ŽILINSKÁ UNIVERZITA V ŽILINE
Fakulta bezpečnostného inžinierstva

Ing. Michal Titko, PhD.
doc. Ing. Ladislav Novák, PhD.
Ing. Michaela Jánošíková, PhD.

PRAKTICKÁ ŠTATISTIKA

ISBN 978-80-554-1814-8

Žilina 2021

Ing. MICHAL TITKO, PhD.
doc. Ing. LADISLAV NOVÁK, PhD.
Ing. MICHAELA JÁNOŠÍKOVÁ, PhD.

PRAKTICKÁ ŠTATISTIKA

Žilinská univerzita v Žiline
EDIS-vydavateľstvo UNIZA
2021

Vzor citácie: Titko, M., Novák, L., Jánošíková, M. 2021. Praktická štatistika. 1. vyd. Žilina: EDIS – vydavateľstvo UNIZA, 2021. 146 s.

Recenzenti: prof. Ing. Jozef Ristvej, PhD., EMBA
Ing. Ján Dvorský, PhD.

Schválila edičná rada ŽU výmerom č. 17/S/2021

ISBN 978-80-554-1814-8

Obsah

1 Úvod do štúdia štatistiky

Andrew Lang o politikovi:

"Používa štatistiku ako opitý človek pouličnú lampu – skôr na podporu než na osvetlenie."

Benjamin Disraeli o lži:

"Sú tri stupne lži - lož, nehanebná lož a štatistika."

1.1 Vývoj a význam pojmu štatistika

Slovu ***„štatistika"*** môže byť priradený najrôznejší obsah a preto existuje taktiež veľa definícií. Slovo štatistika vzniklo z latinského slova „**status**" = „**stav**", resp. „**štát**". Najstaršou štatistikou je „popis štátu", spočívajúci v zobrazení daného zemepisného, hospodárskeho a politického stavu v danom štáte.

Historicky sa štatistika postupne vyvíjala:

- prvé sčítanie obyvateľstva v Egypte – 3 tisíc pred n.l. – výsledky sčítania využívané panovníkmi na vojenské a finančné účely,
- 16. storočie – počiatky politickej aritmetiky – vychádzala z údajov o narodení a úmrtí – na tomto základe bol porovnávaný a pozorovaný číselný vývoj obyvateľstva počas dlhších časových úsekov ako oporný bod pre skúmanie spoločenských javov,
- 18. storočie – obdobie univerzitnej štatistiky – štatistika označená za vedu o štátnych pozoruhodnostiach, t.j. slovný popis územia štátu, jeho obyvateľstva, armády, poľnohospodárstva, obchodu atď. sprevádzaný číselnými údajmi,
- 20. storočie – rozvoj matematickej štatistiky – vychádza z výberových zisťovaní a s využitím teórie pravdepodobnosti formuluje závery o celom súbore údajov.

V súčasnej dobe vnímame štatistiku z dvoch pohľadov. Jednak ide o **praktickú činnosť**, t.j. evidenčná a štatistická prax, a súčasne ide o **vedu**, ktorá sa zaoberá skúmaním hromadných javov.

Štatistika ako praktická činnosť je taktiež označovaná ako **štatistická administratíva** a chápe sa ako:

- **štatistická evidencia** (napr. zber údajov, triedenie, sumarizácia a pod.),
- **inštitúcie,** ktoré štatistickú evidenciu vykonávajú (napr. štatistický úrad, ministerstvá, a pod.),
- **súhrn údajov o nejakej skutočnosti** (štatistika nezamestnanosti, atď.).

Štatistická administratíva sa realizuje prostredníctvom štatistického vykazovania. Ťažisko je v sústave štátnych orgánov, ktoré evidujú potrebné ukazovatele, na čele so Štatistickým úradom Slovenskej republiky. Poskytované informácie slúžia na posudzovanie úrovne sociálneho a ekonomického vývoja príslušných územných celkov alebo celej krajiny.

Štatistika ako vedecká disciplína (teória štatistiky) je disciplínou, ktorá sa zaoberá skúmaním hromadných spoločenských javov a procesov, môže však ísť aj o javy biologické či technické. Aj keď štatistika skúma kvantitatívnu stránku týchto javov, nesmie byť zanedbávaná stránka kvalitatívna. Základným vyjadrovacím prostriedkom je číslo.

Štatistika, chápaná **ako vedecká disciplína,** má dve špecifické časti:

- **ekonomická štatistika –** skúma kvantitatívnu stránku národného hospodárstva a podľa pôsobnosti sa člení na čiastkové zložky (napr. štatistika priemysel, poľnohospodárstvo, obchod, demografia, atď.)
- **všeobecná teória štatistiky –** definuje štatistické pojmy a prezentuje všeobecne platné metódy na kvantitatívne skúmanie hromadných javov. Ide o metódy získavania štatistických údajov, metódy ich spracovania, metódy štatisticko-ekonomickej (biologickej, technickej) analýzy a formy zdieľania výsledkov.

Teória štatistiky v podstate poskytuje metodiku na ten ktorý účel skúmania. Podľa účelu skúmania sa rozdeľuje štatistika na:

- **deskriptívnu (popisnú/opisnú) štatistiku –** používa sa na popis skúmaného štatistického súboru vzhľadom na sledované vlastnosti (napr. demografické informácie o obyvateľoch krajiny; informácie o priemernej mzde v jednotlivých okresoch Slovenska, a pod.). Výsledky sa zverejňujú pomocou absolútnych a relatívnych čísel, jednoznačných tabuliek a grafov,
- **analytickú (relačnú, matematickú, induktívnu) štatistiku** – tá sa pomocou štatistických metód snaží hlbšie analyzovať získané údaje, hľadať súvislosti medzi vlastnosťami štatistického súboru (dáva do vzťahu skúmané premenné), pričom jej cieľom je vyvodzovať závery a podľa možností aj zovšeobecniť výsledky skúmania (napr. *skúmanie vplyvu konkrétnych hnojív na rast (výšku) rastliny (alebo objem plodu); prognóza miery nezamestnanosti na ďalšie obdobia*; a pod.).

Uplatnenie štatistiky ako vednej disciplíny je zrejmé aj v iných vedných odboroch:

- **v aplikovaných vedách –** „-metria“ a „-grafia“ (biometria, dendrometria, ekonometria, chemometria, a pod.),
- **vo vedách so silným štatistickým základom**: sociológia, psychológia, demografia.

1.2 Čo „vie“ a „nevie“ štatistika

Pre štatistiku je typické, že:

- **skúma hromadné javy**,
- **zaoberá sa premenlivými (variabilnými) vlastnosťami** skúmaných subjektov/objektov/javov,
- **pracuje s číslami a vyjadruje sa pomocou čísel** – zaujíma sa predovšetkým o kvantitatívnu stránku skúmaných problémov,
- **používa výpočtovú techniku** na vytváranie a správu štatistických databáz, na hromadné spracovanie údajov a ich analýzu.

Štatistika rieši **úlohy rôzneho stupňa zložitosti**, počínajúc **získavaním údajov** (počet domácností, počet pracovníkov v odvetví XY, objem vývozu, atď.), cez **popis štruktúry** (vekové rozloženie obyvateľov, rozloženie firiem z hľadiska právnej formy podnikania, vyčísľovania čiastkových ukazovateľov v čase a priestore ako je napr. výpočet priemernej mzdy v národnom hospodárstve, výpočet cenovej hladiny spotrebiteľských cien, a pod.), **porovnávanie** takto agregovaných **ukazovateľov v čase alebo priestore** (trend vývoja miezd, zmena hladiny spotrebiteľských cien), **predpovedanie** ich **budúcej úrovne** (tržby v maloobchode v nasledujúcom štvrťroku, vývoz produktu A v ďalšom roku), **meranie závislosti** (závislosť miezd od HDP, závislosť vývozu na kurze meny).

Štatistika zlyháva, ak nemá k dispozícii adekvátne číselné údaje, keď chýba predstava o veľkosti chýb merania a vplyve rôznych sprievodných činiteľov, keď nie je k dispozícii dostatočne rozsiahly súbor prípadov alebo v údajoch chýba premenlivosť (variabilita).

Literatúra

BENČO, J. *Metodológia vedeckého výskumu*. Bratislava: IRIS, 2001. ISBN 80-9018-27-0.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DAŇKO, J. *Úvod do štatistiky: praktikum.* Prešov: Vydavateľstvo Michala Vaška, 2007. ISBN 978-80-7165-597-8.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

GAVORA, P. *Metodologický profil kvantitatívnych výskumných štúdií publikovaných v časopise Pedagogika: porovnanie období 1995–2000 a 2010–2014*. 2015. Pedagogika, roč. 65, č. 4, 2015, s. 372–391.

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

GROFÍK, R. a kol. *Štatistika*. Bratislava: Príroda, 1987.

CHAJDIAK, J. a kol. *Štatistika jednoducho*. Bratislava: STATIS, 2003.

MIKOLAJ, J., VANČO, B. *Štatistika pre manažérov*. Žilina: RVS vydavateľstvo FŠI ŽU v Žiline, 2000.

2 Štatistická terminológia a vyjadrovacie prostriedky

2.1 Štatistická terminológia

Štatistika v mnohých ohľadoch vychádza z teórie pravdepodobnosti, ktorá sa zaoberá štúdiom zákonitostí náhodných javov. Z tohto pohľadu je nevyhnutné definovať náhodný jav, ako aj ďalšie základné termíny, ktoré s problematikou praktickej štatistiky súvisia.

Náhodný jav

Náhodný jav je výsledkom náhodného pokusu. Pričom **náhodný pokus** sa chápe ako dej, činnosť, ktorá sa niekoľkokrát opakuje za rovnakých alebo približne rovnakých podmienok, a ktorej výsledok je neistý, závislý od náhody. **Náhodný jav** je teda jav, udalosť, výsledok, každá skutočnosť, ktorá môže nastať pri uskutočnení pokusu; ak sa opakuje u veľkého počtu prvkov môžeme ho nazývať aj **hromadný náhodný jav**.

Štatistický problém

Z dôvodu komplexnosti náhodných javov v spoločnosti je pre účely štatistického skúmania vhodnejšie hovoriť o **štatistickom probléme.** Štatistickým problémom bývajú problémy spoločenskej praxe, problémy vychádzajúce z teórie (rôzneho zamerania), záujmové témy, pre ktoré dokážeme identifikovať dôležité premenné a predpokladať významné súvislosti medzi nimi (*Vplyv dištančnej výučby na kvalitu výstupov študentov*). Prvky, ktorých sa daný štatistický problém dotýka nazývame **štatistickými jednotkami.**

Štatistická jednotka

Štatistická jednotka je nositeľom štatistickej informácie, je to elementárny prvok hromadného javu. Môžeme ju definovať aj ako **základný objekt/subjekt skúmania**, na ktorom skúmame konkrétne prejavy/vlastnosti určitého hromadného javu. Môže to byť osoba, objekt, región, udalosť a pod. Štatistické jednotky predstavujú:

- **reálne existujúce objekty/subjekty hmotnej povahy:**
  - **ľudia** ako jednotlivci v rôznych pozíciách (zákazníci, voliči, zamestnanci, všeobecne respondenti, a pod.),
  - **živé organizmy a ich skupiny** (zvieratá, rastliny, stáda, porasty, a pod.),
  - **neživé prírodné predmety**,
  - **hmotné výsledky ľudskej činnosti** (výrobky, umelecké diela, a pod.),
- právne, politicky či inak vymedzené **časti spoločenského priestoru** (ekonomické subjekty, hospodárske odvetvia, kraje, štáty, a pod.),
- **nehmotné výsledky ľudskej činnosti** (športové či umelecké výkony, zásahy hasičov, a pod.)
- **javy, udalosti** (požiare, tornáda, narodenia, úmrtia, úrazy, a pod.)
- neopakovateľné **vzorky zo spojitého prostredia** (vzorky atmosféry, vody, pôdy, a pod.)

Štatistický súbor

Množina štatistických jednotiek, ktoré **spoločne tvoria** podklad na skúmanie formulovaného štatistického problému. Je to **množina štatistických jednotiek**, ktoré majú požadované spoločné vlastnosti (domácnosti SR, firmy z jedného odvetvia, výrobky jednej šarže, obce jedného okresu, muži s vysokoškolským vzdelaním, a pod.).

Každý štatistický súbor je určený dvoma základnými atribútmi:

- **obsahom – kvalitou** (kap. 4.2.3)**,**
- **rozsahom – kvantitou** (kap. 4.2.2)**.**

Obsah súboru je vymedzený:

- zoznamom štatistických jednotiek (kto?, čo?) **– explicitne**,
- štatistickými znakmi (vlastnosťami), ktoré chce riešiteľ skúmať na daných štatistických jednotkách **– implicitne**.

Rozsah súboru je určený počtom štatistických jednotiek v štatistickom súbore (počet, množstvo, koľko štatistických jednotiek bolo skúmaných):

- **v popisnej štatistike** sa označuje **rozsah súboru *n*** bez ďalšej špecifikácie,
- **v analytickej (induktívnej) štatistike** sa rozlišuje štatistický súbor **základný (*N*) a výberový (*n*)** súbor.

Základný štatistický súbor (populácia) je štatistický súbor tvorený všetkými štatistickými jednotkami, ktoré doň patria na základe sledovaných vlastností (štatistických znakov) a ich hodnôt/obmien. Príklad: všetci obyvatelia Slovenska s právom voliť, všetci obyvatelia Žilinského kraja, všetci študenti univerzity/fakulty/študijného programu, všetky domácnosti so zatepleným domom na Slovensku, a pod.

Výberový štatistický súbor (vzorka) je tvorený vybranými štatistickými jednotkami (podľa určených kritérií alebo náhodne), ktoré predstavujú podmnožinu základného súboru. Je to skupina štatistických jednotiek, ktoré reálne skúmame. Vybraná vzorka obyvateľov Slovenska s právom voliť, vzorka obyvateľov Žilinského kraja, vybraní študenti univerzity, vybrané domácnosti so zateplenými domami na Slovenku a pod.

Štatistický znak

Štatistický znak predstavuje skúmanú vlastnosť štatistických jednotiek. Riešiteľ ich skúma väčšinou niekoľko, podľa komplexnosti formulovaného štatistického problému.

Klasifikácia štatistických znakov

Základným rozdelením štatistických znakov je rozdelenie na:

- **znaky identifikačné** – z vecného, časového a priestorového hľadiska identifikujú štatistickú jednotku, **rozhodujú o zaradení či nezaradení štatistickej jednotky do štatistického súboru**, nie sú predmetom analýzy; štatistické jednotky sa zhodujú v ich konkrétnej obmene. Napríklad v rámci štatistického problému „*Trávenia voľného času študentmi Žilinskej univerzity na bakalárskom stupni*“, bude štatistickou jednotkou študent a identifikačnými znakmi budú (1) univerzita, na ktorej študent študuje a (2)

stupeň jeho štúdia; na zaradenie do prieskumu musí študent nadobúdať príslušné obmeny daných štatistických znakov (1) Žilinská univerzita a (2) bakalársky stupeň štúdia. Ak štatistická jednotka nespĺňa požadované kritériá nebude zaradená do prieskumu.

- **znaky variabilné** – sú predmetom analýzy a rozhodujú o spôsoboch (metódach) analyzovania získaných údajov

Variabilné štatistické znaky sa ďalej klasifikujú:

- znaky **slovné** (kvalitatívne) – **nominálne** – sa označujú veľkými písmenami zo začiatku abecedy (***A, B, C***),
    - znaky **alternatívne** (dvojné, binárne, dichotomické) – znak nadobúda iba 2 obmeny, alternatívy (pohlavie, očkovaný/neočkovaný, znaky s obmenami odpovedí áno/nie),
    - znaky **množné** – znaky nadobúdajú viac ako 2 obmeny (farba auta, krajina pôvodu, fakulta),
- znaky **číselné** (kvantitatívne) – sa označujú veľkými písmenami z konca abecedy (***X, Y, Z***):
    - **znaky poradové (ordinálne)** – vyjadrujú poradie štatistickej jednotky oproti ostatným, je možné porovnanie medzi jednotkami (školská klasifikácia, akostná trieda, umiestnenie v súťaži),
    - **znaky merateľné — kardinálne** (hmotnosť, počet obyvateľov, mzda).

Merateľné štatistické znaky sa ďalej klasifikujú na:

- **spojité znaky** (reálne čísla) napr. *časové údaje, rozmery, príjmy, výdaje a pod.,*
- **diskrétne znaky** (nespojité, izolované hodnoty, často celočíselné, nezáporné) napr. *počet detí v domácnosti, počet pracovníkov firmy, počet vyrobených výrobkov.*

Štatistické údaje – dáta

Štatistický údaj je **konkrétna hodnota** alebo **obmena**, ktorú môže dosahovať sledovaný štatistický znak.

Číselne vyjadrený údaj (označuje sa malými písmenami ***x, y, z***): cena výrobku je napr. 5€, 10€, 15€ - hovoríme o **hodnotách štatistického znaku**. Hodnoty číselného znaku ***X***, ktoré tvoria štatistický súbor o rozsahu ***n***, sa označujú ako:

$$x_1, x_2,..., x_i,..., x_n,$$ stručne iba x_i,pričom $i=1, 2,...,n$

Slovne vyjadrený údaj (označuje sa malými písmenami ***a, b, c***): dosiahnuté vzdelanie je napr. základné, stredoškolské, vysokoškolské – hovorí sa o **obmenách štatistického znaku.** Obmeny slovného znaku ***A***, ktoré tvoria štatistický súbor o rozsahu ***n***, sa označuje ako:

$$a_1, a_2,..., a_i,..., a_n,$$ stručne iba a_i,pričom $i=1, 2,...,n$

Index ***i*** pritom súvisí s poradím zisťovania.

Príklad:

Hromadný jav resp. **štatistický problém**: výskyt katastrof na Slovensku od roku 2010

Základný štatistický súbor: súbor všetkých katastrof, ktoré sa udiali na Slovensku
Výberový štatistický súbor: napr. vzorka iba prírodných katastrof, ktoré sa udiali na Slovensku (súčasne to nemusia byť ani všetky prírodné katastrofy), od roku 2010

Rozsah súboru: $n = 700$ (počet katastrof výberového súboru)

Štatistické jednotky: jednotlivé katastrofy

Identifikačné štatistické znaky: napr. typ katastrofy (iba prírodné), rok výskytu (od roku 2010), krajina výskytu (iba na Slovensku)

Variabilné štatistické znaky:

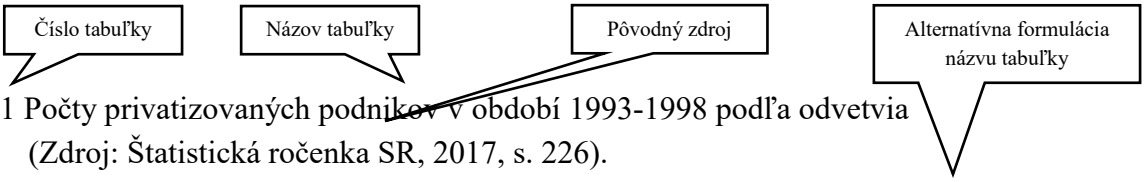
- číselné:
  - spojité: výška škôd, rozsah zasiahnutého územia, trvanie, a pod.
  - nespojité: počet mŕtvych, počet zranených, počet zasahujúcich zložiek, a pod.
- slovné:
  - množné: druh udalosti, miesto výskytu, mesiac výskytu, a pod.
  - alternatívne: vyhlásenie mimoriadnej situácie (áno/nie), potreba evakuácie (áno/nie), a pod.

2.2 Štatistické vyjadrovacie prostriedky

Nevyhnutnou súčasťou štatistickej analýzy je vhodná prezentácia výsledkov, jednak v tabuľkovej podobe ako aj v podobe grafov. Vizuálne spracovanie výsledkov je veľmi dôležitou súčasťou manažérskych zručností a umožňuje prezentujúcemu lepšie prezentovať získané výsledky aj pre publikum, ktoré nemá hlbšie poznatky zo štatistiky. Momentálne sú na trhu k dispozícii rôzne programové prostriedky, ktoré sa sústredia práve na pútavé a vecné zobrazenie získaných výsledkov. Dôležité je, aby vyjadrovacie prostriedky obsahovali minimálne základné prvky, ktoré budú **jednoznačne** vyjadrovať získané výsledky.

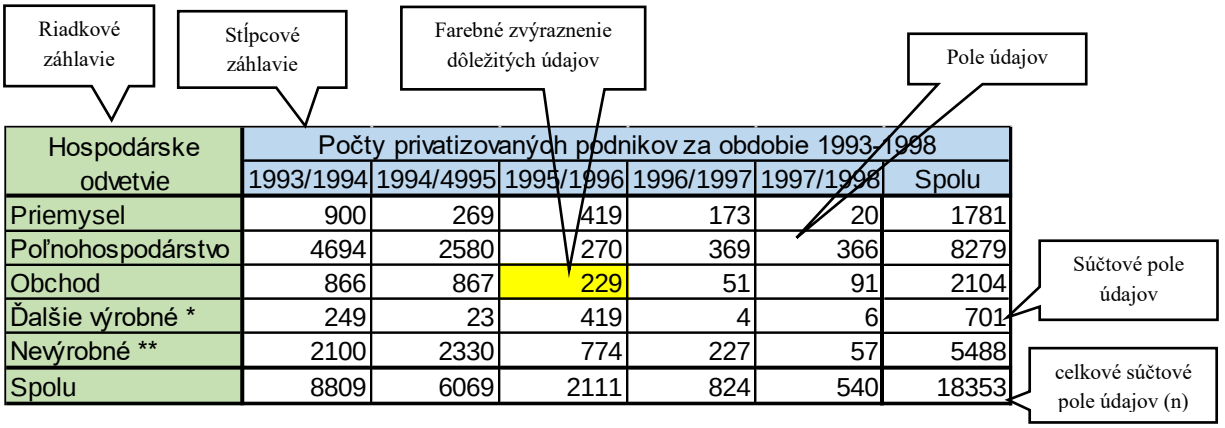
2.2.1 Štatistické tabuľky

Výsledky rôznych štatistických metód je vhodné zobrazovať pomocou konkrétnych tabuliek. Konkrétne príklady budú uvedené v častiach, ktoré sa dotýkajú konkrétnych metód. Štatistické tabuľky by mali obsahovať základné prvky, aby boli tieto tabuľky jednoznačné a s patričným obsahom (Obrázok 1).



Tab. 1 Počty privatizovaných podnikov v období 1993-1998 podľa odvetvia (Zdroj: Štatistická ročenka SR, 2017, s. 226).

Tabuľka 1 Rozdelenie privatizovaných podnikov podľa odvetvia v období 1993-1998



Hospodárske odvetvie	Počty privatizovaných podnikov za obdobie 1993-1998					
	1993/1994	1994/4995	1995/1996	1996/1997	1997/1998	Spolu
Priemysel	900	269	419	173	20	1781
Poľnohospodárstvo	4694	2580	270	369	366	8279
Obchod	866	867	229	51	91	2104
Ďalšie výrobné *	249	23	419	4	6	701
Nevýrobné **	2100	2330	774	227	57	5488
Spolu	8809	6069	2111	824	540	18353

*Poznámky: * zoznam ďalších výrobných podnikov je možné nájsť v prílohe A*
*** zoznam nevýrobných podnikov je možné nájsť v prílohe B*

Obrázok 1 Základné prvky štatistických tabuliek

Dohodnuté značky:

✕	(ležatý krížik)	vyplnenie políčka by bolo nelogické
–	(ležatá čiarka)	žiadny údaj prípad
.	(bodka)	neznámy, nespoľahlivý údaj
0	(nula)	menej než polovica zvolenej mernej jednotky

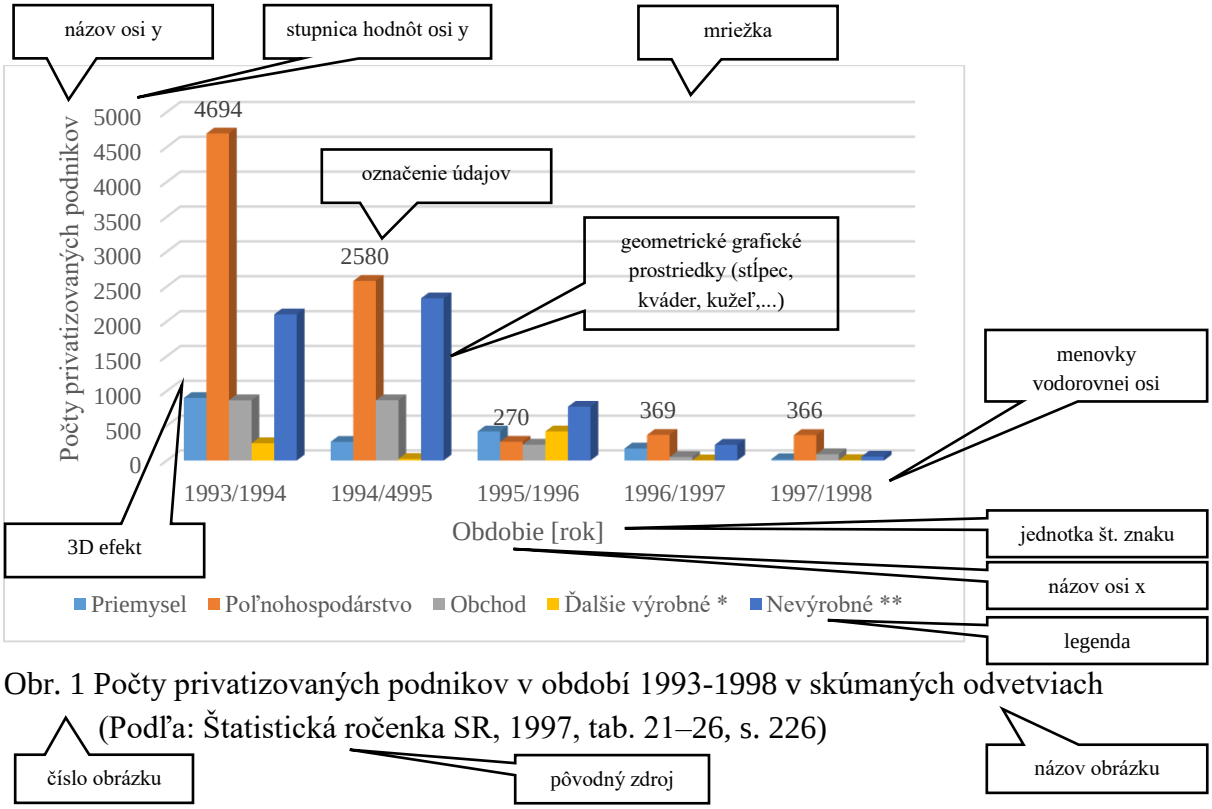
Klasifikácia štatistických tabuliek:

Podľa účelu delíme štatistické tabuľky na:

- tabuľky **prezentačné** (na prezentáciu údajov alebo výsledkov analýzy),
- tabuľky **pracovné** (napr. v zošite v MS Excel),
- tabuľky **dôležitých hodnôt** (už vypočítané konštanty, ktoré sa využívajú na účely rôznych štatistických metód a testovanie hypotéz, napr. logaritmické tabuľky).

2.2.2 Štatistické grafy

Výsledky rôznych štatistických metód je vhodné prezentovať pomocou konkrétnych grafických nástrojov. Konkrétne vhodné typy grafov budú uvedené v častiach, ktoré sa dotýkajú konkrétnych metód. Podobne ako tabuľky aj grafy by mali obsahovať základné prvky, aby bol graf jednoznačný a patrične zobrazoval žiadúce náležitosti (Obrázok 2).



Obr. 1 Počty privatizovaných podnikov v období 1993-1998 v skúmaných odvetviach (Podľa: Štatistická ročenka SR, 1997, tab. 21–26, s. 226)

číslo obrázku
pôvodný zdroj
názov obrázku

Obrázok 2 Základné prvky štatistického grafu

Klasifikácia štatistických grafov

Štatistické grafy sa rozdeľujú:

- podľa účelu na grafy:
  - **prezentačné,**
  - **konštrukčné,**
  - odčítacie — **nomogramy**,
- podľa použitej súradnicovej sústavy na grafy:
  - v **pravouhlej** súradnicovej sústave,
  - v **polárnej** súradnicovej sústave,
  - **ostatné** (nevyžadujú súradnicovú sústavu),
- podľa počtu dimenzií na grafy:
  - **rovinné** (2D),
  - **priestorové** (3D),
- podľa použitých grafických prostriedkov na grafy:
  - bodové, čiarové, stĺpcové, pruhové, kruhové, bublinové, kartogramy, kartodiagramy, piktogramy, atď.
- podľa štatistické analýzy, ku ktorej sa vzťahujú na grafy:
  - porovnávacie, grafy rozloženia, grafy vývoj (posledné dve skupiny sa spoločne nazývajú aj obchodné grafy), grafy vyjadrujúce závislosti medzi javmi, atď.

Literatúra

BENČO, J. *Metodológia vedeckého výskumu*. Bratislava: IRIS, 2001. ISBN 80-9018-27-0.

BUDÍKOVÁ, M, KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. Praha: GRADA, 2010.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

GROFÍK, R. a kol. *Štatistika*. Bratislava: Príroda, 1987.

HINDLS, R., ARLTOVÁ, M., HRONOVÁ, S., MALÁ, I., MAREK, L., PECÁKOVÁ, I., ŘEZANKOVÁ, H. *Statistika v ekonomii*. Praha: Professional Publishing, 2018, ISBN 978-80-88260-09-7.

CHAJDIAK, J. a kol. *Štatistika jednoducho*. Bratislava: STATIS, 2003.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

MIKOLAJ, J., VANČO, B. *Štatistika pre manažérov*. Žilina: RVS vydavateľstvo FŠI ŽU v Žiline, 2000.

SOUČEK, E. *Základy pravděpodobnosti a statistiky*. Pardubice: Univerzita Pardubice, 2005.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

TIRPÁKOVÁ, A., MARKECHOVÁ, D. *Štatistika v praxi*. Nitra: FPV UKF, 2008, ISBN 978-80-8094-283-0.

3 Štatistické skúmanie a štatistický projekt

Štatistické skúmanie sa v praxi realizuje prostredníctvom výskumného resp. **štatistického projektu. Štatistický projekt** je možné popísať pomocou jednotlivých, na seba nadväzujúcich krokov (Obrázok 3):

- **príprava na realizáciu štatistického skúmania** (štatistického projektu) – plánovanie a návrh realizácie štatistického projektu; presné formulácie cieľa a účelu výskumu, štatistického problému, štatistickej jednotky, štatistických otázok alebo hypotéz, štatistických znakov, zostavenie dotazníka, a pod.,
- **štatistické zisťovanie** a **spracovanie získaných údajov** – štatistické zisťovanie je proces získavania štatistických údajov a uskutočňuje sa prostredníctvom štatistických metód a techník; získané štatistické údaje je následne potrebné previesť do takej podoby, v ktorej je možná ich ďalšia analýza (formálna úprava získaných údajov, očistenie a pod.),
- **štatistická analýza** – je používanie konkrétnych štatistických metód na riešenie štatistických otázok s cieľom vyvodzovania záverov.

Štatistická analýza je následne dopĺňaná:

- **formuláciou záverov a interpretáciou výsledkov,**
- **prezentáciou výsledkov,**
- **publikovaním výsledkov alebo aplikáciou získaných výsledkov v praxi,** napr. prijímanie opatrení vyplývajúcich zo záverov štatistického projektu; aplikácia výsledkov v praxi už nie je súčasťou štatistického projektu a väčšinou ani nie je úlohou riešiteľa tohto projektu.

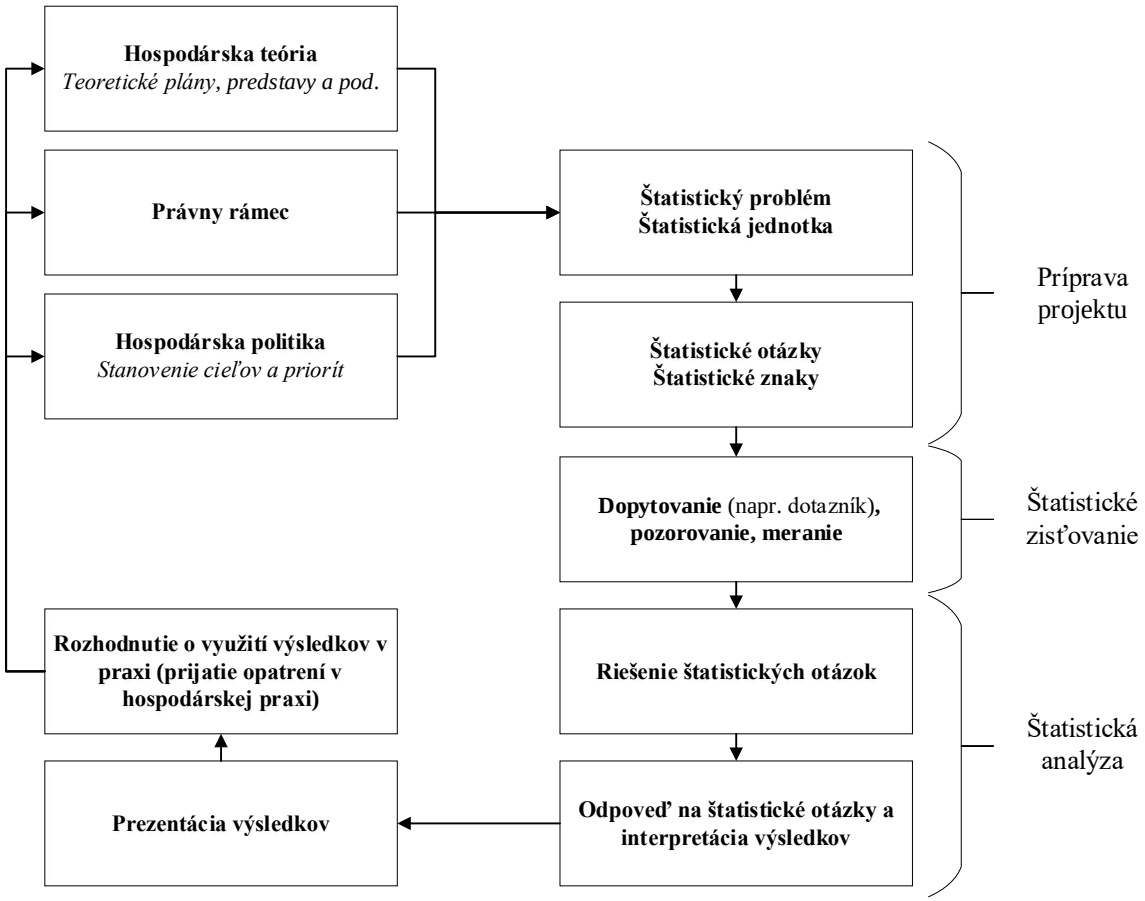
3.1 Metodický postup riešenia štatistického projektu

Metodika (postup) štatistického projektu vychádza zo spomínaných predošlých krokov, ktoré je možné detailnejšie popísať konkrétnymi činnosťami. Tie na seba logicky nadväzujú a nie je možné ich vynechať alebo navzájom prehodiť. Detailnejšie je možné formulovať postup riešenia štatistického projektu takto:

1. Formulácia štatistického problému, ktorý má priamu väzbu na skúmaný odborný, vedný alebo vecný problém. V rámci tohto bodu je nevyhnutné, aby mal riešiteľ prehľad v teórii a praxi danej problematiky a oboznámil sa s výsledkami predošlých výskumov.
2. Určenie štatistickej jednotky. Súčasne je potrebné sa zamýšľať nad rozsahom a obsahom štatistického súboru – požiadavky reprezentatívnosti (viac v kap. 4.2.2, 4.2.3., 5.1.2.).
3. Formulácia štatistických otázok (prípadne hypotéz), ktoré vychádzajú zo štatistického problému a ktorých riešenie bude štatistický problém vysvetľovať. Štatistické otázky musia mať priamy vzťah k štatistickej jednotke.
4. Identifikácia štatistických znakov zo štatistických otázok alebo hypotéz, ktorých obmeny (hodnoty) je potrebné získavať v štatistickom zisťovaní.

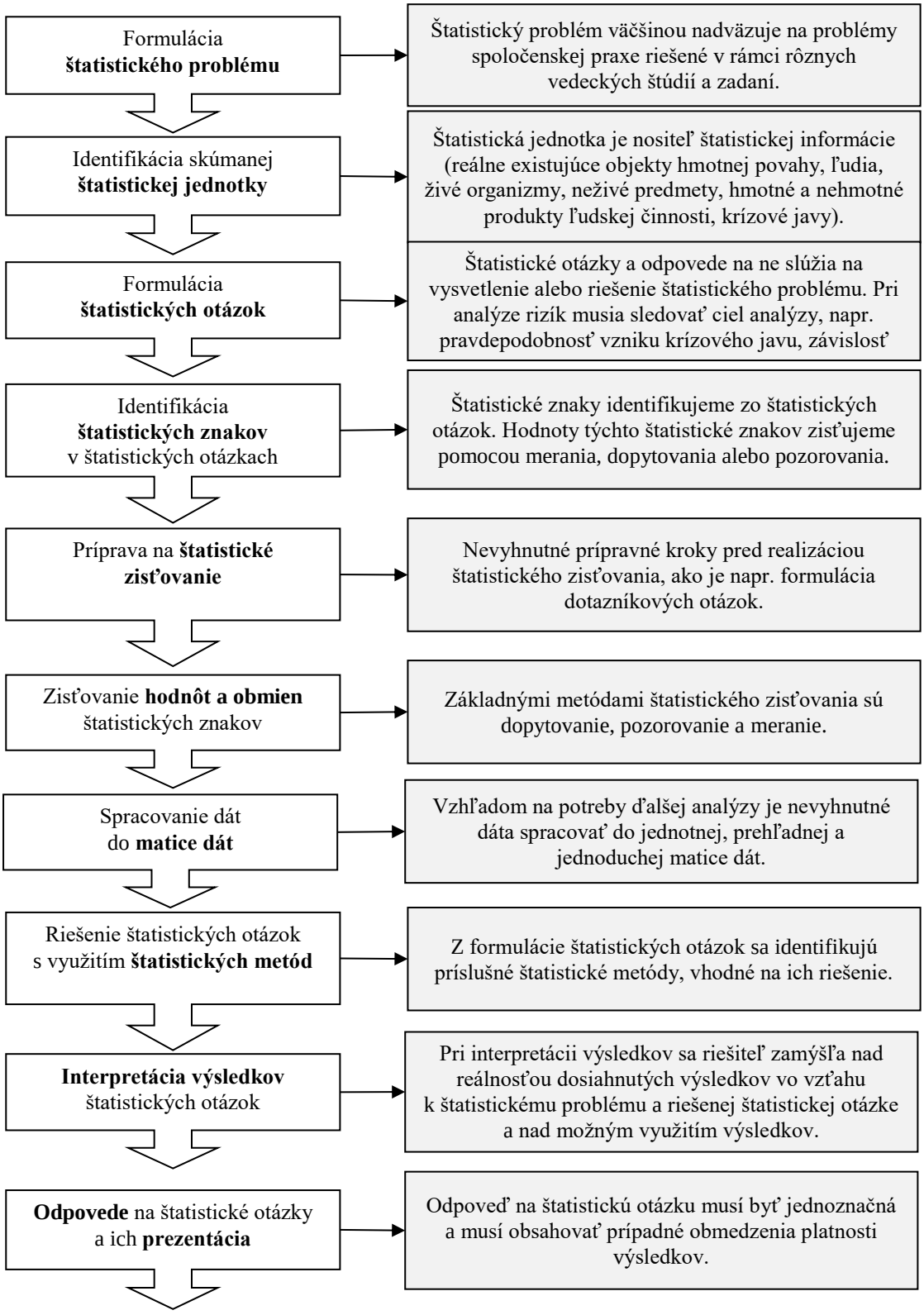
5. Príprava štatistického zisťovania, napr. zostavenie dotazníka (formulácia dotazníkových otázok), ktorý bude zohľadňovať identifikované štatistické znaky z predošlého kroku).
6. Realizácia štatistického zisťovania (napr. distribúcia a zbieranie dotazníkov) s ohľadom na stanovený rozsah a obsahové požiadavky štatistického súboru.
7. Spracovanie získaných štatistických údajov (väčšinou do tzv. matice dát).
8. Riešenie štatistických otázok pomoc konkrétnych štatistických metód, softwaru, grafických a tabuľkových nástrojov.
9. Interpretácia dosiahnutých výsledkov, formulácia záverov a odporúčaní pre prax.
10. Formulácia odpovedí na štatistické otázky a prezentovanie výsledkov.

Obrázok nižšie (Obrázok 3) ilustruje príklad zjednodušenej schémy štatistického projektu v spojitosti s problematikou hospodárskej praxe.



Obrázok 3 Schéma štatistického projektu

Komplexnejšia schéma postupu tvorby a riešenia štatistického projektu je zobrazená nižšie (Obrázok 4).



Obrázok 4 Algoritmus riešenia štatistického projektu

3.2 Príklady základných formulácií štatistického projektu

Príklad 1:

1. Štatistický problém: Dentálna hygiena u študentov.
2. Štatistická jednotka: Študent; rozsah štatistického súboru: n = 300 študentov.
3. Štatistické otázky:
    a. Aký je priemerný počet návštev zubára za rok?
    b. Existuje závislosť medzi počtom kazov študentov a ich nákladmi na dentálnu hygienu?
    c. Aké je rozdelenie študentov podľa preferovanej značky zubnej pasty a pohlavia?
4. Štatistické znaky: počet návštev zubára za rok, počet kazov, náklady na dentálnu hygienu, preferovaná zubná pasta, pohlavie, a pod.
5. Dotazníkové otázky:
    a. Koľkokrát ste tento rok navštívili zubára?
    b. Aký je Váš celkový počet kazov (aj ošetrených)?
    c. Aké sú približne Vaše ročné náklady na dentálnu hygienu?
    d. Aká je Vaša preferovaná značka zubnej pasty (uviesť možnosti)?
    e. Aké je Vaše pohlavie?
6. Metóda štatistického zisťovania: dopytovanie; technika dopytovania: anonymný dotazník.

Body 7. až 10. budú závisieť od získaných údajov a tej ktorej štatistickej otázky, ktorá sa bude riešiť.

Príklad 2:

1. Štatistický problém: Domová kriminalita v obci.
2. Štatistická jednotka: Domácnosť; rozsah štatistického súboru: n = 200 domácností.
3. Štatistické otázky:
    a. Aké je rozdelenie domácností v obci XY podľa počtu vykradnutí domu za posledných 50 rokov a typu zabezpečovacieho zariadenia?
    b. Aká je pravdepodobnosť vykradnutia domu v obci v nasledujúcom období (50 rokov) podľa typu zabezpečovacieho zariadenia?
    c. Existuje závislosť medzi mierou zabezpečenia a chránenou hodnotou v dome?
4. Štatistické znaky: typ zabezpečovacieho zariadenia, počet vykradnutí domu posledných 50 rokov, miera zabezpečenia domu, chránená hodnota v dome, a pod.
5. Dotazníkové otázky:
    a. Aký typ zabezpečovacieho zariadenia používate? (výber zo zoznamu možností)
    b. Koľkokrát bol Váš dom vykradnutý za posledných 50 rokov?
    c. Aká je miera zabezpečenia Vašej domácnosti? (ohodnoťte na stupnici 1 až 5)
    d. Aká je približne celková chránená hodnota vo Vašej domácnosti?
6. Metóda štatistického zisťovania: dopytovanie; technika dopytovania: napr. anonymný dotazník alebo rozhovor (štruktúrovaný).

Body 7. až 10. budú opäť závisieť od získaných údajov a tej ktorej štatistickej otázky, ktorá sa bude riešiť.

Literatúra

BENČO, J. *Metodológia vedeckého výskumu*. Bratislava: IRIS, 2001. ISBN 80-9018-27-0.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

GROFÍK, R. a kol. *Štatistika*. Bratislava: Príroda, 1987.

CHAJDIAK, J. a kol. *Štatistické úlohy a ich riešenie v Exceli*. Bratislava: STATIS, 2005.

CHAJDIAK, J. a kol. *Štatistika jednoducho*. Bratislava: STATIS, 2003.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

MIKOLAJ, J., VANČO, B. *Štatistika pre manažérov*. Žilina: RVS vydavateľstvo FŠI ŽU v Žiline, 2000.

SOUČEK, E. *Základy pravděpodobnosti a statistiky*. Pardubice: Univerzita Pardubice, 2005.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

TIRPÁKOVÁ, A., MARKECHOVÁ, D. *Štatistika v praxi*. Nitra: FPV UKF, 2008, ISBN 978-80-8094-283-0.

4 Príprava štatistického projektu

Kvalitný výskum predpokladá množstvo času stráveného prípravou výskumu a prípravou jeho realizácie. Je potrebné si uvedomiť, že formulovať štatistický problém, identifikovať prvky výskumu (štatistické jednotky), ich rozsah a obsah (štatistický súbor), identifikovať vlastnosti, veličiny, charakteristiky (štatistické znaky), ktoré je potrebné skúmať a zabezpečiť celý proces získavania údajov je prácou, ktorej je v prípravnej fáze projektu potrebné venovať patričnú pozornosť.

V skratke boli spomínané činnosti prípravy štatistického projektu priblížené v predchádzajúcej kapitole. Každá z týchto činností má svoje špecifiká a preto je potrebné ich náležite vysvetliť.

4.1 Formulovanie štatistického problému a hlavnej hypotézy výskumu

Prvotnou fázou štatistického projektu je formulovanie problému, ktorý si podľa riešiteľa vyžaduje ďalšie skúmanie. Môže to urobiť dvoma spôsobmi:

- **formuláciou hlavnej hypotézy výskumu,**
- **formuláciou štatistického problému.**

Forma hypotéz je uprednostňovaná vo vedeckej praxi, pričom aj jeden aj druhý spôsob vedie k používaniu podobných nástrojov štatistickej analýzy. Jeden aj druhý spôsob vyžaduje dostatočnú znalosť problematiky a výskumník by sa mal dobre oboznámiť s výsledkami predošlých výskumov s podobným zameraním. Nedostatok prehľadu v problematike môže spôsobiť viaceré problémy v rôznych bodoch riešenia, ale hlavne pri formulovaní záverov a výsledkov. Pre obidva spôsoby je dôležité mať dostatočne jasne určený **cieľ a účel** štatistického projektu.

Príklady určeného cieľa a účelu skúmania z lekárskej praxe:

Cieľ: Vytipovať osoby, ktoré sa ľahko môžu nakaziť chrípkou.
Účel: Realizovať preventívne opatrenia.

Cieľ: Získať poznatky o účinkoch dvoch liekov užívaných pri liečbe choroby.
Účel: Použiť lepší z dvoch liekov.

Cieľ: Zistiť príčiny vysokej detskej úmrtnosti v danej oblasti.
Účel: Vykonať potrebné zdravotnícke opatrenia.

Pri príprave štatistického projektu a vytváraní presných formulácií štatistického problému, hypotéz výskumu alebo cieľov výskumu, je vhodné, aby príslušný odborník začal spolupracovať s erudovaným štatistikom už na začiatku tohto procesu. Štatistik má byť oboznámený s najdôležitejšími aspektami problému z príslušnej oblasti (napr. lekárskej stránky) a odborník (lekár) musí získať základné štatistické znalosti.

4.1.1 Formulácia hypotéz výskumu

Vedecká prax a kvantitatívny výskum používa na formuláciu teoretického problému prednostne formu **hypotéz**.

Hypotéza je všeobecne:

- očakávanie o charaktere vecí, vyvodených z teórie,
- výrok, alebo tvrdenie o stave subjektu, objektu, javu, alebo o hodnote neznámeho parametra základného súboru,
- tvrdenie, ktoré predpovedá existenciu súvislosti medzi dvoma alebo viacerými štatistickými znakmi.

Výskumník definuje **nulovú hypotézu** (označujeme ju $\mathbf{H_0}$) a **alternatívnu hypotézu** (označujeme $\mathbf{H_1}$). Nulová hypotéza je tvrdenie o jednom (jednorozmerné hypotézy) alebo o viacerých štatistických znakoch populácie (viacrozmerné hypotézy). Je to tvrdenie, ktoré sa považuje za pravdivé. Alternatívna hypotéza je tvrdenie o možnostiach nepokrytých nulovou hypotézou, je disjunktnou oproti nulovej hypotéze. Navzájom sa vylučujú, môže platiť iba jedna. Alternatívna hypotéza sa testuje a buď ju riešiteľ príjme alebo zamietne.

Príklady jednorozmerných a dvojrozmerných hypotéz:

- jednorozmerné:
  - H_0 : automat vydáva 200ml kávy,
  - H_1 : automat nevydáva 200ml kávy,
- dvoj a viacrozmerné:
  - H_0 : výška osoby má vplyv na dĺžku lyží, ktoré používa,
  - H_1 : výška osoby nemá vplyv na dĺžku lyží, ktoré používa.

V spoločenských oblastiach (komplexnejšie problémy) sa prvotne stanoví **hlavná hypotéza** výskumu, ktorá sa ďalej skúma v podobe **pracovných hypotéz**. Nekomplikované problémy (príklady vyššie) si nevyžadujú toto delenie.

Hlavná hypotéza:

*„Čím viac **podnetov** týkajúcich sa krízových situácií osoba dostáva, tým väčšia je pravdepodobnosť, že bude **lepšie pripravená ich zvládať**.“*

V podobe pracovných hypotéz je táto hlavná hypotéza rozdelená na čiastkové tvrdenia, ktoré sa následne riešiteľ snaží prijať alebo vyvrátiť. Príklady nulových hypotéz môžu byť nasledujúce:

H_0: „Čím viac priamych skúseností osoba s krízovými situáciami má, tým je vyššia pravdepodobnosť, že bude mať materiálne zabezpečenie na svoju ochranu.“

H_0: „Čím viac evakuačných cvičení osoba absolvovala, tým lepšie ovláda postupy evakuácie.“

4.1.2 Formulovanie štatistického problému

Praktickejším spôsobom formulácie problému spoločenskej praxe je použitie formy **štatistického problému.** Štatistický problém sa formuluje z dôvodu nedostatočných alebo chýbajúcich informácií o nejakých hromadných javoch resp. nedostatočne objasnených problémoch spoločenskej praxe. Riešiteľ formuluje štatistický problém tak, aby zreteľne vyjadroval podstatu problému, ktorý je potrebné objasniť. Jeho formulácia by mala vychádzať z praktických skúseností, teoretických znalostí a potrieb riešiteľa (alebo potrieb spoločenskej

praxe); napr. lekár dlhoročnou praxou pozoruje, že u niektorých pacientov sa objavujú aj iné symptómy ako boli doposiaľ známe, a dané úvahy chce štatisticky podložiť, hľadať dôvody týchto prejavov a následne odporúčať opatrenia.

Objasniť je možné iba taký štatistický problém, ktorý je možno popísať konkrétnymi štatistickými znakmi a vzťahmi medzi nimi na známom (dostupnom) štatistickom súbore (napr. ľudia užívajúci daný liek).

Príklady:

Dentálna hygiena u študentov.

Preferencie politických strán medzi obyvateľmi SR.

V súvislosti s vyššie formulovanými cieľmi a účelmi z lekárskej praxe môžu byť štatistické problémy formulované takto:

Náchylnosť rôznych skupín obyvateľstva na chorobu XY.

Účinnosť lieku XY pri liečbe choroby Z.

Dôvody vysokej detskej úmrtnosti v oblasti XY.

V prepojení na vyššie uvedenú hlavnú hypotézu je alternatívou formulovať takýto štatistický problém:

Pripravenosť obyvateľstva zvládať krízové situácie.

Je zrejmé, že teoretický problém, je týmto spôsobom popísaný o niečo všeobecnejšie a ponúka širší priestor na skúmanie ako v prípade hlavnej hypotézy.

V spoločenskej praxi sa rozsah formulovania štatistického problému často limituje na skúmanie vplyvov určitých štatistických znakov (nezávislé, vysvetľujúce premenné) na iný skúmaný štatistický znak (závislú, vysvetľovanú premennú), napr. *Vplyv demografických faktorov na predaj konkrétneho výrobku.*

Riešiteľ je pri formulovaní štatistického problému (ale aj hypotéz výskumu) a jeho cieľov často limitovaný svojimi časovými, finančnými alebo personálnymi možnosťami. Vysoké nároky na rôzne zdroje majú hlavne časti získavania údajov a ich analyzovanie. Príliš ambiciózne formulované ciele štatistického projektu tak môžu naraziť na problém ich dosiahnutia, kvôli nedostatočným zdrojom.

4.2 Identifikovanie štatistických jednotiek a štatistického súboru

Po formulovaní štatistického problému alebo hlavnej hypotézy je potrebné určiť štatistickú jednotku a štatistický súbor (jeho rozsah), na ktorom sa bude štatistický problém riešiť, formulovať štatistické otázky a identifikovať taktiež štatistické znaky (kap. 4.3), ktoré bude potrebné skúmať.

4.2.1 Identifikovanie štatistickej jednotky

Identifikácia štatistickej jednotky vychádza z formulovaného štatistického problému. Častokrát je to jednoznačná a jednoduchá úloha, ale niektoré štatistické problémy je možné objasňovať z rôznych pohľadov. Riešiteľ tak pre jeden štatistický problém môže uvažovať o rôznych štatistických jednotkách.

Zjednodušeným príkladom môže byť štatistický problém *„Kriminalita v obciach“*. Štatistickou jednotkou môže byť:

- kriminálny čin – dostupné informácie o jednotlivých kriminálnych činoch, ktoré sa v obciach stali,
- páchateľ – skúmanie vlastností jednotlivých páchateľov a ich dôvodov páchania kriminálnej činnosti,
- obeť kriminálneho činu – skúmanie priamych skúseností obetí kriminálnej činnosti,
- respondent – skúmanie subjektívnych postojov, názorov na mieru kriminality (a bezpečnosti) u náhodných respondentov (nemusia mať priamu skúsenosť s kriminálnym činom).

4.2.2 Rozsah štatistického súboru

Veľkosť vzorky alebo rozsah výberového štatistického súboru je počet respondentov v prieskume, resp. počet pozorovaných štatistických jednotiek alebo počet meraných štatistických jednotiek, ktoré sú zahrnuté do štatistického skúmania. Predstavuje iba časť celej populácie.

Počet štatistických jednotiek, ktorých riešiteľ potrebuje skúmať, závisí od cieľov výskumu resp. štatistického projektu a od toho, ako si chce byť riešiteľ istý svojimi výsledkami. Čím vyššiu spoľahlivosť výsledkov chce riešiteľ dosiahnuť, tým menšiu chybovosť by mal akceptovať.

S výpočtom rozsahu štatistickej vzorky súvisia okrem už uvedených pojmov ešte nasledujúce:

- **miera chyby** (angl. margin of error) **–** percento, ktoré hovorí, do akej miery môže riešiteľ očakávať, že výsledky skúmania budú odrážať názory celej populácie; čím je chyba menšia, tým bližšie je riešiteľ k presnej odpovedi na danej úrovni spoľahlivosti; bežne sa pripúšťa 5% miera chyby (resp. do výpočtu sa používa v tvare 0,05), prípadne pre ešte presnejšie výsledky 3% alebo 1% miera chyby,
- **úroveň spoľahlivosti** (angl. confidence level) **–** percento, ktoré hovorí o tom, nakoľko si môže byť riešiteľ istý, že populácia vyberie odpoveď v rámci určitého rozsahu; napríklad 95% úroveň spoľahlivosti znamená, že si riešiteľ môže byť na 95% istý, že výsledky sa nachádzajú medzi číslami x a y; bežne sa používa 95% úroveň spoľahlivosti, prípadne pre dosiahnutie vyššej spoľahlivosti sa používa 99% úroveň spoľahlivosti.

Výpočet počtu štatistických jednotiek do výberového štatistického súboru (vzorky) je možné vypočítať nasledovným vzorcom:

$$veľkosť\ vzorky = \frac{\frac{z^2 \times p(1-p)}{e^2}}{1 + \frac{z^2 \times p(1-p)}{e^2 N}}$$

kde: *z* je hodnota z – skóre (angl. z-score alebo critical value z) – je to počet štandardných odchýlok, ktorých podiel sa líši od priemeru; vychádza z hodnoty úrovne spoľahlivosti a používa sa do prepočtov namiesto tejto hodnoty (Tabuľka 1),

p je hodnota pravdepodobnosti distribúcie odpovedí – používa sa hodnota 0,5, ak je rovnaká pravdepodobnosť odpovedí v populácií (napr. odpovede áno/nie na nejakú otázku) – platí pre alternatívne štatistické znaky kedy je predpoklad rovnakého rozloženia daného štatistického znaku, resp. sa bežne používa ak riešiteľ nepozná rozloženie štatistického znaku v populácii, tým dostáva najkonzervatívnejšiu a súčasne najväčšiu veľkosť potrebnej vzorky,

e je hodnota miery chyby (uvádza v desiatkovej podobe),

N je veľkosť populácie.

Tabuľka 1 Hodnoty z-skóre pre požadovanú úroveň spoľahlivosti štatistickej vzorky

Požadovaná úroveň spoľahlivosti	z - skóre
80%	1,28
85%	1,44
90%	1,65
95%	1,96
97%	2,17
99%	2,58
99,9%	3,29

Tabuľka 2 Veľkosť vzorky vzhľadom na veľkosť populácie, mieru chyby a úroveň spoľahlivosti

Veľkosť populácie	Úroveň spoľahlivosti 95%				Úroveň spoľahlivosti 99%			
	Miera chyby							
	5%	3%	2,5%	1%	5%	3%	2,5%	1%
100	80	91	94	99	87	95	96	99
500	217	340	377	475	285	394	421	485
1000	278	516	606	906	399	649	727	943
10000	370	964	1332	4899	622	1560	2098	6239
100000	383	1055	1513	8762	659	1815	2585	14227
500000	384	1065	1532	9423	663	1842	2640	16055
1000000	384	1066	1534	9512	663	1845	2647	16317

Požadovaná veľkosť vzorky je väčšia pre menšie prípustné miery chyby alebo vyššiu úroveň požadovanej spoľahlivosti. Ak je veľkosť populácie veľmi veľká (>100000), potom sa veľkosť potrebnej vzorky mení iba minimálne.

Bežné prieskumy *preferencií politických strán*, ktoré sú známe a často prezentované médiami počítajú prevažne s 3% mierou chyby pri 95% hladine spoľahlivosti, čo pre populáciu voličov Slovenskej republiky znamená veľkosť vzorky okolo 1000 respondentov.

Príklad vysvetlenia miery chyby:

Riešiteľ posiela dotazník s otázkou: Či sú rodičia detí vo vašej škole za predĺženie vyučovacieho dňa, a možnosťami odpovede „Áno“ alebo „Nie“. Celkový počet rodičov, ktorý je možné skúmať (veľkosť populácie) je 10 000 a riešiteľovi vyhovuje miera chyby ± 10%. Pomocou prepočtu riešiteľ zistí, že na absolvovanie prieskumu potrebuje približne 95 ľudí (pre úroveň spoľahlivosti 95%). 70% zo 95 opýtaných rodičov odpovedalo, že je za predĺžený vyučovací deň. To znamená, že riešiteľ môže predpokladať, že ak by na prieskum odpovedalo všetkých 10 000 rodičov, 60% až 80% ľudí by bolo za predĺženie vyučovacieho dňa.

Príklad vysvetlenia úrovne (miery) spoľahlivosti:

Úroveň spoľahlivosti riešiteľovi povie, ako spoľahlivé meranie (prieskum) je. 95% úroveň spoľahlivosti znamená, že ak by sa ten istý prieskum mal opakovať 100-krát za rovnakých podmienok, 95-krát zo 100 by výsledky ležali niekde v stanovených hraniciach chyby.

4.2.3 Kvalita štatistického súboru

Vhodný počet respondentov (rozsah štatistického súboru) nezaručuje spoľahlivé výsledky pre formulovaný štatistický problém. Štatistická vzorka by súčasne mala byť dostatočne kvalitná. Kvalita štatistického súboru závisí od spôsobu štatistického zisťovania (kap. 5.1). V praxi sa najčastejšie stretávame s neúplným štatistickým zisťovaním, v ktorom je potrebné spĺňať podmienky reprezentatívnosti – čiže štatistická vzorka by mala čo najvernejšie kopírovať (podobať sa svojim zložením) populácii (viac v kapitole 5.1.2).

Ak dokáže riešiteľ zabezpečiť dostatočný rozsah a kvalitu štatistickej vzorky, dokáže s patričnou spoľahlivosťou interpretovať výsledky svojho skúmania.

4.3 Identifikovanie štatistických znakov

Štatistické znaky sa **často nesprávne** určujú ako prvé a potom sa na tých istých dátach overujú naše predpoklady o povahe skúmaného štatistického problému. Výskumník si stanoví štatistický problém a potom hľadá štatistické znaky, ktoré s daným štatistickým problémom súvisia bez hlbšieho zamyslenia sa nad možnými súvislosťami medzi zvolenými štatistickými znakmi. Problém spočíva najmä v tom, že riešiteľ vytvorí najskôr dotazníkové otázky, často pritom netuší čo je štatistická jednotka, štatistický znak a pod. Potom realizuje zber údajov. Následne rozmýšľa ako so získanými údajmi pracovať a snaží sa vyvodzovať nejaké závery na základe zistených skutočností. V danom momente, ale často zistí, že na formulovanie záverov potrebuje ešte hodnoty ďalšieho štatistického znaku, ktorý, ale k dispozícii nemá (napr. do dotazníkového zisťovania riešiteľ nezahrnul patričnú otázku). Jeho príprava štatistického projektu bola nevhodná/nedostatočná. Riešiteľ tak obmedzuje svoje závery iba na údaje, ktoré má k dispozícii. Takýto prístup veľmi limituje objasnenie formulovaného štatistického problému.

Vhodnejšou cestou je **určiť (identifikovať) štatistické znaky** na základe vopred formulovaných:

- **hypotéz výskumu,**

- **štatistických otázok.**

To si vyžaduje v prípravnej fáze štatistického projektu presne určiť aké skutočnosti potrebuje riešiteľ objasniť, aké vzťahy potrebuje riešiteľ skúmať alebo vplyvy akých štatistických znakov (nezávislé alebo vysvetľujúce premenné) na iné štatistické znaky (závislú premennú, vysvetľovanú premennú) potrebuje riešiteľ objasniť. Súčasne je nevyhnutná znalosť štatistických metód, ktoré štatistická teória poskytuje.

4.3.1 Identifikovanie štatistických znakov prostredníctvom hypotéz

Najbežnejšou cestou kvantitatívneho výskumu je už spomínaná formulácia **hypotéz.** Identifikovanie štatistických znakov pre komplexnejšie teoretické problémy vychádza z formulovanej hlavnej hypotézy.

Príklad:

*„Čím viac **podnetov** týkajúcich sa krízových situácií osoba dostáva, tým väčšia je pravdepodobnosť, že bude **lepšie pripravená ich zvládať**"*

Riešiteľ následne určuje na jednej strane možné podnety, ktoré môže osoba dostávať ohľadom krízových situácií a na druhej strane určuje kedy je osoba pripravená (vlastnosti, ktoré charakterizujú pripravenosť osoby). Týmto spôsobom vlastne identifikuje štatistické znaky.

Pre danú hlavnú hypotézu by mohli byť na strane „podnetov" relevantné štatistické znaky identifikované takto:

- počet absolvovaných evakuačných cvičení (alebo iných cvičení),
- počet skúseností (priame) s pôsobením krízovej situácie v minulosti,
- počet skúseností (nepriamych – sprostredkovaných od známych, rodiny) s pôsobením krízovej situácie v minulosti,
- frekvencia sledovania relácií, čítania informácií o spôsoboch chránenia sa proti účinkom krízových situácií
- atď.

Na strane popisujúcej vlastnosti pripravenosti by mohli byť relevantné štatistické znaky identifikované takto:

- materiálne zabezpečenie proti pôsobeniu konkrétnych krízových situácií (napr. vlastnenie náhradných zdrojov energie, a pod.),
- potravinové zabezpečenie (napr. koľko dní vystačia potraviny),
- poistenie (napr. nehnuteľnosti),
- znalosť postupov prvej pomoci,
- znalosť postupov evakuácie,
- atď.

Riešiteľ dokáže následne formulovať pracovné hypotézy a redukovať ich na tie, ktoré mu umožnia v patričnej miere objasniť hlavnú hypotézu (príklady sú uvedené v kapitole 4.1.1). V praxi je bežnejší a odporúčaný postup, v ktorom si riešiteľ najprv formuluje konkrétne pracovné hypotézy, ktoré musí nevyhnutne testovať (prijať, zamietnuť). Z formulovaných

pracovných hypotéz následne vie riešiteľ jednoducho identifikovať štatistické znaky, ktorých počet bude jednoznačne daný formulovanými hypotézami. Pracovné hypotézy sú preto v tomto zmysle dôležitým nástrojom na optimalizáciu a redukciu potrebných informácií (štatistických znakov) pre prieskum. Súčasne sú aj testom, že výskum je možný, pomáhajú odhadovať rozsah výskumu a v neposlednom rade sú podporou na voľbu metód a techník štatistického zisťovania.

Dôležité je doplniť, že častokrát sú medzi relevantné štatistické znaky uvažované aj demografické (napr. vek, pohlavie,...) a socioekonomické faktory (príjem,...), ktoré môžu taktiež ovplyvniť skúmaný štatistický problém (pre uvedený príklad môžu ovplyvniť pripravenosť človeka zvládať krízové situácie).

4.3.2 Identifikovanie štatistických znakov prostredníctvom štatistických otázok

Prax nevyužíva stále na identifikovanie štatistických znakov hypotézy, pretože nie vždy je potrebná taká hĺbka skúmania teoretického problému. Väčšinou stačí štatistický problém širšie popísať na základe sledovaných štatistických znakov (deskriptívna štatistika) a hľadať iba základné vzťahy medzi nimi. Na tento účel sa využívajú štatistické otázky, ktorých formulácia je previazaná s použitím konkrétnych štatistických metód deskriptívnej a analytickej štatistiky (príklad v kapitole 3.2).

Formulácia takýchto štatistických otázok teda predpokladá znalosť konkrétnych metód a účelu ich použitia. Účel použitia jednotlivých štatistických metód je objasnený v ďalších kapitolách.

Podobne ako pracovné hypotézy, aj riešenie štatistických otázok bližšie objasňuje formulovaný štatistický problém a umožňuje pre riešiteľa identifikovať štatistické znaky, ktorých obmeny/hodnoty potrebuje zistiť.

Príklad štatistického problému:

Pripravenosť obyvateľstva zvládať krízové situácie.

Podobne ako v prípade hypotéz sa pri identifikácii štatistických znakov riešiteľ musí zamyslieť nad viacerými stránkami štatistického problému (v jednoduchších problémoch to nemusí byť pravidlo). Teda čo vlastne znamená pripravenosť človeka zvládať krízovú situáciu (Kedy je pripravený? Aké vlastnosti - štatistické znaky musí spĺňať?) a aké aspekty prostredia (ďalšie štatistické znaky), ktoré ovplyvňujú pripravenosť človeka chce skúmať. V spoločenských problémoch sa väčšinou zohľadňujú aj vybrané socio-ekonomické štatistické znaky (pohlavie, vek, sociálny status, vzdelanie, a pod.), ktoré odlišujú rôzne skupiny obyvateľstva.

Štatistické otázky umožňujú skúmať štatistické znaky individuálne (deskriptívne, popisne), ale aj v kombinácii dvoch štatistických znakov (deskriptívne aj analyticky).

Príklady štatistických otázok (príklady sú uvedené aj v kapitole 3.2):

*Aký je priemerný **počet zažitých krízových situácií** u respondentov?*

*Aká časť respondentov má **absolvované kurzy prvej pomoci** (alebo kurz civilnej ochrany)?*

*Akými **materiálnymi prostriedkami** respondenti najčastejšie disponujú v prípade vzniku krízovej situácie?*

*Aké sú rozdiely v **subjektívnom hodnotení pripravenosti** respondentov na krízové situácie vzhľadom na **pohlavie** respondentov?*

*Existuje závislosť medzi **vekom** respondenta a **počtom zažitých krízových situácií?***

Z daných štatistických otázok je následne možné identifikovať štatistické znaky (potrebné informácie – vyznačené hrubým písmom), ktorých hodnoty a obmeny potrebuje riešiteľ skúmať, teda zahrnúť do štatistického zisťovania.

4.4 Príprava na štatistické zisťovanie

Posledným krokom v rámci prípravnej fázy štatistického projektu je príprava štatistického zisťovania. Z predchádzajúceho kroku sa identifikovali štatistické znaky, ktoré potrebuje riešiteľ skúmať, avšak samotné získavanie obmien a hodnôt týchto štatistických znakov je potrebné následne zabezpečiť. Štatistické zisťovanie je potrebné zabezpečiť z viacerých stránok:

- obsahová stránka (napr. zostavenie dotazníkových otázok a príprava samotného dotazníka v určitej forme),
- materiálna a technická stránka (hardvér a softvér na spracovanie vyplnených dotazníkov; ak je dotazník distribuovaný v tlačenej podobe je potrebné zabezpečiť papier a tlač; ak sa jedná napr. o experimentálne meranie horenia v laboratóriu je potrebné zabezpečiť meracie prostriedky a materiál, ktorý sa bude skúmať, a pod.),
- organizačná stránka (napr. je potrebné naplánovať priebeh experimentu; postup zberu údajov teréne, a pod.),
- personálna stránka (prerozdelenie úloh pre členov riešiteľského kolektívu – zber údajov v teréne môže byť personálne vyťažujúci),
- finančná stránka – všetky spomínané činnosti zväčša vyžadujú aj finančné prostriedky na ich zabezpečenie (od materiálu až po zaplatenie anketárov či spoločnosti, ktorá robí osobný zber údajov na dohodnutom mieste).

Metódy a konkrétne techniky štatistického zisťovania sú priblížené v nasledujúcej kapitole.

Literatúra

COCHRAN, W. G. *Sampling Techniques*. J. Wiley and Sons, NewYork, 1972

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DELL, R.B., HOLLERAN, S., RAMAKRISHNAN, R. *Sample Size Determination*. 2002. ILAR Journal, roč. 43, č. 4, 2002, s. 207–213.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

HAZRA, A. Using the confidence interval confidently. 2017. Journal of Thoracic Disease, roč.9, č.10, s. 4124-4129.

CHAJDIAK, J. *Analýza dotazníkových údajov*. Bratislava: Statis, 2013. ISBN 978-80-85659-76-4.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

PECÁKOVÁ, I. *Statistika v terénních průzkumech*. 2018. 3.vyd. Professional Publishing, Praha. ISBN 978-80-88260-10-3.

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

TAHERDOOST, H. *Determining Sample Size; How to Calculate Survey Sample Size*. 2017. International Journal of Economics and Management Systems, č. 2, s. 237-239.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

5 Získavanie údajov pre štatistický projekt

5.1 Štatistické zisťovanie

Štatistické zisťovanie resp. získavanie údajov pre štatistický projekt spočíva v získavaní, zhromažďovaní a zaznamenávaní údajov o štatistických jednotkách.

Základné rozdelenie spôsobov získavania údajov:

Získavania údajov **podľa zdroja**:

- primárne – získavanie údajov priamo od zdroja,
- sekundárne – sprostredkované údaje napr. z databáz, internetu alebo prepočítané údaje (kap. 5.3).

Získavania údajov **podľa reálnosti situácie**:

- skutočné – údaje získané z reálnej situácie,
- simulované – údaje získané napr. ako výsledok matematickej alebo počítačovej simulácie.

Získavanie údajov **podľa periodicity**:

- priebežné (bežné) – zaznamenanie ihneď, keď skutočnosť nastala (výskyt pracovného úrazu, narodenie dieťaťa, odchod do dôchodku),
- periodické – pravidelne opakované získavanie údajov v stanovených časových periódach (každoročné získavania údajov o úrode, denné meranie teploty),
- jednorázové – získavania údajov organizovane jednorazovo pri zvláštnych prípadoch (sčítanie obyvateľstva, škody po živelnej pohrome, inventúra).

Získavanie údajov **podľa časového hľadiska**:

- okamihové – údaje zisťované k určitému časovému okamihu (početný stav zamestnancov, peňažné prostriedky na účte v banke, objem zásob určitého výrobku,
- intervalové – údaje zisťované za určité obdobie - (výroba určitého produktu za mesiac, vynaložené náklady na cestovanie do práce za týždeň).

Získavanie údajov **podľa stupňa kontroly podmienok pri získavaní údajov**:

- prosté pozorovanie – napr. získavanie údajov značiek prechádzajúcich vozidiel,
- riadený experiment – získavanie údajov resp. výsledkov napr. pri experimentálnych testoch nových zlúčenín látok, experimenty o účinkoch nových liekov, vakcín (kap. 9.5).

Získavanie údajov **podľa rozsahu**:

- úplné (vyčerpávajúce) – vhodné u menej rozsiahlych súborov, podrobné informácie o každej jednotke súboru, spravidla je potrebná dlhšia doba a vyššie náklady na realizáciu získavania údajov; príkladom úplného zisťovania je súpis obyvateľstva (u nás nazývaný "sčítanie obyvateľstva") alebo skúmanie všetkých zamestnancov jednej firmy.

- neúplné (nevyčerpávajúce) – údaje zisťované iba u vybraných jednotiek, využíva sa teória pravdepodobnosti, obvykle je potrebný kratší čas a nižšie náklady na získavanie údajov (oproti úplnému zisťovaniu), môžu byť:
    - reprezentatívne – výberové (zámerný výber) zisťovanie, pričom každá štatistická jednotka musí mať rovnakú pravdepodobnosť, že bude vybraná do štatistického súboru,
    - nereprezentatívne – náhodný výber štatistických jednotiek bez stanovených kritérií výberu.

5.1.1 Úplné štatistické zisťovanie

Úplné zisťovanie má oproti neúplnému tri nesporné **výhody**:

1. Úplné zisťovanie poskytuje presné charakteristiky súboru (súčty a priemery, mieru variability, miery závislosti, indexy vývoja a pod.). Neúplné zisťovanie poskytuje presné charakteristiky iba pre preskúmanú časť súboru; za celý súbor môže poskytnúť iba približné hodnoty týchto charakteristík.
2. Úplné zisťovanie poskytuje nielen informácie o súbore, ale taktiež o každom jednotlivom prvku.
3. Úplné zisťovanie sa stretáva s väčším pochopením u respondentov (osôb, inštitúcií) než zisťovanie neúplné.

Popri týchto prednostiach má však úplné zisťovanie aj niektoré **nevýhody**. Sú to hlavne:

1. Praktická neuskutočniteľnosť takých zisťovaní, ktoré vedú k deštrukcii výrobkov alebo tovarov; veľmi nákladné alebo časovo neúnosne zdĺhavé zisťovanie u rozsiahlych súborov, ktoré by mohlo na časti súboru priniesť dosť nepresné až chybné výsledky, prípadne veľa odmietnutí odpovede – tzv. non-response) a pod.
2. Nehospodárnosť (vysoké náklady vzhľadom k efektu získaných výsledkov), časová náročnosť; príkladom môže byť skúmanie preferencií občanov na konkrétne ponúkané produkty prostredníctvom dotazníka.

Nevýhody úplného zisťovania vedú k tomu, že sa často musí použiť **neúplné** zisťovanie.

5.1.2 Neúplné (výberové) štatistické zisťovanie

Ak je stanovené, že sa určité štatistické zisťovanie uskutoční ako **neúplné**, je potrebné rozhodnúť, akým spôsobom (technikou) bude výber vykonaný. Podkladom pre toto rozhodnutie je predovšetkým veľkosť, štruktúra a stupeň rovnorodosti súboru, ktorý bude podrobený výberovému skúmaniu, ďalej charakteristiky, ktoré chce riešiteľ odhadovať, a odhady (odhadové funkcie), ktoré na ten účel chce použiť. V neposlednom rade sú dôležité finančné, časové, personálne, prípadne iné možnosti riešiteľa.

Pri neúplnom štatistickom zisťovaní sa dopredu (vedome) počíta s tým, že zisťovaniu budú podrobené iba niektoré prvky (štatistické jednotky) populácie, ide o zisťovanie **nevyčerpávajúce alebo čiastkové.**

Hlavnými **technikami** (druhmi) neúplného zisťovania sú:

- anketa,
- metóda základného masívu,
- snehová guľa (snowball),
- úsudkový výber,
- reprezentatívny výber,
- pravdepodobnostný (náhodný) výber.

Anketa / Dotazník

Je taký spôsob štatistického zisťovania, pri ktorom sa určitému okruhu osôb, podnikov, inštitúcií a pod. (štatistickým jednotkám) distribuujú dotazníky s dôkladne zostavenými dotazníkovými otázkami a so žiadosťou o vyplnenie dotazníka a o jeho vrátenie. Na žiadosť o vyplnenie dotazníka reaguje spravidla iba určitá časť dopytovaných – často iba pomerne malá časť zo všetkých dopytovaných (v priemere asi jedna tretina a ešte menej).

Charakteristiky, ako sú priemery, rôzne pomerné hodnoty a iné zhrňujúce charakteristiky štatistického súboru, ktoré sa vypočítajú na základe údajov zo získaných odpovedí, nie je možné považovať za všeobecne platné, pretože medzi vyplnením resp. odmietnutím vyplnenia dotazníka na jednej strane a dopytovanou skutočnosťou na druhej strane býva často úzky vzťah (súvislosť, asociácia). Napríklad dotazník o príjme je často nevyplnený a nevrátený osobami s relatívne vysokými (alebo utajovanými príjmami); dotazník o čitateľských záujmoch často nevrátia čitatelia zábavnej literatúry a pod.

Metóda základného masívu

Ak prebieha skúmaný jav (napr. výroba alebo predaj určitého tovaru) vo veľkých subjektoch (priemyselných závodoch, obchodných domoch), prevažne stačí na získanie hrubého odhadu o objeme alebo kvalite tohto tovaru preveriť iba tieto veľké subjekty a malé subjekty vynechať. Prednosťou tohoto druhu neúplného zisťovania je, že sa ušetrí veľa práce (zistenie hodnoty príslušného štatistického znaku vo veľkom subjekte nebýva oveľa prácnejšie než v malom subjekte), ale pritom sa podchytí prevažná časť skúmaného javu, napr. 80 % produkcie, 96 % obratu a pod. Táto metóda nedovoľuje zovšeobecňovať získané charakteristiky na celý súbor. Vynechané malé subjekty sa preto niekedy podrobia ešte výberovému zisťovaniu.

Technika snehovej gule (snowball)

Niektoré štatistické problémy sú zamerané na veľmi špecifické témy (niekedy nepríjemné témy), ktoré sa dotýkajú iba úzkeho okruhu respondentov, ktorí zväčša nie sú verejne známi, nechcú byť známi, pochádzajú z uzavretej komunity alebo sa len ťažko hľadajú. V takýchto prípadoch teda nepoznáme rozloženie populácie. Príkladom môže byť štatistický problém v oblasti drogovej závislosti, pričom je vhodné priamym spôsobom osloviť konkrétnych potenciálnych respondentov. V tomto prípade je vhodné začať štatistické zisťovanie u jednej osoby, ktorá „odporučí“ ďalších, tí poznajú ďalších, a tak sa môže rozsah súboru zväčšiť na požadovanú úroveň. Pre takýchto respondentov je zväčša nevyhnutné zabezpečiť anonymitu. Môže ísť o skupiny ľudí na hranici zákona („veksláci“ sa medzi sebou poznajú). Podobne to môžu byť priaznivci nejakej politickej strany, ktorých je medzi celou populáciou

ťažko identifikovať bežnými technikami štatistického zisťovania. Princíp tejto techniky spočíva teda v postupnom nabaľovaní (ako snehová guľa) ďalších štatistických jednotiek, ktoré majú požadované vlastnosti skúmané v rámci formulovaného štatistického problému.

Úsudkový alebo zámerný výber

Realizuje ho skúsený riešiteľ (pozor na subjektivitu výberu) a vyberá tie štatistické jednotky, o ktorých sa domnieva, že najlepšie umožnia vykonať zamýšľané štatistické skúmanie. Počet štatistických jednotiek pri úsudkovom výbere býva predom daný finančnými alebo podobnými (napr. pracovnými) možnosťami inštitúcie, ktorá je poverená vykonaním štatistického zisťovania.

Je zrejmé, že na vykonanie úsudkového výberu sú nutné určité predbežné znalosti o skúmanom štatistickom súbore. Ani potom však nie je zaručené, že vybrané štatistické jednotky dobre reprezentujú skúmaný základný súbor. Početné skúsenosti z praxe totiž ukazujú, že i najskúsenejší a úplne nezaujatí znalci majú tendenciu robiť odchýlky (od priemeru) jedným alebo druhým smerom, t.j. dopúšťajú sa systematických chýb. Závažným nedostatkom úsudkového výberu je taktiež nemožnosť objektívne stanoviť presnosť odhadov zostrojených na jeho základe, t.j. vypočítať nejakú priemernú alebo maximálnu chybu odhadu.

Reprezentatívny výber

Ak chce riešiteľ vztiahnuť výsledky získané z neúplného štatistického zisťovania na celú populáciu skúmaných štatistických jednotiek musí dodržať tzv. **požiadavky reprezentatívnosti** resp. uskutočniť **reprezentatívny výber**. Celou populáciou sa myslí súbor štatistických jednotiek, pre ktorý chce riešiteľ vyvodzovať zistené závery štatistického skúmania. Môžu to byť napr. obyvatelia celého štátu alebo konkrétneho mesta, študenti celého štátu alebo jednej univerzity, výrobky z odvetvia priemyslu alebo v jednom podniku, celé územie štátu alebo časť územia a pod.

Reprezentatívny výber spočíva v tom, že z celej populácie sa vyberajú také štatistické jednotky (vzorka), ktoré najlepšie reprezentujú celú populáciu. Výber sa vykonáva podľa dopredu stanovených kritérií, napr. podľa národnosti, vzdelania, veku. Kritéria sa delia na:

- **Identifikačné kritéria** – tie slúžia na prvotnú selekciu štatistických jednotiek a skúmanie ich príslušnosti k celej populácii (používajú sa aj pri iných druhoch výberov vzorky). Na skúmanie študentov Žilinskej univerzity je identifikačným kritériom príslušnosť osloveného študenta k danej univerzite.
- **Reprezentatívne kritéria** – tie slúžia na zabezpečenie reprezentatívnosti výberového súboru. Ako reprezentatívne kritéria je možné používať iba také kritéria, u ktorých je dopredu známe ich rozloženie v celej populácii. Napríklad na základe údajov štatistického úradu riešiteľ pozná rozloženie obyvateľstva podľa vzdelanosti, veku, národnosti; na základe interných záznamov univerzity je možné zistiť rozloženie študentov danej univerzity podľa príslušnosti ku konkrétnej fakulte, pohlavia, veku, a pod.). Je potrebné si uvedomiť, že ak je vzorka reprezentatívna iba z pohľadu jedného kritéria (napríklad štatistického znaku „vek“), tak aj o výsledkoch je možné tvrdiť, že sú reprezentatívne iba z pohľadu štatistického znaku „vek“, nie celkovo.

Na zabezpečenie reprezentatívnosti výberového súboru musí byť rozloženie tohto výberového súboru (vzorky) rovnaké ako je rozloženie celej populácie (prípadne sa jej musí čo najviac podobať). Výber štatistických jednotiek do reprezentatívnej vzorky sa vykonáva najčastejšie nasledujúcimi spôsobmi:

- **Subjektívnym výberom** takých štatistických jednotiek, o ktorých sa riešiteľ domnieva, že sú to typické štatistické jednotky pre daný štatistický súbor, t.j. jednotky s hodnotami skúmaného štatistického znaku blízkymi priemeru alebo s hodnotami modálnymi (najčastejšie sa vyskytujúcimi). Niekedy sa tento spôsob výberu nazýva **typický výber**.
- **Zostavením výberového súboru (kvótny výber)**, v ktorom je rozloženie početností u známeho štatistického znaku (kritérium reprezentatívnosti) totožné s rozdelením v základnom štatistickom súbore. Často sa ako kritérium reprezentatívnosti použije viac štatistických znakov súčasne; volia sa aj štatistické znaky kvalitatívne (slovné). Riešiteľ sa snaží o zhodu rozloženia výberového a základného štatistického súboru. Čím viac reprezentatívnych kritérií je splnených pre vzorku v porovnaní s populáciou, tým je väčší predpoklad, že výsledky budú súčasne reprezentatívne a zároveň presnejšie pre celú populáciu. Na porovnávanie rozloženia vzorky a populácie sa používa percentuálny podiel obmien jednotlivých kritérií; napríklad podľa vekových kategórií, národnosti, typu vzdelania – ak je v populácii 40% vysokoškolsky vzdelaných ľudí, tak z pohľadu reprezentatívnosti by aj vzorka mala obsahovať taký podiel vysokoškolsky vzdelaných ľudí.

Zostavenie výberového súboru sa niekedy nazýva aj **kvótny výber alebo metóda dokonalého prierezu**. Splnenie uvedených zásad, ako ukázali početné skúsenosti z praxe, však nemusia ešte zaručiť, že výber bude dobrým reprezentantom základného štatistického súboru taktiež pre štatistické znaky, ktorých rozloženie v populácii nie je známe alebo je skúmané. Zhoda rozloženia vzorky a populácie v štatistických znakoch vek, pohlavie, povolanie a miesto bydliska (dedina, mesto, veľkomesto) môže byť niekedy nečakane "narušená" rozdielnym rozložením populácie a vzorky podľa najvyššieho stupňa vzdelania. Typickým príkladom kvótneho výberu, ktorý sa snaží o reprezentatívnosť vzorky, je *skúmanie preferencií politických strán*. V daných prieskumoch štatistická vzorka kopíruje zloženie všetkých obyvateľov s volebným právom v zadaných reprezentatívnych kritériách (kvótnych štatistických znakoch). Väčšinou sú to štatistické znaky: vek, pohlavie, vzdelanie, národnosť, región.

V situáciách kedy riešiteľ skúma štatistický problém, ktorý je úplne nový (napr. účinky nového lieku na vybranú skupinu pacientov), na získanie potrebných údajov môže využiť **experimentálnu metódu** (typická hlavne pre spoločenské vedy). Tá umožní stanoviť a vytvoriť nevyhnutné podmienky realizácie experimentu. Typické je kontrolované prostredie realizácie experimentu (na rozdiel od skúmania bežných štatistických problémov), v ktorom sa vopred stanovené štatistické znaky sledujú. Prostredníctvom metód štatistického zisťovania (viď kap. 5.2) sa potrebné údaje (obmeny a hodnoty štatistických znakov) zhromaždia a následne podrobia ďalšej analýze. V podobnej podstate sa realizujú experimenty aj v simulovanom prostredí.

5.2 Metódy získavania údajov

Základnými metódami získavania údajov sú **dopytovanie, pozorovanie a meranie**. Tieto metódy sú vo väčšine **primárnym zdrojom** získaných údajov (údaje sú získavané priamo od štatistických jednotiek).

Na určité účely štatistického skúmania je možné použiť aj tzv. **sekundárne zdroje** (v podobe **databáz**, kombinácie údajov dostupných na internete alebo sprostredkovaných informácii z iných zdrojov). Vo výskumnej praxi sa niekedy používajú aj **kombinácie** spomínaných metód a spôsobov.

5.2.1 Dopytovanie

V rámci dopytovania sa rozlišujú tri dielčie techniky:

- **Výkaz** – je špecifickou technikou dopytovania väčšinou pre štátne štatistické účely (ale aj pre niektoré organizácie); je určený na sledovanie a hodnotenie vybraných ukazovateľov alebo činnosti rôznych subjektov alebo zamestnancov (často ukazovatele výkonnosti); sledované ukazovatele sú určené dopredu a sú pravidelne vykazované k určitému dátumu (príklad: mesačné množstvo vyrobených výrobkov konkrétnymi zamestnancami; množstvo chybových výrobkov; odpracovaný čas zamestnancov, a pod.).
- **Dotazník a anketa** – jeho podstatou sú presne formulované dotazníkové otázky na formulovaný štatistický problém; najčastejšie sa používa na sociálne výskumy; v štátnej štatistike je zamarený na malé podniky, ktoré nemajú vyčleneného pracovníka na vyplňovanie výkazov. Rozdiel medzi dotazníkom a anketou je v podstate v rozsahu a výbere respondentov, pričom pre anketu je typické, že sa snaží o masívnejšiu vzorku (viac odpovedí) bez hlbšej definície danej vzorky.
- **Rozhovor** (interview) – táto technika je používaná pri výberových štatistických zisťovaniach zamarených napríklad na štatistiku domácností. Poznáme:
    - **priamy** – ústny rozhovor (štruktúrovaný; neštruktúrovaný)
    - **nepriamy** – telefonický.

5.2.2 Pozorovanie

Pozorovanie je založené na zisťovaní hodnôt/obmien skúmaných štatistických znakov (často určitého javu) **ľudskými zmyslami.** Nazývame ho aj **prosté pozorovanie. Cieľmi** pozorovania sú zistiť hlavne:

- výskyt sledovaného javu alebo javov v štatistickom súbore,
- trvanie sledovaného javu (ak je prítomný),
- frekvenciu výskytu sledovaného javu (ak sa môže opakovať).

Vedecké pozorovanie sa môže do určitej miery od bežného pozorovania líšiť v účele využitia, obsahu, priebehu a výsledkoch. Základnými vlastnosťami vedeckého pozorovania sú:

- **plánovitosť** – predmet, štruktúra a čas pozorovania sú stanovené vopred, spôsob pozorovania je presne určený, vyskúšaný a nacvičený,

- **systematickosť** – pozorovanie sa neuskutočňuje živelne, ale organizovane, v určitom čase alebo intervale, vo väčšine prípadoch v známom prostredí,
- **objektívnosť** – pozorovanie je čo najmenej ovplyvnené subjektívnymi pocitmi a názormi výskumníka, náhodnosťou a nepresnosťou, typická je eliminácia známych rušivých vplyvov.

Nie každé prostredie poskytuje ideálne možnosti na pozorovanie. Ľahšie sa pozoruje v uzavretých alebo ohraničených priestoroch. Ľahšie sa pozorujú statické subjekty alebo objekty ako tie pohyblivé, a ľahšie sa pozorujú subjekty v obmedzenom počte výskytu súčasne. Príkladom môže byť štatistický problém *„Pozorovanie trávenia času na Vodnom diele v Žiline"* v porovnaní so štatistickým problémom *„Pozorovanie správania sa študentov počas prednášok na konkrétnom predmete"*. Je zrejmé, že v prvom prípade je ťažké ohraničiť množstvo sledovaných subjektov, ich spôsoby trávenia voľného času (beh, bicyklovanie, korčuľovanie, a pod.) a ich dynamiku, čo môže zásadne komplikovať prácu pre pozorovateľa. V tomto zmysle, je prípravná fáza pre túto metódu štatistického zisťovania taktiež veľmi dôležitá.

Realita môže byť teda zložitá, skladá sa z **mnohých prvkov** a viacerých rovín. Paralelne prebiehajú **viaceré činnosti**, prelínajú sa a vzájomne sa ovplyvňujú, zúčastňujú sa na nich obyčajne **viacerí aktéri**, ktorí do nej vstupujú a vystupujú. Je preto veľmi ťažké pozorovať realitu vcelku, **komplexne**.

Aby sme dosiahli stanovený cieľ pozorovania, je potrebné pozorovanie zamerať užšie, koncentrovať sa na **menší výsek z reality**, alebo na jej **segmenty** a tie **sledovať hlbšie**, podrobnejšie a presnejšie. Je potrebné presne vymedziť štatistické znaky, na ktoré sa pozorovanie sústredí. Musia to byť javy, ktoré je možné objektívne zachytiť, registrovať a numericky vyjadriť. Pozorovanie, ktoré má tieto vlastnosti, nazývame **štruktúrované**. V opačnom prípade hovoríme o **neštruktúrovanom** pozorovaní.

Štruktúrované pozorovanie je teda založené na tom, že sa realita rozdelí na menšie javy (základné prejavy), ktoré sa presne pozorujú, zaznamenávajú a vyhodnocujú. Základom sú **pozorované kategórie,** pod ktorými sa rozumejú **javy rovnakých vlastností**.

Príklad:

Výskumník pozoruje pacientov po užití experimentálneho lieku proti konkrétnej chorobe. Stanoví si kategórie, ktoré potrebuje pozorovať a robí si záznamy o výskyte týchto kategórií, frekvencii ich opakovania, trvaní a pod. Napr. *výskyt symptómov a konkrétnych nežiadúcich účinkov, ich intenzita, ich trvanie, zmeny stavu pacienta (zlepšenie, zhoršenie), biologické funkcie pacienta a pod.* Pozorovanie sa často kombinuje s inými metódami získavania údajov ako je dopytovanie (výskumník sa môže opýtať napr. na *subjektívne pocity pacienta* a pod.) alebo meranie (napr. *meranie telesnej teploty, tlaku pacienta*).

Vplyv pozorovateľa na pozorované osoby

S výnimkou prípadov, keď je pozorovateľ nezbadaný (napr. vo verejnom priestore), svojou prítomnosťou vždy ovplyvňuje pozorovanú skutočnosť. Pozorované osoby **si uvedomujú, že**

sú pozorované a viac alebo menej sa odchyľujú od prirodzeného správania sa. Ak je napr. v škole na vyučovacej hodine niekto na pozorovaní, tak je možné, že niektorí žiaci (v domnení, že je sústredenie práve na nich) budú svoje nevhodné správanie obmedzovať, no niekedy môže nastať aj opačná situácia, kde dochádza k istej „hereckej produkcii“, vplyvom ktorej je pozorovaná situácia veľmi skreslená (v negatívnom zmysle).

V takýchto situáciách je žiadúce redukovať (minimalizovať) domnienky plynúce z nevedomosti a objasniť dôvod prítomnosti pozorovateľa, prípadne uistiť pozorované osoby, že nejde o hodnotenie, ani skúšanie, ani inšpekciu, že anonymita je zabezpečená, že nie sú do tohto procesu zainteresované ďalšie osoby, prípadne vykonať pozorovanie viacnásobne alebo nepriamo.

Bežne sa pozorovanie vykonáva priamo, kedy pozorovateľ sleduje priebeh činností osobne na mieste – **priame pozorovanie.** V určitých prípadoch je vhodnejšie použiť **nepriame pozorovanie** (zo záznamu alebo pozorovacími prostriedkami).

Výhody nepriameho pozorovania:

1. nahrávku možno opakovane prehrávať, a dôkladne sa oboznámiť s obsahom,
2. kontrola zaznamenaných hodnôt,
3. z nahrávky možno vystrihnúť len situácie, ktoré výskumníka zaujímajú,
4. dej možno zastaviť a podrobne analyzovať konkrétny obraz (situáciu),
5. nahrávku môže používať niekoľko výskumníkov, ktorí ju môžu použiť na rôzne výskumné ciele,
6. porovnanie pozorovania nezávislými pozorovateľmi,
7. digitálne záznamy umožňujú pružné spracovanie na počítači (automatizácia),
8. nahrávku možno použiť na nácvik pozorovania, na osvojenie si práce s kategóriami pozorovania, a pod.

5.2.3 Meranie

Meranie je založené na získavaní dát pomocou rôznych meradiel, nástrojov a pod. Uskutočňuje sa pri rôznych skúškach materiálov, výrobkov, ale aj pri meraní osobných či športových výkonov (napr. bežecké výkony) a pod. Meranie (zaznamenávanie hodnôt) môže byť priame alebo je tento proces automatizovaný pomocou rôznych čidiel a senzorov. Príkladom môžu byť rôzne laboratórne merania (chemické experimenty, protipožiarne skúšky, odolnosť bezpečnostných systémov, a pod.)

V tomto smere je meranie možné kombinovať s experimentálnou metódou, kedy realizátor experimentu meria definované veličiny a zaznamenáva ich hodnoty v priebehu času a po ich analýze sa snaží prijímať nové závery.

5.3 Databázy a sekundárne zdroje údajov

Databáza je množina štruktúrovaných údajov alebo informácií. Slúži na uloženie informácií takým spôsobom, že počítačový program alebo človek môže použiť špeciálny jazyk na ďalšie

spracovanie a analyzovanie týchto údajov. Vďaka presne určenej štruktúre umožňuje ľahké vyhľadávanie a triedenie údajov aj pri ich veľkom množstve.

Pre databázu sa používajú aj iné názvy: **báza údajov, báza dát, dátová báza**; zriedkavo: **databanka, banka dát, banka údajov.**

Databázy môžu vznikať ako výsledok dlhodobého procesu získavania údajov pomocou výkazov, ktoré majú predom definovanú štruktúru a obsah. Vykazované údaje sa ukladajú do databázy a následne sa môžu použiť na vyhodnotenie sledovaných štatistických znakov.

Databáza môže predstavovať primárny zdroj údajov, ale súčasne môže byť aj sekundárnym zdrojom údajov. Pre užívateľa databázy, ktorý ju súčasne vytvára (napr. určitá podniková databáza) budú získané údaje prvotné, pretože sú získané a spracované týmto užívateľom alebo často aj skupinou užívateľov. V inom prípade, ak má riešiteľ štatistického problému k dispozícii napr. databázu štatistického úradu a chce využiť údaje, ktoré sa tam nachádzajú, pre daného riešiteľa budú tieto údaje predstavovať sekundárny zdroj informácii (riešiteľ nepoužil primárne metódy štatistického zisťovania).

Sekundárne zdroje údajov predstavujú okrem databáz aj rôzne súhrnné správy, štatistické ročenky a pod. Takéto zdroje môžu obsahovať čiastočné štatistické zhrnutia a závery (napr. o nehodovosti v jednotlivých okresoch na Slovensku). Sekundárne zdroje môžu mať taktiež podobu údajov o konkrétnych štatistických jednotkách alebo sledovaných štatistických znakoch – v listinnej forme alebo elektronickej podobe. Príkladom sú správy z konkrétnych zásahov záchranných zložiek, ktoré majú predom určenú štruktúru a obsahovú náplň sledovaných štatistických znakov. Tie je možné následne analyzovať ako štatistický súbor zásahov konkrétnej záchrannej zložky alebo viacerých záchranných zložiek.

5.4 Spracovanie štatistických údajov na ďalšiu analýzu

Vo väčšine prípadov, je potrebné získané údaje ešte ďalej spracovať. Štatistické spracovanie (údajov) je práca so získanými štatistickými údajmi a **prevedenie ich do takej podoby**, v ktorej je možné ich ďalšie analyzovanie. Je to spôsobené požiadavkami softvérov (hlavne štatistických), v ktorých sa získané údaje podrobujú analýze (použitie štatistických metód a štatistických nástrojov).

Bežným príkladom je potreba spracovania získaných údajov od respondentov. V závislosti na technike a forme získavania údajov (priamy rozhovor, telefonický rozhovor, písomný dotazník, online forma) môžu mať získané údaje rôznu podobu, pričom majú rovnakú výpovednú hodnotu. Pre rýchlejšie a efektívnejšie vyhodnotenie (analyzovanie) je však potrebné, aby mali odpovede s rovnakou výpovednou hodnotou, aj rovnakú formu. Veľmi jednoduchý príklad je uvedený v nasledujúcej tabuľke (Tabuľka 3). Ako vidno z tabuľky, spracovanie údajov bude v tomto príklade znamenať, že napr. v rámci štatistického znaku „pohlavie“ je potrebné zjednotiť vyjadrenie jeho obmien. Momentálne sú v danej tabuľke uvedené tri vyjadrenia pre obmenu „mužské pohlavie“ (M, Muž, Chlap) a dve vyjadrenia obmeny „ženské pohlavie“ (Ž, Žena). Tieto obmeny je nutné na ďalšiu analýzu upraviť

do jednotnej podoby napr. pod skratkou „M“ a „Ž“. Podobne to platí aj pre štatistické znaky kraj a vek.

Tabuľka 3 Vstupné údaje (hrubé dáta získané od respondentov)

Respondent	**Pohlavie**	**Kraj**	**Vek [rok]**	**...**
1	M	Trnavský	20	...
2	Žena	BA	32 rokov	...
3	Ž	Bratislava	31	...
4	Muž	Košický	30	...
5	Chlap	Bratislavský kraj	54 r.	...
...	...	...	...	...

Tieto základné problémy vznikajú hlavne v otvorených otázkach v dotazníku. Dajú sa efektívne odstrániť pri online formách dotazníkov a vhodne zvolenými otázkami s priradenými možnosťami odpovede. Otvorená dotazníková otázka môže znieť: *Aké sú hlavné dôvody Vašej nespokojnosti s pracovným prostredím?* alebo *Vyjadrite prosím Váš postoj k vakcinácii.* Prvá otázka predpokladá viacero možností, pričom nie všetky sú riešiteľovi dopredu známe, ale je pravdepodobné, že sa budú opakovať. Druhá dotazníková otázka je natoľko otvorená, že riešiteľ bude mať veľmi ťažkú úlohu spracovať a analyzovať odpovede získané od respondentov. Vhodne zvoliť dotazníkové otázky je preto veľmi dôležitou súčasťou prípravy prieskumu, uľahčí následný zber údajov, ich spracovanie a analýzu.

Literatúra

BENČO, J. *Metodológia vedeckého výskumu*. Bratislava: IRIS, 2001. ISBN 80-9018-27-0.

COCHRAN, W. G. *Sampling Techniques*. J. Wiley and Sons, NewYork, 1972

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

CHAJDIAK, J. *Analýza dotazníkových údajov*. Bratislava: Statis, 2013. ISBN 978-80-85659-76-4.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

PECÁKOVÁ, I. *Statistika v terénních průzkumech*. 2018. 3.vyd. Professional Publishing, Praha. ISBN 978-80-88260-10-3.

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

TIRPÁKOVÁ, A., MARKECHOVÁ, D. *Štatistika v praxi*. Nitra: FPV UKF, 2008, ISBN 978-80-8094-283-0.

6 Triedenie štatistických údajov

Triedenie predstavuje:

- **štatistickú metódu** spracovania údajov o hromadných javoch a procesoch,
- **usporiadanie údajov do skupín** (tried),
- usporadúvanie štatistického súboru **podľa obmien určeného štatistického znaku/znakov (triediaci znak/y)** – logické usporiadanie štatistických jednotiek podľa určených kritérií,
- **určovanie početnosti** (frekvencie) výskytu hodnôt/obmien znaku v súbore – zisťovanie **rozdelenia početnosti**,
- rozdelenie štatistického súboru na určitý počet čiastkových súborov, ktoré vykazujú rovnaké alebo podobné vlastnosti.

Typy triedenia:

- **jednostupňové triedenie** – používa sa tiež **triedenie podľa jedného štatistického znaku –** ak sa pri triedení používa iba jeden triediaci znak (napr. výška študentov) – **jednoduché a skupinové triedenie,**
- **viacstupňové triedenie** – ak sa štatistický súbor triedi súčasne podľa dvoch alebo viacerých štatistických znakov (napr. výška a pohlavie študentov) – **hierarchické, kombinačné).**

Výsledkom všetkých druhov triedenia **je rozdelenie početností** v tabuľkovej a grafickej podobe.

Požiadavky a zásady triedenia:

- každá štatistická jednotka musí byť zaradená do niektorej z vytvorených skupín, resp. aspoň do skupiny „iné“, to znamená, že triedenie musí byt úplné – **zásada úplnosti**,
- triedené (vytvorené) skupiny by mali vyjadrovať podstatné vlastnosti skúmaného súboru, javu, ...,
- triedené (vytvorené) skupiny jednotiek by sa mali vzájomne vylučovať – aby bolo jednoznačné, do ktorej skupiny každú štatistickú jednotku zariadiť – **zásada jednoznačnosti**.

Použité pojmy:

- **triedenie** = rozdelenie štatistických jednotiek do takých skupín (tried), aby čo najlepšie vynikli charakteristické vlastnosti skúmaných javov,
- **triediaci znak/y** = štatistický znak/y, ktorý/é je/sú kritériom triedenia štatistického súboru,
- **trieda** = skupina štatistických jednotiek s rovnakou hodnotou (obmenou) štatistického znaku.

6.1 Jednostupňové triedenie – triedenie podľa jedného štatistického znaku

Základom jednostupňového štatistického triedenia je **usporiadanie hodnôt/obmien štatistického znaku do tried**. Z tohto pohľadu sa pre jednostupňové triedenie rozlišuje:

- rad **neusporiadaný** (pôvodný zistený rad): $x_1, x_2,..., x_i,..., x_n$; index *i* súvisí s poradím zisťovania,
- rad **usporiadaný** (variačný) podľa veľkosti (platí pre číselné znaky): $x_{(1)}, x_{(2)},..., x_{(i)},..., x_{(n)}$, index (*i*) súvisí s veľkosťou hodnôt, pričom $x_{(1)} \leq x_{(2)},... \leq x_{(i)},... \leq x_{(n)}$ pričom $x_{(1)} = x_{min}$, $x_{(n)} = x_{max}$.
- rad **triedený**,
    - jednoduché (prosté) triedenie,
    - skupinové (intervalové) triedenie.

Jednostupňové triedenie predstavuje teda rad triedený.

Výsledky jednostupňového triedenia predstavujú primárne vyjadrenia rôznych typov početností:

- príslušný počet výskytov v súbore **– absolútna početnosť n_i**
- podiel na celkovom rozsahu súboru **– relatívna početnosť p_i**
- súčtový počet od prvej po poslednú triedu **– kumulatívna absolútna početnosť kn_i**
- súčtový podiel od prvej po poslednú triedu **– kumulatívna relatívna početnosť kp_i**

Výsledky triedenia sa prezentujú:

- v tabuľkovej podobe – tabuľky rozdelenia početností,
- v grafickej podobe – stĺpcové a výsekové grafy, polygóny rozdelenia početností.

6.1.1 Jednoduché triedenie

Prosté (jednoduché) triedenie je triedenie realizované podľa každej hodnoty (obmeny) štatistického znaku samostatne. Výhodné je v situácií, keď štatistický znak dosahuje iba **obmedzený počet** hodnôt/obmien.

Typické pre:

- **slovné** (kvalitatívne) štatistické znaky **alternatívne aj množné** (napr. rozdelenie študentov podľa študijného programu, ktorý študujú),
- **číselné** (kvantitatívne) štatistické znaky **s malým počtom obmien** (číselné znaky – **do 15 obmien)** – väčšinou sú to diskrétne (nespojité) štatistické znaky napr. *rozdelenie rodín podľa počtu detí v rodine,* ale v špecifických prípadoch aj spojité štatistické znaky, ktoré vykazujú malý počet hodnôt (malý rozsah štatistického súboru).

Triedenie kvalitatívnych (slovných) znakov

Triedenie **kvalitatívnych (slovných)** znakov sa uskutočňuje podľa **obmien** štatistického znaku. Poradie obmien sa volí **prvotne** podľa nasledujúcich možností:

- obmeny je možné zoradiť podľa významu (napr. podľa úrovne vzdelania)
- obmeny je možné vystupňovať (napr. hodnotenie študentov)
- obmeny zoradíme podľa abecedy,
- obmeny zoradíme náhodne (farby áut),
- obmeny zoradíme náhodne podľa subjektívneho názoru riešiteľa.

Obmeny je možné zoradiť **druhotne** podľa výslednej absolútnej početnosti zostupne alebo vzostupne.

Výsledkom triedenia sú početnosti výskytu daného štatistického znaku, pričom sa jednotlivým obmenám **priradí:**

- príslušný počet výskytov v súbore **– absolútna početnosť n_i**
- podiel na celkovom rozsahu súboru **– relatívna početnosť p_i**
- súčtový počet od prvej po poslednú triedu **– kumulatívna absolútna početnosť kn_i**
- súčtový podiel od prvej po poslednú triedu **– kumulatívna relatívna početnosť kp_i**

Jednoduché triedenie sa prezentuje:

- v tabuľkovej podobe **–** tabuľky rozdelenia početností (tabuľka jednoduchého triedenia),
- v grafickej podobe **–** stĺpcové a výsekové grafy, polygóny rozdelenia početností.

Tabuľka jednoduchého triedenia má štandardnú štruktúru a obsahuje všetky vyššie menované druhy početností. Tabuľka sa doplňuje **súčtovým riadkom**, ktorý slúži na krížovú kontrolu správnosti triedenia. Všeobecný postup tvorby tabuľkovej formy jednoduchého (prostého) triedenia je uvedený v nasledujúcej prehľadnej tabuľke (Tabuľka 4).

Tabuľka 4 Tabuľka jednoduchého triedenia

Trieda	**Triediaci znak**	**Absolútna početnosť**	**Relatívna početnosť**	**Kumulatívna absolútna početnosť**	**Kumulatívna relatívna početnosť**
$\boldsymbol{k}$	$\boldsymbol{x_i}$	$\boldsymbol{n_i}$	$\boldsymbol{p_i}$	$\boldsymbol{kn_i}$	$\boldsymbol{kp_i}$
1	x_1	n_1	p_1	kn_1	kp_1
2	x_2	n_2	p_2	kn_2	kp_2
...	...	...	...	...	...
i	x_i	n_i	$p_i = \frac{n_i}{n}$	$kn_i = \sum_{j=1}^{i} n_j$	$kp_i = \sum_{j=1}^{i} p_j$
...	...	...	...	...	...
k	x_k	n_k	p_k	$kn_k = \sum_{j=1}^{k} n_j = n$	$kp_k = \sum_{j=1}^{k} p_j = 1$
Súčet $\sum$		$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$	×	×

Absolútna početnosť n_i je číslo, ktoré určuje koľko štatistických jednotiek v štatistickom súbore má určitú hodnotu pričom platí:

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

Relatívna početnosť p_i je podiel absolútnej početnosti n_i a rozsahu súboru n

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

pričom platí:

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

alebo alternatívne v percentách (preferovaná voľba pre štatistické projekty):

$$\sum_{i=1}^{k} 100p_i = 100(\%)$$

Kumulatívna (súčtová) absolútna početnosť kn_i udáva postupný súčet početností od 1. triedy až po danú triedu:

$$kn_i = \sum_{j=1}^{i} n_j$$

Kumulatívna (súčtová) relatívna početnosť kp_i udáva postupný podiel početností od 1. triedy až po danú triedu:

$$kp_i = \sum_{j=1}^{i} p_j$$

Alternatívne je možné opäť vyjadrenie v %.

Tabuľka 5 Triedenie študentov podľa záverečného hodnotenia predmetu (klasifikácia)

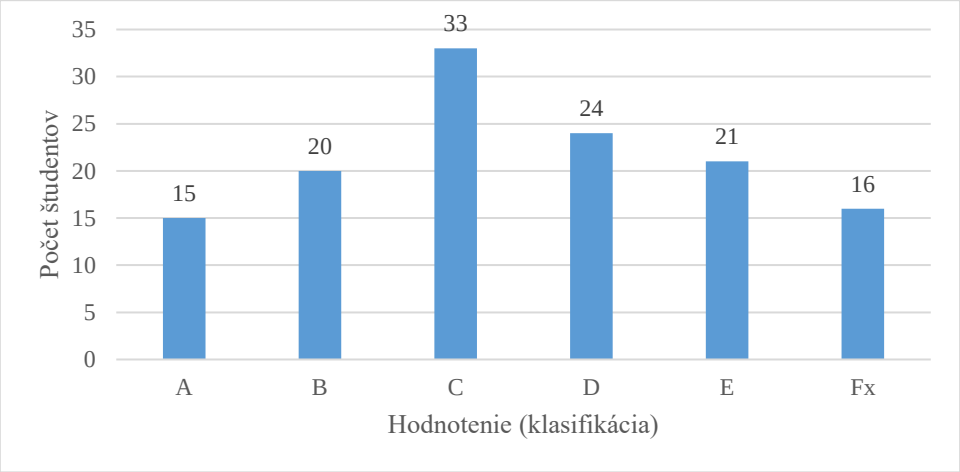
Trieda	Triediaci znak	Absolútna početnosť	Relatívna početnosť	Kumulatívna absolútna početnosť	Kumulatívna relatívna početnosť
k	x_i	n_i	p_i	kn_i	kp_i
Poradové číslo	Hodnotenie (klasifikácia)	Počet študentov	Podiel študentov [%]	Súčtový počet študentov	Súčtový podiel študentov [%]
1	A	15	11,63	15	11,63
2	B	20	15,50	35	27,13
3	C	33	25,58	68	52,71
4	D	24	18,60	92	71,32
5	E	21	16,28	113	87,60
6	Fx	16	12,40	**129**	**100,00**
Spolu ∑	x	**129**	**100,0**	x	x

Tabuľková forma jednoduchého triedenia kvalitatívnych (slovných) znakov sa doplňuje predovšetkým:

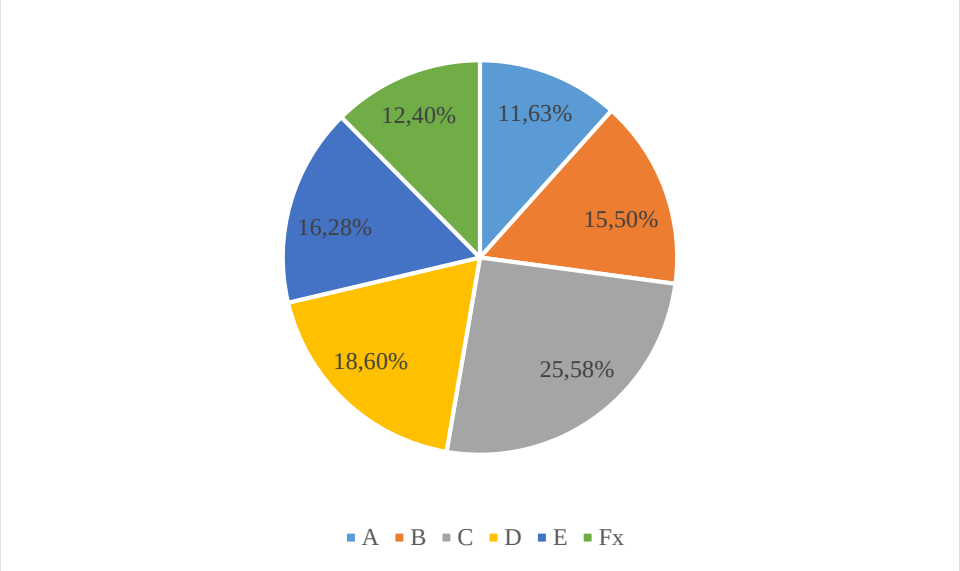
- **stĺpcovým grafom** absolútnych početností – stĺpcové grafy poskytujú jednoduchý a zrozumiteľný spôsob zobrazovania nominálnych a poradových údajov, ktoré sa zaradzujú do tried; početnosť triedy sa zobrazuje ako plocha stĺpca zostrojeného nad príslušným intervalom (triedou) (Obrázok 5),

- **koláčovým grafom** relatívnych početností vyjadrených v % – kruhový výsekový diagram rozdelí kruh na viac častí podľa počtu tried; početnosť triedy je vyjadrená veľkosťou plochy kruhového výseku (Obrázok 6).

Grafy môžu byť spracované aj v podobe nepravých trojrozmerných grafov. Jednotlivé stĺpce alebo výseky je vhodné doplniť konkrétnou hodnotou.

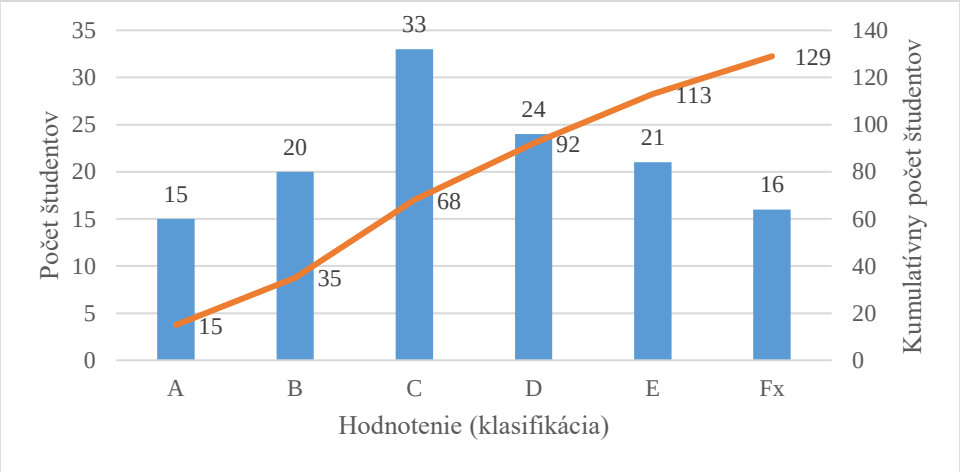


Obrázok 5 Počty študentov podľa dosiahnutej klasifikácie



Obrázok 6 Podiel študentov podľa dosiahnutej klasifikácie

V prípade potreby je možné zobraziť aj **kumulatívne početnosti** napr. na vedľajšej osi stĺpcového grafu (Obrázok 7).



Obrázok 7 Absolútne a kumulatívne početnosti študentov podľa dosiahnutej klasifikácie

Triedenie kvantitatívnych (číselných) štatistických znakov

V prípade, že **kvantitatívny** (číselný) štatistický znak dosahuje málo rôznych hodnôt (počet tried $k < 15$), na zistenie frekvencie výskytu daných hodnôt v štatistickom súbore sa taktiež používa jednoduché triedenie. Bežnejšie spĺňajú túto podmienku diskrétne (nespojité) znaky.

Podobne ako pri slovných štatistických znakoch, číselný štatistický znak s malým počtom hodnôt sa triedi **podľa každej hodnoty znaku** x_i, pričom hodnoty štatistického znaku v tabuľke jednoduchého triedenia sa uvádzajú vo vzostupnom poradí (Tabuľka 6). Hodnoty je možné zoradiť druhotne podľa absolútnej početnosti zostupne alebo vzostupne.

Tabuľka 6 Triedenie rodín podľa počtu detí

Trieda	Triediaci znak	Absolútna početnosť	Relatívna početnosť	Kumulatívna absolútna početnosť	Kumulatívna relatívna početnosť
k	x_i	n_i	p_i	kn_i	kp_i
Poradové číslo	Počet detí	Počet rodín	Podiel rodín	Súčtový počet rodín	Súčtový podiel rodín
1	1	460	46,00	460	46,00
2	2	404	40,40	864	86,40
3	3	101	10,10	965	96,50
4	4	30	3,00	995	99,50
5	5 a viac	5	0,50	1000	100,00
Spolu ∑	x	**1000**	**100,00**	x	x

Spôsob výpočtu a zobrazovania výsledkov ostáva rovnaký ako pre slovné znaky. Opodstatnenosť využitia kumulatívnych početností je pri číselných znakoch vyššia, keďže na rozdiel od slovných znakov (okrem poradových) je usporiadanie hodnôt číselných znakov v zásade vzostupne alebo zostupne a tak jednotlivé triedy na seba nadväzujú.

Tabuľková forma sa opäť doplňuje väčšinou:

- stĺpcovým grafom absolútnych početností alebo,
- koláčovým grafom relatívnych početností vyjadrených v %.

6.1.2 Skupinové triedenie

Skupinové (intervalové) triedenie je rozdelenie štatistických jednotiek podľa hodnôt štatistického znaku zhrnutých do spoločnej triedy (skupiny, **intervalu**) tak, aby čo najlepšie vynikli charakteristické vlastnosti skúmaných javov.

Skupinové (intervalové) triedenie sa používa v prípade, že **číselné štatistické znaky** (spojité i nespojité) dosahujú **veľké množstvo rôznych hodnôt** (nad 15).

Pôvodné (získané) **údaje sa zaraďujú do intervalov** (tried, skupín) a **určujú sa početnosti jednotlivých tried**, čím sa vytvorí rozdelenie početností. Skupinové triedenie spočíva vo vytvorení ***k*** tried (skupín, intervalov) vo variačnom rozpätí ***R*** súboru od minimálnej x_{min} až po maximálnu hodnotu znaku x_{max}. Výsledkom triedenia sú teda početnosti výskytu hodnôt triediaceho znaku z intervalu hodnôt, pričom sa intervalom (triedam) priradí podobne ako pri jednoduchom triedení**:**

- príslušný počet výskytov v súbore – **absolútna početnosť** n_i
- podiel na celkovom rozsahu súboru – **relatívna početnosť** p_i
- súčtový počet od prvej po poslednú triedu – **kumulatívna absolútna početnosť** kn_i
- súčtový podiel od prvej po poslednú triedu – **kumulatívna relatívna početnosť** kp_i

Tvorba intervalov a výpočet početností výskytu hodnôt štatistického znaku v daných intervaloch pozostáva z týchto základných krokov:

1. Určenie **počtu tried** ***k*** – počet tried sa volí intuitívne v rozpätí 6-15 alebo sa vypočíta podľa **Sturgersovho pravidla** – vzorca $k = 1 + 3{,}322 * \log(n)$, kde sa vypočítané číslo **zaokrúhľuje hore**; *n* predstavuje rozsah skúmaného štatistického súboru.
2. Výpočet **variačného rozpätia** ***R*** – výpočet predstavuje rozdiel medzi najväčšou a najmenšou hodnotou variačného radu $R = x_{max} - x_{min}$**.**
3. Výpočet **šírky triedy h –** delenie variačného rozpätie počtom tried $h = R / k$**.** Výsledok sa zaokrúhli podľa pravidiel zaokrúhľovania.
4. Vytvorenie **intervalov hodnôt** – priradenie **dolnej a hornej hranice** jednotlivým triedam. Dolná hranica prvej triedy x_d bude rovná x_{min}. Horná hranica poslednej triedy x_h bude rovná x_{max}. Dbáme na "Nesporné vymedzenie" hraníc jednotlivých tried.
5. **Zaradenie jednotlivých hodnôt** štatistického znaku do príslušného intervalu – riešiteľ dostáva absolútne početnosti (frekvencie) v jednotlivých triedach (odporúča sa používať nástroj „histrogram“ v exceli, ale je možné využiť aj iné nástroje).
6. Výpočet **relatívnych početností** p_i a **kumulatívnych početností** absolútnych kn_i a relatívnych kp_i.
7. Výpočet **stredov tried** x_i $(i = 1, 2, ..., k)$.

Zásady platné pre skupinové triedenie:

- triedy majú **konštantnú šírku** (výnimku tvoria otvorené triedy kvôli výskytu extrémov prípadne posledné horné hranice ak sa šírka triedy musela zaokrúhľovať),
- počet tried ***k*** musí byť v rozmedzí **6 až 15**, počet intervalov, nemá byť ani príliš malý (vedie k hrubému, zjednodušenému pohľadu), ani príliš veľký (robí triedenie neprehľadným),
- všetky hodnoty štatistického znaku, ktoré boli zaradené do príslušnej triedy sa nahradzujú tzv. reprezentatívnou hodnotou, za ktorú sa väčšinou volí stred intervalu x_i, tzv. hodnota triedneho znaku,
- **šírka *h*, dolná hranicu x_d, horná hranicu x_h** sa volí s ohľadom na maximálnu prehľadnosť,
- hranice tried musia mať nesporné (jednoznačné) vymedzenie.

Výsledky sa zapisujú do **tabuľky početností,** ktorá oproti tabuľke jednoduchého triedenia obsahuje naviac prvky (stĺpce) vymedzujúce práve intervaly a to vo forme samotných intervalov, napr. (20;25>, prípadne sa uvádzajú samostatne **hranice intervalov (dolné x_d a horné x_h)**; obmeny triediaceho znaku nahrádzajú **stredy intervalov** (Tabuľka 7).

Tabuľka 7 Základné prvky tabuľky skupinového triedenia

Trieda	**Intervaly hodnôt**		**Stred triedy (triedny znak)**	Absolútna početnosť	Relatívna početnosť	Kumulatívna početnosť	
	dolné hranice tried	**horné hranice tried**				Absolútna	Relatívna
k	x_{di}	x_{hi}	x_i	n_i	p_i	kn_i	kp_i
1	x_{d1}	x_{h1}	x_1	n_1	p_1	kn_1	kp_1
2	x_{d2}	x_{h2}	x_2	n_2	p_2	kn_2	kp_2
...	...	...	...	...	...	...	...
i	x_{di}	x_{hi}	x_i	n_i	$p_i = \frac{n_i}{n}$	$kn_i = \sum_{j=1}^{i} n_j$	$kp_i = \sum_{j=1}^{i} p_j$
...	...	...	...	...	...	...	...
k	x_{dk}	x_{hk}	x_k	n_k	p_k	$kn_k = \sum_{j=1}^{k} n_j = n$	$kp_k = \sum_{j=1}^{k} p_j = 1$
∑	×	×	×	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} p_i = 1$	×	×

Absolútna početnosť n_i v rámci skupinového (intervalového) triedenia je číslo, ktoré určuje koľko jednotiek súboru má hodnotu, ktorá spadá do stanoveného rozpätia príslušnej triedy.

Relatívna početnosť p_i v rámci skupinového (intervalového) triedenia je taktiež ako u jednoduchého triedenia podiel absolútnej početnosti n_i a rozsahu súboru n.

Kumulatívna početnosť (**absolútna, relatívna**) kn_i v rámci skupinového (intervalového) triedenia poskytuje informáciu, koľko štatistických jednotiek súboru resp. aká pomerná časť štatistického súboru má hodnotu štatistického znaku menšiu alebo rovnú ako je horná hranica príslušnej triedy (triedny znak).

Na výpočet príslušných početností je možné použiť vzorce využívané v jednoduchom triedení.

Vymedzenie hraníc intervalov môže mať rôznu úpravu, pričom musí byť zachovaná výpovedná hodnota výsledkov. Možné úpravy sú uvedené v nasledujúcich príkladoch (Tabuľka 8).

Tabuľka 8 Príklady vymedzenia intervalov a ich hraníc

Vymedzenie intervalov ich hraníc				
Odporúčané (vhodné pre výpočty v exceli)	Použiteľné pre celočíselné znaky	Individuálne vymedzenie hraníc intervalu		Opačné použitie uzatvorenia intervalov (nie je kompatibilné s nástrojmi v exceli)
		dolné hranice	horné hranice	
<15; 20>	16—20	<15	20>	<15 až 20)
(20; 25>	21—25	(20	25>	<20 až 25)
(25; 30>	26—30	(25	30>	<25 až 30)
(30; 35>	21—35	(30	35>	<30 až 35)
(35; 40>	36—40	(35	40>	<35 až 40)
(40; 45>	41—45	(40	45>	<40 až 45>

Výsledky skupinového triedenia v tabuľkovej forme sa dopĺňajú v grafickej podobe:

- histogramom alebo polygónom absolútnych početností,
- koláčovým grafom relatívnych početností vyjadrených v %,
- polygónom kumulatívnych absolútnych alebo relatívnych početností.

Histogram skupinového (intervalového) triedenie je stĺpcový graf tvorený pravidelnými rovnobežníkmi, ktorých obsah (aj nulový) je úmerný súčtu hodnôt znaku príslušnej triedy. Základy stĺpcov na osi *x* majú dĺžku vypočítaných intervalov (šírky triedy) *h,* a príslušné výšky majú veľkosť zodpovedajúcu početnosti tried. Základy stĺpcov sú zvyčajne bez medzier vzhľadom na povahu zobrazovania intervalov, ktoré majú spoločné hranice.

Polygón absolútnych početností je možné odvodiť z histogramu. Spojuje triedne znaky (stredy tried) jednotlivých tried (intervalov). Polygón začína a končí na vodorovnej súradnicovej osi v strede susedných prázdnych tried.

Polygón kumulatívnych absolútnych alebo relatívnych početností začína na osi x na dolnej hranici prvej triedy a pokračuje ako spojnica horných hraníc jednotlivých tried.

Príklad použitia skupinového triedenia:

Je potrebné roztriediť súbor **60 zamestnancov** konkrétnej firmy podľa mesačného príjmu.

Usporiadaný (variačný) rad: $x_{(1)} = 450$; $x_{(2)} = 480$;...; $x_{(59)} = 1130$; $x_{(60)} = 1150$

Počet tried sa stanoví výpočtom: $k = 1+3{,}322*\log(60) = 6{,}91$; zaokrúhlime na 7 tried

Variačné rozpätie: $x_{min} = 450$ €, $x_{max} = 1150$ €; potom

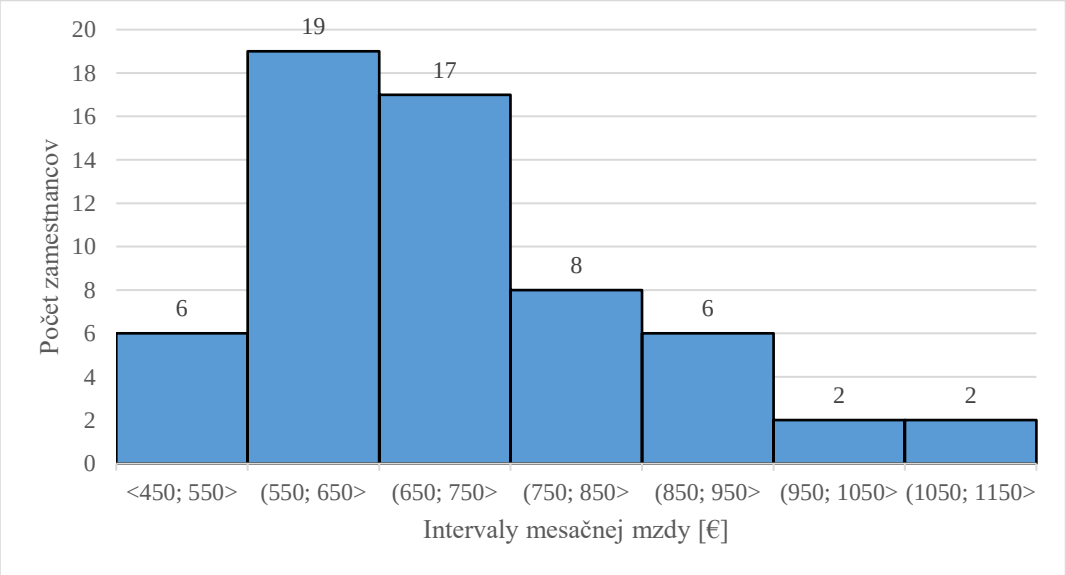
$R = x_{max} - x_{min} = 700$

Šírka triedy $h = (1150 - 450) / 7 = 100$

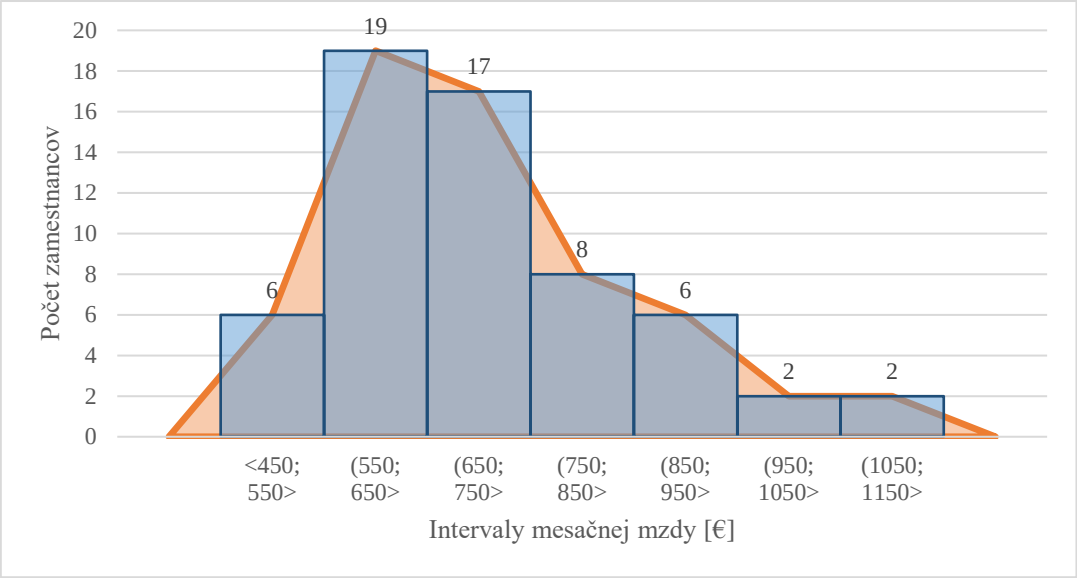
Vytvorené intervaly a početnosť zamestnancov je uvedená v nasledujúcej tabuľke (Tabuľka 9) a grafoch nižšie (Obrázok 8, Obrázok 9, Obrázok 10).

Tabuľka 9 Triedenie zamestnancov podľa výšky mesačného príjmu

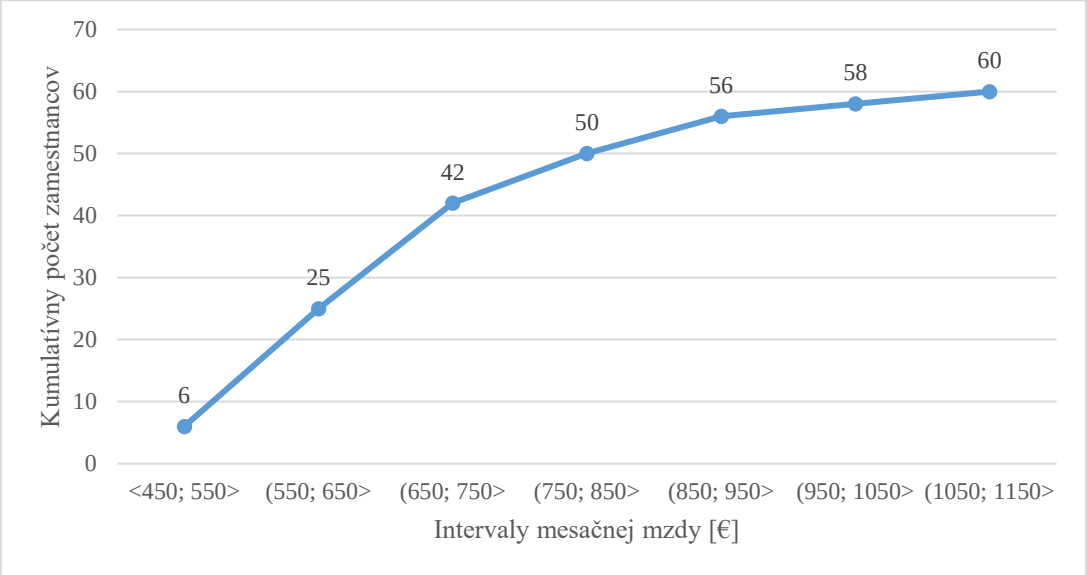
Trieda	Hranice mesačného príjmu [€]	Stred intervalu príjmu [€]	Počet zamestnancov	Podiel zamestnancov [%]	Súčtová početnosť zamestnancov	
					Absolútna	Relatívna [%]
k	$x_d - x_h$	x_i	n_i	p_i	kn_i	kp_i
1	<450; 550>	500	6	10,00	6	10,00
2	(550; 650>	600	19	31,67	25	41,67
3	(650; 750>	700	17	28,33	42	70,00
4	(750; 850>	800	8	13,33	50	83,33
5	(850; 950>	900	6	10,00	56	93,33
6	(950; 1050>	1000	2	3,33	58	96,67
7	(1050; 1150>	1100	2	3,33	60	100,00
∑	✕	✕	60	100	✕	✕



Obrázok 8 Histogram triedenia zamestnancov podľa mesačného príjmu

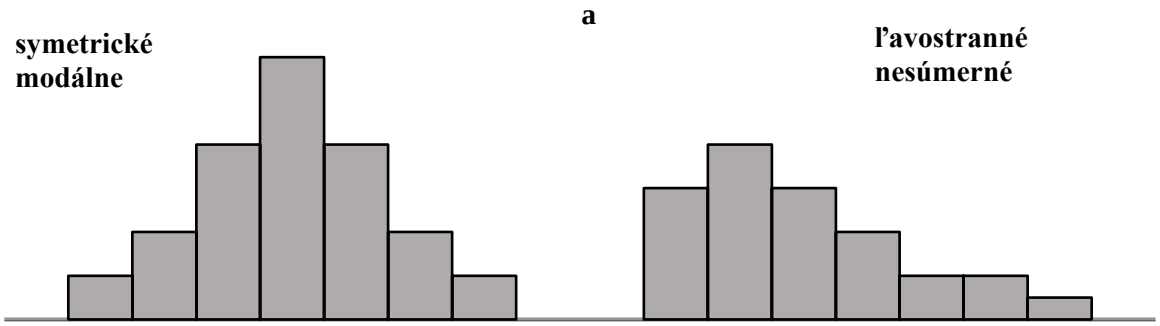


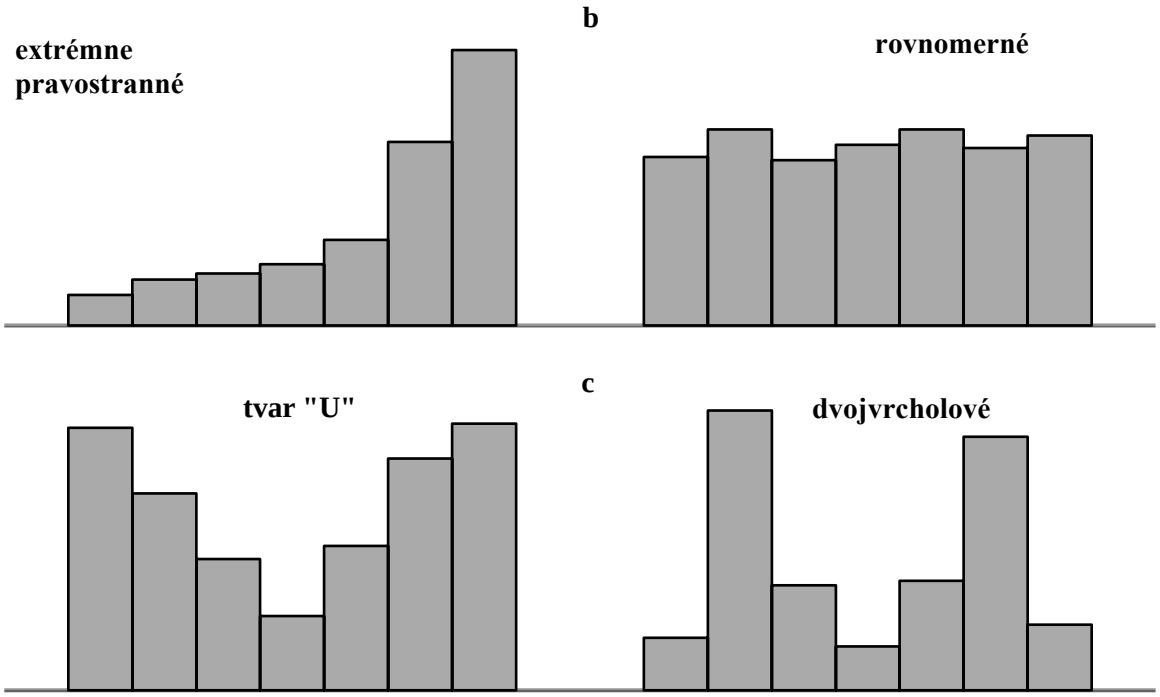
Obrázok 9 Histogram a polygón absolútnych početností zamestnancov podľa mesačného príjmu



Obrázok 10 Kumulatívna početnosť zamestnancov podľa mesačného príjmu

Vzhľadom na povahu skúmaného znaku a jeho hodnôt môžu mať histogramy rôzne typické tvary, príklady sú uvedené nižšie (Obrázok 11a, b, c).





Obrázok 11 Typické tvary histogramov

6.1.3 Extrémy v skupinovom triedení

Pri štatistickom zisťovaní je možné sa stretnúť s prípadmi, keď niektorá hodnota (hodnoty) skúmaného štatistického znaku sa vymyká zo zistených hodnôt smerom nadol alebo nahor (hodnoty v minime alebo maxime). Daná hodnotu alebo hodnoty sa označujú ako extrémne – **extrémy.** Extrémy sa môžu vyskytnúť v rôzne veľkých štatistických súboroch a pre rôzne štatistické číselné znaky (primárne sa však extrémy vyskytujú u znakov s vyšších množstvom rôznych hodnôt). Extrémy je možné identifikovať nie len v rámci skupinového triedenia, ale aj v rámci riešenia iných štatistických úloh. Avšak práve v rámci skupinového triedenia je identifikovanie extrémov relatívne jednoznačné.

Príklady

Ak sú všetci študenti v triede vysokí od 160 do 185 cm, je zrejmé, že študent s výškou 140 alebo 205 bude predstavovať extrém, ktorý môže napr. skresliť priemernú výšku v triede jedným alebo druhým smerom.

Iným príkladom môže byť súbor 1000 mužov, v ktorom sa vyskytol jeden muž, ktorý mal hmotnosť 42 kg a jeden, ktorý mal hmotnosť 200 kg, pričom ďalšie najbližšie hodnoty hmotnosti boli 65 kg a 110 kg.

Usporiadaný rad hmotnosti štatistickej vzorky 1000 mužov vyzerá takto:

$x_{(1)}$=42,0; $x_{(2)}$=65,0;..., $x_{(999)}$=110,0;...; $x_{(1000)}$=200

Na prvý pohľad je možné sa domnievať, že spomínané hodnoty (42 a 200 kg) v podobe lokálnych extrémov by skreslili všetky ďalšie výpočty ohľadom charakteristík úrovne,

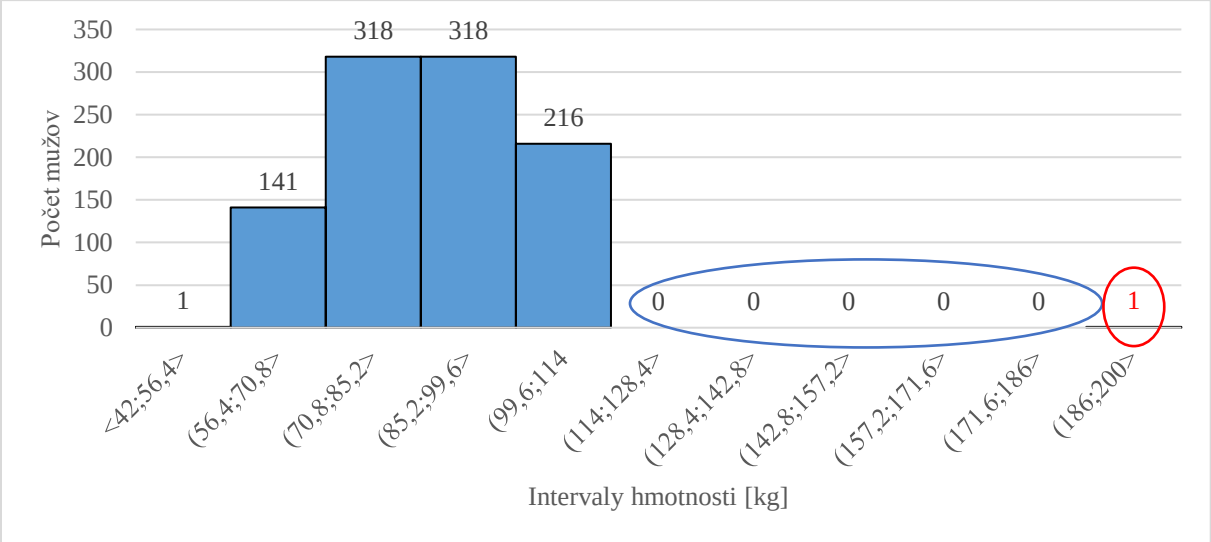
variability, ale aj použitie ďalších metód. Extrém alebo extrémy je však nutné potvrdiť aj výpočtom, nie iba našim odhadom.

Identifikácia extrému

Pri výpočte sa postupuje zo začiatku ako pri skupinovom triedení, až po zistenie absolútnych početností štatistických jednotiek v intervaloch.

Počet tried *(k)*: $k = 1+3{,}322\log(1000) = 11$ (po zaokrúhlení nahor)
Variačné rozpätie: $x_{max} = 200$
$x_{min} = 42$
$x_{max} - x_{min} = 158$
Šírka triedy *(h)*: $h = (x_{max} - x_{min})/k = 14{,}4$ (po zaokrúhlení na 1 desatinné miesto)

Extrémy sa odhalia už po samotnom určení početností, ale najviac evidentné je to v rámci grafického spracovania výsledkov (Obrázok 12).



Obrázok 12 Rozdelenie vzorky mužov podľa hmotnosti (s extrémom)

Nulové intervaly (v modrej elipse) signalizujú, že za nimi bude extrémna hodnota (červeným) u nás je to hodnota 200 kg v intervale (186;200>. Pre určenie extrému by postačoval aj jeden prázdny interval pred extrémnou hodnotou alebo hodnotami. Prázdny interval pred hodnotou alebo hodnotami v maxime (pozor platí aj pre hodnoty v minime) sú teda prvotným predpokladom na určenie extrémov.

Určitú mieru subjektívneho zhodnotenia určenia alebo neurčenia extrémov vzhľadom na skúmanú problematiku je možné uplatniť, ak sa jedná o podstatne vyššie početnosti potenciálnych extrémov. Napríklad by posledný interval vykazoval absolútnu početnosť 7 a pred ním by bol prázdny interval, čo by za bežných okolností stále signalizovalo sedem extrémov. Avšak ak sa súčasne pridá skutočnosť, že nešlo o štatistický súbor veľký 1000 štatistických jednotiek, ale napr. iba 70 štatistických jednotiek, tak je logickejšie prikloniť sa k záveru, že výsledky sú skôr ovplyvnené nevhodným spôsobom výberu štatistických jednotiek do súboru, nereprezentatívnosťou štatistického súboru (čo môže byť stále problém aj pri väčšej vzorke) alebo je možné uvažovať o iných dôvodoch chýbajúcich hodnôt pre nulové intervaly.

Práca s extrémami

Ak riešiteľ objaví extrémne hodnoty, tak použije tento základný postup:

1. pri výpočtoch počtu tried *k* a šírky triedy *h* sa extrémne hodnoty zanedbajú a v tabuľke rozdelenia početnosti sa neuvedie dolná hranica x_d prvej triedy a/alebo horná hranica x_h poslednej triedy – vytvorí sa tzv. **otvorená trieda**,
2. do otvorenej triedy sa k získaným absolútnym početnostiam ešte pripočíta počet odstránených extrémov.

Forma zápisu takto upraveného riešenia (otvorenej triedy) môže byť rôznorodá (viď nasledujúcu tabuľku). Zvolený zápis však vždy hovorí o tom, že v skúmanom štatistickom súbore sa extrém(y) vyskytol/vyskytli. Otvorená dolná (extrém bol v minime) a horná trieda (extrém bol v maxime) sa použije samostatne, ak sa extrém vyskytol iba na jednej strane usporiadaného radu. Ak sa vyskytli extrémy súčasne na oboch stranách usporiadaného radu, tak sa používajú otvorené triedy v prvej aj poslednej triede súčasne. V smere otvorenej triedy sa buď uvedie otvorená hranica alebo sa žiadna zátvorka nepoužije.

Tabuľka 10 Formy zápisu otvorenej triedy

Otvorená trieda v minine				Otvorená trieda v maxime	
do 20>	(do 20>	(menej ako 20>	(20 a menej>	(100 a viac)	(100 a viac

Pre vyššie uvedený príklad ohľadom hmotnosti vyzerajú výpočty po odstránení extrému 200 kg takto:

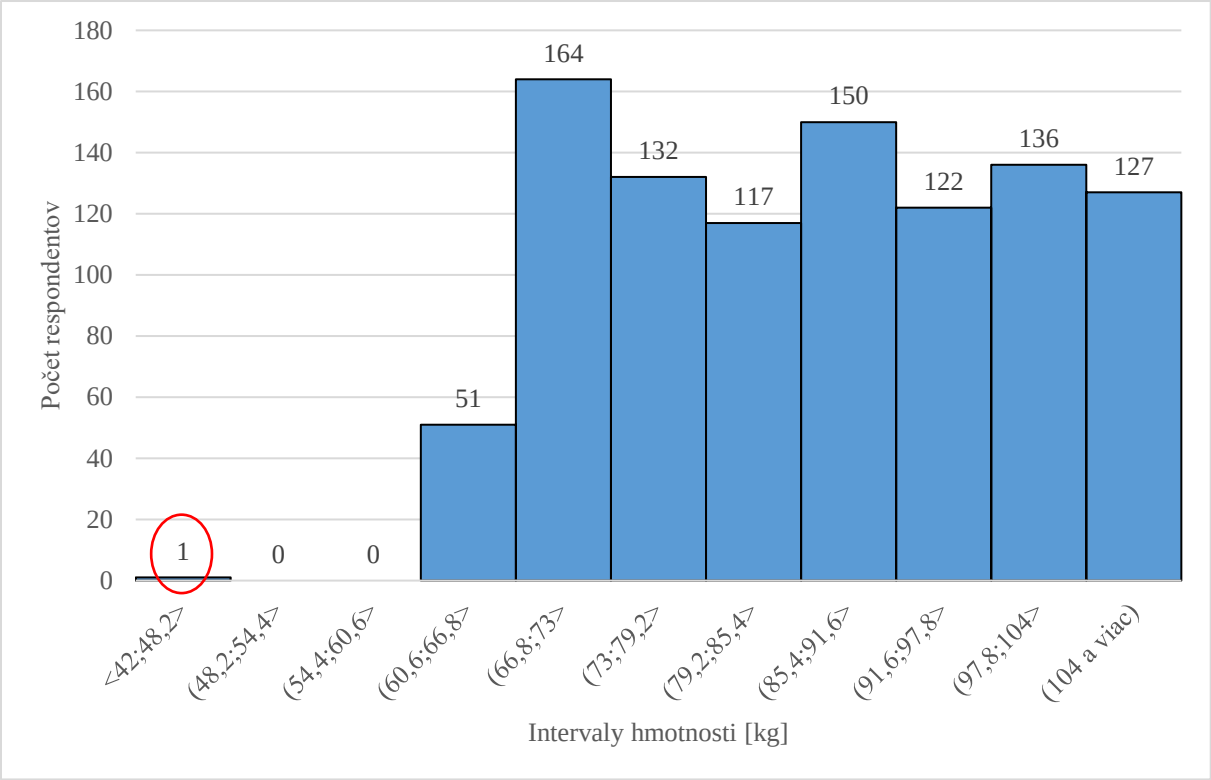
Počet tried *(k)*: $k = 1+3{,}322\log(\mathbf{999}) = 11$ (po zaokrúhlení nahor)

Variačné rozpätie: $x_{max} = \mathbf{110}$

$x_{min} = 42$

$x_{max} - x_{min} = \mathbf{68}$

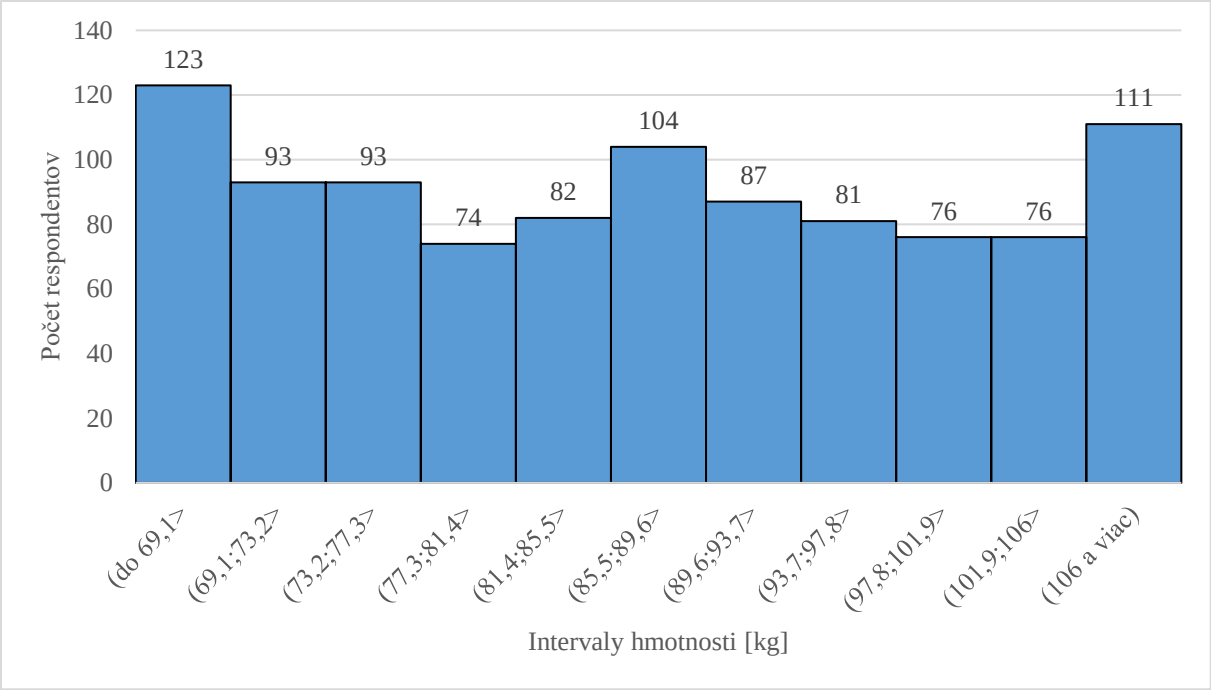
Šírka triedy *(h)*: $h = (x_{max} - x_{min})/k = \mathbf{6{,}2}$ (po zaokrúhlení na 1 desatinné miesto)



Obrázok 13 Rozdelenie vzorky mužov podľa hmotnosti po odstránení extrému 1. rádu (s extrémom 2. rádu)

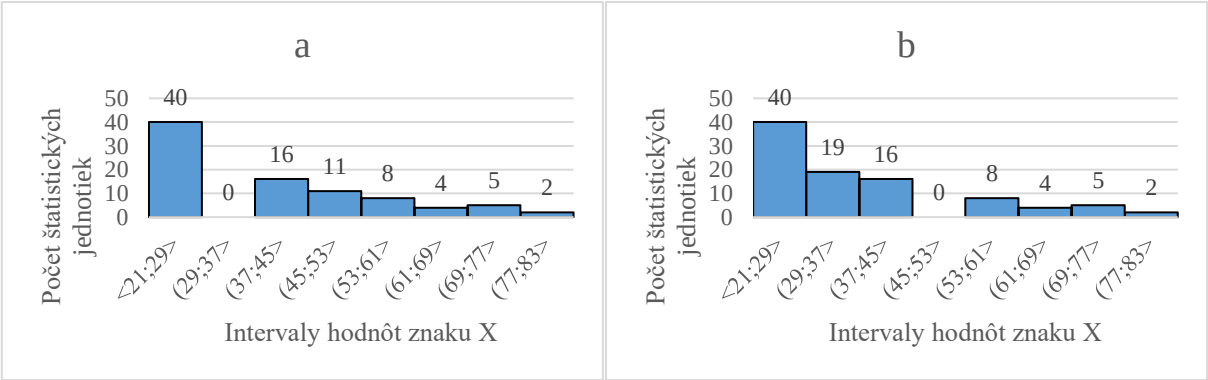
Extrémy poznáme 1. rádu, ale aj druhého rádu prípadne *n*-tého rádu. V prípade, že po odstránení extrémov 1. rádu sú vyššie uvedenou logikou identifikované ďalšie extrémy, je možné postup opakovať a odstránia sa tak aj extrémy 2. rádu. Podobne je možné pokračovať ďalej.

Dané situácie nastavajú zvyčajne v relatívne malých súboroch, ktoré nedostatočne pokrývajú spektrum rozloženia hodnôt daného štatistického znaku v populácii, ale nie je to vylúčené ani v iných príkladoch ako ilustruje uvedený príklad. Výsledné hodnoty uvedeného príkladu by boli výsledné avšak po odstránení extrému 1. rádu sa objavil extrém v minime (42 kg). Z daného dôvodu je potrebné pokračovať v postupe a odstrániť extrém, prepočítať opäť počet tried, variačné rozpätie, šírku triedy a extrém pripočítať do otvorenej triedy. Konečné rozdelenie početnosti očistené od extrémov uvedeného ilustračného príkladu vyzerá nasledovne (Obrázok 14).



Obrázok 14 Rozdelenie vzorky mužov podľa hmotnosti po odstránení extrému 1. a 2. rádu

Ako bolo naznačené vyššie prázdny interval môže signalizovať, že v triede pred ním alebo za ním sa nachádza extrém. V praxi sa však objavujú aj príklady kedy to nemusí platiť. Pre ilustráciu sú uvedené aj príklady rozdelenia početností, ktoré obsahujú prázdny interval, avšak štatistický súbor neobsahuje extrémy (Obrázok 15). V niektorých prípadoch môže ísť o chyby štatistického zisťovania, počas ktorého mohla byť opomenutá určitá skupina štatistických jednotiek, prípadne ide o objektívny dôvod a daný štatistický znak nemôže dosahovať hodnoty v stanovenom intervale – to platí v niektorých prípadoch iba pre diskrétne štatistické znaky. V praxi sa častejšie stretávame s prvým problémom, ktorý môže výrazne skresliť aj ostatné výsledky.



Obrázok 15 Príklady histrogramov (a, b) s prázdnymi intervalmi bez extrémov

Otvorenú triedu je možné použiť i v prípade, že sa extrém nevyskytol, ale riešiteľ chce zdôrazniť, že sa v spojitom prostredí môžu pri skúmaní rozsiahlejšieho súboru vyskytnúť hodnoty nižšie alebo vyššie než boli hodnoty zistené. V tomto prípade je, ale problematické správne stanovenie dolnej hranice 1. triedy x_d resp. hornej hranice poslednej triedy x_d, ako aj stredu tried.

6.2 Viacstupňové triedenie – triedenie podľa dvoch a viacerých štatistických znakov

Inak nazývané tiež triedenie v kombinácií dvoch alebo viacerých znakov. Ide o štatistickú metódu, ktorej výsledkom je taktiež logické usporiadanie štatistických jednotiek do skupín (tried), pričom zaradenie do skupiny je podmienené splnením viacerých kritérií (zvyčajne dvoch) súčasne. Štatistický súbor sa tak rozdelí na určitý počet čiastkových súborov, ktoré vykazujú rovnaké vlastnosti.

Výsledky takéhoto triedenia predstavujú primárne vyjadrenie týchto typov početností:

- príslušný počet výskytov v súbore **– absolútna početnosť n_i**
- podiel na celkovom rozsahu súboru **– relatívna početnosť p_i**

Kumulatívne početnosti sa pre tento druh v zásade neurčujú, keďže v celom rozsahu nie je možné robiť kumuláciu všetkých kombinácii postupne.

Základné rozdelenie hovorí o dvoch kategóriách triedenia podľa viacerých znakov:

- triedenie hierarchické,
- triedenie kombinačné.

Hierarchické triedenie

Predstavuje triedenie podľa ľubovoľného počtu znakov, robené v ľubovoľnom poradí. Vo vnútri tried jedného znaku sú vytvárané triedy ďalšieho (podriadeného) znaku.

Napr. študenti sú najskôr klasifikovaní podľa počtu absolvovaných skúšobných termínov a vo vnútri každého termínu sú klasifikovaní podľa dosiahnutej klasifikácie (známky). Je možné triediť/postupovať aj v opačnom poradí triedených znakov.

Typickým výsledkom triedenia je hierarchický strom **– dendrogram** (evolučný strom).

Kombinačné triedenie

Súčasné triedenie podľa ľubovoľného počtu znakov. Taktiež je možné voliť rôzne poradie. Pre účely tejto publikácie budeme uvažovať hlavne o bežne používaných formách kombinačného triedenia, a teda o **triedení v kombinácií dvoch štatistických znakov.**

Podľa potreby je možné kombinovať rôzne znaky (2x číselný znak, 2x slovný znak, kombinácia 1x číselný a 1x slovný). Najbežnejšími formami sú:

- triedenie v kombinácii dvoch číselných znakov,
- triedenie v kombinácii dvoch slovných znakov.

Typickým výsledkom triedenia sú kombinačné tabuľky, ktoré odrážajú práve povahu (charakteristiky) triedených štatistických znakov. Podľa charakteru triedených štatistických znakov sa rozlišujú tabuľky:

- **korelačná tabuľka –** triedenie podľa dvoch číselných znakov,
- **kontingenčná tabuľka –** triedenie podľa dvoch slovných znakov (aspoň 1 musí byť množný – využíva sa aj pre kombináciu 1 číselného a 1 slovného štatistického znaku),
- **asociačná tabuľka –** triedenie podľa dvoch alternatívnych slovných znakov.

Kombinačné triedenie je možné považovať za východiskový bod pre skúmanie závislosti medzi dvoma štatistickými znakmi – platí hlavne pre triedenie v kombinácii dvoch slovných štatistických znakov (kontingenčná resp. asociačná závislosť).

Napr. respondenti sú súčasne triedení podľa stupňa dosiahnutého vzdelania a ich preferovanej televíznej stanice. Výsledky tohto triedenia je možné ďalej použiť napr. pri skúmaní závislosti preferovanej televíznej stanice na vzdelaní respondentov (viac v kap. 9.4).

6.2.1 Triedenie v kombinácii dvoch číselných štatistických znakov

Pri malom počte štatistických jednotiek je základom triedenia číselných (kvantitatívnych) štatistických znakov pracovná (základná) tabuľka, do ktorej sa zaznamenávajú hodnoty štatistických znakov pre všetky štatistické jednotky od $i = 1$ až po $i = n$.

Tabuľka 11 Základná pracovná tabuľka pre triedenie v kombinácií dvoch číselných znakov

Štatistická jednotka	**Hodnoty štatistických znakov**	
	Znak x_i	**Znak y_i**
1	x_1	y_1
2	x_2	y_2
...	...	...
n	x_n	y_n

V tejto podobe ide iba o záznam výsledkov zisťovania za *n* členný štatistický súbor.

Pri veľkom rozsahu údajov je pracovná tabuľka nepraktická a neprehľadná. Výhodnejšie je v tejto situácii použiť tzv. **korelačnú tabuľku** (Tabuľka 12, Tabuľka 14, Tabuľka 15), v ktorej sú uvedené početnosti kombinácií hodnôt oboch štatistických znakov (absolútne alebo relatívne). Ak ide o nezávislé štatistické znaky je možné premenné ***X*** a ***Y*** v tabuľke a v grafe zameniť. Ak je však možné identifikovať jeden štatistický znak ako nezávislý a druhý ako závislý použijú sa hodnoty nezávislého znaku ako záhlavie stĺpcov, pričom nezávislá premenná sa označí ako ***X*** (viac v časti týkajúcej sa korelačnej úlohy a regresie – kap. 0, kap. 10).

Tabuľka 12 Všeobecný vzhľad korelačnej tabuľky

Znak Y	**Znak X**				**Σ**
	x_1	x_2	...	x_i	
y_1	$n_{x_1y_1}$	$n_{x_2y_1}$	...	$n_{x_iy_1}$	n_{y_1}
y_2	$n_{x_1y_2}$	$n_{x_2y_2}$	...	...	n_{y_2}
...	...	...	...	...	...
y_k	$n_{x_1y_k}$	...	...	$n_{x_iy_k}$	n_{y_k}
Σ	n_{x_1}	n_{x_2}	...	n_{x_i}	n

Príklad:

O 10 rodinách sú k dispozícii údaje o štatistických znakoch: veľkosť bytu (premenná *X*) vyjadrená počtom obytných miestností a počet detí v rodine (premenná *Y*)

Tabuľka 13 Počty detí a počty obytných miestností skúmaných rodín

Rodina	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.
Počet obytných miestností	2	3	3	3	1	2	3	2	4	4
Počet detí v rodine	1	1	0	2	0	1	2	0	3	2

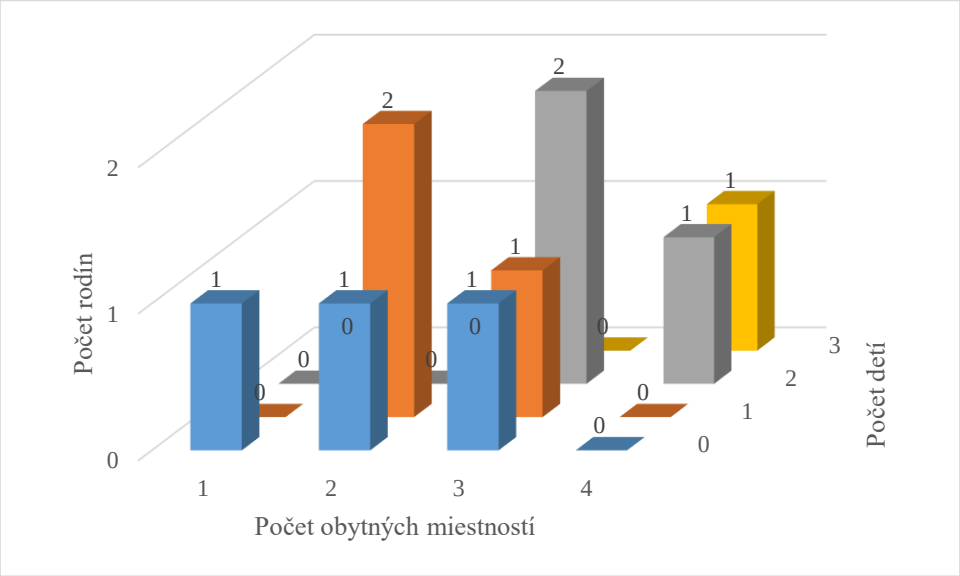
Tabuľka 14 Počet rodín podľa počtu detí a počtu obytných miestností

Počet detí	Počet obytných miestností				**Spolu**
	1	2	3	4	
0	1	1	1	-	3
1	-	2	1	-	3
2	-	-	2	1	3
3	-	-	-	1	1
Spolu	1	3	4	2	10

Tabuľka 15 Podiel rodín podľa počtu detí a počtu obytných miestností

Počet detí	Počet obytných miestností				**Spolu**
	1	2	3	4	
0	10%	10%	10%	-	30%
1	-	20%	10%	-	30%
2	-	-	20%	10%	30%
3	-	-	-	10%	10%
Spolu	10%	30%	40%	20%	100%

Graficky sa zobrazujú primárne absolútne početnosti kombinácii hodnôt skúmaných štatistických znakov, ale podľa potreby je možné zobraziť aj relatívne početnosti. Na grafické zobrazenie údajov z korelačnej tabuľky sa používa najčastejšie **pseudo 3-D graf** (Obrázok 16), kde sú na osiach *x* a *z* hodnoty štatistických znakov a na osi *y* počty štatistických jednotiek.



Obrázok 16 Rozdelenie rodín podľa počtu detí a počtu obytných miestností (pseudo 3-D priestorový graf)

V prípade, že je **počet hodnôt** niektorého číselného (kvantitatívneho) štatistického znaku **veľký**, musia byť konkrétne **hodnoty** nahradené **skupinami** (intervalmi). Tvorba (výpočet) intervalov je identická ako u skupinového triedenia (kap. 6.1.2)

6.2.2 Triedenie v kombinácii dvoch slovných štatistických znakov

Na triedenie štatistických jednotiek podľa dvoch štatistických slovných (kvalitatívnych) znakov sa používa rovnaký postup ako u číselných znakov, s výnimkou, že sa ani pri veľkom počte obmien nevyužíva tvorba intervalov.

Poradie obmien štatistických znakov v záhlaví stĺpcov a riadkov sa riadi rovnakými zásadami ako u jednoduchého triedenia.

Základom riešenia štatistických otázok, ktoré vyžadujú triedenie v kombinácii dvoch slovných znakov je vytvorenie **kontingenčnej alebo asociačnej tabuľky** (podľa charakteru skúmaných slovných štatistických znakov – viď rozdelenie v kap. 6.2).

Príklad na triedenie v kombinácií dvoch slovných štatistických znakov použitím kontingenčnej tabuľky:

V rámci skúmania štatistického problému, ktorý sa týka dopravných nehôd na Slovensku, môže riešiteľa zaujímať napríklad rozdelenie dopravných nehôd vzhľadom na štatistické znaky „druh komunikácie kde sa nehoda stala“ a „úmrtie pri nehode“ (či niekto zahynul pri nehode alebo nie). Riešenie spočíva vo vytvorení kontingenčných tabuliek absolútnych (Tabuľka 16) a relatívnych početností (Tabuľka 17) výskytu kombinácií všetkých obmien skúmaných štatistických znakov.

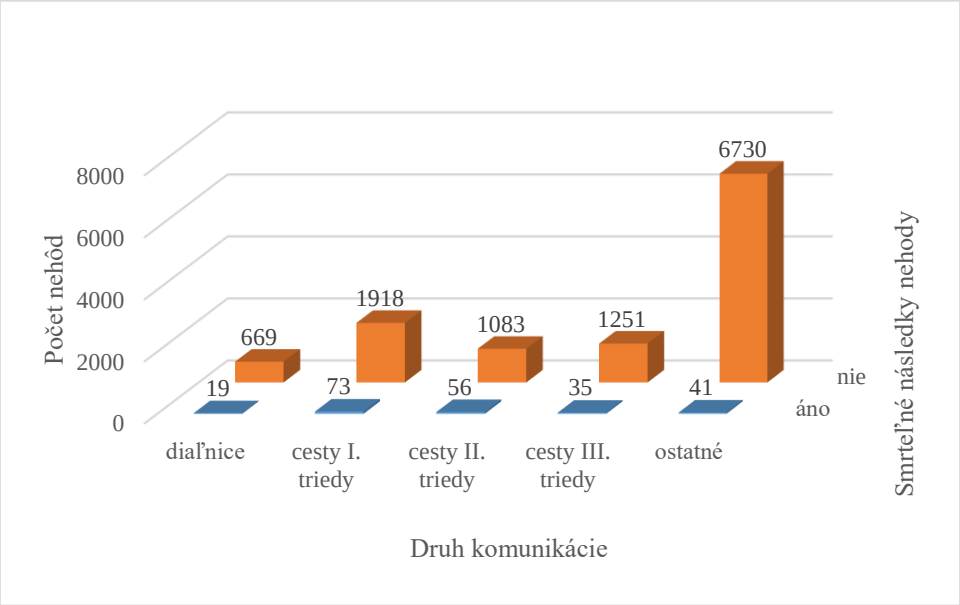
Tabuľka 16 Počet dopravných nehôd v roku 2020 podľa druhu komunikácie kde sa stali a smrteľných následkov (PPZ SR, 2020)

Úmrtie pri nehode	**Druh komunikácie**					Spolu
	diaľnice	cesty I. triedy	cesty II. triedy	cesty III. triedy	ostatné	
áno	19	73	56	35	41	224
nie	669	1918	1083	1251	6730	11651
Spolu	688	1991	1139	1286	6771	11875

Tabuľka 17 Podiel dopravných nehôd v roku 2020 podľa druhu komunikácie kde sa stali a smrteľných následkov (PPZ SR, 2020)

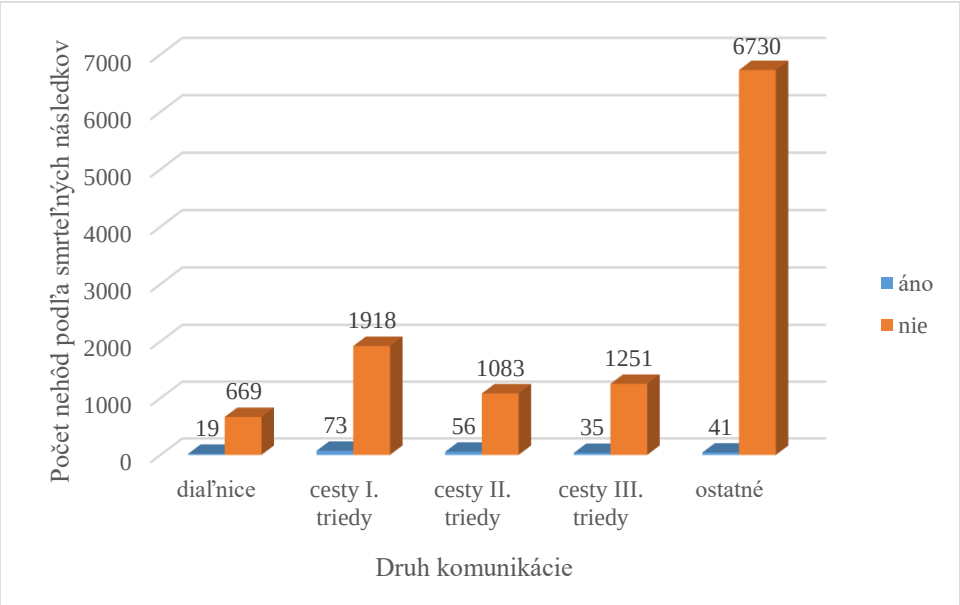
Úmrtie pri nehode	**Druh komunikácie**					Spolu
	diaľnice	cesty I. triedy	cesty II. triedy	cesty III. triedy	ostatné	
áno	0,16	0,61	0,47	0,29	0,35	1,89
nie	5,63	16,15	9,12	10,53	56,67	98,11
Spolu	5,79	16,77	9,59	10,83	57,02	100,00

Pri triedení dvoch slovných štatistických znakov sa využívajú rovnaké grafické nástroje ako pri triedení v kombinácií dvoch číselných znakov. Najčastejšie sa používa **pseudo 3-D graf** (Obrázok 17), kde sú na osiach *x* a *z* obmeny štatistických znakov a na osi *y* počty štatistických jednotiek.



Obrázok 17 Rozdelenie dopravných nehôd v roku 2020 podľa smrteľných následkov a druhu komunikácie kde sa stali (pseudo 3-D priestorový graf)

Spôsobov zobrazenia pre triedenie dvoch štatistických znakov je však podstatne viac. Pre niektoré účely a vyjadrenia sa viac hodí priestorový graf skupinový – porovnanie hodnôt medzi viacerými kategóriami (Obrázok 18), skladaný (porovnanie častí celku) alebo aj 100% skladaný graf (porovnanie percentuálnych podielov, ktorými časti prispievajú k celkovému súčtu v rámci kategórie).



Obrázok 18 Počty dopravných nehôd v roku 2020 podľa smrteľných následkov a druhu komunikácie kde sa stali (skupinový priestorový graf)

Príklad na triedenie v kombinácií dvoch slovných štatistických znakov použitím asociačnej tabuľky:

Špeciálnym prípadom triedenia podľa dvoch štatistických znakov je **asociačná tabuľka**, ktorú používame v prípade, že oba **štatistické znaky dosahujú iba dve obmeny – alternatívy.**

Na vzorke zamestnancov (n = 450) konkrétneho podniku „XY“ je záujmom riešiteľa zistiť aké je rozdelenie týchto zamestnancov podľa štatistických (alternatívnych) znakov:

- štatistický znak A: očkovanie zamestnancov (zamestnanec bol alebo nebol očkovaný),
- štatistický znak B: ochorenie zamestnancov (zamestnanec bol alebo nebol v ďalšom sledovanom období chorý).

Tabuľka 18 Triedenie zamestnancov podľa účinnosti očkovania

Očkovanie zamestnancov	**Ochorenie zamestnancov**		Σ
	bol chorý	nebol chorý	
bol očkovaný	12	323	335
nebol očkovaný	53	62	115
Σ	65	385	450

Podobne ako v predošlom príklade sa ešte počíta relatívna početnosť a používajú príslušné grafické zobrazenia výsledkov.

V praxi je možné triediť aj podľa troch a viacerých štatistických znakov. Táto problematika nie je záujmom týchto skrípt.

Literatúra

BENČO, J. *Metodológia vedeckého výskumu*. Bratislava: IRIS, 2001. ISBN 80-9018-27-0.

BUDÍKOVÁ, M, KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. Praha: GRADA, 2010

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

GROFÍK, R. a kol. *Štatistika*. Bratislava: Príroda, 1987.

HINDLS, R., ARLTOVÁ, M., HRONOVÁ, S., MALÁ, I., MAREK, L., PECÁKOVÁ, I., ŘEZANKOVÁ, H. *Statistika v ekonomii.* Praha: Professional Publishing, 2018, ISBN 978-80-88260-09-7.

CHAJDIAK, J. a kol. *Štatistické úlohy a ich riešenie v Exceli*. Bratislava: STATIS, 2005.

CHAJDIAK, J. *Analýza dotazníkových údajov*. Bratislava: Statis, 2013.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

PREZÍDIUM POLICAJNÉHO ZBORU SR. Vyhodnotenie dopravno-bezpečnostnej situácie za 12 mesiacov 2020. 2020. Dostupné na: https://www.minv.sk/swift_data/source/policia/dopravna_policia/dn/prezentacie_dbs/2020/Vyhodnotenie%20DBS%20za%20rok%202020%20def..pdf

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

TIRPÁKOVÁ, A., MARKECHOVÁ, D. *Štatistika v praxi*. Nitra: FPV UKF, 2008, ISBN 978-80-8094-283-0.

7 Základný štatistický rozbor

Základný štatistický rozbor spočíva **vo výpočtoch a prezentácii číselných charakteristík** štatistického súboru hodnôt skúmaného číselného (kvantitatívneho) štatistického znaku. Číselné charakteristiky sú číselné hodnoty, ktoré zhustením údajov súboru **súhrne charakterizujú základné vlastnosti súboru** z hľadiska skúmaného štatistického znaku. Z tohto pohľadu tvorí základný štatistický rozbor základ **popisnej štatistiky** (popisuje štatistický súbor).

Príklad: *rozbor škôd po povodniach – súhrnne (číselne) popísať aké boli škody.*

Súhrnné charakteristiky základného štatistického rozboru:

- **absolútna úroveň** (poloha) → stredné hodnoty (charakteristiky úrovne),
- **variabilitu** (premenlivosť) → miera variácie (charakteristiky variability),
- **nesúmernosť** (šikmosť) → miera nesúmernosti (šikmosti),
- **špicatosť** (koncentráciu) → miera špicatosti (koncentrácie).

Najčastejšie sa používajú **charakteristiky úrovne a charakteristiky variability.**

7.1 Charakteristiky úrovne (polohy)

Úroveň (poloha) je najnákladnejšou a najjednoduchšou vlastnosťou štatistických údajov. Úroveň meriame pomocou charakteristík úrovne.

Základnými charakteristikami **úrovne** (**polohy**) sú tzv. **stredné hodnoty:**

- priemery – $\overline{x}$,
- medián – $\tilde{x}$,
- modus – $\hat{x}$.

7.1.1 Priemery

Priemery sú stredné hodnoty, ktoré vychádzajú zo všetkých hodnôt štatistického znaku (variačného radu).

V štatistike sa rozlišuje:

- všeobecný druh priemeru – priemer **mocninový**,
- priemer **aritmetický** $\overline{x}$,
- priemer **kvadratický** $\overline{x}_{Q,}$
- priemer **harmonický** $\overline{x}_{H,}$
- priemer **geometrický** $\overline{x}_{G.}$

Rozlišujú sa dve formy výpočtu:

- **prostá forma** je využívaná u netriedených hodnôt znaku (obvykle ide o málo rozsiahle súbory),
- **vážená forma** je využívaná u hodnôt, ktoré sú triedené (pri rozdelení početností resp. intervalovom rozdelení početností).

Všeobecné vlastnosti priemerov:

- sú funkciami všetkých hodnôt variačného radu,
- ležia vždy medzi minimálnou a maximálnou hodnotou variačného radu,
- ak sa zmení ktorákoľvek z hodnôt variačného radu, zmení sa aj priemer rovnakým smerom (nie však o rovnakú hodnotu!).

Špecifické vlastnosti priemerov:

Aritmetický priemer – je založený na stálosti súčtu hodnôt.

$$x_1 + x_2 + \cdots + x_n = \bar{x} + \bar{x} + \cdots + \bar{x}$$

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

Kvadratický priemer – je založený na stálosti súčtu štvorcov hodnôt.

$$x_1^2 + x_1^2 + \cdots + x_1^2 = \bar{x}_Q^2 + \bar{x}_Q^2 + \cdots + \bar{x}_Q^2$$

$$\sum_{i=1}^{n} x_i^2 = n\bar{x}_Q^2$$

Harmonický priemer – je založený na stálosti súčtu prevrátených hodnôt.

$$\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n} = \frac{1}{\bar{x}_H} + \frac{1}{\bar{x}_H} + \cdots + \frac{1}{\bar{x}_H}$$

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

Geometrický priemer – je založený na stálosti súčinu hodnôt.

$$x_1 \cdot x_2 \cdot \ldots \cdot x_n = \bar{x}_G \cdot \bar{x}_G \cdot \ldots \cdot \bar{x}_G$$

$$\prod_{i=1}^{n} x_i = \bar{x}_G^n$$

Mocninový priemer

Priemery je možné charakterizovať všeobecným vzorcom ako s-tú odmocninu z aritmetického priemeru s-tých mocnín hodnôt číselného (kvantitatívneho) štatistického znaku.

Mocninový priemer stupňa ***s*** v **prostej forme** (netriedené údaje):

$$\bar{x}_s = \sqrt[s]{\frac{1}{n}\sum_{i=1}^{n} x_i^s}$$

Mocninový priemer stupňa ***s*** vo **váženej forme** (triedené údaje):

$$\bar{x}_s = \sqrt[s]{\frac{1}{n}\sum_{i=1}^{k} x_i^s n_i}$$

kde: x_i – hodnota znaku

n – rozsah súboru

k – počet tried

s – stupeň mocninového priemeru (celé číslo).

Pre priemery platí:

Aritmetický priemer ($s = 1$)

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

Kvadratický priemer ($s = 2$)

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

Harmonický priemer ($s = -1$)

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

Geometrický priemer ($s \to 0$)

$$pomocou\ log = \sqrt[n]{\prod x_i} = \bar{x}_G$$

Aritmetický priemer

Na meranie priemeru sa najviac používa aritmetický priemer, ktorý zjednodušene označujeme ako priemer. Aritmetický priemer **je mocninový priemer stupňa 1 (*s* = 1).**

Aritmetický priemer by sa nemal brať do úvahy, keď:

- je rozdelenie viacvrcholové,
- rozdelenie je asymetrické,
- okrajové triedy sú otvorené,
- výber obsahuje extrémne málo prvkov.

Prostá (jednoduchá) forma aritmetického priemeru

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

Vážená forma aritmetického priemeru

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

kde: $n_1 + n_2 + \cdots + n_k = n$

Pri váženej forme výpočtu môžu byť použité nielen absolútne početnosti, ale aj početnosti relatívne.

Príklad:

Štatistický súbor má rozsah n = 12. Hodnoty skúmaného kvantitatívneho štatistického znaku majú hodnoty: : 4, 4, 8, 8, 8, 9, 9, 9, 9, 9, 11, 14.

Prostá forma výpočtu:

$$\bar{x} = \frac{4 + 4 + 8 + 8 + 8 + 9 + 9 + 9 + 9 + 9 + 11 + 14}{12} = \frac{102}{12} = 8{,}5$$

Vážená forma výpočtu:

$$\bar{x} = \frac{4 \cdot 2 + 8 \cdot 3 + 9 \cdot 5 + 11 \cdot 1 + 14 \cdot 1}{2 + 3 + 5 + 1 + 1} = \frac{102}{12} = 8{,}5$$

Vážená forma musí byť uplatňovaná pri výpočte aritmetického priemeru:

- z rozdelenia početností či intervalového rozdelenia početností,
- z čiastkových priemerov,
- z pomerných čísel alebo percent.

Príklad:

Známky žiakov z matematiky v určitej triede sú uvedené v nasledujúcej tabuľke (Tabuľka 19).

Tabuľka 19 Rozdelenie žiakov podľa známky z matematiky

Štatistický znak	Známka	1	2	3	4	5
Početnosť	Počet známok (žiakov)	14	6	5	4	1

Priemerná známka sa vypočíta ako vážený priemer:

$$\bar{x} = \frac{14 \cdot 1 + 6 \cdot 2 + 5 \cdot 3 + 4 \cdot 4 + 1 \cdot 5}{14 + 6 + 54 + 1} = \frac{62}{30} = 2{,}0\overline{6}$$

Vlastnosti aritmetického priemeru

Súčet hodnôt všetkých hodnôt štatistického znaku (x_i) je rovnaký ako súčin aritmetického priemeru a rozsahu štatistického súboru *n*:

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

Súčet rozdielov jednotlivých hodnôt štatistického znaku (x_i) a ich aritmetického priemeru je rovný nule :

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

Súčet štvorcov rozdielov jednotlivých hodnôt štatistického znaku od ich aritmetického priemeru je minimálny :

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

Pripočítanie (odpočítanie) rovnakej konštanty k (od) každej hodnoty štatistického znaku (x_i) má za následok zvýšenie (zníženie) aritmetického priemeru o túto konštantu.

$$\frac{\sum(x_i \pm a)}{n} = \bar{x} \pm a$$

Násobením (delením) všetkých hodnôt štatistického znaku (x_i) rovnakou konštantou sa aritmetický priemer zvýši (zníži) konštantne-krát.

$$\frac{\sum x_i c}{n} = c \cdot \bar{x}$$

Násobením (delením) všetkých početností rovnakou konštantou sa aritmetický priemer nemení.

$$\frac{\sum x_i n_i c}{\sum n_i c} = \bar{x}$$

Aritmetický priemer súčtu (rozdielu) hodnôt dvoch štatistických znakov je rovný súčtu (rozdielu) ich aritmetických priemerov :

$$\overline{x_i \pm y_i} = \bar{x} + \bar{y}$$

Kvadratický priemer

Pre ***s* = 2** bude kvadratický priemer v prostej forme:

$$\bar{x}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

resp. vo váženej forme:

$$\bar{x}_k = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2\, n_i}$$

Kvadratický priemer sa obyčajne používa vo fyzikálnych aplikáciách.

Harmonický priemer

Pre ***s* = -1** bude harmonický priemer v prostej forme:

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

resp. vo váženej forme:

$$\bar{x}_h = \frac{\sum_{i=1}^{k} n_i}{\sum_{i=1}^{k} \frac{n_i}{x_i}}, pre\ x_i > 0$$

Harmonický priemer sa často používa na charakteristiku hodnôt, ktoré predstavujú napríklad výkonové limity, alebo na priemerovanie štatistických znakov, ktoré majú charakter rozmerných či bezrozmerných **pomerových čísel**, pričom váha je veličina z **čitateľa** zlomku — napr. *výpočet priemernej rýchlosti* (dráha/čas), kde váhami sú dráhy; *výpočet priemernej chorobnosti* (počet chorých/počet všetkých), kde váhami sú počty chorých; *výpočet priemernej marže zisku* (zisk/tržba), kde váhami sú zisky. Ak je váhou veličina z menovateľa zlomku, používa sa aritmetický priemer.

Geometrický priemer

Pre ***s* = 0** bude geometrický priemer v prostej forme:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i}, pre\ x_i > 0$$

Geometrický priemer sa často používa v ekonomických a obchodných výpočtoch ako ukazovateľ rastu alebo podielu (zisku). Používa sa taktiež na priemerovanie bezrozmerných rastových charakteristík reťazených (previazaných) v čase (koeficienty rastu, reťazové indexy), kedy celková zmena je daná ako súčin čiastkových zmien — napr. *priemerná mesačná inflácia vypočítaná z údajov za niekoľko po sebe nasledujúcich mesiacov* (cenová hladina daného mesiaca/cenová hladina predchádzajúceho mesiaca).

Vzhľadom k tomu, že mocninový priemer stupňa ***s*** je **neklesajúcou** funkciou čísla ***s***, platilo by pri výpočte z rovnakých údajov $\bar{x}_h \leq \bar{x}_g \leq \bar{x} \leq \bar{x}_k$.

Nevýhody priemerov

V štatistickej praxi sa najčastejšie ako charakteristika polohy používa aritmetický priemer, lebo **závisí od všetkých pozorovaných hodnôt**. Je však veľmi citlivý na to, ak sú niektoré hodnoty znaku extrémne veľké alebo malé. **Extrémne hodnoty** môžu spôsobiť, že priemer nie je najlepšou charakteristikou polohy. V takýchto prípadoch treba použiť inú charakteristiku polohy.

Príklad:

Počet vymeškaných hodín u žiakov za školský rok: 5, 6, 110, 8, 10, 5, 9, 12, 6. Hodnota aritmetického priemeru je 19 hodín, pričom je zrejmé, že takmer všetci žiaci majú počet vymeškaných hodín pod 10 hodín alebo tesne okolo tejto hodnoty. Je zjavné, že u jedného žiaka je počet vymeškaných hodín oproti ostatným žiakom extrémny. Pri výpočte priemeru môžeme za určitých okolností túto hodnotu vynechať, alebo rozdeliť štatistický súbor na menšie časti.

7.1.2 Medián a kvantily

Medián ako charakteristika polohy (absolútnej úrovne) sa radí medzi tzv. **kvantily.** Kvantily rozdeľujú **vzostupne usporiadaný** variačný rad hodnôt v určitom pomere početností (napr. v pomere k dvom, čo znamená, že rozdeľujú rad na 2 rovnaké časti).

Označuje sa percentuálne (napr. 50% kvantil) alebo v rozmedzí od 0 po 1 (napr. $x_{0,50}$).

Príklady kvantilov:

- **Medián** 50% kvantil ($x_{0,50}$) – rozdeľuje štatistický súbor na dve polovice,
- **Kvartily** 25%, 50% a 75% kvantil ($x_{0,25}$, $x_{0,50}$, $x_{0,75}$),
- **Decily** 9 kvantilov ($x_{0,10}$, $x_{0,20}$,..., $x_{0,90}$),
- **Percentily** 99 kvantilov, ktoré rozdeľujú súbor na 100 dielov po 1 %,
- **Oktily** a pod.

Rozhodujúce je **poradie hodnôt vo variačnom rade**, nie ich samotná hodnota.

Medián $\tilde{x}$ sa chápe ako **prostredná hodnota** usporiadaného radu hodnôt štatistického znaku, ktorý rozdeľuje vzostupne usporiadaný variačný rad práve v pomere k dvom. Ako taký delí tento rad na dve rovnaké časti (rovnaký počet členov radu – štatistických jednotiek).

Príklad:

Pri **nepárnom počte členov** variačného radu je medián prostredný člen. V usporiadanom rade hodnôt 5, 7, 8, 8, 8, 11, **12**, 15, 17, 17, 20, 20, 22 je medián číslo 12. Keďže je 13 členov radu, tak hodnota práve 7. v poradí je hodnotou mediánu (šesť hodnôt je pred 7. členom a šesť hodnôt je za 7. členom radu). Je možné použiť aj jednoduchý výpočet:

$$\tilde{x} = \frac{n+1}{2} = \frac{13+1}{2} = 7.\check{c}len = 12$$

V prípade **párneho počtu členov** variačného radu je mediánom priemer hodnôt dvoch prostredných členov radu. V usporiadanom rade hodnôt 5 ,7, 8, 8, 8, 11, **12, 15,** 17, 17, 20, 20, 22, 25 je medián číslo 13,5, pretože máme 14 členov radu a práve toto číslo je priemerom medzi 7. a 8. hodnotou radu (sedem hodnôt je pred mediánom a sedem hodnôt je za mediánom). Je možné použiť aj jednoduchý výpočet:

$$\tilde{x} = \frac{\frac{n}{2} + (\frac{n}{2}+1)}{2} = \frac{\frac{14}{2} + (\frac{14}{2}+1)}{2} = \frac{7.\check{c}len + (8.\check{c}len)}{2} = \frac{12+15}{2} = 13{,}5$$

Zisťovanie mediánu **nezávisí na hodnotách** skúmaného štatistického znaku**,** závisí na rozdelení početnosti hodnôt štatistického znaku.

7.1.3 Modus

Modus je charakterizovaný ako **najpočetnejšia hodnota** kvantitatívneho štatistického znaku skúmaného štatistického súboru, je to teda hodnota, ktorá sa **v štatistickom súbore vyskytuje najčastejšie**. Zisťovanie modusu závisí na rozdelení početnosti hodnôt štatistického znaku.

Príklad:

Napr. v súbore s hodnotami znaku 8, 9, 9, 12, 12, 12, **15, 15, 15, 15, 15**, 16, 16, 17, 17, 17, 20 je najčastejšou hodnotou 15, takže modus je rovný tomuto číslu; $\hat{x} = 15$.

Ak existujú dve navzájom **nesusediace** hodnoty s relatívne najväčšími početnosťami, tak obe tieto hodnoty uvádzame ako modus. V takomto prípade hovoríme, že rozdelenie je **bimodálne** (dvojvrcholové).

Pre údaje triedené **skupinovým triedením** určujeme **modálny interval** (typickú triedu) pričom modálnym intervalom je interval s najväčšou početnosťou (Tabuľka 20). Pri určitom zjednodušení môžeme za modus považovať strednú hodnotu (triedny znak) tohto intervalu.

Tabuľka 20 Rozdelenie domácností podľa mesačných príjmov so zvýraznením modálneho intervalu mesačných príjmov

Trieda	Hranice mesačných príjmov [€]	Stred príjmovej hranice [€]	Počet domácností	Relatívny počet domácností [%]	Kumulatívna početnosť domácností	
					absolútna	relatívna [%]
1.	(do 1000>	750	12	15,0	12	15,0
2.	**(1000; 1500>**	**1250**	**32**	**40,0**	**44**	**55,0**
3.	(1500; 2000>	1750	20	25,0	64	80,0
4.	(2000; 2500>	2250	8	10,0	72	90,0
5.	(2500; 3000>	2750	6	8,0	78	98,0
6.	(3000 a viac)	3250	2	2,0	80	100,0
Σ	-	-	80	100,00	-	-

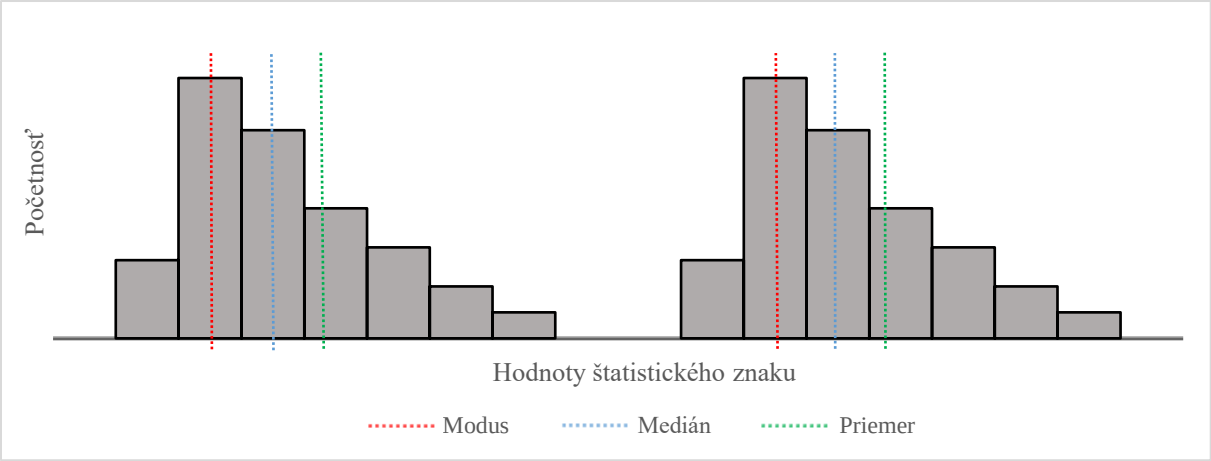
Graficky je možné modus odvodiť z rozdelenia početnosti spojitého štatistického znaku. Prislúcha hodnote znaku pod vrcholom frekvenčnej krivky.

7.1.4 Vzťahy medzi charakteristikami úrovne

Vzájomná poloha aritmetického priemeru, mediánu a modusu hovorí o **asymetrii** rozdelenia hodnôt štatistického znaku:

- ak je $\bar{x} = \tilde{x} = \hat{x}$, ide o **súmerné** (symetrické) rozdelenie hodnôt štatistického znaku,
- ak je $\hat{x} < \tilde{x} < \bar{x}$, ide o **ľavostrannú** (kladnú) asymetriu,
- ak je $\bar{x} < \tilde{x} < \hat{x}$, ide o **pravostrannú** (zápornú) asymetriu.

Nasledujúci príklad ukazuje dva triedené údajové rady, ktoré sa líšia iba polohou na osi *x* (Obrázok 19).

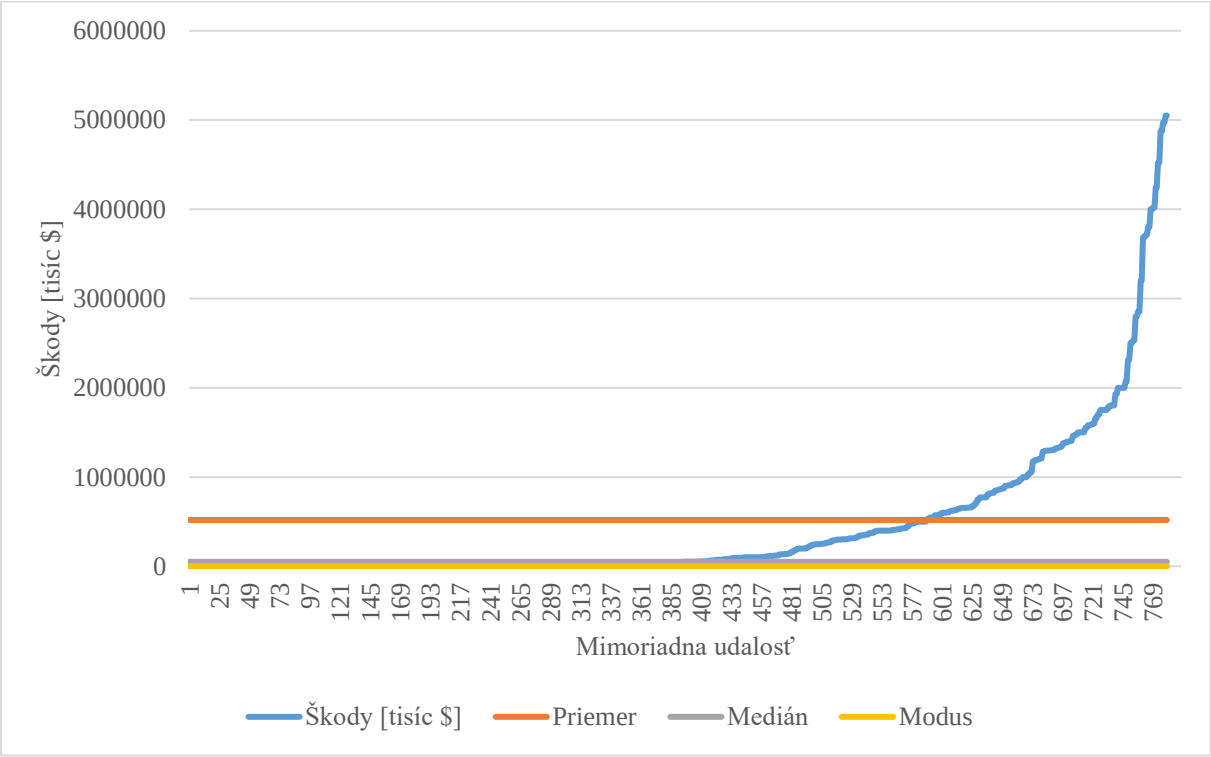


Obrázok 19 Dva štatistické súbory odlišujúce sa úrovňou (polohou)

Pre porovnanie stredných hodnôt súboru je možné použiť aj čiarový graf (Obrázok 20), ktorý obsahuje jednotlivé vzostupne usporiadané hodnoty skúmaného štatistického znaku a jeho vybrané charakteristiky.

Príklad:

Na vzorke mimoriadnych udalostí (n = 778), ktoré sa udiali vo svete v období od roku 2010 (podľa databázy EM-DAT) boli skúmané základné charakteristiky polohy pre štatistický znak „škody po mimoriadnych udalostiach“. Ich možné grafické spracovanie je zobrazené na nasledujúcom obrázku (Obrázok 20).



Obrázok 20 Vzostupne usporiadané škody po jednotlivých mimoriadnych udalostiach a ich stredné hodnoty

Je zrejmé, že najčastejšie boli mimoriadne udalosti s nulovými škodami. V skúmanom štatistickom súbore je však aj množstvo extrémne veľkých škôd. Dané skutočnosti ovplyvňujú priemer a medián a preto je viditeľný značný rozdiel medzi týmito charakteristikami. Dané výsledky naznačujú heterogenitu skúmaného štatistického súboru, čo však lepšie popisujú charakteristiky variability.

7.2 Charakteristiky variability

Variabilita = premenlivosť, kolísavosť (napr. v časovom rade) číselných údajov.

Charakteristiky variability sú významnou skupinou jednorozmerných súhrnných číselných charakteristík. Zatiaľ čo stredné hodnoty dávajú informáciu o absolútnej úrovni, nič nevypovedajú o premenlivosti (rozdielnosti) údajov.

Charakteristiky variability rozširujú tieto informácie tým, že **charakterizujú premenlivosť (variabilitu) skúmaného kvantitatívneho znaku** v danom súbore.

Príklad:

Dve skupiny študentov píšu rovnaký test s nasledujúcimi výsledkami:

- 1. skupina: 65; 66; 67; 68; 71; 73; 74; 77; 77; 77,
- 2. skupina: 42; 54; 58; 62; 67; 77; 77; 85; 93; 100.

Výpočet charakteristík úrovne ukazuje rovnaké hodnoty pre obe skupiny $\overline{x}$ = 71,5; $\tilde{x}$ = 72; $\hat{x}$ = 77.

Na prvý pohlaď nie je z uvedených charakteristík viditeľný rozdiel medzi skupinami. Podrobnejšie skúmanie, ale ukáže, že rozdiel medzi skupinami je v tom, že výsledky 2. skupiny sú omnoho rozptýlenejšie (rozkolísanejšie) – **variabilnejšie**. Variabilita hodnôt štatistického znaku je jednou z charakteristík, voči ktorým sú hlavne priemery necitlivé.

Čím je väčšia variabilita, tým je takisto väčšia hodnota charakteristík variability. **Homogénnejšie súbory majú** teda menšiu variabilitu a to má vo svojom dôsledku vplyv na **lepšiu výpovednú schopnosť stredných hodnôt**.

S variabilitou je možné sa stretnúť:

- medzi rôznymi štatistickými jednotkami toho istého štatistického súboru,
- u jednej štatistickej jednotky v rôznych časových intervaloch alebo okamžikoch štatistického zisťovania,
- u jednej štatistickej jednotky pri opakovanom štatistickom zisťovaní rovnakej konštantnej hodnoty — náhodné chyby merania,
- defekty v údajoch — hrubé chyby, heterogenita údajov.

Zatiaľ čo prvé dva body reprezentujú prirodzenú variabilitu, ktorej zdrojom je rôznosť podmienok v priestore a čase, ďalšie dva predstavujú chybovú variabilitu, ktorej prítomnosť v údajoch je nežiadúca.

Prehľad charakteristík variability:

- variačné rozpätie $\boldsymbol{R_V}$,

- kvantilové odchýlky (kvartilová Q, decilová D, percentilová P),
- priemerná odchýlka absolútna $\overline{\boldsymbol{d}}$**,**
- priemerná odchýlka relatívna $\overline{\boldsymbol{d}}'$**,**
- rozptyl (variácia) $\boldsymbol{\sigma_{x^2}}$**,** $\boldsymbol{s_{x^2}}$,
- smerodajná odchýlka $\boldsymbol{\sigma_x}$**,** $\boldsymbol{s_x}$,
- variačný koeficient $\boldsymbol{V_x}$**.**

Základnými charakteristikami variability, významnými hodnotami sú:

- **variačné rozpätie,**
- **rozptyl,**
- **smerodajná odchýlka.**

7.2.1 Variačné rozpätie

Variačné rozpätie je rýchla, jednoduchá, ale iba orientačná charakteristika variability, ktorá je založená na informácii o rozdiele medzi maximálnou a minimálnou hodnotou súboru.

$$\boldsymbol{R_V = x_{max} - x_{min}}$$

Pri použití variačného rozpätia je potrebné si vždy uvedomovať, že hodnoty minima a maxima v súbore môžu mať charakter náhodných extrémov a tým neprimerane zväčšovať riešiteľovu predstavu o miere variability v skúmanom súbore.

7.2.2 Priemerné odchýlky

Priemerná odchýlka je založená na rozdiele medzi nameranými hodnotami štatistického znaku a určitej stanovenej hodnoty**.** Veličinami, ku ktorým sa priemerné odchýlky spravidla vzťahujú sú medián a aritmetický priemer.

Priemerné odchýlky je možné stanoviť **v absolútnej, ako aj v relatívnej podobe**. **Relatívna priemerná odchýlka** môže nadobúdať hodnoty od 0 % do 100 %. Čím viac sa blíži odchýlka k nule, tým viac je skúmaný štatistický súbor homogénny, a teda s väčšou presnosťou a vierohodnosťou je ho možné popísať pomocou priemeru, modusu alebo mediánu.

Priemerná absolútna odchýlka $\overline{\boldsymbol{d}}$ okolo aritmetického priemeru je **definovaná ako absolútna** hodnota odchýlok jednotlivých hodnôt štatistického súboru x_i od aritmetického priemeru $\overline{x}$.

V prípade netriedených súborov má podobu:

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

V prípade triedených súborov má tvar:

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

Pri výpočte priemernej absolútnej odchýlky okolo mediánu používame v uvedených vzorcoch hodnotu mediánu.

7.2.3 Rozptyl

Rozptyl je definovaný ako **aritmetický priemer zo štvorcov odchýlok jednotlivých hodnôt od aritmetického priemeru** (priemerná štvorcová odchýlka okolo aritmetického priemeru).

V prípade netriedených súborov má podobu:

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

V prípade triedených súborov má tvar:

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

Väčšinou sa však stretneme s **výberovým rozptylom s_{x^2}.** V praxi máme k dispozícii len výberovú vzorku údajov zo základného štatistického súboru daného štatistického znaku. Preto sa používa modifikovaný vzorec rozptylu pre počet štatistických jednotiek n <30.

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

Vlastnosti rozptylu:

- rozptyl je **nezáporný**,
- rozptyl konštanty je rovný nule,
- rozptyl je najmenšia priemerná štvorcová odchýlka (viď vlastnosti priemeru).

Príklad:

100 študentov robilo test, za ktorý mohli získať maximálne 10 bodov. Nasledujúca tabuľka (Tabuľka 21) sumarizuje všetky výsledky a súčasne ponúka všetky dôležité čiastkové výpočty na lepšie pochopenie výpočtu rozptylu.

Tabuľka 21 Príklad na čiastkové výpočty na získanie rozptylu

Výsledok testu [bod]	**Počet výsledkov (študentov)**	**Priemerné hodnotenie z testu [bod]**	**"Vzdialenosť" od priemeru [bod]**	**Druhá mocnina vzdialenosti [bod^2]**	**Súčin počtu výsledkov a druhej mocniny vzdialenosti [bod^2]**
10	5	5,89	4,11	16,89	84,46
9	6	5,89	3,11	9,67	58,03
8	19	5,89	2,11	4,45	84,59
7	17	5,89	1,11	1,23	20,95
6	18	5,89	0,11	0,01	0,22
5	11	5,89	-0,89	0,79	8,71
4	6	5,89	-1,89	3,57	21,43

3	4	5,89	-2,89	8,35	33,41
2	4	5,89	-3,89	15,13	60,53
1	7	5,89	-4,89	23,91	167,38
0	3	5,89	-5,89	34,69	104,08
Σ	100				643,79

Rozptyl sa následne vypočíta ako priemer súčtu štvorcových odchýlok: σ_{x^2} = 643,79/100 = 6,44.

7.2.4 Smerodajná odchýlka

Rozptyl sám o sebe nie je dobre interpretovateľnou veličinou, pretože výsledok je daný v štvorcových merných jednotkách. Preto sa pri hodnotení variability **dáva prednosť druhej odmocnine rozptylu tzv. smerodajnej odchýlke σ_x resp. s_x**

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

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

Jej rozmer **zodpovedá rozmeru údajov**, **je vždy väčšia než priemerná absolútna odchýlka od aritmetického priemeru.**

Príklad:

V nadväznosti na predchádzajúci príklad, ktorý sa týkal výsledkov testov študentov stačí pre výpočet smerodajnej odchýlky iba odmocniť hodnotu rozptylu: $\sigma_x = \sqrt{6{,}44} = 2{,}54$.

7.2.5 Variačný koeficient

O variabilite súboru rozhoduje riešiteľ porovnaním smerodajnej odchýlky s aritmetickým priemerom. Čím väčší rozdiel, tým väčšia variabilita. Počíta sa tzv. **variačný koeficient:**

$$v_x = \frac{\sigma_x}{\bar{x}} \text{ pre } \bar{x} \neq 0$$

$$v_x = \frac{\sigma_x}{\bar{x}} * 100\ [\%]$$

Variačný koeficient patrí medzi relatívne miery variability, pretože nevyjadruje variabilitu v pôvodných merných jednotkách, ale pomerom smerodajnej odchýlky a aritmetického priemeru. Obvykle sa tento pomer prezentuje v percentách. Potom tento pomer udáva, z koľko percent sa v priemere odchyľujú jednotlivé hodnoty od aritmetického priemeru. Odporúča sa používať iba pre kladné hodnoty štatistického znaku.

Pre rôzne hodnoty variačného koeficientu je zaužívaná nasledovná **stupnica variability**:

- neveľká: 0 % – 4%,
- normálna: 5% – 44%,
- veľká: 45% – 64%,
- veľmi veľká: 65% – 84%,
- extrémna: 85% – 104%,
- anomálna: 105% a viac percent.

Pre symetrické rozdelenia početnosti by nemala hodnota variačného koeficientu presahovať 50%.

Variačný koeficient sa často používa na porovnanie variability dvoch štatistických súborov vzhľadom na skúmaný štatistický znak.

Príklad:

Porovnávame variabilitu príjmov v rôznych hospodárskych odvetviach. V potravinárskom priemysle boli zistené hodnoty $\overline{x}$ = 1278 €; σ_x=212 €; v hutníckom priemysle $\overline{x}$ = 1742 €; σ_x=230 €. Hodnoty variačných koeficientov sú nasledujúce (potravinársky priemysel v_{x_P}; hutnícky priemysel v_{x_H}):

$$v_{x_P} = \frac{212}{1278} = 16{,}7\%$$

$$v_{x_H} = \frac{230}{1742} = 13{,}2\%$$

Vypočítané hodnoty variačných koeficientov potvrdzujú, že v štatistickom súbore mesačných príjmov pracovníkov v potravinárskom priemysle je variabilita väčšia ako v hutníckom priemysle.

Literatúra

BUDÍKOVÁ, M, KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. Praha: GRADA, 2010.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

EM-DAT. The International Disaster Database. Dostupné na: https://public.emdat.be/

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

GROFÍK, R. a kol. *Štatistika*. Bratislava: Príroda, 1987.

HINDLS, R., ARLTOVÁ, M., HRONOVÁ, S., MALÁ, I., MAREK, L., PECÁKOVÁ, I., ŘEZANKOVÁ, H. *Statistika v ekonomii*. Praha: Professional Publishing, 2018, ISBN 978-80-88260-09-7.

CHAJDIAK, J. a kol. *Štatistické úlohy a ich riešenie v Exceli*. Bratislava: STATIS, 2005.

CHAJDIAK, J. a kol. *Štatistika jednoducho*. Bratislava: STATIS, 2003.

CHAJDIAK, J. *Analýza dotazníkových údajov*. Bratislava: Statis, 2013. ISBN 978-80-85659-76-4.

KOVAČKA, M., KONTEŠOVÁ, O. *Štatistické metódy*. 2. vyd. Bratislava: Slovenské vydavateľstvo technickej literatúry, 1962.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

PECÁKOVÁ, I. *Statistika v terénních průzkumech*. 2018. 3.vyd. Professional Publishing, Praha. ISBN 978-80-88260-10-3.

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

TIRPÁKOVÁ, A., MARKECHOVÁ, D. *Štatistika v praxi*. Nitra: FPV UKF, 2008, ISBN 978-80-8094-283-0.

VARGA, Š. *Matematická štatistika.* Bratislava: STU, 2012, ISBN 978-80-2273-789-0.

8 Pravdepodobnosť

Pravdepodobnosť (hovorovo *šanca*; značka je *P* z anglického *probability*) je **hodnota vyčíslujúca istotu resp. neistotu** výskytu určitej udalosti. K pojmu pravdepodobnosť dospejeme zovšeobecnením a abstraktným vyjadrením empirických skúseností z rôznych oblastí ľudskej činnosti.

Pravdepodobnostnú hodnotu nadobúdajú **náhodné javy (premenné)**. Je to teda číselná miera možnosti, že nastane náhodný jav.

Príbuzným pojmom k pravdepodobnosti **je pojem šanca**, ktorý vyjadruje pomer početností výskytu priaznivých a nepriaznivých prípadov — zatiaľ čo **pravdepodobnosť,** že padne napr. znak na minci **je 0,5**, **šanca**, že nastane tento jav **je 1:1,** prípadne v relatívnych hodnotách **50:50.**

Intuitívne (laické) chápanie pravdepodobnosti je spojené s každodennou ľudskou činnosťou, keď hovoríme napr. o počasí (na 100 % bude pršať) alebo o šanci, že stihneme autobus MHD, keď sme zaspali alebo o našej šanci vyhrať niektorú z cien na zakúpenom žrebe.

Pravdepodobnosťou sa zaoberá a aj ju skúma **teória pravdepodobnosti.** Teória pravdepodobnosti je **odvetvie matematiky**, ktoré umožňuje **podľa pravdepodobnosti určitých náhodných udalostí nájsť iné náhodné udalosti**, ktoré nejakým spôsobom súvisia s prvými. Vo vede existuje niekoľko koncepcií práce s náhodou (**napr. teória hromadnej obsluhy**). Pravdepodobnosť je iba jednou z nich.

8.1 Základné pojmy teórie pravdepodobnosti

8.1.1 Náhodný jav a náhodný experiment

Nech je definovaný komplex podmienok, za ktorých je sledovaná možnosť vzniku nejakého javu (napr. premena vody na paru pri danej teplote, tlaku a pod.).

Potom:

Jav, **ktorý** za týchto podmienok **nemôže nikdy nastať,** sa nazýva **jav nemožný** (napr. premena vody na paru pri normálnom tlaku a teplote 10 °C).

Jav, ktorý za týchto podmienok nutne **musí nastať,** sa nazýva **jav istý** (napr. premena vody na paru pri normálnom tlaku a teplote 100 °C).

Jav, ktorý i pri striktnom dodržaní podmienok **môže, ale nemusí nastať, prípadne nastáva s rôznou intenzitou,** sa nazýva **javom náhodným.**

Náhodný jav teda nie je určený komplexom podmienok, ale o jeho vzniku (či jav nastane) či o tom, že nenastane spolurozhoduje **náhoda**.

Nemožný jav sa označuje symbolom ***V***, **istý jav** symbolom ***I***, pre **náhodné javy** sú vyhradené veľké písmená zo začiatku latinskej abecedy, napr. ***A, B, C***, eventuálne $A_1, A_2, A_3, \ldots\ldots A_n$ a pod.

Každý dej, ktorý v sebe obsahuje prvok náhody a ktorého výsledok má charakter náhodného javu, nazývame **náhodný experiment**. Tento pojem je v pravdepodobnosti **široko využívaný** i tam, kde sa v „bežnom živote“ nevyskytuje (teda nielen *hádzanie kockou alebo streľba do terča, ale i výroba výrobku, liečenie pacienta, atď.*). Príklady o hode mincou či kockou, streľba do terča a pod. sa vzhľadom k jednoduchosti a všeobecnej znalosti podmienok týchto náhodných experimentov používajú celkom bežne.

Základnou jednotkou, ktorá označuje výsledok náhodného experimentu je **možný prípad (alebo elementárny jav)**. Ak skončí náhodný experiment tým, že nastane nejaký náhodný jav ***A***, hovoríme, že nastal **priaznivý prípad pre jav *A*** (v opačnom prípade nastal **nepriaznivý prípad**).

Zložený jav je výsledok náhodného pokusu, ktorý je možné ďalej rozložiť (napr. výsledok, že padne párne číslo možno ďalej rozložiť na elementárne javy: padne dvojka, padne štvorka, padne šestka).

8.1.2 Elementárny jav, základný priestor javov, opačné javy

Každý z možných výsledkov náhodného experimentu sa nazýva možným prípadom alebo elementárnym javom.

Elementárne javy sa označujú symbolmi $e_1, e_2, \ldots$.

Elementárny jav má tieto vlastnosti:

- je nezlučiteľný (disjunktný) s ľubovoľným iným elementárnym javom – ak padne na kocke napr. číslo 1, nemôže súčasne padnúť žiadne iné číslo,
- tvorí úplnú skupinu javov (jeden z nich musí nutne nastať) – nejaké číslo medzi 1 a 6 padnúť musí.
- môže nastať práve jedným spôsobom, je nerozložiteľný – každé číslo na kocke môže padnúť iba jedným spôsobom.

Množina všetkých elementárnych javov určitého náhodného experimentu sa nazýva **základný priestor javov** a označuje sa Ω.

Do základného priestoru javov patrí vždy jav istý *I* a taktiež jav nemožný *V*.

Ak je elementárnych javov konečný počet, bude posledný z nich e_n a hovoríme o konečnom základnom priestore javov.

Elementárnych javov môže byť aj nekonečne veľa. Ak k očíslovaniu všetkých elementárnych javov postačia prirodzené čísla, označujeme ich počet za sčítateľne nekonečný, inak ide o nespočetne nekonečný počet.

Príklady určenia počtu elementárnych javov

Hod mincou

Základný priestor javov je tvorený javom istým (padne hlava alebo znak), javom nemožným (nepadne ani hlava ani znak) a dvoma elementárnymi javmi (padne hlava, padne znak). Celkový počet javov je $2^2 = 4$.

Hod kockou

Základný priestor javov je tvorený javom istým, nemožným a šiestimi elementárnymi javmi a 56 zloženými javmi (napr. padne párne číslo, padne číslo menšie než 4, atď.). Celkový počet javov je $2^6 = 64$.

Všimnime si, že v oboch prípadoch je exponent u čísla 2 tvorený počtom elementárnych javov. Celkový počet javu konečného je rovný 2^n, kde n je počet elementárnych javov.

Opačné javy

Pri hádzaní kockou bude javom *E „padnutie párneho čísla*“, ktorý obsahuje elementárne javy e_2, e_4, e_6.

Opačným javom *E´ = „padnutie nepárneho čísla“* bude taký, ktorý obsahuje všetky ostávajúce elementárne javy e_1, e_3, e_5 .

Opačné javy tvoria úplnú skupinu nezlučiteľných (disjunktných) javov.

8.2 Definície pravdepodobnosti

Pravdepodobnosť má viac, takmer, ekvivalentných definícií.

8.2.1 Klasická definícia pravdepodobnosti (Pierre Simone de Laplace)

Je to číselná miera možnosti, že nastane náhodný jav.

$$P(A) = \frac{m}{n}$$

kde: A = jav, ktorého pravdepodobnosť sa skúma,

m = počet relevantných (priaznivých) prípadov (výsledkov pokusu), že nastane jav A,

n = počet všetkých možných prípadov (všetkým možných výsledkov pokusu).

Príklad 1.: Hod kockou

Pri hode nepoškodenou kockou z dostatočnej výšky na rovnú pevnú plochu existuje rovnaká možnosť, že padne ktorékoľvek číslo od 1 do 6. Pravdepodobnosť jedného výsledku pokusu (napr. že padne číslo 6) zo všetkých možných možností je vyjadrená klasickou pravdepodobnosťou pomerom $P(A)$=1/6=0,16667.

Príklad 2.: Športka

Akú možnosť (šancu) uhádnutia 5 čísel má hráč, ktorý na tikete *Športky* označil 6 čísel z ponúkaných 49?

Existuje 258 spôsobov ako uhádnuť práve 5 čísel zo 6 bez ohľadu na poradie a pritom neuhádnuť práve jedno číslo z ostávajúcich 43. Exaktne je to vyjadrené súčinom kombinačných čísel:

$$\binom{6}{5}\binom{43}{1} = \frac{6!}{5!\,(6-5)!}43 = 258$$

Teda 258 je priaznivých prípadov.

Vyjadrime relatívnu početnosť pozitívnych prípadov medzi všetkými možnými prípadmi, ktorých je toľko, koľko šestíc sa dá vytvoriť zo 49 čísel:

$$\binom{49}{6} = \frac{49!}{6!\,(49-6)!} = 13983816$$

Relatívna početnosť priaznivých prípadov medzi všetkými možnými je teda:

$$\frac{258}{13983816} = 0{,}0000184$$

Toto číslo je zároveň pravdepodobnosťou výhry 5 čísel z piatich.

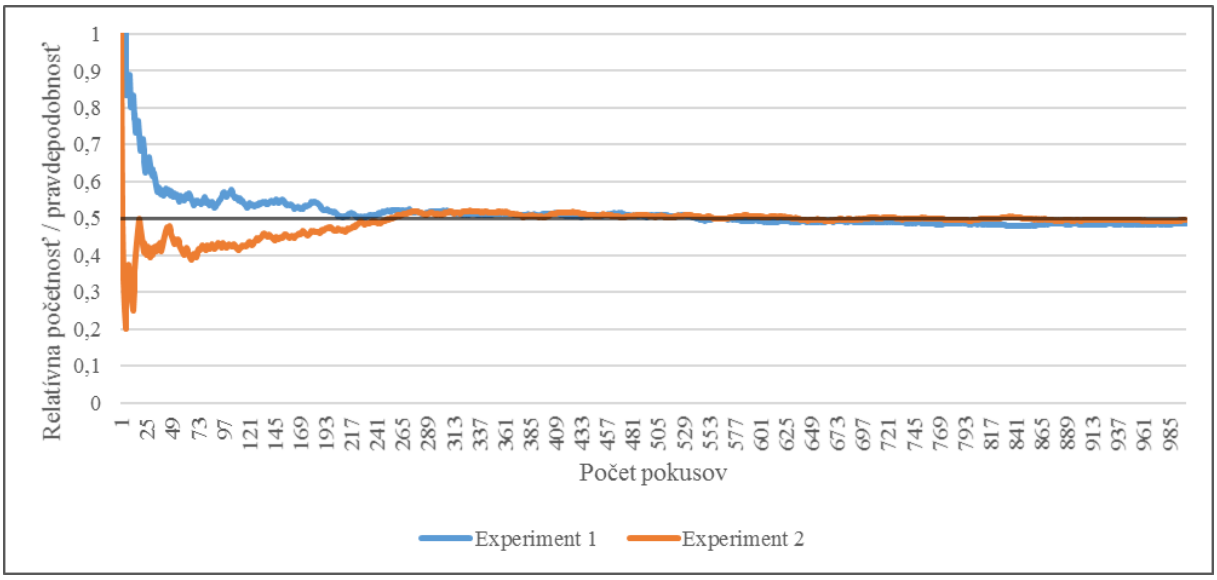
8.2.2 Štatistická definícia pravdepodobnosti (Richard von Mises)

Táto definícia je spojená s pojmom relatívnej početnosti, kedy pri malom počte pokusov má relatívna početnosť náhodný charakter. S rastúcim počtom pokusov sa však stabilizuje a približuje k určitému číslu (pravdepodobnosti). Je to akési upresnenie klasickej definície. Pri mnohonásobnom opakovaní "pokusu" sa relatívna početnosť javu blíži ku konkrétnej hodnote, pravdepodobnosti, limite; t.j. k pomeru počtu priaznivých prípadov/počet všetkých možných prípadov.

$$P(A) = \lim_{n\to\infty} \frac{m}{n}$$

Príklad: Hod mincou

Pri hode nepoškodenou mincou z dostatočnej výšky na rovnú pevnú plochu existuje rovnaká možnosť, že padne ktorákoľvek z dvoch strán mince. Avšak pri experimente či to naozaj tak je, sa môže stať, že viackrát padne jedna strana a pri použití klasickej pravdepodobnosti by sa mohlo zdať, že pravdepodobnosť je naklonená na jednu alebo druhú stranu mince. Predpoklad, že existuje rovnaká možnosť, že padne ktorákoľvek z dvoch strán mince platí, aj ak sa tento experiment za rovnakých podmienok opakuje. V tomto prípade sa rovnaká možnosť padnutia každej strany prejaví (približne) rovnakou absolútnou aj relatívnou početnosťou padnutia každej strany mince. Očakávame, že pri rastúcom počte pokusov sa bude relatívny počet ustaľovať na hodnote 0,5.



Obrázok 21 Pravdepodobnostné zobrazenie padnutia „znaku“ na minci pri 1000 pokusoch pre 2 individuálne experimenty

8.2.3 *Pravdepodobnosť ako miera dôvery (Thomas Bayes)*

Táto, na prvý pohlaď veľmi nevedecká, definícia hovorí, že **pravdepodobnosť je číslo medzi 0 a 1, ktorá je mierou pre našu vieru v realizáciu** (vznik) nejakého javu alebo viera v pravdivosť nejakého tvrdenia.

Pod realizáciou javu tu môžeme mať na mysli napríklad výhru domácich vo futbale, pod pravdivosťou nejakého tvrdenia zase pravdepodobnosť, že hmotnosť Saturnu sa nachádza v nejakom dopredu zvolenom intervale. Zvlášť druhá možnosť je pri fyzikálnom výskume častá (experimentálne hľadáme hodnoty rôznych konštánt, hmotností elementárnych častíc, atď.). So Saturnom totiž nemôžeme urobiť viac pokusov - jeho hmotnosť je daná a v danom intervale buď leží, alebo neleží - problém je iba v tom, že my nepoznáme odpoveď.

8.2.4 *Axiomatická definícia (Andrej Nikolajevič Kolmogorov)*

Pravdepodobnosť P(A) náhodnej udalosti A je v tomto prípade reálna funkcia, ktorá každej náhodnej udalosti A priradí určité číslo P(A), pričom platia tieto axiómy:

- $P(A) \geq 0$
- $P(E)=1$, keď E je istá udalosť
- keď $A_1, A_2, \ldots, A_n$ je postupnosť disjunktných náhodných udalostí (nemôžu súčasne nastať). Pravdepodobnosť, že aspoň jedna z nich pri náhodnom pokuse nastane, sa teda rovná súčtu ich pravdepodobnosti:
- $P(A_1 \cup A_2 \cup \ldots \cup A_n \cup \ldots) = P(A_1) + P(A_2) + \ldots + P(A_n) + \ldots$

Vlastnosti:

- P leží vždy medzi 0 a 1 (resp. vyjadrené v percentách: medzi 0% a 100%),
- P javu nemožného je 0 (0%), P javu istého je 1 (100%),
- súčet jednotlivých P všetkých možných prípadov je 1 (100%),

- pravdepodobnosť, že nastane len *A*, alebo iba *B*, alebo *A* i *B* súčasne = *P*, že nastane *A* + *P*, že nastane *B* - *P*, že nastane *A* i *B* súčasne.

Z predchádzajúcich tvrdení je zrejmé, že pravdepodobnosť je možné vyjadriť štyrmi základnými spôsobmi:

- klasicky - ako číslo v intervale <0 – 1>,
- v percentách [%],
- pomerom, napr. 1 : 1,
- zlomkom, napr. ½.

Pravdepodobnosť, že pri hode mincou padne hlava je teda možné identicky vyjadriť ako **0,5 ~ 50% ~ 1:1~ ½.**

V prepojení na pravdepodobnosť je dôležité spomenúť pojem riziko. Riziko je pravdepodobnosť toho, že nastane iný než očakávaný výsledok. Riziko teda nemožno stotožňovať iba s hrozbou nejakej straty či neúspechu. Teória pravdepodobnosti má práve preto veľké využitie aj v analýze a hodnotení rizík.

8.2.5 Sebaklam hazardného hráča

Ako sebaklam hazardného hráča označujeme jeho **mylný dojem**, že šanca vyhrať (pravdepodobnosť výhry), **stúpa alebo klesá na základe predchádzajúcich výsledkov**.

V lotérii Superlotto je potrebné uhádnuť 6 čísel z 51 pre maximálnu výhru. Vyhrávate aj keď uhádnete iba 5, 4 alebo aj 3 čísla. Samozrejme nižšiu sumu. Znie to jednoducho, ale aká je pravdepodobnosť (šanca vyhrať), že sa to stane?

Pravdepodobnosť uhádnutia:

- všetkých 6 čísel je 1 ku 18 009 460,
- pre 5 z 6 je 1 k 66 702,
- pre 4 z 6 je 1 k 1 213, a
- pre 3 z 6 je 1 ku 63. Teda šanca niečo vyhrať je jedna k šesťdesiat.

Možno si ešte stále niekto myslí, že môže pomôcť šťastiu tým, že si zvolí čísla, ktoré neboli vybrané v minulých kolách, alebo čísla, ktoré vychádzajú častejšie. Práve v tejto chvíli tak prepadáte falošnej **ilúzii hazardného hráča. Pravdepodobnosť je vždy rovnaká, bez ohľadu na čísla, ktoré vyhrali v minulosti.**

Tejto ilúzii bežne prepadávajú hráči, ktorí napríklad, stavia na rulete na červenú, keď sa predtým trikrát objavila čierna. Pravdepodobnosť, že sa znova objaví čierna ostáva rovnaká bez ohľadu na to, aká farba sa objavila predtým.

8.3 Pravdepodobnosť a štatistika

Pravdepodobnosť sa vzťahuje v štatistických projektoch na očakávaný výskyt obmien / hodnôt:

- jedného štatistického znaku,
- dvoch štatistických znakov,

- troch a viacerých štatistických znakov.

Základná štatistická metóda na skúmanie pravdepodobnosti štatistických znakov je triedenie štatistických údajov uvedené v kapitole 6.

Ak sa hľadá pravdepodobnosť výskytu:

- jednotlivých obmien jedného štatistického znaku, používa sa ako základná štatistická metóda jednoduché triedenie,
- skupín (intervalov) hodnôt jedného štatistického znaku, používa sa ako základná štatistická metóda skupinové triedenie,
- kombinácie obmien dvoch štatistických znakov, používa sa ako základná štatistická metóda triedenie podľa dvoch štatistických znakov,
- troch a viacerých štatistických znakov, používa sa triedenie podľa viacerých znakov.

Na vyjadrenie pravdepodobnosti sa používa relatívna početnosť v pravdepodobnostnom tvare, ktorá sa môže pohybovať v intervale <0,1>. Pravdepodobnostné vyjadrenie relatívnej početnosti je bezrozmerné.

Hodnota vyjadruje, **s akou pravdepodobnosťou je možné očakávať výskyt štatistickej jednotky** s príslušnou obmenou/hodnotou štatistického znaku **v budúcnosti**.

Základnou podmienkou skúmania pravdepodobnosti je reprezentatívnosť štatistického súboru, použitého na jej zisťovanie (kapitola 5.1.2).

8.3.1 Pravdepodobnosť pri jednoduchom triedení

Jednoduché triedenie štatistických jednotiek je možné použiť na výpočet pravdepodobnosti výskytu obmien:

- kvalitatívnych (slovných) znakov,
- kvantitatívnych diskrétnych (nespojitých) číselných znakov s malou obmenou hodnôt štatistického znaku.

Triedenie sa uskutočňuje podľa jednotlivých obmien štatistického znaku. Na triedenie štatistického znaku používame triediacu tabuľku (napr. Tabuľka 4, Tabuľka 22), ktorej tvorba má štandardný postup, zásady a štruktúru ako už bolo uvedené v kapitole 6.1.1.

Príklad:

Príklad na určenie pravdepodobnosti výsledku hodnotenia študentov konkrétneho predmetu je uvedený nižšie (Tabuľka 22). Podmienkou správnej interpretácie je reprezentatívnosť výberového súboru 129 študentov.

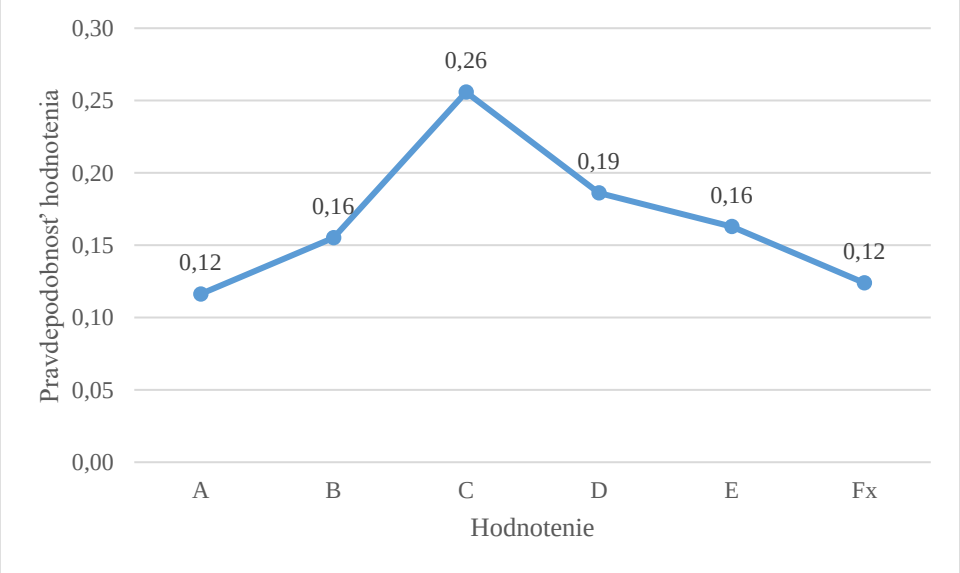
Tabuľka 22 Pravdepodobnosť výsledku hodnotenia študentov konkrétneho predmetu

Trieda	Triediaci znak	Absolútna početnosť	Relatívna početnosť	Kumulatívna absolútna početnosť	Kumulatívna relatívna početnosť
k	x_i	n_i	p_i	kn_i	kp_i

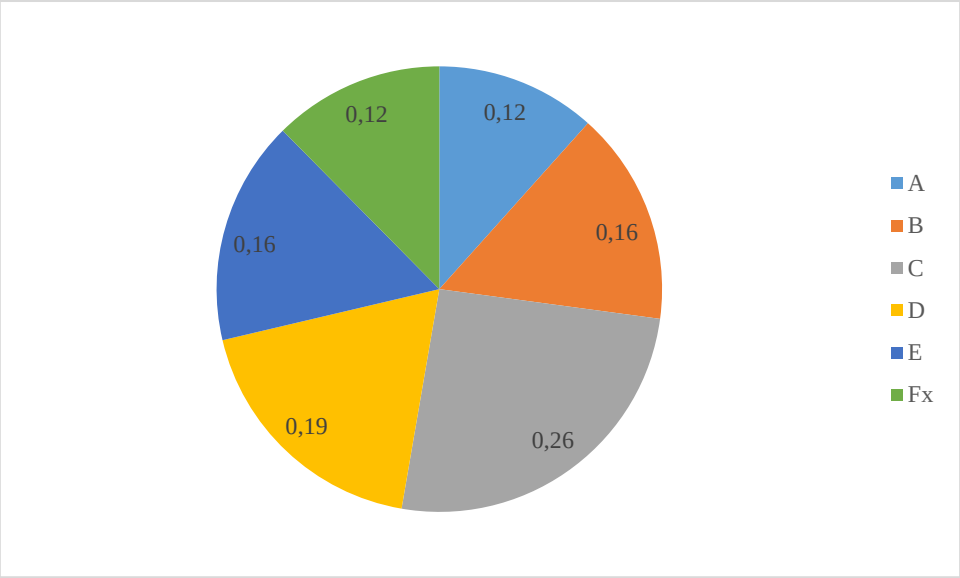
Poradové číslo	Hodnotenie	Počet študentov	Pravdepodobnosť hodnotenia	Súčtový počet študentov	Súčtová pravdepodobnosť hodnotenia
1	A	15	0,12	15	0,12
2	B	20	0,16	35	0,27
3	C	33	0,26	68	0,53
4	D	24	0,19	92	0,71
5	E	21	0,16	113	0,88
6	Fx	16	0,12	**129**	**1,00**
Spolu ∑	x	**129**	**1,0**	x	x

Tabuľková forma výsledkov sa pri zobrazení pravdepodobnosti doplňuje väčšinou:

- spojnicovým grafom relatívnych početností (Obrázok 22),
- koláčovým (kruhovým výsekovým) grafom relatívnych početností (Obrázok 23).



Obrázok 22 Pravdepodobnosť výsledného hodnotenia študenta pre konkrétny predmet (spojnicový graf)



Obrázok 23 Pravdepodobnosť výsledného hodnotenia študenta pre konkrétny predmet (koláčový graf)

Jednotlivé spojnice alebo výseky je vhodné doplniť konkrétnou hodnotou.

8.3.2 Pravdepodobnosť pri skupinovom (intervalovom) triedení

Skupinové (intervalové) triedenie štatistických jednotiek je možné použiť na výpočet pravdepodobnosti v prípade kvantitatívnych (číselných) štatistických znakov, ktoré majú spravidla viac ako 15 rozličných hodnôt.

Pôvodné údaje sa zaradzujú do tried (skupín, intervalov) a zisťuje sa absolútna početnosť výskytu štatistických jednotiek, ktorá je následne prepočítaná do podoby relatívnej početnosti v jednotlivých triedach a je vyjadrená pravdepodobnostne.

Príklad:

Pravdepodobnosť výskytu zamestnancov v jednotlivých príjmových skupinách je uvedená v nasledujúcej tabuľke (Tabuľka 23). Podmienkou správnej interpretácie by mala znova byť reprezentatívnosť výberu domácností do skúmanej vzorky.

Tabuľka 23 Pravdepodobnosť výskytu zamestnancov v jednotlivých príjmových skupinách

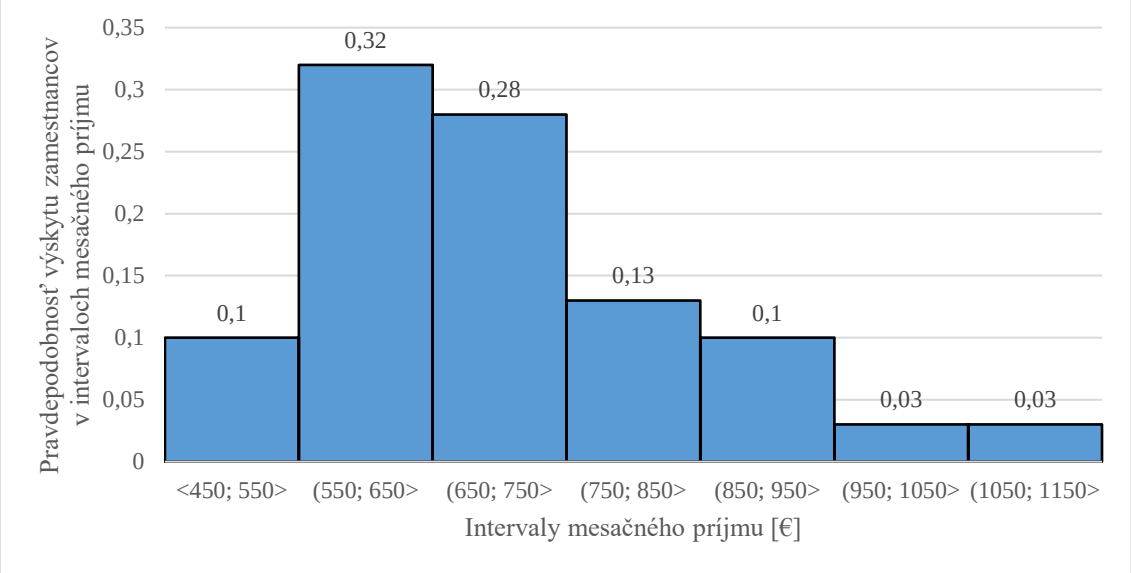
Trieda	Hranice mesačného príjmu [€]	Stred intervalu príjmu [€]	Počet zamestnancov	Pravdepodobnosť budúceho príjmu nových zamestnancov	Súčtová početnosť zame-stnancov	Súčtová pravdepodobnosť výskytu zamestnancov v intervaloch mesačného príjmu
k	$x_d - x_h$	x_i	n_i	P_i	kn_i	kP_i
1	<450; 550>	500	6	0,10	6	0,10
2	(550; 650>	600	19	0,32	25	0,42
3	(650; 750>	700	17	0,28	42	0,70
4	(750; 850>	800	8	0,13	50	0,83

5	(850; 950>	900	6	0,10	56	0,93
6	(950; 1050>	1000	2	0,03	58	0,97
7	(1050; 1150>	1100	2	0,03	60	1,00
∑	×	×	60	1	×	×

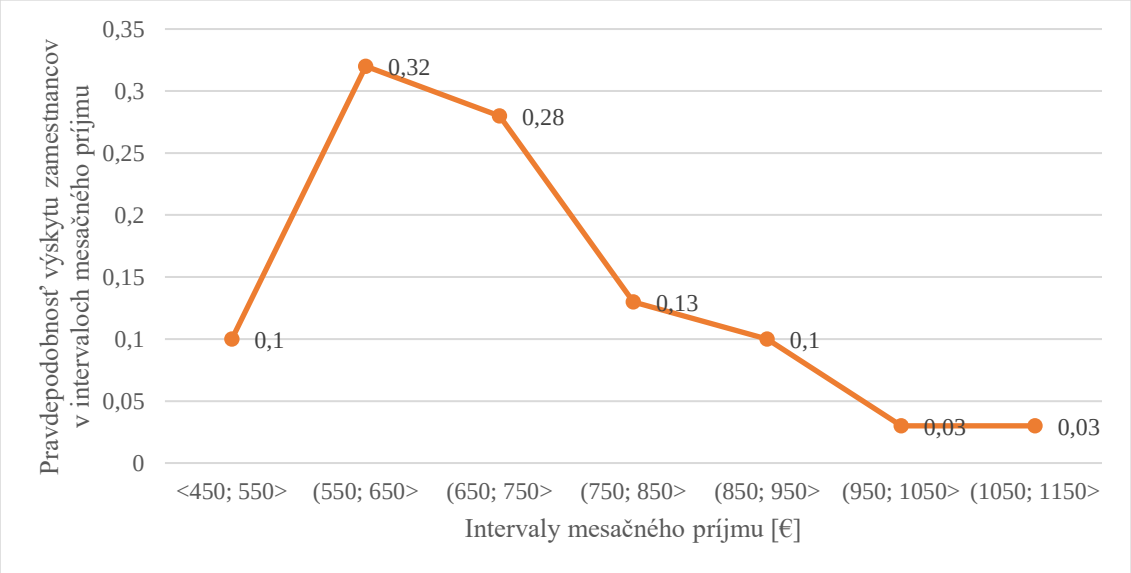
Tabuľková forma sa pri zobrazení pravdepodobnosti doplňuje:

- histogramom (Obrázok 24),
- bodovým grafom relatívnych početností (Obrázok 25),
- koláčovým (výsekovým) grafom relatívnych početností (Obrázok 26).

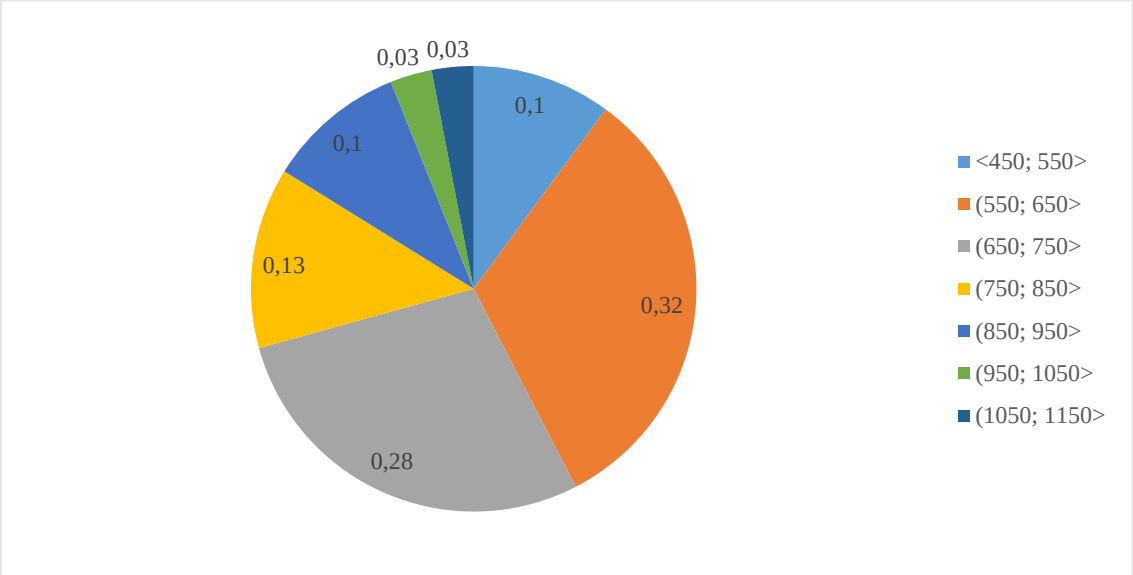
Histogram pravdepodobnosti je stĺpcový graf tvorený pravidelnými rovnobežníkmi, ktorý priraďuje príslušnú pravdepodobnosť celému intervalu (skupine). Základy stĺpcov na osi *x* majú dĺžku intervalov (šírku triedy) *h*, pre všetky rovnakú a príslušné výšky stĺpcov majú veľkosť zodpovedajúcu vypočítanej pravdepodobnosti.



Obrázok 24 Pravdepodobnosť výskytu zamestnancov v jednotlivých príjmových skupinách



Obrázok 25 Pravdepodobnosť výskytu zamestnancov v jednotlivých príjmových skupinách vztiahnutá na stredy príjmových skupín



Obrázok 26 Zobrazenie pravdepodobnosti výskytu zamestnancov v príjmových skupinách

8.3.3 Pravdepodobnosť pri triedení podľa dvoch štatistických znakov

Triedenie podľa dvoch štatistických znakov je možné použiť na výpočet pravdepodobnosti výskytu jednotlivých kombinácii ich obmien. Výsledkom triedenia sú kombinačné tabuľky (uvedené v kap. 6.2).

Príklad 1:

Príklad na pravdepodobnosť výskytu rodín podľa počtu detí a počtu miestností v byte je uvedený v nasledujúcej korelačnej tabuľke (Tabuľka 24) a vychádza z údajov príkladu na triedenie v kombinácii dvoch štatistických znakov (Tabuľka 14). Podmienkou správnej interpretácie by mala byť opäť reprezentatívnosť výberu rodín.

Tabuľka 24 Pravdepodobnosť výskytu rodín podľa počtu detí a bytových miestností

Počet detí	Počet obytných miestností				**Spolu**
	1	2	3	4	
0	0,10	0,12	0,08	-	0,30
1	-	0,18	0,12	-	0,30
2	-	-	0,20	0,10	0,30
3	-	-	-	0,10	0,10
Spolu	0,10	0,30	0,40	0,20	1,00

V prípade, že je počet obmien niektorého číselného (kvantitatívneho) štatistického znaku veľký, musia byť konkrétne obmeny **nahradené skupinami** (intervalmi). Tvorba intervalov je identická ako u skupinového triedenia.

Príklad 2:

Aká je pravdepodobnosť vykradnutia domu/bytu v obci resp. v meste vzhľadom na typ zabezpečenia? Tento príklad je uvedený v nasledujúcich kontingenčných tabuľkách (Tabuľka 25, Tabuľka 26).

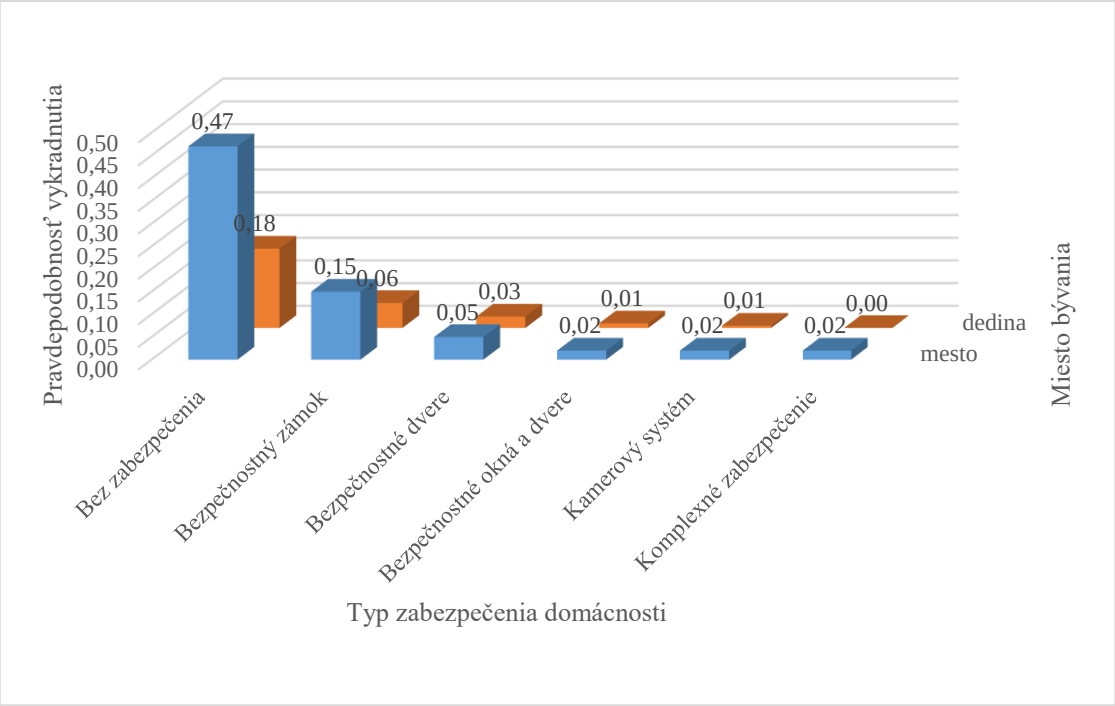
Tabuľka 25 Počty vykradnutia domácnosti vzhľadom na miesto bývania a typ zabezpečenia domácnosti

Typ zabezpečenia	Miesto bývania		Spolu
	mesto	dedina	
Bez zabezpečenia	94	35	129
Bezpečnostný zámok	30	11	41
Bezpečnostné dvere	10	5	15
Bezpečnostné okná a dvere	4	2	6
Kamerový systém	4	1	5
Komplexné zabezpečenie	4	0	4
Spolu	146	54	200

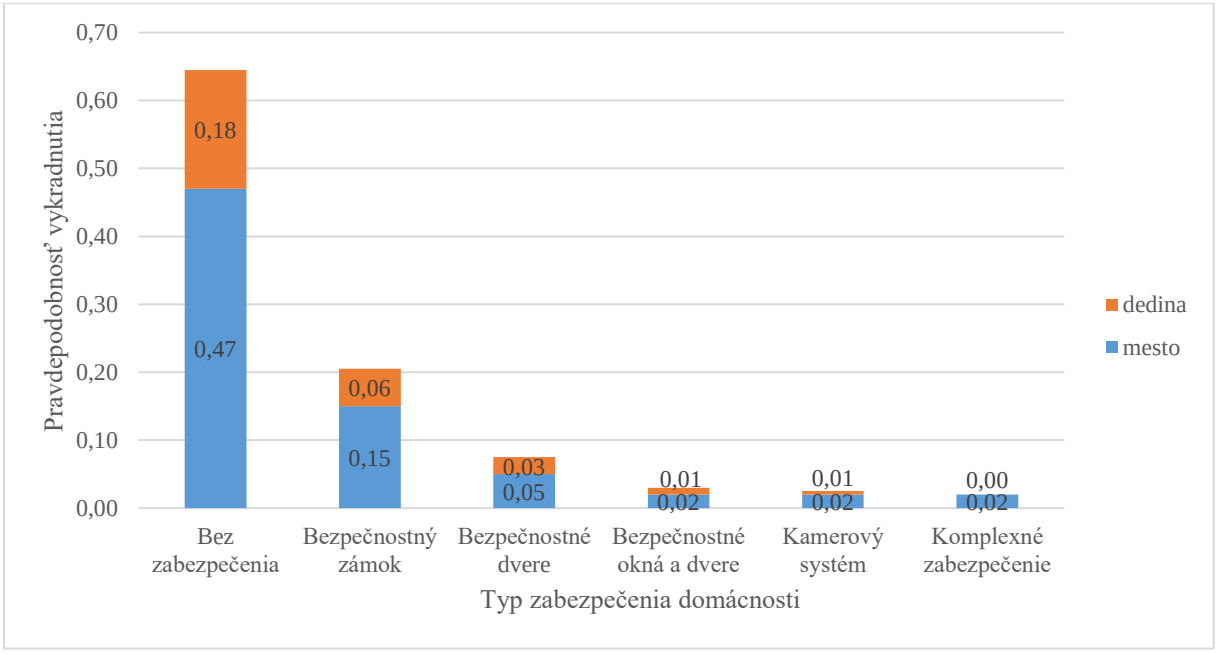
Tabuľka 26 Pravdepodobnosť vykradnutia domácnosti vzhľadom na miesto bývania a typ zabezpečenia domácnosti

Typ zabezpečenia	Miesto bývania		Spolu
	mesto	dedina	
Bez zabezpečenia	0,47	0,18	0,65
Bezpečnostný zámok	0,15	0,06	0,21
Bezpečnostné dvere	0,05	0,03	0,08
Bezpečnostné okná a dvere	0,02	0,01	0,03
Kamerový systém	0,02	0,01	0,03
Komplexné zabezpečenie	0,02	0,00	0,02
Spolu	0,73	0,27	1,00

Na zobrazenie pravdepodobnosti je praktické použiť stĺpcový pseudo 3D graf (Obrázok 27) alebo skladaný stĺpcový graf (Obrázok 28).



Obrázok 27 Pravdepodobnosť vykradnutia domácnosti vzhľadom na miesto bývania a typ zabezpečenia domácnosti (priestorový graf)



Obrázok 28 Pravdepodobnosť vykradnutia domácnosti vzhľadom na miesto bývania a typ zabezpečenia domácnosti (skladaný graf)

8.4 Rozdelenie pravdepodobnosti v štatistike

Jednou z úloh štatistiky je odhad (výpočet) hodnôt štatistického znaku x_i, ktoré sa nachádzajú:

- **medzi hodnotami získanými štatistickým zisťovaním**, alebo sa
- nachádzajú **mimo variačné rozpätie R_V** štatistického súboru.

Tato úloha je typická v prípadoch, kedy by skúmanie všetkých štatistických jednotiek bolo zdĺhavé, neekonomické alebo fyzicky nemožné. Zrejmé je to napríklad z frekvenčnej tabuľky (Tabuľka 27) pre skupinové (intervalové) triedenie mesačných príjmov zamestnaných ľudí v konkrétnom okrese. Všetkých zamestnancov v okrese je podstatne viac, ale na vzorke respondentov (n = 60) získame základný prehľad aj o rozdelení ostatných zamestnaných ľudí podľa ich príjmu.

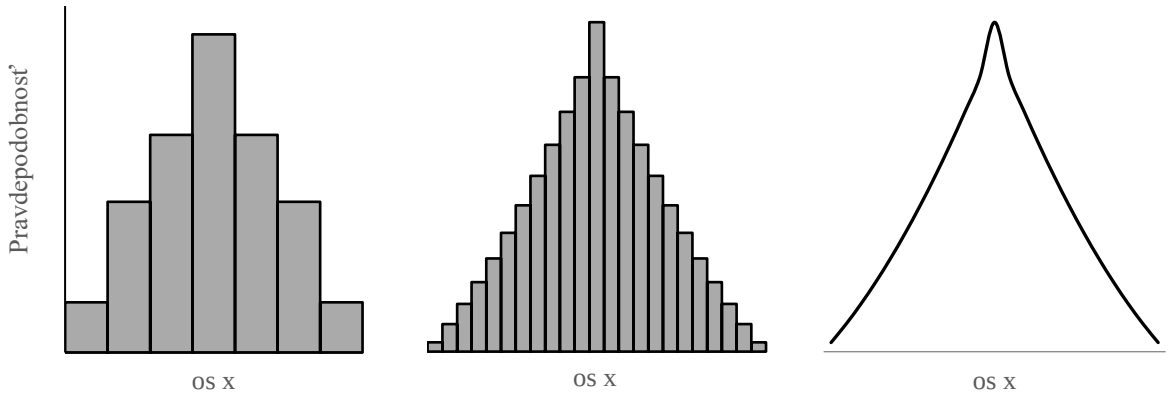
Tabuľka 27 Pravdepodobnosť budúceho príjmu nových zamestnancov

Trieda	Hranice mesačného príjmu [€]	Stred intervalu príjmu [€]	Počet zame-stnancov	Pravdepodobnosť budúceho príjmu nových zamestnancov	Súčtová početnosť zamestnancov	Súčtová pravdepodobnosť budúceho príjmu nových zamestnancov
k	x_d - x_h	x_i	n_i	P_i	kn_i	kP_i
1	(do 550>	500	6	0,10	6	0,10
2	(550; 650>	600	19	0,32	25	0,42
3	(650; 750>	700	17	0,28	42	0,70
4	(750; 850>	800	8	0,13	50	0,83
5	(850; 950>	900	6	0,10	56	0,93
6	(950; 1050>	1000	2	0,03	58	0,97
7	(1050 a viac)	1100	2	0,03	60	1,00
∑	×	×	60	100	×	×

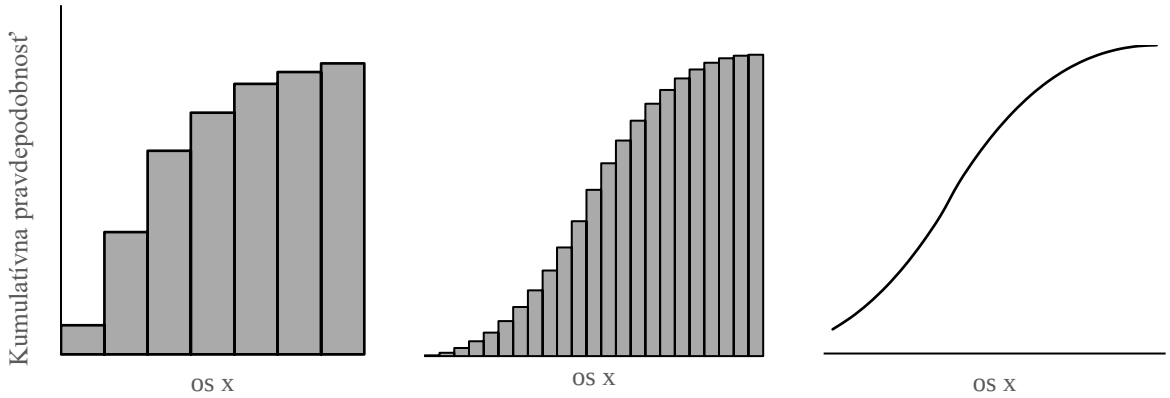
Úlohami môžu byť v tomto prípade napríklad:

- odhad (výpočet) relatívnej početnosti pre príjem 350 €,- alebo 1 500 €,- na jednej alebo druhej strane otvorených intervalov,
- odhad (výpočet) počtu zamestnancov a ich príjmov vyskytujúcich sa s relatívnou početnosťou (pravdepodobnosťou) na úrovni 0,50.

Najjednoduchšia možnosť ako vyriešiť vyššie uvedený prípad by bolo vykonať aproximáciu pre najbližšie sa vyskytujúce hodnoty zľava a sprava. Lepšie riešenie tejto úlohy vychádza zo štatistickej definície pravdepodobnosti, ktorá považuje relatívnu početnosť p_i zároveň za pravdepodobnosť výskytu príslušnej hodnoty štatistického znaku. **Riešenie** tejto úlohy bude spočívať v hľadaní takých funkcií, ktoré by dokázali s dostatočnou presnosťou **popísať priebeh relatívnej početnosti** hodnôt štatistického znaku. Jedná sa o nájdenie tzv. **funkcie rozdelenia pravdepodobnosti náhodnej premennej (náhodnej veličiny),** ktorou je výskyt (početnosti) hodnôt štatistického znaku (Obrázok 29, Obrázok 30). Viac v kapitole 8.4.2.



Obrázok 29 Prechod od histogramu relatívnych početností k frekvenčnej funkcii



Obrázok 30 Prechod od kumulatívnej početnosti k distribučnej funkcii

8.4.1 Diskrétna a spojitá náhodná veličina

Náhodná veličina (náhodná premenná) je taká veličina, ktorá vplyvom náhodných (stochastických) okolností **nadobúda vždy jednu z množstva možných hodnôt**.

Poznáme:

- **spojitú náhodnú veličinu** – tá môže nadobúdať ľubovoľné reálne hodnoty,
- **diskrétnu náhodnú veličinu** – tá nadobúda iba izolované hodnoty.

Pre **diskrétnu** náhodnú veličinu je typická izolovanosť jej hodnôt. Tato veličina dokonca **často nadobúda hodnoty z oboru celých alebo prirodzených čísel** (počet chybných výrobkov vo výrobnej sérii, počet porúch zariadení v určitom časovom intervale, a pod.).

8.4.2 Rozdelenie pravdepodobnosti náhodnej veličiny

O vzťahu medzi hodnotami štatistického znaku (diskrétneho alebo spojitého) a ich početnosťou nás informuje **rozdelenie štatistického súboru** (viď viac v kap. 8.5 a 8.6). Funkcia, ktorá každej hodnote náhodnej veličiny priraďuje príslušnú pravdepodobnosť, sa nazýva **rozdelenie pravdepodobnosti**, tiež **zákon rozdelenia pravdepodobnosti**.

Pravdepodobnostné správanie sa náhodných veličín je možné popísať mnohými spôsobmi. Najobvyklejšími sú:

- **popis frekvenčnej (pravdepodobnostnej) funkcie** alebo **funkcie hustoty pravdepodobnosti**, ktorej tvar podáva obraz o dôležitých vlastnostiach rozdelenia,
- **popis distribučnej funkcie**.

Oba spôsoby popisu charakterizujú rozdelenie náhodných veličín úplne, teda napr. ak majú dve veličiny rovnaké distribučné funkcie, majú aj rovnaké rozdelenie a naopak.

Zákon rozdelenia pravdepodobnosti je teda funkcia ***p(x)***, ktorá priraďuje každej hodnote x_i diskrétnej náhodnej veličiny X príslušnú pravdepodobnosť $p_i = p_{x_i}$. Pritom súčet pravdepodobnosti musí byť:

- pre konečný počet hodnôt x_i

$$\sum_{i=1}^{n} p_i = 1$$

- pre nekonečný počet hodnôt x_i

$$\sum_{i=1}^{\infty} p_i = 1$$

Pri **diskrétnych náhodných veličinách** je možné túto pravdepodobnosť **priradiť každej konkrétnej hodnote x_i.**

Pri **spojitých náhodných veličinách** je možné túto pravdepodobnosť **priradiť iba určitému intervalu nenulovej dĺžky.**

Rozdelenie pravdepodobnosti náhodnej veličiny je možné formálne prezentovať:

- tabuľkou,
- vzorcom,
- graficky.

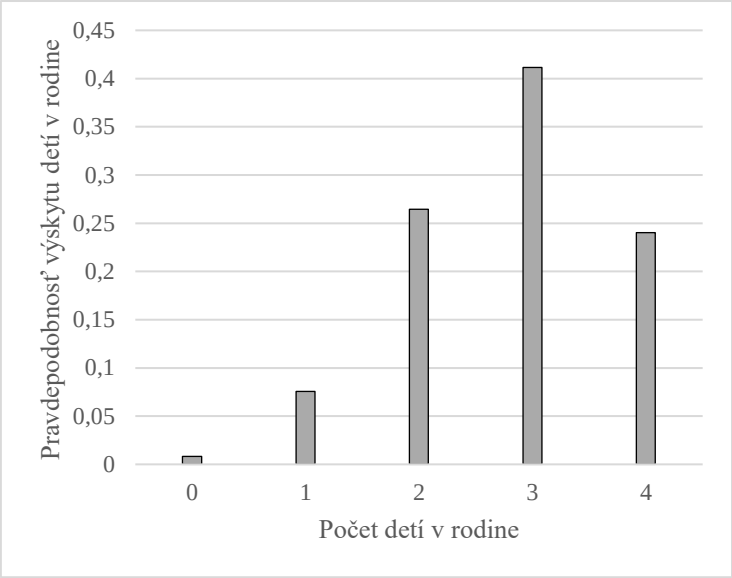
8.4.3 Frekvenčná funkcia

Príklad na diskrétnu náhodnú veličinu:

Diskrétnu náhodnú veličinu môže predstavovať počet detí zistených na skúmanej vzorke rodín. Na základe vykonaného štatistického zisťovania riešiteľ získa zjednodušenú frekvenčnú tabuľku (Tabuľka 28) a hodnoty frekvenčnej funkcie.

Tabuľka 28 Frekvenčná tabuľka počtu rodín podľa počtu detí v rodine

Počet detí x_i	0	1	2	3	4	Celkom
Relatívna početnosť = Pravdepodobnosť *p(x)*	0,0081	0,0756	0,2646	0,4116	0,2401	1



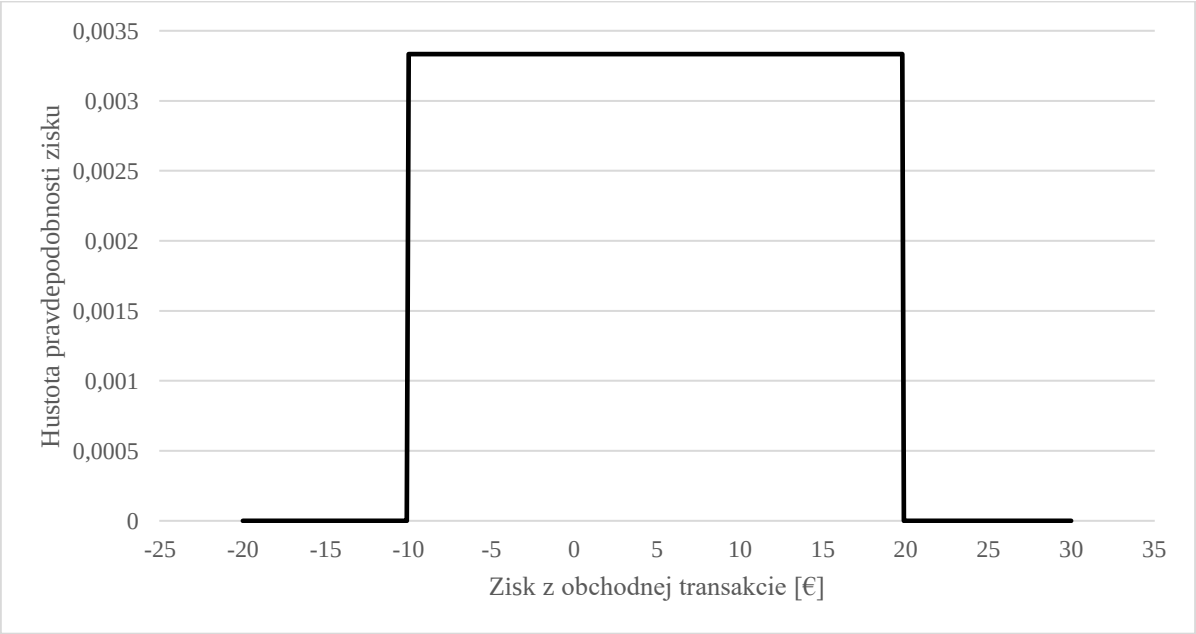
Obrázok 31 Rozdelenie pravdepodobnosti počtu detí v rodine

Príklad na spojitú náhodnú veličinu:

Považujme zisk z obchodnej transakcie za spojitú náhodnú veličinu ***X***, ktorej výskyt na intervale < -10 (t. j. strata 10 tis. €) až + 20 (zisk 20 tis. €)> je **rovnako možný - rovnomerný**.

Tabuľkové vyjadrenie v tomto prípade neprichádza do úvahy, pretože reálnych čísel je v intervale <-10;+20> nekonečne veľa. Ak sa vyjadrí aká časť jednotkovej pravdepodobnosti istého javu pripadá na jednotku intervalu možných hodnôt spojitej veličiny, potom je možné hovoriť o hustote pravdepodobnosti (Obrázok 32).

V uvedenom prípade (rovnako možný výskyt) $f(x) = \begin{cases} \frac{1}{30} \text{ pre } -10 \leq x \leq 20 \\ 0 \end{cases}$



Obrázok 32 Rozdelenie hustoty pravdepodobnosti zisku

Pre spojitú náhodnú veličinu je taktiež možné použiť príklad uvedený na začiatku tejto kapitoly (Obrázok 29).

8.4.4 Distribučná funkcia

Distribučná funkcia náhodnej veličiny (diskrétnej či spojitej) je funkcia $\boldsymbol{F(x) = P(X \leq x)}$, **t. j. pravdepodobnosť, že náhodná veličina nepresiahne ľubovoľnú hodnotu** x **z oboru možných hodnôt náhodnej veličiny.**

Príklad na diskrétnu náhodnú veličinu:

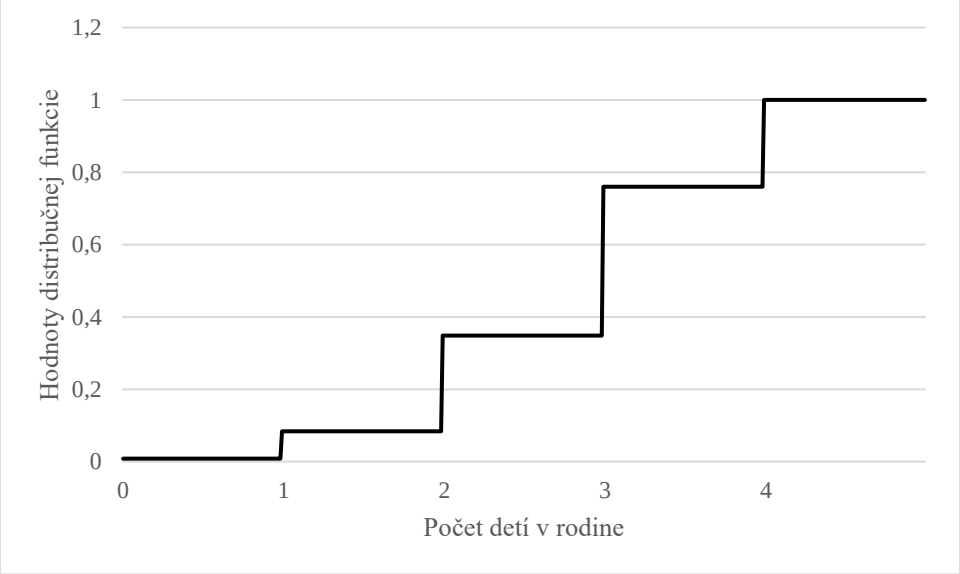
Do zjednodušenej frekvenčnej tabuľky (Tabuľka 29) sa **doplní kumulatívna relatívna početnosť**, ktorú je možné nazvať aj súčtovou pravdepodobnosťou a predstavuje vlastné **hodnoty distribučnej funkcie.**

Tabuľka 29 Relatívna a kumulatívna relatívna početnosť rodín podľa počtu detí v rodine

Počet detí x	0	1	2	3	4
Relatívna početnosť = Pravdepodobnosť P(x)	0,0081	0,0756	0,2646	0,4116	0,2401
Kumulatívna relatívna početnosť = Súčtová pravdepodobnosť F(x)	0,0081	0,0837	0,3483	0,7599	1,0000

Tvar distribučnej funkcie:

$$F(x) = \sum \binom{4}{x} 0{,}7^x 0{,}3^{4-x} \; pre \; x = 0{,}1, \ldots, 4$$



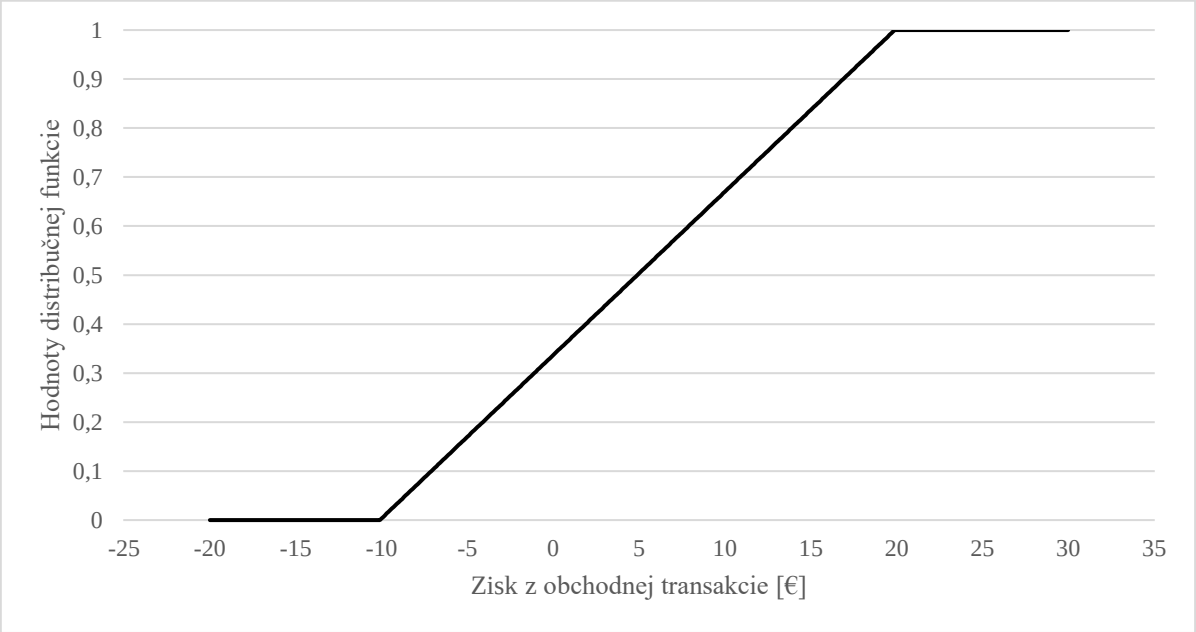
Obrázok 33 Distribučná funkcia pre pravdepodobnosť počtu detí v rodine

U diskrétnej náhodnej veličiny je distribučná funkcia definovaná ako kumulatívny súčet hodnôt funkcie pravdepodobnosti.

Príklad na spojitú náhodnú veličinu:

Distribučná funkcie prechádza bodmi [-10; 0] a [20; 1], medzi ktorými lineárne rastie, a teda:

$$F(x) = \begin{cases} 0 & pre\ x < -10 \\ \frac{x+10}{30} & pre -10 \leq x \leq 20 \\ 1 & pre\ x > 20 \end{cases}$$



Obrázok 34 Distribučná funkcia pravdepodobnosti zisku

U spojitej náhodnej veličiny reprezentuje distribučná funkcia plochu pod čarou hustoty pravdepodobnosti, ktorá je sprava ohraničená hodnotou *x*.

Vlastnosti distribučnej funkcie:

- distribučná funkcia je pravdepodobnosť, ktorej hodnoty sa pohybujú v intervale $0 \leq F(x) \leq 1$,
- distribučná funkcie je neklesajúca (pravdepodobnosť je nezáporná), a teda pre $x_1 < x_2$ bude $F(x_2) \geq F(x_1)$ a $P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1)$,
- $F(-\infty) = 0; F(+\infty) = 1$
- pre diskrétnu náhodnú veličinu je distribučná funkcia funkciou sprava spojitou, pre spojitú náhodnú veličinu je distribučná funkcia funkciou spojitou na intervale možných hodnôt náhodnej veličiny.

Vlastnosti hustoty pravdepodobnosti

Hustota pravdepodobnosti je deriváciou neklesajúcej funkcie a ako taká:

- $f(x) \geq 0$ (hustota pravdepodobnosti nie je pravdepodobnosť!),
- plocha pod čiarou hustoty pravdepodobnosti je pravdepodobnosť istého javu, a teda:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

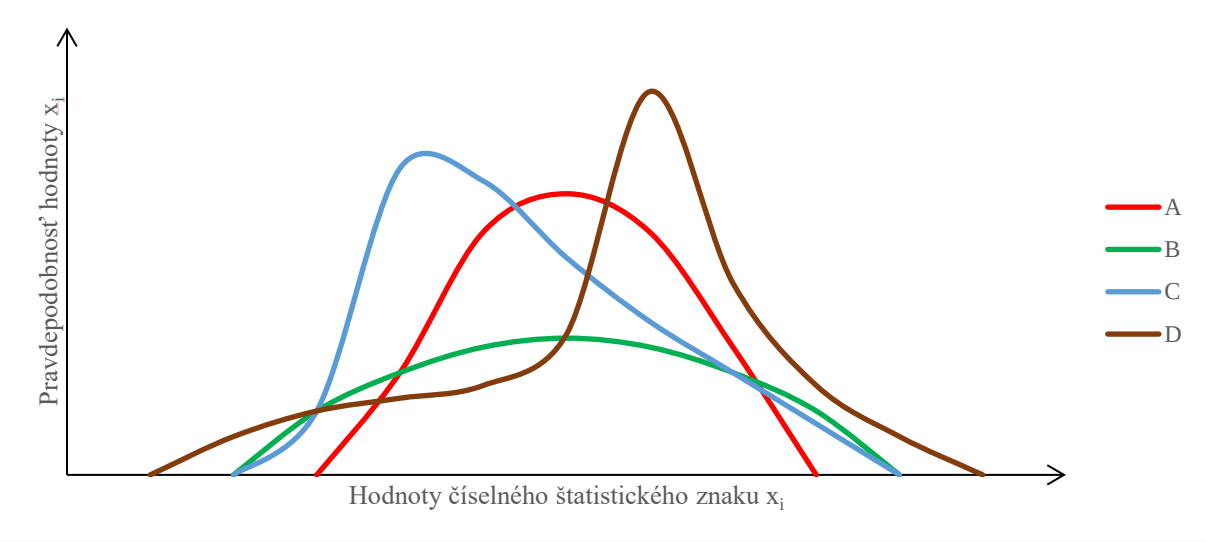
- pre $P(x_1 \leq X \leq x_2)$ platí:

$$\int_{x_1}^{x_2} f(x)dx$$

- pre dostatočne malé Δx je $P(x \leq X \leq X + \Delta x) = f(x) \cdot \Delta x$. Tento súčin (plocha obdĺžnika pod čiarou hustoty) sa nazýva pravdepodobnostný element.

Distribučná funkcia, pravdepodobnostná funkcia a hustota pravdepodobnosti popisujú rozdelenie pravdepodobnosti ľubovoľnej náhodnej veličiny úplne jednoznačne. Konštanty v rovniciach týchto funkcií sa nazývajú **parametre rozdelenia.**

Väčšina náhodných dejov, ale nemá rozloženie pravdepodobnosti rovnaké v celom rozsahu skúmaného intervalu ako v príklade zisku z obchodnej transakcie. **Najväčšiu pravdepodobnosť majú väčšinou hodnoty vyskytujúce sa okolo charakteristík úrovne** (polohy) – **mediánu, modusu a priemeru**. Grafické vyjadrenie pravdepodobnostnej funkcie alebo funkcie hustoty pravdepodobnosti má potom rôzny, väčšinou zvonovitý tvar, niekedy zošikmený sprava alebo zľava.



Obrázok 35 Rôzne tvary rozdelenia pravdepodobnosti hodnôt štatistického znaku

8.4.5 *Aproximácia reálnych náhodných veličín*

Rozdelenia niektorých náhodných veličín sú získavané na **základe teoretických znalostí** o týchto veličinách. Pre väčšinu fyzikálnych veličín nie sú typy a parametre rozdelenia dostatočne presne známe, a preto sú k ich určeniu používané **empirické údaje získané na základe merania, testov alebo skúšok.**

Toto empirické rozdelenie býva **aproximované vhodným teoretickým rozdelením**, ktorého typ býva zvolený buď na základe znalostí teoretických princípov vedúcich k zvoleniu rozdelenia alebo na základe vlastností vzorky získaných údajov.

V situáciách, kedy nie je k dispozícii dostatočne rozsiahla vzorka, je **treba typ rozdelenia odhadnúť**. Pri voľbe typu rozdelenia hrá rolu aj jeho predpokladané budúce použitie. Preto

býva často použité **normálne rozdelenie**, ktorého vlastnosti sú dobre známe (viď nasledujúcu kapitolu).

8.5 Základné typy rozdelenia diskrétnej náhodnej veličiny

Diskrétne rozdelenie pravdepodobnosti sa používa vtedy, ak pre všetky hodnoty náhodnej premennej X je možné určiť pravdepodobnosti $P(X = x_i) = p(x_i)$, a súčasne platí:

$$\sum_{i=1}^{n} p(x_i) = 1$$

Rozoznávame viac typov rozdelenia diskrétnej náhodnej premennej. Medzi základné typy patrí **alternatívne, rovnomerné, binomické, Poisssonovo, geometrické, hypergeometrické,** atď.

8.5.1 Alternatívne rozdelenie A(p)

Alternatívne rozdelenie, inak nazývané aj Bernoulliho rozdelenie, je založené na princípe, že niektoré náhodné pokusy môžu mať iba dva rôzne výsledky:

- pokus je úspešný,
- pokus je neúspešný.

Príslušná náhodná veličina ***X*** sa potom nazýva alternatívna (dvojdobá, nulovo-jednotková).

Tato náhodná veličina teda nadobúda iba dve hodnoty:

- 1 – v prípade priaznivého výsledku pokusu (nastane jav ***A***),
- 0 – v prípade nepriaznivého výsledku pokusu.

Obor hodnôt teda obsahuje dva prvky M = {0,1}. Používa sa označenie:

$$P(A) = P(X = 1) = p$$

$$P(A) = P(X = 0) = 1 - p$$

Definícia

$$X \sim A(p)$$

Náhodná veličina ***X*** s pravdepodobnostnou funkciou $P(X) = 0$, $P(X = 1) = p$ $(0 < p < 1)$ má alternatívne rozdelenie pravdepodobnosti ***A(p)*** s parametrom ***p***.

Príkladom môže byť skúmanie pravdepodobnosti výskytu konkrétnej farby kvetu pri krížení dvoch farieb (červené ruže, biele ruže). Ak sa podľa genetických zákonov očakáva, že väčšina bude červených lebo červený gén prevláda (napr. nech to platí pre 75% prípadov). Aká je napr. pravdepodobnosť, že pri ďalších 100 nových krížených rastlinách nebude asi jedna ruža biela.

8.5.2 Rovnomerné rozdelenie Ro(m)

Rovnomerné rozdelenie pravdepodobnosti priraďuje všetkým hodnotám náhodnej veličiny **rovnakú pravdepodobnosť**. Rovnomerné rozdelenie má svoju diskrétnu aj spojitú podobu. Rovnomerné rozdelenie predstavuje najjednoduchší prípad diskrétneho rozdelenia.

Diskrétne rovnomerné rozdelenie popisuje náhodnú veličinu, ktorá môže nadobúdať ***n*** hodnôt s rovnakou pravdepodobnosť, pričom sa predpokladá, že vzdialenosti medzi jednotlivými hodnotami náhodnej veličiny sú rovnaké (*1/n*).

Definícia

$$X \sim Ro(m)$$

Nech náhodná premenná ***X*** nadobúda hodnoty *x* = 1,2,..., *m* s pravdepodobnosťami $P(X = x) = \frac{1}{m}$ pre x = 1,2,...,m. Hovoríme, že náhodná premenná **X** má rovnomerné diskrétne rozdelenie s parametrom ***n***.

Typickým príkladom rovnomerného rozdelenia je hod kockou, kedy pravdepodobnosť padnutia každého z čísel je 1/6.

8.5.3 Binomické rozdelenie Bi(n, p)

Predpoklady vzniku náhodnej veličiny s binomickým rozdelením:

- pravdepodobnosť výskytu javu ***A*** v jedinom pokuse $P(A) = p$,
- je uskutočnených ***n*** pokusov,
- pokusy sú nezávislé, t. j. pravdepodobnosť výskytu javu ***A*** v dvoch pokusoch je $p \cdot p = p^2$, v troch pokusoch $p \cdot p \cdot p = p^3$, atď.

Pritom počet pokusov ***n*** nesmie byť príliš veľký a pravdepodobnosť ***p*** nie je blízka nule ani jednotke. Vo všetkých uvedených prípadoch diskrétnej náhodnej veličiny X nadobúda hodnoty *x=0, 1, 2, ..., n.*

Definícia

$$X \sim Bi(n, p)$$

Náhodná veličina ***X*** má binomické rozdelenie **Bi (n, p)** práve vtedy, keď má pravdepodobnostná funkcia rovnicu:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

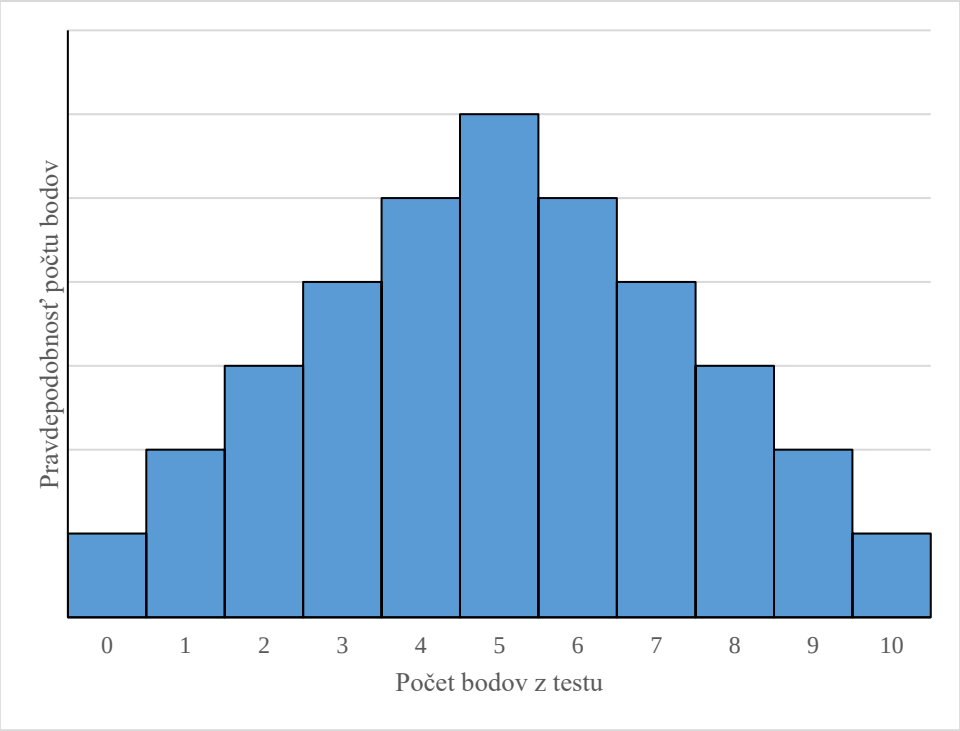
kde *x* = 0,1,...,*n*

n je počet pokusov

p je pravdepodobnosť úspešnosti v každom pokuse

Príklad

Študenti dostali test s desiatimi otázkami s možnosťami odpovedí iba „áno“ „nie“, ktorý bol nečakaný a na ktorý sa nikto nepripravoval. Študenti teda hádali správne odpovede. Aká je pravdepodobnosť, že uhádnu určitý počet otázok (napr. 2 alebo 7)?



Obrázok 36 Grafické znázornenie hustoty pravdepodobnosti binomického rozdelenia (pravdepodobnosť počtu bodov z náhodného testu)

Klasickými príkladmi pre binomické rozdelenie pravdepodobnosti sú aj nezávislé opakované pokusy tzv. výber s opakovaním, kedy vybraná štatistická jednotka je pred ďalším ťahom vrátená a zamiešaná medzi ostatné.

Príklady ďalších diskrétnych náhodných veličín s binomickým rozdelením:

- počet úspešných (neúspešných) zásahov pri *n* výstreloch do terča,
- počet vyrobených dobrých (zlých) výrobkov pri výrobe *n* kusov,
- počet chlapcov (dievčat) v rodine s *n* deťmi,
- počet ziskových (stratových) investícií medzi *n* investíciami, a iné.

8.5.4 Poissonovo rozdelenie Po(λ)

Popisuje náhodné rozdelenie objektov (udalostí) v jednotke priestoru či času, t. j. také, že každý bod v priestore (čase) má rovnakú pravdepodobnosť, že môže obsahovať daný objekt a výskyt objektu v danom bode nemá žiadny vplyv na výskyt akéhokoľvek objektu v rovnakom či akomkoľvek inom bode priestoru (času).

Základné vlastnosti Poissonovho rozdelenia:

- nezávislosť,
- jednotlivosť,
- homogenita.

Definícia

$$X \sim Po(\lambda)$$

Náhodná veličina ***X*** má Poissonovo rozdelenie ***Po(λ)*** práve vtedy, keď má pravdepodobnostná funkcia rovnicu

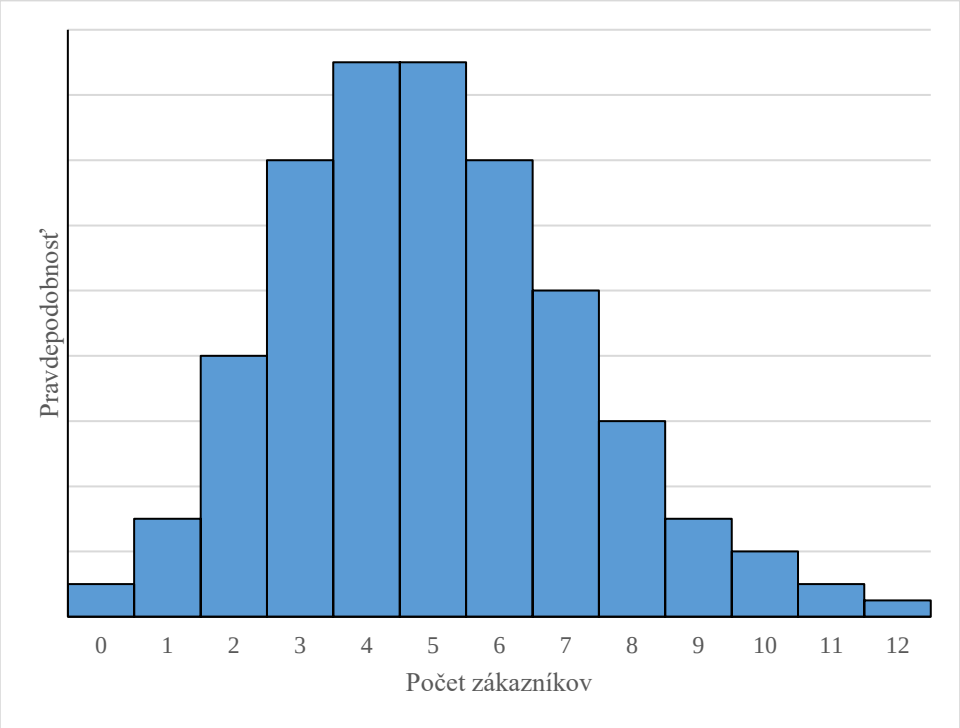
$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Číslo λ sa nazýva parameter. V tomto prípade platí, že parameter $\lambda = s_x^2 = \overline{x}$.

Príklad

Existuje rad náhodných veličín, ktorý sa týmto zákonom riadi, napr:

- počet zákazníkov v obchode za jednotku času (Obrázok 37),
- počet chýb na jednom výrobku,
- počet preklepov na stránke textu,
- počet smrteľných dopravných nehôd za deň,
- počet mimoriadnych udalostí v Žilinskom kraji za rok,
- počet telefónnych hovorov na linku 112 za jednotku času a pod.



Obrázok 37 Poissonovo rozdelenie pravdepodobnosti počtu zákazníkov

8.6 Základné typy rozdelenia spojitej náhodnej veličiny

Medzi základné typy rozdelenia spojitej náhodnej veličiny patrí napríklad rozdelenie **rovnomerné, exponenciálne, normálne (Gaussovo), Erlangovo** atď.

8.6.1 Rovnomerné rozdelenie R(a,b)

Rovnomerné rozdelenie je typické pre náhodné premenné, ktoré majú rovnakú možnosť nadobudnúť ktorúkoľvek hodnotu z nejakého konečného intervalu.

Definícia

$$X \sim R(a, b)$$

Rovnomerné rozdelenie na intervale **(*a*,*b*)**, kde $-\infty < a < b < \infty$, má vo všetkých bodoch daného intervalu konštantnú hustotu pravdepodobnosti, ktorú je možné vyjadriť vzťahom:

$$f(x) = \begin{cases} \frac{1}{b-a} & pre\ x \in (a, b) \\ 0 & pre\ x \notin (a, b) \end{cases}$$

Mimo interval *(a, b)* je teda hustota pravdepodobnosti nulová.

Základný príklad a grafické spracovanie rovnomerného rozdelenia pravdepodobnosti boli uvedené v predošlých častiach (kap. 8.4.3, 8.4.4).

Ďalšie **príklady** spojitých náhodných veličín s rovnomerným rozdelením:

- chyby pri zaokrúhľovaní čísel,
- doba, ktorá uplynie od náhodne zvoleného okamihu do nastúpenia javu, ktorý sa pravidelne opakuje v časovom intervale,
- dráha, ktorú je potrebné prejsť z náhodne zvoleného bodu do cieľa,
- ľubovoľná spojitá veličina z určitého intervalu, o ktorého správaní sa na tomto intervale nie je nič bližšie známe (núdzové riešenie v prípade neznalosti skutočného rozdelenia).

8.6.2 Exponenciálne rozdelenie E(λ)

Toto rozdelenie má spojitá náhodná veličina ***X***, ktorá predstavuje dobu čakania do nastúpenia (Poissonovského) náhodného javu, alebo dĺžku intervalu (časového alebo dĺžkového) medzi takými dvoma javmi (napr. doba čakania na obsluhu, vzdialenosť medzi dvoma poškodenými miestami na ceste). Závisí na parametri $\boldsymbol{\lambda}$, čo je prevrátená hodnota strednej hodnoty doby čakania do nastúpenia sledovaného javu.

Definícia

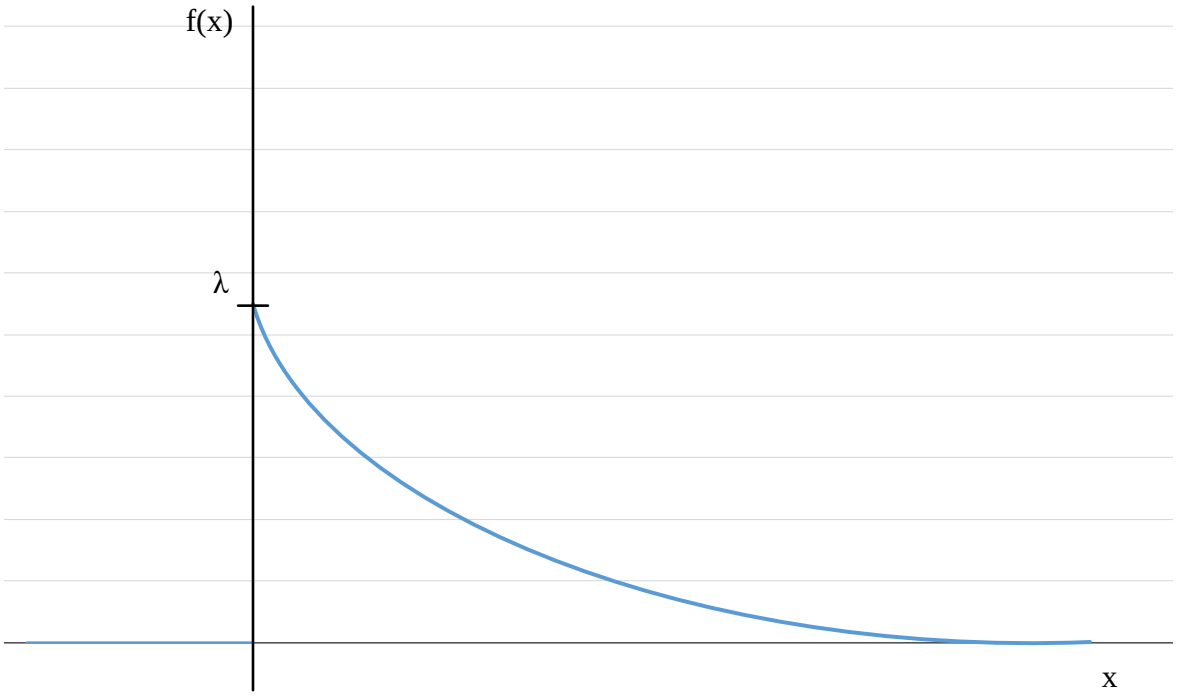
$$X \sim E(\lambda)$$

Náhodná veličina ***X*** má exponenciálne rozdelenie **E(λ)** práve vtedy, keď má hustota pravdepodobnosti rovnicu:

$$f(x) = \begin{cases} \lambda . e^{-\lambda x} & pre\ x \in (0, \infty) \\ 0 & pre\ x \leq 0 \end{cases}$$

Príklady spojitých náhodných veličín s exponenciálnym rozdelením (Obrázok 38):

- doba, ktorá uplynie od náhodne zvoleného okamihu do nastúpenia javu, ktorý sa pravidelne opakuje v časovom intervale,
- dráha, ktorú je potrebné prejsť z náhodne zvoleného bodu do cieľa,
- ľubovoľná spojitá veličina z určitého intervalu, o ktorého správaní sa na tomto intervale nie je nič bližšie známe (núdzové riešenie v prípade neznalosti skutočného rozdelenia).



Obrázok 38 Graf hustoty pravdepodobnosti exponenciálneho rozdelenia

8.6.3 Normálne rozdelenie N (μ,σ^2)

Označované tiež ako všeobecné normálne rozdelenie sa univerzálne používa k aproximácii (k približnému vyjadreniu) rozdelenia pravdepodobnosti veľkého množstva náhodných veličín v biológii, technike, ekonómii atď. Je veľmi dôležité, pretože:

- sa vyskytuje najčastejšie,
- veľa iných rozdelení sa mu približuje,
- rad iných rozdelení sa ním dá nahradiť.

Predpoklady vzniku náhodnej veličiny s normálnym rozdelením:

- spojitá náhodná veličina sa vytvára pod vplyvom mnohých činiteľov,
- jednotlivé činitele sú vzájomne nezávislé,
- žiadny z činiteľov nemá na výsledok rozhodujúci vplyv.

Na hodnoty náhodnej veličiny nekladieme žiadne obmedzenie, t. j. $-\infty < x < +\infty$.

Definícia

$$X \sim N(\mu, \sigma^2)$$

Náhodná veličina ***X*** má normálne rozdelenie ***N(μ,σ^2)*** práve vtedy, keď má hustota pravdepodobnosti rovnicu:

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

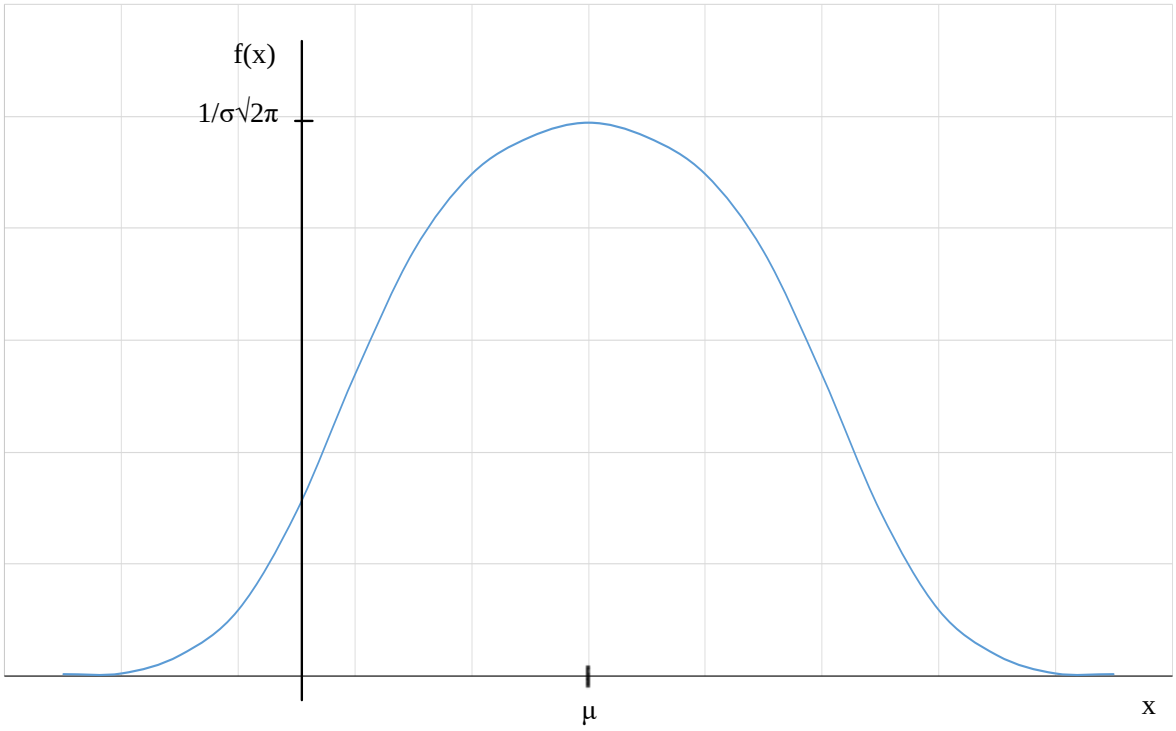
kde: $\pi = 3{,}141$ – matematická konštanta,

$e = 2,781$ – matematická konštanta,

μ a $\sigma > 0$ – parametre normálneho rozdelenia – konštanty určujúce polohu krivky na osi x (μ – vážený priemer) a jej odchýlenie pozdĺž osi x (σ – smerodajná odchýlka), t. j. priemerná hodnota a miera variability.

Ak poznáme parametre μ, σ je normálne rozdelenie plne určené.

Grafické znázornenie normálneho rozdelenia je dané symetrickou jednovrcholovou hustotou, ktorá je zvonovitého tvaru a nikde nepretína vodorovnú os (Obrázok 39). Plocha pod krivkou hustoty normálneho rozdelenia je rovná jednej. Pravdepodobnosť, že náhodná veličina nadobudne hodnoty z určitého intervalu, je rovná ploche pod hustotou nad týmto intervalom. Napríklad pre interval s hranicami $\mu - 1,96\sigma$ a $\mu + 1,96\sigma$ má tato plocha veľkosť 0,95. Náhodná veličina X nadobúda teda hodnoty z tohoto intervalu s 95% pravdepodobnosťou a iba s 5% pravdepodobnosťou ležia jej hodnoty mimo uvedený interval.



Obrázok 39 Graf hustoty pravdepodobnosti normálneho rozdelenia (Gaussova krivka resp. Gaussova-Laplaceova krivka)

Pre veličinu ***X*** s normálnym rozdelením je možné histogram početností veľkého počtu ***n*** nezávislých pozorovaní vyrovnať krivkou hustoty pravdepodobnosti.

Príklady spojitých náhodných veličín s normálnym rozdelením:

- náhodné chyby fyzikálnych (všeobecne akýchkoľvek) meraní,
- veličiny vznikajúce pod vplyvom balistických zákonov (výsledky streľby),
- znaky v biologických populáciách podliehajúci zákonom genetiky,
- všeobecne – náhodné veličiny vznikajúce ako súčty či priemery iných náhodných veličín (spojitých ale aj diskrétnych) s ľubovoľným rozdelením.

8.6.4 Normované normálne rozdelenie N(0,1)

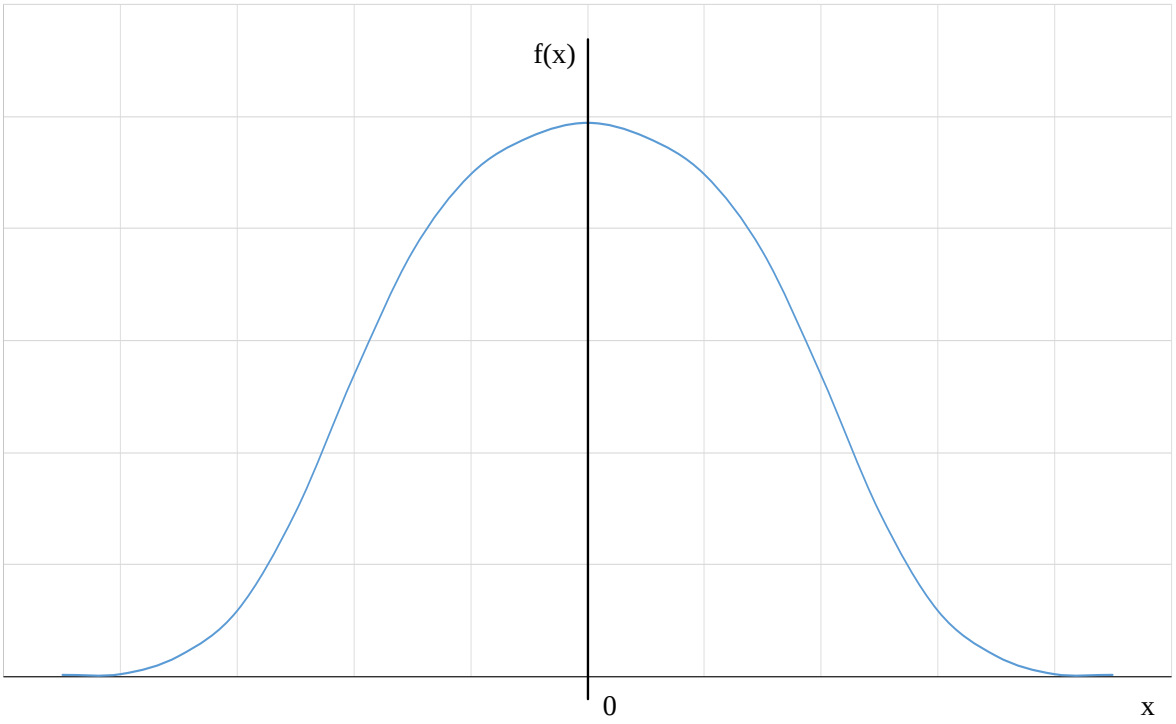
Normované normálne rozdelenie je špeciálny prípad všeobecného normálneho rozloženia, kde $\mu = 0$, $\sigma^2 = 1$.

Definícia

$$X \sim N(0,1)$$

Náhodná veličina má normované normálne rozdelenie, ak má hustota pravdepodobnosti hodnotu:

$$f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}x^2} \qquad pre\ x \in (-\infty, \infty)$$



Obrázok 40 Graf hustoty pravdepodobnosti normovaného normálneho rozdelenia

8.6.5 Niektoré ďalšie rozdelenia

Weibullovo rozdelenie *W(λ,c)*

$$X \sim \mathrm{W}(\lambda, c)$$

Toto rozdelenie má spojitá náhodná veličina, ktorá predstavuje dobu života (bezporuchovosti) technických zariadení, ktorým nevyhovuje exponenciálne rozdelenie pravdepodobnosti. Teda tam, kde sa prejavuje mechanické opotrebenie alebo únava materiálu. Parameter **λ** závisí na použitom materiáli, namáhaní a podmienkach používania ($\lambda > 0$; $c > 0$).

Funkcia hustoty pravdepodobnosti:

$$f(x) = \begin{cases} 0 & pre\ x \leq 0 \\ \dfrac{c.x^{c-1}}{\lambda^c} \cdot e^{-\left(\frac{x}{\lambda}\right)^c} & pre\ x > 0 \end{cases}$$

pre $c = 1$ dostaneme exponenciálne rozdelenie $E(\lambda)$

Pearsonovo rozdelenie $\chi^2(n)$

$X \sim \chi^2(\mathrm{n})$ (čítame chí kvadrát s *n* stupňami voľnosti)

Ak *n* nezávislých veličín $X_1,...,X_n$ má rozdelenie $N(0,1)$, tak veličina $X = X_1^2 + X_2^2 + \ldots + X_n^2$ má Pearsonovo rozdelenie.

Funkcia hustoty pravdepodobnosti:

$$f(x) = \begin{cases} \dfrac{x^{\frac{x}{2}-1} \cdot e^{-\frac{x}{2}}}{2^{\frac{n}{2}} \cdot \Gamma\left(\frac{n}{2}\right)} & pre\ x > 0 \\ 0 & pre\ x \leq 0 \end{cases}$$

Studentovo rozdelenie *t(n)*

$$X \sim t(n)$$

Ak sú X_1, X_2 dve nezávislé náhodné premenné, kde X_1 sa riadi rozložením $N(0,1)$ a X_2 rozložením $\chi^2(n)$, tak náhodná veličina $T = \frac{x_1}{\sqrt{x_2}} \cdot \sqrt{n}$ má Studentovo rozloženie s *n* stupňami voľnosti.

Funkcia hustoty pravdepodobnosti:

$$\boldsymbol{f(x) = \frac{1}{\sqrt{n\pi}} \cdot \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{\frac{n+1}{2}} \qquad pre\ x \in (-\infty, \infty)}$$

Literatúra

BEDFORD, T., COOKE, R. *Probalistic Risk Analysis: Foundations and Methods*. 7. vyd. Cambridge: Cambridge Press, 2011.

BÍLKOVÁ, D., BUDINSKÝ, P. A V. VOHÁNKA. *Pravděpodobnost a statistika.* Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, s.r.o., 2009, ISBN 978-80-7380-224-0.

BUDÍKOVÁ, M, KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. Praha: GRADA, 2010.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

FELLER, W. *An introduction to probability theory and its applications*. I. a II., New York: J. Wiley, 1970.

GIBILISCO, S. *Statistika bez předchozích znalostí*. Brno: Computer Press, 2012.

GROFÍK, R. a kol. *Štatistika*. Bratislava: Príroda, 1987.

HAINING, R. *Spatial data analysis for social and enviromental sciences*. London: OUP, 1991.

HINDLS, R., ARLTOVÁ, M., HRONOVÁ, S., MALÁ, I., MAREK, L., PECÁKOVÁ, I., ŘEZANKOVÁ, H. *Statistika v ekonomii*. Praha: Professional Publishing, 2018, ISBN 978-80-88260-09-7.

LAMOŠ, F., POTOCKÝ, R. *Pravdepodobnosť a matematická štatistika – Štatistické analýzy*. 1989. Alfa, Bratislava.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

NÁNÁSIOVÁ, O., KOHNOVÁ, S. *Štatistika a pravdepodobnosť. Základy matematickej štatistiky a teórie pravdepodobnosti*. Bratislava: STU, 2016, ISBN 978-80-2274-527-7.

OTIPKA, P., ŠMAJSTRLA, V. *Pravděpodobnost a statistika*. 2008. VŠB-TU, Ostrava. Dostupné na: http://homel.vsb.cz/~oti73/cdpast1/

RIEČAN, B.. Pravdepod*obnosť a matematická štatistika*. 1984Alfa, Bratislava.

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

SOUČEK, E. *Základy pravděpodobnosti a statistiky*. Pardubice: Univerzita Pardubice, 2005.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

VARGA, Š. *Matematická štatistika.* Bratislava: STU, 2012, ISBN 978-80-2273-789-0.

ŽEŽULA, I. *Základy pravdepodobnosti a štatistiky.* 2015. UPJŠ, Košice. Dostupné na: https://umv.science.upjs.sk/zezula/stgjax/

9 Skúmanie závislosti v štatistike

Skúmanie závislostí v štatistike sa zaoberá predovšetkým **vzájomnou závislosťou štatistických znakov viacrozmerných štatistických súborov**. Závislosti pritom môžu byť napríklad **pevné, voľné, jednostranné, obojstranné, príčinné, zdanlivé** atď.

9.1 Pevná a voľná závislosť

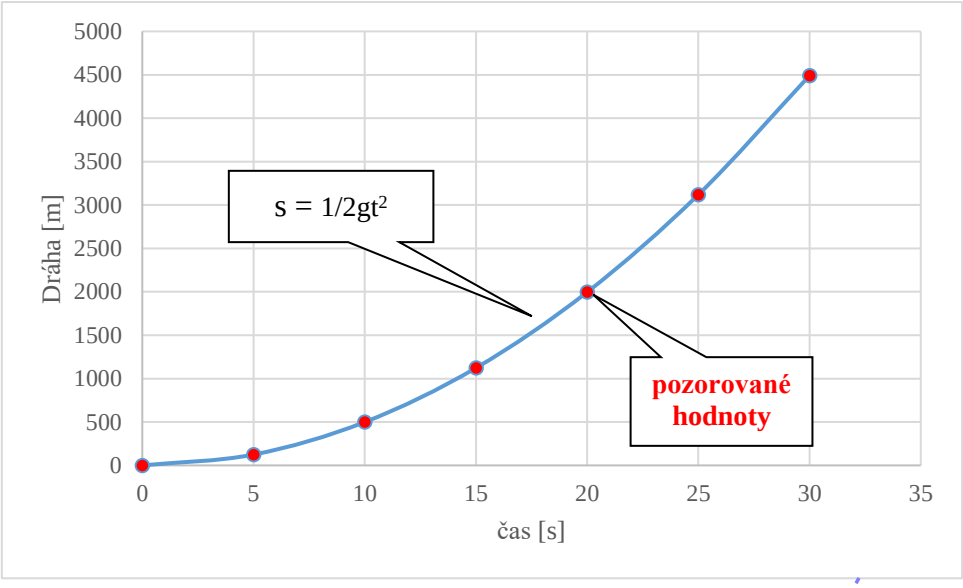
Pre pochopenie závislosti je potrebné poznať predovšetkým pevnú a voľnú závislosť.

9.1.1 Pevná závislosť

Pevná závislosť sa obvykle vyskytuje pri niektorých prírodných javoch, kedy **zmena** jedného javu spôsobuje zmenu javu druhého, a to **v presne zodpovedajúcej intenzite**. Napríklad: *dĺžka kovovej tyče je vo funkčnom vzťahu závislá od teploty, v geometrii plocha štvorca funkčne závisí na jeho strane a pod.*

Príklad:

Pevná (funkčná, deterministická) závislosť — voľný pád.



Obrázok 41 Pevná závislosť dráhy na čase pri voľnom páde

Pre pevnú závislosť platí:

- pozorovanými hodnotami je možné presne preložiť spojitú krivku o známej rovnici,
- prípadné odchýlky od krivky sú spôsobené iba chybami merania,
- počet nameraných hodnôt neovplyvňuje presnosť záverov,
- situáciu je možné kedykoľvek presne opakovať.

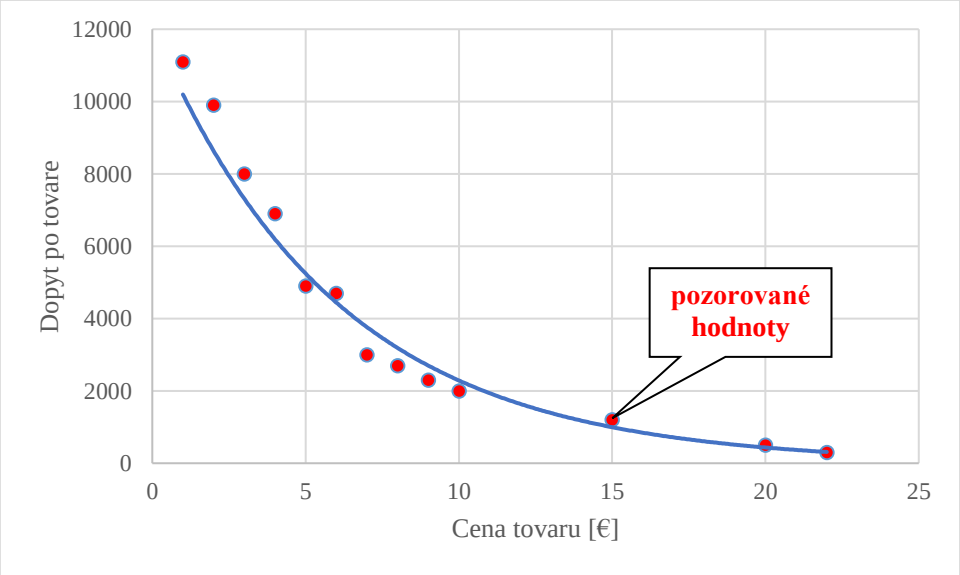
9.1.2 Voľná závislosť

Niektoré javy môžu byť na sebe **závislé iba voľne**, napr. *závislosť výnosu plodiny na spotrebe hnojiva, závislosť dopytu na cene tovaru a pod.* Aj tu sa prejaví závislosť, avšak vzťah je viac či menej voľný.

Zmena jedného javu **podmieňuje** úroveň iného javu iba **s určitou pravdepodobnosťou** a takisto intenzita zmeny druhého javu môže byť rôzna. Túto závislosť môžeme skúmať iba pri väčšom množstve skúmaných štatistických jednotiek/pokusov.

Príklad:

Voľná (stochastická) závislosť — trhový dopyt:



Obrázok 42 Voľná závislosť medzi dopytom a cenou tovaru

Pre voľnú závislosť platí:

- všetkými pozorovanými hodnotami nie je možné preložiť krivku,
- odchýlky od „ideálneho priebehu“ závislosti sú dané individuálnymi zvláštnosťami jednotlivých prípadov,
- informácie o závislosti sa spresňujú s pribúdajúcim počtom prípadov,
- situáciu sa nikdy nepodarí znovu presne reprodukovať.

Predmetom záujmu štatistiky je **predovšetkým** voľná závislosť, ktorá je typická pre sociálne, ekonomické i mnohé iné vysoko komplikované systémy a javy. V rámci tohoto skúmanie je ale možné odhaliť aj pevné závislosti.

9.2 Klasifikácia štatistických závislostí

Podľa druhu štatistických znakov je možné závislosti členiť nasledovne:

- **závislosť medzi kvantitatívnymi štatistickými znakmi** – **korelačná závislosť –** napr. *vzťah medzi spotrebou krmiva a dosahovaným prírastkom u zvierat, medzi dĺžkou klasu pšenice a počtom zŕn v klase, medzi výnosom plodiny na strane jednej a spotrebou hnojiva*, a pod.
- **závislosť medzi slovnými štatistickými znakmi – asociačná závislosť a kontingenčná závislosť** (kap. 9.4).

Všetky závislosti je možné rozdeliť na závislosti:

- **príčinné,** kde vystupuje:

- jeden jav ako **príčina** — „**nezávislá**“ **premenná (*X*),**
- druhý jav ako **účinok** — „**závislá**“ **premenná (*Y*).**

- **zdanlivé** (neexistuje logická spojitosť medzi dvoma štatistickými znakmi; medzi príčinou a účinkom).

Štatistika skúma príčinné voľné závislosti.

Každá závislosť **číselných štatistických znakov** má dva vzájomne neoddeliteľné atribúty (vlastnosti), ktoré je vhodné skúmať:

- **intenzitu (silu, mieru) závislosti –** skúma **korelácia**,
- **priebeh (vývoj) závislosti –** skúma **regresia.**

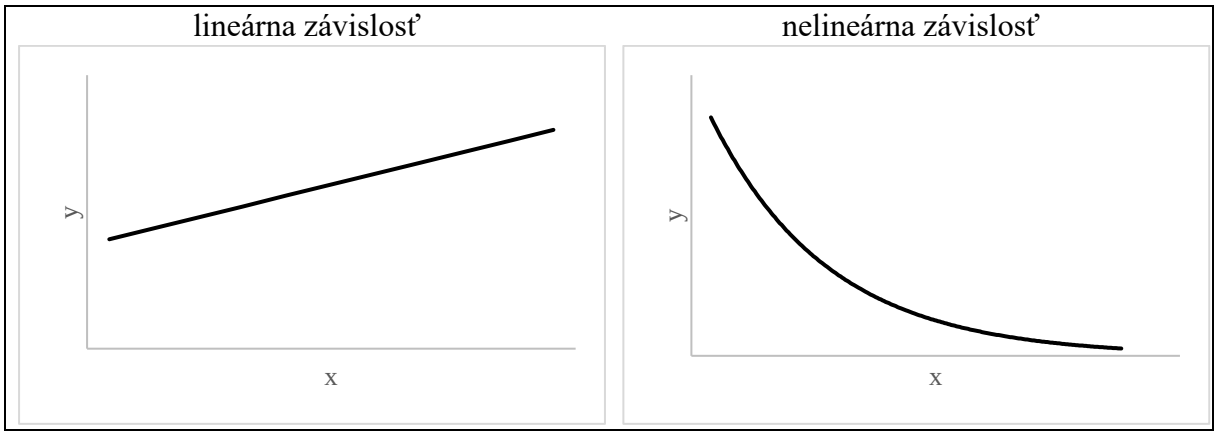
Pre skúmanie závislosti medzi dvoma **slovnými (kvalitatívnymi) štatistickými znakmi** platí, že štatistika skúma iba mieru závislosti medzi týmito štatistickými znakmi (**mieru kontingencie** alebo **mieru asociácie**).

Príčinné závislosti číselných štatistických znakov je možné **klasifikovať** z rôznych hľadísk:

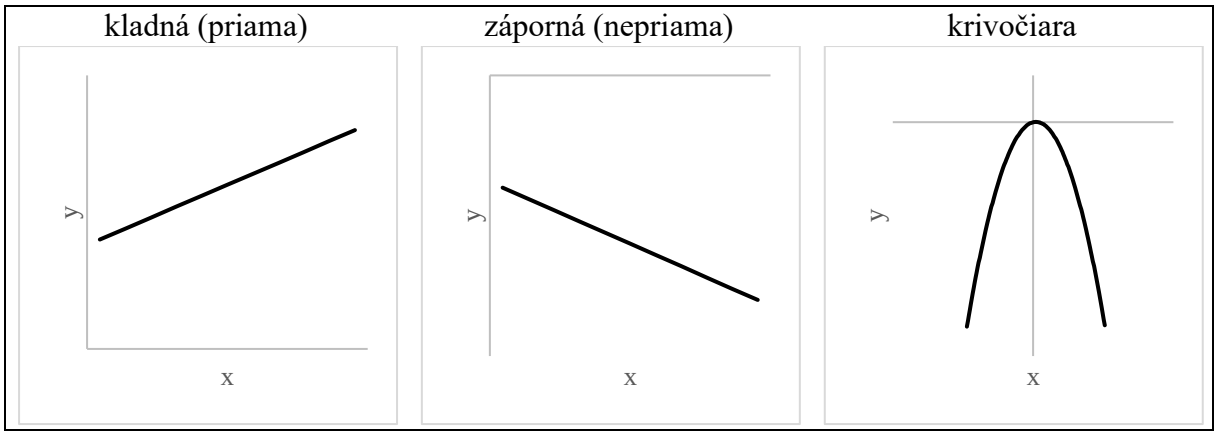
- na závislosti **jednostranné** a závislosti **obojstranné** (vzájomné),
- na závislosti **priamočiare** a závislosti **krivočiare**,
- niektoré (najmä priamočiare) na závislosti **pozitívne** a závislosti **negatívne** (toto hľadisko má iba okrajový význam),
- podľa matematických (regresných) funkcií použitých na skúmanie priebehu závislosti na závislosti **lineárne** a závislosti **nelineárne**,
- podľa počtu príčin (nezávislých premenných) na závislosti **párové** (jednoduché, s jednou nezávislou premennou) a závislosti **mnohonásobné** (s najmenej dvoma súčasne pôsobiacimi nezávislými premennými), atď.

V praxi sa väčšina úloh obmedzuje iba na **párové a lineárne alebo krivočiare** závislosti.

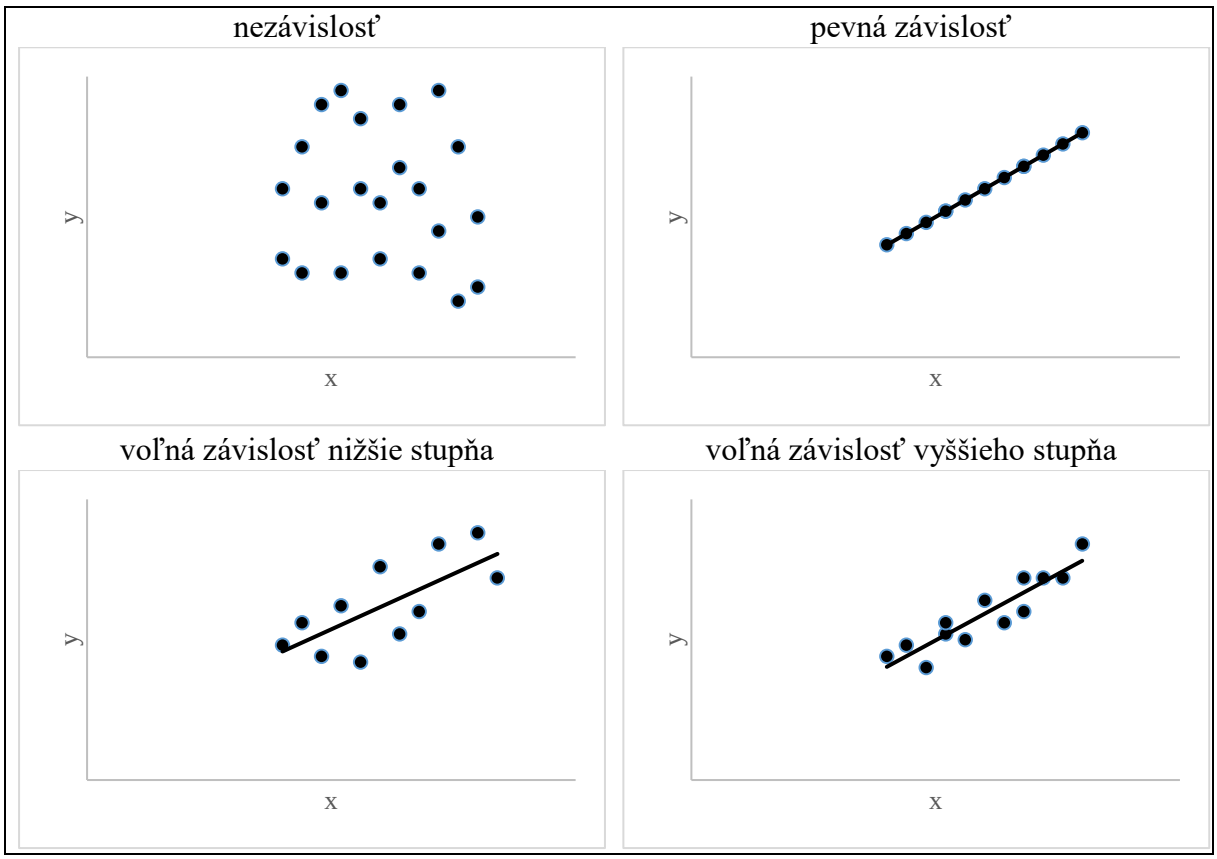
Druhy korelačných závislostí sú zobrazené na nasledujúcich obrázkoch (Obrázok 43, Obrázok 44, Obrázok 45).



Obrázok 43 Príklady korelačnej závislosti podľa typu regresnej funkcie



Obrázok 44 Príklady korelačnej závislosti podľa smeru regresnej funkcie



Obrázok 45 Príklady korelačnej závislosti stupňa závislosti (korelácie) štatistických znakov

9.3 Korelačná závislosť – korelačná analýza

Korelačná analýza skúma korelačnú závislosť medzi **kvantitatívnymi** (číselnými) štatistickými znakmi.

Pri skúmaní korelačnej závislosti (korelačnej analýze) sa rozoznávajú dva základné pojmy:

- **korelácia** – miera, stupeň (tesnosť) závislosti – riešená **korelačnou úlohou**,
- **regresia** – priebeh závislosti prostredníctvom matematickej funkcie, zmena závislej pramennej podľa nezávislej pramennej – riešená **regresnou úlohou** (kap. 10.1).

Základom pre skúmanie korelačnej závislosti medzi dvoma číselnými štatistickými znakmi (korelácie aj regresie) je základná - dátová tabuľka (Tabuľka 30), do ktorej sa zaznamenávajú hodnoty štatistických znakov pre všetky štatistické jednotky od $i = 1$ až po $i = n$.

Tabuľka 30 Základná - dátová tabuľka na skúmanie závislosti

Štatistická jednotka	Hodnoty štatistických znakov	
	Znak x_i (nezávislá)	Znak y_i
1	x_1	y_1
2	x_2	y_2
...	...	...
n	x_n	y_n

V tejto podobe **ide súčasne o záznam výsledkov zisťovania** za ***n*** členný štatistický súbor.

Korelačná úloha

Korelačná úloha spočíva v skúmaní **tesnosti korelačného vzťahu** medzi dvomi štatistickými znakmi (premennými) X a Y. Závislosť znamená, že hodnoty jedného štatistického znaku ovplyvňujú hodnoty druhého štatistického znaku.

Mierou korelácie je **koeficient alebo index korelácie *r (Peasonov korelačný koeficient),*** ktorý nadobúda hodnoty od ***-1*** do ***1***. Udáva stupeň (úroveň, tesnosť) vzťahu (závislosti) posudzovaných štatistických znakov.

$$\boldsymbol{r =< -1; +1 >}$$

Silu (mieru, intenzitu) **lineárnej** závislosti je možné odvodiť z absolútnej hodnoty koeficientu korelácie. Ak sú **hodnoty oboch štatistických znakov nezávislé**, bude **korelácia blízka nule**. Čím je bližšie **absolútna hodnota** k hodnote **1**, tým je **silnejšia závislosť** medzi skúmanými štatistickými znakmi. Znamienko, teda pozitívna závislosť (+) alebo negatívna závislosť (-), určuje smer závislosti takto:

- **r > 0 (pozitívna závislosť)** = medzi premennými je priamy vzťah,
- **r < 0 (negatívna závislosť)** = medzi premennými je nepriamy vzťah.

Podľa hodnoty indexu (koeficientu) korelácie sa určuje **miera závislosti**.

Pozitívna závislosť:

- r = <0 – 0,2> jedná sa o žiadnu alebo veľmi slabú závislosť,
- r = (0,2 – 0,4> jedná sa o slabú závislosť,
- r = (0,4 – 0,6> jedná sa o priemernú závislosť,
- r = (0,6 – 0,8> jedná sa o silnú závislosť,
- r = (0,8 – 1,0> jedná sa o veľmi silnú závislosť.

Negatívna závislosť:

- r = <0; -0,2> jedná sa o žiadnu alebo veľmi slabú závislosť,
- r = (-0,2; -0,4> jedná sa o slabú závislosť,
- r = (-0,4; -0,6> jedná sa o priemernú závislosť,

- r = (-0,6; -0,8> jedná sa o silnú závislosť,
- r = (-0,8; -1,0> jedná sa o veľmi silnú závislosť.

V prípade, že náhodné veličiny X a Y sú kvantitatívne náhodné veličiny je pre konkrétne hodnoty (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) **Pearsonov korelačný koeficient** daný vzťahom:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Používa sa aj jednoduchšie vyjadrenie **korelačného koeficientu:**

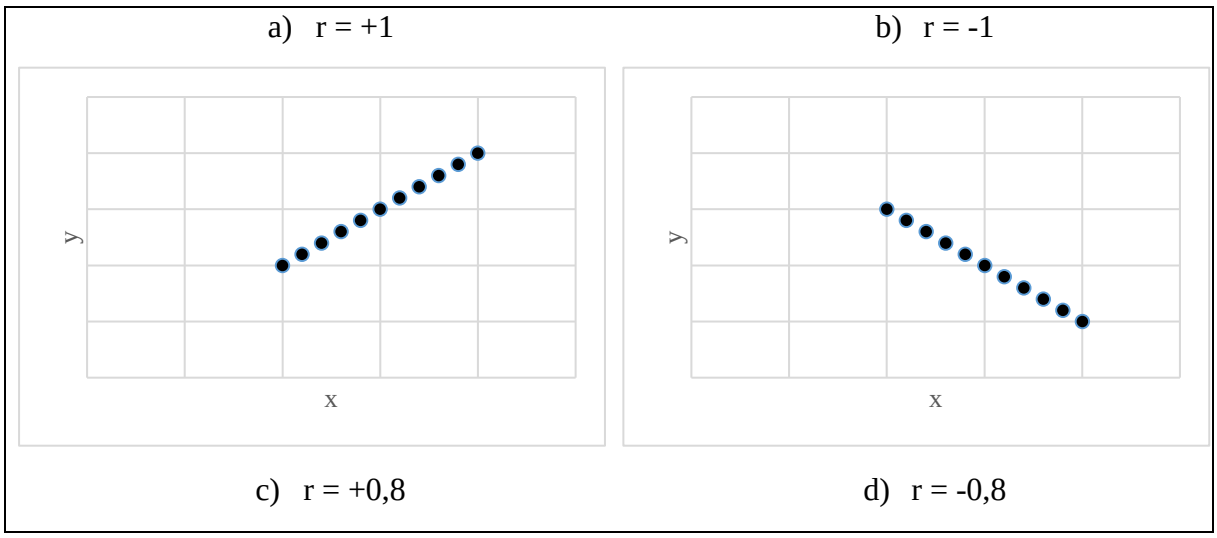
$$r = \frac{s_{xy}}{s_x s_y}$$

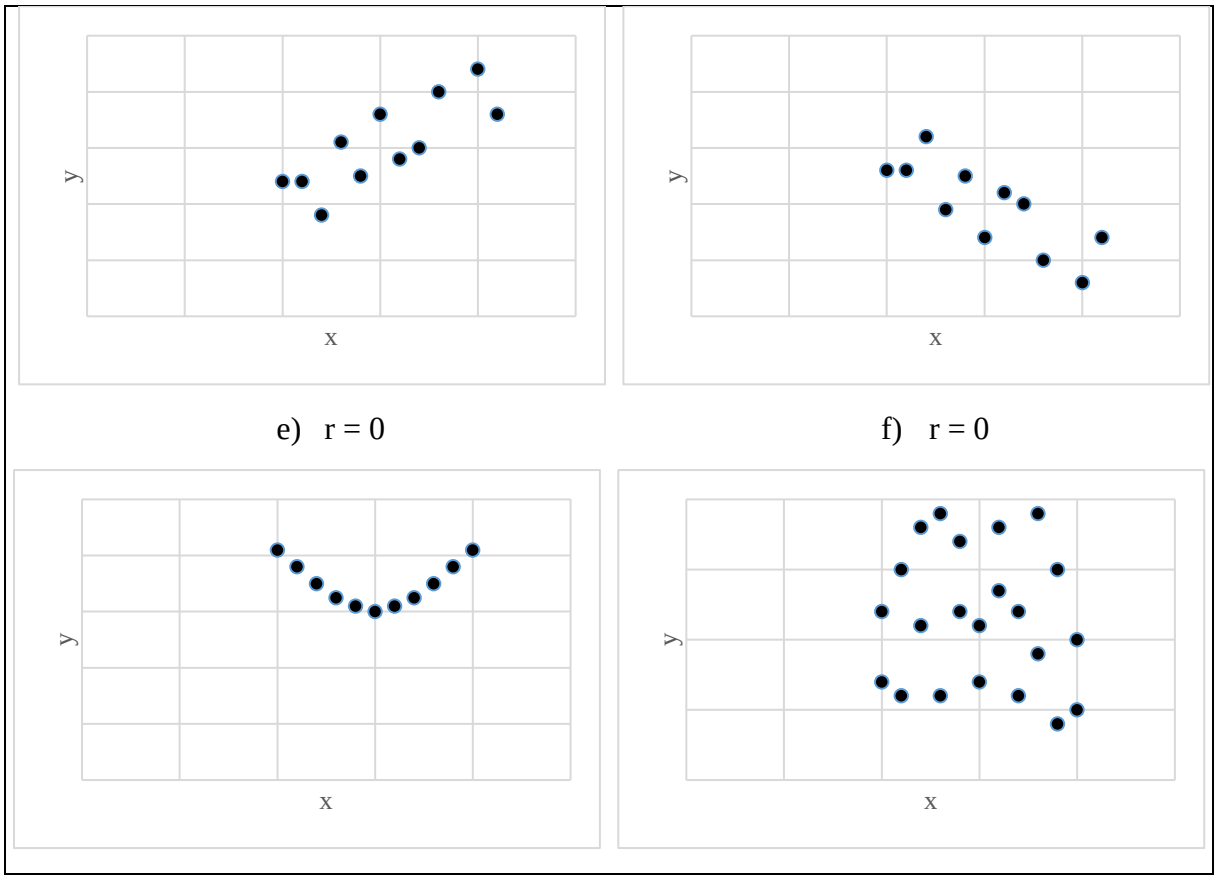
kde: s_x je smerodajná odchýlka premennej X,
s_y je smerodajná odchýlka premennej Y,
s_{xy} je takzvaná kovariancia (alebo spoločný rozptyl) premenných X a Y.

Spoločný rozptyl premenných X a Y je možné vyjadriť vzťahom:

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

Lineárnu závislosť dvoch štatistických znakov je možné skúmať napr. vynesením hodnôt **premenných (štatistických znakov) do grafu**. Rôzne miery korelácie je možné vidieť na ďalších grafoch (Obrázok 46). Zvlášť dôležitý je graf e), ktorý evidentne vykazuje silnú závislosť medzi premennými X a Y. Závislosť medzi danými premennými však nie je lineárna a keďže Pearsonov koeficient korelácie skúma lineárnu závislosť je koeficient *r* rovný nule. To však nevylučuje inú závislosť. Pre uvedený príklad sa jedná o parabolickú závislosť.





Obrázok 46 Príklady korelácií

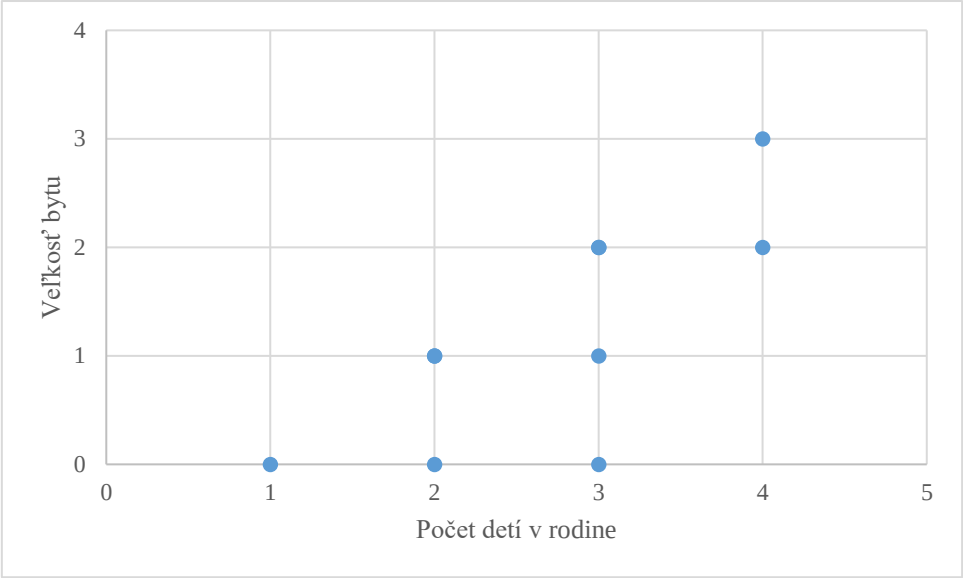
V rámci korelačnej úlohy sa **nestanovuje** rovnica priamky závislosti (to je úlohou lineárnej regresie – kap. 10). Vyjadrenie lineárneho vzťahu medzi premennými si je možné predstaviť pomyslenou priamkou, ktorá sa preloží bodmi grafu. Miera lineárneho vzťahu záleží od veľkosti odchýlok jednotlivých bodov grafu od tejto pomyselnej priamky. Vzhľadom na veľkosť odchýlok je následne možné odhadnúť mieru tohoto vzťahu. Väčšie odchýlky indikujú nižšiu mieru závislosti.

Príklad na korelačnú úlohu:

Je možné použiť príklad, ktorý bol prvotne zameraný na triedenie rodín podľa počtu detí a veľkosti bytu (viď kap. 6.2.1; Tabuľka 13, Tabuľka 14). Rozdiel je v tom, že riešiteľa momentálne nezaujíma samotná početnosť kombinácií daných štatistických znakov, ale fakt, či medzi počtom detí v rodine (nezávislá premenná ***x***) a veľkosťou bytu, ktorý je vyjadrený počtom obytných miestností (závislá premenná ***y***), existuje závislosť a aká je miera tejto závislosti.

Základom korelačnej úlohy je výpočet koeficientu korelácie. Pre uvedený príklad je koeficient korelácie $r = 0,75$. Čo znamená, že medzi uvedenými štatistickými znakmi je vysoká (pozitíva) lineárna závislosť. Z uvedeného vyplýva, že s narastajúcim počtom detí narastá aj počet obytných miestností (veľkosť bytu) rodiny. Sila závislosti určuje silu tohto tvrdenia (vzťahu).

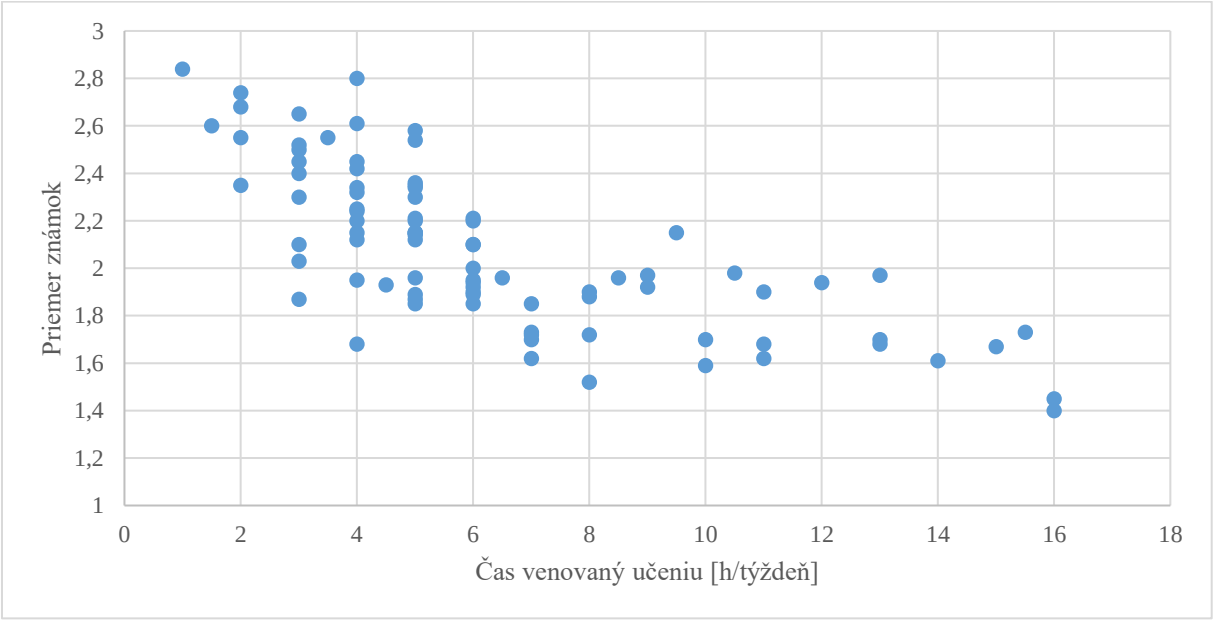
Prostriedkom **grafickej prezentácie** závislostí číselných znakov je **korelačný bodový graf** (Obrázok 47). Body v grafe predstavujú jednotlivé štatistické jednotky, ktorým zodpovedajú obmeny príslušných štatistických znakov na osiach x a y.



Obrázok 47 Korelačný bodový graf na skúmanie závislosti medzi počtom detí v rodine a veľkosťou ich bytu

Poznámka: Keď sa vyskytne viac štatistických jednotiek z rovnakými hodnotami štatistických znakov, body sa v bodovom korelačnom grafe prekrývajú.

V uvedenom príklade boli použité diskrétne číselné štatistické znaky (štatistické znaky, ktoré dosahujú iba izolované hodnoty). Na grafe je zrejmé, že existuje iba niekoľko možných kombinácií obmien štatistických znakov (iba niekoľko bodov v grafe). Skúmať takúto závislosť z grafického znázornenia je tak čiastočne obmedzené množstvom možných kombinácií medzi danými číselnými štatistickými znakmi. Závislosť dvoch číselných štatistických znakov je názornejšia u **spojitých** číselných štatistických znakov a z tohto pohľadu vhodnejšia na skúmanie. Viď nasledujúci príklad zameraný na skúmanie závislosti medzi časom, ktorý študenti venujú učeniu a ich dosiahnutými výsledkami (priemer za školský rok).



Obrázok 48 Korelačný bodový graf na skúmanie závislosti medzi časom, ktorý študenti venujú učeniu a ich priemeru známok za daný školský rok

Pre uvedený príklad je koeficient korelácie $r = -0,73$, čo znamená, že medzi uvedenými štatistickými znakmi je vysoká negatívna lineárna závislosť. Z uvedeného vyplýva, že žiak, ktorý venuje viac času príprave a učeniu, dosahuje lepšie výsledky v škole (dosahuje nižší priemer). Negatívna závislosť znamená, že s rastúcimi hodnotami nezávislého štatistického znaku (čas venovaný učeniu), klesajú hodnoty závislej premennej (priemer).

9.4 Závislosť medzi slovnými štatistickými znakmi

V praxi sa často stretávame taktiež s úlohami, ktorých cieľom je zistiť či sú dve **slovné** náhodné veličiny (štatistické znaky) nezávislé alebo medzi nimi existuje závislosť. Riešiteľa môže napríklad zaujímať *či je farba očí určitej populácie a farba ich vlasov v určitom vzťahu alebo nie; či má očkovanie vplyv na výskyt ochorenia pre ďalšie obdobie* a pod.

Súčasne môže riešiteľa zaujímať či prípadná závislosť je silná alebo slabá. Podobne ako u číselných štatistických znakov je tak potrebné **skúmať silu (tesnosť, intenzitu)** tejto **závislosti**. Na tento účel slúžia rôzne koeficienty (viď nasledujúce kapitoly), ktoré nadobúdajú hodnoty z intervalov **<0,1> prípadne <-1;1>.** Absolútna hodnota týchto koeficientov hovorí o sile závislosti (bližšie k 1 je závislosť silnejšia) a ak je možné vyjadriť koeficient aj zápornou hodnotou, tak to naznačuje smer danej závislosti. Kladná hodnota hovorí o priamej závislosti a záporná hodnota o nepriamej závislosti.

Stupnica sily závislosti medzi slovnými štatistickými znakmi:

- <0 – 0,3> jedná sa o veľmi slabú až slabú závislosť,
- (0,3 – 0,7> jedná sa o priemernú závislosť,
- (0,7 – 0,9> jedná sa o silnú závislosť,
- (0,9 – 1,0> jedná sa o veľmi silnú závislosť.

Niekedy sa používa na vyjadrenie sily vzťahu medzi dvoma slovnými štatistickými znakmi aj stupnica pre korelačnú závislosť (kap. 0).

Vzhľadom na charakter skúmaných slovných štatistických znakov (alternatívne alebo množné – teda rozdiel v množstve obmien skúmaných štatistických znakov) sa rozdeľuje skúmanie miery závislosti medzi slovnými štatistickými znakmi na:

- **asociačnú závislosť** – závislosť medzi kvalitatívnymi **alternatívnymi** štatistickými znakmi napr. *vzťah medzi postrekom stromov (bol použitý postrek/nebol použitý postrek) a červivosťou ovocia (červivé ovocie/nečervivé ovocie)*,
- **kontingenčnú závislosť** – závislosť medzi kvalitatívnymi štatistickými znakmi **množnými** (prípadne jeden množný a jeden alternatívny) – napr. *citlivosť rôznych druhov zvierat na niektoré stresové podnety, vplyv rôznych technológií na výnos jednotlivých druhov obilnín*.

9.4.1 Asociačná závislosť

Asociačná závislosť je teda závislosť medzi dvoma kvalitatívnymi **alternatívnymi (dvojnými, dichotomickými) štatistickými znakmi**.

Pokiaľ sa skúma jednostranná závislosť medzi danými štatistickými znakmi – skúma sa či štatistický znak A ovplyvňuje obmeny štatistického znaku B, na zistenie miery takejto asociačnej závislosti je možné použiť napríklad **asociačný koeficient (koeficient asociácie):**

$$Q_{ab} = \frac{n_{ab} \cdot n_{\alpha\beta} - n_{a\beta} \cdot n_{\alpha b}}{n_{ab} \cdot n_{\alpha\beta} + n_{a\beta} \cdot n_{\alpha b}}$$

Pre jeho správne použitie je potrebné získať početnosti kombinácii výskytu prvého a druhého štatistického znaku ako je uvedené nižšie (Tabuľka 32). Predtým je však potrebné zvoliť vhodný zápis skúmaných obmien štatistických znakov do základnej tabuľky (Tabuľka 31). Ak to povaha štatistického znaku umožňuje (pre alternatívny štatistický znak „pohlavie“ (muž/žena) nie je možné určiť tzv. prítomnosť štatistického znaku), prvé sa zapisujú obmeny vyjadrujúce prítomnosť alternatívneho štatistického znaku. Napríklad štatistický znak „*skúsenosť so zemetrasením*“ môže mať obmeny „áno“ respondent má skúsenosť so zemetrasením a „nie“ respondent nemá skúsenosť so zemetrasením. Ako prvá sa teda do asociačnej tabuľky zapisuje obmena „áno“, ktorá popisuje prítomnosť skúsenosti so zemetrasením. Podobne to môže byť s chorobou, očkovaním a pod. Poradie obmien štatistických znakov v asociačnej tabuľke ovplyvňuje korektnosť asociačného koeficientu, hlavne v zmysle jeho pozitívnej alebo negatívnej hodnoty a tým pádom ovplyvňuje jeho následnú interpretáciu.

Tabuľka 31 Základná - dátová tabuľka na skúmanie asociačnej závislosti

	prítomnosť znaku	neprítomnosť znaku
znak **A**	**a**	**α**
znak **B**	**b**	**β**

Tabuľka 32 Všeobecná asociačná tabuľka

znak **A**	znak **B**		Spolu
	b	**β**	
a	n_{ab}	$n_{a\beta}$	n_a
α	$n_{\alpha b}$	$n_{\alpha\beta}$	n_α
Spolu	n_b	n_β	n

Príklad:

K dispozícii je reprezentatívny štatistický súbor vybraných zamestnancov o rozsahu n = 450. Na danom štatistickom súbore skúmame závislosť medzi očkovaním zamestnancov a následným výskytom ochorenia u týchto zamestnancov. Alternatívne štatistické znaky sú popísané nasledovne:

- štatistický znak A: očkovanie zamestnancov (zamestnanec bol alebo nebol očkovaný),
- štatistický znak B: ochorenie zamestnancov (zamestnanec bol alebo nebol v ďalšom sledovanom období chorý).

Tabuľka 33 Absolútne početnosti kombinácie obmien štatistických znakov očkovanie zamestnancov a ochorenie zamestnancov

Očkovanie zamestnancov (A)		Ochorenie zamestnancov (B)		Spolu
		bol chorý	nebol chorý	
		b	**β**	
bol očkovaný	**a**	12	323	335
nebol očkovaný	**α**	53	62	115
Spolu		65	385	450

Výpočet koeficientu asociácie:

$$Q_{ab} = \frac{n_{ab} \cdot n_{\alpha\beta} - n_{a\beta} \cdot n_{\alpha b}}{n_{ab} \cdot n_{\alpha\beta} + n_{a\beta} \cdot n_{\alpha b}} = \frac{12 \cdot 62 - 323 \cdot 53}{12 \cdot 62 + 323 \cdot 53} = -0{,}92$$

Koeficient asociácie vykazuje vysoký stupeň účinnosti očkovania, pretože absolútna hodnota koeficientu asociácie je blízka hodnote 1. Negatívny vzťah poukazuje na fakt, že „prítomnosť“ očkovania u respondentov spôsobí opačne (nepriamy vzťah) „neprítomnosť“ ochorenia.

Spomínaný koeficient asociácie je primárne určený na skúmanie jednostrannej závislosti medzi alternatívnymi štatistickými znakmi (napr. očkovanie môže mať vplyv na následné ochorenie, naopak to neplatí). V prípadoch ak existuje obojstranná závislosť medzi skúmanými štatistickými znakmi je možné použiť na výpočet miery tejto závislosti tzv. **koeficient korelácie** pre alternatívne štatistické znaky:

$$R_{ab} = \frac{n_{ab} \cdot n_{\alpha\beta} - n_{a\beta} \cdot n_{\alpha b}}{\sqrt{n_a \cdot n_\alpha \cdot n_b \cdot n_\beta}}$$

Uvedený koeficient môže dosahovať rovnaké hodnoty ako koeficient asociácie.

9.4.2 Kontingenčná závislosť

Kontingenčná závislosť skúma závislosť medzi **kvalitatívnymi štatistickými znakmi** (pričom **aspoň 1 je množný**).

Kontingenčná závislosť sa určuje na základe výpočtu **koeficientov kontingencie** (v angličtine sa častejšie uvádzajú ako koeficienty asociácie, ako všeobecný pojem pre všetky závislosti medzi slovnými štatistickými znakmi).

Na výpočet koeficientov kontingencie je zväčša potrebné poznať hodnotu χ^2 (tzv. **chí kvadrát**). Pre kontingenčnú tabuľku s početnosťami O_{ij} a očakávanými početnosťami E_{ij} sa definuje veličina chí-kvadrát vzťahom

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

kde: *r je* počet obmien 1. štatistického znaku (počet riadkov)
c je počet obmien 2. štatistického znaku (počet stĺpcov)
O_{ij} sú pozorované (z angl. „observed“) hodnoty
E_{ij} sú očakávané (z angl. „expected“) hodnoty

Príklad pre výpočet chí kvadrátu:

Pre účely spracovanie štatistického projektu a postupného výpočtu je uvedený príklad na skúmanie závilosti medzi preferenciou konkrétneho výrobku vybranými zákazníkmi (n = 441) vo vybraných krajinách.

Prvou úlohou je získať pozorované hodnoty O_{ij}, ktoré predstavujú absolútne početnosti zistené **z kombinačného triedenia dvoch slovných štatistických znakov** (Tabuľka 34, prípadne pre iný príklad Tabuľka 16).

Tabuľka 34 Pozorované hodnoty kombinácie obmien štatistických znakov (empirické početnosti) – rozdelenie zákazníkov podľa krajiny a preferovaného výrobku

Krajina	**Výrobok**				a_i
	A	B	C	D	
Francúzsko	72	45	12	23	152
Španielsko	45	10	16	56	127
Portugalsko	7	10	56	89	162
b_j	124	65	84	168	441

Druhou úlohou je získať očakávané hodnoty E_{ij}. Na ich výpočet sa používa prepočet:

$$E_{ij} = \frac{a_i \cdot b_j}{n}$$

kde: a_i sú súčty absolútnych početností v jednotlivých riadkoch (Tabuľka 34)
b_j sú súčty absolútnych početností v jednotlivých stĺpcoch (Tabuľka 34)

Pre hodnotu E_{11} (kombinácia výrobok A a krajina Francúzsko) platí výpočet:

$$E_1 = \frac{152 \cdot 124}{441} = 42{,}74$$

Tabuľka 35 Očakávané hodnoty kombinácie obmien štatistických znakov (teoretické početnosti)

Krajina	**Výrobok**				a_i
	A	B	C	D	
Francúzsko	42,74	22,40	28,95	57,90	152
Španielsko	35,71	18,72	24,19	48,38	127
Portugalsko	45,55	23,88	30,86	61,71	162
b_j	124	65	84	168	441

Treťou úlohou je porovnať empirické a teoretické početnosti, z čoho riešiteľ získa hodnotu chí kvadrátu. Jednotlivé porovnania daných početností sa zaznamenávajú do tabuľky, tzv. testovacie kritérium chí kvadrátu (Tabuľka 36).

Tabuľka 36 Testovacie kritérium chí kvadrátu

Krajina	**Výrobok**				a_i
	A	B	C	D	
Francúzsko	20,03	22,79	9,93	21,04	73,79
Španielsko	2,42	4,06	2,77	1,20	10,45
Portugalsko	32,63	8,07	20,49	12,06	73,24
b_j	55,08	34,92	33,19	34,30	157,48

Výpočet chí kvadrátu pre daný príklad:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(O_{ij}-E_{ij}\right)^2}{E_{ij}} = \frac{(72-42{,}74)^2}{42{,}74} + \frac{(45-22{,}40)^2}{22{,}40} + \cdots + \frac{(89-61{,}71)^2}{61{,}71} = 157{,}48$$

Hodnota χ^2 slúži ako hodnota na testovanie resp. porovnanie danej hodnoty s tabuľkovou hodnotou kritického oboru pre príslušný stupeň voľnosti a zvolenú hladinu významnosti výsledku α (α sa zväčša stanovuje na hodnotu 0,05 alebo 0,01, resp. 95% spoľahlivosť alebo 99% spoľahlivosť). Stupeň voľnosti je pritom súčin medzi počtom obmien jedného štatistického znaku mínus 1 a počtom obmien druhého štatistického znaku mínus 1; pre uvedený príklad je stupeň voľnosti rovný (4-1)*(3-1) = 6. V podstate sa jedná o test hypotézy o nezávislosti štatistických znakov. Nízke hodnoty χ^2 hovoria v prospech hypotézy, vysoké hodnoty v neprospech hypotézy. Daná problematika nie je súčasťou týchto skrípt.

Kým teda riešiteľ pristúpi k samotnému meraniu sily (tesnosti) závislosti medzi slovnými štatistickými znakmi prostredníctvom konkrétnych koeficientov, mal by okrem vyššie uvedených krokov otestovať hypotézu o nezávislosti týchto štatistických znakov (ak je hypotéza o nezávislosti je zamietnutá pristupuje k meraniu sily závilosti). Pre účely týchto skrípt sú skúmané štatistické znaky považované za závislé.

Silu závislosti medzi slovnými štatistickými znakmi najčastejšie vyjadruje:

- **Tschuprowov** (Čuprovov) **koeficient** kontingencie – **T,**
- **Cramerov** koeficient – **V,**
- **koeficient kontingencie** – **C**.

Tschuprowov (Čuprovov) koeficient kontingencie – T

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(r-1)(s-1)}}}$$

kde: *r* – počet obmien 1. štatistického znaku (počet riadkov)
s – počet obmien 2. štatistického znaku (počet stĺpcov)

Daný koeficient nadobúda iba kladné hodnoty: **T = <0 – 1>**

Cramerov koeficient – V

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1, s-1)}}$$

Daný koeficient nadobúda iba kladné hodnoty: **V = <0 – 1>**

Koeficient kontingencie – C

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Daný koeficient nadobúda iba kladné hodnoty: **C = <0 – 1>**

Pre vyššie uvedený príklad sú hodnoty koeficientov kontingencie nasledujúce: *T=V=0,38; C=0,51.* Medzi preferovaným výrobkom a krajinou je priemerná závislosť.

9.5 Skúmanie príčinnej závislosti prostredníctvom klasického experimentu

Ako bolo naznačené vyššie, skúmanie závislosti medzi štatistickými znakmi neskúma stále príčinné (kauzálne) závislosti. Z tohto pohľadu má **experiment** alebo **experimentálna metóda** medzi metódami kvantitatívneho výskumu kľúčové postavenie. Experimentálna metóda ako jediná z výskumných metód vie dokázať **kauzálne (príčinné) dôsledky** pôsobenia určitých javov, procesov, štatistických znakov. Pomocou iných štatistických metód sa zisťujú prevažne popisné charakteristiky javov alebo ich vzájomné vzťahy, s istotou však nie je možné potvrdiť, že tieto vzťahy sú kauzálne.

Kauzalita je teda jadrom experimentu. Podstatu kauzality je možné vyjadriť vzťahom: **jav A → dôsledok D.** Na to, aby sa mohla zistiť kauzalita javu A, tento jav musí **predchádzať** jeho dôsledkom. Jav A musí existovať pred dôsledkom D.

Príklady:

Nový liek → vyliečenie (vyliečenie nastalo vplyvom lieku)

Športový tréning → rekord (rekord nastal vplyvom športového tréningu)

Zvýšenie motivácie žiaka učiť sa → lepšie učebné výsledky (lepšie učebné výsledky vznikli vplyvom zvýšenia motivácie)

Na dokázanie kauzálnej súvislosti medzi javom A a dôsledkom D sa najčastejšie používa tzv. **klasický experiment.** Klasický experiment spočíva v skúmaní dvoch skupín súčasne (tzv. experimentálna skupina S1 a kontrolná skupina S2), pričom dané skupiny sú sledované pred vystavením javu A (inak tiež experimentálnej premennej) a po vystavení experimentálnej premennej. Skúmanie takto rozdielnych skupín má **predísť** mylným záverom, ktoré môžu súvisieť s **vplyvom inej neznámej premennej** (iného javu než je jav A) ak by sa skúmala iba experimentálna skupina. Teda, aby sa predišlo náhodnému pozitívnemu alebo negatívnemu výsledku. K tomuto účelu pomáha aj kontrolované prostredie, v ktorom sa experiment realizuje a jeho vopred stanovené podmienky. Príkladom môže byť testovanie nových liekov, pričom kontrolná skupina nedostáva žiadne lieky alebo dostáva tzv. placebo.

Literatúra

BEDFORD, T., COOKE, R. *Probalistic Risk Analysis: Foundations and Methods*. 7. vyd. Cambridge: Cambridge Press, 2011.

BUDÍKOVÁ, M, KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. Praha: GRADA, 2010.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

EVANS, J. D. *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole Publishing: Pacific Grove, CA, USA, 1996.

FELLER, W. *An introduction to probability theory and its applications*. 1970. I. a II., New York: J. Wiley.

HINDLS, R., ARLTOVÁ, M., HRONOVÁ, S., MALÁ, I., MAREK, L., PECÁKOVÁ, I., ŘEZANKOVÁ, H. *Statistika v ekonomii*. Praha: Professional Publishing, 2018, ISBN 978-80-88260-09-7.

KOVAČKA, M., KONTEŠOVÁ, O. *Štatistické metódy*. 2. vyd. Bratislava: Slovenské vydavateľstvo technickej literatúry, 1962.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

ŠOLTÉS, E. *Regresná a korelačná analýza*. 2008. Iura Edition, spol. s r. o., člen skupiny Wolters Kluwer, Bratislava, ISBN 978-80-8078-163-7.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

VARGA, Š. *Matematická štatistika.* Bratislava: STU, 2012, ISBN 978-80-2273-789-0.

10 Predpovedanie – aplikácia regresnej úlohy

Ako už bolo uvedené v kapitole 9.3 korelačná analýza skúma korelačnú závislosť medzi kvantitatívnymi (číselnými) znakmi. Okrem miery (sily) korelačnej závislosti sa prostredníctvom **regresnej úlohy** skúma taktiež tzv. **regresia**, teda **priebeh závislosti medzi dvoma číselnými štatistickými znakmi.**

10.1 Koeficient determinácie

Pomerne vysoká hodnota koeficientu korelácie (napr. r = 0,75) znamená, že medzi premennými X a Y je vysoká vzájomná lineárna závislosť, ale to neznamená, že medzi premennými existuje aj vysoká príčinná závislosť, pretože môže existovať ďalšia premenná napr. Z, od ktorej je premenná Y taktiež lineárne závislá a ktorou sa lepšie vysvetlí variabilita hodnôt premennej Y. Stupeň príčinnej závislosti premenných X a Y určuje koeficient determinácie R^2.

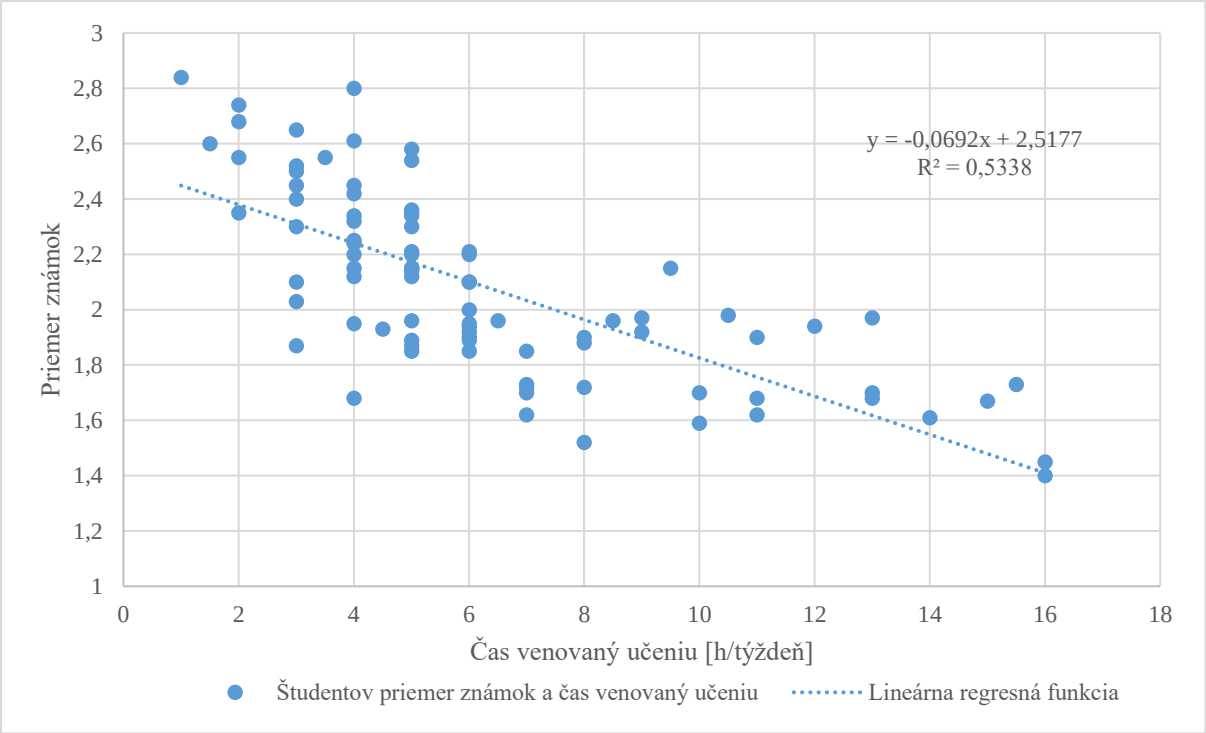
Stupeň príčinnej závislosti premennej Y od premennej X vyjadruje **koeficient determinácie (regresný koeficient, koeficient spoľahlivosti),** definovaný ako druhá mocnina koeficientu korelácie *r*. Označuje sa r^2 alebo $\boldsymbol{R^2}$.

Interpretácia koeficientu determinácie vychádza z analýzy variability (rozptylu) závislej premennej Y , ktorú by mala do značnej miery vysvetliť variabilita nezávisle premennej X. To platí za predpokladu, že od hodnôt premennej X lineárne závisí veľkosť hodnôt Y. Hodnota R^2 určuje percento celkovej variability závisle premennej, ktorá je vysvetlená lineárnou regresnou funkciou prostredníctvom vysvetľujúcej (nezávislej) premennej.

Koeficient determinácie pre lineárnu závislosť charakterizuje spoľahlivosť prípadnej predpovede, preto sa tiež nazýva **koeficient spoľahlivosti**. Udáva ako presne zodpovedajú predpokladané (očakávané) hodnoty, vyjadrené **regresnou funkciou** – **trendovou spojnicou** (trend, vývoj, smer, vyrovnanie meraných veličín), skutočným dátam (Obrázok 49). Inými slovami ako spoľahlivo (ako presne) vysvetľujú hodnoty premennej X (vysvetľujúca premenná) hodnoty premennej Y (vysvetľovaná premenná).

Príklad:

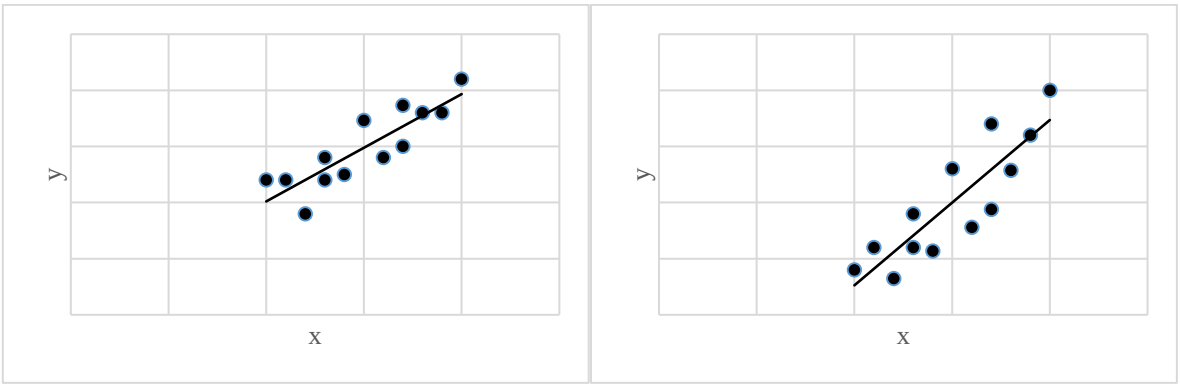
Vychádzajúc z príkladu v kap. 9.3 (závislosť medzi časom, ktorý študenti venujú učeniu a ich výsledkami za daný školský rok), je možné v grafe zobraziť trendovú spojnicu a jej rovnicu a taktiež koeficient determinácie (Obrázok 49).



Obrázok 49 Model lineárnej regresnej funkcie (lineárna trendová spojnica) a vyjadrenie jeho spoľahlivosti (R^2) pre skúmanie vzťahu medzi časom, ktorý študenti venujú učeniu a ich priemeru známok za daný školský rok

Pre daný príklad je koeficient korelácie r = -0,73, potom R^2 = 0,53, čo znamená, že 53% variability premennej Y sa dá vysvetliť lineárnym vzťahom s premennou X (regresnou priamkou). 47% variability premennej Y zostalo nevysvetlenej lineárnym vzťahom s premennou X. Je možné uvažovať resp. skúmať či neexistuje vhodnejší model na popísanie uvedeného vzťahu (namiesto lineárnej závislosti uvažovať o inej – nelineárnej závislosti – viď kap. 10.5).

Aj keď koeficient korelácie r má úzky vzťah k regresnému koeficientu R^2, neinformuje riešiteľa (čo sa často mýli) o sklone regresnej priamky. Pre regresné priamky s rôznymi sklonmi môže byť koeficient korelácie rovnaký (Obrázok 50), prípadne z opačného pohľadu môžu rôzne hodnoty koeficientu korelácie prislúchať regresným priamkam rovnakého sklonu.



Obrázok 50 Vzťahy medzi premennými X a Y s rovnakými koeficientami korelácie (r = 0,86) a rôznymi vyrovnávajúcimi priamkami

10.2 Regresná úloha

Regresná úloha korelačnej analýzy má za cieľ **popísať priebeh** skúmaného vzťahu číselných štatistických znakov a **použiť jej výsledky pri prognózach (predpovedaní) vývoja daného vzťahu.**

Regresná úloha (analýza) je označenie pre štatistickú metódu, pomocou ktorej sa teda **odhaduje hodnota náhodnej veličiny *y*** (tzv. závislej premennej, cieľovej premennej alebo vysvetľovanej premennej) **na základe poznania inej veličiny *x* (veličín)** (nezávisle premenných, regresorov alebo vysvetľujúcich premenných).

Ide o to, aby sa vyjadril **priebeh korelačnej závislosti** t. j. vyjadrenie zmien závislej premennej *y* na zmenách nezávislej premennej *x*. Tento vzťah sa nazýva **regresia**. Regresiu popisujú **regresné funkcie** (kap. 10.3). **Výsledkom regresnej úlohy (analýzy)** je teda **nájdenie** matematickej (regresnej) **funkcie**, ktorá čo najlepšie zobrazuje (popisuje, zachytáva) **priebeh závislosti** kvantitatívnych štatistických znakov.

Regresná úloha **meria všeobecne známe voľné závislosti** v konkrétnych podmienkach, môžeme však pomocou nej mapovať a vyhľadávať (príp. „objavovať“) aj doposiaľ neznáme závislosti.

Príklad:

Príkladom z bežného života môže byť situácia, ak sa ráno snaží osoba odhadnúť, približnú teplotu aká bude cez deň (regresant) na základe predchádzajúcej znalosti počasia a toho, aké je vonku počasie v danom momente (dva regresory).

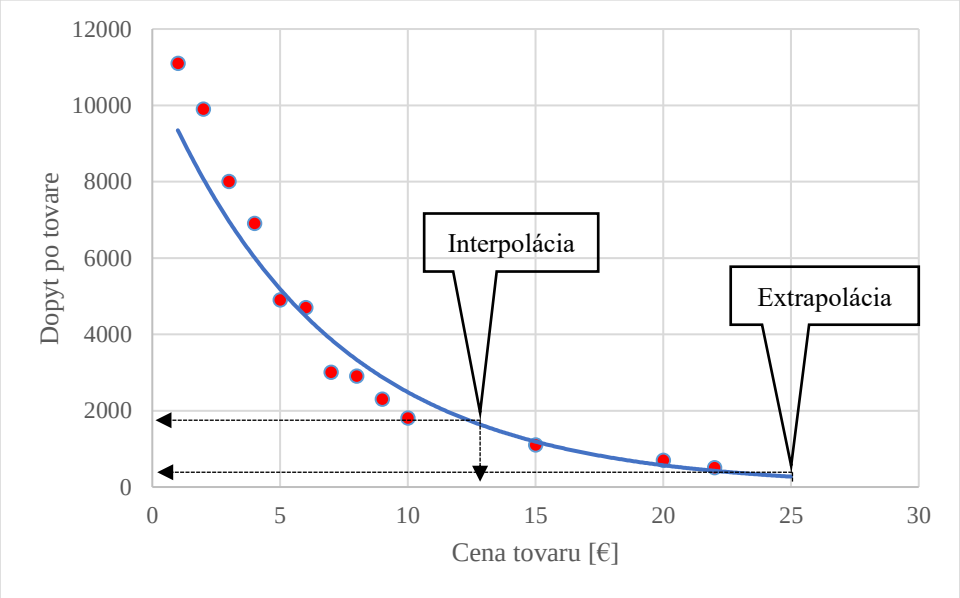
Príklad skutočnej regresnej analýzy v praxi je napríklad *odhadovanie očakávanej pooperačnej dĺžky života pacientov trpiacich rakovinou*.

Na základe údajov o zdravotnom stave väčšieho počtu pacientov, napríklad veľkosť a typ nádorov, vek pacientov a pod. (regresory) a dĺžky ich života po operácii (regresant), je možné pomocou vhodného typu regresnej analýzy (v tomto prípade obvykle tzv. Coxovej regresie) stanoviť vzorec, s ktorého pomocou bude možné u nového pacienta na základe znalosti jeho zdravotného stavu odhadnúť strednú hodnotu očakávanej doby prežitia v prípade operácie. Ak je naviac k dispozícií podobná analýza pre pacientov liečených konzervatívne, je možné tomuto novému pacientovi odporučiť, ktorý spôsob liečby mu v danej situácii dáva nádej na dlhšie prežitie.

Zmysel má zaoberať sa podrobnejšie iba závislosťami s vyššou intenzitou $R^2 = <0{,}7\text{-}0{,}99>$, u ktorých je rozptyl pozorovaných hodnôt okolo regresnej funkcie malý. Regresná funkcia v takýchto prípadoch popisuje priebeh závislosti spoľahlivo a umožňuje napr.:

- **interpoláciu** – odhad (predpovedanie) hodnôt závislej premennej *y* pre nenamerané hodnoty nezávislej premennej *x* **vo vnútri** intervalu meraných hodnôt,
- **extrapoláciu** – odhad (predpovedanie) hodnôt závislej premennej *y* na základe priebehu regresnej funkcie **mimo interval** nameraných hodnôt nezávislej premennej *x*,

- **odhad obtiažne merateľných hodnôt** závislej premennej *y* na základe ľahko merateľných hodnôt nezávislej premennej *x*.



Obrázok 51 Interpolácia a extrapolácia hodnôt na základe regresnej funkcie

10.3 Základné typy regresných funkcií a ich aplikácia

Pri regresnej analýze sa používajú rôzne typy regresných funkcií resp. trendových spojníc, ktoré môžu mať rôzny tvar. Najčastejšie sa používajú regresné funkcie:

- lineárna (regresná priamka),
- exponenciálna (regresná exponenciála),
- mocninová,
- logaritmická,
- polynomická (napr. regresná parabola).

Regresné funkcie sa používajú na vyrovnávanie získaných bodov korelačného grafu a následne na predpovedanie neznámych hodnôt závislej premennej.

10.3.1 Lineárna regresná funkcia

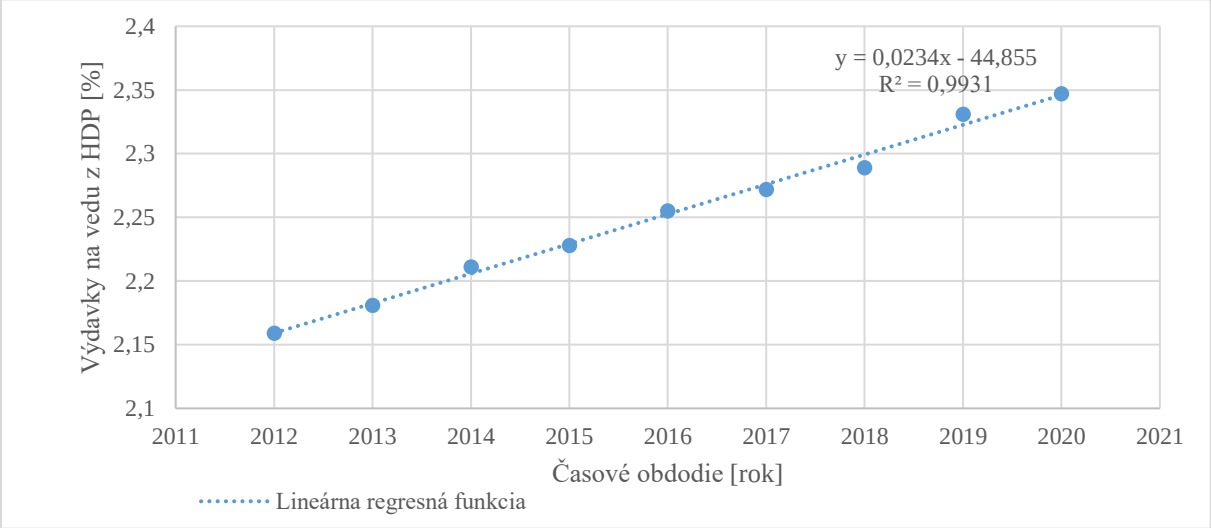
Lineárna regresná funkcia (trendová spojnica) je prispôsobená priamka používaná pri jednoduchých lineárnych množinách údajov. Vzťah medzi štatistickými znakmi je lineárny, ak priebeh dátových bodov (body kombinácií obmien štatistických znakov v grafe) pripomína priamku. Lineárna trendová spojnica obvykle zobrazuje, že hodnoty závislej premennej *y* sa menia (rastú alebo klesajú) konštantnou mierou:

$$y = b_0 + b_1 x$$

Príklad:

Riešiteľ skúma vývoj výdajov na vedu z HDP Slovenska (percentuálny podiel) za vybrané časové obdobie. Na skúmanie (vyrovnanie dátových bodov regresnou funkciou) použil riešiteľ

lineárnu funkciu (lineárnu trendovú spojnicu). Inak je možné použiť taktiež označenie lineárny model vývoja výdavkov na vedu z HDP za vybrané časové obdobie (Obrázok 52).



Obrázok 52 Vývoj výdavkov na vedu z HDP Slovenska za sledované časové obdobie (vyrovnanie lineárnou regresnou funkciou)

10.3.2 Mocninová regresná funkcia

Mocninová regresná funkcia (trendová spojnica) je krivka používaná pri údajoch porovnávajúcich stúpajúce hodnoty namerané v určitých intervaloch. Napríklad zrýchlenie auta v intervaloch po 1 sekunde. Mocninovú trendovú spojnicu nie je možné vytvoriť, ak dáta obsahujú nulové alebo záporné hodnoty.

$$y = b_0 x_i^{b_1}$$

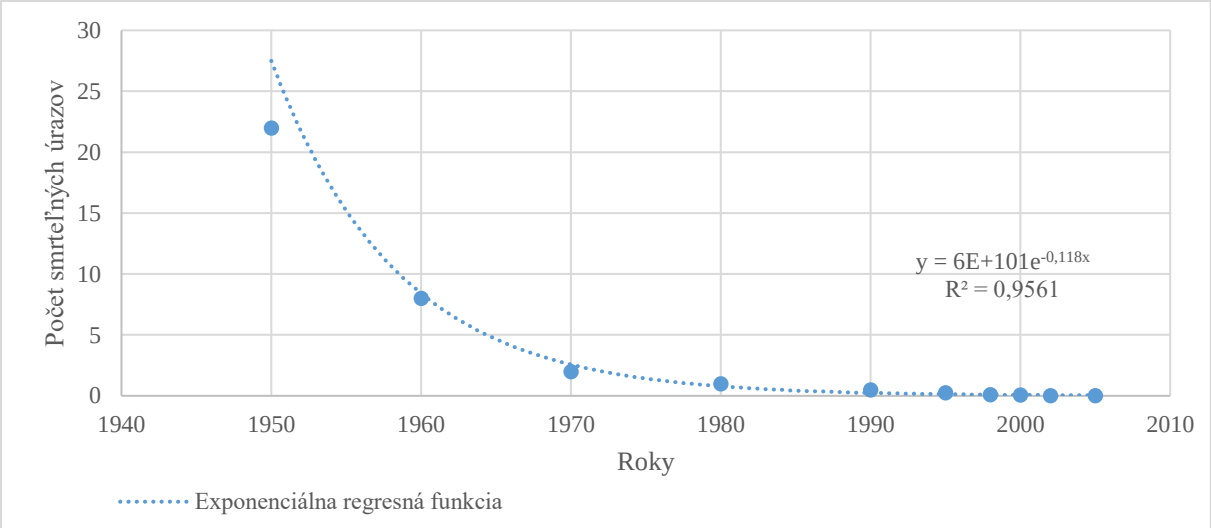
10.3.3 Exponenciálna regresná funkcia

Exponenciálna regresná funkcia (trendová spojnica) je krivka, ktorá sa používa v prípade, že hodnoty údajov stúpajú alebo klesajú v stále väčších krokoch. Túto spojnicu nie je možné vytvoriť ak dáta obsahujú nulové alebo záporné hodnoty.

$$y = b_0 e^{b_1 x_i}$$

Príklad:

Bol skúmaný počet smrteľných úrazov pripadajúcich na 1 milión nalietaných kilometrov v rozmedzí rokov 1950 až 2005. Súbor bol vyrovnaný exponenciálnou funkciou.



Obrázok 53 Vývoj smrteľných úrazov v leteckej doprave na 1 mil. nalietaných kilometrov za roky 1950 - 2005

Iný príklad môže predstavovať štatistický súbor vytvorený z údajov o mimoriadnych udalostiach vo svete. Vzhľadom na sledované obdobie sa počty mimoriadnych udalostí exponenciálne zvyšujú. Podobne to môže byť so škodami po mimoriadnych udalostiach.

10.3.4 Logaritmická regresná funkcia

Logaritmická regresná funkcia (trendová spojnica) je prispôsobená krivka používaná pri vývoji údajov, ktoré rýchlo stúpajú alebo klesajú a postupne sa ich hodnoty vyrovnávajú. Pri logaritmickej trendovej spojnici je možné použiť kladné i záporné hodnoty.

$$y = b_0 \ln(x_i) + b_1$$

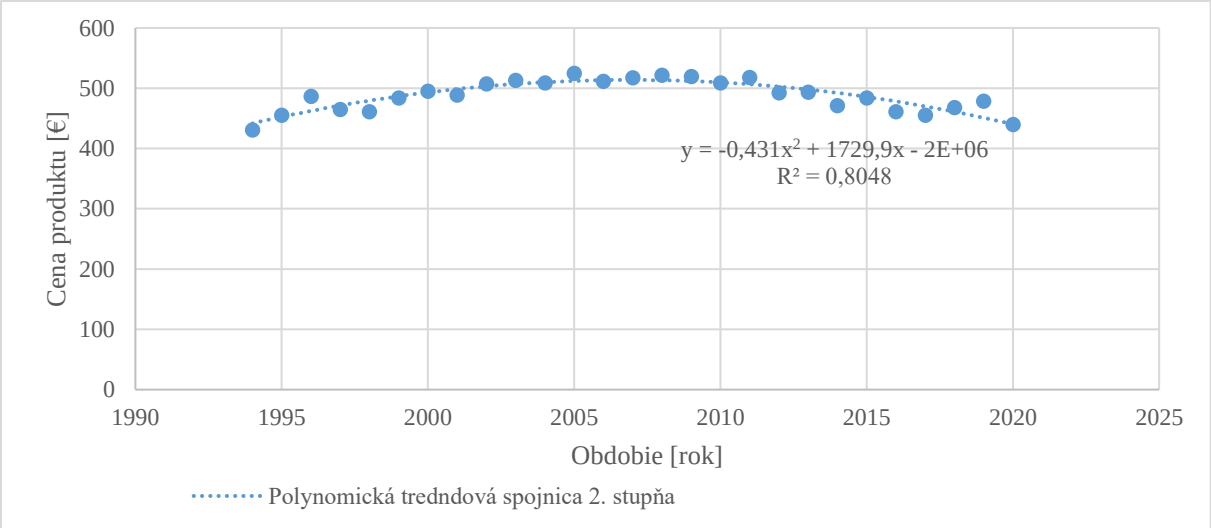
10.3.5 Polynomická regresná funkcia

Polynomická regresná funkcia (trendová spojnica) je krivka používaná pri údajoch, ktoré kolíšu a nedajú sa teda aproximovať jednoduchšou funkciou. Stupeň polynómu *n* môže byť určený počtom kolísaní v dátach alebo počtom zakrivení (maxim a minim) v krivke. Stupeň 2 má obvykle jeden vrchol. Stupeň 3 má obvykle jeden alebo dva vrcholy. Stupeň 4 má obvykle až tri vrcholy.

$$y = b_0 + b_1 x + b_2 x^2 + \cdots + b_n x^n$$

Príklad:

Vyrovnanie ceny vybraného produktu za sledované časové obdobie polynomickou funkciou (Obrázok 54).



Obrázok 54 Vývoj ceny vybraného produktu v čase (vyrovnanie polynomickou regresnou funkciou 2. stupňa)

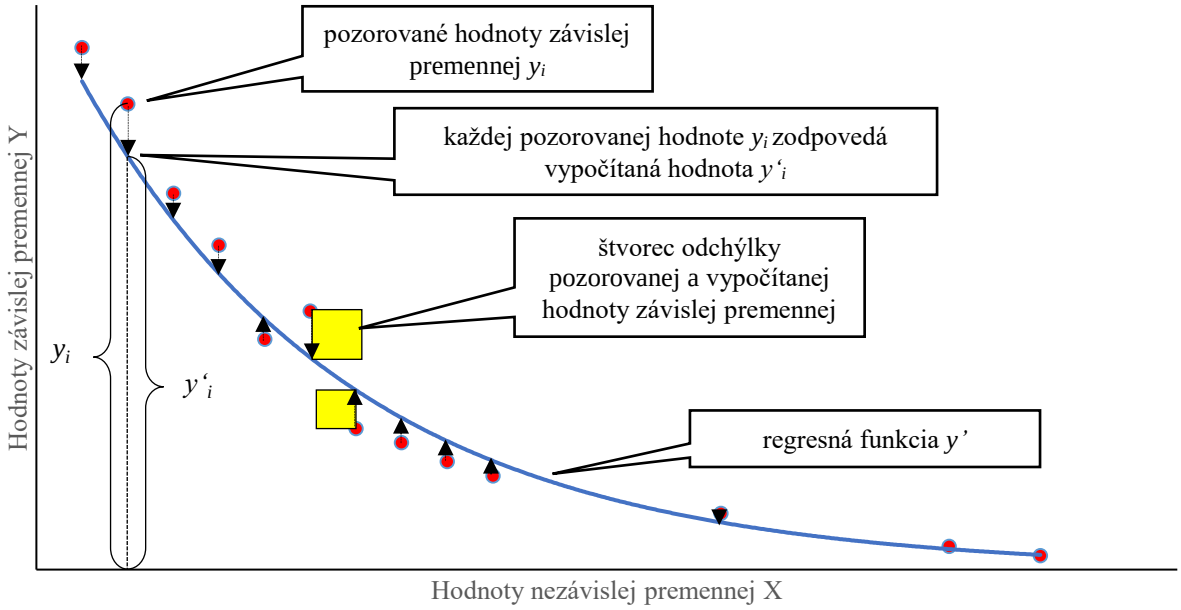
10.4 Metóda najmenších štvorcov

Metóda najmenších štvorcov je univerzálnou metódou stanovovania (odhadu) parametrov $b_0, b_1,..., b_m$ funkcie, ktoré nahrádzajú pôvodne namerané hodnoty y_i závislej premennej Y.

Znamená to, že sa **hľadá funkcia**, ktorá má súčet štvorcov odchýlok meraných údajov od teoretických čo najmenší. V **geometrickej predstave** to znamená, že hľadáme takú krivku, ktorá čo najtesnejšie prilieha k jednotlivým bodom.

Funkcia tejto krivky by mala byť čo najjednoduchšia, aby sa dala ľahko používať k výpočtu ďalších potrebných hodnôt. Táto funkcia sa nazýva **regresná funkcia**.

Výber typu funkcie (t. j. napr. mocninová, lineárna, exponenciálna a pod.) je v kompetencií riešiteľa úlohy. Metóda najmenších štvorcov potom nájde parametre resp. „najlepšiu“ funkciu vopred zvoleného typu.



Obrázok 55 Grafické znázornenie metódy kritéria najmenších štvorcov

Metóda najmenších štvorcov minimalizuje súčet štvorcov odchýlok pozorovaných (nameraných) hodnôt závislej premennej a zvolenej regresnej funkcie. Spočíva teda v hľadaní takej regresnej funkcie, pre ktorú bude platiť vzťah:

$$\sum_{i=1}^{n}(y_i - y_i')^2 = min$$

Platí pre funkcie lineárne aj nelineárne, jednoduché aj viacnásobné.

Ak je rozsah súboru rovný n, je kritérium najmenších štvorcov:

$$\sum_{i=1}^{n}(y_i - y_i')^2 = \sum_{i=1}^{n}\left[y_i - \sum_{j=0}^{m} b_j f_j(x_i)\right]^2 \rightarrow min$$

Dá sa preukázať, že ak vyhovuje určitá funkcia kritériu najmenších štvorcov, spĺňa automaticky tiež (súčet kladných a záporných odchýlok okolo regresnej funkcie sa kompenzuje):

$$\sum_{i=1}^{n}(y_i - y_i') = 0$$

Táto podmienka však regresnú funkciu neurčuje jednoznačne. Existuje jediná regresná funkcia zvoleného typu, ktorá pre konkrétne dáta vyhovuje podmienke najmenších štvorcov.

10.5 Praktický postup riešenia regresnej úlohy

1. Definovanie nezávislej x a závislej premennej y a zistenie ich hodnôt niektorou zo známych štatistických metód (pozorovanie, meranie, dopytovanie).
2. Zostrojenie korelačného (bodového) grafu (využitie softvérových nástrojov).

3. Vloženie dostupných regresných grafov, funkcií a koeficientov spoľahlivosti R^2 (využitie softvérových nástrojov).
4. Výber najvhodnejšej regresnej funkcie s použitím exaktného, logického a vizuálneho prístupu (pohľadu):
    - **exaktný prístup** pre výber regresnej funkcie spočíva vo výbere funkcie s najvyšším koeficientom spoľahlivosti R^2; tento koeficient je možné interpretovať tiež ako pravdepodobnosť predpovede,
    - **logický prístup** pre výber regresnej funkcie spočíva vo výbere funkcie, ktorej priebeh najlepšie vystihuje vzťah nezávislej a závislej premennej; ide napríklad o situáciu, kedy musí regresná funkcia vychádzať nevyhnutne z bodu [0, 0],
    - **vizuálny** (praktický) **prístup** spočíva vo vizuálnom vzájomnom posúdení vložených regresných funkcií z hľadiska ich praktickej aplikácie pri predpovedi.
5. Grafická predpoveď – odhad, pre zvolenú hodnotu nezávislej premennej (využitie softvérových nástrojov).
6. Matematická predpoveď – výpočet, pre zvolenú hodnotu nezávislej premennej dosadením hodnoty nezávislej premennej *x* do vybranej regresnej funkcie.

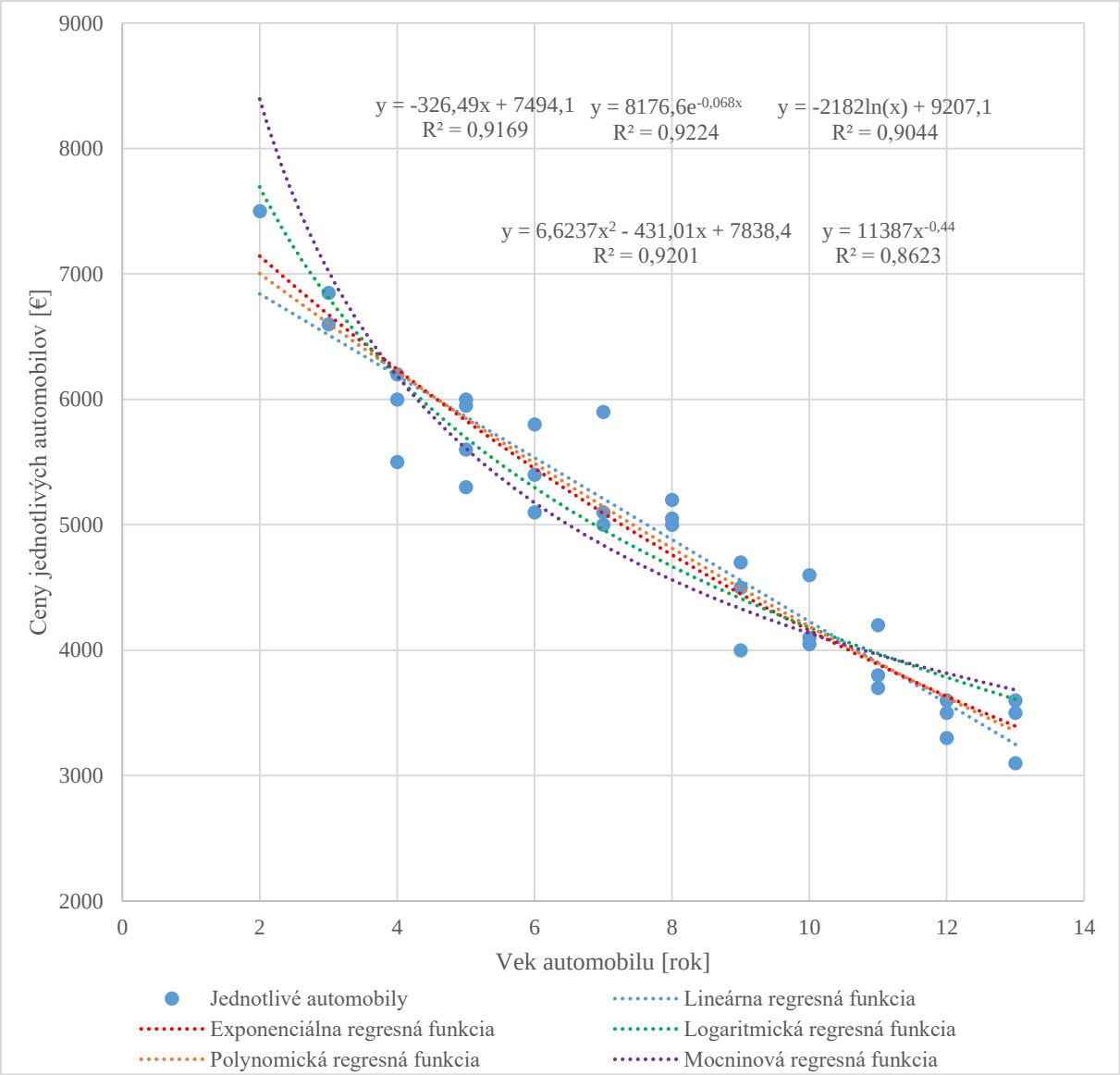
Poznámka: Hodnotu nezávislej premennej *x* môžeme pri extrapolácií zvoliť v rôznej vzdialenosti od pôvodných údajov. Teória nerieši, či je koeficient spoľahlivosti stále rovnaký bez ohľadu na túto vzdialenosť. Logická úvaha však vedie k hypotéze, že čím je zvolená hodnota na osi x vzdialenejšia od pôvodných údajov, tým by spoľahlivosť predpovede závislej premennej mala klesať.

Príklad 1:

Na základe štatistických údajov získaných pozorovaním v autobazároch nám regresná úloha umožňuje skúmať vývoj cien automobilov (napr. konkrétneho typu a značky) vzhľadom na ich vek (Obrázok 56) a predpovedať ceny ďalších automobilov prostredníctvom výberu vhodnej regresnej funkcie.

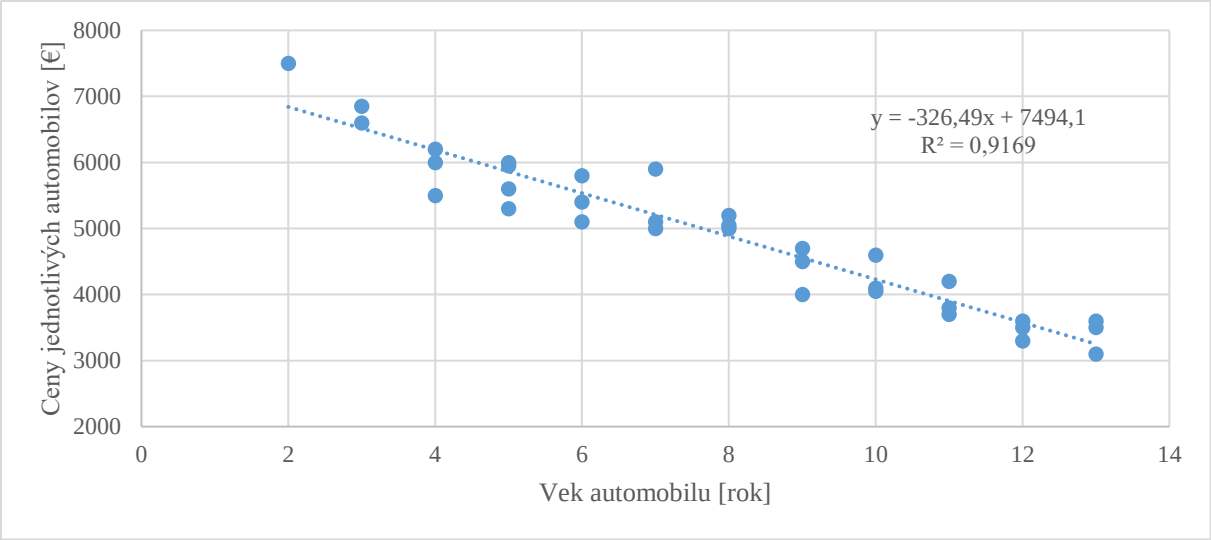
Štatistická otázka pre daný príklad: Akú cenu ojazdených automobilov je možné očakávať po 15 rokoch ich prevádzky?

Nezávislá pramenná *x* je vek automobilov. Závislá pramenná *y* je cena automobilov v autobazároch.

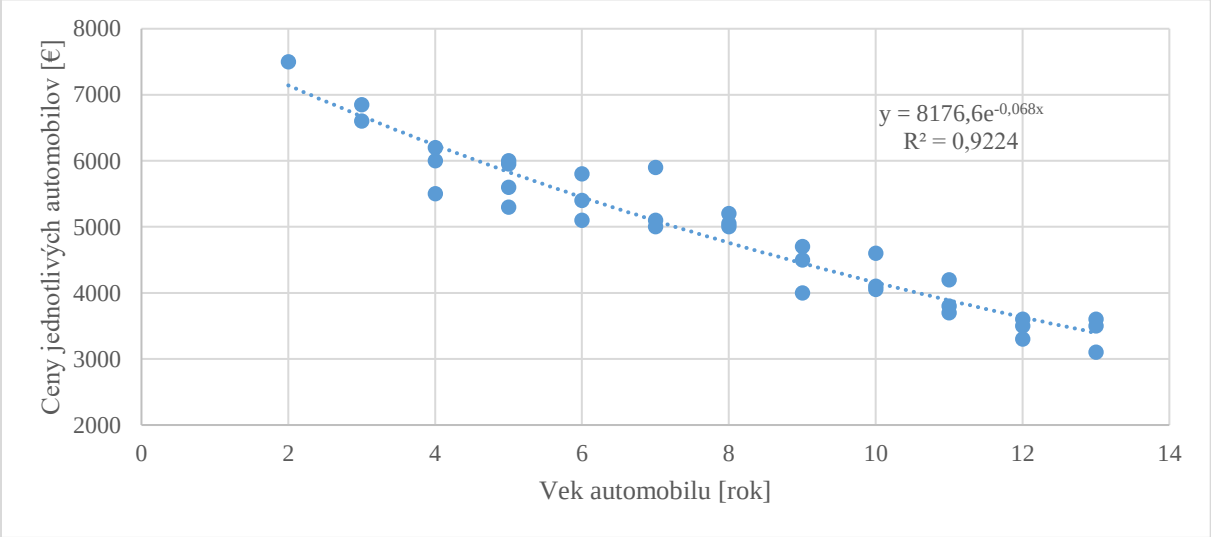


Obrázok 56 Korelačný graf vzťahu veku a ceny automobilov s vloženými regresnými čiarami

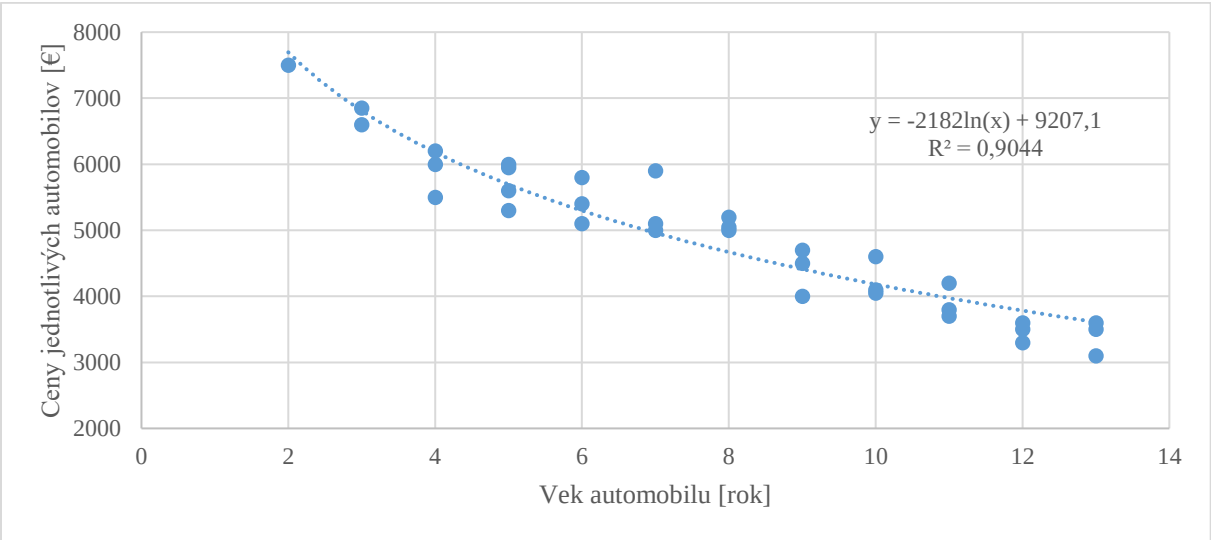
Vloženie viacerých (dostupných) regresných funkcií môže obmedziť prehľadnosť v korelačnom grafe a sťažiť následné posúdenie vhodnosti regresných funkcií pre účely predpovede. Niekedy je preto vhodné regresné funkcie vkladať do samostatných korelačných grafov (viď nasledujúce grafy) a posúdiť ich vhodnosť pre účely predpovede samostatne a až následne porovnaním s ostatnými.



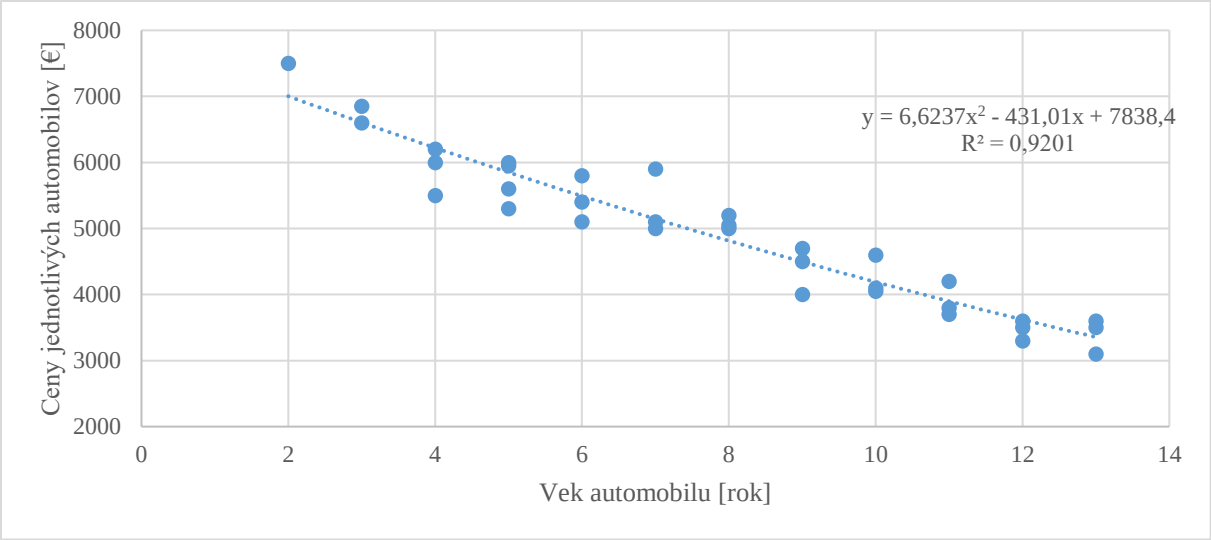
Obrázok 57 Lineárny model vývoja vzťahu veku a ceny automobilov



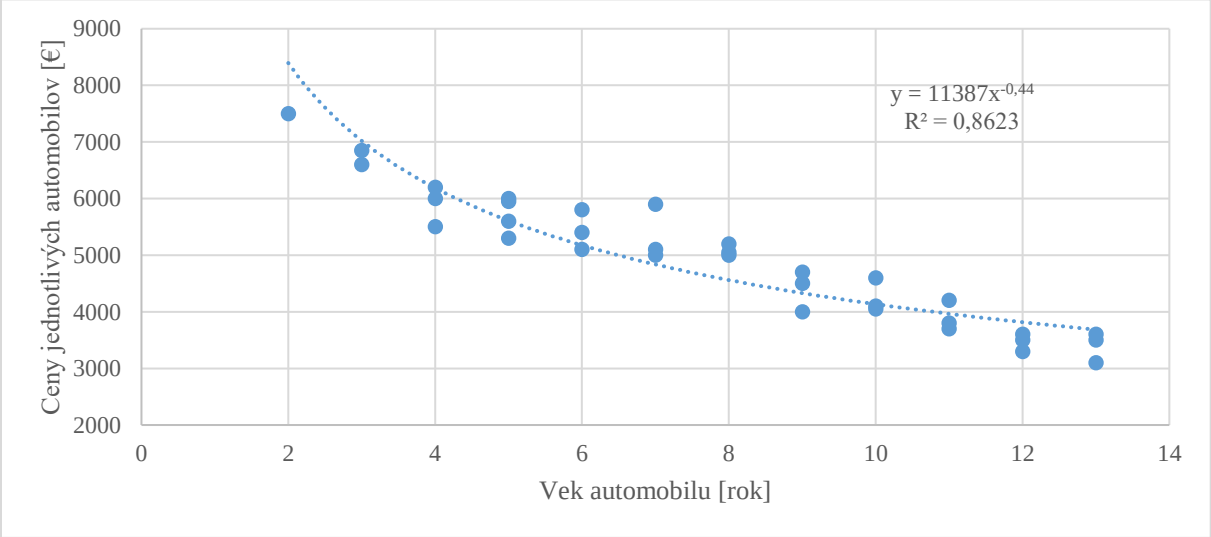
Obrázok 58 Exponenciálny model vývoja vzťahu veku a ceny automobilov



Obrázok 59 Logaritmický model vývoja vzťahu veku a ceny automobilov



Obrázok 60 Polynomický model (2. stupňa) vývoja vzťahu veku a ceny automobilov



Obrázok 61 Mocninový model vývoja vzťahu veku a ceny automobilov

Výber regresnej funkcie:

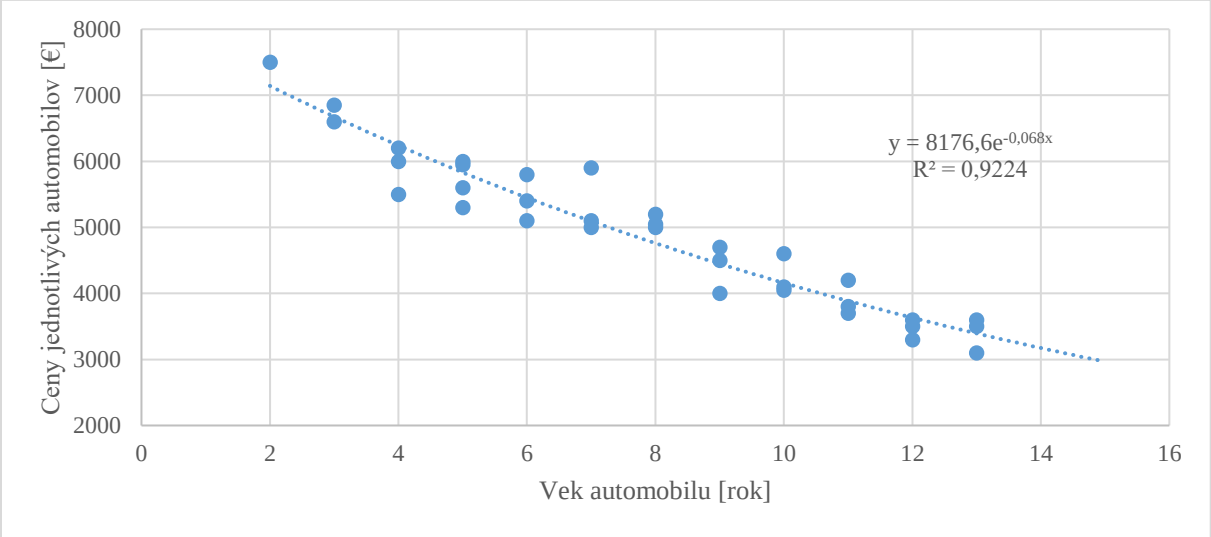
Z exaktného pohľadu je najspoľahlivejšia exponenciálna regresná funkcia s koeficientom spoľahlivosti $R^2 = 0{,}9224$.

Z logického pohľadu je najvýhodnejšia logaritmická funkcia s $R^2 = 0{,}9044$, ktorá predovšetkým pri mladších automobiloch lepšie koreluje s množinou údajov korelačného grafu. Keďže predpoveď je zamýšľaná pre staršie automobily, tak aj ostatné funkcie okrem lineárnej sú logicky prijateľné. Zjavne nelogická je lineárna funkcia, ktorá bude zjavne pri vyšších hodnotách veku pretínať os x.

Z vizuálneho (praktického) pohľadu bude zaujímavé skúmať ceny automobilov zhruba od 3 do 10 rokov veku. V tomto intervale sú prijateľné predpovede podľa všetkých regresných funkcií. Pri predpovedi ceny starších automobilov je potrebné vylúčiť lineárnu regresnú funkciu. Pri predpovedi cien mladších automobilov je vhodné vylúčiť mocninovú funkciu, ktorá na danom úseku príliš prudko stúpa.

Na základe predchádzajúcich úvah sa prikloníme k predpovedi podľa exponenciálnej regresnej funkcie $y = 8176{,}6e^{-0{,}068x}$. Vhodnosť daného modelu je vhodné ešte testovať (používa sa napr. F-test alebo iné), aby sme pri určitej hladine spoľahlivosti (prevažne sa používa α=0,05) mohli tvrdiť, že model bol zvolený správne (testovanie modelov nie je súčasťou týchto skrípt).

Grafická predpoveď:



Obrázok 62 Predpoveď ceny automobilu vo veku 15 rokov exponenciálnou regresnou funkciou

Matematická predpoveď:

$y = 8176{,}6e^{-0{,}068x} = 8176{,}6e^{-0{,}068*15} = 2948{,}22$ €

Grafická a matematická predpoveď je pre uvedený príklad vysoko spoľahlivá, keďže koeficient spoľahlivosti $R^2 = 0{,}9224$. Výklad hodnoty R^2 môže byť rôzny – koeficient spoľahlivosti prípadnej predpovede, udáva ako presne zodpovedajú predpokladané (očakávané) hodnoty, vyjadrené regresnou funkciou – trendovou spojnicou (trend, vývoj, smer, vyrovnanie meraných veličín), skutočným dátam. Trendová spojnica je najspoľahlivejšia v prípade, že sa hodnota indexu R^2 spoľahlivosti blíži alebo rovná hodnote 1, vtedy predpokladané hodnoty lepšie odrážajú skutočnosť. Koeficient regresie je možné vyjadriť aj pravdepodobnostne – cenu ojazdeného automobilu je možné vo vybranom modeli odhadnúť s pravdepodobnosťou 0,9224. Odhadnutá cena bude správna u 9 z 10 ďalších automobilov.

Presnosť regresnej funkcie je priamo závislá na rozsahu súboru. Presnosť matematického predpovedania je úmerná veľkosti korelačnej závislosti.

Príklad 2:

Experimentálne meranie objemu škodlivých látok v spalinách pri rôznych teplotách horenia je príkladom na využitie odhadu (dopočítania) objemu škodlivých látok pri vyšších teplotách. Tento spôsob je podstatne rýchlejší a lacnejší ako určovanie objemu škodlivín chemickým rozborom pre všetky záujmové teploty. Nezávislou premennou x, ktorá je súčasne kontrolovaná riešiteľom, je teplota horenia a závislou premennou y je objem škodlivej látky v spalinách.

Literatúra

BEDFORD, T., COOKE, R. *Probalistic Risk Analysis: Foundations and Methods*. 7. vyd. Cambridge: Cambridge Press, 2011.

BUDÍKOVÁ, M, KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. Praha: GRADA, 2010.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

EVANS, J.D. *Straightforward Statistics for the Behavioral Sciences*; Brooks/Cole Publishing: Pacific Grove, CA, USA, 1996.

FELLER, W. *An introduction to probability theory and its applications*. 1970. I. a II., New York: J. Wiley.

HINDLS, R., ARLTOVÁ, M., HRONOVÁ, S., MALÁ, I., MAREK, L., PECÁKOVÁ, I., ŘEZANKOVÁ, H. *Statistika v ekonomii*. Praha: Professional Publishing, 2018, ISBN 978-80-88260-09-7.

KOVAČKA, M., KONTEŠOVÁ, O. *Štatistické metódy*. 2. vyd. Bratislava: Slovenské vydavateľstvo technickej literatúry, 1962.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

ŠOLTÉS, E. *Regresná a korelačná analýza*. 2008. Iura Edition, spol. s r. o., člen skupiny Wolters Kluwer, Bratislava, ISBN 978-80-8078-163-7.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.

VARGA, Š. *Matematická štatistika.* Bratislava: STU, 2012, ISBN 978-80-2273-789-0.

11 Časové rady

Špecifickými štatistickými údajmi sú **časové rady**, pomocou ktorých je možné skúmať **dynamiku javov v čase**. Časovým radom (dynamický rad, vývojový rad) sa chápu **v čase usporiadané** číselné (kvantitatívne) údaje.

Časové rady sú určené predovšetkým:

- na sledovanie a vyhodnocovanie zmien, ku ktorým dochádza vo vývoji skúmaných javov v závislosti na čase,
- na analýzu príčin, ktoré na tieto javy pôsobili a ovplyvňovali ich správanie sa v minulosti,
- na predvídanie ich budúceho vývoja.

Hodnoty časového radu sa označujú symbolom Y_t, kde t predstavuje čas. Odhadnutá hodnota časového radu sa označuje $\widehat{Y_t}$.

Množinu hodnôt časového radu až do časového bodu t značíme $Y_1, Y_2, \ldots, Y_{t-1}, Y_t$. Ak riešiteľ pracuje s viacerými časovými radmi naraz, používa pre ich označenie ďalšie písmená z konca abecedy – *Z, X* atď.

V matematickom vyjadrení je časový rad časovou postupnosťou pozorovaných hodnôt číselného štatistického znaku $y_1, y_2, \ldots, y_t, \ldots, y_n$, , pre y_t, $t=1, 2,\ldots, n$, kde n je dĺžka časového radu.

Rozdiel $n - t$ sa nazýva vek pozorovania vyjadrený v rôznych (požadovaných) časových jednotkách.

Časové rady môžu byť spojité a nespojité. Veľa radov, ktoré majú nespojitý charakter, sa často prevádza na rady spojité sčítaním, priemerovaním a pod. Často sa tak deje pri ekonomických časových radoch. Napríklad *objem výroby v podniku* (zaujíma nás výroba za mesiac, štvrťrok), *priemerná denná teplota, tlak* a pod.

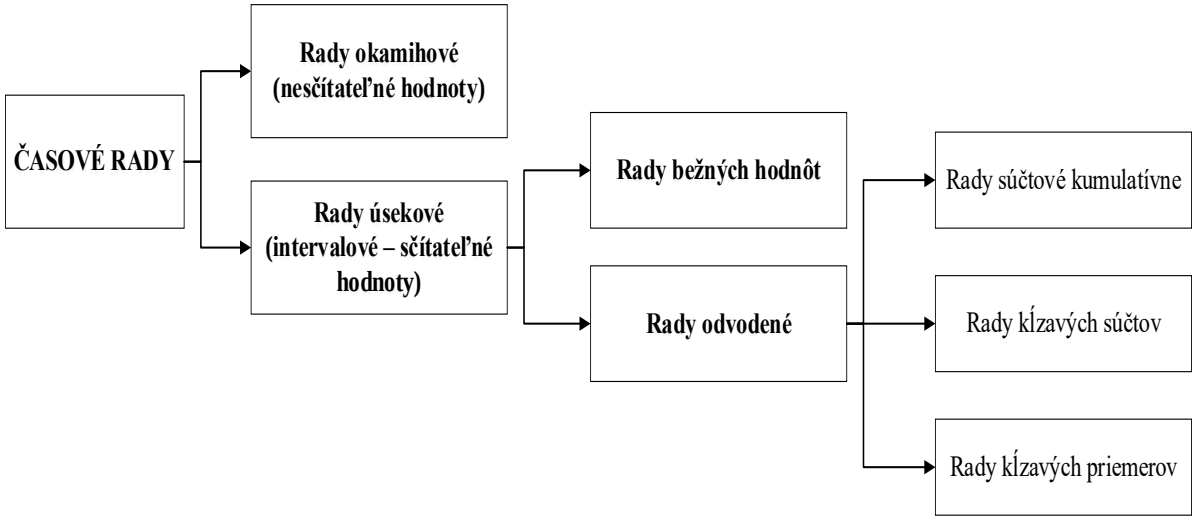
Problémy časových radov:

Pri spracovaní údajov vo forme časových radov je možné sa stretnúť s množstvom problémov:

- problémy s voľbou časových bodov pozorovanie,
- problémy s kalendárom,
    - rôzna dĺžka mesiacov,
    - rôzny počet víkendov v mesiaci,
    - rôzny počet pracovných dní v mesiaci,
    - pohyblivé sviatky,
- problémy s dĺžkou časových radov,
- problémy (ne)porovnateľnosti údajov,

11.1 Klasifikácia časových radov

Základné delenie časových radov poskytuje nasledujúca schéma:



Obrázok 63 Klasifikácia časových radov

Údaje **okamihových časových radov** sa vzťahujú vždy k určitému časovému okamihu napr. počet pracovníkov k prvému dňu v jednotlivých mesiacoch, stav zásob materiálu k 1.1. v jednotlivých rokoch, údaje o teplote vzduchu. Ide o nesčítateľné hodnoty.

Údaje **úsekových časových radov** sa vzťahujú vždy k určitému časovému úseku. Veľkosť údajov je v priamej závislosti s dĺžkou časových úsekov, napr. počty výrobkov v jednotlivých mesiacoch roku, počet narodených detí v jednotlivých rokoch. Typické je sčítanie (kumulovanie) údajov. Ide o sčítateľné hodnoty.

Údaje **časových radov bežných hodnôt** predstavujú neupravené hodnoty získané za daný časový úsek.

Údaje **odvodených časových radov** predstavujú upravené bežné hodnoty do súčtových alebo priemerných hodnôt za určité časové úseky:

- časové rady **súčtové** (kumulatívne) – umožňujú sledovať postupné narastanie ukazovateľov od prvého časového úseku až po posledný,
- časové rady **kĺzavých súčtov** – hodnoty ukazovateľa za obdobie pozostávajúce z určitého počtu čiastkových úsekov, pričom každý ďalší prírastok (úhrn) v rade priberá údaj najnovšieho úseku a vypúšťa údaj najstaršieho úseku,
- časové rady **kĺzavých priemerov** – rady kĺzavých súčtov podelené počtom úsekov, za ktoré sú kĺzavé súčty počítané.

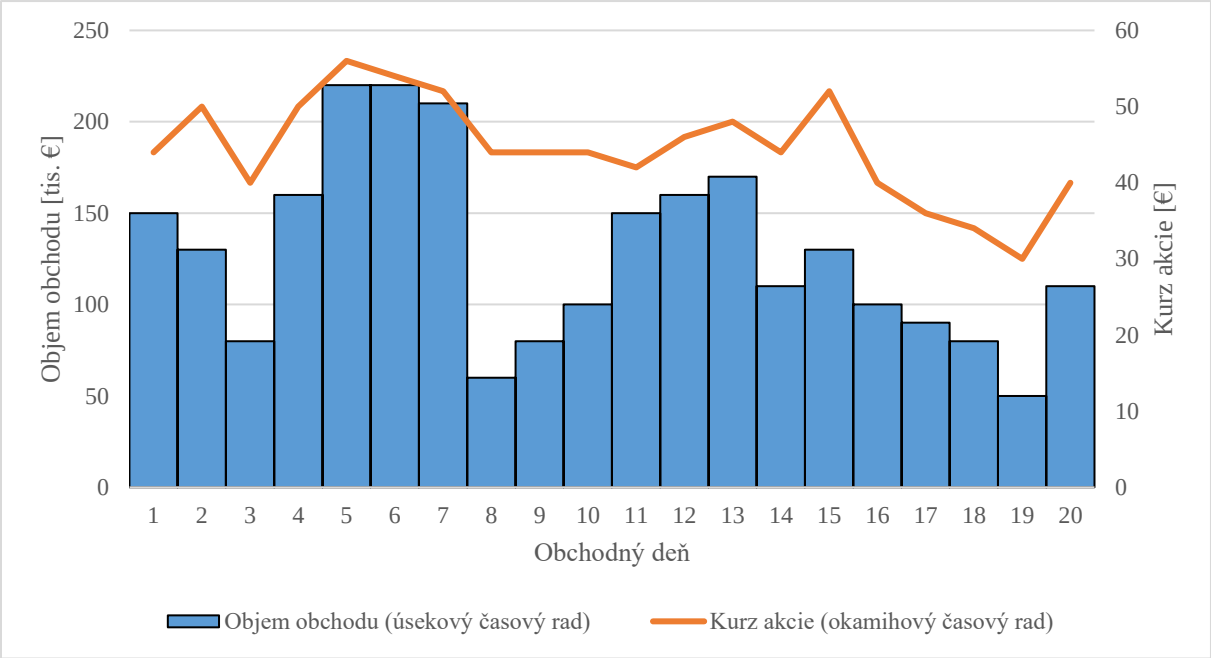
Príklady:

Časové rady pomerných veličín - napr. *plnenie plánu v jednotlivých mesiacoch, produktivita práce dosiahnutá v jednotlivých rokoch,*

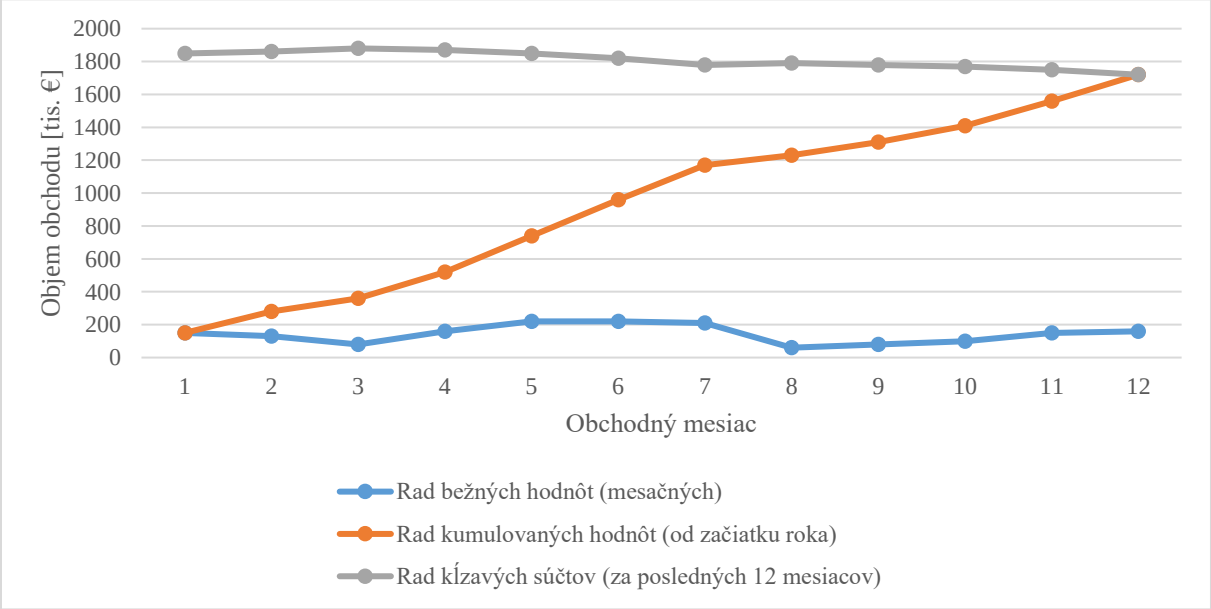
Časové rady priemerných veličín - napr. *priemerná mzda pracovníkov v jednotlivých rokoch, priemerná spotreba mäsa na jedného obyvateľa v jednotlivých rokoch.*

V ekonomickej oblasti sú typické napr. úsekové a okamihové časové *rady denných, týždenných, mesačných, štvrťročných, ročných údajov.*

Pri grafickom znázorňovaní úsekových časových radov sa používajú najmä stĺpcové grafy, stupňovité čiarové a spojnicové grafy, typické je vynášanie hodnôt nad stredmi úsekov (Obrázok 64). Z kombinácie radu bežných hodnôt, kumulovaného radu a radu kĺzavých súčtov sa zostavuje tzv. Z – diagram (Obrázok 65).



Obrázok 64 Kurz akcií a objem obchodu za 20 obchodných dní (príklad na okamihový a úsekový časový rad)



Obrázok 65 Z diagram pre objem obchodovaných akcií

11.2 Analýza časových radov

Cieľom analýzy časových radov je väčšinou **konštrukcia vhodného modelu** vývoja časového radu. Ak je riešiteľ schopný zostrojiť kvalitný model, umožní mu to porozumieť mechanizmu, na ktorého základe „vznikajú“ hodnoty časového radu, a porozumieť podmienkam, ktoré vznik týchto hodnôt ovplyvňujú. To umožňuje tieto podmienky následne ovplyvňovať a v niektorých prípadoch ovplyvniť aj vývoj časového radu. Ďalším veľmi častým cieľom je **konštrukcia predpovedí** na základe zistených údajov a zostrojeného modelu.

Pri klasickej analýze časových radov sa vychádza z predpokladu, že každý časový rad môže obsahovať štyri zložky:

- trend – T_t,
- sezónnu zložku – S_t,
- cyklickú zložku – C_t,
- náhodnú zložku – ε_t.

Trend je všeobecná tendencia vývoja skúmaného javu za dlhé obdobie. Je výsledkom dlhodobých a stálych procesov. Trend môže byť rastúci, klesajúci alebo môže existovať rad bez trendu.

Sezónna zložka je pravidelne sa opakujúca odchýlka od trendovej zložky. Perióda tejto zložky je menšia než celková veľkosť sledovaného obdobia.

Cyklická zložka udáva kolísanie okolo trendu v dôsledku dlhodobého cyklického vývoja (používané skôr v makroekonomických úvahách).

Náhodná (stochastická) zložka sa nedá popísať žiadnou funkciou času. "Zostáva" po vylúčení trendu, sezónnej a cyklickej zložky.

Najčastejšie sa pri analýze časového radu predpokladá aditívny model popisu správania sa radu. Predpokladá sa, že jednotlivé zložky vývoja sa sčítajú y_t , takže platí:

$$y_t = T_t + S_t + C_t + \varepsilon_t$$

Rôzne modifikácie modelov vzniknú, keď sa niektorá zložka z úvah vypustí.

Analýza zložky ktoréhokoľvek typu sa vykonáva v podstate klasickou regresnou analýzou. Podstatný rozdiel je iba v tom, že nezávislá premenná x, je v tomto prípade časová premenná a je možné ju v podstate ľubovoľne vyjadriť v akýchkoľvek časových jednotkách s ľubovoľným začiatkom. Prioritne sa analyzuje trendová a sezónna zložka (nasledujúce kapitoly).

11.2.1 Analýza trendovej zložky časových radov

Analýza trendovej zložky je zrejme najdôležitejšia časť analýzy časových radov. Podobne ako pri regresnej analýze, aj tu sa využívajú trendové spojnice (regresné funkcie) na analýzu vývoja vzťahu medzi závislou a nezávislou premennou, pričom pri časových radoch je nezávislou premennou stále čas. Na analýzu trendovej zložky časového radu sa bežne používajú už spomínané trendové funkcie, no v praxi sa potvrdilo, že pri výbere trendových funkcií si riešiteľ

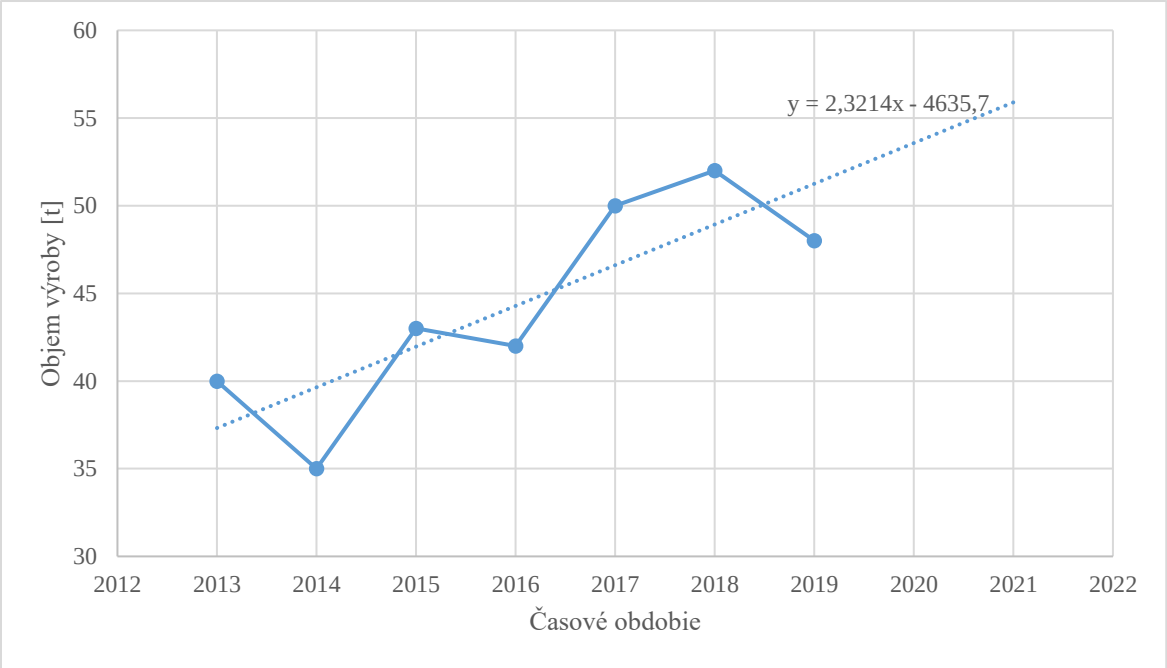
väčšinou vystačí s úzkou ponukou funkcií. Najčastejšie používané trendové funkcie pre skúmanie trendovej zložky časových radov sú sumarizované v nasledujúcej tabuľke.

Tabuľka 37 Najčastejšie trendové funkcie pre analýzu trendovej zložky časových radov

lineárny trend	$y_t = b_0 + b_1 t$	Parameter b_1 predstavuje prírastok hodnoty y pripadajúci na jednotkovú zmenu časovej premennej.
polynomický trend	$y_t = b_0 + b_1 t + b_2 t^2 + \cdots + b_k t^k$	Umožňuje nájsť trendovú funkciu, ktorá má vrchol/y (prípadne extrém).
exponenciálny trend	$y_t = b_0 b_1^t$	Parameter b_1 predstavuje priemerný prírastok hodnôt y_t. (tie sa správajú ako členy geometrickej postupnosti).
modifikovaný exponenciálny trend	$y_t = k + b_0 b_1^t$	Funkcia má vodorovnú asymptotu a dá sa pomocou nej ľahko modelovať vývoj javov, ktoré vychádzajú z obmedzených údajov o raste a u ktorých existuje určitá hranica nasýtenia, daná napr. záujmom o určitý výrobok alebo jeho potrebou.
logistický trend	$y_t = \frac{1}{k + b_0 b_1^t}$	Krivka má tri úseky, prvý je charakteristický pozvoľným vzostupom, druhý má v okolí inflexného bodu prudký rast a tretí úsek má určitú vrcholovú stagnáciu (nasýtenie). Uvedený tvar je jeden z mnohých rôznych funkčných predpisov popisujúcich krivku s charakteristickým priebehom v tvare písmena S.
Gompertzová krivka	$y_t = k b_0^{b_1^t}$	Krivka s podobným esovitým priebehom ako pri logistickom trende, ale na rozdiel od nej je asymetrická.

Príklad na určenie trendu časového radu (celkový smer vývoja):

Riešiteľ skúma objem výroby v podniku A za predchádzajúce obdobie v rokoch 2013-2019 (Obrázok 66) za účelom odhadu objemu výroby pre ďalšie obdobie. Z grafu je zrejmý lineárny vývoj, a preto bol zvolený lineárny trend vývoja časového radu.



Obrázok 66 Objem výroby v podniku A za sledované obdobie s prognózou jeho vývoja na ďalšie 2 roky

11.2.2 Analýza sezónnej zložky časového radu

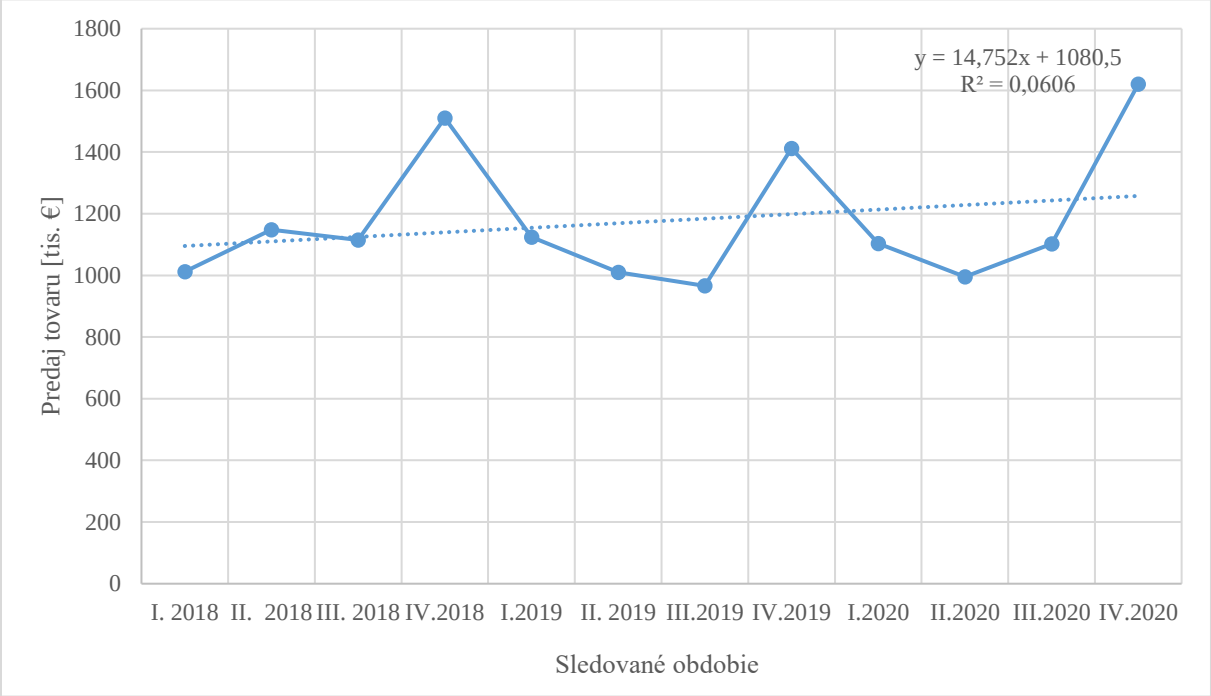
Analýza sezónnej zložky sa často vykonáva až po očistení údajov od trendovej zložky. Ide o určenie časového úseku, po ktorého uplynutí majú dáta opäť rovnakú hodnotu prípadne sú ovplyvnené trendovou a náhodnou zložkou.

Pre štúdium sezónnej zložky sa používa niekoľko typov modelov. V ekonomických modeloch býva spravidla zrejmá veľkosť periódy (štvrťrok, mesiac), v iných prípadoch je nutné aj túto dĺžku odhadovať (v hydrogeológií napr. pri výške hladiny spodných vôd). Používa sa tu tiež harmonická analýza, ktorá modeluje priebeh údajov pomocou niekoľkých členov Fourierového radu. Parametre sa určujú použitím numerických metód.

Príklad na sezónnosť (sezónna zložka) časového radu:

Riešiteľ skúma predaj konkrétneho tovaru v priebehu troch rokov (štvrťročne) a snaží sa vysvetliť značné rozdiely medzi niektorými obdobiami (Obrázok 67). Na zobrazenom časovom rade je zrejmé opakovanie výkyvu predaja (zvýšený predaj) v poslednej štvrtine každého roku. Daný príklad ilustruje sezónnu zložku časového radu, ktorý môže byť spôsobený napríklad vianočnými sviatkami, kde sa predaj niektorého tovaru podstatne zvyšuje oproti ostatným obdobiam. Tento fakt je potrebné zohľadniť pri predpovediach ďalšieho vývoja. Ako vidno z koeficientu spoľahlivosti (R^2=0,0606) pre vloženú lineárnu regresnú funkciu, spoľahlivosť predpovede je veľmi nízka a daná funkcia by bola nepoužiteľná pre všetky obdobia v roku (hlavne pre „sezónu“). Daná skutočnosť sa rieši zohľadnením sezónnej zložky napr. prostredníctvom tzv. sezónneho indexu (porovnanie skutočnej hodnoty za dané obdobie a vyrovnanej hodnoty vypočítanej prostredníctvom rovnice regresnej funkcie); prípadne

určitými technikami priemerovania alebo zanedbaním sezónneho obdobia pri odhadoch ďalších „nesezónnych“ období.



Obrázok 67 Vývoj predaja konkrétneho tovaru za sledované obdobie (dôraz na sezónnu zložku časového radu)

Literatúra

BEDFORD, T., COOKE, R. *Probalistic Risk Analysis: Foundations and Methods*. 7. vyd. Cambridge: Cambridge Press, 2011.

BUDÍKOVÁ, M, KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami*. Praha: GRADA, 2010.

CSÁMPAI, O. *Elementárium kvantitatívneho výskumu*. Trnava: Oliva, 2013.

DISMAN, M. *Jak se vyrábí sociologická znalost*. Praha: Karolinum, 2018.

FELLER, W. *An introduction to probability theory and its applications*. 1970. I. a II., New York: J. Wiley.

HINDLS, R., ARLTOVÁ, M., HRONOVÁ, S., MALÁ, I., MAREK, L., PECÁKOVÁ, I., ŘEZANKOVÁ, H. *Statistika v ekonomii*. Praha: Professional Publishing, 2018, ISBN 978-80-88260-09-7.

MARKECHOVÁ, D., STEHLÍKOVÁ, B., TIRPÁKOVÁ, A. *Štatistické metódy a ich aplikácie*. 2011. Univerzita Konštantína Filozofa v Nitre. ISBN 978-80-80948-07-8.

ŘEZANKOVÁ, H. *Analýza dat dotazníkových šetření*. 4. vyd. Praha: Professional Publishing, 2017.

STEHLÍKOVÁ, B. *Nový Metodologický prístup k prognózovaniu demografických časových radov* = a new methodological approach to the demographic time series projection. In Acta Economica et Informatica, roč. 7, 2004, č. 2, s. 49-52, ISSN 1335-2571.

ŠOLTÉS, E. *Regresná a korelačná analýza*. 2008. Iura Edition, spol. s r. o., člen skupiny Wolters Kluwer, Bratislava, ISBN 978-80-8078-163-7.

TEREK, M. *Interpretácia štatistiky a dát*. 5.vyd. Košice: EQUILIBRIA, 2017.
VARGA, Š. *Matematická štatistika.* Bratislava: STU, 2012, ISBN 978-80-2273-789-0.

Za odbornú náplň tohto vydania zodpovedá odborná redaktorka doc. Ing. Katarína Buganová, PhD.

Autori Ing. Michal Titko, PhD., doc. Ing. Ladislav Novák, PhD.,
Ing. Michaela Jánošíková, PhD.

Názov: PRAKTICKÁ ŠTATISTIKA

Vydala Žilinská univerzita v Žiline v EDIS-vydavateľstve UNIZA v roku 2021
ako svoju 4732. publikáciu
Vydanie prvé
Náklad online
AH/VH 12,17/12,54
ISBN 978-80-554-1814-8

Rukopis vo vydavateľstve neprešiel redakčnou ani jazykovou úpravou.

www.edis.uniza.sk