, aj keď vrcholy nie sú rovnako vysoké – tzv. bimodálne rozdelenia,
- pri niektorých typoch nemá zmysel hovoriť o mode (keď krivka rozdelenia početností nemá nijaký vrchol) – tzv. antimodálne rozdelenia.

Určenie modusu zo skupinového rozdelenia početností

Pri skupinovom rozdelení početností sa priamo určí iba **modálny interval**, t.j. interval s maximálnou početnosťou. O polohe modusu v najpočetnejšom intervale sa predpokladá, že jeho vzdialenosť od dolnej, resp. hornej hranice je úmerná rozdielom medzi najväčšou početnosťou a početnosťou predchádzajúceho, príp. nasledujúceho intervalu. Polohu modusu v modálnom intervale je možné určiť dvoma spôsobmi:

a) výpočtom:

1. pre rovnako veľké intervaly

$$\hat{x} = s_{Mo} - \frac{h_{Mo}}{2} \cdot \frac{n_{Mo-1} - n_{Mo+1}}{n_{Mo+1} + n_{Mo-1} + 2 \cdot n_{Mo}},$$

s_{Mo} – stred modálneho intervalu,

h_{Mo} – šírka modálneho intervalu,

n_{Mo} – početnosť modálneho intervalu,

n_{Mo-1} – početnosť intervalu nachádzajúceho sa pred modálnym intervalom,

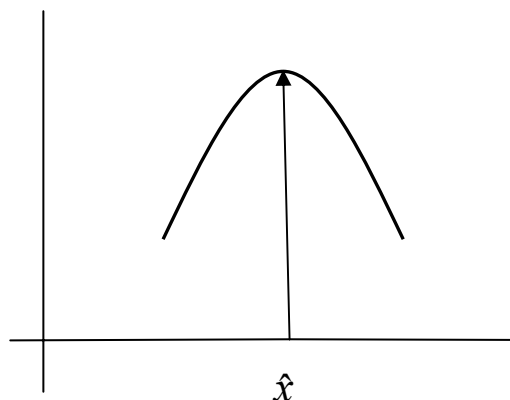
n_{Mo+1} – početnosť intervalu nachádzajúceho sa za modálnym intervalom.

2. pre nerovnako veľké intervaly

$$\hat{x} = s_{Mo} - \frac{h_{Mo}}{2} \cdot \frac{n_{Mo-1} \cdot h_{Mo-1} - n_{Mo+1} \cdot h_{Mo+1}}{n_{Mo+1} \cdot h_{Mo+1} + n_{Mo-1} \cdot h_{Mo-1} + 2 \cdot n_{Mo} \cdot h_{Mo}},$$

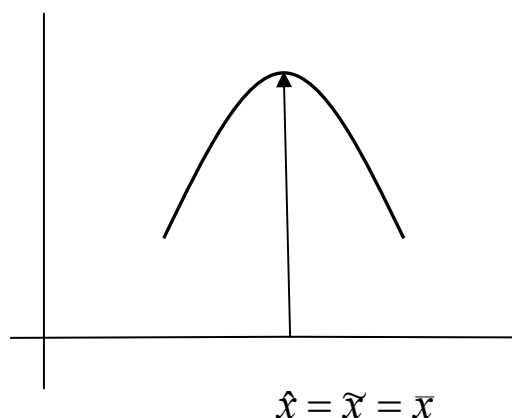
- jednotlivé početnosti intervalov sa násobia šírkou týchto intervalov,

b) graficky - lokalizáciou vrcholu v grafe, t.j. odčítaním hodnoty znaku s najvyššou početnosťou.

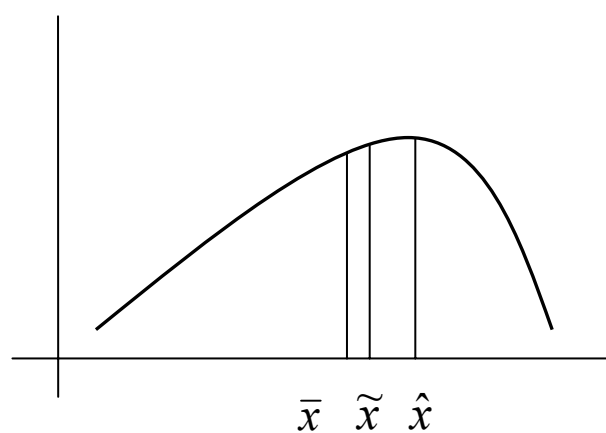
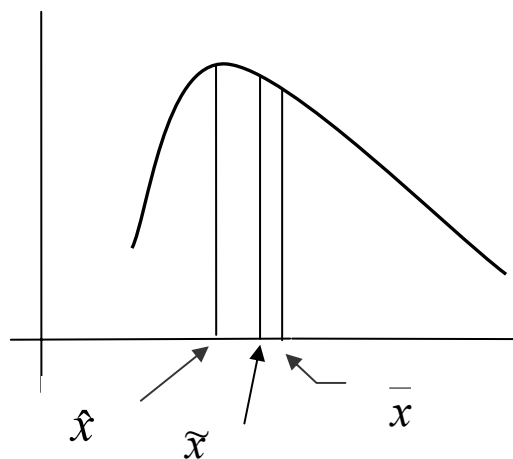


Vzťahy medzi strednými hodnotami

Vzdialenosť medzi aritmetickým priemerom, mediánom a modusom určitým spôsobom charakterizuje asymetriu štatistického súboru a môže sa stať východiskom pri konštrukcii mier asymetrie. Čím je rozloženie početností súmernejšie, tým menej sa líši úroveň aritmetického priemeru, mediánu a modusu. Ak je rozdelenie početností v súbore úplne symetrické vzhľadom na určitú hodnotu znaku $x_i=a$, ktorá sa v súbore vyskytuje najčastejšie, potom:



Ak je rozdelenie asymetrické, aritmetický priemer je vzhľadom na modus posunutý smerom k zošikmej časti krivky rozdelenia početností a medián je približne v tretine vzdialenosti od aritmetického priemeru smerom k modusu:



Pre približné určenie modusu v štatistickom súbore platí vzťah: $\hat{x} = \bar{x} - 3 \cdot (\bar{x} - \tilde{x})$.

Riešené úlohy:

1. Zo skupinového rozdelenia početností vypočítajte aritmetický priemer metódou vhodne zvoleného počiatku.

x_i	n_i
10-19	2
20-29	22
30-39	16
40-49	8

Riešenie:

x_i	n_i	s_i	$s_i - a$	$z_i = \frac{s_i - a}{h}$	$z_i n_i$
10-19	2	14,5	-20	-2	-4
20-29	22	24,5	-10	-1	-22
30-39	16	34,5	0	0	0
40-49	8	44,5	10	1	8
50-59	2	55,5	20	2	4

Poznámka: Číslo 34,5 si vyberieme za konštantu a (Možné je vybrať si akékoľvek číslo zo stĺpca s_i , avšak najvýhodnejšie je zvoliť si číslo v strede stĺpca). Šírku intervalu vyrátame pomocou vzorca $h = x_{\max} - x_{\min} + 1$.

Vypočítame si $\bar{z} = \frac{\sum_{i=1}^5 z_i \cdot n_i}{n} = \frac{-14}{50} = -0,28$ a vrátime sa k výpočtu $\bar{x} = \bar{z} \cdot h + a = -0,28 \cdot 10 + 34,5 = 31,7$.

2. Jeden výrobok zhotovuje v dielni niekoľko robotníkov. Jeden z nich ho vyrobí za 4 minúty, iní 10 za 5 minút, ďalší 12 za 6 minút, 5 za 10 minút a 6 za 12 minút. Vypočítajte priemerný počet minút potrebných na zhotovenie výrobku.

Riešenie:

x_i (čas v minútach)	n_i (počet robotníkov)
4	1
5	10
6	12
10	5
12	6

Nakoľko medzi hodnotou znaku a výsledným javom je nepriamy vzťah, použijeme vzorec pre harmonický priemer (čím šikovnejší je robotník, tým kratší čas potrebuje na zhotovenie výrobku).

$$\bar{x}_h = \frac{\sum_{i=1}^5 n_i}{\sum_{i=1}^5 \frac{n_i}{x_i}} = \frac{n_1 + n_2 + \dots + n_5}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_5}{x_5}} = \frac{1 + 10 + 12 + 5 + 6}{\frac{1}{4} + \frac{10}{5} + \frac{12}{6} + \frac{5}{10} + \frac{6}{12}} \approx 6,5$$

Odpoveď: Robotník potrebuje v priemere 6,5 minúty na zhotovenie výrobku.

3. V tabuľke sú uvedené informácie o počte obyvateľov, rozlohe a hustote zaľudnenia v okresoch Trnavského kraja za rok 1995. Vypočítajte priemernú hustotu zaľudnenia v Trnavskom kraji, ak poznáte:

- a) hustotu zaľudnenia a rozlohu územia,
- b) hustotu zaľudnenia a počet obyvateľov,
- c) počet obyvateľov a rozlohu územia.

Okresy	Počet obyvateľov	Rozloha okresu (v km ²)	Hustota zaľudnenia (obyvateľov/km ²)
Trnava	126 333	741	170,5
Dunajská streda	111 100	1075	103,3
Galanta	93 808	641	146,3
Hlohovec	45 595	267	170,8
Piešťany	64 079	381	168,2
Senica	60 502	684	88,5
Skalica	46 550	359	129,6

Spracované podľa: Korec, et. al. 1997

Riešenie:

Pri riešení uvedenej úlohy vychádzame zo vzťahu $w = \frac{y}{x}$, kde y predstavuje číselnú jednotku, ktorej veľkosť sa vyjadruje na jednotku x – základ pomerného čísla, w je pomerné číslo.

a) ak poznáme priame váhy (základ pomerného čísla - v našom prípade ho predstavuje rozloha okresu), využijeme vzťah váženého aritmetického priemeru:

$$\bar{w} = \frac{\sum_{i=1}^7 w_i x_i}{\sum_{i=1}^7 x_i} = \frac{547914,5}{4148} \approx 132,1$$

Rozloha okresu (v km ²) - x _i	Hustota zaľudnenia (obyvateľov/km ²) - w _i	w _i · x _i
741	170,5	126340,5
1075	103,3	111047,5
641	146,3	93778,3
267	170,8	45603,6
381	168,2	64084,2
684	88,5	60534
359	129,6	46526,4
Σ=4148		Σ=547914,5

b) ak sú dané len nepriame váhy (čitateľ pomerných čísel), je vhodné využiť nasledovný vzťah:

$$\bar{w} = \frac{\sum_{i=1}^7 y_i}{\sum_{i=1}^7 \frac{y_i}{w_i}} = \frac{547967}{4148,4} \approx 132,1$$

Počet obyvateľov - y _i	Hustota zaľudnenia (obyvateľov/km ²) - w _i	$\frac{y_i}{w_i}$
126 333	170,5	741,0
111 100	103,3	1075,5
93 808	146,3	641,2
45 595	170,8	266,9
64 079	168,2	381,0
60 502	88,5	683,6
46 550	129,6	359,2
Σ=547 967		Σ=4148,4

c) v prípade, že poznáme čitateľa a menovateľa pomerných čísel, možno priemer vypočítať automaticky:

$$\bar{w} = \frac{\sum_{i=1}^7 y_i}{\sum_{i=1}^7 x_i} = \frac{547967}{4148} \approx 132,1.$$

4. Nasledujúca tabuľka poukazuje na spotrebu elektrickej energie za roky 1991-1997. Zistite priemerný ročný koeficient rastu spotreby elektrickej energie za celé sledované obdobie 1991-1997 ak viete, že spotreba elektrickej energie v roku 1992 poklesla vzhľadom na rok 1991 o 2% a v roku 1997 vzrástla vzhľadom na rok 1996 o 3%.

Rok	Spotreba energie (v mil. kWh)
1991	1235
1992	nepoznáme
1993	1348
1994	1212
1995	1281
1996	1300
1997	nepoznáme

Riešenie:

Najprv si vypočítame ročné koeficienty rastu. Ak poznáme hodnoty spotreby elektrickej energie pre dva po sebe nasledujúce roky, potom ročný koeficient rastu vypočítame ako

$x_i = \frac{x_{i+1}^{\circ}}{x_i^{\circ}}$, kde x_i° predstavuje spotrebu elektrickej energie v určitom roku a x_{i+1}° spotrebu

elektrickej energie v roku po ňom nasledujúcom. V prípade roku 1992 síce hodnotu spotreby elektrickej energie nepoznáme, no ak uvažíme, že spotreba elektrickej energie v roku 1991 bola 100% a v roku 1992 už len 98%, potom pre ročný koeficient rastu x_1 za sledované

obdobie 1992/91 platí: $x_1 = \frac{98}{100} = 0,98$. Obdobne by sme mohli vypočítať aj hodnotu ročného

koeficientu rastu pre obdobie 1997/96: $x_6 = \frac{103}{100} = 1,03$. Nakoľko v roku 1992 sa spotreba

elektrickej energie znížila o 2% na úroveň 1210,3 mil.kWh, pre ročný koeficient rastu za sledované obdobie 1993/92 platí: $x_2 = \frac{1348}{1210,3} = 1,114$. Výslednú hodnotu priemerného

ročného koeficientu rastu vyrátame zo vzorca pre geometrický priemer, kde n bude predstavovať počet sledovaných období, hodnoty x_i° ročné koeficienty rastu a hodnoty $n_1, n_2, \dots, n_5$ budú v danom prípade rovné 1, nakoľko priemerný ročný koeficient rastu sme sledovali stále v rámci obdobia jedného roka.

Obdobie	Ročný koeficient rastu
1992/91	0,98
1993/92	1,114
1994/93	0,899
1995/94	1,057
1996/95	1,015
1997/96	1,03

Poznámka: Ak by sme nemali údaje o ročnom koeficiente rastu za sledované obdobie 1992/91 a 1993/92, ale len údaje o ročnom koeficiente rastu x_1 za obdobie 1993/91, potom by sme za n_1 dosadili hodnotu 2.

$$\bar{x}_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_k^{n_k}} = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_6} = \sqrt[6]{0,98 \cdot 1,114 \cdot 0,899 \cdot 1,057 \cdot 1,015 \cdot 1,03} \approx 1,014$$

Odpoveď: Priemerný ročný koeficient rastu za sledované obdobie 1991-1997 mal hodnotu 1,014.

5. V roku 1995 vyťažila ťažobná spoločnosť 400 000 ton železnej rudy a v roku 1999 až 450 000 ton. Vypočítajte priemerný ročný koeficient rastu ťažby rúd za celé sledované obdobie 1995-1999. O koľko percent priemerne ročne sa zmenila ťažba rúd?

Riešenie:

Nakoľko ide o výpočet tempa rastu, použijeme vzorec pre výpočet geometrického priemeru:

$$\bar{x}_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_k^{n_k}} = \sqrt[n]{\prod_{i=1}^k x_i^{n_i}}.$$

Rok	Ťažba železnej rudy (v tonách)
1995	400 000
1996	nepoznáme
1997	nepoznáme
1998	nepoznáme
1999	450 000

Označme ťažbu v jednotlivých rokoch ako $x_1^\circ, x_2^\circ, x_3^\circ, x_4^\circ, x_5^\circ$. Potom pre priemerný ročný koeficient rastu platí:

Obdobie	Ročný koeficient rastu
1996/95	$\frac{x_2^\circ}{x_1^\circ} = x_1$
1997/96	$\frac{x_3^\circ}{x_2^\circ} = x_2$
1998/97	$\frac{x_4^\circ}{x_3^\circ} = x_3$
1999/98	$\frac{x_5^\circ}{x_4^\circ} = x_4$

$$\bar{x}_g = \sqrt[n]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_k^{n_k}} = \sqrt[4]{x_1 \cdot x_2 \cdot \dots \cdot x_4} = \sqrt[4]{\frac{x_2^\circ}{x_1^\circ} \cdot \frac{x_3^\circ}{x_2^\circ} \cdot \frac{x_4^\circ}{x_3^\circ} \cdot \frac{x_5^\circ}{x_4^\circ}} = \sqrt[4]{\frac{x_5^\circ}{x_1^\circ}} = \sqrt[4]{\frac{450000}{400000}} \approx 1,03$$

1,03-1,0=0,03, t.j. 3%.

Odpoveď: Priemerný ročný koeficient rastu bol 1,03. Ťažba rúd stúpala o 3%.

6. Z výsledkov bodového ohodnotenia žiakov na konci semestra vypočítajte

a) medián,

b) modus.

Body	Počet žiakov
10	2
15	5
17	6
19	8
25	3

Riešenie:

a) Nakoľko rozsah súboru je párny, pre výpočet mediánu použijeme nasledovný vzorec:

$$\tilde{x} = \frac{x_r + x_{r+1}}{2}, \quad \text{kde} \quad r = \frac{n}{2}.$$

Body (x_i)	počet žiakov (n_i)	N_i
10	2	2
15	5	7
17	6	13
19	8	21
25	3	24

$$r = \frac{n}{2} = \frac{24}{2} = 12, \text{ teda medián má hodnotu:}$$

$$\tilde{x} = \frac{x_r + x_{r+1}}{2} = \frac{x_{12} + x_{13}}{2} = \frac{17 + 17}{2} = 17.$$

b) Modus je štatistická jednotka, ktorej absolútna početnosť je najväčšia, t.j. 19.

7. V tabuľke sú uvedené informácie o počte detí, ktoré navštevujú záujmový krúžok:

a) vypočítajte priemerný vek detí, ktoré navštevujú záujmový krúžok,

b) vypočítajte vek detí, v ktorom je návštevnosť krúžku najväčšia,

c) vypočítajte vek, do ktorého navštívi krúžok polovica všetkých detí.

Vek detí	Počet detí
(5-7>	5
(7-9>	10
(9-11>	5
(11-13>	4
(13-15>	3

Riešenie:

a) Cieľom je vypočítať aritmetický priemer. Vek detí si označíme ako x_i , počet detí ako n_i .

Stred intervalu vypočítame pomocou vzorca $s_i = \frac{x_{\max} + x_{\min}}{2}$.

Vek detí (x_i)	Počet detí (n_i)	s_i	$s_i n_i$	N_i
(5-7>	5	6	30	5
(7-9>	10	8	80	15
(9-11>	5	10	50	20
(11-13>	4	12	48	24
(13-15>	3	14	42	27

Aritmetický priemer: $\bar{x} = \frac{\sum_{i=1}^k s_i n_i}{n} = \frac{250}{27} \approx 9,26.$

Odpoveď: Priemerný vek detí, ktoré navštevujú záujmový krúžok je 9,26 roka.

b) Úloha vedie k výpočtu modusu. Najprv si musíme určiť modálny interval. Modálnym intervalom je interval, ktorého absolútna početnosť je najväčšia, teda interval (7-9>. Následne vypočítame šírku intervalu. Nakoľko ide o matematické spojité intervaly, kde hodnota x_i môže predstavovať akékoľvek číslo z oboru reálnych čísel, šírku intervalu určíme ako $h = x_{\max} - x_{\min}$.

modálny interval

Vek detí (x_i)	Počet detí (n_i)	s_i	N_i
(5-7>	5	6	5
(7-9>	10	8	15
(9-11>	5	10	20
(11-13>	4	12	24
(13-15>	3	14	27

Modus:

$$\hat{x} = s_{Mo} - \frac{h_{Mo}}{2} \cdot \frac{n_{Mo-1} - n_{Mo+1}}{n_{Mo+1} + n_{Mo-1} + 2 \cdot n_{Mo}} = 8 - \frac{2}{2} \cdot \frac{5-5}{5+5+2 \cdot 10} = 8$$

Odpoveď: Návštevnosť krúžku je najväčšia vo veku 8 rokov.

c) V prípade, že máme vypočítať vek, do ktorého navštívi krúžok polovica všetkých detí, ide o určenie hodnoty mediánu. Nakoľko opäť pracujeme s matematickými spojitými intervalmi, kde hodnota x_i môže predstavovať akékoľvek číslo z oboru reálnych čísel, šírku intervalu určíme ako $h = x_{\max} - x_{\min}$. Mediánovým intervalom bude ten, v ktorom sa nachádza

relatívna kumulatívna početnosť 50%, teda ten interval, v ktorom sa nachádza 50% štatistických jednotiek súboru.

Vek detí (x_i)	Počet detí (n_i)	N_i	f_i	F_i
(5-7>	5	5	0,1852	0,1852
(7-9>	10	15	0,3704	0,5556
(9-11>	5	20	0,1852	0,7408
(11-13>	4	24	0,1481	0,8889
(13-15>	3	27	0,1111	1,00

n – rozsah súboru

Medián:

$$\tilde{x} = a + h \cdot \frac{0,5 - \sum_{i=1}^{r-1} f_i}{f_{\tilde{x}}} = 7 + 2 \cdot \frac{0,5 - 0,1852}{0,3704} \approx 8,7$$

Ak použijeme vzorec pre absolútne početnosti, mediánový interval by sme určili pomocou vzorca $k = \frac{1}{2}(n+1) = \frac{1}{2}(27+1) = 14$. Vypočítané k udáva poradie štatistickej jednotky, ktorá mi predstavuje medián (teda medián je štrnástá štatistická jednotka v poradí). Z uvedeného vyplýva, že medián sa nachádza v intervale (7-9>.

$$\tilde{x} = a + h \cdot \frac{\frac{n+1}{2} - \sum_{i=1}^{r-1} n_i}{n_{\tilde{x}}} = 7 + 2 \cdot \frac{14 - 5}{10} = 8,8$$

Poznámka: Uvedená odchýlka medzi výpočtom hodnoty mediánu z absolútnych a relatívnych početností nastala v dôsledku zaokrúhľovania čísel.

Odpoveď: Polovica všetkých detí navštívi krúžok do veku 8,8 roka.

Neriešené úlohy:

1. Zo skupinového rozdelenia početnosti vypočítajte aritmetický priemer metódou vhodne zvoleného počiatku.

x_i	n_i
10-14	3
15-19	6
20-24	8
25-29	4

2. Do vodnej nádrže vtekajú tri rieky. Keby do prázdnej nádrže vtekala každá rieka osobitne, potom prvá by ju naplnila za 300 dní, druhá za 400 dní a tretia za 600 dní. Vypočítajte priemerný počet dní potrebných na naplnenie nádrže riekou.

3. V tabuľke sú informácie o rozlohe osevných plôch v jednotlivých krajoch SR, produkcii obilia v tonách a priemerných výnosoch obilia z 1 hektára v tonách. Vypočítajte priemerný výnos obilia z 1 ha osevnej plochy v rámci celej Slovenskej republiky, ak poznáte:

a) rozlohu osevnej plochy a výnos v tonách z 1 ha,

b) produkciu obilia a výnos v tonách z 1 ha,

c) produkciu obilia a rozlohu osevnej plochy.

Kraj	Produkcia obilia (v tonách)	Rozloha osevnej plochy (v ha)	Výnos v tonách z 1 ha
Bratislavský	151 112	39 558,1	3,82
Trnavský	716 028	157 368,8	4,55
Trenčiansky	166 438	43 915,0	3,79
Nitriansky	1 167 111	253 719,8	4,60
Žilinský	54 363	17 650,3	3,08
Banskobystrický	234 793	74 537,5	3,15
Prešovský	138 007	49 464,9	2,79
Košický	300 952	96 769,1	3,11

Spracované podľa: Správa o poľnohospodárstve a potravinárstve v Slovenskej republike 2007 (stav za rok 2006) - (Zelená správa, 2007) a Štatistická ročenka regiónov Slovenska 2007 (údaje sú za rok 2006).

4. Vypočítajte priemerný ročný koeficient rastu obyvateľstva SR v období 1992-1997. Určte, aký bol priemerný ročný koeficient prírastku obyvateľstva a o koľko percent sa priemerne ročne zmenil počet obyvateľov (teda priemerné ročné tempo prírastku obyvateľstva) .

Rok	Počet obyvateľov (v mil.)
1992	5,314
1993	5,336
1994	5,356
1995	5,368
1996	5,379
1997	5,388

Spracované podľa: Štatistická ročenka Slovenskej republiky.

5. Z uvedených hodnôt 1,2,3,3,1,1,1,5,5,5,5,8,8,9,9,1,1 vypočítajte:

a) medián,

b) modus.

6. Určte medián a modus v súbore 50 študentov strednej školy. Sledovaným znakom je ich priemerná hmotnosť v kg.

Priemerná hmotnosť v kg	Počet študentov
(50-60>	10
(60-70>	12
(70-80>	6
(80-90>	14
(90-100>	8

5 Kvantily

Kvantily

Kvantily sú hodnoty znaku, ktoré rozdeľujú usporiadaný štatistický súbor na α rovnakých častí. Označujeme ich Q_k^α , kde

α predstavuje zvolené číslo, na koľko častí sa má rozdeliť štatistický súbor,

k poradie kvantilu ($k = 1, 2, \dots, \alpha-1$).

Ak $k = 1$, ide o dolný kvantil, ak $k = \alpha-1$, ide o horný kvantil.

Ak $\alpha = 2$, kvantil nazývame **medián**,

$\alpha = 4$, kvantily nazývame **kvartily**,

$\alpha = 6$, kvantily nazývame **sextily**,

$\alpha = 10$, kvantily nazývame **decily**,

$\alpha = 100$, kvantily nazývame **percentily**.

Výpočet kvantilov

Spôsob výpočtu kvantilov z empirického štatistického súboru závisí od toho, či v štatistickom súbore poznáme hodnotu sledovaného znaku u každej štatistickej jednotky alebo nie.

a) Individuálne rozdelenie početností

Ak sú známe jednotlivé hodnoty štatistického znaku, potom hľadaným kvantilom je

$\frac{k}{\alpha} \cdot n$ za predpokladu, že $\frac{k}{\alpha} \cdot n$ je celé číslo. Ak $\frac{k}{\alpha} \cdot n$ nie je celé číslo, potom príslušný kvantil nájdeme pomocou aritmetického priemeru tých hodnôt štatistického súboru, medzi ktoré daná hodnota padla.

b) Skupinové (intervalové) rozdelenie početností

Ak máme intervalové rozdelenie početností, väčšinou nemožno určiť hodnotu kvantilu priamo. Bezprostredne je možné určiť len interval, do ktorého skúmaný kvantil padne.

Kvantilovými intervalmi budú tie, pre ktoré $N_i = \frac{k}{\alpha} \cdot n$ alebo $F_i = \frac{k}{\alpha}$. Hľadané hodnoty

kvantilov možno vypočítať pomocou vzorcov pre **absolútne**

$$Q_k^\alpha = a + h \cdot \frac{\left(\frac{k}{\alpha} \cdot n\right) + 0,5 - \sum_{i=1}^{r-1} n_i}{n_Q}, \text{ respektívne relatívne početnosti. } Q_k^\alpha = a + h \cdot \frac{\frac{k}{\alpha} - \sum_{i=1}^{r-1} f_i}{f_Q},$$

kde r označuje poradie kvantilového intervalu,

a dolnú hranicu kvantilového intervalu,

$\sum_{i=1}^{r-1} n_i$ absolútnu kumulatívnu početnosť po kvantilový interval,

$\sum_{i=1}^{r-1} f_i$ relatívnu kumulatívnu početnosť po kvantilový interval,

n_Q absolútnu početnosť kvantilového intervalu,

f_Q relatívnu početnosť kvantilového intervalu.

Riešené úlohy:

1. Z výsledkov bodového ohodnotenia žiakov na konci semestra vypočítajte:

a) dolný kvartil,

b) horný kvartil.

Body	Počet žiakov
10	2
15	5
17	6
19	8
25	3

Riešenie:

a) Poradie dolného kvartilu Q_1^4 vypočítame z výrazu $\frac{k}{\alpha} \cdot n$. Nakoľko $\frac{k}{\alpha} \cdot n = \frac{1}{4} 24 = 6$ je celé číslo, dolný kvartil bude 6. Štatistickou jednotkou v poradí, teda $Q_1^4 = 15$.

b) Poradie horného kvartilu Q_3^4 určíme z výrazu $\frac{k}{\alpha} \cdot n$. Keďže $\frac{k}{\alpha} \cdot n = \frac{3}{4} 24 = 18$,

horný kvartil bude 18. Štatistickou jednotkou v poradí, teda $Q_3^4 = 19$.

2. V tabuľke sú uvedené informácie o počte detí, ktoré navštevujú záujmový krúžok. Vypočítajte vek, do ktorého navštívi krúžok štvrtina všetkých detí.

Vek detí	Počet detí
(5-7>	5
(7-9>	10
(9-11>	5
(11-13>	4
(13-15>	3

Úloha vedie k výpočtu dolného kvartilu. Využijeme vzorec pre relatívne početnosti. Dolnokvartilový interval je ten, v ktorom sa nachádza 25% štatistických jednotiek súboru, teda interval (7-9>. Je to druhý interval v poradí, preto $r=2$.

dolnokvartilový interval

Vek detí (x_i)	Počet detí (n_i)	N_i	f_i	F_i
(5-7>	5	5	0,1852	0,1852
(7-9>	10	15	0,3704	0,5556
(9-11>	5	20	0,1852	0,7408
(11-13>	4	24	0,1481	0,8889
(13-15>	3	27	0,1111	1,00

n – rozsah súboru

$$Q_1^4 = a + h \cdot \frac{0,25 - \sum_{i=1}^{r-1} f_i}{f_{Q_1^4}} = 7 + 2 \cdot \frac{0,25 - 0,1852}{0,3704} \approx 7,3.$$

Odpoveď: Štvrtina všetkých detí navštevuje krúžok do veku 7,3 roka.

Neriešené úlohy:

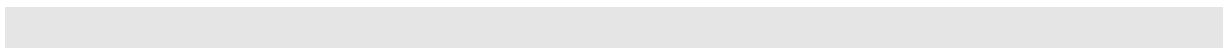
1. Z uvedených hodnôt vypočítajte: 1,2,3,3,1,1,1,5,5,5,5,8,8,9,9,1,1.

a) dolný kvartil

b) horný kvartil.

2. Určte dolný a horný kvartil v súbore 50 študentov strednej školy. Sledovaným znakom je ich priemerná hmotnosť v kg.

<i>Priemerná hmotnosť v kg</i>	<i>Počet študentov</i>
<i>(50-60></i>	<i>10</i>
<i>(60-70></i>	<i>12</i>
<i>(70-80></i>	<i>6</i>
<i>(80-90></i>	<i>14</i>
<i>(90-100></i>	<i>8</i>



6 Miery variability

Štatistické súbory sú charakterizované nielen svojimi strednými hodnotami, ale aj variabilitou, premenlivosťou výskytu hodnôt. Znamená to, že i keď dva štatistické súbory majú rovnaké stredné hodnoty, môže ísť o odlišné súbory. Príčiny variability sú rôzne (vecné, časové a priestorové). Mnohé miery variability vyjadrujú, akým spôsobom sú usporiadané konkrétne hodnoty znaku okolo stredných hodnôt. Skúmané hodnoty znaku môžu byť okolo strednej hodnoty usporiadané veľmi voľne (t.j. rozptýlene) až veľmi tesne (t.j. hodnoty sú rozmiestnené v bezprostrednom susedstve so strednou hodnotou). Miery variability sú tiež odchýlky od niektorej strednej hodnoty alebo odchýlky medzi jednotlivými hodnotami. Využívajú sa aj na vyjadrenie chyby merania, vyplývajúcej z nemožnosti zistiť akúkoľvek premennú veličinu s absolútnou presnosťou.

ABSOLÚTNE MIERY VARIABILITY

- umožňujú charakterizovať variabilitu v štatistickom súbore

Variačné rozpätie

- je rozdiel medzi maximálnou a minimálnou hodnotou skúmaného znaku. Je závislé od krajných hodnôt štatistického radu, teda je necitlivé voči ostatným hodnotám.

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

Kvantilové rozpätie

- ide o rozdiel medzi horným a dolným kvantilom:

$$R_Q = Q_{\alpha-1}^{\alpha} - Q_1^{\alpha}.$$

Pri výpočte je potrebné uviesť o aký typ kvantilu ide, napr. kvartilové rozpätie, decilové rozpätie a podobne. Kvartilové rozpätie obsahuje polovicu všetkých jednotiek štatistického súboru. Polovičná hodnota kvartilového rozpätia sa nazýva **kvartilová odchýlka**:

$$Q = \frac{Q_3^4 - Q_1^4}{2}.$$

Odchýlka

- najčastejšie používaná miera variability

- je definovaná ako absolútna hodnota rozdielu hodnoty štatistického znaku a určitej strednej hodnoty, z ktorej bola vypočítaná.

Definujeme: a) odchýlku od aritmetického priemeru: $|d_i| = |x_i - \bar{x}|$,

b) odchýlku od mediánu $|d_i| = |x_i - \tilde{x}|$,

c) odchýlku od modusu $|d_i| = |x_i - \hat{x}|$.

Súčet jednotlivých odchýlok hodnôt od aritmetického priemeru je rovný nule, preto sa tieto odchýlky vyjadrujú v tvare absolútnej hodnoty.

Priemerná odchýlka

a) priemerná odchýlka od aritmetického priemeru

- je to vlastne jednoduchý aritmetický priemer jednotlivých absolútnych odchýlok od aritmetického priemeru

jednoduchá forma:

$$\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n} = \frac{\sum_{i=1}^n |d_i|}{n}$$

vážená forma:

$$\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^k |x_i - \bar{x}| \cdot n_i}{n} = \frac{\sum_{i=1}^k |d_i| \cdot n_i}{n}$$

- ak hodnotíme opakované merania na tom istom objekte, priemerná odchýlka od aritmetického priemeru sa označuje ako **absolútna stredná chyba**

b) priemerná odchýlka od mediánu

jednoduchá forma:

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

vážená forma:

$$\bar{d}_{\tilde{x}} = \frac{\sum_{i=1}^k |x_i - \tilde{x}| \cdot n_i}{n} = \frac{\sum_{i=1}^k |d_i| \cdot n_i}{n}$$

c) priemerná odchýlka od modusu

jednoduchá forma:

$$\bar{d}_{\hat{x}} = \frac{\sum_{i=1}^n |x_i - \hat{x}|}{n} = \frac{\sum_{i=1}^n |d_i|}{n}$$

vážená forma:

$$\bar{d}_{\hat{x}} = \frac{\sum_{i=1}^k |x_i - \hat{x}| \cdot n_i}{n} = \frac{\sum_{i=1}^k |d_i| \cdot n_i}{n}$$

Rozptyl, variancia s_x^2 , $\text{var } x$, σ^2

- nazýva sa aj priemerná štvorcová (kvadratická) odchýlka,
- charakterizuje vzájomnú odlišnosť jednotlivých hodnôt znaku a zároveň aj varianciu jednotlivých hodnôt znaku od priemeru,
- je definovaný ako aritmetický priemer štvorcov odchýlok jednotlivých hodnôt znaku od ich aritmetického priemeru.

jednoduchá forma:

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

- pri jeho výpočte nedochádza ku vzájomnej kompenzácii kladných a záporných odchýlok od aritmetického priemeru, pretože sa počíta so štvorcom odchýlok, ktoré majú vždy kladnú hodnotu

vážená forma:

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

Rozptyl možno vypočítať aj iným spôsobom, a to ako rozdiel aritmetického priemeru zo štvorcov hodnôt znaku a štvorca ich aritmetického priemeru:

jednoduchá forma:

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

vážená forma:

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

Smerodajná (štandardná) odchýlka s_x , σ ,

Je druhou odmocninou rozptylu. Nazýva sa tiež absolútnou chybou merania.

jednoduchá forma:

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

vážená forma:

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

V geoinformatike existujú aplikácie, pri ktorých sa vykonáva odhad skúmanej veličiny pomocou matematickej, interpolačnej funkcie (Hofierka, 2003). V prípade prvkov (bodov), pre ktoré poznáme aj skutočné, správne hodnoty, môžeme využiť štatistické ukazovatele chyby, ktoré majú blízky vzťah k odchýlke a smerodajnej odchýlke.

Priemerná absolútna chyba (anglicky MAE - mean absolute error):

$$MAE = \frac{1}{n} \sum_{i=1}^n |x_i^* - x_i|.$$

Stredná kvadratická chyba, reziduálna variabilita (anglicky RMSE - root mean square error):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i^* - x_i)^2},$$

kde x_i^* je odhadovaná hodnota a x_i je skutočná hodnota.

RELATÍVNE (POMERNÉ) MIERY VARIABILITY

- umožňujú porovnávať variabilitu v rôznych štatistických súboroch,
- zvyčajne charakterizujú variabilitu vzhľadom na aritmetický priemer,
- najčastejšie používané relatívne miery variability sú relatívna priemerná odchýlka a variačný koeficient.

Relatívna priemerná odchýlka

- je to podiel absolútnej priemernej odchýlky a tej strednej hodnoty, od ktorej boli počítané odchýlky,
- najčastejšie sa používa **relatívna priemerná odchýlka od aritmetického priemeru**,
- ide o podiel priemernej odchýlky a aritmetického priemeru

$$\delta_x = \overline{d'_x} = \frac{\overline{d_x}}{\bar{x}}$$

- vyjadruje sa v %,
- ak hodnotíme opakované merania na tom istom objekte relatívna priemerná odchýlka od aritmetického priemeru sa označuje aj ako **relatívna (pomerná) stredná chyba**.

Variačný koeficient

- je to podiel smerodajnej odchýlky a aritmetického priemeru daného znaku

$$V_x = \frac{s_x}{\bar{x}},$$

- často sa vyjadruje v %.

Riešené úlohy:

1. V tabuľke sú informácie o priemerných mesačných prietokoch rieky Morava na vodomernej stanici Kopčany a rieky Bodrog na vodomernej stanici Streda nad Bodrogom v roku 2003. Vypočítajte pomernú priemernú odchýlku a variačný koeficient. Napíšte, ktorý z uvedených vodných tokov mal v danom roku vyrovnanjší prietok?

Mesiac	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.
Rieka/stanica												
Morava/stanica Kopčany	115,2	43,4	83,3	73,6	46,0	19,3	13,5	8,0	7,1	18,0	16,3	30,5
Bodrog/stanica Streda nad Bodrogom	75,3	54,1	167,0	166,8	67,6	42,7	32,9	28,6	27,9	65,1	72,6	52,1

Spracované podľa: www.shmu.sk/File/kvant_pr_03/Morava_Q03.pdf,
www.shmu.sk/File/kvant_pr_03/Bodrog_Q03.pdf

Riešenie:

Najprv si vypočítame pomernú priemernú odchýlku pre rieku Morava/stanica Kopčany. Pomocné údaje sú uvedené v tabuľke.

x_i	$x_i - \bar{x}$	$ d_i = x_i - \bar{x} $	$(x_i - \bar{x})^2$
115,2	75,68	75,68	5727,46
43,4	3,88	3,88	15,05
83,3	43,78	43,78	1916,69
73,6	34,08	34,08	1161,45
46	6,48	6,48	41,99
19,3	-20,22	20,22	408,85
13,5	-26,02	26,02	677,04
8,0	-31,52	31,52	993,51
7,1	-32,42	32,42	1051,06
18,0	-21,52	21,52	463,11
16,3	-23,22	23,22	539,17
30,5	-9,02	9,02	81,36

$\delta_x = \frac{\bar{d}_{\bar{x}}}{\bar{x}}$, kde $\bar{x} = \frac{\sum_{i=1}^{12} x_i}{n} = \frac{474,2}{12} \approx 39,52$ a $\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^{12} |d_i|}{n} = \frac{327,84}{12} \approx 27,32$. Následne určíme veľkosť pomernej priemernej odchylky: $\delta_x = \frac{\bar{d}_{\bar{x}}}{\bar{x}} = \frac{27,32}{39,52} = 0,6913$. $\delta_x = 69,13\%$. Variačný

koeficient vypočítame podľa vzťahu $V_x = \frac{s_x}{\bar{x}} 100$, kde smerodajná odchýlka nadobúda

hodnotu $s_x = \sqrt{\frac{\sum_{i=1}^{12} (x_i - \bar{x})^2}{n}} = \sqrt{\frac{13076,74}{12}} = \sqrt{1089,73} \approx 33,01$. Výsledná hodnota variačného

koeficientu bude: $V_x = \frac{s_x}{\bar{x}} 100 = \frac{33,01}{39,52} 100 \approx 83,53\%$.

Odpoveď: Pomerná priemerná odchýlka pre rieku Morava/Kopčany bola 69,13% a veľkosť variačného koeficientu 83,53 %.

Poznámka: Uvedené informácie o priemerných mesačných prietokoch nebolo nutné usporiadať do tvaru variačného radu. Nakoľko aritmetický priemer použitý pri výpočte zadanej úlohy je závislý od všetkých hodnôt štatistického súboru, usporiadaním hodnôt by sa hodnota variačného koeficientu nezmenila.

Druhá časť úlohy smeruje k výpočtu pomernej priemernej odchýlky pre rieku Bodrog/stanica Streda nad Bodrogom. Pomocné údaje sú uvedené v tabuľke.

x_i	$x_i - \bar{x}$	$ d_i = x_i - \bar{x} $	$(x_i - \bar{x})^2$
75,3	4,24	4,24	17,98
54,1	-16,96	16,96	287,64
167,0	95,94	95,94	9204,48
166,8	95,74	95,74	9166,15
67,6	-3,46	3,46	11,97
42,7	-28,36	28,36	804,29
32,9	-38,16	38,16	1456,19
28,6	-42,46	42,46	1802,85
27,9	-43,16	43,16	1862,79
65,1	-5,96	5,96	35,52
72,6	1,54	1,54	2,37
52,1	-18,96	18,96	359,48
spolu	x	394,94	25 011,71

Poznámka: x – neuvedený údaj.

$\delta_x = \frac{\bar{d}_{\bar{x}}}{\bar{x}}$, kde $\bar{x} = \frac{\sum_{i=1}^{12} x_i}{n} = \frac{852,7}{12} \approx 71,06$ a $\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^{12} |d_i|}{n} = \frac{394,94}{12} \approx 32,91$. Následne určíme veľkosť pomernej priemernej odchýlky: $\delta_x = \frac{\bar{d}_{\bar{x}}}{\bar{x}} = \frac{32,91}{71,06} = 0,4632$. $\delta_x = 46,32\%$. Variačný koeficient vypočítame podľa vzťahu $V_x = \frac{s_x}{\bar{x}} 100$, kde smerodajná odchýlka nadobúda hodnotu $s_x = \sqrt{\frac{\sum_{i=1}^{12} (x_i - \bar{x})^2}{n}} = \sqrt{\frac{25011,71}{12}} = \sqrt{2084,31} \approx 45,65$. Výsledná hodnota variačného koeficientu bude: $V_x = \frac{s_x}{\bar{x}} 100 = \frac{45,65}{71,06} 100 \approx 64,24\%$.

Odpoveď: Pomerná priemerná odchýlka pre rieku Bodrog/stanica Streda nad Bodrogom bola 46,32 %. Veľkosť variačného koeficientu bola 64,24 %. Z porovnania hodnôt variačného koeficientu u riek Morava a Bodrog vyplýva, že rieka Bodrog mala v danom roku 2003 vyrovnanejší prietok (prejavila sa u nej menšia variabilita hodnôt).

2. Zhodnoťte variabilitu v súbore študentov, u ktorých bola sledovaná telesná hmotnosť (v kg). Výsledky interpretujte.

Hmotnosť v kg (x_i)	Počet študentov (n_i)
(50,60>	5
(60,70>	7
(70,80>	8
(80,90>	9
(90,100>	2
(100,110>	1

Riešenie:

a) Aritmetický priemer:

Pri výpočte aritmetického priemeru postupujeme podľa vzorca $\bar{x} = \frac{\sum_{i=1}^6 s_i n_i}{n}$, kde s_i predstavujú stredy matematických spojitých intervalov, ktoré vypočítame ako $s_i = \frac{x_{\min} + x_{\max}}{2}$, kde $x_{\min}$ a

x_{max} predstavujú krajné hodnoty matematických spojitých intervalov. Následne hodnotu aritmetického priemeru vypočítame zo vzťahu: $\bar{x} = \frac{\sum_{i=1}^6 s_i n_i}{n} = \frac{2390}{32} \approx 74,69$.

b) Priemerná odchýlka od aritmetického priemeru:

$$\bar{d}_{\bar{x}} = \frac{\sum_{i=1}^6 |d_i| n_i}{n} = \frac{332,48}{32} = 10,39.$$

c) Relatívna priemerná odchýlka od aritmetického priemeru:

$$\bar{\delta}_{\bar{x}} = \frac{\bar{d}_{\bar{x}}}{\bar{x}} = \frac{10,39}{74,69} \approx 0,1391.$$

Odpoveď: Priemerná odchýlka tvorí 13,91 % z hodnoty aritmetického priemeru.

Poznámka: Čím väčšia je hodnota pomernej (relatívnej) priemernej odchýlky, tým je štatistický súbor variabilnejší.

d) Rozptyl:

$$s_x^2 = \frac{\sum_{i=1}^6 (s_i - \bar{x})^2 n_i}{n} = \frac{5296,875}{32} = 165,53.$$

e) Smerodajná odchýlka:

$$s_x = \sqrt{s_x^2} = \sqrt{165,53} = 12,87.$$

f) Variačný koeficient:

$$V_x = \frac{s_x}{\bar{x}} 100 = \frac{12,87}{74,69} 100 = 17,23\%.$$

Odpoveď: Smerodajná odchýlka tvorí 17,23 % z hodnoty aritmetického priemeru. Štatistický súbor sa vyznačuje pomerne malou variabilitou.

Poznámka: Variačný koeficient u homogénnych súborov s nízkou hodnotou variability štatistického znaku dosahuje hodnoty 5-20 %. V prípade súborov s extrémne asymetrickým rozdelením, teda s vysokou variabilitou hodnôt znaku početností môže nadobúdať hodnotu blízku 100 %.

g) Variačné rozpätie:

Variačné rozpätie je definované ako rozdiel medzi maximálnou a minimálnou hodnotou skúmaného štatistického znaku. Závisí od krajných hodnôt štatistického radu, teda je necitlivé voči všetkým ostatným hodnotám. Určíme ho zo vzťahu:

$R_x = x_{max} - x_{min} = 110 - 50 = 60$, kde x_{max} predstavuje maximálnu a x_{min} minimálnu hodnotu skúmaného štatistického radu.

Hmotnosť v kg x_i	Počet štud. n_i	Stred int. s_i	N_i	$s_i \cdot n_i$	$ d_i = s_i - \bar{x} $	$ d_i n_i$	$(s_i - \bar{x})^2$	$(s_i - \bar{x})^2 \cdot n_i$
(50,60>	5	55	5	275	19,69	98,45	387,6961	1938,4805
(60,70>	7	65	12	455	9,69	67,83	93,8961	657,2727
(70,80>	8	75	20	600	0,31	2,48	0,0961	0,7688
(80,90>	9	85	29	765	10,31	92,79	106,2961	956,6649
(90,100>	2	95	31	190	20,31	40,62	412,4961	824,9922
(100,110>	1	105	32	105	30,31	30,31	918,6961	918,6961
suma	x	x	x	2390	x	332,48	x	5296,875

Poznámka: x – neuvedený údaj.

7 Momenty polohy, miery šikmosti, špicatosti a koncentrácie, teoretické rozdelenia náhodných premenných

MOMENTY POLOHY

Základom mnohých opisných charakteristík (akými sú napr. aritmetický priemer či rozptyl) sú momenty. **Všeobecný moment r -tého stupňa** možno definovať vzťahom:

$\varphi_{x,r} = \frac{1}{n} \cdot \sum_{i=1}^k (x_i - a)^r \cdot n_i$ pre absolútne početnosti, resp. $\varphi_{x,r} = \sum_{i=1}^k (x_i - a)^r \cdot f_i$ pre relatívne početnosti, kde x_i sú hodnoty znaku, n_i absolútne početnosti, f_i relatívne početnosti, n rozsah súboru, k počet tried, a ľubovoľne zvolené číslo a r stupeň momentu.

V prípade, že do vzorca všeobecného momentu dosadíme $a=0$, dostaneme **začiatočný moment r -tého stupňa** v tvare: $v_{x,r} = \frac{1}{n} \cdot \sum_{i=1}^k x_i^r \cdot n_i$ pre absolútne početnosti a $v_{x,r} = \sum_{i=1}^k x_i^r \cdot f_i$ pre relatívne početnosti. Ak $r=1$, dostaneme vzorec pre aritmetický priemer, ktorý je začiatočným momentom prvého stupňa.

Centrálny moment r -tého stupňa (resp. r -ty centrálny moment) je odvodený od všeobecného momentu, ak $a = \bar{x}$. Pre absolútne početnosti tak dostávame vzorec $\mu_{x,r} = \frac{1}{n} \cdot \sum_{i=1}^k (x_i - \bar{x})^r \cdot n_i$, pre relatívne $\mu_{x,r} = \sum_{i=1}^k (x_i - \bar{x})^r \cdot f_i$. Ak $r=1$, absolútnym centrálnym momentom prvého stupňa je priemerná odchýlka od aritmetického priemeru $\mu_{x,1} = \frac{1}{n} \cdot \sum_{i=1}^k |x_i - \bar{x}| \cdot n_i$, ak $r=2$ je centrálnym momentom druhého stupňa rozptyl $\mu_{x,2} = \frac{1}{n} \cdot \sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i = s_x^2$.

Pri výpočtoch zložitejších charakteristík je vhodné definovať pojem **normovanej premennej** t_i . Je definovaná ako veľkosť dochýlky hodnoty znaku x_i od aritmetického priemeru $\bar{x}$ prepočítaná na veľkosť smerodajnej odchýlky s_x : $t_i = \frac{x_i - \bar{x}}{s_x}$. Pre centrálny

moment normovanej premennej platí vzťah $\mu_{t,r} = \frac{1}{n} \cdot \sum_{i=1}^k (t_i - \bar{t})^r \cdot n_i$.

Vzťahy medzi začiatočným a centrálnym momentom normovanej premennej (Gregorová, Fillová 2004):

1. Prvý začiatočný moment normovanej premennej t sa rovná 0:

$$\bar{t} = v_{1,t} = \frac{1}{n} \sum_{i=1}^k t_i^1 \cdot n_i = \bar{t} = 0.$$

2. Centrálna a začiatočné momenty normovanej premennej sa rovnajú:

$$\mu_{t,r} = \frac{1}{n} \sum_{i=1}^k (t_i - \bar{t})^r \cdot n_i = \frac{1}{n} \sum_{i=1}^k (t_i - 0)^r \cdot n_i = \frac{1}{n} \sum_{i=1}^k t_i^r \cdot n_i = v_{t,r}.$$

3. Druhý centrálny moment $\mu_{t,2}$ a druhý začiatkový moment $v_{t,2}$ normovanej premennej t sa rovná 1.

$$\mu_{t,2} = \frac{1}{n} \sum_{i=1}^k (t_i - \bar{t})^2 \cdot n_i = \frac{1}{n} \sum_{i=1}^k (t_i - 0)^2 \cdot n_i = v_{t,2} = 1.$$

4. Začiatkové i centrálné momenty normovanej premennej t sa dajú vyjadriť pomocou začiatkových a centrálnych momentov pôvodnej premennej x .

$$v_{t,r} = \mu_{t,r} = \frac{\mu_{x,r}}{s_x^r} = \frac{\frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^r \cdot n_i}{s_x^r}.$$

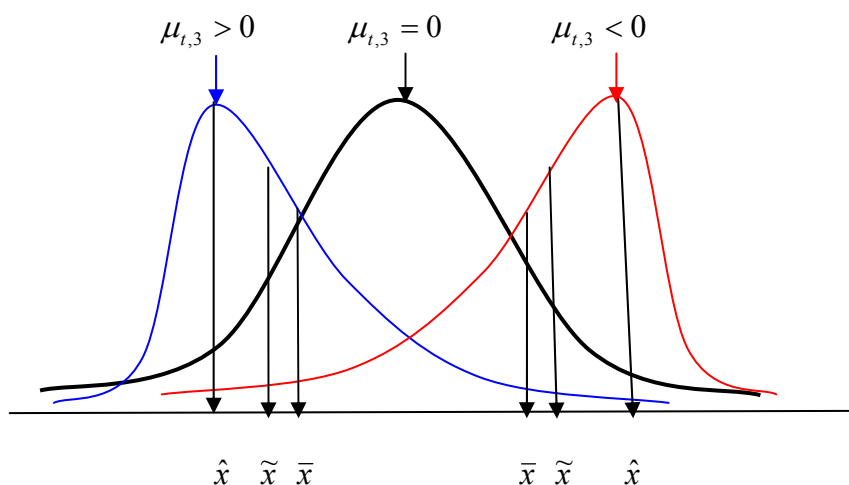
MIERY ŠIKMOSTI

Ak dve rozdelenia početností majú rovnaké stredné hodnoty a rozptyly, to ešte neznamená, že tieto rozdelenia sú rovnaké. Pri asymetrických rozdeleniach dochádza ku koncentrácii hodnôt do určitej oblasti - vľavo alebo vpravo od strednej hodnoty. Pre numerické vyjadrenie tvaru rozdelenia početností sa využívajú miery šikmosti (asymetrie). Na vyjadrenie zošikmenia sa používa **koefficient šikmosti (asymetrie)**. Koefficient šikmosti nazývame aj **tretím centrálnym momentom normovanej premennej t** . Spolu s mierami polohy (strednými hodnotami) a variability patrí k základným popisným charakteristikám štatistického súboru.

Koefficient šikmosti vyjadruje smer a stupeň asymetrie rozdelenia premennej.

Možno ho vypočítať nasledovným spôsobom: $\gamma_1 = \mu_{t,3} = \frac{\mu_{x,3}}{s_x^3} = \frac{\frac{1}{n} \sum_{i=1}^k (x_i - \bar{x})^3 \cdot n_i}{s_x^3}.$

Graf 4 Miera šikmosti



Spracované podľa:

<http://www.fem.uniag.sk/ACADEMIC/depart/ksov/predmety/statistika/subory/prednaska22.ppt>

Kladná hodnota šikmosti (ľavostranná šikmost') vyjadruje, že priemer je väčší ako medián, teda väčšina hodnôt je menšia ako priemer. Záporná hodnota šikmosti (pravostranná šikmost') znamená, že priemer je menší ako medián, a teda väčšina hodnôt je väčšia ako priemer. Ak je koeficient šikmosti rovný 0, ide o symetrické rozdelenie, teda priemer a medián sa rovnajú (graf 4).

Mieru šikmosti možno definovať i pomocou **Pearsonovej a kvartilovej miery šikmosti**. **Pearsonova miera šikmosti** je definovaná ako rozdiel medzi aritmetickým priemerom a modusom prepočítaný na veľkosť smerodajnej odchýlky:

$$\check{S}_p = \frac{\bar{x} - \hat{x}}{s_x}.$$

Pre úplne symetrické rozdelenia početnosti platí, že $\bar{x} = \hat{x}$, teda $\check{S}_p = 0$, pre ľavostranne asymetrické rozdelenia početnosti je $\bar{x} > \hat{x}$, teda $\check{S}_p > 0$, pre pravostranne asymetrické rozdelenia početností je $\bar{x} < \hat{x}$, z čoho vyplýva, že $\check{S}_p < 0$.

Kvartilová miera šikmosti vychádza z faktu, že v prípade symetrického rozdelenia početností je medzi dolným kvartilom a mediánom rovnaký počet štatistických jednotiek ako medzi horným kvartilom a mediánom, čo možno napísať v tvare: $(\tilde{x} - Q_1^4) = (Q_3^4 - \tilde{x})$.

Pri asymetrických rozdeleniach početnosti daný vzťah neplatí, čo sa využíva pre výpočet šikmosti, ktorá je vyjadrená nasledovným vzťahom:

$$\check{S}_K = \frac{(Q_3^4 - \tilde{x}) - (\tilde{x} - Q_1^4)}{(Q_3^4 - \tilde{x}) + (\tilde{x} - Q_1^4)} = \frac{Q_3^4 + Q_1^4 - 2\tilde{x}}{2Q}, \text{ kde } \check{S}_K \in \langle -1, 1 \rangle.$$

Hodnota Q predstavuje kvartilovú odchýlku $\frac{Q_3^4 - Q_1^4}{2}$. Pri ľavostranne asymetrických rozdeleniach početností sú hodnoty $\check{S}_K > 0$, pri pravostranne asymetrických rozdeleniach $\check{S}_K < 0$.

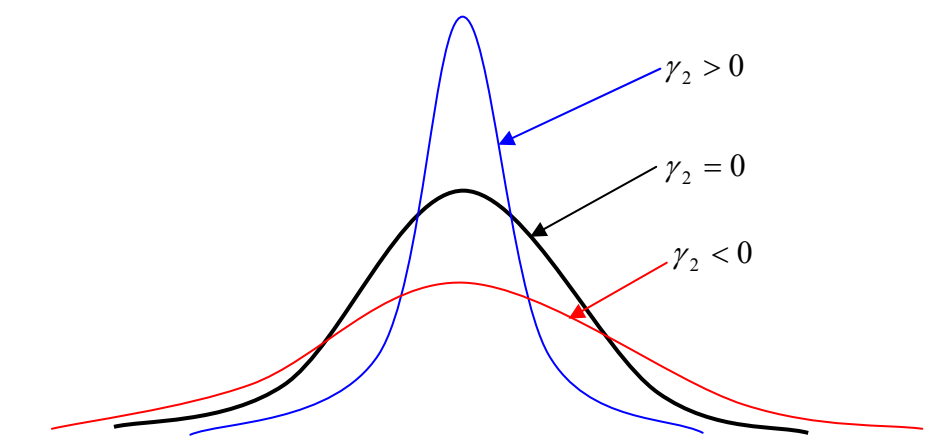
MIERY ŠPICATOSTI

Koncentrácia hodnôt okolo strednej hodnoty sa meria pomocou koeficientu špicatosti. Špicatosť rozdelenia sa porovnáva so špicatosťou normálneho (Laplaceovo-Gaussovho) rozdelenia. Keďže pri normálnom rozdelení sa špicatosť rovná 3 a chceme, aby sa miera špicatosti rovnala 0, od hodnoty koeficientu špicatosti jednoducho odpočítame číslo 3.

Koeficient špicatosti nazývame aj **štvrtým centrálnym momentom normovanej premennej t zmenšený o 3**:

$$\gamma_2 = \mu_{t,4} - 3 = \frac{\mu_{x,4}}{s_x^4} - 3 = \frac{\frac{1}{n} \cdot \sum_{i=1}^k (x_i - \bar{x})^4 \cdot n_i}{s_x^4} - 3.$$

Graf 5 Miera špicatosti



Spracované podľa:

<http://www.fem.uniag.sk/ACADEMIC/depart/ksov/predmety/statistika/subory/prednaska22.ppt>

V prípade unimodálnych rozdelení, u ktorých je špicatosť väčšia ako 0, je výskyt extrémnych hodnôt častejší. Takéto rozdelenia majú špicatejší vrchol. Unimodálne rozdelenia, ktorých špicatosť je menšia ako 0, teda výskyt extrémnych hodnôt je menej častý ako u normálneho rozdelenia, sú plochejšie (graf 5).

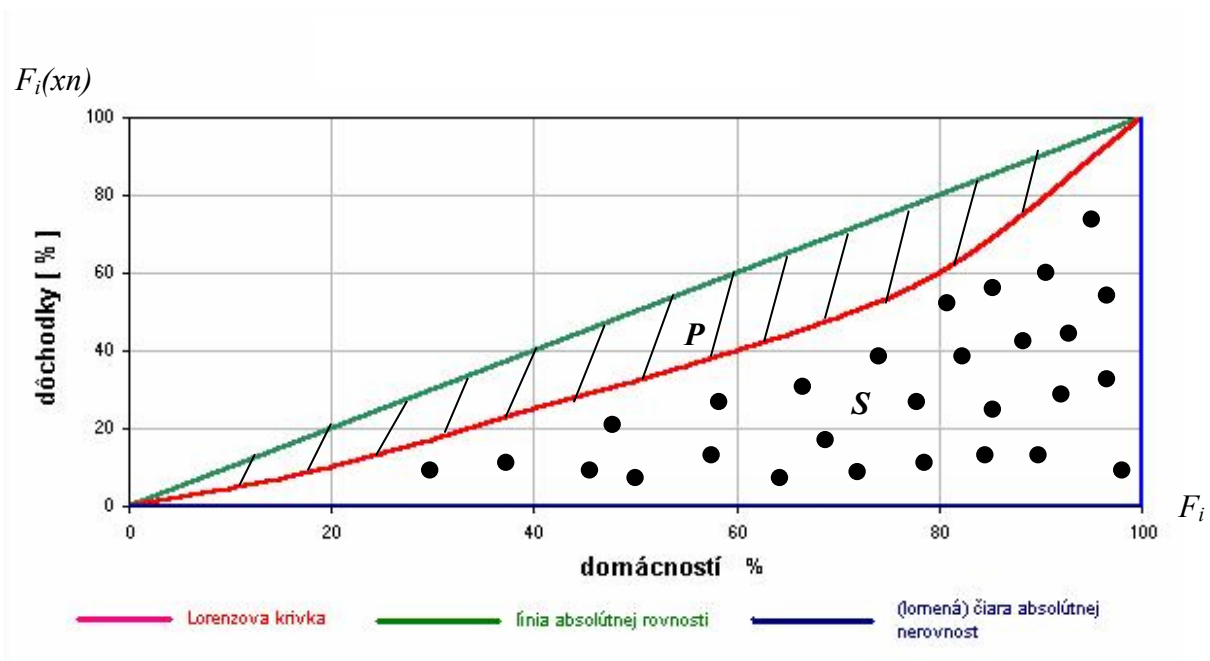
MIERY KONCENTRÁCIE

Koncentráciu hodnôt štatistického znaku možno vyjadriť Lorenzovou krivkou, kde sa na os x -ovú sa nanášajú relatívne kumulatívne početnosti F_i a na os y -ovú relatívne kumulatívne podiely skupín (tried) na celkovom úhrne, ktoré označujeme $F_i(xn)$ resp. Z_i . Pre relatívny

triedny úhrn $f_i(xn)$ platí:
$$f_i(xn) = \frac{x_i n_i}{\sum_{i=1}^k x_i n_i}.$$

Uvedená metóda znázorňovania koncentrácie štatistického znaku bola po prvýkrát predstavená v roku 1905 Max O. Lorenzom pre prezentáciu rozdelenia dôchodkov. Autor na os x -ovú zaznačil kumulatívne percento domácností, na os y -ovú kumulatívne percento získaného dôchodku. Pospájaním x -ových a y -ových súradníc vznikla krivka (Lorenzova krivka) znázorňujúca stupeň rovnosti alebo nerovnosti pri rozdeľovaní dôchodkov (graf 6). Čím viac sa uvedená krivka odchyľovala od čiar s uhlom 45° , tým väčšia bola nerovnomernosť pri rozdeľovaní dôchodkov. Ideálna Lorenzova krivka, ktorá je vedená pod 45 -stupňovým uhlom, predstavuje absolútnu rovnosť v dôchodkoch, z čoho vyplýva, že napríklad 10 % domácností poberá 10 % dôchodkov v spoločnosti. V prípade, že Lorenzova krivka je totožná s osou x a v bode 100 % na osi x sa mení na kolmicu na os x , môže dôjsť k situácii, že 1 % domácností poberá 100 % dôchodkov v spoločnosti, čo je absolútny nepomer. V praxi sa však stretávame väčšinou s Lorenzovou krivkou, ktorá sa nachádza medzi dvoma extrémami, teda v našom prípade medzi ideálnou Lorenzovou krivkou a krivkou, ktorá znázorňuje absolútnu nerovnosť v rozdeľovaní dôchodkov. Z uvedeného vyplýva, že čím je skutočná Lorenzova krivka bližšie k ideálnej Lorenzovej krivke, tým je rozdelenie dôchodkov v ekonomike rovnomernejšie a naopak. Obdobne by sme mohli analyzovať aj iné spoločensko-ekonomické javy, ako napr. koncentráciu obyvateľstva vzhľadom na rozlohu územia a podobne.

Graf 6 Lorenzova krivka



Spracované podľa: <http://www.avoarm.sk/Lorenz.htm>

Veľkosť koncentrácie určitého javu možno číselne vyjadriť pomocou **koeficientu koncentrácie**. Možno ho definovať ako podiel plochy medzi Lorenzovou krivkou a diagonálou (P) a plochy pravouhlého trojuholníka tvoreného diagonálou, osou x a kolmicou spustenou na os x z bodu $[1,1]$, resp. $[100\%, 100\%]$ (graf 6).

$$K_K = \frac{P}{T} = \frac{0,5 - S}{0,5} = 1 - 2S, \text{ kde}$$

P je plocha ohraničená diagonálou a Lorenzovou krivkou,

T plocha trojuholníka rovnajúca sa hodnote 0,5,

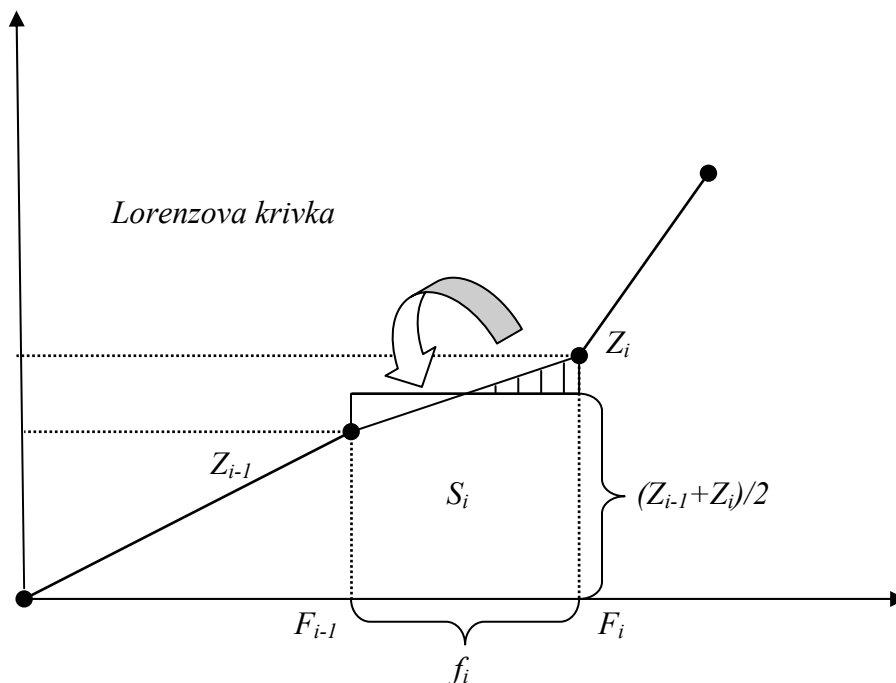
S plocha ohraničená Lorenzovou krivkou, x -ovou osou a kolmicou spustenou v bode $[1,1]$ resp. $[100\%, 100\%]$.

Plochu S vypočítame ako súčet obsahov obdĺžnikov s rozmermi: šírka f_i , výška rovná $(Z_{i-1} + Z_i)/2$, pretože obdĺžnik s takýmito rozmermi má rovnaký obsah ako lichobežník, ktorý vznikne pod jednou časťou Lorenzovej krivky (graf 7). Výraz $(Z_{i-1} + Z_i)/2$ predstavuje stred jednej časti Lorenzovej krivky, $Z_0=0$ je začiatok Lorenzovej krivky, Z_i kumulatívny podiel prvej až i -tej triedy na celkovom úhrne hodnôt znaku v štatistickom súbore. Ak do vzorca $K_K = \frac{P}{T} = \frac{0,5 - S}{0,5} = 1 - 2S$ dosadíme za S výraz $\sum_{i=1}^k \frac{f_i(Z_{i-1} + Z_i)}{2}$, mieru koncentrácie možno vyjadriť i nasledovne:

$$K_K = 1 - \sum_{i=1}^k f_i(Z_{i-1} + Z_i).$$

Uvedená miera koncentrácie nadobúda hodnoty z intervalu (0, 1). Čím sa hodnoty blížia viac k 1, tým je koncentrácia výraznejšia.

Graf 7 Schematické vyjadrenie miery koncentrácie



Spracované podľa: Gregorová, Fillová 2004

Pomer koncentrácie P_K je aritmetickým priemerom z relatívnych rozdielov medzi kumulatívnymi relatívnymi početnosťami v triedach a kumulatívnymi podielmi skupín na úhrne hodnôt znaku v štatistickom súbore:

$$P_K = \frac{\sum_{i=1}^{k-1} (F_i - Z_i)}{\sum_{i=1}^{k-1} F_i}, \text{ kde } P_K \in (0, -1).$$

Čím viac sa hodnota koeficientu P_K blíží k 0, tým je koncentrácia menšia.

VYBRANÉ TEORETICKÉ ROZDELENIA NÁHODNÝCH PREMENNÝCH

Náhodná veličina (jav) je premenná, ktorá môže nadobúdať rôzne hodnoty alebo hodnoty z rôznych intervalov v závislosti na náhode. Náhodné veličiny budeme označovať X , ich konkrétne hodnoty: $x_i, i=1, 2 \dots n$. Na štatistické znaky (premenné) sa môžeme pozerať ako na:

- diskrétné** -nadobúdajú izolované, väčšinou celočíselné hodnoty,
- spojité** -môžu nadobúdať ľubovoľné hodnoty z ohraničeného alebo neohraničeného intervalu.

Zákon rozdelenia náhodnej veličiny je pravidlo, ktoré každej hodnote náhodnej veličiny priradí pravdepodobnosť nadobudnutia danej hodnoty v prípade diskretných náhodných veličín, alebo množine hodnôt z každého intervalu priradí pravdepodobnosť nadobudnutia hodnôt z intervalov v prípade spojitých náhodných veličín.

Určovanie pravdepodobnosti vychádza z klasickej matematickej metódy, kde pravdepodobnosť, že určitý jav nastane, definujeme ako $P(A) = \frac{m}{n}$, kde m je počet žiadanych výsledkov a n počet možných výsledkov. V niektorých prípadoch je vhodné určenie pravdepodobnosti, že nastane určitý jav, i pomocou frekvencie výskytu žiadanych výsledkov (n_i), ktoré nastali pri realizácii daného javu. Pravdepodobnosť, že nastane určitý jav A , v tomto prípade vypočítame podľa vzorca $P(A) = \frac{n_i}{n}$, kde n predstavuje počet možných výsledkov a n_i sú početnosti nastatia žiadanych výsledkov.

Pri sledovaní náhodných javov sa možno stretnúť s nasledovnými pojmami:

- a) **istý jav** – ide o taký jav, ktorý nastane za každých okolností (označujeme ho U),
- b) **nemožný jav** je taký, ktorý nemôže nastať za žiadnych okolností a pri žiadnom pokuse (označenie $\emptyset$),
- c) **opačný jav k javu A** je taký, ktorý nastane práve vtedy a len vtedy, ak nenastane jav A (označenie $\bar{A}$).

Ak A, B sú dva náhodné javy, potom:

- a) **súčtom (zjednotením)** javov A, B nazývame jav C taký, ktorý nastane práve vtedy a len vtedy, ak nastane aspoň jeden z javov A, B

$$A \cup B = C,$$

- b) **diferenciou (rozdielom)** javov A, B nazveme taký jav C , ktorý nastane práve vtedy a len vtedy, ak nastane A a súčasne nenastane B

$$A - B = C,$$

- c) **súčinom (prienikom)** javov A, B nazývame taký jav C , ktorý nastane práve vtedy a len vtedy, ak súčasne nastanú javy A aj B .

$$A \cap B = C.$$

Vlastnosti náhodných javov:

- | | | |
|--|---|--|
| a) $A \cap \bar{A} = \emptyset$ | h) $A \cap \emptyset = \emptyset$ | o) $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ |
| b) $\bar{\bar{A}} = A$ | i) $A \cap A = A$ | p) $(A \cup B) \cup C = A \cup (B \cup C)$ |
| c) $\bar{\bar{A}} = A$ | j) $A \cup B = B \cup A$ | q) $(A \cap B) \cap C = A \cap (B \cap C)$ |
| d) $A = B \Leftrightarrow \bar{A} = \bar{B}$ | k) $A \cap B = B \cap A$ | r) $\overline{A_1 \cup A_2 \cup \dots \cup A_n} = \bar{A}_1 \cap \bar{A}_2 \dots \cap \bar{A}_n$ |
| e) $A \cup \emptyset = A$ | l) $\overline{A \cap B} = \bar{A} \cup \bar{B}$ | s) $\overline{A_1 \cap A_2 \cap \dots \cap A_n} = \bar{A}_1 \cup \bar{A}_2 \dots \cup \bar{A}_n$ |
| f) $A \cap U = A$ | m) $\overline{A \cup B} = \bar{A} \cap \bar{B}$ | t) $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ |

$$g) A \cup \bar{A} = U \quad n) A - B = A \cap \bar{B}$$

Zlúčiteľné náhodné javy sú také, ktoré môžu za istých okolností nastať súčasne. Ich prienik je rôzny od prázdnej množiny. **Nezlúčiteľné náhodné javy** nenastanú pri žiadnom pokuse. Ak jav A je závislý na jave B , tak pravdepodobnosť nastatia javu A za predpokladu, že jav B už nastal, nazývame **podmienenou pravdepodobnosťou** a označujeme ako $P(A/B)$:

$P(A/B) = \frac{m(A \cap B)}{m(B)}$, kde $m(A \cap B)$ predstavuje počet vyhovujúcich možností pre $A \cap B$ a $m(B)$ je počet priaznivých možností pre B .

Ak A, B sú dva náhodné javy, ktoré sú:

- a) **nezávislé** $\Rightarrow P(A \cap B) = P(A) \cdot P(B)$,
- b) **nezlúčiteľné, ale nezávislé** $\Rightarrow P(A \cup B) = P(A) + P(B) - P(A) \cdot P(B)$,
- c) **závislé** $\Rightarrow P(A \cap B) = P(B) \cdot P(A/B)$, resp. $P(B \cap A) = P(A) \cdot P(B/A)$.

Pravdepodobnosť javu $\bar{A}$ (opačného k javu A) vypočítame ako: $P(\bar{A}) = 1 - P(A)$.
 $P(A) + P(\bar{A}) = P(U) = 1$.

Bayesova veta:

Ak $A_1, A_2, \dots, A_n$ sú nezlúčiteľné javy tvoriace úplný systém javov a ak jav A je jav, ktorý nastane spolu s jedným z javov A_i , potom $P(A_i/A) = \frac{P(A_i) \cdot P(A/A_i)}{P(A)}$, kde $i=1, 2, \dots, k$ a

$$P(A) = \sum_{i=1}^k P(A_i) \cdot P(A/A_i).$$

DISKRÉTNÁ NÁHODNÁ PREMENNÁ

Rozdelenie pravdepodobnosti diskkrétnej náhodnej premennej X je funkcia $P(X)$, ktorá každej diskkrétnej náhodnej veličine X priradí určitú pravdepodobnosť $p(x)$, s ktorou nadobudne túto hodnotu.

$$P(X = x) = p(x).$$

Distribučná funkcia $F(x)$ vyjadruje pravdepodobnosť toho, že

$$P(X < x) = \sum_{i=1}^n p(x_i) \text{ pre } x_i < x.$$

SPOJITÁ NÁHODNÁ PREMENNÁ

Náhodná premenná X sa nazýva spojité, ak existuje taká nezáporná, reálna integrovateľná funkcia $f(x)$, že pre každé $x \in R$ platí:

$$F(x) = P(X < x) = \int_{-\infty}^x f(l) dl.$$

V prípade, že distribučná funkcia $F(x)$ má pre každé x spojitú deriváciu $f(x) = F'(x)$, budeme veličinu X nazývať spojitou náhodnou veličinou, $f(x)$ hustotou pravdepodobnosti spojitej náhodnej veličiny X v bode x a $F(x)$ jej distribučnou funkciou.

Vlastnosti distribučnej funkcie:

1. $F(x)$ je neklesajúca funkcia.
2. Pre každé $x \in R$ platí: $0 \leq F(x) \leq 1$.
3. $F(x)$ je zľava spojitá.
4. $\lim_{x \rightarrow -\infty} F(x) = 0$, $\lim_{x \rightarrow \infty} F(x) = 1$.
5. $P(a \leq X < b) = F(b) - F(a)$, kde $a, b \in R$, $a < b$.

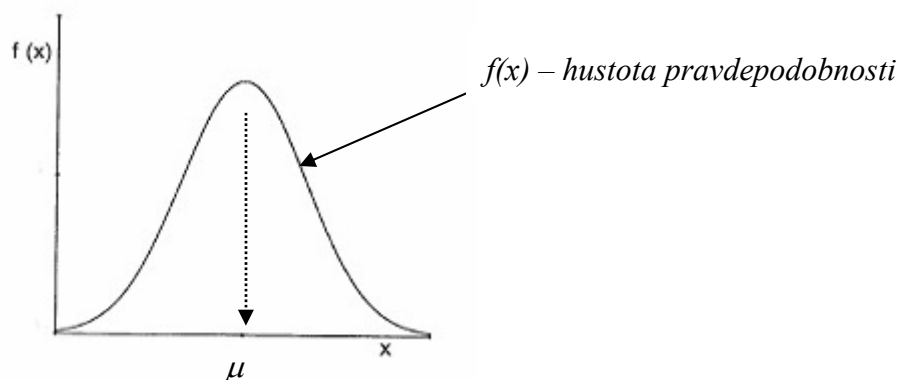
VYBRANÉ ROZDELENIA NÁHODNÝCH PREMENNÝCH

Normálne rozdelenie (Laplace-Gaussovo) je najdôležitejším rozdelením spojitej náhodnej premennej. Možno ho očakávať u takých náhodných premenných, na hodnoty ktorých pôsobí veľké množstvo vzájomne nezávislých vplyvov. Spojitá náhodná premenná X , ktorá má normálne rozdelenie, nadobúda hodnoty z intervalu $(-\infty, \infty)$.

Náhodná premenná X má normálne rozdelenie s parametrami μ a σ^2 , $\mu \in R$, $\sigma > 0$, ak jej hustota pravdepodobnosti má tvar (graf 8):

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, kde x je ľubovoľné reálne číslo, μ je stredná hodnota rozdelenia, σ^2 je rozptyl rozdelenia, e je Eulerovo číslo ($e \approx 2,7182818$) a π je Ludolfovo číslo ($\pi \approx 3,1415927$).

Graf 8 Hustota pravdepodobnosti normálneho rozdelenia



Vlastnosti normálneho rozdelenia:

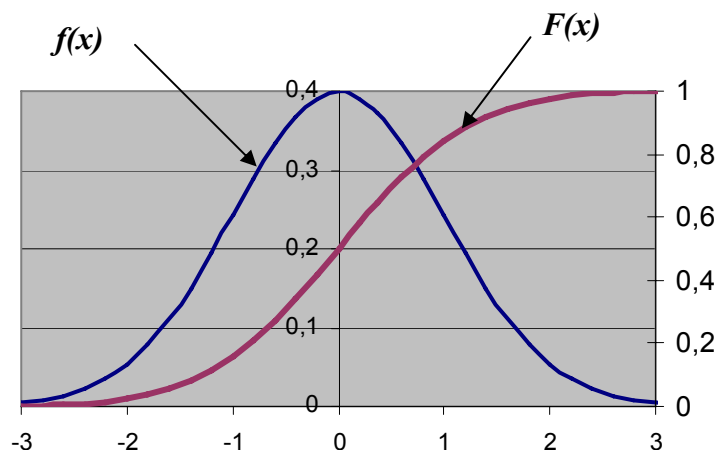
1. Krivka hustoty pravdepodobnosti normálneho rozdelenia má zvonovitý tvar a asymptoticky sa približuje k osi x . Jej definičný obor je $(-\infty, \infty)$. Je súmerná podľa osi rovnobežnej s osou y , ktorá prechádza vrcholom normálneho rozdelenia.
2. Aritmetický priemer pre normálne rozdelenie hustoty pravdepodobnosti sa rovná modusu a ten sa rovná mediánu ($\mu = \bar{x} = \hat{x} = \tilde{x}$).
3. Koeficient šikmosti normálneho rozdelenia $\gamma_1 = 0$, koeficient špicatosti $\gamma_2 = 3$ (resp. 0, ak berieme do úvahy, že 3 odpočítavame).

Distribučná funkcia normálneho rozdelenia je definovaná vzťahom (graf 9):

$$F(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^x e^{-\frac{(l-\mu)^2}{2\sigma^2}} dl.$$

Distribučná funkcia priradí každej hodnote x obsah plochy pod krivkou hustoty normálneho rozdelenia v intervale od $(-\infty, x >$. V prípade, že obsah plochy pod krivkou hustoty pravdepodobnosti je na intervale $(-\infty, x_0 >$ rovný 70 % celej plochy pod krivkou hustoty normálneho rozdelenia na intervale $(-\infty, x >$, tak hodnota distribučnej funkcie $F(x)$ v bode x_0 je rovná 0,7 ($F(x_0) = 0,7$).

Graf 9 Hustota pravdepodobnosti a distribučná funkcia normálneho rozdelenia



Spracované podľa: <http://eldum.phil.muni.cz/mod/resource/view.php?id=1764>

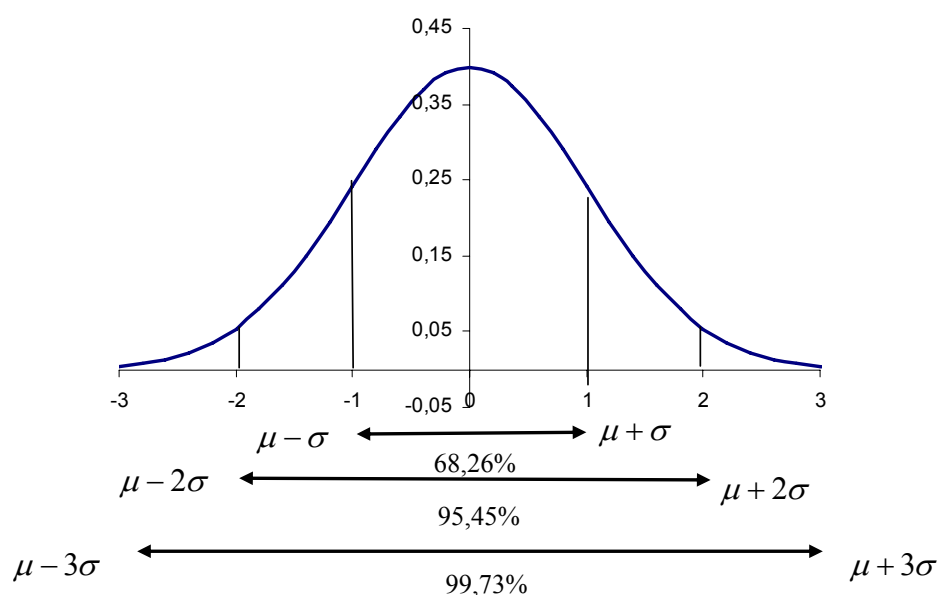
Transformáciou náhodnej premennej x z normálneho rozdelenia na normovanú náhodnú premennú t podľa vzťahu $t = \frac{x - \mu}{\sigma}$ dostaneme **normované normálne rozdelenie** s parametrami $\mu = 0$ a $\sigma^2 = 1$. Hustota pravdepodobnosti normovanej premennej t

s normálnym rozdelením je: $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$, kde $t \in R$. Jej distribučná funkcia má

$$\text{tvar: } F(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^t e^{-\frac{l^2}{2}} dl.$$

Hodnoty distribučnej funkcie $F(x)$ sa určia pomocou distribučnej funkcie $F(t)$. Pre výpočet hodnôt distribučnej funkcie normovanej náhodnej premennej $F(t)$ s normálnym rozdelením platí: $F(-t) = 1 - F(t)$, pre jej hustotu pravdepodobnosti $f(-t) = f(t)$.

Graf 10 Krivka hustoty pravdepodobnosti normovaného normálneho rozdelenia



Spracované podľa:

<http://www.fem.uniag.sk/cvicenia/ksov/obtulovic/BIO%C5%A0TATISTIKA/prednaska3.ppt>.

Ak plochu pod krivkou hustoty normálneho rozdelenia s parametrami (μ, σ^2) považujeme za jednotkovú (rovnú 100%), potom platí, že (graf 10):

a) v intervale $(\mu - \sigma, \mu + \sigma)$ je 68,26% štatistických jednotiek súboru,

$$P(\mu - \sigma < X < \mu + \sigma) = 0,6826,$$

$\Rightarrow$ po transformácii na normálne rozdelenie dostávame

$$P(-1 < t < 1) = 0,6826,$$

b) v intervale $(\mu - 2\sigma, \mu + 2\sigma)$ je približne 95,45% štatistických jednotiek súboru,

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0,9545,$$

$\Rightarrow$ po transformácii na normálne rozdelenie dostávame

$$P(-2 < t < 2) = 0,9545,$$

c) v intervale $(\mu - 3\sigma, \mu + 3\sigma)$ je približne 99,73% štatistických jednotiek súboru

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0,9973,$$

$\Rightarrow$ po transformácii na normálne rozdelenie dostávame

$$P(-2 < t < 2) = 0,9973.$$

Hodnoty distribučnej funkcie normovaného normálneho rozdelenia (ide o obsah plochy pod krivkou hustoty pravdepodobnosti normovaného normálneho rozdelenia) sa v praxi používajú v menšej miere (tab. 14). Obvykle sa využívajú tabuľky kritických hodnôt z_p normovaného normálneho rozdelenia, ktoré pracujú už s vopred určenou pravdepodobnosťou p výskytu sledovaného javu (tab. 15). V prípade, že sa sledovaný jav vyskytuje s pravdepodobnosťou napr. 0,65, znamená to, že 65% plochy pod krivkou hustoty normovaného normálneho rozdelenia sa nachádza po kritickú hodnotu $z_{0,65} = 0,385$. Pravdepodobnosť, že premenná z nadobudne hodnotu menšiu ako 0,385 je 65%: $P(z < z_{0,65}) = P(z < 0,385) = 0,65$. Ak sú hodnoty $p < 0,5$, potom $z_p = -z_{1-p}$.

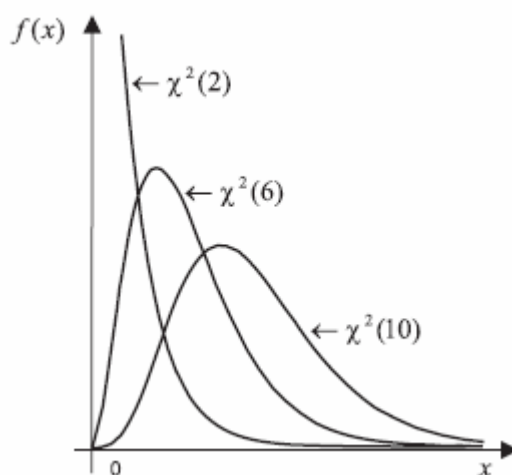
Pre rozdelenie spojitých náhodných premenných možno využiť aj χ^2 (**Pearsonovo**) **rozdelenie**. Ak $X_1, X_2, \dots, X_n$ sú nezávislé náhodné premenné, z ktorých každá má normované normálne rozdelenie s parametrami $\mu = 0$ a $\sigma^2 = 1$, potom možno vytvoriť náhodnú premennú $\chi^2 = X_1^2 + X_2^2 + \dots + X_n^2$. Uvedené rozdelenie spojitých náhodných premenných nazývame χ^2 rozdelením s n stupňami voľnosti².

Stredná hodnota χ^2 rozdelenia je rovná počtu stupňov voľnosti n , rozptyl jej dvojnásobnej hodnote ($2n$). Koeficient šikmosti $\gamma_1 = \frac{8}{n}$, koeficient špicatosti $\gamma_2 = \frac{12}{n}$. Nakoľko n je vždy kladné číslo, koeficienty šikmosti a špicatosti pre χ^2 rozdelenie sú vždy kladné, a teda graf hustoty χ^2 rozdelenia je vždy zošikmený doľava (ľavostranne asymetrické rozdelenie). Pri menších hodnotách n je špicatosť výraznejšia, s rastúcim počtom stupňov voľnosti je krivka hustoty χ^2 rozdelenia plochejšia (graf 11).

Podobne ako pri normovanom normálnom rozdelení, aj u χ^2 rozdelenia sa vo väčšej miere namiesto hodnôt distribučnej funkcie využívajú tabuľky kritických hodnôt χ^2 rozdelenia s vopred určenou pravdepodobnosťou pre n stupňov voľnosti (tab. 16).

² Parameter n udáva počet nezávislých náhodných premenných X_i , z ktorých premenná χ^2 vznikla, preto ho nazývame počet stupňov voľnosti.

Graf 11 Hustota pravdepodobnosti χ^2 rozdelenia pre 2, 6 a 10 stupňov voľnosti



Spracované podľa: <http://fstroj.uniza.sk/kam/pobocikova/pdf/TEXT12.PDF>

Studentovo rozdelenie (t -rozdelenie). Majme náhodnú premennú X , ktorá má normované normálne rozdelenie $N(0,1)$ a náhodnú premennú χ^2 , ktorá má χ^2 rozdelenie s n stupňami voľnosti (tab. 17). Ak náhodné premenné X a χ^2 sú navzájom nezávislé, potom možno vytvoriť náhodnú premennú $t = \frac{X}{\sqrt{\frac{\chi^2}{n}}}$. Uvedené rozdelenie nazývame Studentovo alebo

t -rozdelenie. Krivka hustoty pravdepodobnosti Studentovho rozdelenia je plochejšia ako u normálneho rozdelenia, avšak s rastúcou hodnotou n sa jej špicatosť zväčšuje. Graf hustoty pravdepodobnosti uvedeného rozdelenia je jednovrcholový, súmerný okolo strednej hodnoty $E(t)$.

Pre diskkrétne náhodné veličiny možno použiť **binomické rozdelenie pravdepodobnosti**, ktorého predmetom skúmania je alternatívny jav s alternatívami A a B . V prípade, že jav A nastane s pravdepodobnosťou p a jav B s pravdepodobnosťou $q=1-p$, tak pravdepodobnosť, že jav A nastane x -krát pri vykonaní n nezávislých pokusov bude:

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

Uvedený vzťah nazývame Bernoulliho schéma. Distribučnú funkciu binomického rozdelenia určuje funkcia $F(x) = \sum_{x_i < x} \binom{n}{x_i} p^{x_i} \cdot q^{n-x_i}$, kde $x_i = 0, 1, 2, \dots, n$.

Binomické rozdelenie pravdepodobnosti výskytu javu má tvar paličkového grafu. Závisí od počtu pokusov n a hodnoty p . Čím je počet pokusov väčší, tým je binomické rozdelenie súmernejšie.

Opisné charakteristiky binomického rozdelenia:

stredná hodnota $E(X)=n.p$, disperzia (rozptyl) $D(X)=n.p.q$, koeficient šikmosti $\gamma_1 = \frac{q-p}{\sqrt{n.p.q}}$, koeficient špicatosti $\gamma_2 = \frac{1-6.p.q}{n.p.q}$.

V prípade, že máme n nezávislých pokusov a pravdepodobnosť, že nastane jav v jednom pokuse, je malá $p \leq 0,1$ a naopak, počet pokusov je veľký ($n > 30$), namiesto binomického rozdelenia využívame **Poissonovo rozdelenie** (tab. 18). Diskrétna náhodná veličina X má Poissonovo rozdelenie, ak pravdepodobnostná funkcia je tvaru $P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda}$, kde $\lambda = n.p$. Parameter Poissonovho rozdelenia λ nadobúda hodnoty z intervalu $(0, \infty)$. Grafickým znázornením Poissonovho rozdelenia pravdepodobnosti je paličkový graf, ktorý je pre väčšie λ hodnoty súmernejší a plochejší.

Opisné charakteristiky Poissonovho rozdelenia: stredná hodnota $E(X)=\lambda$, disperzia $D(X)=\lambda$, koeficient šikmosti $\gamma_1 = \frac{1}{\sqrt{\lambda}}$, koeficient špicatosti $\gamma_2 = \frac{1}{\lambda}$.

Riešené úlohy:

1. V tabuľke sú uvedené informácie o priemernom úhrne zrážok (v mm) za obdobie 1931-1960 podľa klimateckej stanice Michalovce (123 m n. m.). Pomocou normovanej premennej vypočítajte hodnotu koeficientu šikmosti a koeficientu špicatosti. Na ich základe posúďte asymetriu a špicatosť rozdelenia. Medzivýsledky zaokrúhľujte na tretie desatinné miesto, konečné výsledky pre šikmosť a špicatosť na dve desatinné miesta, porovnajte ich s normálnym rozdelením.

Obdobie	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.
Michalovce	41	38	29	38	56	79	67	67	63	52	55	52

Spracované podľa: SHMÚ Košice

Riešenie:

Nakoľko šikmosť možno vypočítať ako tretí centrálny moment normovanej premennej a špicatosť ako štvrtý centrálny moment normovanej premennej zmenšený o hodnotu 3, kde

$$t_i = \frac{x_i - \bar{x}}{s_x}, \text{ výpočet bude smerovať najprv k určeniu hodnoty aritmetického priemeru}$$

a smerodajnej odchýlky už pomocou známych vzorcov z kapitol 4 a 6. Pre výpočet uvedených charakteristík využijeme nasledovnú tabuľku:

x_i	n_i	$x_i n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$	$t_i = \frac{x_i - \bar{x}}{s_x}$	$t_i^3 \cdot n_i$	$t_i^4 \cdot n_i$
29	1	29	-24,083	579,991	579,991	-1,720	-5,088	8,752
38	2	76	-15,083	227,497	454,994	-1,077	-2,498	2,691
41	1	41	-12,083	145,999	145,999	-0,863	-0,643	0,555
52	2	104	-1,083	1,173	2,346	-0,077	-0,001	0,000
55	1	55	1,917	3,675	3,675	0,137	0,003	0,000
56	1	56	2,917	8,509	8,509	0,208	0,009	0,002
63	1	63	9,917	98,347	98,347	0,708	0,355	0,251
67	2	134	13,917	193,683	387,366	0,994	1,964	1,952
79	1	79	25,917	671,691	671,691	1,851	6,342	11,739
x	$\Sigma=12$	x	x	x	$\Sigma=2352,918$	x	$\Sigma=0,443$	$\Sigma=25,942$

Poznámka: x – neuvedený údaj.

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{n} = \frac{637}{12} \approx 53,083, \quad s_x^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i}{n} = \frac{2352,918}{12} \approx 196,077, \quad s_x = 14,003.$$

Koeficient šikmosti:

$$\gamma_1 = \frac{1}{n} \sum_{i=1}^k t_i^3 \cdot n_i = \frac{0,443}{12} \approx 0,04$$

Koeficient špicatosti:

$$\gamma_2 = \left(\frac{1}{n} \sum_{i=1}^k t_i^4 \cdot n_i \right) - 3 = \frac{25,942}{12} - 3 \approx -0,84.$$

Odpoveď: Koeficient šikmosti $\gamma_1 > 0$, teda ide o ľavostranne asymetrické rozdelenie početností. Koeficient špicatosti $\gamma_2 < 0$, čo znamená, že rozdelenie početností je plochejšie ako u normálneho rozdelenia.

2. Vypočítajte Pearsonovu a kvartilovú mieru šikmosti z riešeného príkladu 2 v kapitole 6. Výsledky interpretujte.

Riešenie:

Pre výpočet Pearsonovej a kvartilovej miery šikmosti využijeme už vypočítané hodnoty aritmetického priemeru $\bar{x} = 74,69$ a smerodajnej odchýlky $s_x = 12,87$. Ďalšie hodnoty ako modus, medián, dolný a horný kvartil si dopočítame.

x_i	n_i	N_i	f_i	F_i
(50,60>	5	5	0,156	0,156
(60,70>	7	12	0,219	0,375
(70,80>	8	20	0,250	0,625
(80,90>	9	29	0,281	0,906
(90,100>	2	31	0,63	0,969
(100,110>	1	32	0,031	1

$$\hat{x} = s_{Mo} - \frac{h_{Mo}}{2} \cdot \frac{n_{Mo-1} - n_{Mo+1}}{n_{Mo-1} + n_{Mo+1} + 2n_{Mo}} = \frac{80+90}{2} - \frac{10}{2} \cdot \frac{8-2}{8+2+2 \cdot 9} = 85 - 5 \cdot \frac{6}{28} \approx 83,93$$

$$\tilde{x} = a + h \cdot \frac{0,5 - \sum_{i=1}^{r-1} f_i}{f_{\tilde{x}}} = 70 + 10 \cdot \frac{0,5 - 0,375}{0,250} = 70 + 10 \cdot \frac{0,125}{0,250} = 75$$

$$Q_1^4 = a + h \cdot \frac{0,25 - \sum_{i=1}^{r-1} f_i}{f_{Q_1^4}} = 60 + 10 \cdot \frac{0,25 - 0,156}{0,219} \approx 64,29$$

$$Q_3^4 = a + h \cdot \frac{0,75 - \sum_{i=1}^{r-1} f_i}{f_{Q_3^4}} = 80 + 10 \cdot \frac{0,75 - 0,625}{0,281} \approx 84,45$$

Pearsonova miera šikmosti: $\check{S}_p = \frac{\bar{x} - \hat{x}}{s_x} = \frac{74,69 - 83,93}{12,87} \approx -0,718$

Kvartilová miera šikmosti:

$$\check{S}_K = \frac{(Q_3^4 - \tilde{x}) - (\tilde{x} - Q_1^4)}{(Q_3^4 - \tilde{x}) + (\tilde{x} - Q_1^4)} = \frac{(84,45 - 75) - (75 - 64,29)}{(84,45 - 64,29)} = \frac{9,45 - 10,71}{20,16} = \frac{-1,26}{20,16} \approx -0,063$$

Odpoveď: Pearsonova i kvartilová miera šikmosti nadobúdajú zápornú hodnotu, teda ide o pravostranne asymetrické rozdelenie početností.

3. V tabuľke sú uvedené informácie o počte detí, intervale, v ktorom sa pohybuje výška ich mesačného vreckového a o celkovom počte EUR, ktoré dostanú mesačne deti v každej intervalovej skupine. Vypočítajte koeficient koncentrácie a pomer koncentrácie. Načrtnite Lorenzovu krivku a výsledky interpretujte.

<i>Vreckové v EUR</i>	<i>Počet detí</i>	<i>Celková výška mesačného vreckového pre daný interval</i>
(5-7>	5	30
(7-9>	10	84
(9-11>	5	50
(11-13>	4	47
(13-15>	3	44

Riešenie:

x_i	n_i	C	f_i	F_i	z_i	Z_i	$\frac{(Z_{i-1} + Z_i)}{2}$	$\frac{(Z_{i-1} + Z_i)}{2} \cdot f_i$
(5-7>	5	30	0,185	0,185	0,118	0,118	0,059	0,0109
(7-9>	10	84	0,371	0,556	0,329	0,447	0,2825	0,1048
(9-11>	5	50	0,185	0,741	0,196	0,643	0,545	0,1008
(11-13>	4	47	0,148	0,889	0,184	0,827	0,735	0,1088
(13-15>	3	44	0,111	1	0,173	1	0,9135	0,1014
x	$\Sigma=27$	$\Sigma=255$	X	x	x	x	x	$\Sigma=0,4267=S$

Vysvetlivky: x_i - vreckové v EUR, n_i - počet detí, C - celková výška mesačného vreckového pre daný interval, f_i - relatívna početnosť detí, F_i - kumulatívna relatívna početnosť detí, z_i - podiel danej skupiny (výšky vreckového v danom intervale) na celkovej mesačnej výške vreckového, Z_i - kumulatívny podiel danej skupiny (výšky vreckového v danom intervale) na celkovej mesačnej výške vreckového

Koeficient koncentrácie:

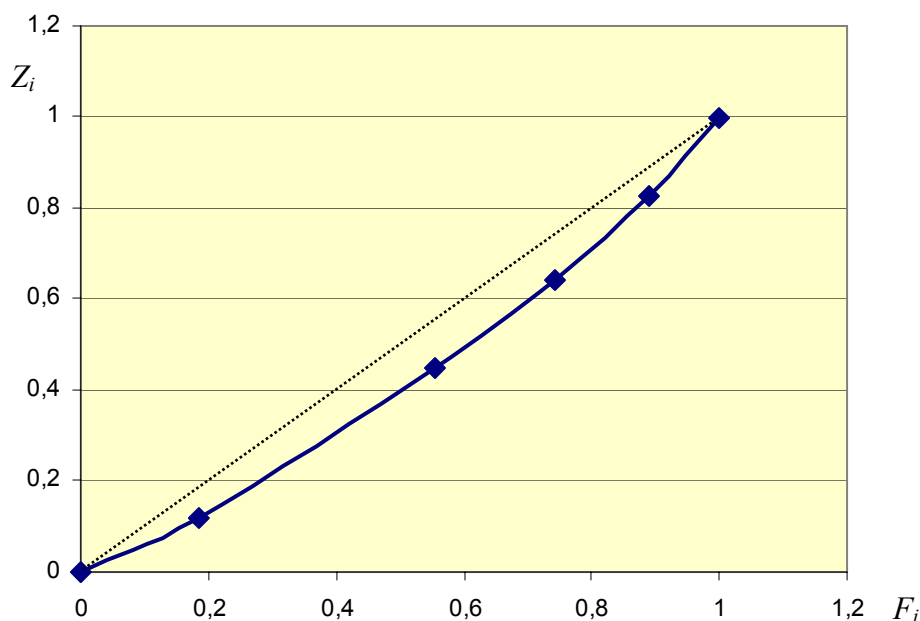
$$K_K = 1 - 2S = 1 - 0,8534 = 0,1466.$$

Pomer koncentrácie:

$$P_K = \frac{\sum_{i=1}^{k-1} (F_i - Z_i)}{\sum_{i=1}^{k-1} F_i} = \frac{(0,185 - 0,118) + (0,556 - 0,447) + (0,741 - 0,643) + (0,889 - 0,827)}{0,185 + 0,556 + 0,741 + 0,889} =$$

$$= \frac{0,067 + 0,109 + 0,098 + 0,062}{2,371} = \frac{0,336}{2,371} \approx 0,1417.$$

Lorenzova krivka:



Odpoveď: Koeficient koncentrácie má hodnotu 14,66%, pomer koncentrácie je rovný 0,1417. Z uvedeného možno usúdiť, že sledovaný jav sa vyznačuje len slabou koncentráciou. Koncentráciu možno interpretovať aj na základe zostrojenej Lorenzovej krivky. Ako si môžeme v grafe všimnúť, hodnoty sledovaného znaku sú rozdelené pomedzi jednotky približne rovnomerne. Plocha medzi Lorenzovou krivkou a diagonálou je pomerne malá, a teda sledovaný jav je slabو koncentrovaný.

4. Hádžeme kockou. Aká je pravdepodobnosť, že

- a) pri jednom hode padne 2,
- b) pri dvoch po sebe nasledujúcich hodoch padnú 6,
- c) pri dvoch po sebe nasledujúcich hodoch padne súčet 8.

Riešenie:

Pravdepodobnosť, že nastane uvedený jav, vypočítame ako podiel počtu žiadanych výsledkov m k počtu možných výsledkov n .

a) Pri jednom hode padne 2: $P(A) = \frac{m}{n} = \frac{1}{6}$.

b) Pri dvoch hodoch padne po sebe 6: $P(A).P(B) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$

$\Rightarrow$ pravdepodobnosť padnutia 6 pri prvom hode kockou je: $P(A) = \frac{m}{n} = \frac{1}{6}$,

$\Rightarrow$ pravdepodobnosť padnutia 6 pri druhom hode kockou je: $P(B) = \frac{m}{n} = \frac{1}{6}$.

c) Pri dvoch hodoch padne súčet 8: $P = \frac{m}{n} = \frac{5}{36}$,

$\Rightarrow m$ - počet priaznivých možností: 2 6

6 2

3 5

5 3

4 4

$\Rightarrow n$ - počet všetkých možností: je $6 \cdot 6 = 36$.

5. Zistite, či platí:

$$(A \cup B) - B = A - (A \cap B).$$

Riešenie:

$$(A \cup B) - B = A - (A \cap B)$$

$$(A \cup B) \cap \bar{B} = A \cap (\bar{A} \cup \bar{B})$$

$$(A \cap \bar{B}) \cup (B \cap \bar{B}) = (A \cap \bar{A}) \cup (A \cap \bar{B})$$

$$(A \cap \bar{B}) = (A \cap \bar{B})$$

$$A - B = A - B$$

6. V 3 hranatých akváriách plávajú po 3 zlaté rybičky a 1 biela. V 2 okrúhlych akváriách po 1 zlatej rybke a 2 bielych. Aká je pravdepodobnosť, že zlatá rybka bola vylovená z hranatého akvária a aká z okrúhleho?

Riešenie:

Riešenie uvedeného príkladu smeruje k výpočtu podmienenej pravdepodobnosti:

$$P(A_i / A) = \frac{P(A_i) \cdot P(A / A_i)}{P(A)}, \text{ kde } P(A) = \sum_{i=1}^2 P(A_i) \cdot P(A / A_i).$$

Označenie:

$P(A)$ – pravdepodobnosť vytiahnutia zlatej rybky

$P(A_1)$ – pravdepodobnosť výberu hranatého akvária,

$P(A_2)$ – pravdepodobnosť výberu okrúhleho akvária,

$P(A/A_1)$ – vybrali sme si hranaté akvárium, zaujíma nás pravdepodobnosť vytiahnutia zlatej rybky z tohto akvária,

$P(A/A_2)$ – vybrali sme si okrúhle akvárium, zaujíma nás pravdepodobnosť vytiahnutia zlatej rybky z tohto akvária.

$$P(A_1) = \frac{3}{5}, \quad P(A_2) = \frac{2}{5}, \quad P(A/A_1) = \frac{3}{4}, \quad P(A/A_2) = \frac{1}{3},$$

$$P(A) = \sum_{i=1}^2 P(A_i)P(A/A_i) = P(A_1).P(A/A_1) + P(A_2).P(A/A_2) = \frac{3}{5} \cdot \frac{3}{4} + \frac{2}{5} \cdot \frac{1}{3} = \frac{9}{20} + \frac{2}{15} = \frac{27+8}{60} = \frac{35}{60} = \frac{7}{12}$$

$$P(A_1/A) = \frac{P(A_1).P(A/A_1)}{P(A)} = \frac{\frac{3}{5} \cdot \frac{3}{4}}{\frac{7}{12}} = \frac{\frac{9}{20}}{\frac{7}{12}} = \frac{27}{35}$$

$$P(A_2/A) = \frac{P(A_2).P(A/A_2)}{P(A)} = \frac{\frac{2}{5} \cdot \frac{1}{3}}{\frac{7}{12}} = \frac{\frac{2}{15}}{\frac{7}{12}} = \frac{8}{35}$$

Poznámka: V uvedenej úlohe stačilo vypočítať pravdepodobnosť vytiahnutia zlatej rybky len z hranatého akvária, potom od 1 odpočítať hodnotu $\frac{27}{35}$, čím by sme dostali pravdepodobnosť vytiahnutia zlatej rybky z okrúhleho akvária.

Odpoveď: Pravdepodobnosť vytiahnutia zlatej rybky z hranatého akvária je $\frac{27}{35}$, z okrúhleho akvária $\frac{8}{35}$.

7. V škatuli je 10 kariet, z toho dve esá. Zistite pravdepodobnosť, s akou možno vytiahnuť eso pri troch náhodných ťahoch (esá do škatule po ťahu vraciame). Zostrojte graf rozdelenia pravdepodobnosti a graf distribučnej funkcie.

Riešenie:

Označme si A náhodný jav – vytiahnutie esá, x - jednotlivé hodnoty, ktoré môže premenná A nadobúdať, t. j. z troch ťahov môže padnúť eso 0-krát, 1-krát, 2-krát alebo 3-krát. Pre výpočet hľadanej pravdepodobnosti využijeme binomické rozdelenie pravdepodobnosti:

$$P(A = x) = \binom{n}{x} p^x \cdot q^{n-x} = \frac{n!}{x!(n-x)!} p^x \cdot q^{n-x},$$

kde p predstavuje pravdepodobnosť, s akou nastane žiadany jav (vytiahnutie esa):

$P(A) = \frac{2}{10}$ a q pravdepodobnosť, že nenastane žiadany jav (t. j., že eso nevytiahnem):

$$P(A) = \frac{8}{10}.$$

a) Riešime pravdepodobnosť, že eso nevytiahneme ani raz ($x=0$):

$$P(A=0) = \binom{3}{0} p^0 \cdot q^3 = \binom{3}{0} \left(\frac{2}{10}\right)^0 \cdot \left(\frac{8}{10}\right)^3 = \frac{3!}{0!(3-0)!} \cdot 1 \cdot \frac{512}{1000} = 0,512.$$

b) Riešime pravdepodobnosť, že eso vytiahneme raz ($x=1$):

$$P(A=1) = \binom{3}{1} p^1 \cdot q^2 = \binom{3}{1} \left(\frac{2}{10}\right)^1 \cdot \left(\frac{8}{10}\right)^2 = \frac{3!}{1!(3-1)!} \cdot \frac{2}{10} \cdot \frac{64}{100} = 0,384.$$

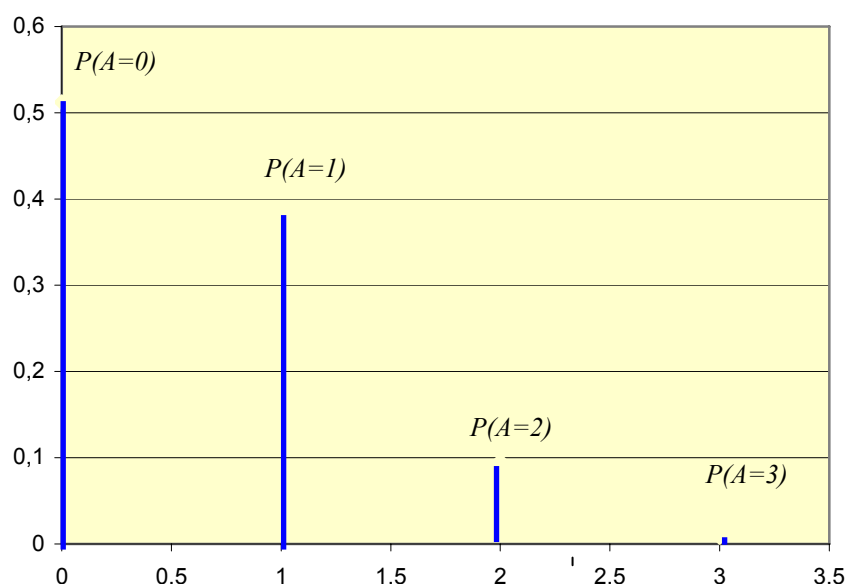
c) Riešime pravdepodobnosť, že eso vytiahneme 2-krát ($x=2$):

$$P(A=2) = \binom{3}{2} p^2 \cdot q^1 = \binom{3}{2} \left(\frac{2}{10}\right)^2 \cdot \left(\frac{8}{10}\right)^1 = \frac{3!}{2!(3-2)!} \cdot \frac{4}{100} \cdot \frac{8}{10} = 0,096.$$

d) Riešime pravdepodobnosť, že eso vytiahneme 3-krát ($x=3$).

$$P(A=3) = \binom{3}{3} p^3 \cdot q^0 = \binom{3}{3} \left(\frac{2}{10}\right)^3 \cdot \left(\frac{8}{10}\right)^0 = \frac{3!}{3!(3-3)!} \cdot \frac{8}{1000} \cdot 1 = 0,008.$$

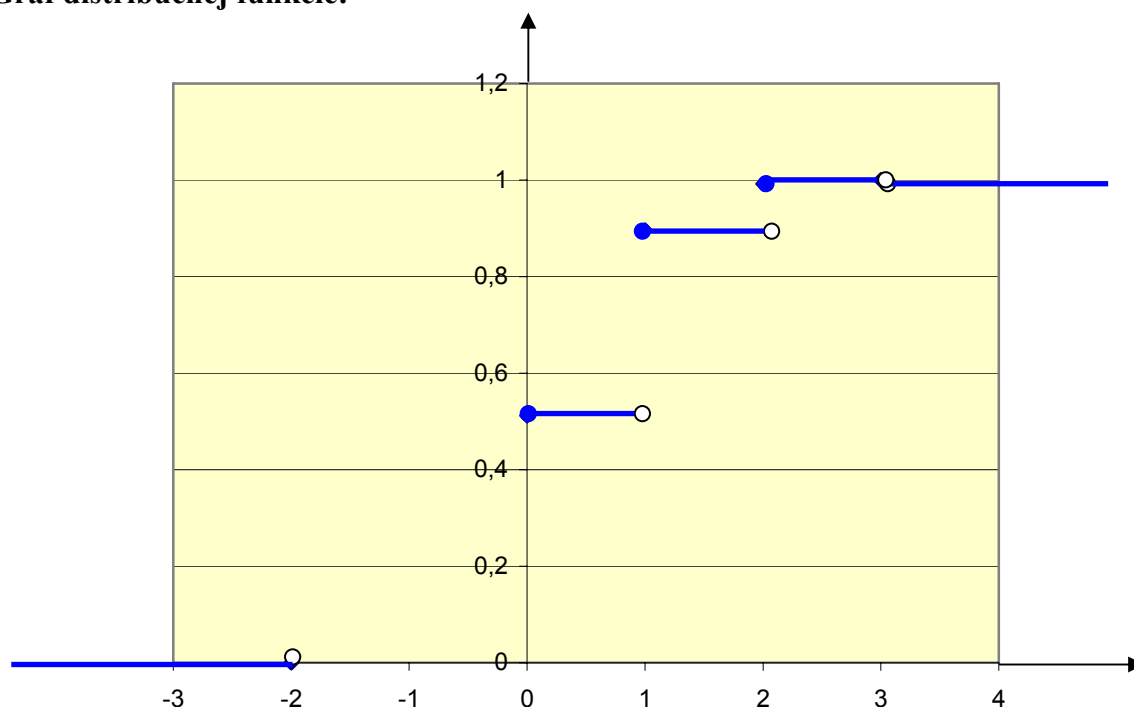
Graf rozdelenia pravdepodobnosti:



Nakoľko distribučná funkcia je definovaná pre všetky reálne čísla, musíme preskúmať nasledovné možnosti:

- a) V prípade, že x nadobúda hodnoty z intervalu $(-\infty, 0)$, pravdepodobnosť, že pri troch ťahoch vytiahneme eso menej ako 0-krát je 0 (nemôžem ho ťahať -1, resp. -2 krát pod.).
- b) Ak by x nadobúdalo hodnoty z intervalu $[0, 1)$, tak jediná reálna možnosť je tá, že z troch ťahov by sme eso nevytiahli ani raz, teda pravdepodobnosť $P(A=0)=0,512$. Distribučnú funkciu na uvedenom intervale by sme definovali nasledovne:
 $F(x) = P(A < 1) = P(A = 0) = 0,512$.
- c) Pre x patriace do intervalu $[1, 2)$ by už pripadali do úvahy dve možnosti, a to, že eso nevytiahneme ani raz alebo ho vytiahneme práve raz, čo možno zapísať v tvare:
 $F(x) = P(A < 2) = P(A = 0) + P(A = 1) = 0,512 + 0,384 = 0,896$.
- d) Pre x patriace do intervalu $[2, 3)$ môžu nastať až 3 možnosti. Eso nemusíme vytiahnuť vôbec, môžeme ho vytiahnuť raz alebo 2-krát:
 $F(x) = P(A < 3) = P(A = 0) + P(A = 1) + P(A = 2) = 0,896 + 0,096 = 0,992$.
- e) Posledným intervalom, ktorý je potrebné preskúmať, je interval $[3, \infty)$. Pre x patriace do tohto intervalu môžu nastať až tri možnosti: eso nevytiahneme ani raz, eso vytiahneme raz, 2-krát alebo 3-krát. Distribučná funkcia pre uvedený interval má tvar:
 $F(x) = P(A < \infty) = P(A = 0) + P(A = 1) + P(A = 2) + P(A = 3) = 0,992 + 0,008 = 1,000$.

Graf distribučnej funkcie:



7. Vo výrobnnej prevádzke pripadajú na 1000 výrobkov 4 chybné výrobky. Určte pravdepodobnosť, že pri 100 náhodne vybraných výrobkoch bude najviac 1 chybný.

Riešenie:

Nakoľko pravdepodobnosť, že jav nastane v jednom pokuse je veľmi malá ($p = \frac{4}{1000} = 0,004$) a zároveň počet pokusov je veľký ($n=1000$), využijeme Poissonove

rozdelenie pravdepodobnosti $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$, kde $\lambda = n \cdot p$. Označme si x jednotlivé hodnoty, ktoré môže jav X nadobúdať. Pravdepodobnosť, že pri náhodne vybraných 100 výrobkoch bude najviac 1 chybný, možno zapísať v tvare: $P(X < 1) = P(X = 0) + P(X = 1)$.

$$\lambda = n \cdot p = 100 \cdot 0,004 = 0,4.$$

$$P(X = 0) = \frac{0,4^0}{0!} e^{-0,4} \approx 0,670.$$

$$P(X = 1) = \frac{0,4^1}{1!} e^{-0,4} \approx 0,268.$$

$$P(X < 1) = P(X = 0) + P(X = 1) = 0,607 + 0,268 = 0,875.$$

Odpoveď: Pravdepodobnosť, že pri 100 náhodne vybraných výrobkoch bude 1 chybný, je 0,875.

8. Predpokladajme, že priemerný počet dní bez slnečného svitu v jednotlivých mesiacoch má v určitej klimatickej stanici normálne rozdelenie s parametrami $\mu = 130$ a smerodajnou odchýlkou $\sigma = 9$. Určte pravdepodobnosť, že v náhodne zvolenom mesiaci bude

a) menej ako 140 slnečných dní,

b) viac než 135 slnečných dní.

Riešenie:

Náhodnú premennú x_0 si prevedieme na normovanú premennú t_0 :

$$a) t_0 = \frac{x_0 - \mu}{\sigma} = \frac{140 - 130}{9} = \frac{10}{9} \approx 1,11.$$

$$F(x_0) = F(t_0), P(x < x_0) = P(t < t_0).$$

$$P(x < 140) = P(t < 1,11) = 0,86650.$$

Odpoveď: Pravdepodobnosť, že v náhodne zvolenom mesiaci bude menej ako 140 slnečných dní je 86,65 %.

$$b) t_0 = \frac{x_0 - \mu}{\sigma} = \frac{135 - 130}{9} = \frac{5}{9} \approx 0,56.$$

$$F(x_0) = F(t_0), P(x > x_0) = P(t > t_0).$$

$$P(x > 135) = P(t > 0,56) = 1 - P(t < 0,56) = 1 - 0,71226 = 0,28774.$$

Odpoveď: Pravdepodobnosť, že v náhodne zvolenom mesiaci bude viac ako 135 slnečných dní, je 28,774 %.

9. Predpokladajme, že priemerná júlová teplota vo vybranej klimatickej stanici má normálne rozdelenie so strednou hodnotou $\mu = 15,5$ a smerodajnou odchýlkou $\sigma = 2$. Vypočítajte, aká je pravdepodobnosť, že pri náhodne vybranom dni bude v danom období kolísť jeho teplota v rozpätí 12,5-18,5 °C?

Riešenie:

Najprv si prevedieme transformáciu náhodnej premennej x na normovanú náhodnú premennú t podľa vzťahu: $t = \frac{x - \mu}{\sigma}$. Nakoľko premenná x môže nadobúdať hodnoty z intervalu $\langle 12,5; 18,5 \rangle$, vypočítame normovanú premennú t pre dolnú a hornú hranicu intervalu.

Normovaná premenná t pre dolnú hranicu intervalu:

$$t_d = \frac{12,5 - 15,5}{2} = -1,5 \quad \text{Normovaná premenná } t \text{ pre hornú hranicu intervalu:}$$

$$t_h = \frac{18,5 - 15,5}{2} = 1,5.$$

Z tabuľky distribučnej funkcie následne určíme hodnotu distribučnej funkcie v bode 2: $F(t_h) = F(1,5) = 0,933$. Nakoľko platí, že $F(-t_d) = 1 - F(t_d)$, hodnota distribučnej funkcie v bode -1,5 sa bude rovnať: $F(-1,5) = 1 - F(1,5) = 1 - 0,933 = 0,067$.

Vzhľadom na to, že pravdepodobnosť $P(12,5 \leq x \leq 18,5) = F(1,5) - F(-1,5) = 0,933 - 0,067 = 0,866$.

Odpoveď: Pravdepodobnosť, že pri náhodne vybranom júlovom dni bude priemerná teplota vzduchu v požadovanom intervale, je 86,6%.

Neriešené úlohy:

1. V tabuľke sú uvedené informácie o žiakoch, ktorí navštevujú záujmové krúžky. Pomocou normovanej premennej vypočítajte hodnotu koeficientu šikmosti a koeficientu špicatosti. Na ich základe posúďte asymetriu a špicatosť rozdelenia. Medzivýsledky zaokrúhľujte na tretie desatinné miesto, konečné výsledky pre šikmosť a špicatosť na dve desatinné miesta.

Počet krížkov x_i	Počet žiakov n_i
1	8
2	12
3	4
4	3

2. Z nasledujúcej tabuľky vypočítajte Pearsonovu a kvartilovú mieru šikmosti, určte, o aký typ rozdelenia početnosti ide.

x_i	n_i
(5-9>	2
(10-14>	3
(15-19>	8
(20-24>	5
(25-29>	1

3. V tabuľke sú uvedené informácie o počte študujúcich v zahraničí v jednotlivých obciach. Vypočítajte koeficient koncentrácie a pomer koncentrácie. Načrtnite Lorenzovu krivku a výsledky interpretujte.

Počet študujúcich	Počet obcí	Počet študujúcich v jednotlivých obciach
1-4	5	17
5-9	3	22
10-14	6	64
15-19	4	66
20-24	3	67

4. Zistite, či nasledovný jav je nemožný: $(A \cup B) \cap (\bar{A} \cup B) \cap (A \cup \bar{B}) \cap (\bar{A} \cup \bar{B})$.

5. Hra pozostáva z troch kôl. Hráč absolvuje všetky tri kolá bez ohľadu na to, či v niektrom kole vyhral alebo nie. Pravdepodobnosť výhry v prvom kole hry je $\frac{1}{5}$, v druhom kole $\frac{1}{4}$ a v treťom kole $\frac{2}{7}$. Určte pravdepodobnosť, že hráč v treťom kole nevyhral, ak vyhral celkovo dve kolá hry.

6. V klobúku je 5 bielych a 3 čierne guľičky. Zistite pravdepodobnosť, s akou vytiahnem z klobúka bielu guľičku, ak ťahám guľičky z klobúka dvakrát (Vytiahnutá guľička sa pri druhom ťahu vracia späť do klobúka).

7. Uvažujme rodiny s tromi deťmi. Nech náhodná veličina X znamená počet chlapcov v rodine. Určte rozdelenie pravdepodobnosti a distribučnú funkciu, ak viete, že na 100 000 narodených detí pripadá 49 000 chlapcov.

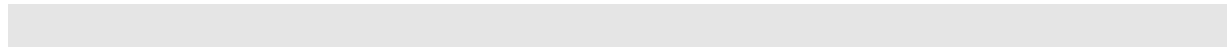
8. Vo vybranom regióne žije 1500 žien europidnej rasy a 3 ženy negroidnej rasy. Určte pravdepodobnosť, že pri náhodnom výbere 80 žien bude medzi nimi najviac 1 žena negroidnej rasy.

9. Predpokladajme, že výška príjmov má vo vybranej skupine študentov normálne rozdelenie so strednou hodnotou $\sigma = 110$ EUR a smerodajnou odcýlkou $\mu = 67$ EUR. Určte pravdepodobnosť, že náhodne vybraný študent má:

a) menej ako 50 EUR,

b) viac ako 120 EUR,

c) od 80 do 150 EUR.



8 Časové rady a miery rastu

Štatistický rad predstavuje určitú postupnosť hodnôt znaku, ktoré môžu byť určitým spôsobom usporiadané. Ak je toto usporiadanie realizované na základe časového výskytu hodnôt znaku, hovoríme o **časových radoch**.

Časové rady môžu byť vytvorené priamo zistiteľnými znakmi (absolútnymi ukazovateľmi) alebo odvodenými. V prípade, že je časový rad tvorený absolútnymi veličinami viažúcimi sa k určitému časovému okamžiku, hovoríme o **okamžikových časových radoch**. Ak sú absolútne veličiny vzťahované na určitý časový interval (obdobie), ide o **intervalové časové rady**. Časové rady vytvorené z odvodených ukazovateľov môžu byť buď **časovými radmi priemerných veličín** (klzávých priemerov) alebo **relatívnych pomerných veličín** (napr. radov indexov, hektárových výnosov).

OKAMŽIKOVÉ ČASOVÉ RADY

Vlastnosti okamžikových časových radov:

- sú spojité v čase,
- hodnota ukazovateľa je len funkciou času, nezávisí od dĺžky intervalu,
- okamžité ukazovatele za niekoľko po sebe idúcich časových intervalov nesčítavame,
- charakteristickou okamžikovou veličinou je **chronologický priemer** vyjadrujúci priemer sledovaného javu v čase:

a) v prípade, že je dĺžka časových intervalov (t_{i-1}, t_i) konštantná, potom

$$\bar{x}_{ch} = \frac{\frac{x_1 + x_2}{2} + \frac{x_2 + x_3}{2} + \dots + \frac{x_{n-1} + x_n}{2}}{n-1} = \frac{\frac{x_1}{2} + x_2 + x_3 + \dots + x_{n-1} + \frac{x_n}{2}}{n-1}, \text{ kde}$$

$x_1, x_2, \dots, x_n$ sú okamžikové ukazovatele pre n časových okamžikov označených ako $t_1, t_2, \dots, t_n$,

c) ak dĺžka časových intervalov (t_{i-1}, t_i) nie je konštantná, potom

$$\bar{x}_{ch} = \frac{\frac{x_1 + x_2}{2} \cdot h_1 + \frac{x_2 + x_3}{2} \cdot h_2 + \dots + \frac{x_{n-1} + x_n}{2} \cdot h_{n-1}}{h_1 + h_2 + \dots + h_{n-1}} = \frac{\frac{1}{2} \cdot \sum_{i=1}^{n-1} (x_i + x_{i+1}) h_i}{\sum_{i=1}^{n-1} h_i},$$

$x_1, x_2, \dots, x_n$ sú okamžikové ukazovatele pre n časových okamžikov označených ako $t_1, t_2, \dots, t_n$ a $h_1, h_2, \dots, h_n$ dĺžky časových intervalov,

- graficky sa okamžikové časové rady obvykle znázorňujú pomocou spojnicových diagramov, v ktorých je os úsečiek časovou osou.

INTERVALOVÉ ČASOVÉ RADY

Vlastnosti intervalových časových radov:

- jednotlivé hodnoty sa vzťahujú k časovým intervalom,
- hodnota intervalového ukazovateľa závisí na dĺžke intervalu (s rastúcou hodnotou dĺžky intervalu sa zväčšuje aj ukazovateľ),
- pred analýzou intervalových časových radov je vhodné prepočítať hodnoty intervalových ukazovateľov na jednotný interval (táto požiadavka vyplýva z praxe, kde často pracujeme s prirodzenými časovými obdobiami, ako napr. mesiac, rok, ktoré môžu mať rôznu dĺžku),
- za dlhšie časové obdobie možno intervalové ukazovatele prezentovať formou súčtových (kumulatívnych) radov.

Typy intervalových časových radov:

Ak si označíme hodnotu intervalového ukazovateľa zistenú za časový interval (t_0, t_i) , kde $1 \leq i \leq n$ symbolom Q_i a priradíme ho časovému okamžiku t_i , dostaneme nasledovný intervalový rad:

časový okamžik: $t_0 \quad t_1 \quad t_2 \quad \dots \quad t_{n-1} \quad t_n$

hodnota ukazovateľa: $Q_0 \quad Q_1 \quad Q_2 \quad \dots \quad Q_{n-1} \quad Q_n$

kde $Q_0=0$ a $Q_i > Q_{i-1}$. Ak hodnotu intervalového ukazovateľa zistenú za časový interval (t_{i-1}, t_i) označíme ako q_i a priradíme ju stredu daného časového intervalu, ktorý označíme ako t_i^* , dostaneme intervalový časový rad tvaru:

stred intervalu: $t_0^* \quad t_1^* \quad t_2^* \quad \dots \quad t_{n-1}^* \quad t_n^*$

hodnota ukazovateľa: $q_0 \quad q_1 \quad q_2 \quad \dots \quad q_{n-1} \quad q_n$

kde $q_i = q_{i-1}$, $q_i > q_{i-1}$ alebo $q_i < q_{i-1}$. Časový rad ukazovateľov q_i nazývame **intervalovým radom bežných hodnôt**. Rad ukazovateľov Q_i je vyjadrením súčtového **kumulatívneho radu**: $Q_i = q_1 + q_2 + \dots + q_{i-1} + q_i$.

Iným typom súčtového radu je **rad kľzavých úhrnov**. Využíva sa pre porovnanie hodnôt v odpovedajúcich si časových intervaloch, t.j. pre porovnanie radu hodnôt v sledovanom období s úrovňou radu v predošlom období. Ak hodnoty kľzavých úhrnov vzrastajú, veľkosť sledovaných ukazovateľov je vyššia ako v predchádzajúcom období. Podrobnejšie vysvetlenie výpočtu kľzavých úhrnov je súčasťou vzorového príkladu 1.

Všetky uvedené časové rady (rad bežných hodnôt, kumulovaných hodnôt -súčtový rad, rad kľzavých úhrnov) možno vhodne znázorniť pomocou **Z-diagramu**. Ide o kombináciu troch druhov časových radov v jednom grafe:

- a) znázornenie bežných hodnôt - priebeh vývoja daného javu v bežnom období,
- b) znázornenie kumulovaných hodnôt - vývoj skúmaného javu od začiatku sledovaného obdobia,
- c) znázornenie kľzavých úhrnov - vývoj javu v bežnom období oproti minulému obdobiu.

CHARAKTERISTIKA ČASOVÝCH RADOV

Hlavným cieľom analýzy časových radov je:

- odhalenie zákonitosti ich vývoja,
- zistenie príčin tohto vývoja,
- určenie ďalšej prognózy vývoja.

Podľa toho, či dané časové rady vykazujú určitý trend vo vývoji, môžeme hovoriť o časových radoch **bez trendu** alebo **s trendom**. Časové rady sa môžu vzájomne líšiť dĺžkou cyklu i amplitúdou výkyvu. V prípade niektorých časových radov možno pozorovať kolísanie hodnôt okolo určitej úrovne v ročných, štvrťročných, mesačných, prípadne denných intervaloch, pričom toto kolísanie má v každom roku rovnaký charakter. Ide o tzv. **sezónne kolísania**. Vo všeobecnosti možno konštatovať, že časové rady môžu obsahovať trend, periodické (sezónne, krátkodobé alebo cyklické) kolísanie alebo náhodné kolísanie. Z uvedeného vychádzame aj pri popise vývojového charakteru časových radov, kde charakterizujeme sezónne alebo cyklické kolísanie či intenzitu náhodného kolísania, kde plynulý vývoj časového radu je ovplyvnený náhodnými vplyvmi, ktoré spôsobujú menšie výkyvy ukazovateľov okolo trendu (Brázdil, et. al. 1995).

Základné informácie o časových radoch získavame pomocou popisných charakteristík, ktoré umožňujú použiť rôzne typy trendových funkcií k popísaniu analyzovaného časového radu. Patria tu:

a) absolútne miery rastu (absolútny prírastok, zrýchlenie absolútneho prírastku),

b) relatívne miery rastu (koeficient rastu, koeficient prírastku).

Absolútny prírastok a_i definujeme ako $x_i - x_{i-1}$. Vyjadruje prírastok hodnoty časového radu v časovom okamžiku t_i , kde $2 \leq i \leq j$. Ak zistíme, že hodnoty v časovom rade absolútnych prírastkov sú v podstate konštantné, možno daný časový rad považovať za rad s lineárnym trendom. V uvedenom prípade môžeme pre charakteristiku daného časového radu využiť **priemerný absolútny prírastok**, ktorý definujeme ako:

$$\bar{a}_i = \frac{\sum_{i=2}^j a_i}{j-1} = \frac{\sum_{i=2}^j (x_2 - x_1) + (x_3 - x_2) + \dots + (x_j - x_{j-1})}{j-1} = \frac{x_j - x_1}{j-1}.$$

Ak sa v časovom rade absolútnych prírastkov prejaví rastúca alebo klesajúca tendencia, je potrebné vypočítať hodnotu **zrýchlenia absolútneho prírastku** z_i : $z_i = a_i - a_{i-1} = (x_i - x_{i-1}) - (x_{i-1} - x_{i-2}) = x_i - 2x_{i-1} + x_{i-2}$.

V prípade, že poznáme hodnoty pôvodného časového radu, môžeme definovať **koeficient rastu**: $r_i = \frac{x_i}{x_{i-1}}$. Znázorňuje sa desatinným číslom alebo percentom. Vyjadruje,

o koľko percent vzrástla hodnota časového radu v okamžiku t_i v porovnaní s časovým okamžikom t_{i-1} . Ak hodnoty časového radu rastú stále a približne rovnako, môžeme vypočítať **priemerný koeficient rastu**, ktorý je geometrickým priemerom jednotlivých

hodnôt koeficientov rastu: $\bar{r}_i = \sqrt[j-1]{r_1 \cdot r_2 \cdot \dots \cdot r_{j-1}} = \sqrt[j-1]{\frac{x_2}{x_1} \cdot \frac{x_3}{x_2} \cdot \dots \cdot \frac{x_j}{x_{j-1}}} = \sqrt[j-1]{\frac{x_j}{x_1}}.$

Ďalšou relatívnou mierou rastu je **koeficient prírastku** p_i . Získame ho jednoduchým odpočítaním 1 (resp. 100 %) od každej hodnoty koeficientu rastu r_i : $p_i = r_i - 1$ (tab. 8).

Tab. 8 Výpočet základných mier rastu

t_i	x_i	a_i	z_i	r_i		p_i	
1960	4	-	-	-	-	-	-
1965	8	4	-	2	200%	1	100%
1970	12	4	0	1,5	150%	0,5	50%
1975	13	1	-3	1,1	108%	0,1	8%
1980	11	-2	-3	0,8	85%	-0,2	-15%
1985	14	3	5	1,3	127%	0,3	27%
1990	16	2	-1	1,1	114%	0,1	14%

Poznámka: V uvedenom príklade sú $a_i = x_i - x_{i-1}$, $r_i = \frac{x_i}{x_{i-1}}$, $p_i = r_i - 1$ definované pre $2 \leq i \leq j$, z_i je definované pre $3 \leq i \leq j$.

ZISŤOVANIE VÝVOJOVÝCH TRENDOV ČASOVÝCH RADOV

Ako sme už vyššie spomínali, hlavnými javmi, ktoré je potrebné sledovať pri analýze časových radov, sú trend vývoja, periodické a náhodné kolísanie. Pre určenie trendu vývoja, a tým aj umožnenie **extrapolácie časového radu**³, sa využíva **vyrovnávanie (zhladzovanie) časového radu**.

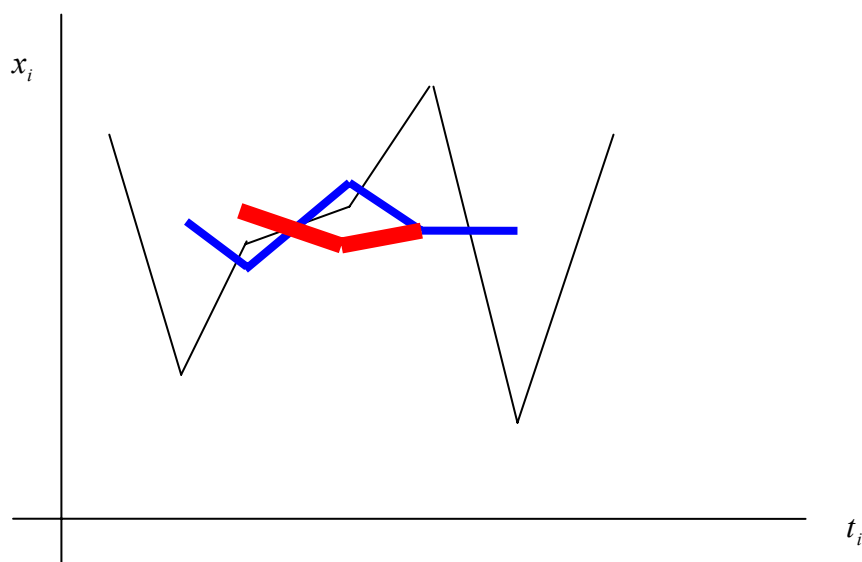
a) **Mechanické zhladzovanie časového radu** - pôvodné hodnoty nahrádzame priemernými, pričom nepožadujeme možnosť extrapolácie časového radu. Najčastejšie sa k mechanickému zhladzovaniu (odstráneniu lokálnych výkyvov v časovom rade) používajú **kľzavé (pohyblivé) priemery** (graf 12). Tie nahrádzajú empirické hodnoty jednoduchými alebo váženými priemermi (prevažne aritmetickými), ktoré vypočítame z hodnoty x_i a určitého počtu hodnôt, ktoré bezprostredne predchádzajú danej hodnote a z rovnakého počtu hodnôt, ktoré sa nachádzajú za uvedenou hodnotou (tab. 9 a riešený príklad 3). Ak je počet hodnôt párny, pristupujeme k **centrovaniu kľzavých priemerov**, ktoré spočíva vo výpočte priemeru z každých dvoch susedných kľzavých priemerov. Nevýhodou je to, že krajné hodnoty tak ostávajú v časovom rade nevyrovnané.

³ **Extrapolácia** vyjadruje odhad parametrov v priestore mimo zadaných hodnôt. **Interpolácia** je odhad parametrov medzi známymi hodnotami.

Tab. 9 Kľzavé priemery z 3 a z 5 hodnôt

x_i	Kľzavý priemer z 3 hodnôt	Kľzavý priemer z 5 hodnôt
12	-	-
8	10	-
10	9,7	10,8
11	11,3	9,6
13	10	10,4
6	10,3	-
12	-	-

Graf 12 Schematické znázornenie zhladzovania časového radu pomocou kľzavých priemerov



b) Pri **analytickom vyrovnávaní časového radu** sa snažíme vyjadriť časový rad nejakou analytickou funkciou, ktorá najlepšie vystihuje vývojový trend. Vo všeobecnosti ju možno zapísať v tvare $y = f(t, \alpha_1, \alpha_2, \dots, \alpha_n)$, kde t je časová premenná a $\alpha_1, \alpha_2, \dots, \alpha_n$ príslušné parametre analytickej funkcie. Vzniknutá regresná funkcia môže predstavovať:

- lineárny (priamkový) trend, kde prírastok, resp. úbytok javu v čase je pravidelný: $y_i = a + b \cdot t_i$, kde y_i predstavuje závislú premennú, t_i nezávislé premennú (čas) a a, b sú parametre,
- nelineárny (mocninový, exponenciálny, hyperbolický, parabolický 2. stupňa – kvadratický trend, vývojový trend vo forme logistickej krivky, ..).

Riešené úlohy:

1. Z údajov uvedených v tabuľke vypočítajte:

- a) kumulované hodnoty,
- b) kľzavé úhrny,
- c) výpočty znázorníte graficky vo forme Z-diagramu.

	<i>Predaj počítačov v ks v roku</i>	
<i>Mesiac</i>	<i>rok 2004</i>	<i>rok 2005</i>
<i>Január</i>	7	10
<i>Február</i>	8	4
<i>Marec</i>	9	5
<i>Apríl</i>	12	8
<i>Máj</i>	7	7
<i>Jún</i>	3	6
<i>Júl</i>	8	5
<i>August</i>	10	7
<i>September</i>	4	4
<i>Október</i>	8	7
<i>November</i>	6	7
<i>December</i>	9	4

Riešenie:

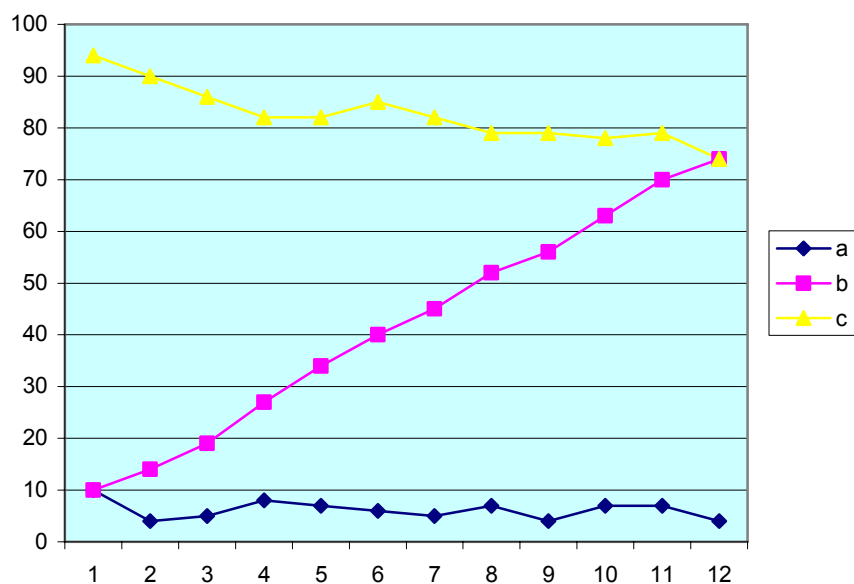
a) Kumulované hodnoty počítame pre aktuálny rok (v danom príklade pre rok 2005). Ak si hodnoty v stĺpci za rok 2005 označíme ako n_i , potom kumulatívne hodnoty N_i určíme

$$\text{ako } N_i = \sum_{i=1}^k n_i.$$

b) Pri výpočte kľzavých úhrnov postupujeme nasledovne: Najprv si určíme sumárnu hodnotu počítačov pre rok 2004. Následne si vypočítame rozdiel medzi počtom kusov predaných počítačov v roku 2005 a 2004. Hodnotu kľzavého úhrnu pre mesiac január určíme tak, že k celkovému počtu počítačov v roku 2004 pripočítame rozdiel hodnôt v počte počítačov v januári 2005 a 2004 a vpíšeme ju do stĺpca kľzavých úhrnov za mesiac január, t.j. $(91+3=94)$. V prípade nasledujúceho mesiaca február využijeme už hodnotu kľzavého priemeru z mesiaca január, ku ktorej pripočítame rozdiel v počte kusov počítačov vo februári 2005 a 2004 $(94-4=90)$. Výslednú hodnotu zapíšeme do stĺpca kľzavých úhrnov pre mesiac február a postup opakujeme. Pri výpočtoch kľzavých úhrnov pre nasledujúce mesiace využívame stále hodnotu kľzavého úhrnu za predošlý mesiac.

	Predaj počítačov v kusoch v roku		Rozdiel	Kumulované hodnoty N_i	Kľzavé úhrny
Mesiac	rok 2004	rok 2005 n_k			
Január	7	10	3	10	94
Február	8	4	-4	14	90
Marec	9	5	-4	19	86
Apríl	12	8	-4	27	82
Máj	7	7	0	34	82
Jún	3	6	3	40	85
Júl	8	5	-3	45	82
August	10	7	-3	52	79
September	4	4	0	56	79
Október	8	7	-1	63	78
November	6	7	1	70	79
December	9	4	-5	74	74
Suma	91	-	-	-	

c) Z-diagram



Legenda: a-priebeh vývoja daného javu v roku 2005, b- znázornenie kumulovaných hodnôt pre rok 2005, c-vykreslenie vývoja javu v roku 2005 oproti roku 2004 prostredníctvom kľzavých úhrnov.

2. Z údajov o počte obyvateľov vypočítajte:

a) absolútny prírastok,

b) zrýchlenie absolútneho prírastku,

c) koeficient rastu

d) koeficient prírastku. Koeficienty rastu a prírastku zaokrúhľujte na 2 desatinné miesta.

Rok	Počet obyvateľov
1870	1061
1880	1524
1890	1325
1900	1305
1910	1472
1920	1438
1930	1428
1940	1500

Riešenie:

Označme si časové obdobia ako t_i , hodnoty štatistického znaku ako x_i , hodnoty absolútneho prírastku a_i , zrýchlenia absolútneho prírastku z_i , koeficientu rastu r_i a koeficientu prírastku p_i . Hľadané miery rastu vypočítame pomocou nasledovných vzorcov:

$$a_i = x_i - x_{i-1}, z_i = a_i - a_{i-1} = (x_i - x_{i-1}) - (x_{i-1} - x_{i-2}) = x_i - 2x_{i-1} + x_{i-2}, r_i = \frac{x_i}{x_{i-1}}, p_i = r_i - 1,$$

kde $2 \leq i \leq j$ pre a_i, p_i, r_i a $3 \leq i \leq j$ pre z_i .

t_i	x_i	a_i	z_i	r_i	r_i (v %)	p_i	p_i (v %)
1870	1061	-	-	-	-	-	-
1880	1524	463	-	1,44	144	0,44	44
1890	1325	-199	-662	0,87	87	-0,13	-13
1900	1305	-20	179	0,98	98	-0,02	-2
1910	1472	167	187	1,13	113	0,13	13
1920	1438	-34	-201	0,98	98	-0,02	-2
1930	1428	-10	24	0,99	99	-0,01	-1
1940	1500	72	82	1,05	105	0,05	5

3. Graficky znázorníte zmeny v zrážkových pomeroch vybraného územia pomocou reálnych hodnôt a trojročných kĺzavých priemerov.

Rok	Priemerný ročný úhrn zrážok v mm
1950	700
1955	650
1960	625
1965	730
1970	680
1975	655
1980	730
1985	695
1990	600
1995	705
2000	680

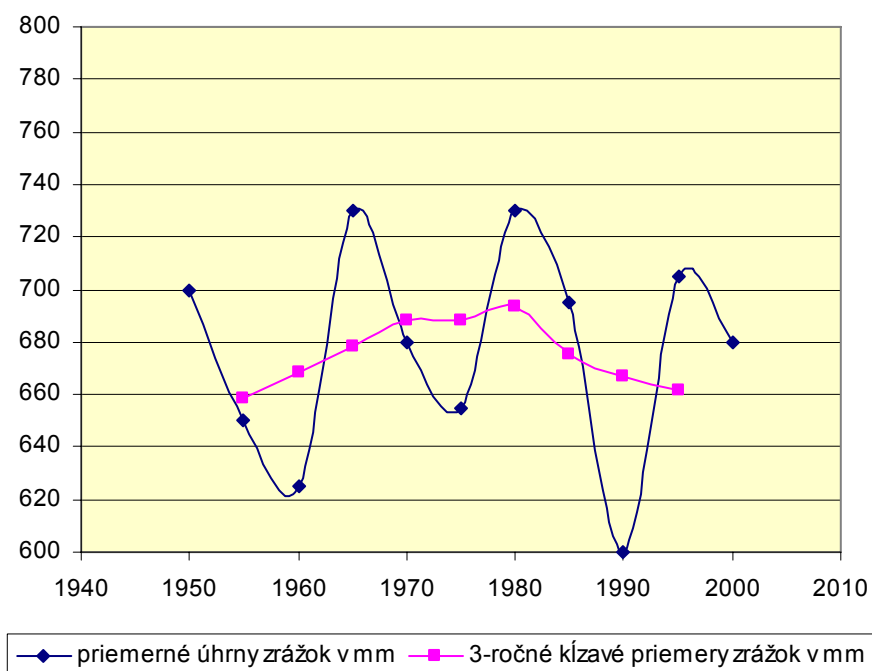
Riešenie:

Zadanú tabuľku si doplníme o 3-ročné kĺzavé priemery, ktoré si vypočítame ako aritmetický priemer z troch po sebe nasledujúcich priemerných ročných úhrnov zrážok. Pre roky 1950, 1955 a 1960 bude hodnota kĺzavého priemeru: $(700+650+625)/3=658,3$, ktorú napíšeme k roku 1955 do stĺpca 3-ročných kĺzavých priemerov. Nasledujúci kĺzavý priemer počítame z priemerných ročných úhrnov zrážok pre roky 1955, 1960 a 1965, pričom výslednú hodnotu zapíšeme k roku 1960 do stĺpca 3-ročných kĺzavých úhrnov. Uvedený postup opakujeme.

Rok	Priemerný ročný úhrn zrážok v mm	3- ročné kĺzavé priemery
1950	700	-
1955	650	658,3
1960	625	668,3
1965	730	678,3
1970	680	688,3
1975	655	688,3
1980	730	693,3
1985	695	675
1990	600	666,7
1995	705	661,7
2000	680	-

Poznámka: Hodnoty 3-ročných kĺzavých úhrnov zrážok pre roky 1950 a 2000 nie je možné z tabuľky vypočítať.

Graf: Priemerné ročné úhrny zrážok a 3-ročné klzavé priemery zrážok (v mm)



Neriešené úlohy:

1. Z údajov o priemernom úhrne zrážok podľa vybranej zrážkomernej stanice (v mm) vypočítajte:

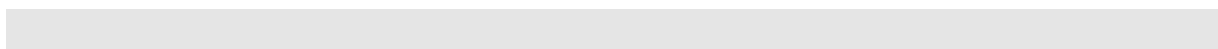
- kumulované hodnoty,
- klzavé úhrny,
- výpočty znázorníte graficky vo forme Z-diagramu.

Obdobie	I.	II.	III.	IV.	V.	VI.	VII.	VIII.	IX.	X.	XI.	XII.
Rok 1990	41	38	29	38	56	79	67	67	63	52	55	52
Rok 1991	44	33	26	35	57	83	67	65	68	55	51	50

2. Z údajov o počte kusov dobytku vypočítajte:

- absolútny prírastok,
- zrýchlenie absolútneho prírastku,
- koeficient rastu
- koeficient prírastku. Koeficienty rastu a prírastku zaokrúhľujte na 2 desatinné miesta.

<i>Rok</i>	<i>Počet kusov dobytka</i>
1985	600
1986	685
1987	700
1988	724
1989	720
1990	581
1991	505
1992	467



VÝBEROVÉ SKÚMANIE

V niektorých prípadoch je pomerne problematické získať údaje o všetkých jednotkách štatistického súboru, prípadne je rozsah základného súboru príliš veľký. Z tohto dôvodu pracujeme s výberovým štatistickým súborom, ktorý predstavuje reprezentatívnu podmnožinu základného štatistického súboru. Reprezentatívnosť vypovedá o miere zhody štruktúry výberového súboru podľa určitého znaku so základným súborom. Jednotlivé **charakteristiky (parametre) základného súboru odhadujeme na základe vypočítaných charakteristík výberového súboru**. Odhadnuté charakteristiky potom vystihujú parametre základného súboru len s určitou pravdepodobnosťou. Pre označenie skutočných a odhadnutých parametrov budeme používať nasledovné označenia:

- skutočné hodnoty parametrov základného súboru: stredná hodnota μ , rozptyl σ^2 , smerodajná odchýlka σ ,
- skutočné hodnoty parametrov výberového súboru: stredná hodnota $\bar{x}$, rozptyl s_x^2 , smerodajná odchýlka s_x .
- odhadnuté hodnoty parametrov základného súboru: stredná hodnota $\hat{\mu}$, rozptyl $\hat{\sigma}^2$, smerodajná odchýlka $\hat{\sigma}$.

Neznáme parametre základného súboru môžeme odhadovať dvoma spôsobmi:

- a) bodovým odhadom, kde neznámy parameter odhadujeme pomocou jednej hodnoty,
- b) intervalovým odhadom, pomocou ktorého určujeme interval, v ktorom bude ležať odhadovaný parameter.

Intervalový odhad parametrov je vhodnejší ako bodový, nakoľko rozsah intervalu udáva i presnosť s akou môžeme počítať.

BODOVÝ ODHAD

Pre bodový odhad aritmetického priemeru platí vzťah $\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$, teda odhad priemerov základného súboru je rovný výberovému súboru. Priemer výberového súboru je tak **neskresleným odhadom**⁴ priemeru základného súboru. Pri výpočte smerodajnej odchýlky vychádzame z odchýlky jednotlivých hodnôt znaku od ich aritmetického priemeru. Nakoľko pre výpočet aritmetického priemeru sme použili n hodnôt: $\sum_{i=1}^n x_i = n \cdot \bar{x}$, potom možno pre

⁴ Neskreslenými odhadmi základného súboru nazývame tie, ktorých stredná hodnota sa rovná skutočnej hodnote tohto parametra základného súboru. V prípade, že máme k dispozícii viacero výberových súborov z jedného základného súboru a každý z nich by sme použili pre odhad parametra základného súboru, tak skutočná hodnota parametra základného súboru by bola priemerom z týchto odhadnutých hodnôt.

výpočet odhadovanej smerodajnej odchýlky použiť len $n-1$ nezávislých odchýlok, tzv. **stupňov voľnosti**, lebo posledná odchýlka musí dopĺňať súčet odchýlok do 0.

Pre bodový odhad smerodajnej odchýlky platí vzťah: $\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$.

Ak si uvedený vzťah dáme do pomeru so vzťahom pre výpočet skutočnej hodnoty smerodajnej odchýlky pre výberový súbor, dostávame:

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

Pre odhadovaný rozptyl základného súboru platí: $\hat{\sigma}^2 = s_x^2 \cdot \frac{n}{n-1}$.

Rozptyl a smerodajnú odchýlku výberového súboru teda nemožno považovať za neskreslené odhady rozptylu a smerodajnej odchýlky základného súboru.

Základné vlastnosti bodového odhadu:

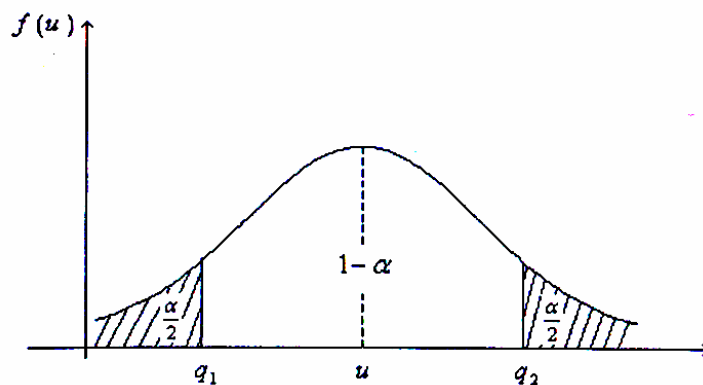
- 1. konzistencia (preukázateľnosť)** – odhad je konzistentný, ak rozdiel medzi odhadnutou hodnotou parametra základného súboru a skutočnou hodnotou parametra základného súboru sa znižuje s rastúcim rozsahom výberového súboru,
- 2. neskreslenosť (nevychýlenosť)** – neskreslené odhady parametra základného súboru sú tie, ktorých stredná hodnota výberovej charakteristiky sa rovná skutočnej hodnote parametra základného súboru,
- 3. výdatnosť** – výdatným odhadom parametra základného súboru nazývame tú charakteristiku, ktorej rozptyl je zo všetkých možných výberových charakteristík najmenší.

INTERVALOVÝ ODHAD

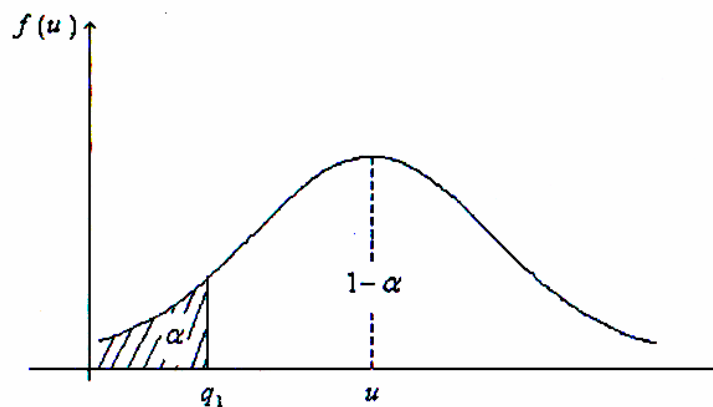
Skutočné hodnoty výberových charakteristík základného súboru sa výber od výberu menia, a tým aj odhady výberových charakteristík základného súboru. Z tohto dôvodu je potrebné stanoviť odchýlky odhadnutých parametrov základného súboru od skutočných pomocou tzv. **intervalov spoľahlivosti** (tab. 10). Intervalovým odhadom parametra základného súboru nazývame odhad pomocou číselného intervalu $\langle q_1, q_2 \rangle$, v ktorom sa odhadovaný parameter u nachádza s určitou pravdepodobnosťou (graf 13). Interval $\langle q_1, q_2 \rangle$ nazývame intervalom spoľahlivosti, hodnoty q_1, q_2 sú **medze spoľahlivosti**. Interval $(-\infty, q_1)$ a (q_2, ∞) predstavujú **kritický obor**. Plochu pod krivkou hustoty pravdepodobnosti medzi medzami spoľahlivosti interpretujeme ako **oblasť prijatia**, ostané oblasti pod krivkou hustoty pravdepodobnosti sú **oblasťami zamietnutia**, resp. **kritickými oblasťami**. Pravdepodobnosť, že vybraný parameter u padne do intervalu spoľahlivosti $\langle q_1, q_2 \rangle$, je $1 - \alpha$, čo možno zapísať $P(q_1 \leq u \leq q_2) = 1 - \alpha$. Parameter u je v uvedenom prípade ohraničený zdola i zhora, teda ide o **obojsmerný interval spoľahlivosti**. Číslo $1 - \alpha$ nazývame **spoľahlivosť odhadu**. Hodnota α - **riziko odhadu** predstavuje pravdepodobnosť

toho, že parameter u nie je z intervalu spoľahlivosti. Pravdepodobnosť, že parameter u nadobudne hodnotu väčšiu alebo rovnú q_1 zapíšeme ako $P(q_1 \leq u) = 1 - \alpha$. Obdobne pravdepodobnosť, že parameter u nadobudne hodnotu menšiu alebo rovnú q_2 , vyjadríme vzťahom $P(q_2 \geq u) = 1 - \alpha$. Nakoľko interval spoľahlivosti je v týchto prípadoch ohraničený len z jednej strany, hovoríme o **jednostrannom intervale spoľahlivosti** (grafy 14,15).

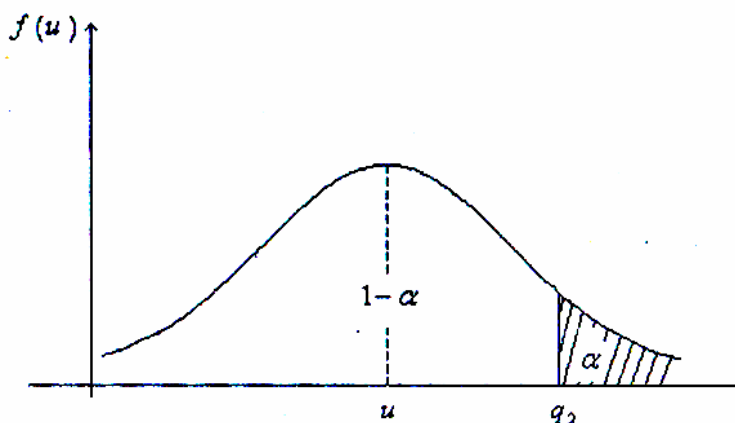
Graf 13 Obojstranný interval spoľahlivosti



Graf 14 Jednostranný zdola ohraničený interval spoľahlivosti



Graf 15 Jednostranný zhora ohraničený interval spoľahlivosti



Spracované podľa: Gregorová, Fillová 2004 (graf 13,14,15)

Tab. 10 Najčastejšie používané intervaly spoľahlivosti

Násobky smerodajnej odchýlky	Oblasť prijatia / zamietnutia	
	prijatie	zamietnutie
$\pm 1,960$	95 %	5 %
$\pm 2,567$	99 %	1 %
$\pm 3,291$	99,9 %	0,1 %

Spracované podľa: Brázdil, et. al. 1995

Intervalový odhad parametra strednej hodnoty μ pre veľké súbory ($n > 30$)

Pravdepodobnosť, že vybraný parameter μ predstavujúci strednú hodnotu základného súboru padne do intervalu spoľahlivosti, je: $P(q_1 \leq \mu \leq q_2) = 1 - \alpha$. Interval spoľahlivosti určíme pomocou výpočtu dolnej a hornej medze spoľahlivosti nasledovne:

a) ak **poznaáme rozptyl** základného súboru σ^2 , potom hodnoty q_1, q_2 vypočítame pomocou nasledovných vzorcov: $q_1 = \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$, $q_2 = \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$,

b) v prípade, že **hodnotu rozptylu základného súboru nepoznáme**, využijeme pre výpočet medzí spoľahlivosti q_1, q_2 odhadnutú hodnotu rozptylu s_x pre výberový súbor:

$$q_1 = \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{s_x}{\sqrt{n-1}} \text{ a } q_2 = \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{s_x}{\sqrt{n-1}}.$$

Intervalový odhad parametra strednej hodnoty μ pre malé súbory ($n \leq 30$)

Ak je rozsah výberového súboru malý, potom pre intervalový odhad parametra μ využijeme Studentovo rozdelenie s $(n-1)$ stupňami voľnosti. Pravdepodobnosť, že vybraný parameter μ bude z intervalu $\langle q_1, q_2 \rangle$, možno zapísať v tvare: $P(q_1 \leq \mu \leq q_2) = 1 - \alpha$.

a) Keď je **hodnota rozptylu σ^2 známa**, potom hodnoty q_1, q_2 určíme zo vzťahov:

$$q_1 = \bar{x} - t_{1-\frac{\alpha}{2};(n-1)} \cdot \frac{\sigma}{\sqrt{n}}, \quad q_2 = \bar{x} + t_{1-\frac{\alpha}{2};(n-1)} \cdot \frac{\sigma}{\sqrt{n}}.$$

b) V prípade, že **hodnotu rozptylu σ^2 nepoznáme**, potom pre výpočet medzí spoľahlivosti q_1, q_2 použijeme odhadnutú hodnotu rozptylu s_x pre výberový súbor:

$$q_1 = \bar{x} - t_{1-\frac{\alpha}{2};(n-1)} \cdot \frac{s_x}{\sqrt{n-1}}, \quad q_2 = \bar{x} + t_{1-\frac{\alpha}{2};(n-1)} \cdot \frac{s_x}{\sqrt{n-1}}.$$

Intervalový odhad parametra σ^2 základného súboru

Ak predpokladáme normálne rozdelenie základného súboru, interval spoľahlivosti pre rozptyl základného súboru možno zapísať v tvare: $P(q_1 \leq \sigma^2 \leq q_2) = 1 - \alpha$. Hodnoty medzí spoľahlivosti pre $(n-1)$ stupňov voľnosti určíme z tabuliek χ^2 rozdelenia.

$$q_1 = \frac{(n-1) \cdot s_x^2}{\chi_{1-\alpha/2; (n-1)}^2}, \quad q_2 = \frac{(n-1) \cdot s_x^2}{\chi_{\alpha/2; (n-1)}^2}.$$

Poznámka: Medze intervalu spoľahlivosti pre odhad parametra σ základného súboru vypočítame odmocnením medzí spoľahlivosti pre odhad parametra σ^2 základného súboru.

TESTOVANIE HYPOTÉZ

Štatistickými hypotézami nazývame domnienky (predpoklady) o parametroch základného súboru alebo o jeho rozdelení, prípadne o parametroch viacerých základných súborov, ktoré možno overiť štatistickými metódami. Postup, ktorým overujeme platnosť hypotézy, nazývame **testovanie hypotéz**.

Podstata testovania hypotéz spočíva v tom, že naformulujeme **nulovú hypotézu** H_0 ⁵, ktorá sa po overení prijme alebo zamietne. **Cieľom testovania** je teda rozhodnúť, či daná hypotéza je pravdivá alebo nepravdivá. Ak je hypotéza pravdivá, potom ju prijímame, ak nie, tak ju zamietame.

Pravidlo, na základe ktorého rozhodneme, či testovanú hypotézu zamietneme alebo prijímame, nazývame **štatistický test**.

Pri testovaní hypotéz pracujeme s určitou pravdepodobnosťou omylu, ktorú nazývame **hladinou významnosti** α . Chápeme pod ňou pravdepodobnosť, že náhodná odchýlka prekročí danú hodnotu – tzv. **kritickú hodnotu**. Ak $\alpha \leq 0,05$, tak možno povedať, že pravdepodobnosť omylu je 5% (5 zo 100), ak by sme α zvolili $\leq 0,001$, potom by pravdepodobnosť omylu bola 0,1%, teda mohli by sme povedať, že skúmaný štatistický jav je veľmi vysoko signifikantný (má vysokú štatistickú významnosť).

Testovacia charakteristika je číslo, ktoré získavame pri overovaní hypotéz. Každú testovaciu charakteristiku porovnávame s kritickou hodnotou. Na základe výsledkov ich porovnania následne rozhodneme, či danú hypotézu prijmem alebo zamietneme. Kritické hodnoty testovacích kritérií pre najčastejšie používané hladiny významnosti a pre rôzne rozsahy výberov sú uvedené v štatistických tabuľkách.

Pri testovaní nulovej hypotézy sa môžeme dopustiť dvoch druhov chýb. Prvý druh chyby súvisí s tým, že zamietneme správnu hypotézu (chyba prvého druhu) alebo tým, že prijmem nesprávnu hypotézu (chyba druhého druhu). Chybu prvého druhu možno obmedziť vopred stanovenou hladinou významnosti. Čím je hladina významnosti menšia, tým menšia je pravdepodobnosť, že sa tejto chyby dopustíme. Zároveň sa tým zvyšuje možnosť chyby druhého druhu (nesprávnym prijatím nulovej hypotézy).

⁵ **Nulová hypotéza** je základná hypotéza, kde predpokladáme nulový rozdiel medzi parametrami. Jej platnosť sa overuje testom významnosti.

Postup testovania štatistickej hypotézy

1. Formulácia nulovej hypotézy H_0 , príp. alternatívnej hypotézy H_A .
2. Voľba hladiny významnosti α .
3. Vypočet výberových a testovacích charakteristík.
4. Rozhodnutie o prijatí, resp. zamietnutí H_0 .

Výsledok testovania:

a) Zamietnutie hypotézy H_0 neznamená, že testovaná nulová hypotéza *neplatí*, ale že jej platnosti neveríme, pretože výsledok testu poskytol objektívny dôvod neveriť jej. V ďalšom postupe uvažujeme, že H_0 neplatí a platí H_A

b) Ak nezamietneme hypotézu H_0 , neznamená to jej prijatie. Výsledok testu neukázal tak veľkú nezhodu medzi zistenou skutočnosťou a testovanou hypotézou, ktorá by dala dostatočný dôvod na zamietnutie hypotézy.

Klasifikácia testovacích kritérií

1. Podľa zamerania

- napr. testy o stredných hodnotách (aritmetický priemer, medián), testy o vybraných mierach variability (napr. rozptyloch, kvartilových odchýlkach...), testy zhody teoretického a empirického rozdelenia a pod.

2. Zloženia

- **parametrické** (empirické údaje majú normálne rozdelenie a rozsah štatistického súboru je väčší než 30 pozorovaní),

- **neparametrické** (normalita rozdelenia nie je splnená, v súbore je napr. veľa rovnakých hodnôt, niektoré hodnoty sú evidentne odľahlé alebo rozdelenie početností je síce súmerné, no má tvar U).

3. Podľa vzťahu výberových súborov

- **párové** (pracujeme s dvojicou hodnôt, napr. vstupný a výstupný test, úvodný a záverečný ceremoniál),

- **nepárové** (napr. 1.B trieda a 1. C. trieda).

4. **Podľa formulácie alternatívnych hypotéz** (sú to hypotézy, kde predpokladáme nerovnosť parametrov základného súboru)

- **jednostranné** $H_1: Q < Q_0$ resp. $H_2: Q > Q_0$,

- **obojsstranné** $H_1: Q \neq Q_0$,

Q, Q_0 sú parametre základného súboru.

Riešené príklady:

1. Určte 99 % interval spoľahlivosti pre návštevnosť divadelného krúžku, ak viete, že pri náhodnom výbere 100 žiakov vybranej základnej školy tvorila priemerná návštevnosť divadelného krúžku 32 %. Uvedený príklad vyriešte za predpokladu, že návštevnosť divadelného krúžku má normálne rozdelenie s hodnotou rozptylu základného súboru 49.

Riešenie:

Základný súbor - tvoria ho všetci žiaci danej školy,

- rozptyl základného súboru $\sigma^2 = 49$,

- smerodajná odchýlka základného súboru je $\sigma = 7$.

Výberový súbor: - rozsah výberového súboru je $n=100$,

- priemerná návštevnosť krúžku je $\bar{x} = 32$.

Najprv si určíme z tabuliek kritických hodnôt normálneho rozdelenia (tab. 15) určíme hodnotu z_p pre $\alpha = 0,01$: $z_p = z_{1-\frac{\alpha}{2}} = z_{1-\frac{0,01}{2}} = z_{0,995} = 2,576$.

$$q_1 = \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 32 - 2,576 \cdot \frac{7}{\sqrt{100}} = 32 - 2,576 \cdot 0,7 = 30,1968$$

$$q_2 = \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} = 32 + 2,576 \cdot \frac{7}{\sqrt{100}} = 32 + 2,576 \cdot 0,7 = 33,8032$$

Hľadaný interval spoľahlivosti: $P(q_1 \leq \mu \leq q_2) = 1 - \alpha$

$$P(30,1968 \leq \mu \leq 33,8032) = 0,99$$

Odpoveď: S 99 % pravdepodobnosťou možno tvrdiť, že počet žiakov, ktorí navštevujú divadelný krúžok, sa pohybuje v intervale $\langle 30,1968; 33,8032 \rangle$.

2. Určte pravdepodobnosť toho, že priemerná výška u študentov sa bude pohybovať v hraniciach $\pm 2,1$ cm, ak viete, že v prípade 101 náhodne vybraných študentov bola priemerná výška 174,9 cm s rozptylom $s_x^2 = 121$ cm.

Riešenie: Nakoľko nepoznáme hodnotu rozptylu pre základný súbor, budeme vychádzať zo vzorca pre výpočet dolnej medze spoľahlivosti $q_1 = \bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{s_x}{\sqrt{n-1}}$, ktorý si upravíme tak,

aby sme z neho mohli vyjadriť kritickú hodnotu normovaného normálneho rozdelenia z_p .

$$z_p = z_{1-\frac{\alpha}{2}} = \frac{\sqrt{(n-1) \cdot (\bar{x} - q_1)}}{s} = \frac{\sqrt{100 \cdot 2,1}}{11} = \frac{10,2,1}{11} \approx 1,909 \Rightarrow p \approx 0,972$$

$$p = 1 - \frac{\alpha}{2} \Rightarrow 1 - \alpha = 2p - 1 \Rightarrow 1 - \alpha = 2 \cdot 0,972 - 1 = 0,944$$

Odpoveď: S 94,4% pravdepodobnosťou možno tvrdiť, že priemerná výška u študentov sa bude pohybovať v intervale $174,9 \text{ cm} \pm 2,1 \text{ cm}$.

3. Na 10 farmách bol zistený nasledovný počet kusov oviec: 100, 525, 630, 420, 100, 149, 854, 956, 546, 169. Určte 95 % interval spoľahlivosti pre priemerný počet oviec na farme.

Riešenie:

Vypočítame si výberové charakteristiky $\bar{x}$ a s_x^2 :

$$\bar{x} = \frac{\sum_{i=1}^k x_i n_i}{n} = \frac{4449}{10} = 444,9; s_x^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 \cdot n_i}{n} = \frac{881694,9}{10} = 88169,49 \Rightarrow s \approx 296,933.$$

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot n_i$
100	2	200	-344,9	118956,01	237912
149	1	149	-295,9	87556,81	87556,81
169	1	169	-275,9	76120,81	76120,81
420	1	420	-24,9	620,01	620,01
525	1	525	80,1	6416,01	6416,01
546	1	546	101,1	10221,21	10221,21
630	1	630	185,1	34262,01	34262,01
854	1	854	409,1	167362,81	167362,8
956	1	956	511,1	261223,21	261223,21
Σ	$\Sigma=10$	$\Sigma=4449$			$\Sigma=881694,9$

V tabuľke Studentovho rozdelenia nájdeme hodnotu $t_{1-\frac{\alpha}{2}}$ pre $\alpha = 0,05$ a pre $(n-1)=9$ stupňov voľnosti: $t_{1-\frac{0,05}{2}} = t_{0,975} = 2,262$. Dosadením uvedenej hodnoty do vzťahu

$$\bar{x} - t_{1-\frac{\alpha}{2};(n-1)} \cdot \frac{s_x}{\sqrt{n-1}} \leq \mu \leq \bar{x} + t_{1-\frac{\alpha}{2};(n-1)} \cdot \frac{s_x}{\sqrt{n-1}} \text{ dostávame}$$

$$444,9 - 2,262 \cdot \frac{296,933}{\sqrt{9}} \leq \mu \leq 444,9 + 2,262 \cdot \frac{296,933}{\sqrt{9}} \Rightarrow 221,013 \leq \mu \leq 668,787.$$

Odpoveď: Priemerný počet oviec sa s 95 % spoľahlivosťou pohybuje v intervale $449,9 \pm 223,887$.

4. V skupine 10 žiakov sme vykonali vstupné a výstupné testy z geografie. Po absolvovaní vstupného testu sme žiakom pridali navyše 2 vyučovacie hodiny geografie týždenne. Na hladine významnosti $\alpha = 0,05$ zistíte, či sa zvýšenie časovej dotácie prejavilo v bodovom ohodnotení žiakov pri absolvovaní výstupného testu štatisticky významne.

Poradie žiaka	Vstupný test	Výstupný test
1.	11	14
2.	12	12
3.	5	14
4.	9	13
5.	18	14
6.	15	20
7.	13	10
8.	15	15
9.	17	13
10.	11	8

Riešenie:

Pri testovaní hypotéz musíme najskôr zistiť, ktorý test je vhodné použiť. V uvedenom príklade pracujeme s párovými hodnotami – vstupný a výstupný test, ktoré sú závislé. Zadanie príkladu napovedá, že pôjde o neparametrický test⁶ - Wilcoxonov test.

Použitie Wilcoxonovho testu:

Wilcoxonov neparametrický test pre párové hodnoty pracuje s poradiami hodnôt. Ako testovacie kritérium W berie do úvahy menší zo súčtov kladných, respektíve záporných poradií hodnôt. W_α predstavuje kritickú hodnotu pre n -počet nenulových rozdielov kladných, respektíve záporných poradií hodnôt.

⁶ Neparametrické testy používame vtedy, ak je normalita rozdelenia evidentne nespĺnená (napr. ak je v súbore veľa rovnakých hodnôt, niektoré hodnoty sú evidentne odľahlé alebo rozdelenie početností je síce súmerné, no má tvar U).

Rozhodnutie: $W < W_{\alpha} \Rightarrow H_0$ zamietame,

$W > W_{\alpha} \Rightarrow H_0$ prijímame.

Formulácia nulovej hypotézy H_0 : Predpokladáme, že sa u žiakov pridanie 2 hodín geografie neprejavilo na výsledkoch bodového ohodnotenia výstupného testu štatisticky významne.

Poradie žiaka	Vstupný test	Výstupný test	Diferencia	Kladné poradia	Záporné poradia
1.	11	14	-3		2.
2.	12	12	0		
3.	5	14	-9		8.
4.	9	13	-4		4,5.
5.	18	12	6	7.	
6.	15	20	-5		6.
7.	13	10	3	2.	
8.	15	15	0		
9.	17	13	4	4,5.	
10.	11	8	3	2.	
Spolu:				15,5	20,5

Určenie poradia: Nakoľko u troch žiakov sa prejavil 3-bodový rozdiel medzi vstupným a výstupným testom, musíme najskôr vypočítať aritmetický priemer 1., 2. a 3. poradia:

$\frac{1+2+3}{3} = 2$. Podľa toho, či je pred číslom 3 v stĺpci diferencia kladné alebo záporné

znamienko, napíšeme výslednú hodnotu poradia (2.) do príslušného stĺpca kladných alebo záporných poradií hodnôt. Nakoľko ďalší dvaja žiaci mali medzi vstupným a výstupným testom 4-bodový rozdiel a prvé tri miesta sú už obsadené, počítam aritmetický priemer

z dvoch poradií (4. a 5. poradia): $\frac{4+5}{2} = 4,5$. Vypočítané poradie následne vpišem podľa

vyššie spomenutých pravidiel do príslušného stĺpca poradií. Uvedený postup opakujeme so všetkými hodnotami.

Rozhodnutie: Ako testovacie kritérium W používa Wilcoxonov test menší zo súčtov kladných, respektíve záporných poradií hodnôt. Z toho vyplýva, že $W=15,5$. W_{α} predstavuje kritickú hodnotu pre n -počet nenulových rozdielov. Nakoľko máme 10 hodnôt a 2 nulové rozdiely, $W_{\alpha}=3$ (tab.19). $W > W_{\alpha} \Rightarrow H_0$ prijímame.

Odpoveď: Napriek pridaniu 2 hodín geografie výsledky výstupného testu u žiakov nepreukázali oproti vstupnému testu štatisticky výrazný rozdiel v ich bodovom ohodnotení.

5. Po zadaní toho istého testu v dvoch triedach sme získali nasledovné bodové výsledky:

Počet bodov	1. trieda	2. trieda
0-2	5	4
3-5	10	12
6-8	25	30
9-11	15	6
nad 12	5	8

Na hladine významnosti $\alpha = 0,05$ zistite, či tieto triedy patria do toho istého základného súboru.

Výber testu:

Uvedený príklad je testom zhody rozdelení, lebo rozdeľuje výsledky pozorovania do určitých tried (porovnávame bodové ohodnotenie v 2 triedach) a skúma mieru ich zhody. Použijeme Kolmogorov Smirnovov test pre 2 výbery.

Použitie Kolmogorovho-Smirnovovho testu: Kolmogorov-Smirnovov test významnosti zhody dvoch rozdelení pracuje s kumulatívnymi početnosťami. Možno ním testovať zhodu medzi empirickým rozdelením a teoretickým daným testovanou hypotézou. Dá sa použiť aj v prípade, že porovnávame dve skupiny (2 výbery) navzájom, t.j. či tieto výbery pochádzajú z toho istého základného súboru alebo nie. Pri testovaní postupujeme nasledovne:

- zvolíme si hladinu významnosti α ,
- určíme, či ide o test zhody rozdelenia pre jeden výber alebo dva výbery,

Kolmogorov-Smirnovov test pre dva nezávislé výbery

a) ak $n_1, n_2 \leq 40$ a $n_1 = n_2$, potom ako testovacie kritérium použijeme $D_2 = \max |N_{1j} - N_{2j}|$, kde N_{1j} a N_{2j} predstavujú absolútne kumulatívne početnosti, v tabuľkách kritických hodnôt si vyhľadáme pre dané n a zvolenú hladinu významnosti α príslušnú kritickú hodnotu D_α ,

- ak $D_2 \geq D_\alpha \Rightarrow H_0$ zamietame a tvrdíme, že dva nezávislé výbery sa líšia štatisticky významne,

b) ak $n_1, n_2 > 40$ alebo $n_1 \neq n_2$, potom ako testovacie kritérium použijeme $D_2 = \max |F_{1j} - F_{2j}|$, kde F_{1j} a F_{2j} predstavujú relatívne kumulatívne početnosti a pre

hladinu významnosti α si vypočítame príslušnú kritickú hodnotu D_α podľa nasledovných vzorcov: $D_\alpha = D_{0,05} = 1,36 \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}$ alebo $D_\alpha = D_{0,01} = 1,63 \cdot \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}}$, kde $n = \frac{n_1 \cdot n_2}{n_1 + n_2}$, ak $D_2 \geq D_\alpha \Rightarrow H_0$ zamietame a tvrdíme, že sa empirické a teoretické rozdelenie sa veľmi líši.

Riešenie:

Využijeme Kolmogorov-Smirnovov test pre dva nezávislé výbery pre $n_1, n_2 > 40$.

Formulácia nulovej hypotézy H_0 : Medzi výsledkami testov v prvej a druhej triede nie je štatisticky významný rozdiel.

Počet bodov	n_{1j}	n_{2j}	N_{1j}	N_{2j}	f_{1j}	f_{2j}	F_{1j}	F_{2j}	$ F_{1j} - F_{2j} $
0-2	5	4	5	4	0,083	0,067	0,083	0,067	0,016
3-5	10	12	15	16	0,167	0,2	0,25	0,267	0,017
6-8	25	30	40	46	0,417	0,5	0,667	0,767	0,1
9-11	15	6	55	52	0,25	0,1	0,917	0,867	0,05
nad 12	5	8	60	60	0,083	0,133	1	1	0

$$D_2 = \max |F_{1j} - F_{2j}| = 0,1$$

Pre $n = \frac{60 \cdot 60}{60 + 60} = 30$ pri zvolenej hodnote $\alpha = 0,05$ sa $D_\alpha = 0,242$ (tab. 20)

$$D_2 < D_\alpha \Rightarrow H_0 \text{ prijímame}$$

Odpoveď: Medzi výsledkami testu v prvej a druhej triede nie je štatisticky významný rozdiel.

Neriešené úlohy:

1. Vybrali sme 100 respondentov, u ktorých sme sledovali výskyt chrípky. V uvedenej vzorke respondentov sa choroba prejavila v priemere u 30 % z nich. Z dlhodobého pozorovania však vieme, že smerodajná odchýlka má hodnotu 2,5. Určte 95 % interval spoľahlivosti pre výskyt chrípky.

2. V prípade 16 náhodne vybraných študentov sme zaznamenali nasledovné hodnoty telesnej výšky: 165, 166, 168, 169, 170, 172, 173, 174, 175, 176, 176, 177, 178, 178, 179, 182 cm. Určte 95 % interval spoľahlivosti pre telesnú výšku.

3. V tabuľke sú uvedené výsledky o počte bodov zo zápasov u vybraného basketbalového mužstva počas dvoch sezón. Zistite, či sa v prípade daného basketbalového mužstva prejavil štatisticky významný rozdiel medzi prvou a druhou sezónou, ak viete, že hráči venovali tréningu počas druhej sezóny väčší počet hodín. Zápas sa hral počas oboch sezón s tými istými mužstvami. Poradie zápasov bolo v oboch sezónach rovnaké. Predpokladajme, že u iných basketbalových mužstiev sa počet hodín tréningu počas oboch sezón nemenil.

Zápasy	Počet bodov zo zápasov získaných v sezóne 2007/08	Počet bodov zo zápasov získaných v sezóne 2008/09
1.	30	60
2.	55	55
3.	77	82
4.	66	80
5.	76	70
6.	58	76
7.	60	90
8.	70	49
9.	60	79
10.	87	86

10 Závislosti

Hromadné javy sa nevyskytujú v realite izolovane, ale vždy v určitých vzťahoch s inými javmi. Jedným z cieľov je teda poznanie **závislosti javov**. Ak pri výskume prvkov štatistického súboru je predmetom nášho záujmu **vzťah viacerých štatistických znakov**, hovoríme o **viacrozmerných štatistických súboroch**. Ak skúmame **vzťah dvoch štatistických znakov** (štatistický súbor sa skladá z dvojíc hodnôt $[X_i, Y_i]$ týkajúcich sa sledovanej udalosti, kde každú hodnotu možno spracovať samostatne), ide o **dvojrozmerné štatistické súbory**.

Podľa stupňa závislosti jedného javu od druhého rozoznávame dva základné druhy závislosti:

1. pevná závislosť – zmena jedného javu spôsobí zmenu javu druhého, a to v úplne presne zodpovedajúcej intenzite. Priebeh tejto závislosti je možné charakterizovať určitou matematickou funkciou, preto sa označuje ako **funkčná závislosť**. Vyskytuje sa u väčšiny prírodných javov – napr. závislosť objemu telesa od teploty. Objem telesa je v tomto prípade **závislou premennou** (závisí od teploty), teplota **nezávislou**.

Ide o **pevnú (funkčnú) závislosť** $t = f(p)$, t - závislá premenná, p - nezávislá premenná.

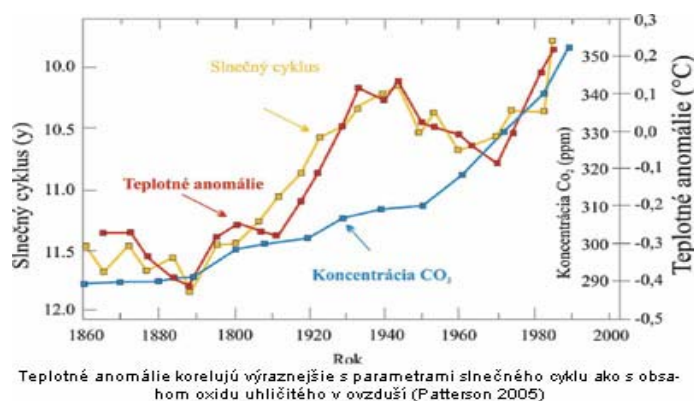
2. voľná závislosť – jeden jav nám podmieňuje jav druhý len s určitou pravdepodobnosťou a rôznou intenzitou.

Pri tomto type závislosti je možné charakterizovať len teoretický priebeh a pevnosť tejto závislosti. Vyskytuje sa omnoho častejšie ako pevná závislosť, a to najmä pri spoločenských a biologických javoch (jazda pod vplyvom alkoholu zvyšuje pravdepodobnosť dopravnej nehody, avšak vodič túto dopravnú nehodu nemusí spôsobiť). Takáto závislosť je označovaná aj ako **stochastická (pravdepodobnostná) závislosť**.

ZÁVISLOSŤ MEDZI KVANTITATÍVNymi ZNAKMI

Korelačná závislosť vyjadruje závislosť medzi kvantitatívnymi znakmi, t. j. vplyv zmeny nezávisle premenných na zmenu závisle premenných veličín (graf 16). Pri skúmaní korelačnej závislosti charakterizujeme **regresiu** (viac v kapitole 11) a **koreláciu**.

Graf 16 Porovnanie teplotných anomálií, slnečného cyklu a koncentrácie CO₂



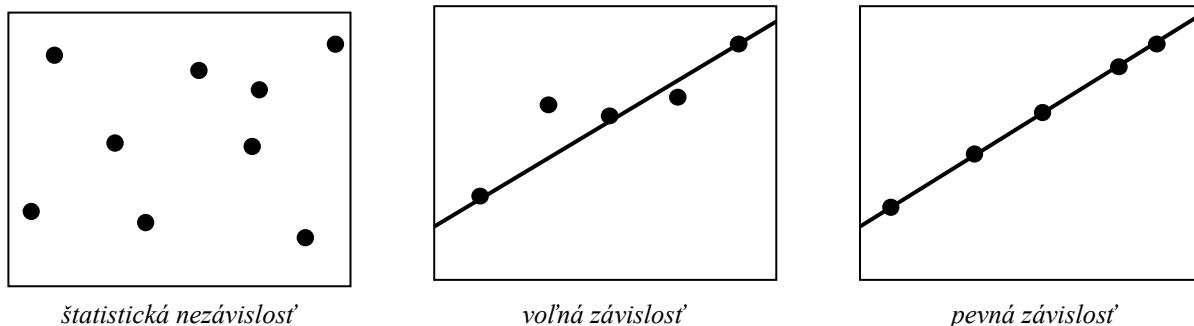
Spracované podľa: http://www.quark.sk/images/clanky/2005124/05125_obzalovany2.jpg

Pri štatistickom skúmaní závislosti riešime dve základné úlohy:

1. Zistujeme **priebeh závislosti** (riešime regresnú úlohu), t. j. odhadujeme funkčný vzťah f , podľa ktorej sa mení hodnota závislej premennej Y_i pri zmene hodnôt nezávisle premennej X_i (viac v kapitole 11).
2. Snažíme sa posúdiť **tesnosť (silu závislosti)**, t. j. určiť charakteristiky, ktoré informujú o tom, do akej miery vplyvajú nezávislé premenné X_i na variabilitu závislej premennej Y_i (riešime korelačnú úlohu).

Korelácia charakterizuje tesnosť (stupeň) závislosti. Vzhľadom na to, aká silná je závislosť medzi jednotlivými štatistickými znakmi, možno hovoriť o štatistickej nezávislosti, voľnej alebo pevnej závislosti (obr. 2).

Obr. 2 Stupeň korelačnej závislosti



Spracované podľa: <http://www.fhvp.unipo.sk/cvt/statistika/zlacka/geoinfo6.pdf>

Silu lineárnej závislosti medzi premennými, t. j. mieru vplyvu variability nezávisle premenných na závislú premennú možno vyjadriť pomocou **Pearsonovho koeficientu korelácie** r_{xy} . Ten môže nadobúdať hodnoty z intervalu od $\langle -1, 1 \rangle$. Ak je $r_{xy} < 0$, ide o zápornú lineárnu závislosť, ak $r_{xy} > 0$, kladnú lineárnu závislosť. Keď $r_{xy} = 0$, ide o lineárnu nezávislosť. V prípade, že hodnoty sa pohybujú v intervale $\langle 0,7; 1 \rangle$, ide o **vysokú (silnú) priamu (kladnú) lineárnu závislosť**. Ak sú hodnoty z intervalu $\langle -1; -0,7 \rangle$, ide o **vysokú (silnú) nepriamu (zápornú) lineárnu závislosť**. Ak sú hodnoty z intervalu $\langle -0,5; 0,5 \rangle$ a pritom nie sú blízke nule, ide o **slabú priamu alebo slabú nepriamu lineárnu závislosť**.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2 \right]}}, \quad \text{kde } r_{xy} \in \langle -1, 1 \rangle$$

Ak $r_{xy} = -1$ alebo $r_{xy} = 1$, ide o funkčnú závislosť. Je potrebné však poznamenať, že medzi premennými nemusí existovať len lineárny vzťah. Existujú aj iné typy závislosti, napr. kvadratická, exponenciálna a podobne.

Koeficient determinácie je definovaný ako druhá mocnina koeficientu korelácie. Označujeme ho r_{xy}^2 . Reprezentuje proporciu spoločného rozptylu, teda na koľko % zmena jednej premennej ovplyvní druhú premennú. Koeficient determinácie nám teda hovorí, aké percento variability závisle premennej sa dá vysvetliť variabilitou nezávisle premennej.

Pri riešení korelačných úloh často využívame **korelačnú tabuľku**. Záhlavie tvoria hodnoty prvého znaku, legendu hodnoty druhého znaku. Do políčok sa vpisujú početnosti (tab. 11).

Tab. 11 Korelačná tabuľka

Znak Y Znak X	y_1	y_2	...	y_l	Spolu „ X “
x_1	n_{11}	n_{12}	...	n_{1l}	n_{1x_1}
x_2	n_{21}	n_{22}	...	n_{2l}	n_{2x_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	...	n_{kl}	n_{kx_k}
Spolu „ Y “	$n_{y_1 1}$	$n_{y_2 2}$	...	$n_{y_l l}$	n

Druhy korelačnej závislosti:

1. Podľa počtu kvantitatívnych znakov, variabilitu ktorých skúmame:

a) párovú závislosť - skúmame závislosť medzi dvoma premennými X a Y (napr. vzťah medzi rýchlosťou auta a prejdenou dráhou), teda uvažujeme iba o jednej nezávislej premennej,

b) viacnásobnú závislosť - predmetom záujmu je závislosť medzi premennou Y a viacerými premennými $X_1, \dots, X_m$ (Y je závislá premenná a $X_1, \dots, X_m$ sú nezávislé premenné),

b) parciálna korelácia - súvislosť medzi dvoma premennými Y a X_1 s vylúčením akéhokoľvek vplyvu ďalších premenných $X_2, \dots, X_m$.

2. Podľa typu regresnej funkcie:

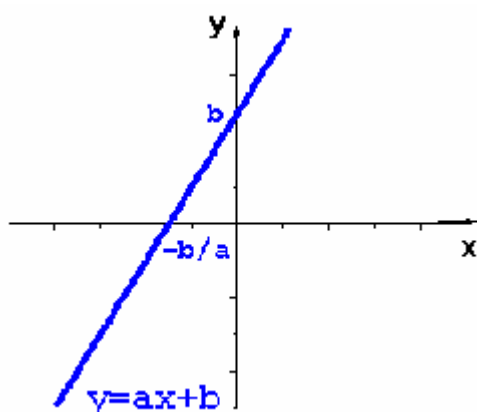
a) lineárna korelačná závislosť - zmeny jedného štatistického znaku sú približne lineárne závislé na zmenách druhého štatistického znaku (grafy 17, 19). Ak ide o jednoduchú lineárnu závislosť, grafom je priamka.

b) nelineárna korelačná závislosť - zmeny jedného a druhého štatistického znaku nie sú na sebe lineárne závislé (grafy 18, 20). Jednoduchú nelineárnu závislosť je možné vyjadriť pomocou regresnej krivky (paraboly, hyperboly, exponenciálnej funkcie).

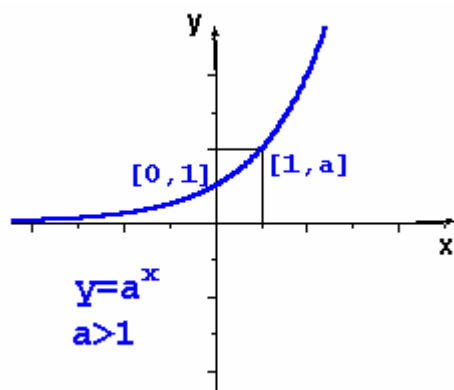
3. Podľa smeru zmien kvantitatívnych znakov:

a) kladná (pozitívna) - s rastom (poklesom) hodnôt jedného štatistického znaku rastú (klesajú) hodnoty i druhého štatistického znaku (grafy 17, 18)

Graf 17 Lineárna pozitívna korelačná závislosť



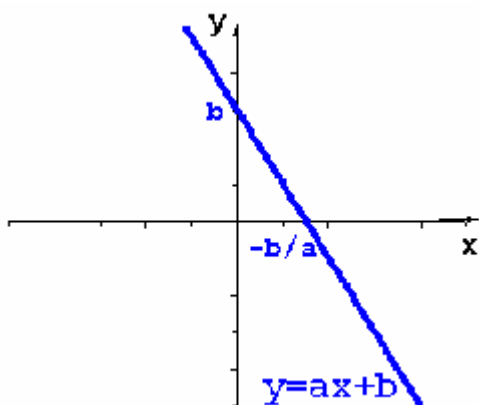
Graf 18 Nelineárna pozitívna korelačná závislosť



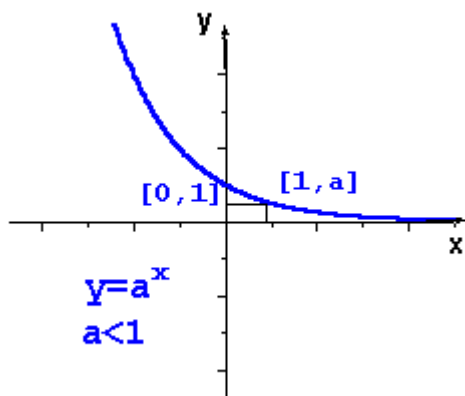
Spracované podľa: http://sis.science.upjs.sk/matematika/kniznica_vp/funkcie/Linear.htm,
http://sis.science.upjs.sk/matematika/kniznica_vp/funkcie/Kvadrat.htm

b) záporná (negatívna) - s rastom hodnôt jedného štatistického znaku klesajú hodnoty druhého štatistického znaku (grafy 19, 20)

Graf 19 Lineárna negatívna korelačná závislosť



Graf 20 Nelineárna negatívna korelačná závislosť



Spracované podľa: http://sis.science.upjs.sk/matematika/kniznica_vp/funkcie/Linear.htm,
http://sis.science.upjs.sk/matematika/kniznica_vp/funkcie/Kvadrat.htm

ZÁVISLOSŤ MEDZI KVALITATÍVNymi ZNAKMI

Závislosť medzi kvalitatívnymi štatistickými znakmi možno vyjadriť pomocou **asociácie** a **kontingencie**.

ASOCIÁCIA

Asociácia vyjadruje **závislosť medzi dvomi alternatívnymi kvalitatívnymi znakmi**. Základom pre jej výskum tvorí asociačná tabuľka. V jej v záhlaví sú uvedené štatistické znaky a ich obmeny, vnútro asociačnej tabuľky tvoria početnosti (tab. 12).

Tab. 12 Asociačná tabuľka

Znak B Znak A	b_1	b_2	Spolu „A“
a_1	n_{11}	n_{12}	n_{1a_1}
a_2	n_{21}	n_{22}	n_{2a_2}
Spolu „B“	n_{b_11}	n_{b_22}	n

Druhy asociačnej závislosti:

- a) **kladná asociácia** – s rastom (poklesom) početnosti štatistického znaku A rastú (klesajú) početnosti štatistického znaku B,
- b) **záporná asociácia** – s rastom (poklesom) početnosti štatistického znaku A klesajú (rastú) početnosti štatistického znaku B,
- c) **asociačná nezávislosť** – štatistické znaky sú vzájomne nezávislé.

Intenzita asociácie pre dvojrozmerné štatistické súbory:

Intenzitu asociácie možno vypočítať pomocou koeficientu asociácie alebo upraveným koeficientom korelácie. Oba koeficienty môžu nadobúdať hodnoty z intervalu (-1,1). Pri nezávislosti sa ich hodnoty blížia k 0. Pri kladnej asociácii sú ich hodnoty kladné, pri zápornej asociácii záporné. Čím sa hodnoty blížia viac k 1, resp. k -1, tým je asociácia silnejšia.

Koeficient asociácie

$$Q_{AB} = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{n_{11} \cdot n_{22} + n_{12} \cdot n_{21}},$$

Koeficient korelácie kvalitatívnych znakov

$$R_{AB} = \frac{n \cdot n_{11} - n_{b_11} \cdot n_{1a_1}}{\sqrt{n_{1a_1} \cdot n_{2a_2} \cdot n_{b_11} \cdot n_{b_22}}}$$

KONTINGENCIA

Kontingencia predstavuje **závislosť medzi množnými kvalitatívnymi znakmi**. (Pojem kontingencia možno použiť i v prípade, že skúmame napr. vzťah medzi kvalitatívnym množným znakom a kvalitatívnym alternatívnym znakom). Schematické znázornenie kontingenčného vzťahu vyjadruje kontingenčná tabuľka (tab. 13). Obdobne ako pri korelačnej a asociačnej tabuľke i tu tvoria záhlavie štatistické znaky a ich obmeny, v tele tabuľky sú informácie o absolútnych početnostiach, ktoré nadobúdajú.

Tab. 13 Kontingenčná tabuľka

Znak A \ Znak B	b_1	b_2	...	b_l	Spolu „A“
a_1	n_{11}	n_{12}	...	n_{1l}	n_{1a_1}
a_2	n_{21}	n_{22}	...	n_{2l}	n_{2a_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_k	n_{k1}	n_{k2}	...	n_{kl}	n_{ka_k}
Spolu „B“	n_{b_1}	n_{b_2}	...	n_{b_l}	n

Intenzita kontingencie:

Intenzitu kontingencie možno vyjadriť napr. pomocou Pearsonovho koeficientu kontingencie, Čuprovho koeficientu kontingencie, resp. Cramerovho koeficientu. Základom všetkých uvedených koeficientov tvorí **štvorcová kontingencia**, označovaná ako χ^2 . Jej veľkosť možno vypočítať z nasledovného vzťahu:

$$\chi^2 = n \cdot \left[\left(\sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij})^2}{(n_{ai})(n_{bj})} \right) - 1 \right],$$

kde n predstavuje rozsah súboru, k počet obmien znaku A, l počet obmien znaku B a n_{ij} absolútne početnosti, ktoré môžu obmeny uvedených znakov A, B nadobúdať (pozri kontingenčnú tabuľku). Následne možno vypočítať intenzitu kontingencie podľa jedného z uvedených vzťahov:

Pearsonov koeficient

kontingencie

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

Čuprov koeficient

kontingencie

$$K = \frac{\chi^2}{n\sqrt{(k-1)(l-1)}},$$

Cramerov koeficient

kontingencie

$$V = \sqrt{\frac{\chi^2}{n \cdot h}},$$

kde n predstavuje rozsah súboru, χ^2 štvorcovú kontingenciu, k počet obmien znaku A, l počet obmien znaku B a h minimum z $(k-1)$ a $(l-1)$.

Pearsonov koeficient kontingencie môže nadobúdať hodnoty z intervalu $(0,1)$. Čím viac sa jeho hodnota blíži k 0, tým je závislosť medzi kvalitatívnymi znakmi slabšia. Pri úplnej závislosti sa jeho hodnoty blížia k 1. Nevýhodou uvedeného kontingenčného vzťahu je, že maximálna hodnota Pearsonovho koeficientu kontingencie je limitovaná veľkosťou kontingenčnej tabuľky (v tabuľke 2x2 je to 0,707). Z tohto dôvodu je vhodné ako doplnujúce kritérium využiť Cramerov koeficient kontingencie, ktorý môže nadobúdať hodnoty od 0 (nezávislosť) po 1 (pevná závislosť). Pre jeho interpretáciu možno využiť aj škálu pre korelačný koeficient, kde koreláciu menšiu ako 0,1 možno považovať za triviálnu, od 0,1

do 0,3 za malú, od 0,31 do 0,5 za strednú a nad 0,5 za veľkú. V prípade Čuprovho koeficientu kontingencie vyjadruje hodnota 0 neexistenciu závislosti. Ak obidva znaky majú rovnaký počet obmien, pohybuje sa hodnota koeficientu od 0 do 2. Pri nerovnakom počte obmien nedosiahne koeficient hodnotu 1 ani pri úplnej kontingencii.

Riešené úlohy:

1. Zmerajte stupeň asociácie medzi vybavením domácností murovaným jadrom (MJ) a plastovými oknami (PO).

MJ \ PO	mám	nemám
mám	70	42
nemám	30	8

Riešenie:

Asociačnú tabuľku si najprv doplníme, následne vypočítame intenzitu asociácie pomocou koeficientu asociácie a koeficientu korelácie kvalitatívnych znakov:

MJ \ PO	b ₁ mám	b ₂ nemám	Spolu A
a ₁ mám	70	42	112
a ₂ nemám	30	8	38
Spolu B	100	50	150

$$Q_{AB} = \frac{n_{11} \cdot n_{22} - n_{12} \cdot n_{21}}{n_{11} \cdot n_{22} + n_{12} \cdot n_{21}} = \frac{70 \cdot 8 - 42 \cdot 30}{70 \cdot 8 + 42 \cdot 30} = \frac{560 - 1260}{560 + 1260} = \frac{-700}{1820} \approx -0,38$$

$$R_{AB} = \frac{n \cdot n_{11} - n_{b_1} \cdot n_{1a_1}}{\sqrt{n_{1a_1} \cdot n_{2a_2} \cdot n_{b_1} \cdot n_{b_2}}} = \frac{150 \cdot 70 - 100 \cdot 112}{\sqrt{112 \cdot 38 \cdot 100 \cdot 50}} = \frac{10500 - 11200}{\sqrt{21280000}} = \frac{-700}{4613,025} \approx -0,15$$

Odpoveď: Vybavenie domácností murovaným jadrom je nepriamo úmerne závislé od vybavenia domácností plastovými oknami. Vzhľadom na to, že výsledné hodnoty intenzity asociácie sú blízke 0, možno konštatovať, že ide o slabú asociačnú závislosť.

2. Pomocou údajov z kontingenčnej tabuľky vypočítajte stupeň kontingencie medzi druhom absolvovanej strednej školy (DŠ)- znak A a prijatím do zamestania (PZ)- znak B.

DŠ \ PZ	prijali	neprijali
SOŠ	15	13
gymnázium	20	12
iné školy	5	2

Riešenie:

Kontingenčnú tabuľku si najprv doplníme, následne vypočítame intenzitu kontingencie pomocou Pearsonovho, Čuprovho a Cramerovho koeficientu kontingencie.

<i>DŠ \ PZ</i>	b ₁ prijali	b ₂ neprijali	Spolu A
a ₁ SOŠ	15	13	28
a ₂ gymnázium	20	12	32
a ₃ iné školy	5	2	7
Spolu B	40	27	67

$$\begin{aligned}\chi^2 &= n \cdot \left[\left(\sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij})^2}{(n_{ai})(n_{bj})} \right) - 1 \right] = \\ &= n \cdot \left[\frac{n_{11}^2}{n_{1a_1} \cdot n_{b_1}} + \frac{n_{12}^2}{n_{1a_1} \cdot n_{b_2}} + \frac{n_{21}^2}{n_{2a_2} \cdot n_{b_1}} + \frac{n_{22}^2}{n_{2a_2} \cdot n_{b_2}} + \frac{n_{31}^2}{n_{3a_3} \cdot n_{b_1}} + \frac{n_{32}^2}{n_{3a_3} \cdot n_{b_2}} - 1 \right] = \\ &= 67 \cdot \left[\frac{15^2}{28 \cdot 40} + \frac{13^2}{28 \cdot 27} + \frac{20^2}{32 \cdot 40} + \frac{12^2}{32 \cdot 27} + \frac{5^2}{7 \cdot 40} + \frac{2^2}{7 \cdot 27} - 1 \right] \approx 0,9416\end{aligned}$$

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{0,9416}{67 + 0,9416}} \approx 0,12$$

$$K = \frac{\chi^2}{n \sqrt{(k-1)(l-1)}} = \frac{0,9416}{67 \cdot \sqrt{2 \cdot 1}} \approx 0,01$$

$$V = \sqrt{\frac{\chi^2}{n \cdot h}} = \sqrt{\frac{0,9416}{67 \cdot 1}} \approx 0,12$$

Odpoveď: Medzi druhom absolvovanej strednej školy a prijatím do zamestnania sa prejavil len malý stupeň kontingenčnej závislosti.

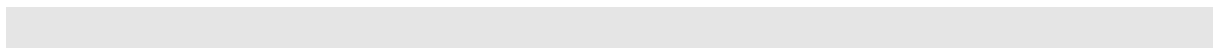
Neriešené úlohy:

1. Zmerajte stupeň asociácie medzi vlastníctvom auta u mužov a žien. Informácie sú uvedené v tabuľke.

<i>vlastníctvo auta</i> <i>pohlavie</i>	<i>vlastním</i>	<i>nevlastním</i>
žena	33	15
muž	70	20

2. Pomocou údajov z kontingenčnej tabuľky vypočítajte stupeň kontingencie medzi typom ukončenej školy (UŠ) a zamestnaním (Z).

<i>UŠ</i> <i>Z</i>	<i>číšnik</i>	<i>chyžná</i>	<i>recepčný</i>
SOU s maturitou	5	8	0
SOŠ	5	2	10
vysoká škola	2	0	8



11 Regresná analýza

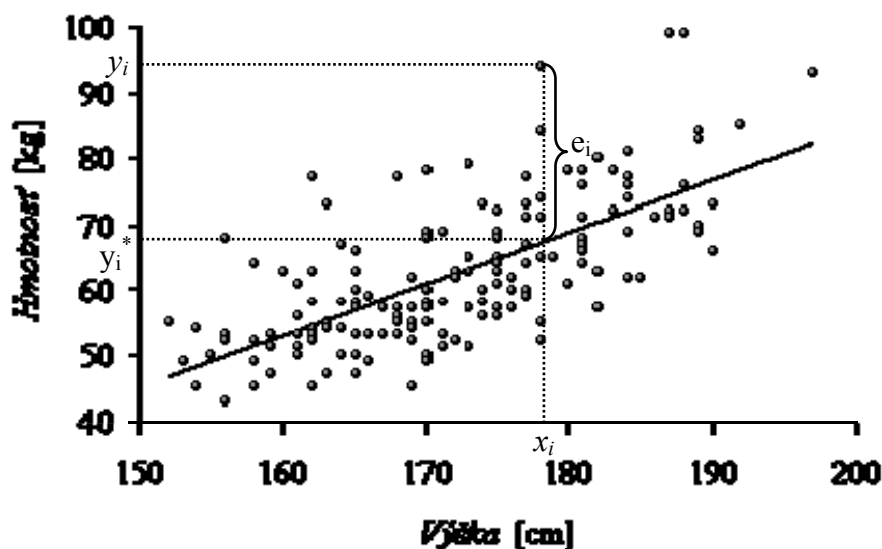
Regresia, regresné modely

Pri riešení regresnej úlohy sa snažíme vyjadriť vzťah medzi dvoma premennými (závisle a nezávisle premennou) pomocou matematickej funkcie, ktorú nazývame **regresná funkcia**.

Jednoduchá lineárna regresia

Najjednoduchšou formou korelácie je lineárna korelácia medzi dvoma kvantitatívnymi znakmi. Najčastejšou metódou odhadu lineárnych regresných funkcií je **metóda najmenších štvorcov**.

Graf 21 Lineárna regresia



Cieľom metódy je **odhadnúť parametre regresnej funkcie** tak, aby bolo možné preložiť danými bodmi regresnú priamku, ktorá by najlepšie vystihovala priebeh závislosti (graf 21).

$$y_i^* = a + bx_i,$$

kde

y_i^* - závislá premenná (označuje odhadnuté hodnoty ležiace na regresnej priamke),

x_i - nezávislá premenná,

b - regresný koeficient (smernica priamky).

Teoretické (ideálne) hodnoty ležia na priamke a sú označené ako y_i^* . **Empirické hodnoty (skutočné)** sú hodnotami závislej premennej y_i . Z rôznych možností, ktorými je

možné preložiť regresnú priamku, je najvhodnejšia tá alternatíva, pri ktorej sa súčet odchýlok empirických hodnôt od teoretických bude rovnať 0.

$$\sum_{i=1}^n (y_i - y_i^*) = 0$$

$$\sum_{i=1}^n (y_i - y_i^*)^2 = \min.$$

Rozdiel medzi skutočnou hodnotou y_i a vypočítanou (odhadnutou) hodnotou y_i^* je reziduálna odchýlka e_i .

$$e_i = y_i - y_i^*$$

Pretože rovnica priamky má tvar $y_i^* = a + bx_i$, možno príslušnú funkciu prepísať

$$f(a, b) = \sum_{i=1}^n (y_i - y_i^*)^2 \Rightarrow f(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2 \rightarrow \min.$$

Parametre regresnej priamky sa získavajú metódou najmenších štvorcov. Podstatou tejto metódy je minimalizácia súčtu kvadratických odchýlok teoretických a skutočných hodnôt premennej y . Ak predpokladaná funkčná závislosť (funkcia) má prvé parciálne derivácie, tak odhad koeficientov a , b je výsledkom riešenia systému lineárnych (normálnych) rovníc:

$$\frac{\partial f(a, b)}{\partial a} = 0, \quad \frac{\partial f(a, b)}{\partial b} = 0.$$

$$a = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}, \quad b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

b - *smernica priamky (regresný koeficient)* – udáva, o koľko sa v priemere zmení závislá premenná y_i^* pri zmene nezávisle premennej x_i . Zároveň vyjadruje určitý sklon regresnej priamky.

V prípade, že $b > 0$, regresná priamka je rastúca, hovoríme o *priamej závislosti*.

V prípade, že $b < 0$, regresná priamka je klesajúca, hovoríme o *nepriamej závislosti*.

Ak $b = 0$ *závislosť nejestvuje*.

a – *lokujúca konštanta* – vyjadruje, akú hodnotu bude mať závislá premenná y_i^* , ak nezávislá premenná x_i bude mať hodnotu 0. Pri grafickom znázornení regresnej priamky nám lokujúca konštanta a udáva bod, v ktorom priamka pretne os y . Jej hlavný zmysel spočíva v tom, že posúva regresnú priamku v priestore.

Niekedy je vhodné vykonať transformáciu nelineárnych regresných funkcií na lineárnu regresiu.

Napr. $y = ax^b$ môžeme transformovať na funkciu: $\log y = \log a + b \cdot \log x$.

V prípade progresívnej alebo regresívnej závislosti je potrebné použiť kvadratickú (polynomicnú) regresiu v tvare: $y = a + bx + cx^2$.

$$a = \frac{1}{D} \begin{vmatrix} \sum y & \sum x & \sum x^2 \\ \sum xy & \sum x^2 & \sum x^3 \\ \sum x^2 y & \sum x^3 & \sum x^4 \end{vmatrix} \quad b = \frac{1}{D} \begin{vmatrix} n & \sum y & \sum x^2 \\ \sum x & \sum xy & \sum x^3 \\ \sum x^2 & \sum x^2 y & \sum x^4 \end{vmatrix}$$

$$c = \frac{1}{D} \begin{vmatrix} \sum n & \sum x & \sum y \\ \sum x & \sum x^2 & \sum xy \\ \sum x^2 & \sum x^3 & \sum x^2 y \end{vmatrix} \quad D = \begin{vmatrix} n & \sum x & \sum x^2 \\ \sum x & \sum x^2 & \sum x^3 \\ \sum x^2 & \sum x^3 & \sum x^4 \end{vmatrix}$$

Viacnásobná lineárna regresia

Používa sa pri štúdiu lineárnej závislosti medzi dvoma a viacerými premennými. Najdôležitejším kritériom linearity je (Pearsonov) korelačný koeficient r_{xy} .

$$y = a + b_1 x_1 + b_2 x_2.$$

Nelineárne regresné modely

Často ide o nejakú empirickú alebo fyzikálnu závislosť. Na rozdiel od lineárnej regresie sa pre odhad minima používajú iteračné metódy (minimum súčtu štvorcov odchýlok).

Riešené úlohy:

1. V tabuľke sú uvedené informácie o veku a počte dní dovolenky u 10 pracovníkov. Predpokladajme, že medzi počtom dní dovolenky a vekom pracovníka je lineárna závislosť. Vypočítajte parametre regresnej priamky a uveďte stupeň korelácie medzi štatistickými znakmi. Pomocou koeficientu determinácie určte, koľko percent variability závislej premennej sa dá vysvetliť variabilitou nezávislej premennej. Vypočítajte súhrnnú hodnotu štvorcov reziduálnych odchýlok medzi skutočným a odhadnutým počtom dní dovolenky.

Riešenie:

Vek (x_i)	Počet dní (y_i)	x_i^2	$x_i y_i$	y_i^*	$y_i - y_i^*$	$e_i^2 = (y_i - y_i^*)^2$	y_i^2
18	1	324	18	1,1	-0,1	0,01	1
20	1	400	20	1,3	-0,3	0,09	1
30	2	900	60	2,3	-0,3	0,09	4
40	2	1600	80	3,3	-1,3	1,69	4
50	4	2500	200	4,3	-0,3	0,09	16
$\Sigma=158$	$\Sigma=10$	$\Sigma=5724$	$\Sigma=378$			$\Sigma=1,97$	

$$a = \frac{\sum_{i=1}^5 x_i^2 \sum_{i=1}^5 y_i - \sum_{i=1}^5 x_i \sum_{i=1}^5 x_i y_i}{n \sum_{i=1}^5 x_i^2 - \left(\sum_{i=1}^5 x_i \right)^2} = \frac{5724 \cdot 10 - 158 \cdot 378}{5 \cdot 5724 - 158^2} = \frac{57240 - 59724}{28620 - 24964} = \frac{-2484}{3656} \approx -0,7$$

$$b = \frac{n \cdot \sum_{i=1}^5 x_i y_i - \sum_{i=1}^5 x_i \sum_{i=1}^5 y_i}{n \sum_{i=1}^5 x_i^2 - \left(\sum_{i=1}^5 x_i \right)^2} = \frac{5 \cdot 378 - 158 \cdot 10}{5 \cdot 5724 - 158^2} = \frac{1890 - 1580}{28620 - 24964} = \frac{310}{3656} \approx 0,1$$

Rovnica regresnej priamky má tvar: $y_i^* = -0,7 + 0,1x_i$. Odhadnuté hodnoty y_i^* vypočítame jednoduchým dosadením hodnôt x_i do rovnice regresnej priamky, kde $i=1,2,...5$. Súhrnú hodnotu štvorcov reziduálnych odchýlok Σe^2 získame ako sumu štvorcov rozdielov skutočných a odhadnutých dní absencie (pozri tabuľku vyššie). Stupeň korelácie medzi štatistickými znakmi určíme pomocou Pearsonovho korelačného koeficientu. Jeho umocnením dostaneme koeficient determinácie. Pomocné výpočty sú uvedené v tabuľke.

$$r_{xy} = \frac{\sum_{i=1}^5 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^5 (x_i - \bar{x})^2 \sum_{i=1}^5 (y_i - \bar{y})^2 \right]}} = \frac{62}{\sqrt{731,2 \cdot 6}} \approx 0,94, \quad r_{xy}^2 = 0,94^2 \approx 0,88$$

$$\bar{x} = \frac{\sum_{i=1}^5 x_i}{n} = \frac{158}{5} \approx 31,6 \quad \bar{y} = \frac{\sum_{i=1}^5 y_i}{n} = \frac{10}{5} = 2$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
18	1	-13,6	-1	184,96	1	13,6
20	1	-11,6	-1	134,56	1	11,6
30	2	-1,6	0	2,56	0	0
40	2	8,4	0	70,56	0	0
50	4	18,4	2	338,56	4	36,8
$\Sigma=158$	$\Sigma=10$			$\Sigma=731,2$	$\Sigma=6$	$\Sigma=62$

Odpoveď: Medzi vekom pracovníkov a počtom dní absencie sa prejavuje silná korelačná závislosť. Z koeficientu determinácie vyplýva, že 88 % variability závislej premennej (počet dní dovolenky) sa dá vysvetliť variabilitou nezávislej premennej (vek pracovníka).

Neriešené úlohy:

1. Predpokladajme, že vzťah medzi príjmami a výdavkami za potraviny by sme mohli vyjadriť prostredníctvom regresnej priamky. Vypočítajte jej parametre a uveďte stupeň korelácie medzi štatistickými znakmi. Pomocou koeficientu determinácie určte, koľko percent variability závislej premennej sa dá vysvetliť variabilitou nezávislej premennej. Vypočítajte súhrnú hodnotu štvorcov reziduálnych odchýlok medzi skutočnými a odhadnutými výdavkami za potraviny. Pomocné údaje sú uvedené v tabuľke.

Príjmy v EUR	Výdavky v EUR
300	100
500	200
650	250
800	300
1000	380

12 Analýza viacrozmerných štatistických údajov

Viacrozmerné údaje sa často vyskytujú pri skúmaní vlastností krajiny. Napríklad sa sledujú fyzikálne a chemické vlastnosti vody vo vodných tokoch, jazerách alebo sa sledujú rôzne charakteristiky obyvateľstva na určitom území (veková, príjmová, vzdelanostná a náboženská štruktúra).

Účelom viacrozmernej analýzy je najmä skúmanie vzťahov medzi jednotlivými atribútmi, vlastnosťami štatistického súboru.

Charakteristiky viacrozmerných údajov

Údaje sú charakterizované **zdrojovou maticou** ($n * m$):

$$X = \begin{bmatrix} x_1 \\ \dots \\ x_i \\ \dots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n,1} & \dots & x_{n,j} & \dots & x_{n,m} \end{bmatrix}$$

kde riadky n_i predstavujú jednotlivé **objekty** a stĺpce m_j ich vlastnosti, premenné. Jednotlivé objekty sú teda tvorené vektormi $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,m})$. Premenné môžu byť buď kvantitatívne alebo kvalitatívne (napr. rodinný stav osoby).

Keďže merané vlastnosti majú niekedy rôzne jednotky, ktoré sa aj rádovo líšia, ich hodnoty bývajú niekedy ešte pred štatistickým spracovaním škálované centrovaním (od hodnôt sa odčítava ich aritmetický priemer) alebo normovaním (hodnoty sa predelia smerodajnou odchýlkou).

Najjednoduchšie vyjadrenie podobnosti objektov (napr. vlastnosti auta, jeho zaradenie do triedy áut) je pomocou **miery vzdialenosti** objektov v m -rozmernom priestore premenných.

Štandardným typom vzdialenosti je geometrická (euklidovská) vzdialenosť:

$$d_E(\mathbf{x}_k, \mathbf{x}_l) = \sqrt{\sum_{j=1}^m (x_{kj} - x_{lj})^2}$$

Dva objekty sú si tým bližšie (podobnejšie), čím je vzdialenosť menšia. Používajú sa aj iné definície vzdialenosti (Mahalanobisova, Hammingova). Pri týchto mierach sa neuvažuje o závislosti medzi medzi premennými (Meloun, Militký, 2002).

Trieda, zhľuk objektov - množina objektov, ktoré majú rovnaké alebo aspoň podobné vlastnosti. Podobnosť, blízkosť objektov z hľadiska ich vlastností posudzujeme podľa miery vzdialenosti objektov v m -rozmernom priestore premenných.

Miera podobnosti dvoch objektov alebo premenných sa môže vyjadriť pomocou Pearsonovho korelačného koeficientu. Objekty sú si tým podobnejšie, čím je korelačný koeficient bližší jednotke. Na základe podobnosti objektov je možné objekty zgrupovať do tried (zhlukov) a tie tiež analyzovať obdobným spôsobom. Čím je väčšia vzdialenosť medzi triedami, tým je podobnosť menšia.

Všeobecný postup analýzy viacrozmerných údajov

Postup analýzy viacrozmerných údajov závisí od typu údajov a od toho, aké informácie chceme z údajov získať. Vo všeobecnosti postup pozostáva z týchto krokov (Meloun, Militký, 2002):

1. Analýza závislosti a nezávislosti údajov. Snažíme sa zistiť, či je možné analyzované premenné rozdeliť na závislé a nezávislé premenné a zistiť počet závisle premenných.
2. Ak medzi premennými existujú aj závisle premenné, určuje sa vzťah medzi závisle premennými a vhodná kombinácia nezávisle premenných. Ak sú všetky nezávislé, aplikujú sa metódy určenia vzájomných väzieb, t.j. realizuje sa simultánna analýza všetkých premenných.
3. Z hľadiska použitia konkrétnych metód je tiež dôležitý počet závisle premenných a v akej forme sú premenné. Pri číselných údajoch hovoríme o metrickej škále, pri nečíselných údajoch (kvalitatívnych) o nemetrickej škále.

Výber konkrétnej metódy analýzy údajov závisí od cieľa analýzy a počtu a typu závisle, resp. nezávisle premenných.

Charakteristiky viacrozmerných premenných

Každá vlastnosť, premenná v j-tom stĺpci, sa môže charakterizovať pomocou strednej hodnoty a pomocou rozptylu. Mieru intenzity vzťahu medzi dvoma premennými vyjadruje **kovariancia** $\text{cov}(x, y)$.

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

Vlastnosti kovariancie:

- znamienko ukazuje na trend stochastickej väzby medzi i-tým a j-tým stĺpcom matice,
- pre nekorelované veličiny sa rovná nule,
- je mierou intenzity lineárnej závislosti.

Nevýhodou kovariancie je, že jej hodnoty závisia od mierky, v ktorej sú vyjadrené premenné. Z toho dôvodu sa na vyjadrenie závislosti používa Pearsonov párový korelačný koeficient, ktorý nie je ovplyvnený jednotkami premenných.

$$r_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\left[\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i \right)^2 \right] \left[\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right]}}$$

Na základe týchto charakteristík vyjadrujeme kovariančnú alebo korelačnú maticu rádu $m \times n$. Korelačná matica má na diagonále jednotky a mimodiagonálne prvky sú jednotlivé párové korelačné koeficienty. Kovariačná i korelačná matica sú symetrické.

Podobne ako pri jednorozmerných súboroch, aj pri viacrozmerných je dôležité poznanie rozdelenia početností. Pokiaľ rozdelenie početností nie je typu normálneho rozdelenia (napríklad obsahuje rôzne extrémne, vybočujúce hodnoty), je vhodné použiť také charakteristiky a štatistické testy, ktoré nie sú príliš citlivé k týmto hodnotám.

Určenie štruktúry a väzieb v premenných a objektoch

Pred samotnou aplikáciou vhodnej metódy viacrozmernej štatistickej analýzy je potrebné vykonať *prieskumnú analýzu údajov*, ktorá umožňuje (Meloun, Militký, 2002):

- posúdiť *podobnosť objektov* pomocou rozptylových a symbolových grafov,
- nájsť *vybočujúce objekty*, resp. ich premenné,
- stanoviť, či je možné použiť predpoklad *lineárnych väzieb*,
- overiť *predpoklady o údajoch* (normalita, nekorelovanosť, homogenita).

Jednotlivé metódy na určenie vzájomných väzieb vo viacrozmerných štatistických súboroch sa delia podľa toho, či sa analyzuje štruktúra a väzby medzi premennými alebo štruktúra väzby v objektoch zdrojovej matice rozmeru $n \times m$. V prípade hľadania štruktúry v *premenných* v metrickej škále sa využívajú metódy *faktorovej analýzy*, *analýzy hlavných komponentov* a *zhluková analýza*. V prípade hľadania štruktúr v *objektoch* v metrickej škále sa využíva *zhluková analýza* a v nemetrickej škále *korešpondenčná analýza*.

Väčšina metód viacrozmernej štatistickej analýzy umožňuje *spracovanie lineárnych viacerozmerných modelov*, v ktorých sa závisle premenné považujú za lineárnu kombináciu nezávisle premenných, resp. väzby medzi premennými sú lineárne. V mnohých prípadoch sa tiež uvažuje o normalite metrických premenných.

Určením štruktúry a vzájomných väzieb medzi premennými, ale i medzi objektami sa zaoberajú metódy redukcie premenných na tzv. latentné premenné⁷: metóda *analýzy hlavných komponentov* (*Principal Component Analysis - PCA*) a metóda *faktorovej analýzy* (*Factor Analysis - FA*). Dôležitou metódou určenia vzájomných väzieb medzi premennými je aj *kanonická korelačná analýza*, ktorá sa používa na skúmanie závislosti medzi dvoma skupinami premenných, pričom jedna skupina sa považuje za premenné nezávislú a druhá za skupinu závislých premenných.

Hľadaním štruktúry a vzájomných väzieb v objektoch sa zaoberajú klasifikačné metódy viacrozmernej štatistickej analýzy. *Klasifikačné metódy* sú postupy, pomocou ktorých sa jeden objekt zaradí do jednej z existujúcich tried (*diskriminačná analýza DA*) alebo pomocou nich je možné neusporiadanú skupinu objektov usporiadať do niekoľkých vnútorne súrodých tried-zhlukov (zhluková, klastrová analýza, Cluster Analysis - CLU). Klastrová analýza sa v geografii využíva najmä v oblasti regionálnej analýzy na zníženie počtu subregiónov a ich začlenenie do menšieho počtu typov regiónov. Môže byť hierarchická alebo nehierarchická.

⁷ Latentné premenné sú priamo nepozorovateľné premenné, ktoré sa dajú opísať a merať len sprostredkované, ale majú schopnosť vecnej interpretácie

Analýza hlavných komponentov (PCA)

Cieľom tejto metódy je odhalenie kľúčového, nosného javu a odfiltrovanie šumu v údajoch. Prakticky to znamená, že sa snažíme transformovať pôvodné premenné x_j , $j=1, \dots, m$ do menšieho počtu i , tzv. latentných premenných y_i . Tieto premenné majú vhodnejšie vlastnosti, je ich menej, vystihujú takmer celú *premenlivosť* pôvodných premenných a sú vzájomne nekorelované (korelačný koeficient medzi latentnými premennými $y_1, \dots, y_n$ je 0).

Latentné premenné sa v tejto metóde nazývajú *hlavné komponenty* a sú lineárnou kombináciou pôvodných premenných: *prvý hlavný komponent* y_1 popisuje najväčšiu časť premenlivosti, čiže rozptylu pôvodných údajov, *druhý hlavný komponent* y_2 zas najväčšiu časť rozptylu neobsiahnutého v y_1 . Matematicky povedané, *prvý hlavný komponent* je takou lineárnou kombináciou vstupných premenných, ktorá zahŕňa najväčšiu premenlivosť medzi všetkými lineárnymi kombináciami. Má tvar (Meloun, Militký, 2002):

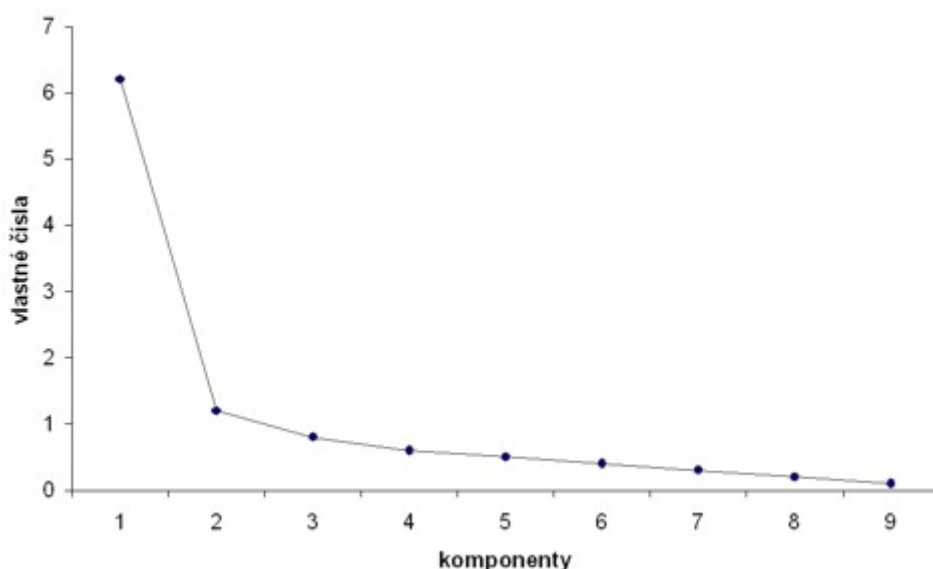
$$y_1 = \sum_{j=1}^m v_{1j} x_j = \mathbf{v}_1^T \mathbf{x},$$

kde objekt x_j obsahuje premenné $x_1, x_2, \dots, x_m$. Pre vektor koeficientov $\mathbf{v}_1^T = (v_{11}, \dots, v_{1m})^T$ platí, že premenlivosť vyjadrená rozptylom komponentu y_1 je maximálna. Vektor koeficientov $\mathbf{v}_i$, ktorý maximalizuje rozptyl premennej y_i , zodpovedá tzv. vlastnému vektoru a vlastnému číslu kovariančnej matice pôvodných údajov. Veľkosť vlastných čísel určuje význam a počet hlavných komponentov. Podobne sú konštruované aj ďalšie komponenty. Súčet rozptylov všetkých hlavných komponentov sa rovná súčtu rozptylov vstupných, pôvodných premenných. Ak je súčet prvých najväčších podielov dostatočne blízky 1 (100 %) - zvyčajne stačí 80-90 %, stačí zobrať do úvahy len týchto zopár hlavných komponentov na vysvetlenie variability pôvodných premenných. Zvyčajne to býva 2-5 komponentov. Tvorí *model hlavných komponentov*. Strata informácie spôsobená redukciou komponentov (rozmerov) sa nazýva tiež chybou modelu PCA. Výber počtu komponentov predstavuje samotný model hlavných komponentov PCA a jeho kľúčovou časťou je vysvetlenie použitých hlavných komponentov, ich pomenovanie a vysvetlenie vzťahu k pôvodným premenným.

Výsledok analýzy hlavných komponentov sa znázorňuje najmä graficky. *Indexový graf úpätia vlastných čísel* (nazýva sa tiež sutinový graf, scree plot) znázorňuje relatívnu početnosť jednotlivých vlastných čísel. Hlavné komponenty alebo faktory sú oddelené zreteľným zlomovým miestom na grafe a súradnica x tohto bodu je hľadaná hodnota indexu (graf 22). *Graf váh (záťaž) komponentov* zobrazuje váhy prvých dvoch hlavných komponentov, pričom sa porovnávajú vzdialenosti medzi premennými. Krátka vzdialenosť medzi dvoma premennými znamená silnú koreláciu. Tento graf môžeme tiež využiť aj na pochopenie, ako mierou prispievajú jednotlivé pôvodné premenné do hlavných komponentov. *Rozpyľový diagram komponentného skóre* sa používa na identifikáciu odláhlých objektov, identifikáciu trendov, tried a zhlukov objektov, objasnenie podobnosti objektov a podobne. Zobrazuje komponentné skóre, čiže hodnoty zvyčajne prvých dvoch hlavných komponentov všetkých objektov. Dokonalý rozptyl objektov v rovine oboch hlavných komponentov vedie k rozlíšeniu objektov pomocou y_1 a y_2 . V rovine je možné ľahko nájsť zhluk vzájomne podobných, ale aj odlišných objektov. Obvykle sa vyberajú diagramy, v ktorom porovnávame prvý hlavný komponent y_1 s druhým y_2 , tretím y_3 alebo štvrtým komponentom y_4 . Nie je potrebné analyzovať všetky kombinácie diagramov, ale držíme sa prvého hlavného komponentu, v ktorom je najväčšia miera variability údajov. Na základe polohy objektov od

začiatku môžeme posúdiť typickosť objektov (sú blízko začiatku, alebo naopak, ďaleko - sú to extrémny). Objekty blízko seba sú si podobné, ďaleko od seba nie sú si podobné. Pomocou diagramu je možné zistiť existenciu osamelých objektov alebo naopak, sústredených v zhlukoch s podobnými vlastnosťami. Zhluky je možné pomenovať a graficky načrtnúť možný vzťah medzi zhlukmi.

Graf 22 Indexový graf úpätia vlastných čísel



Pri aplikácii metódy hlavných komponentov sa stretávame s rôznymi problémami. Napríklad analyzované údaje neobsahujú predpokladanú informáciu, takže použitie metódy nemalo zmysel. Kritický je aj správny výber počtu hlavných komponentov, ktoré vystihnú daný jav. V analyzovanom štatistickom súbore sa môžu objaviť odľahlé objekty. Môžu byť prejavom chýb merania, ale môžu byť zdrojom zaujímavej informácie. Preto pri odstraňovaní odľahlých hodnôt je potrebné postupovať opatrne. Pred samotnou aplikáciou metódy pomocou softvéru je potrebná dôsledná úvaha nad analyzovaným javom a problém musí byť správne a presne definovaný. Údajové súbory je zvyčajne potrebné škálovať alebo aspoň centrovat'. Následne sa z pomocou indexového grafu úpätia vlastných čísel určí vhodný počet hlavných komponentov, vypočítajú sa vlastné vektory hlavných komponentov a komponentné váhy. V tejto fáze sa vyberie vhodný počet hlavných komponentov a tak sa vytvára model PCA. Pomocou diagramov sa vyšetrujú vlastnosti a úspešnosť modelu (Meloun, Militký, 2002).

V geografii sa metóda PCA využíva najmä na identifikáciu skupín korelovaných premenných a na zjednodušenie viacrozmerných údajových súborov redukciou vzájomne korelovaných premenných. Metóda vyžaduje viac pozorovaní (meraní) jednotlivých premenných, úplne minimálny počet je 10, avšak optimálny počet je niekoľko sto meraní.

Faktorová analýza (FA)

Podobne ako metóda hlavných komponentov, patrí aj faktorová analýza medzi metódy redukcie počtu pôvodných premenných. Vo faktorovej analýze predpokladáme, že každú vstupujúcu premennú môžeme vyjadriť ako lineárnu kombináciu neveľkého počtu spoločných skrytých faktorov a jediného chybového faktora. Na rozdiel od komponentnej analýzy sa pri faktorovej analýze snažíme vysvetliť závislosť premenných. K nevýhodám metódy patrí najmä nevyhnutnosť zadania *počtu spoločných faktorov* ešte pred realizáciou vlastnej analýzy.

Model faktorovej analýzy má v maticovej podobe tento tvar (Meloun, Militký, 2002):

$$\mathbf{x} = \mathbf{L}\mathbf{f} + \boldsymbol{\varepsilon}$$

resp.

$$x_1 = l_{11}f_1 + l_{12}f_2 + \dots l_{1m}f_m + \varepsilon_1,$$

$$x_2 = l_{21}f_1 + l_{22}f_2 + \dots l_{2m}f_m + \varepsilon_2,$$

.....

$$x_n = l_{n1}f_1 + l_{n2}f_2 + \dots l_{nm}f_m + \varepsilon_n,$$

kde $f_1, f_2, \dots, f_m$ sú faktory, ktoré vyvolávajú koreláciu medzi premennými a $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ sú chybové faktory, ktoré prispievajú k rozptylu jednotlivých premenných. Koeficienty l_{ik} nazývame faktorovými záťažami i-tého objektu a k-tého spoločného faktoru f_k , a predstavujú prvky matice faktorových záťaží.

Každý objekt pozorovaných premenných $\mathbf{x} = (x_1, x_2, \dots, x_m)$ má svoju korelačnú alebo kovariančnú maticu.

Kovariančnú maticu $\mathbf{C}$ vektora vstupných premenných (stĺpcov zdrojovej matice) môžeme zapísať v tvare základnej faktorovej vety:

$$\mathbf{C} = \mathbf{L}\mathbf{L}^T + \mathbf{G}^2,$$

kde $\mathbf{L}$ je matica faktorových záťaží a $\mathbf{L}\mathbf{L}^T$ je kovariančná matica vektora $\mathbf{L}\mathbf{f}$ a $\mathbf{G}^2$ je matica jedinečností. Matica jedinečností je kovariančnou maticou chybových faktorov. Jej diagonálne prvky tvoria rozptyly jednotlivých stĺpcov zdrojovej matice. Platí vzťah:

$$\mathbf{S}^2 = \mathbf{H}^2 + \mathbf{G}^2,$$

kde $\mathbf{S}^2$ je diagonálna matica rozpylov faktorov. Premennivosť každého faktoru, vyjadrenú stĺpcom zdrojovej matice, je možné rozdeliť na súčet dvoch zložiek: komunalitu $\mathbf{H}^2$, ktorá predstavuje variabilitu spoločnú pre všetky faktory a jedinečnosť $\mathbf{G}^2$, ktorá predstavuje časť variability, t.j. rozptylu zodpovedajúcej premennej. Štvorec komunality je suma faktorových záťaží faktorov. Jedinečnosť obsahuje časť, ktorá predstavuje experimentálnu chybu merania a špecifickú variabilitu, ktorú sa nedá vysvetliť ani chybou experimentu a ani spoločnými faktormi.

Metóda hlavných komponentov je vlastne špeciálnym prípadom faktorovej analýzy, kde $\mathbf{S}^2 = \mathbf{H}^2$, t.j. pomocou hlavných komponentov je možné variabilitu zdrojovej matice úplne, bez zvyšku reprodukovať. Vtedy hovoríme tiež o úplnej komponentnej analýze. Ak pomocou hlavných komponentov je možné reprodukovať len podstatnú časť variability, hovoríme o neúplnej komponentnej analýze.

Porovnanie FA a PCA:

1. Obe metódy nemá zmysel použiť, ak pôvodné premenné x nie sú korelované. FA potom nemá čo objasniť a PCA povedie k hlavným komponentom totožným s pôvodnými premennými.

2. FA sa pokúša objasniť kovarianciu a koreláciu pôvodných premenných pomocou niekoľkých spoločných faktorov. PCA objasňuje len rozptyl pôvodných premenných.

3. PCA: keď zvýšime počet použitých premenných A o 1 na $A+1$, pôvodné komponenty sa nezmenia. FA: keď pridáme ďalší faktor, ostatné faktory sa podstatne zmenia.

4. PCA: výpočet je priamočiary, jednoduchý. FA: výpočet je komplexnejší a existuje viacero postupov.

Tab. 14 Hodnoty distribučnej funkcie normovaného normálneho rozdelenia vybraných kladných hodnôt t

t	$F(t)$	t	$F(t)$	t	$F(t)$	t	$F(t)$	t	$F(t)$	t	$F(t)$
0,00	0,50000	0,30	0,61791	0,60	0,72575	0,90	0,81594	1,20	0,88493	1,50	0,93319
0,01	0,50399	0,31	0,62172	0,61	0,72907	0,91	0,81859	1,21	0,88686	1,51	0,93448
0,02	0,50798	0,32	0,62552	0,62	0,73237	0,92	0,82121	1,22	0,88877	1,52	0,93574
0,03	0,51197	0,33	0,62930	0,63	0,73565	0,93	0,82381	1,23	0,89065	1,53	0,93699
0,04	0,51595	0,34	0,63307	0,64	0,73891	0,94	0,82639	1,24	0,89251	1,54	0,93822
0,05	0,51994	0,35	0,63683	0,65	0,74215	0,95	0,82894	1,25	0,89435	1,55	0,93943
0,06	0,52392	0,36	0,64058	0,66	0,74537	0,96	0,83147	1,26	0,89617	1,56	0,94062
0,07	0,52790	0,37	0,64431	0,67	0,74857	0,97	0,83398	1,27	0,89796	1,57	0,94179
0,08	0,53188	0,38	0,64803	0,68	0,75175	0,98	0,83646	1,28	0,89973	1,58	0,94295
0,09	0,53586	0,39	0,65173	0,69	0,75490	0,99	0,83891	1,29	0,90147	1,59	0,94408
0,10	0,53983	0,40	0,65542	0,70	0,75804	1,00	0,84134	1,30	0,90320	1,60	0,94520
0,11	0,54380	0,41	0,65910	0,71	0,76115	1,01	0,84375	1,31	0,90490	1,61	0,94630
0,12	0,54776	0,42	0,66276	0,72	0,76424	1,02	0,84614	1,32	0,90658	1,62	0,94738
0,13	0,55172	0,43	0,66640	0,73	0,76730	1,03	0,84850	1,33	0,90824	1,63	0,94845
0,14	0,55567	0,44	0,67003	0,74	0,77035	1,04	0,85083	1,34	0,90988	1,64	0,94950
0,15	0,55962	0,45	0,67364	0,75	0,77337	1,05	0,85314	1,35	0,91149	1,65	0,95053
0,16	0,56356	0,46	0,67724	0,76	0,77637	1,06	0,85543	1,36	0,91309	1,66	0,95154
0,17	0,56749	0,47	0,68082	0,77	0,77935	1,07	0,85769	1,37	0,91466	1,67	0,95254
0,18	0,57142	0,48	0,68439	0,78	0,78230	1,08	0,85993	1,38	0,91621	1,68	0,95352
0,19	0,57535	0,49	0,68793	0,79	0,78524	1,09	0,86214	1,39	0,91774	1,69	0,95449
0,20	0,57926	0,50	0,69146	0,80	0,78814	1,10	0,86433	1,40	0,91924	1,70	0,95543
0,21	0,58317	0,51	0,69497	0,81	0,79103	1,11	0,86650	1,41	0,92073	1,71	0,95637
0,22	0,58706	0,52	0,69847	0,82	0,79389	1,12	0,86864	1,42	0,92220	1,72	0,95728
0,23	0,59095	0,53	0,70194	0,83	0,79673	1,13	0,87076	1,43	0,92364	1,73	0,95818
0,24	0,59483	0,54	0,70540	0,84	0,79955	1,14	0,87286	1,44	0,92507	1,74	0,95907
0,25	0,59871	0,55	0,70884	0,85	0,80234	1,15	0,87493	1,45	0,92647	1,75	0,95994
0,26	0,60257	0,56	0,71226	0,86	0,80511	1,16	0,87698	1,46	0,92786	1,76	0,96080
0,27	0,60642	0,57	0,71566	0,87	0,80785	1,17	0,87900	1,47	0,92922	1,77	0,96164
0,28	0,61026	0,58	0,71904	0,88	0,81057	1,18	0,88100	1,48	0,93056	1,78	0,96246
0,29	0,61409	0,59	0,72240	0,89	0,81327	1,19	0,88298	1,49	0,93189	1,79	0,96327

Spracované podľa: http://web.tuke.sk/sjf-kama/ulohy/Statistika-tabulky/norm_r.pdf

Tab. 15 Kritické hodnoty normovaného normálneho rozdelenia

P	z_p	P	z_p	P	z_p	P	z_p
0,50	0,000	0,75	0,674	0,950	1,645	0,975	1,96
0,51	0,025	0,76	0,706	0,951	1,655	0,976	1,977
0,52	0,050	0,77	0,739	0,952	1,665	0,977	1,995
0,53	0,075	0,78	0,772	0,953	1,675	0,978	2,014
0,54	0,100	0,79	0,806	0,954	1,685	0,979	2,034
0,55	0,126	0,80	0,842	0,955	1,695	0,980	2,054
0,56	0,151	0,81	0,878	0,956	1,706	0,981	2,075
0,57	0,176	0,82	0,915	0,957	1,717	0,982	2,097
0,58	0,202	0,83	0,954	0,958	1,728	0,983	2,12
0,59	0,228	0,84	0,994	0,959	1,739	0,984	2,144
0,60	0,253	0,85	1,036	0,960	1,751	0,985	2,17
0,61	0,279	0,86	1,080	0,961	1,762	0,986	2,197
0,62	0,305	0,87	1,126	0,962	1,774	0,987	2,226
0,63	0,332	0,88	1,175	0,963	1,787	0,988	2,257
0,64	0,358	0,89	1,227	0,964	1,799	0,989	2,29
0,65	0,385	0,90	1,282	0,965	1,812	0,990	2,326
0,66	0,412	0,905	1,311	0,966	1,825	0,991	2,366
0,67	0,440	0,91	1,341	0,967	1,838	0,992	2,409
0,68	0,468	0,915	1,372	0,968	1,852	0,993	2,457
0,69	0,496	0,92	1,405	0,969	1,866	0,994	2,512
0,70	0,524	0,925	1,440	0,970	1,881	0,995	2,576
0,71	0,553	0,93	1,476	0,971	1,896	0,996	2,652
0,72	0,583	0,935	1,514	0,972	1,911	0,997	2,748
0,73	0,613	0,94	1,555	0,973	1,927	0,998	2,878
0,74	0,643	0,945	1,598	0,974	1,943	0,999	3,09

Spracované podľa: Gregorová, Fillová 2004

Tab. 16 Kritické hodnoty χ^2 rozdelenia

	0,0005	0,001	0,005	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	-	-	-	-	-	0,0039	0,0158	2,71	3,84	5,02	6,63	7,88	10,8	12,1
2	0,001	0,002	0,0100	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21	10,6	13,8	15,2
3	0,0153	0,0243	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3	12,8	16,3	17,7
4	0,0639	0,0908	0,207	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3	14,9	18,5	20
5	0,158	0,21	0,412	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1	16,7	20,5	22,1
6	0,299	0,381	0,676	0,872	1,24	1,64	2,2	10,6	12,6	14,4	16,8	18,5	22,5	24,1
7	0,485	0,598	0,989	1,24	1,69	2,17	2,83	12	14,1	16	18,5	20,3	24,3	26
8	0,71	0,857	1,34	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1	22	26,1	27,9
9	0,972	1,15	1,73	2,09	2,7	3,33	4,17	14,7	16,9	19	21,7	23,6	27,9	29,7
10	1,26	1,48	2,16	2,56	3,25	3,94	4,87	16	18,3	20,5	23,2	25,2	29,6	31,4
11	1,59	1,83	2,6	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7	26,8	31,3	33,1
12	1,93	2,21	3,07	3,57	4,4	5,23	6,3	18,5	21	23,3	26,2	28,3	32,9	34,8
13	2,31	2,62	3,57	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7	29,8	34,5	36,5
14	2,7	3,04	4,07	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1	31,3	36,1	38,1
15	3,11	3,48	4,6	5,23	6,26	7,26	8,55	22,3	25	27,5	30,6	32,8	37,7	39,7
16	3,54	3,94	5,14	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32	34,3	39,3	41,3
17	3,98	4,42	5,7	6,41	7,56	8,67	10,1	24,8	27,6	30,2	33,4	35,7	40,8	42,9
18	4,44	4,9	6,26	7,01	8,23	9,39	10,9	26	28,9	31,5	34,8	37,2	42,3	44,4
19	4,91	5,41	6,84	7,63	8,91	10,1	11,7	27,2	30,1	32,9	36,2	38,6	43,8	46
20	4,4	5,92	7,43	8,26	9,59	10,9	12,4	28,4	31,4	34,2	37,6	40	45,3	47,5
21	5,9	6,45	8,03	8,9	10,3	11,6	13,2	29,6	32,7	35,5	38,9	41,4	46,8	49
22	6,4	6,98	8,64	9,54	11	12,3	14	30,8	33,9	36,8	40,3	42,8	48,3	50,5
23	6,92	7,53	9,26	10,2	11,7	13,1	14,8	32	35,2	38,1	41,6	44,2	49,7	52
24	7,45	8,08	9,89	10,9	12,4	13,8	15,7	33,2	36,4	39,4	43	45,6	51,2	53,5
25	7,99	8,65	10,5	11,5	13,1	14,6	16,5	34,4	37,7	40,6	44,3	46,9	52,6	54,9
26	8,54	9,22	11,2	12,2	13,8	15,4	17,3	35,6	38,9	41,9	45,6	48,3	54,1	56,4
27	9,09	9,8	11,8	12,9	14,6	16,2	18,1	36,7	40,1	43,2	47	49,6	55,5	57,9
28	9,66	10,4	12,5	13,6	15,3	16,9	18,9	37,9	41,3	44,5	48,3	51	56,9	59,3
29	10,2	11	13,1	14,3	16	17,7	19,8	39,1	42,6	45,7	49,6	52,3	58,3	60,7
30	10,8	11,6	13,8	15	16,8	18,5	20,6	40,3	43,8	47	50,9	53,7	59,7	62,2

Spracované podľa: Gregorová, Fillová 2004

Tab. 17 Kritické hodnoty Studentovho (t) - rozdelenia

n \ p	0,9	0,95	0,975	0,99	0,995
1	3,078	6,314	12,706	31,821	63,657
2	1,886	2,920	4,303	6,965	9,925
3	1,638	2,353	3,182	4,541	5,841
4	1,533	2,132	2,776	3,747	4,604
5	1,476	2,015	2,571	3,365	4,032
6	1,440	1,943	2,447	3,143	3,707
7	1,415	1,895	2,365	2,998	3,499
8	1,397	1,860	2,306	2,896	3,355
9	1,383	1,833	2,262	2,821	3,250
10	1,372	1,812	2,228	2,764	3,169
11	1,363	1,796	2,201	2,718	3,106
12	1,356	1,782	2,179	2,681	3,055
13	1,350	1,771	2,160	2,650	3,012
14	1,345	1,761	2,145	2,624	2,977
15	1,341	1,753	2,131	2,602	2,947
16	1,337	1,746	2,120	2,583	2,921
17	1,333	1,740	2,110	2,567	2,898
18	1,330	1,734	2,101	2,552	2,878
19	1,328	1,729	2,093	2,539	2,861
20	1,325	1,725	2,086	2,528	2,845
21	1,323	1,721	2,080	2,518	2,831
22	1,321	1,717	2,074	2,508	2,819
23	1,319	1,714	2,069	2,500	2,807
24	1,318	1,711	2,064	2,492	2,797
25	1,316	1,708	2,060	2,485	2,787
26	1,315	1,706	2,056	2,479	2,779
27	1,314	1,703	2,052	2,473	2,771
28	1,313	1,701	2,048	2,467	2,763
29	1,311	1,699	2,045	2,462	2,756
30	1,310	1,697	2,042	2,457	2,750

Spracované podľa: Gregorová, Fillová 2004

Tab. 18 Vybrané hodnoty Poissonovho rozdelenia

x					λ			
	0,1	0,2	0,3	0,4	0,5	1,0	1,5	2,0
0	0,904837	0,818731	0,740818	0,670320	0,606531	0,367879	0,223130	0,135335
1	0,090484	0,163746	0,222245	0,268128	0,303265	0,367879	0,334695	0,270671
2	0,004524	0,016375	0,033337	0,053626	0,075816	0,183940	0,251021	0,270671
3	0,000151	0,001092	0,003334	0,007150	0,012636	0,061313	0,125510	0,180447
4	0,000004	0,000055	0,000250	0,000715	0,001580	0,015328	0,047067	0,090224
5	-	0,000002	0,000015	0,000057	0,000158	0,003066	0,014120	0,036089
6	-	-	0,000001	0,000004	0,000013	0,000511	0,003530	0,012030
7	-	-	-	-	0,000001	0,000073	0,000756	0,003437
8	-	-	-	-	-	0,000009	0,000142	0,000859
9	-	-	-	-	-	0,000001	0,000024	0,000191
10	-	-	-	-	-	-	0,000004	0,000038
11	-	-	-	-	-	-	-	0,000007
12	-	-	-	-	-	-	-	0,000001

Spracované podľa: Gregorová, Fillová 2004

Tab. 19 Kritické hodnoty W_α ($\alpha=0,05$) jednovýberového Wilcoxonovho testu pre n-počet nenulových rozdielov

n	W_α	n	W_α	n	W_α	n	W_α	n	W_α	n	W_α
6	0	16	29	26	98	36	208	46	361	56	557
7	2	17	34	27	107	37	221	47	378	57	579
8	3	18	40	28	116	38	235	48	396	58	602
9	5	19	46	29	126	39	249	49	415	59	625
10	8	20	52	30	137	40	264	50	434	60	648
11	10	21	58	31	147	41	279	51	453	61	672
12	13	22	65	32	159	42	294	52	473	62	697
13	17	23	73	33	170	43	310	53	494	63	721
14	21	24	81	34	182	44	327	54	514	64	747
15	25	25	89	35	195	45	343	55	536	65	772

Spracované podľa: http://home.zcu.cz/~sediva/pse/pse_pr11.pdf

Tab. 20 Kritické hodnoty D_α Kolmogorovho-Smirnovho testu

$n \backslash \alpha$	0,05	0,01	$n \backslash \alpha$	0,05	0,01
1	0,975	0,995	21	0,287	0,344
2	0,842	0,929	22	0,281	0,337
3	0,708	0,829	23	0,275	0,330
4	0,624	0,734	24	0,269	0,323
5	0,563	0,669	25	0,264	0,317
6	0,519	0,617	26	0,259	0,311
7	0,483	0,576	27	0,254	0,305
8	0,454	0,542	28	0,250	0,300
9	0,430	0,513	29	0,246	0,295
10	0,409	0,489	30	0,242	0,290
11	0,391	0,468	31	0,238	0,285
12	0,375	0,449	32	0,234	0,281
13	0,361	0,432	33	0,231	0,277
14	0,349	0,418	34	0,227	0,273
15	0,338	0,404	35	0,224	0,269
16	0,327	0,392	36	0,221	0,265
17	0,318	0,381	37	0,218	0,262
18	0,309	0,371	38	0,215	0,258
19	0,301	0,361	39	0,213	0,255
20	0,294	0,352	40	0,210	0,252
			nad 40	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Spracované podľa: Spišiak 1997

Zoznam použitej literatúry

- ANDĚL, J. (1978). Matematická statistika. Bratislava, SNTL/Alfa, 346 s.
- BAČÍK, M. (2007). Základy štatistiky pre geografov. Banská Bystrica, FPV UMB, 122 s.
- BAKYTOVÁ, H. a kol. (1975). Základy štatistiky. Bratislava, Alfa. 392 s.
- BRÁZDIL, R., et. al. (1995). Statistické metody v geografii. Brno, Marykova univerzita, 177 s.
- CYHELSKÝ, L. (1967a). Statistika v příkladech. Praha, SNTL, 227 s.
- CYHELSKÝ, L. (1974b). Úvod do teórie popisné štatistiky. Praha a Bratislava, STNL a Alfa, 288 s.
- ČERMÁK, V. (1968). Statistika II. Praha, STNL, 205 s.
- GREGOROVÁ, G., FILLOVÁ, V. (2004). Štatistické metody v geografii. Bratislava, PRIF UK, 117 s.
- HOFIERKA, J. (2003). Geografické informačné systémy a diaľkový prieskum Zeme. Vysokoškolské učebné texty. Prešov, FHPV PU. 116 s.
- KOREC, P., et. al. (1997). Kraje a okresy Slovenska. Nové administratívne členenie. Bratislava: Q111, 392 s.
- LAMOŠ, F., POTOCKÝ, R. (1989). Pravdepodobnosť a matematická štatistika. Bratislava, Alfa, 344 s.
- MELOUN, M, MILITKÝ, J. (2002). Kompendium statistického zpracování dat. Praha, Academia, 766 s.
- NIŽŇANSKÝ, B. (2000). Základy geoinformatiky. Vysokoškolské učebné texty. Prešov, FHPV PU. 232 s.
- NOVÁKOVÁ, G. (2008). Základy štatistiky pre geografov. Bratislava, PRIF UK, 220 s.
- PAVLÍK, Z., KÜHNL, K. (1981). Úvod do kvantitatívnych metod pro geography. Učebné texty. Praha, SPN, 267 s.
- ROGERSON, P. A. (2001). Statistical Methods for Geography. SAGE Publications, 248 s.
- SVOBODA, H. (1977). Moderní statistika. Praha, Svoboda, 352 s.
- SPIŠIAK, P. (1997). Cvičenia z vybraných kvantitatívnych a prognostických metod v geografii. Vysokoškolské učebné texty. Bratislava, UK, 146 s.
- Správa o poľnohospodárstve a potravinárstve v Slovenskej republike 2007 (stav za rok 2006), (Zelená správa, 2007). Bratislava, Ministerstvo pôdohospodárstva SR.
- Štatistická ročenka regiónov Slovenska (2007). Bratislava, ŠÚ SR, 382 s.
- Štatistická ročenka Slovenskej republiky (1998). Bratislava, ŠÚ SR, 750 s.

Internetové zdroje:

http://www.rajcany.sk/geograficke_udaje_o_obci/geograficke_udaje_o_obci_soubory/image010.gif

<http://eldum.phil.muni.cz/mod/resource/view.php?id=1764>

<http://fstroj.uniza.sk/kam/pobocikova/pdf/TEXT12.PDF>

<http://kosice.upjs.sk/~zezula/stg/k14.pdf>

<http://rimarcik.com/navigator/metody.html>

http://sis.science.upjs.sk/matematika/kniznica_vp/funkcie/Linear.htm

http://sis.science.upjs.sk/matematika/kniznica_vp/funkcie/Kvadrat.htm

http://web.tuke.sk/sjf-kama/ulohy/Statistika-tabulky/norm_r.pdf

<http://www.avozarm.sk/Lorenz.htm>

<http://www.fem.uniag.sk/ACADEMIC/depart/ksov/predmety/statistika/subory/asociacie.pdf>

<http://www.fem.uniag.sk/ACADEMIC/depart/ksov/predmety/statistika/subory/prednaska22.ppt>

<http://www.fem.uniag.sk/cvicenia/ksov/obtulovic/BIO%C5%A0TATISTIKA/prednaska3.ppt>

<http://www.fhvp.unipo.sk/cvt/statistika/zlacka/geoinfo6.pdf>

http://www.quark.sk/images/clanky/2005124/05125_obzalovany2.jpg

http://www.shmu.sk/File/kvant_pr_03/Bodrog_Q03.pdf

http://www.shmu.sk/File/kvant_pr_03/Morava_Q03.pdf

http://www.uniag.sk/SKOLA/rvv/doc/ddiz/2007/martina_chudikova.pdf

http://home.zcu.cz/~sediva/pse/pse_pr11.pdf

Zoznam grafov, tabuliek a obrázkov

Grafy:

Graf 1 Rozdelenie obyvateľov obce Vinné z hľadiska produktivity (rok 2001)	11
Graf 2 Histogram a polygón	18
Graf 3 Veková pyramída	19
Graf 4 Miera šikmosti	65
Graf 5 Miera špicatosti	67
Graf 6 Lorenzova krivka	68
Graf 7 Schematické vyjadrenie miery koncentrácie	69
Graf 8 Hustota pravdepodobnosti normálneho rozdelenia	72
Graf 9 Hustota pravdepodobnosti a distribučná funkcia normálneho rozdelenia	73
Graf 10 Krivka hustoty pravdepodobnosti normovaného normálneho rozdelenia	74
Graf 11 Hustota pravdepodobnosti χ^2 rozdelenia pre 2, 6 a 10 stupňov voľnosti	76
Graf 12 Schematické znázornenie zhladzovania časového radu pomocou kľazavých priemerov	94
Graf 13 Obojstranný interval spoľahlivosti	103
Graf 14 Jednostranný zdola ohraničený interval spoľahlivosti	103
Graf 15 Jednostranný zhora ohraničený interval spoľahlivosti	103
Graf 16 Porovnanie teplotných anomálií, slnečného cyklu a koncentrácie CO_2	114
Graf 17 Lineárna pozitívna korelačná závislosť	117
Graf 18 Nelineárna pozitívna korelačná závislosť	117
Graf 19 Lineárna negatívna korelačná závislosť	118
Graf 20 Nelineárna negatívna korelačná závislosť	118
Graf 21 Lineárna regresia	124
Graf 22 Indexový graf úpätia vlastných čísel	133

Tabuľky:

Tab. 1 Rozdelenie obyvateľov obce Vinné podľa pohlavia a národnosti (rok 2001)	11
Tab. 2 Vývoj počtu obyvateľov obce Vinné od roku 1828 do roku 1948	12
Tab. 3 Miera zamestnanosti vo vybraných štátoch Európy (rok 1999)	12
Tab. 4 Rozvody podľa veku manželov v roku 2004	12
Tab. 5 Výsledky písomnej preverky zo slovenského jazyka	13
Tab. 6 Rozdelenie študentov 1. ročníka externého štúdia odboru GRR podľa ich veku	14
Tab. 7 Frekvenčná tabuľka meraní ročných úhrnov zrážok	17
Tab. 8 Výpočet základných mier rastu	93
Tab. 9 Kľazavé priemery z 3 a z 5 hodnôt	94
Tab. 10 Najčastejšie používané intervaly spoľahlivosti	104
Tab. 11 Korelačná tabuľka	116
Tab. 12 Asociačná tabuľka	119
Tab. 13 Kontingenčná tabuľka	120
Tab. 14 Hodnoty distribučnej funkcie normovaného normálneho rozdelenia vybraných kladných hodnôt t	136
Tab. 15 Kritické hodnoty normovaného normálneho rozdelenia	137
Tab. 16 Kritické hodnoty χ^2 rozdelenia	138
Tab. 17 Kritické hodnoty Studentovho (t) - rozdelenia	139
Tab. 18 Vybrané hodnoty Poissonovho rozdelenia	140
Tab. 19 Kritické hodnoty W_α ($\alpha = 0,05$) jednovýberového Wilcoxonovho testu pre n -počet nenulových rozdielov	140
Tab. 20 Kritické hodnoty D_α Kolmogorovho-Smirnovho testu	141

Obrázky:

Obr. 1 Databázová tabuľka štatistického softvéru SPSS	15
Obr. 2 Stupeň korelačnej závislosti	115

ZÁKLADY ŠTATISTICKÝCH METÓD V GEOGRAFII

vysokoškolský učebný text pre vybrané študijné programy na FHPV PU

Autori: RNDr. Monika Ivanová, PhD., Doc. Mgr. Jaroslav Hofierka, PhD.
Recenzenti: Doc. PhDr. RNDr. Martin Boltížiar, PhD.
Mgr. Mária Majherová, PhD.
Korektúra: autori
Vydala: Prešovská univerzita v Prešove, Fakulta humanitných a prírodných vied
Náklad: 300 výtlačkov
Rozsah diela: 144 strán
AH: 9,85
Formát: A4
Vydanie: prvé, 2009
Sadzba: autori
Tlač: Grafotlač Prešov

ISBN 978 – 80 – 555 - 0091-1