UNIVERZITA KONŠTANTÍNA FILOZOFA V NITRE
FAKULTA PRÍRODNÝCH VIED A INFORMATIKY

EKONOMICKÁ ŠTATISTIKA

Milan Maroš

NITRA 2024

Ekonomická štatistika

Milan Maroš

2024

Ekonomická štatistika

Edícia Prírodovedec č. 862

Autor:

PaedDr. Milan Maroš, PhD.

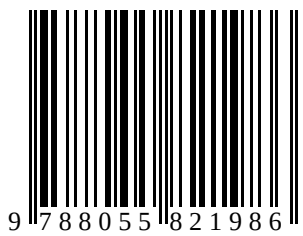
Recenzenti:

doc. RNDr. Jana Špirková, PhD.

doc. Ing. Ivan Holúbek, PhD.

Publikácia bola vytvorená v rámci projektu KEGA č. 029UKF-4/2022 Príprava štruktúry a obsahu predmetu, učebnice a e-learningového kurzu "Podnikateľské zručnosti" pre neekonomické študijné programy s fokusom na online prostredie.

ISBN 978-80-558-2198-6



Obsah

ÚVOD

Štatistika je veda, ktorá je založená na získavaní, spracovaní a vyhodnotení dát, preto sa využíva napríklad v biológii, chémii, pedagogike, medicíne, marketingovom výskume a mnohých ďalších oblastiach. V tejto publikácii sa zameriavame predovšetkým na jej uplatnenie v ekonomickej sfére. Čitateľ tu môže nájsť základy z oblasti metodológie výskumu, výberu a spracovania údajov, tvorby hypotéz, spôsobov ich testovania a ďalších tém. Hoci sa vo všetkých kapitolách nachádzajú rôzne teoretické formulácie k jednotlivým témam, opomenutá nebola ani praktická využiteľnosť. Preto tu čitateľ nájde aj veľké množstvo príkladov so vzorovým riešením. Publikácia si nekladie za cieľ poskytnúť vyčerpávajúce informácie z oblasti štatistiky, či ekonomickej štatistiky. Zaoberá sa vybranými dôležitými témami tak, aby čitateľ získal základné zručnosti a spôsobilosti pre využitie štatistiky v ekonomickej praxi. Rozsah bol koncipovaný tak, aby záujemcom postačoval na ucelenú predstavu o použití štatistiky vo výskume a zároveň ich nezneistil príliš podrobným obsahom. Dôraz bol kladený aj na prácu s počítačom a využitie programu Microsoft Excel, čo výrazne urýchli výpočty. Všetky príklady sú uvádzané pre verziu Excel 2016, ktorá patrí v súčasnosti medzi najpoužívanejšie. Pre správne pochopenie postupov riešenia sa od čitateľa očakávajú aspoň základné znalosti práce s týmto softvérom. Napriek tomu, že požiadavky na znalosti čitateľa v oblasti práce s Excelom a všeobecne v oblasti digitálnej gramotnosti nie sú veľké, vyšší stupeň znalostí umožní rýchlejšie pochopenie problematiky.

V publikácii nájdu vysokoškolskí študenti vhodný učebný materiál predovšetkým k predmetu s názvom ekonomická štatistika, ale využiť sa dá aj pri ďalších, ktoré s témou aspoň čiastočne súvisia. Kniha je určená nielen študentom vysokých škôl, ale aj odbornej verejnosti, poslucháčom vzdelávacích kurzov a všetkým, ktorí sa pri svojom štúdiu alebo odbornej činnosti stretávajú s problematikou ekonomickej štatistiky, spracovania a vyhodnotenia dát. Pre uľahčenie pochopenia spracovanej problematiky sa na začiatku každej kapitoly nachádzajú kľúčové slová, ktoré obsahujú najdôležitejšie pojmy k preberanej téme. Na konci každej kapitoly čitateľ nájde kontrolné otázky a úlohy na riešenie, aby si mohol preveriť úroveň nadobudnutých vedomostí.

Cieľom tejto publikácie je predovšetkým poskytnúť informácie o postupoch riešenia štatistických úloh v Exceli. Problematika ekonomickej štatistiky je veľmi rozsiahla. Keďže texty v knihe sú primárne orientované na aplikačnú časť štatistiky a využitie v praxi, často chýba podrobné odvodenie postupov, dôkazy vzorcov, či vyčerpávajúce definovanie všetkých pojmov. Z hľadiska rozsahu by to ani nebolo možné, existuje však veľké množstvo literatúry (napríklad aj v prehľade literatúry na konci tejto knihy), kde prípadní záujemcovia môžu nájsť všetky ďalšie informácie. Pri

niektorých odborných pojmoch je v zátvorke uvádzaný aj anglický preklad, aby sa čitateľ mohol ľahšie zorientovať pri prípadnom štúdiu cudzojazyčnej literatúry.

Poďakovanie na záver patrí všetkým tým, ktorí prispeli ku skvalitneniu publikácie. Predovšetkým recenzentom za starostlivé prečítanie a pripomienky, ktoré boli zapracované do predkladanej publikácie a pomohli vylepšiť pôvodné znenie rukopisu. Napriek tomu sa v tejto publikácii môžu vyskytovať nedostatky, tak ako v každom diele, ktoré vytvoril človek. Budem preto vďačný každému za spätnú väzbu a upozornenie na prípadné chyby.

Autor

1 VÝSKUM A ŠTATISTIKA

Kľúčové slová: veda, výskum, prieskum, kvalitatívny a kvantitatívny výskum, ekonomická štatistika, štatistický znak

Ľudstvo od počiatku svojej existencie potrebovalo poznávať prostredie a zákonitosti okolo seba. Zhromažďovanie informácií a získavanie nových poznatkov je zvyčajne veľmi zložitý proces, ktorý môže byť realizovaný viacerými metódami a postupmi. Ľudia môžu napríklad vychádzať z rôznych **tradícií**, ktoré keď sa často opakujú, zvyšuje sa ich hodnovernosť v očiach ľudí. Je zaujímavé, že niektorí ľudia často trvajú na tradičných poznatkoch, aj keď majú k dispozícii fakty, ktoré dokazujú niečo iné. Iní ľudia zasa preferujú prijímanie určitých poznatkov iba vtedy, ak ich vysloví nejaká osobnosť, ktorá je pre nich **autoritou**. To má pre poznávanie väčší význam ako tradície, rozhodne by to však nemala byť najdôležitejšia metóda. Niekedy je pre človeka kritériom pravdivosti, keď sa poznatky „**zhodujú s rozumom**“. Často sa vtedy používa argument „veď to dá rozum“. Je však veľmi problematické presne vymedziť, čo znamená „zhodovať sa s rozumom“. Ako najlepší spôsob bádania sa javí použitie **vedeckého prístupu**, kedy sa nové poznatky získavajú bez ohľadu na názory, postoje či želania bádateľa. Pri správnom vedeckom poznávaní je činnosť vedcov kontrolovaná veľmi prísnymi mechanizmami. Je preto takmer vylúčené, aby sa uplatnili ich osobné názory alebo emócie. Preto sa s vedeckým poznávaním často spája pojem „**objektivita**“ (Chráska, 2016).

1.1 Charakteristika a typy výskumu

Základom pre vznik vedeckého myslenia boli už síce poznatky a procesy v období staroveku, ale v tom čase všetko patrilo pod nerozčlenenú disciplínu – filozofiu. Moderné vnímanie vedy datujeme až od 17. storočia, kedy sa začali vymedzovať samostatné vedné odbory tak, ako ich poznáme dnes. Ešte dôležitejšie, ako členenie na vedné odbory, je skutočnosť, že v modernej vede musia bádatelia svoje tvrdenia overovať. **Veda** je teda súbor overených poznatkov, pomocou ktorých sa dajú vysvetliť rôzne javy objektívnej reality (sveta okolo nás). V rámci vedy sa vytvárajú **vedecké teórie**, ktoré je možné uplatniť v praxi napríklad na predpovedanie (predikciu) alebo zmenu rôznych javov. Podľa Gavoru (2010) je vedecká teória:

- systém overených a usporiadaných poznatkov,
- určitá abstrakcia, ktorá platí pre vymedzený súbor javov.

Výroky a tvrdenia sú vo vedeckých teóriách formulované odbornou terminológiou. Každá vedecká teória musí byť publikovaná, pretože verejná dostupnosť pre ostatných je v podstate podmienkou jej existencie. Na tvorbe vedeckých teórií pracujú zvyčajne odborníci s príslušným vzdelaním a kvalifikáciou. Takúto vedeckú

činnosť nazývame **výskum**. **Prieskum** je nižšia forma empirického skúmania ako výskum. Napríklad realizáciou prieskumu trhu alebo zisťovaním názorov na nejaký problém sa síce získavajú terénne údaje, ale nebuduje sa týmto spôsobom vedecká teória.

Výskum môžeme deliť z rôznych hľadísk. Napríklad na **základný výskum**, ktorý sa sústredí hlavne na teoretické otázky a **aplikovaný výskum**, ktorý je zameraný na prax. Čitateľ sa tiež môže stretnúť s pojmom **empirický výskum** (jeho podstata spočíva v zbieraní aj spracovaní údajov z terénu) a jeho opakom tzv. výskumom „pri stole“ (z angličtiny desk study). Podľa toho, akým spôsobom sú výskumné dáta ďalej používané, rozlišujeme **kvalitatívny** a **kvantitatívny** výskum, čo je jedno z najviac používaných členení. Sú založené na odlišnej filozofii, majú rozdielne ciele a každý z nich má svoje výhody aj nevýhody. Napríklad, ak sa chce výskumník dozvedieť preferencie ľudí v oblasti marketingu, zvolí pravdepodobne kvantitatívne metódy. Ak ho však zaujímali životné osudy ľudí alebo ich každodenné správanie, potom sa javia vhodnejšie kvalitatívne metódy (Silverman, 2005).

Kvalitatívny výskum

V kvalitatívnom výskume sa zdôrazňujú subjektívne aspekty jednania ľudí (Chráska, 2016). Kvalitatívny výskum je zvyčajne realizovaný dlhodobo a intenzívne, vyžaduje si pružné reagovanie na situácie, ktoré vznikajú v teréne a aj preto sa **pracuje iba s malými skupinami** osôb (zvyčajne 10 až 20). Cieľom výskumu potom pochopiteľne **nie je štatistické zovšeobecnenie** údajov. Nepoužívajú sa tu štandardizované a algoritmizované postupy. Skúmané osoby sa tu zvyčajne nazývajú participanti, prípadne účastníci alebo informanti. Nezvykne sa tu používať slovo respondenti, ktoré je typické skôr pre kvantitatívny výskum. Okrem ľudí sa v tomto type výskumu môžu skúmať aj neživé produkty človeka - textové, vizuálne produkty alebo prvky materiálnej kultúry (Čukan a kol., 2017). **Cieľom** je predovšetkým preniknutie do hĺbky problému, **tvorba hypotéz** a **tvorba novej teórie,** kvalitatívny výskum má teda exploračný charakter.

Kvantitatívny výskum

Má predovšetkým verifikačný charakter, to znamená, že **verifikuje**, prípadne falzifikuje už **existujúcu teóriu**. Výskumník sa snaží pracovať nestranne, bez subjektívnych postojov a usiluje sa získať čo najväčšie množstvo údajov tak, aby mohol výsledky **zovšeobecniť** na čo najväčšiu skupinu subjektov. V kvantitatívnom výskume (s výnimkou deskriptívneho) sa formulujú hypotézy vyvodené z existujúcej teórie, ktoré sa následne testujú (Čukan a kol., 2017). Cieľom je získať exaktné a objektívne overiteľné dáta o skúmanej problematike. Základom pre tento cieľ je meranie, čiže postup získavania údajov vyjadrených numericky (Gavora a kol., 2010).

1.2 Ekonomická štatistika

Vedecký výskum by sa len ťažko mohol realizovať bez korektnej práce s dátami. Veda, ktorá sa zaoberá zberom, analýzou a interpretáciou údajov, sa nazýva **štatistika**. Využíva sa preto snáď vo všetkých vedeckých disciplínach od prírodných vied, techniky až po sociálne a humanitné vedy. Štatistiku využívajú aj štátni úradníci, výrobcovia, či obchodníci (Magnello – Van Loon, 2010).

Táto publikácia je zameraná predovšetkým na **ekonomickú štatistiku**. Ekonómovia môžu využiť štatistiku na modelovanie ekonomických procesov, predikciu vývoja, ale aj pri tvorbe rôznych odporúčaní pre hospodársku politiku. V bankovníctve sa aj pomocou štatistických metód rozhoduje o poskytnutí úveru, v poisťovníctve o výške poistného, či odhaľovaní poistných podvodov (Rimarčík, 2007). Denne nás obklopujú rôzne ekonomické informácie, preto sú príklady v tejto publikácii zamerané práve na túto oblasť. Každý ekonóm by mal primerane rozumieť aj štatistickým metódam a ich výsledky by mal vedieť korektne interpretovať. Štatistika síce za neho nikdy neurobí žiadne dôležité rozhodnutie, ale môže mu pri rozhodovaní výrazne pomôcť (Hindls a kol., 2007).

Na začiatok je vhodné vysvetliť niektoré **základné štatistické pojmy**:

Hromadný jav – je taký jav, ktorý sa vyskytuje u veľkého počtu objektov.

Štatistický súbor – množina skúmaných objektov.

Rozsah súboru – počet prvkov štatistického súboru.

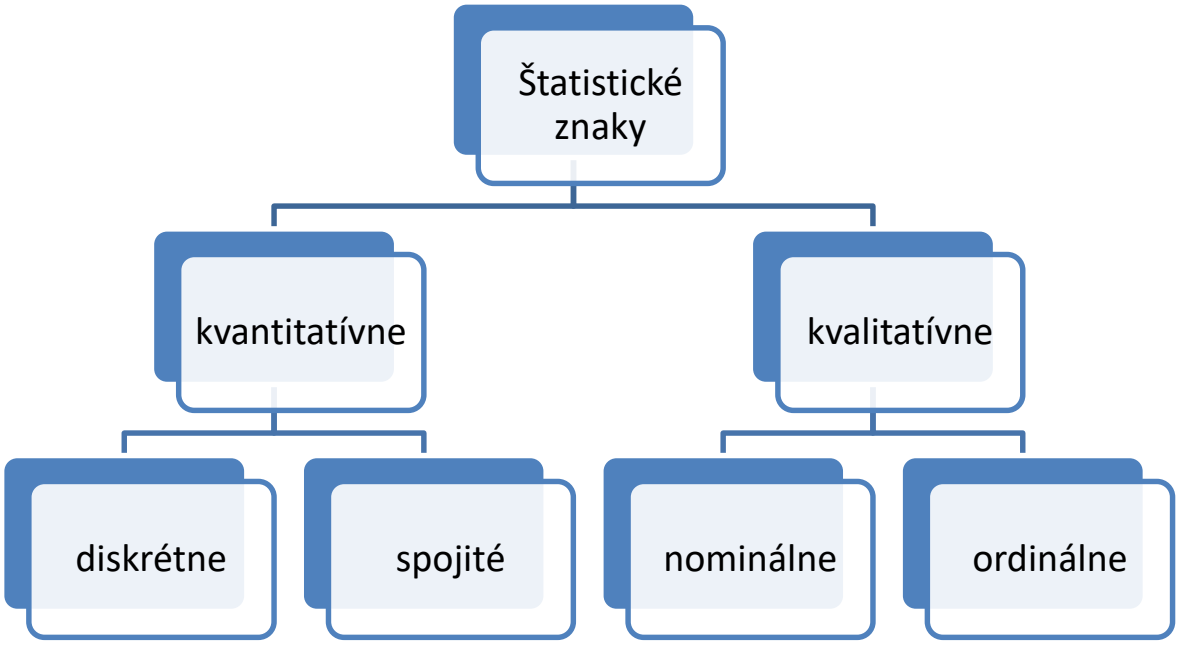
Štatistické jednotky – prvky štatistického súboru.

Štatistické znaky (premenné) – vlastnosti, ktoré sa vyskytujú na štatistických jednotkách. Napríklad hmotnosť, mzda, počet rokov praxe, farba očí, vek, pohlavie.

Rôzne hodnoty štatistického znaku sa nazývajú **obmeny**. Ak napríklad zistíme pri 100 respondentoch ich krvnú skupinu, získame 100 hodnôt štatistického znaku, ale iba štyri obmeny A, B, AB a 0. Štatistické znaky môžeme deliť z rôznych hľadísk. Jedno z najčastejších členení je podľa toho, ako sa dajú vyjadriť ich obmeny. Niektoré sa vyjadrujú číselne (napríklad počet rokov praxe), iné sa môžu vyjadriť slovne (napríklad vzdelanie môže byť stredoškolské, vysokoškolské, atď.). Takéto rozlišovanie premenných má svoje dôvody. Pre číselné premenné sa totiž používajú iné štatistické metódy a postupy ako pre premenné vyjadrené slovne.

Štatistické znaky môžeme deliť na **kvantitatívne** (vyjadrené číslom) a **kvalitatívne** (vyjadrené slovne). Kvalitatívne znaky sa často označujú aj ako **kategoriálne**. Kategoriálny znak môže niekedy vzniknúť aj z kvantitatívneho, napríklad rozdelením hodnôt do intervalov. Kvalitatívne znaky delíme na **nominálne** a **ordinálne** (poradové). Pri nominálnych znakoch vieme povedať iba to, že ich obmeny sa

navzájom vylučujú. Príkladom ordinálnych znakov môže byť dosiahnuté vzdelanie. Môžeme povedať, že vysokoškolské vzdelanie je vyššie ako stredoškolské, ale nevieme vyjadriť o koľko. Kvantitatívne znaky delíme na **spojité** (môžu nadobúdať v rámci nejakého intervalu ľubovoľné hodnoty) a **diskrétne** (sú nespojité, môžu nadobúdať iba určité hodnoty, najčastejšie celé nezáporné čísla).



Obr. 1.1 Schematické zobrazenie členenia štatistických znakov (zdroj: vlastné spracovanie na základe Hindls a kol., 2007)

Pre názornejšiu predstavu je na obrázku 1.1 grafické znázornenie delenia štatistických znakov.

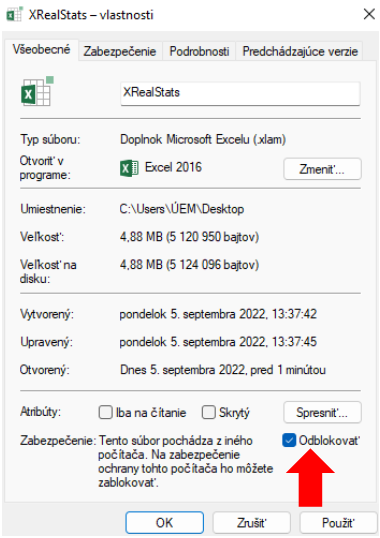
1.3 Štatistika a Excel

Pre uľahčenie a zrýchlenie výpočtov je v príkladoch so vzorovým riešením použitý program Excel. Tento softvér bol vybratý z dôvodu vysokej dostupnosti a rozšírenosti u väčšiny užívateľov, preto by čitatelia nemali mať problém s praktickou aplikáciou poznatkov. Všetky postupy sú uvádzané pre verziu Excel 2016, ktorá patrí v súčasnosti medzi najpoužívanejšie.

Excel má v sebe zabudovaných viacero užitočných štatistických funkcií, ale chýbajú niektoré zložitejšie procedúry. Napríklad testy normality, neparametrické testy a tak ďalej. Dajú sa však doinštalovať rôzne doplnky, ktoré značne rozšíria výpočtové možnosti. Jedným z bezplatných doplnkov je aj **Real Statistics** Resource Pack

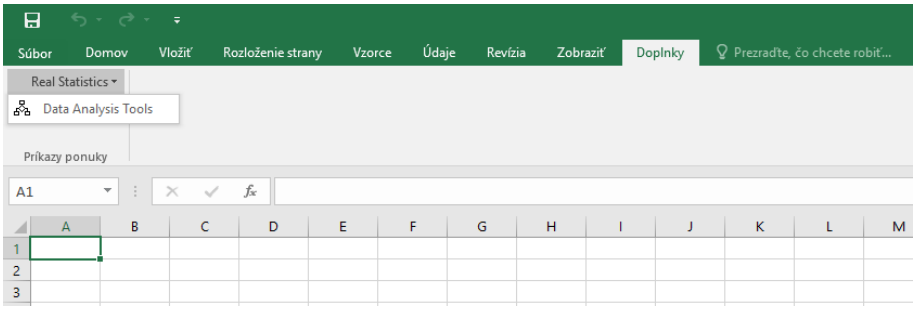
software (Copyright 2013 – 2023) Charles Zaiontz (Zaiontz, 2023), ktorý je použitý v ďalších kapitolách. **Postup inštalácie** tohto doplnku je nasledovný:

- V internetovom prehliadači navštívime webovú stránku pre stiahnutie inštalačného súboru: https://www.real-statistics.com.
- V záhlaví stránky vyberieme z hlavného menu položku „Free Download“ a potom „Resource Pack“.
- Dole v texte nájdeme hypertextový odkaz s názvom „Real Statistics Resource Pack for Excel 2010/2013/2016/2019/2021/365“, ktorý by mal byť zvýraznený červeným písmom. V prípade požiadavky na inú verziu Excelu, prípadne na iný operačný systém ako je Windows, je potrebné nájsť iný hypertextový odkaz.
- Po kliknutí na vyššie uvedený odkaz stiahneme do počítača súbor **XRealStats.xlam** a uložíme ho napríklad na pracovnú plochu.
- Súbor potrebujeme presunúť do adresára, ktorý je skrytý, preto ho musíme najskôr zobraziť. **Zobrazíme** teda v počítači **skryté súbory a priečinky**: Ovládací panel – Vzhľad a prispôsobenie – Zobraziť skryté súbory a priečinky – Zobrazovať skryté súbory, priečinky a jednotky (klikneme na „použiť“ a potom „OK“).
- Premiestnime súbor XRealStats.xlam do adresára: C:\Users\user-name\AppData\Roaming\Microsoft\AddIns. Táto cesta môže byť u každého trochu iná. Napríklad namiesto „Users“ môže byť „Používatelia“ a namiesto „user-name“ názov počítača, resp. konta.
- Klikneme pravým tlačidlom na súbor XRealStats.xlam a vo **vlastnostiach** zaškrtneme políčko „**Odblokovať**“ (na obrázku 1.2 je vyznačené červenou šípkou). Potom „použiť“ a „OK“.



Obr. 1.2 Odblokovanie súboru pri inštalácii (zdroj: vlastné spracovanie)

- Spustíme Excel a v hlavnom menu klikneme na Súbor – Možnosti – Doplnky – Spustiť. Treba zaškrtnúť položky „**Doplnok riešiteľ**“ a „**Xrealstats**“. Na stránke tvorcu doplnku je uvedené, že treba mať zaškrtnuté aj položky „Analytické nástroje“ a „Analytické nástroje - VBA“. Po kliknutí na „OK“ bude počítač chvíľu pracovať, je potrebné ho nechať a počkať. Real Statistics nájdeme v Exceli v hlavnom menu pod položkou „**Doplnky**“ (obr. 1.3).



Obr. 1.3 Spustenie Real Statistics cez položku „Doplnky“ (zdroj: vlastné spracovanie)

- Na záver uvedieme počítač do pôvodného stavu, čiže zrušíme zobrazovanie skrytých priečinkov: Ovládací panel – Vzhľad a prispôsobenie – Zobraziť skryté súbory a priečinky – **Nezobrazovať** skryté súbory, priečinky a jednotky (klikneme na „použiť“ a potom „OK“).

Kontrolné otázky:

1. Akými metódami môže byť realizované zhromažďovanie informácií a získavanie nových poznatkov?
2. Aký spôsob bádania a získavania nových poznatkov je nezávislý na názoroch a emóciách?
3. Aký je rozdiel medzi výskumom a prieskumom?
4. Aký je rozdiel medzi základným a aplikovaným výskumom?
5. Čo je to empirický výskum?
6. Aké sú odlišnosti medzi kvantitatívnym a kvalitatívnym výskumom?
7. Čím sa zaoberá štatistika a v akých oblastiach sa používa?
8. Na čo je zameraná ekonomická štatistika a aké sú príklady jej využitia?
9. Aký je rozdiel medzi kvalitatívnymi a kvantitatívnymi štatistickými znakmi (premennými)? Ako ich ďalej delíme?
10. Aký je rozdiel medzi hodnotami štatistického znaku a obmenami?

Úlohy na riešenie:

1. Diskutujte o etických otázkach spojených s vedeckým výskumom. Napríklad na tému spravodlivá odmena a motivácia účastníkov výskumu.

2. Diskutujte o tom, ako sa tradície a vedecký prístup líšia pri získavaní poznatkov.

3. Analyzujte výraz „zhodovať sa s rozumom“. Zdôvodnite, prečo je dôležité kriticky posudzovať informácie a nebrať ich automaticky ako pravdivé.

4. Zisťovali sme, koľko detí majú zamestnanci podniku. Získali sme dáta od desiatich pracovníkov: 1; 2; 0; 0; 1; 0; 2; 1; 0; 0. Čo sú v tomto prípade obmeny štatistického znaku?

   [0; 1; 2]

5. V obchode majú na sklade 14 tričiek o veľkosti S, 10 tričiek o veľkosti M, 105 tričiek o veľkosti L a 11 tričiek o veľkosti XL. O aký štatistický znak sa v tomto prípade jedná?

   [kvalitatívny – ordinálny]

6. Krajina vyslala na olympiádu 17 futbalistov, 10 atlétov a 2 tenistov. O aký štatistický znak sa v tomto prípade jedná?

   [kvalitatívny – nominálny]

7. Aký typ štatistického znaku bol použitý v úlohe číslo 4, kde sme zisťovali počet detí u zamestnancov?

   [kvantitatívny – diskrétny]

2 DESKRIPTÍVNA ŠTATISTIKA

Kľúčové slová: deskriptívna štatistika, triedenie štatistických údajov, početnosti, charakteristiky polohy, charakteristiky variability, histogram, box plot

Výskumník vo svojej práci zvyčajne zhromaždí veľké množstvo údajov, ktoré je potrebné ďalej spracovať, lebo neposkytujú dostatočné informácie o skúmanom súbore a sú neprehľadné. Vyjadrenie a prezentovanie údajov v prehľadnej forme je úlohou **deskriptívnej štatistiky**. V rôznej literatúre ju nájdeme označenú aj ako **popisná** alebo **opisná štatistika**. Za týmto účelom sa v deskriptívnej štatistike využívajú:

- tabuľky,
- štatistické charakteristiky,
- grafy.

2.1 Triedenie štatistických údajov s využitím tabuliek

Jednou z možností, ako sprehľadniť údaje, je usporiadať ich do skupín, ktoré nazývame **triedy**. Usporiadanie hodnôt znaku do tried sa nazýva **triedenie**. Keď sa pri triedení používa jeden triediaci znak, hovoríme o *jednostupňovom* triedení (prípadne jednorozmernom), ak sa súbor triedi súčasne podľa dvoch alebo viacerých štatistických znakov, hovoríme o *dvojstupňovom* resp. *viacstupňovom* triedení (Tirpáková – Malá, 2007).

Pri triedení je nutné dodržať dve základné zásady. Zásada **úplnosti** znamená, že každá hodnota je zaradená do niektorej triedy. Zásada **jednoznačnosti** znamená, že každá hodnota je zaradená práve do jednej triedy. Pri spracovaní kvalitatívnych alebo kvantitatívnych znakov s malým počtom obmien môžeme triedenie znázorniť pomocou tzv. **tabuľky rozdelenia početností**, niekedy označovanej ako **frekvenčná tabuľka** (*angl. frequency table*). V tejto tabuľke môžeme uvádzať niekoľko druhov početností:

- **Absolútne početnosti** – počet štatistických jednotiek u ktorých sa vyskytla príslušná hodnota.
- **Relatívne početnosti** – podiel štatistických jednotiek, u ktorých sa príslušná hodnota vyskytla na celkovom počte. Tento podiel je možné vyjadriť hodnotou od 0 do 1 alebo ako percentuálny podiel (vtedy sú možné hodnoty od 0 % do 100 %). Pre teoretickú analýzu je lepšie používať prvú možnosť, pre praktickú prezentáciu výsledkov sa častejšie používa percentuálny podiel.
- **Kumulatívne početnosti** – početnosť výskytu možných hodnôt od mínus nekonečno po analyzovanú hodnotu. Môžu byť absolútne alebo relatívne.

Príklad 2.1 Turisti hodnotili spokojnosť so zájazdom na stupnici od 1 do 5 podobne ako známkovanie v škole. Celkove bolo 30 turistov a spokojnosť bola nasledovná: 1, 4, 5, 2, 3, 3, 4, 2, 2, 5, 1, 3, 3, 4, 3, 2, 5, 3, 1, 3, 2, 3, 3, 4, 5, 3, 2, 2, 3, 4. Zostavte tabuľku rozdelenia početností tak, aby obsahovala absolútne početnosti, relatívne početnosti v percentách a kumulatívne početnosti.

Riešenie: Znak „známka“ je ordinálny, počet jeho obmien je 5 a môžeme ich usporiadať do poradia vo frekvenčnej tabuľke, ktorá je na obrázku 2.1.

	Spokojnosť turistov so zájazdom					
	1	**2**	**3**	**4**	**5**	**Spolu**
absolútna početnosť	3	7	11	5	4	30
relatívna početnosť v %	10,00	23,33	36,67	16,67	13,33	100
kumulatívna početnosť	3	10	21	26	30	-

Obr. 2.1 Tabuľka rozdelenia početností z príkladu 2.1 (zdroj: vlastné spracovanie)

Pri malom počte znakov, ako je to aj v tomto príklade, môžeme hodnoty v tabuľke vypočítať manuálne. Dá sa však využiť aj program Excel, čo oceníme hlavne pri spracovaní väčšieho množstva údajov. Na súčty v poslednom stĺpci je vhodná funkcia *SUM*. Pri výpočte relatívnych početností a následnom kopírovaní vzorca medzi bunkami je potrebné dať pozor na to, aby sa v menovateli zlomku použil celkový počet znakov stále z tej istej bunky. To dosiahneme pomocou tzv. absolútneho adresovania, keď danú bunku zafixujeme pridaním znaku $ (dolár). Tvorba riešenia pomocou funkcií a výpočtov medzi bunkami je výhodná aj preto, že v prípade zmeny vstupných údajov nie je potrebné všetko znovu prepočítať, ale program to urobí automaticky.

Pri diskrétnych znakoch s veľkým počtom obmien alebo pri spojitých znakoch sa frekvenčná tabuľka vytvára pomocou **intervalového triedenia**. Na vytvorenie tried rozdelíme interval tvorený najmenšou a najväčšou hodnotou znaku na niekoľko disjunktných intervalov. Ich počet môže výskumník stanoviť podľa potreby sám tak, aby výsledný zápis bol čo najprehľadnejší. S realizáciou intervalového triedenia môže významne pomôcť Excel, čo bude zrejmé v ďalšom texte o grafickom znázornení (tvorba histogramu). V priebehu niekoľkých sekúnd si na počítači môžeme vyskúšať triedenie s rôznym počtom intervalov a potom vybrať to najvhodnejšie.

2.2 Štatistické charakteristiky

Opis skúmaného súboru pomocou tabuliek a grafov je síce väčšinou prehľadný a má svoje výhody, ale zvyčajne obsahuje pomerne veľa dát. Preto sa výskumníci často snažia nahradiť dáta jednou alebo niekoľkými hodnotami, ktoré by reprezentovali celý súbor. Tieto hodnoty sa nazývajú **štatistické charakteristiky**, pričom ich najčastejšie delíme na **charakteristiky polohy** (alebo aj miery polohy, stredné hodnoty, miery centrálnej tendencie) a **charakteristiky variability** (alebo aj miery variability, miery rozptylu).

2.2.1 Charakteristiky polohy

Charakteristiky polohy predstavujú v istom zmysle typickú hodnotu znaku skúmaného súboru. Pomocou jedného číselného ukazovateľa umožňujú vytvorenie určitej predstavy o celom súbore. Niekedy ich nazývame aj **miery polohy**, **stredné hodnoty** alebo **miery centrálnej tendencie** *(angl. central tendency)*. V ďalšom texte budú podrobnejšie vysvetlené najznámejšie stredné hodnoty.

Aritmetický priemer *(angl. arithmetic mean)* vypočítame ako súčet všetkých hodnôt vydelený ich počtom. Poznáme viac druhov priemerov (napríklad harmonický alebo geometrický), ale najznámejší je práve aritmetický.

Geometrický priemer *(angl. geometric mean)* sa používa vtedy, ak je medzi hodnotami znaku multiplikatívny vzťah. To znamená, že každá nasledujúca hodnota je násobkom predchádzajúcej. Označíme ho $\bar{x}_G$ a vypočíta sa

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

kde $x_1, x_2, \ldots x_n$ sú hodnoty premennej X. V ekonomickej oblasti sa používa napríklad na výpočet priemerného rastu inflácie alebo iných indikátorov v danom časovom období.

Medián *(angl. median)* je hodnota, ktorá rozdeľuje usporiadaný súbor na dve časti s rovnakým počtom prvkov. Výhodou aritmetického priemeru v porovnaní s mediánom je to, že závisí od všetkých hodnôt, aj preto je to najčastejšie používaná stredná hodnota. Aritmetický priemer je však citlivý na extrémne hodnoty. Výhodou mediánu je práve skutočnosť, že extrémne hodnoty ho neovplyvňujú. Nevýhodou mediánu je, že na jeho výpočet sa využíva jedna, prípadne dve hodnoty, bez ohľadu na veľkosť skúmaného súboru.

Modus *(angl. mode)* je hodnota, ktorá sa v súbore vyskytuje najčastejšie (má najväčšiu absolútnu početnosť). Využitie modusu je spomedzi spomínaných stredných hodnôt najmenšie. Dôvodov je niekoľko, ale problematické je najmä to, že často neexistuje, alebo ich existuje viacero.

Kvantil je hodnota *k*-tej časti, ak je súbor rozdelený na *n* rovnakých častí, pričom sú hodnoty premennej zoradené od najmenšej po najväčšiu. Medián je najpoužívanejší kvantil. Existuje viacero druhov kvantilov: **kvartily** (3 hodnoty, ktoré delia usporiadaný súbor na 4 rovnako početné časti), **kvintily** (4 hodnoty deliace usporiadaný súbor na 5 častí), decily (9 hodnôt deliacich usporiadaný súbor na 10 častí), **vingtily** (19 hodnôt deliacich usporiadaný súbor na 20 častí) a **percentily** (99 hodnôt, ktoré rozdeľujú usporiadaný súbor na 100 rovnako početných častí). Medián je zároveň 2. kvartil, 5. decil a 50. percentil. Uvedieme príklad interpretácie percentilov v ekonómii. Ak sa 25. percentil mesačného príjmu rovná 700 €, znamená to, že 25 % najchudobnejších členov súboru má mesačný príjem, ktorý nepresahuje 700 €. Jednoducho povedané, v tomto prípade 25 % ľudí s najnižšími príjmami zarába maximálne 700 € mesačne.

V **Exceli** môžeme charakteristiky polohy vypočítať pomocou funkcií **AVERAGE** (aritmetický priemer), **GEOMEAN** (geometrický priemer), **MEDIAN** (medián), **MODE** (modus), **QUARTILE** (kvartil). Na výpočet viacerých štatistických charakteristík je možné využiť aj doplnok Real Statistics v záložke **Desc**, prvá položka **Descriptive Statistics and Normality**.

Príklad 2.2 V istej firme pracuje 15 zamestnancov. U každého z nich sme skúmali mesačnú mzdu, pričom sme zistili nasledovné údaje v eurách: 850, 900, 1000, 900, 1250, 1100, 1030, 1400, 1350, 1020, 990, 900, 1480, 1220, 900. Vypočítajte aritmetický priemer, medián a modus mesačných miezd.

Riešenie: Na výpočet požadovaných charakteristík môžeme použiť v Exceli funkcie *AVERAGE*, *MODE.SNGL* a *MEDIAN*, ako je znázornené na obrázku 2.2.

Arit. priemer:	=AVERAGE(A1:A15)
Modus:	=MODE.SNGL(A1:A15)
Medián:	=MEDIAN(A1:A15)

Arit. priemer:	1086,00
Modus:	900
Medián:	1020

Obr. 2.2 Výpočet a výsledky k príkladu 2.2 (zdroj: vlastné spracovanie)

V ľavej časti obrázku 2.2 vidíme zápis funkcií do buniek (vstupné dáta sa nachádzajú v bunkách A1 až A15). Priemerná mzda zamestnancov podniku je 1086 €. Modus by sme mohli interpretovať napríklad tak, že najbežnejšia (najtypickejšia) mzda zamestnancov firmy je 900 €. Pri interpretácii mediánu môžeme povedať, že počet zamestnancov, ktorí majú mzdu nižšiu alebo rovnú 1020 €, je rovnaký, ako počet zamestnancov, ktorí majú mzdu vyššiu alebo rovnú 1020 €.

2.2.2 Charakteristiky variability

Charakteristiky variability, miery variability *(angl. variability)* alebo aj miery rozptylu vyjadrujú rozdielnosti v údajoch. Sú to kolísania hodnôt okolo centrálnej charakteristiky (priemeru). V ďalšom texte budú podrobnejšie vysvetlené najznámejšie charakteristiky variability.

Variačné rozpätie *(angl. range)* je rozdiel medzi najväčšou a najmenšou hodnotou v skúmanom súbore. Výpočet je veľmi jednoduchý, ale variačné rozpätie je len zriedka používaná charakteristika, lebo je výrazne ovplyvnená extrémnymi hodnotami.

Medzikvartilové rozpätie *(angl. interquartile range)* je rozdiel medzi tretím a prvým kvartilom (alebo medzi 75. a 25. percentilom). Je to oblasť stredných 50 % hodnôt usporiadaného súboru a túto charakteristiku neovplyvňujú extrémne hodnoty.

Rozptyl *(angl. variance)* je priemer štvorcov odchýlok hodnôt od aritmetického priemeru $\bar{x}$:

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

Ak by rozdiely neboli umocnené, tak kladné a záporné hodnoty by sa pri sčitovaní vzájomne vykrátili. Po umocnení je však sťažená interpretácia tejto charakteristiky, lebo nie je vyjadrená v rovnakých jednotkách ako pôvodné údaje.

Smerodajná odchýlka *(angl. standard deviation)* alebo tiež štandardná odchýlka je odmocnina z rozptylu.

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

Vďaka odmocňovaniu je výsledok opäť v pôvodných jednotkách. Smerodajnú odchýlku interpretujeme ako kolísanie hodnôt znaku okolo hodnoty aritmetického priemeru. Konkrétne môžeme povedať, že **väčšina** hodnôt bola v rozmedzí $\bar{x} \pm s$. Ak by sme napríklad vypočítali, že priemerná mzda v nejakom odvetví je 980 € a smerodajná odchýlka 45 €, tak môžeme povedať, že väčšina pracovníkov má mzdu vo výške (980 ± 45) €.

Variačný koeficient *(angl. coefficient of variation)* je podiel smerodajnej odchýlky a priemeru. Je to relatívna miera variability a po vynásobení číslom 100 dostaneme vyjadrenie v percentách. Môžeme ním porovnávať aj variabilitu súborov, ktoré majú iné jednotky merania alebo rôzne priemery. Variačný koeficient vyjadruje, koľko percent z priemernej hodnoty predstavuje smerodajná odchýlka.

V **Exceli** môžeme aj niektoré charakteristiky variability vypočítať pomocou funkcií, napríklad **VAR** (rozptyl) a **STDEV** (smerodajná odchýlka). Ďalšie sa dajú vypočítať pomocou jednoduchých operácií s bunkami. Podobne ako pri charakteristikách polohy, môžeme na výpočet niektorých charakteristík variability využiť aj doplnok Real Statistics v záložke **Desc**, prvá položka **Descriptive Statistics and Normality**.

Na začiatku každej kvantitatívnej analýzy je vhodné urobiť základný štatistický rozbor. Ten by mal pozostávať z určenia dôležitých štatistík, ako napríklad rozsah súboru, minimum, maximum, priemer, medián a smerodajná odchýlka. Tieto opisné štatistiky dávajú stručnú a pomerne jasnú predstavu o rozdelení súboru analyzovaných hodnôt.

Príklad 2.3 Vypočítajte a interpretujte smerodajnú odchýlku k údajom z príkl. 2.2.

Riešenie: Na výpočet smerodajnej odchýlky môžeme použiť v Exceli funkciu *STDEV*, ako je znázornené na obrázku 2.3.

Sm. odchýlka:	=STDEV.P(A1:A15)		Sm. odchýlka:	197,78

Obr. 2.3 Výpočet a výsledky k príkladu 2.3 (zdroj: vlastné spracovanie)

V ľavej časti obrázku 2.3 vidíme zápis funkcie do bunky (vstupné dáta sa opäť nachádzajú v bunkách A1 až A15). Vypočítaná smerodajná odchýlka je 197,78 €. Mohli by sme to interpretovať tak, že väčšina zamestnancov podniku má mzdu vo výške (1086 ± 197,78) €. Konkrétne v tomto príklade má v uvedenom intervale mzdu 11 pracovníkov z celkového počtu 15.

Poznámka: V príklade 2.3 bola namiesto funkcie *STDEV* použitá funkcia *STDEV.P*. Je to preto, lebo sme počítali smerodajnú odchýlku miezd všetkých zamestnancov podniku a považovali sme ich teda za základný súbor. Ak by sme mali k dispozícii iba dáta časti zamestnancov a považovali by sme ich za vzorku zo všetkých zamestnancov celého podniku, počítali by sme tzv. výberovú smerodajnú odchýlku pomocou funkcie *STDEV.S* (rovnaký výsledok dostaneme použitím funkcie *STDEV*). V interpretácii by sme potom použili slovo *odhad*, keďže by sme mali k dispozícii iba dáta za výberový súbor a nie celý základný súbor. Táto problematika je rozobratá neskôr. Čitateľ sa o pojmoch *výberový súbor*, *základný súbor*, *odhad* a ďalších, dozvie viac v nasledujúcich častiach publikácie, predovšetkým v kapitolách „*Úvod do inferenčnej štatistiky*“ a „*Bodový a intervalový odhad*“.

2.3 Grafické znázornenie

Ďalším veľmi dobrým spôsobom, ako znázorniť výsledky alebo namerané údaje, je graf. Grafické zobrazenie je zvyčajne veľmi názorné a umožňuje lepšiu predstavu a interpretáciu. Pre vytváranie grafov je vhodné využiť Excel, ktorý ponúka veľké množstvo rôznych typov grafov. Medzi najčastejšie používané grafy patria stĺpcové, histogramy, spojnicové a kruhové.

Stĺpcový graf sa používa na znázornenie absolútnych aj relatívnych početností z tabuľky rozdelenia početností. Využíva sa pre kvantitatívne znaky s konečným počtom obmien alebo pre kvalitatívne znaky. Na vodorovnej osi sú spravidla obmeny znaku a na zvislej ich početnosti.

Príklad 2.4 Vytvorte stĺpcový graf z frekvenčnej tabuľky príkladu 2.1 o hodnotení spokojnosti turistov so zájazdom.

Riešenie: V Exceli vyznačíme všetkých 5 buniek s obmenami v prvom riadku a tiež všetky absolútne početnosti pod týmito bunkami. Potom vložíme stĺpcový graf cez záložku *Vložiť* v hlavnom menu. Následne môžeme zmeniť názov grafu a pomocou zelenej ikony v tvare plus (vedľa grafu) môžeme označiť vodorovnú aj zvislú os (obrázok 2.4).

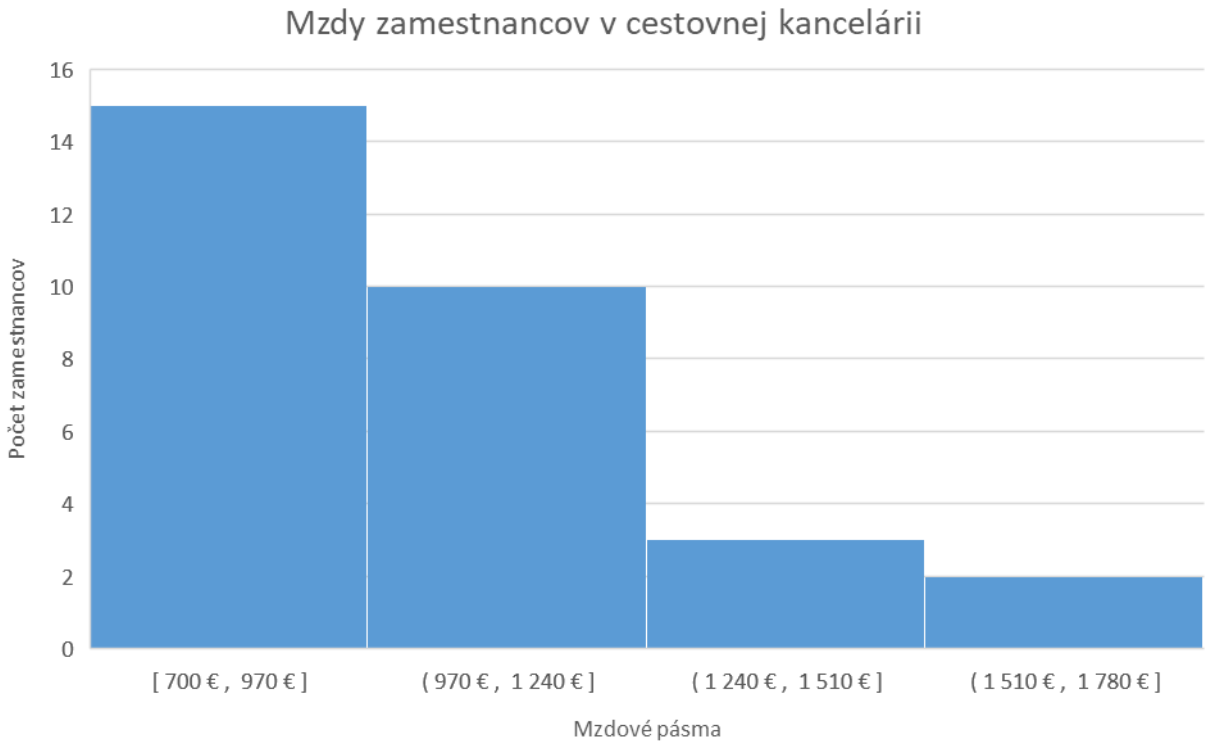
Obr. 2.4 Stĺpcový graf (zdroj: vlastné spracovanie)

Histogram je špeciálny typ stĺpcového grafu, ktorý sa používa na znázornenie početností pri intervalovom triedení. Vodorovná os obsahuje intervaly (triedy) a zvislá os početnosti v príslušných intervaloch. Na rozdiel od stĺpcového grafu

histogram nemá medzi stĺpcami medzery, sú tesne vedľa seba a základne stĺpcov sa rovnajú šírke intervalov. Realizácia tohto grafu v Exceli môže byť využitá aj na tvorbu frekvenčnej tabuľky pomocou intervalového triedenia.

Príklad 2.5 Vo firme pracuje 30 zamestnancov, pričom ich hrubé mzdy v eurách sú nasledovné: 1030, 950, 865, 700, 1780, 1225, 1235, 1005, 995, 840, 1530, 1340, 1000, 990, 700, 895, 942, 950, 1100, 1090, 1250, 960, 910, 840, 1380, 960, 1115, 850, 940, 720. Vytvorte histogram pre uvedené dáta.

Riešenie: Vstupné dáta prepíšeme v Exceli do stĺpca a potom označíme myšou všetky bunky s údajmi. Cez záložku *Vložiť* v hlavnom menu vyberieme medzi grafmi *histogram*. Podobne ako v predošlom príklade môžeme zmeniť názov grafu a pomocou zelenej ikony v tvare plus (vedľa grafu) aj vodorovnú a zvislú os (obrázok 2.5).



Obr. 2.5 Histogram (zdroj: vlastné spracovanie)

Počítač automaticky vytvoril mzdové pásma, ktoré je možné využiť aj na tvorbu tabuľky rozdelenia početností pomocou intervalového triedenia. Ak nám nevyhovuje aktuálny počet intervalov (v tomto prípade štyri), nie je problém to zmeniť. Klikneme v Exceli pravým tlačidlom myši do grafu a vyberieme položku *Formátovať rad údajov*. Rozbalíme položku *Možnosti radu* a vyberieme *Vodorovná*

Os kategórií. Aktivujeme položku *Počet priehradiek*, pričom vo vedľajšom okienku je vložené číslo 4, čo je aktuálny počet intervalov. Zmeníme ho na požadovanú hodnotu a počítač automaticky vypočíta nové hodnoty a upraví graf.

Spojnicový graf (polygón) vzniká pospájaním jednotlivých súradnicových bodov úsečkami. Využitie má pri kvantitatívnych znakoch s konečným počtom obmien a aj pri kvalitatívnych znakoch.

Príklad 2.6 Vytvorte spojnicový graf z frekvenčnej tabuľky príkladu 2.1 o hodnotení spokojnosti turistov so zájazdom.

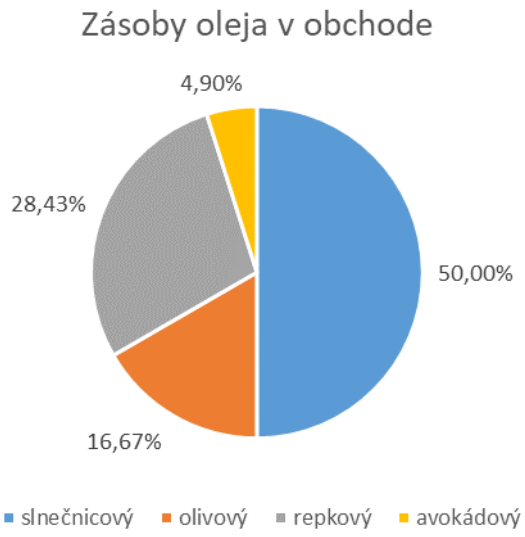
Riešenie: V Exceli vyznačíme všetkých 5 buniek s obmenami v prvom riadku a tiež všetky absolútne početnosti pod týmito bunkami. Potom cez záložku *Vložiť* v hlavnom menu vyberieme *Čiarový graf*. Podobne ako v predošlých príkladoch môžeme označiť vodorovnú aj zvislú os a tiež názov grafu (obrázok 2.6).

Obr. 2.6 Spojnicový graf (zdroj: vlastné spracovanie)

Kruhový (koláčový, výsekový) **graf** plochou kruhu znázorňuje celý súbor a jednotlivými výsekmi obmeny znaku. Je vhodný na vizualizáciu relatívnych početností, ktoré nemajú veľký počet obmien. Často zobrazuje kvalitatívne znaky, ako sú napríklad preferencie, zloženie populácie, rozdelenie výdavkov a podobne. Veľkosti stredových uhlov pri jednotlivých výsekoch sú úmerné relatívnym početnostiam.

Príklad 2.7 Obchod má v sklade nasledovné počty litrových balení oleja: 204 kusov slnečnicového, 68 kusov olivového, 116 kusov repkového a 20 kusov avokádového. Vytvorte kruhový graf, ktorý znázorňuje percentuálny podiel jednotlivých typov oleja na celkových zásobách.

Riešenie: V Exceli vytvoríme frekvenčnú tabuľku s relatívnymi početnosťami v percentách. Potom vyznačíme všetky bunky so slovným opisom obmien a tiež relatívne početnosti. Vyberieme *Koláčový graf* cez záložku *Vložiť* v hlavnom menu. Graf pomenujeme a pomocou zelenej ikony v tvare plus môžeme pridať označenia údajov (obrázok 2.7).



Obr. 2.7 Kruhový graf (zdroj: vlastné spracovanie)

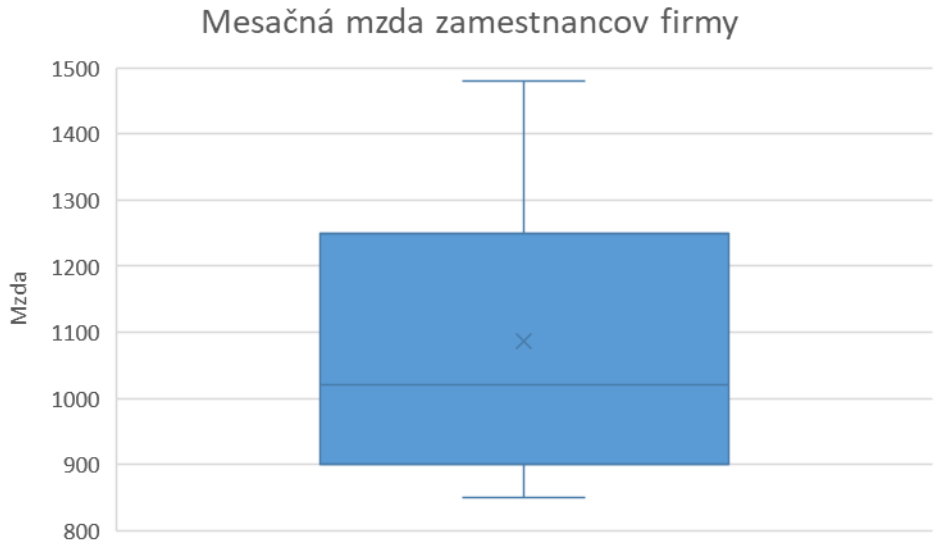
Veľkosť výseku kruhového grafu je úmerná podielu danej kategórie na celkovom súčte. Napríklad, ak má kategória 50 % podiel, jej výsek bude zaberať polovicu kruhu.

Krabicový graf vizualizuje viaceré charakteristiky. Môžeme sa stretnúť aj s označením krabičkový graf, škatuľový graf a hlavne **box plot**, čo je síce výraz z angličtiny, ale používa sa aj v slovenčine. Graf sa v štatistike veľmi často používa. Existuje v rôznych variáciách, ktoré ovplyvňujú jeho interpretáciu. V základnej verzii horizontálna čiara znázorňuje medián (2. kvartil), horná časť krabice 3. kvartil a dolná časť 1. kvartil, horná čiarka maximum a dolná čiarka minimum. Výška krabice je určená medzikvartilovým rozpätím. Celý graf je možné umiestniť vertikálne ale aj horizontálne. Okrem zobrazenia základných štatistických charakteristík umožňuje identifikovať aj odľahlé hodnoty (*angl. outliers*), ktoré môžu skresľovať výsledky

analýzy. Staršie verzie Excelu priamu tvorbu tohto grafu neobsahovali, avšak od verzie 2016 je už v programe zabudovaný.

Príklad 2.8 Vytvorte krabicový graf pre údaje z príkladu 2.2.

Riešenie: Vstupné dáta prepíšeme v Exceli do stĺpca a potom označíme myšou všetky bunky s údajmi. Cez záložku *Vložiť* v hlavnom menu vyberieme medzi grafmi *škatuľový*. Podobne ako v predošlých príkladoch môžeme označiť vodorovnú aj zvislú os, názov a upraviť škálu pre zobrazenie údajov (obrázok 2.8).



Obr. 2.8 Krabicový graf (zdroj: vlastné spracovanie)

Kontrolné otázky:

1. Čo je úlohou a cieľom popisnej štatistiky?
2. Ako by ste charakterizovali absolútne, relatívne a kumulatívne početnosti?
3. Kedy používame intervalové triedenie hodnôt?
4. Aké charakteristiky polohy poznáte a ktorá je najpoužívanejšia?
5. Akú nevýhodu má aritmetický priemer napríklad v porovnaní s mediánom?
6. Čo vyjadrujú charakteristiky variability?
7. Čo je to histogram?
8. Čo je to polygón?
9. Aké vlastnosti má box plot?

Úlohy na riešenie:

1. Vytvorte kruhový graf na základe nasledovných údajov o obľúbenosti druhov ovocia medzi 100 respondentmi. Jablká: 35 respondentov, banány: 20 respondentov, pomaranče: 15 respondentov, jahody: 12 respondentov, ostatné: 18 respondentov.

2. Vo firme pracuje 20 programátorov, pričom ich hrubé mzdy v eurách sú nasledovné: 2500, 3200, 1800, 2700, 4100, 2400, 3500, 2100, 2900, 1900, 3800, 2600, 2300, 3100, 1700, 2800, 3400, 2200, 3000, 2000. Vytvorte histogram pre uvedené dáta.

3. Na základe údajov o mesačných výdavkoch na potraviny v 50 domácnostiach (v eurách) vytvorte krabicový graf. Dáta: 250, 380, 190, 320, 450, 270, 410, 230, 350, 200, 420, 290, 260, 340, 180, 300, 370, 240, 360, 210, 430, 280, 390, 200, 310, 400, 250, 330, 190, 290, 440, 260, 350, 220, 380, 170, 340, 410, 230, 370, 190, 430, 270, 280, 310, 270, 350, 440, 450, 400.

4. Zisťovali sme cenu tej istej čokolády v 20 rôznych obchodoch, zozbierali sme nasledovné údaje v eurách: 3,05; 2,95; 3,25; 3,55; 3,5; 3,8; 2,9; 3,4; 3,3; 3; 3,1; 3,25; 3,8; 3,65; 3,4; 4,05; 3,5; 3,15; 3,8; 2,9. Vypočítajte aritmetický priemer, medián a modus ceny tejto čokolády v rôznych obchodoch.

   [Aritmetický priemer: 3,37; medián: 3,35; modus: 3,8]

3 PRAVDEPODOBNOSŤ

Kľúčové slová: pravdepodobnosť, náhodná premenná, distribučná funkcia, funkcia hustoty, rozdelenie pravdepodobnosti, normálne rozdelenie

Z deskriptívnej štatistiky vychádza tzv. inferenčná (induktívna) štatistika, ktorá rieši mnohé zaujímavé otázky a problémy. Ešte pred podrobnejším pohľadom na inferenčnú štatistiku, je potrebné venovať sa teórii pravdepodobnosti, s ktorou je táto oblasť štatistiky úzko spojená. Teória pravdepodobnosti je v súčasnosti podrobne rozpracovaná a veľmi rozsiahla. Keďže to však nie je hlavným zameraním tejto publikácie, budeme sa jej venovať iba stručne. Načrtneme určité základy potrebné k lepšiemu porozumeniu inferenčnej štatistiky, ktorej sa budeme venovať v ďalších kapitolách.

3.1 Základné pojmy

V praxi sa často môžeme stretnúť s tým, že výsledky určitých procesov alebo činností nevieme s istotou predpovedať ani pri zachovaní rovnakých experimentálnych podmienok. Takáto činnosť sa v teórii pravdepodobnosti nazýva **náhodný pokus**. Ako príklad náhodného pokusu môže slúžiť hod mincou, hod kockou a tak ďalej. Z hľadiska štatistiky je dôležitá ešte jedna vlastnosť a to je **hromadnosť**. Znamená to, že náhodný pokus je možné za rovnakých podmienok neobmedzene opakovať.

Každý náhodný pokus má nejakú množinu možných výsledkov. Túto množinu nazývame **základný priestor** a v literatúre sa zvyčajne označuje ako Ω (prípadne E). Konkrétny výsledok náhodného pokusu sa nazýva **elementárny jav** (elementárna udalosť). Písmenami gréckej abecedy $\omega_1, \omega_2, \omega_3, \ldots$ označujeme jednotlivé výsledky náhodného pokusu. Elementárne udalosti sa nedajú ďalej rozložiť, ale môžeme ich skladať na tzv. **náhodné javy** (náhodné udalosti). Náhodné javy sa zvyčajne označujú veľkými písmenami A, B, C, atď. Priestor všetkých náhodných javov sa označuje S.

Po tomto úvode sa dostávame ku kľúčovému pojmu tejto kapitoly a tým je samotná pravdepodobnosť. **Pravdepodobnosť** javu je istá miera očakávania, že skúmaný jav nastane. V histórii sa vyvinulo viacero pohľadov na to, ako presne pravdepodobnosť definovať. Existuje napríklad geometrická definícia, ktorou sa ale detailne nebudeme zaoberať. Čitateľ si v prípade záujmu vie nájsť viac informácií v rôznej inej literatúre (napríklad Somorčík – Teplička, 2015). Podrobnejšie sa budeme venovať klasickej definícii pravdepodobnosti.

Klasická definícia pravdepodobnosti (autorom je Laplace): Pravdepodobnosť javu A je definovaná takto: $P(A) = \frac{m}{n}$, kde m je počet výsledkov priaznivých udalosti A a n je počet všetkých možných výsledkov. Definíciu je možné použiť iba vtedy, ak sú

všetky výsledky náhodného pokusu rovnako možné. Okrem toho, klasická definícia pravdepodobnosti je *aprioristická* (lat. *a priori* znamená nezávisle od skúsenosti, resp. bez predchádzajúcich skúseností), čiže bez vykonania náhodného pokusu umožňuje vypočítať pravdepodobnosť. Ak by sme v praxi začali realizovať nejaký náhodný pokus, napríklad hod kockou, zrejme by frekvencia čísel bola odlišná od teoreticky vypočítanej hodnoty. Preto je zaujímavé zadefinovať tzv. štatistickú pravdepodobnosť, ktorá je *aposterioristická* (lat. *a posteriori* znamená na základe skúsenosti). To znamená, že umožňuje vypočítať pravdepodobnosť až po viacnásobnom vykonaní náhodného pokusu a určení relatívnej početnosti.

Štatistickú definíciu pravdepodobnosti v roku 1919 vytvoril Richard von Mises: Pravdepodobnosť javu A je číslo $P(A)$, ku ktorému sa približuje relatívna početnosť javu A, ak počet opakovaní náhodného pokusu neohraničene narastá, teda $P(A) = \lim_{n\to\infty} \frac{m(n)}{n}$. Číslo $\frac{m(n)}{n}$ nazývame relatívna početnosť javu A, ak náhodný pokus opakujeme n-krát a z toho jav A nastane $m(n)$-krát. Relatívna početnosť je len odhadom pravdepodobnosti $P(A)$. Čím je počet pokusov vyšší, tým je odhad lepší.

Axiomatickú definíciu pravdepodobnosti v roku 1933 vytvoril Kolmogorov. Doteraz uvedené definície mali svoje nedostatky (špecifiká). Preto sa v súčasnosti používa tzv. axiomatická definícia pravdepodobnosti. Zahŕňa všetky ostatné definície a je vytvorená aj pre prípad, že by sme uvažovali o nekonečnom počte výsledkov náhodného pokusu. Znenie definície je nasledovné:

Nech Ω je ľubovoľná neprázdna množina. Pod **pravdepodobnosťou** budeme rozumieť zobrazenie $P: S \to R$ definované na σ-algebre S podmnožín množiny Ω s týmito vlastnosťami:

1. $P(\Omega) = 1$.
2. $P(A) \geq 0$ pre každé $A \in S$.
3. Pre ľubovoľnú postupnosť $\{A_n\}_{n=1}^{\infty}$ po dvoch disjunktných množín z S platí $P(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n)$.

Systém S podmnožín neprázdnej množiny Ω sa nazýva σ-algebra, ak spĺňa nasledovné vlastnosti:

(i) $\Omega \in S$;
(ii) ak $A \in S$, tak $A' \in S$;
(iii) ak $A_1, A_2, \ldots \in S$, tak $\bigcup_{n=1}^{\infty} A_n \in S$.

Trojicu (Ω, S, P) nazývame pravdepodobnostný priestor. Je to matematický model náhodného pokusu s nasledovnými vlastnosťami: Ω je množina všetkých výsledkov náhodného pokusu, S je systém všetkých udalostí, ktorých pravdepodobnosť možno určiť a $P(A)$ je pravdepodobnosť udalosti A.

3.2 Náhodná premenná a rozdelenie pravdepodobnosti

Náhodná premenná je taká premenná, ktorej hodnotu určuje výsledok náhodného pokusu. Náhodnú premennú teraz zadefinujeme matematicky presne. Nech (Ω, S, P) je pravdepodobnostný priestor. Reálnu funkciu $X: \Omega \to R$ nazývame **náhodnou premennou**, ak pre každé $c \in R$ platí $\{\omega; X(\omega) < c\} \in S$.

Poznáme dva základné typy náhodných premenných.

Diskrétne náhodné premenné: Môžu nadobúdať konečný počet hodnôt, alebo nekonečný počet hodnôt, ktoré sa dajú očíslovať prirodzenými číslami 1, 2, 3, atď. V matematike takúto nekonečnú množinu hodnôt nazývame spočítateľná. Môže teda nadobúdať len izolované hodnoty. Napríklad počet turistov v krajine, počet chybných výrobkov, počet skrachovaných firiem a tak ďalej.

Spojité náhodné premenné: Môžu nadobúdať nekonečný počet hodnôt z istého intervalu alebo zjednotenia intervalov. Príkladom spojitej náhodnej premennej môže byť výška alebo hmotnosť náhodne vybraného človeka, doba čakania na autobus a tak ďalej.

Vzťah medzi hodnotami náhodnej premennej a pravdepodobnosťou ich výskytu je možné zhrnúť pomocou tzv. **rozdelenia pravdepodobnosti**. Pomenovania náhodných premenných sa prenášajú aj na pomenovanie rozdelenia ich pravdepodobnosti a tiež ich môžeme deliť na diskrétne a spojité. (Švábová a kol., 2022). Existujú síce aj náhodné premenné a ich rozdelenia pravdepodobnosti, ktoré sú kombináciou oboch typov, ale to nebude predmetom nášho záujmu.

Diskrétne rozdelenia pravdepodobnosti opisujú rozdelenie pravdepodobnosti diskrétnych náhodných premenných, ktoré nadobúdajú hodnoty s nenulovou pravdepodobnosťou. Priradenie pravdepodobnosti ku konkrétnym možným hodnotám diskrétnej náhodnej premennej sa nazýva **rozdelenie pravdepodobnosti** alebo **pravdepodobnostná funkcia**. Súčet pravdepodobností nastania všetkých možných hodnôt sa rovná jednej.

Spojité rozdelenia pravdepodobnosti necharakterizujeme pravdepodobnostnou funkciou, lebo hodnôt je nekonečne veľa (nespočítateľne veľké množstvo) a pravdepodobnosť, že nastane nejaká konkrétna hodnota sa rovná nule. Preto sa tu používa tzv. **funkcia hustoty** *(density function, probability density function)*. Hustota pravdepodobnosti spojitej náhodnej premennej X umožňuje vypočítať pravdepodobnosť, že náhodná premenná nadobudne hodnotu z určitého intervalu (a, b). Je to obsah plochy pod grafom tejto funkcie na príslušnom intervale, čo môžeme matematicky zapísať takto: $P(a < X < b) = \int_a^b f(x)dx$. Funkcia hustoty musí byť na uvedenom intervale spojitá a celá plocha pod krivkou sa musí rovnať jednej, čo môžeme vyjadriť nasledovne: $\int_{-\infty}^{\infty} f(x)dx = 1$.

Distribučná funkcia *(distribution function, cumulative distribution function)* náhodnej premennej X vyjadruje pravdepodobnosť, že náhodná premenná X nadobudne hodnotu menšiu ako x, čo zapisujeme $F(x) = P(X < x)$. Distribučnou funkciou môžeme opísať diskrétnu aj spojitú náhodnú premennú. Chajdiak (2010) uvádza, že distribučná funkcia je v teórii pravdepodobnosti teoretickým ekvivalentom kumulatívnej relatívnej početnosti.

Základnými charakteristikami náhodnej premennej sú stredná hodnota a rozptyl. **Stredná hodnota** premennej X sa zvyčajne označuje $E(X)$. V literatúre sa často môžeme stretnúť aj s pomenovaním **očakávaná hodnota** *(expected value)*. Okolo strednej hodnoty sú sústredené hodnoty náhodnej premennej. Môžeme tiež povedať, že okolo strednej hodnoty kolíšu hodnoty náhodnej premennej pri opakovaných realizáciách náhodného pokusu, resp. k strednej hodnote sa blíži dlhodobý priemer realizácií náhodnej premennej X.

V prípade diskrétnej náhodnej premennej sa stredná hodnota vypočíta nasledovne:

$$E(X) = \sum_{i=1}^{n} x_i . P(X = x_i).$$

Je to vlastne vážený aritmetický priemer hodnôt náhodnej premennej X, pričom váhami sú ich pravdepodobnosti. Keďže pravdepodobnosti sú už v relatívnom vyjadrení, nedelíme celkovým počtom hodnôt, ako to bežne býva v prípade výpočtu aritmetického priemeru.

V prípade spojitej náhodnej premennej sa stredná hodnota vypočíta nasledovne:

$$E(X) = \int_{-\infty}^{\infty} x . f(x) dx.$$

Rozptyl *(dispersion)* náhodnej premennej X je najčastejšie používanou charakteristikou jej variability. Môžeme sa tiež stretnúť s názvami disperzia alebo variancia. Je definovaný vzťahom:

$$D(X) = E[(X - E(X))^2],$$

prípadne praktickejším tvarom na výpočet:

$$D(X) = E(X^2) - E^2(X).$$

V prípade diskrétnej náhodnej premennej je $E(X^2) = \sum_{i=1}^{n} x_i^2 . p_i$ a v prípade spojitej náhodnej premennej je $E(X^2) = \int_{-\infty}^{\infty} x^2 . f(x) dx$.

Nevýhodou rozptylu je jeho problematická interpretácia. Na tento účel sa preto využíva **smerodajná odchýlka** σ, ktorú vypočítame ako druhú odmocninu rozptylu, čo môžeme zapísať takto: $\sigma(X) = \sqrt{D(X)}$.

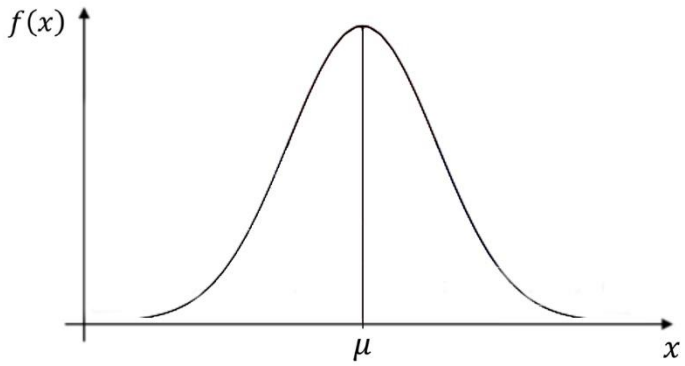
3.3 Normálne rozdelenie pravdepodobnosti

Ako sme už spomínali, rozdelenia pravdepodobnosti môžeme členiť na diskrétne (alternatívne, binomické, rovnomerné, Poissonovo, atď.) a spojité (Studentovo t-rozdelenie, chí-kvadrát rozdelenie, atď.). Podrobný popis konkrétnych rozdelení je možné nájsť v príslušnej literatúre. My sa budeme zaoberať normálnym rozdelením, ktoré je kľúčové z hľadiska ďalších kapitol. Normálne rozdelenie patrí medzi spojité rozdelenia, ale v praxi sa ním často aproximujú aj diskrétne rozdelenia, ak nadobúdajú dostatočne veľké množstvo hodnôt. Teraz ho podrobnejšie charakterizujeme.

Ak má náhodná premenná *X* hustotu pravdepodobnosti vyjadrenú funkciou

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

hovoríme, že má **normálne rozdelenie pravdepodobnosti**. Niekedy sa môžeme stretnúť aj s označením Gaussovo rozdelenie, lebo graf hustoty pravdepodobnosti normálneho rozdelenia sa nazýva **Gaussova krivka** (obr. 3.1).



Obr. 3.1 Gaussova krivka (zdroj: vlastné spracovanie)

Hustota normálneho rozdelenia má **zvonovitý tvar** a je to **symetrická** funkcia podľa priamky $x = \mu$. Normálne rozdelenie má jediný modus (maximálna hodnota hustoty), to znamená, že je **unimodálne**. **Modus**, **medián** a **aritmetický priemer** majú tú istú hodnotu (**sú totožné**). Oba **konce** normálnej distribúcie **sa asymptoticky blížia k nule**. To znamená, že sú stále bližšie k nulovej hustote pravdepodobnosti, nikdy však samotnú nulovú hodnotu nedosiahnu (Ferjenčík, 2009). Tvar zvonovitej krivky normálneho rozdelenia je ovplyvnený **dvoma kľúčovými parametrami**: μ a σ^2. Parameter μ môže nadobúdať ľubovoľné hodnoty vrátane záporných. Určuje polohu krivky a zároveň indikuje bod na osi x, kde krivka dosahuje svoje maximum. Šírku „zvonu“ určuje parameter σ^2, ktorý je vždy kladný. Niektorí autori ako druhý

parameter používajú σ. Parameter μ predstavuje tiež strednú hodnotu $E(X)$ náhodnej premennej X. Teda môžeme písať:

$$E(X) = \mu.$$

Pre rozptyl normálneho rozdelenia platí:

$$D(X) = \sigma^2.$$

Normálne rozdelenie má v štatistike významné postavenie a jeho dôležitosť sa ukáže v ďalších kapitolách. Mnohé zo štatistických testov, ktoré budú popísané neskôr, sú založené na predpoklade normality rozloženia dát.

V **Exceli** používame funkciu **NORM.DIST** na výpočet hodnôt distribučnej funkcie a funkcie hustoty. Funkcia **NORM.INV** (inverzná distribučná funkcia) pre zadanú pravdepodobnosť p počíta zodpovedajúcu hodnotu x (p x 100 % kvantil).

Príklad 3.1 Čas potrebný na vypracovanie posudku úradníkom k žiadosti o dotáciu sa riadi normálnym rozdelením s priemerom 100 minút a smerodajnou odchýlkou 25 minút. Vypočítajte, koľko percent žiadostí bude pre úradníka veľmi náročných a ich posudzovanie mu zaberie viac ako 2 hodiny (120 minút) na jednu žiadosť?

Riešenie: Na výpočet použijeme v Exceli funkciu *NORM.DIST*, ako je znázornené na obrázku 3.2.

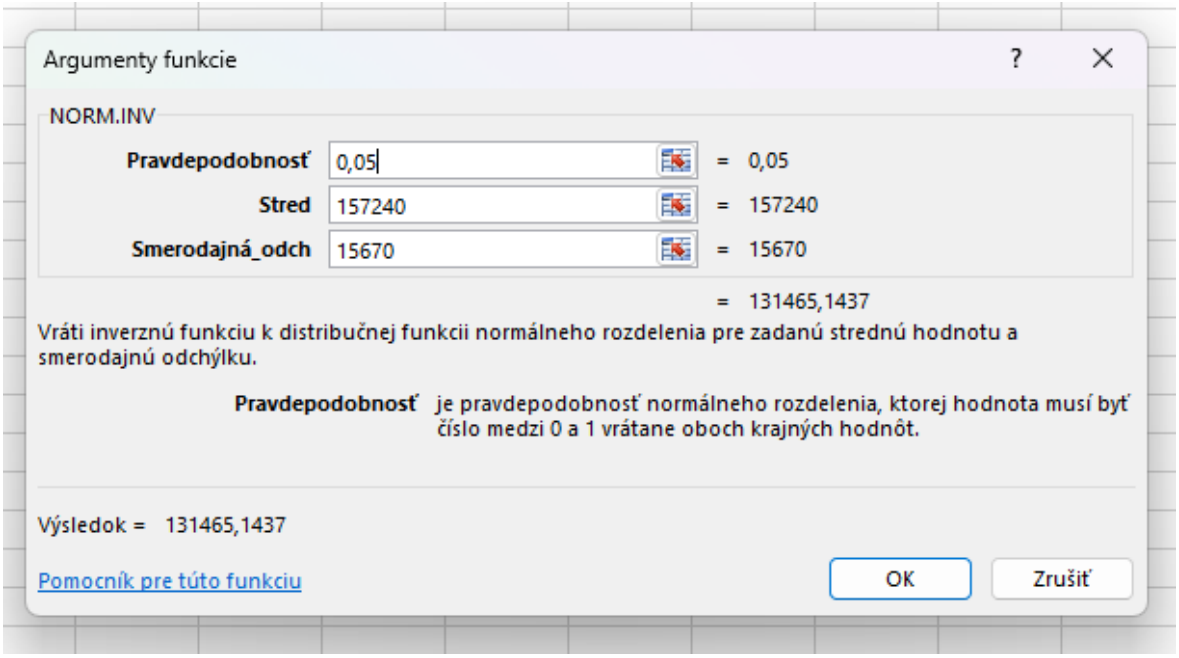
P(X≤120)=	=NORM.DIST(120;100;25;1)	P(X≤120)=	0,7881
P(X>120)=	=1-0,7881	P(X>120)=	0,2119

Obr. 3.2 Výpočet a výsledky k príkladu 3.1 (zdroj: vlastné spracovanie)

V ľavej časti obrázka vidíme zápis do buniek, v pravej časti obrázka sú výsledky. V dialógovom okne funkcie *NORM.DIST* je potrebné zadať 4 údaje. Prvý údaj je hraničná hodnota 120 minút, ktorá rozdeľuje normálne rozdelenie na dve časti. Druhý údaj je stredná hodnota (100), ďalej je potrebné zadať smerodajnú odchýlku (25) a na záver vložíme číslo 1 (TRUE) pretože nás zaujíma distribučná funkcia. Vypočítaná hodnota 0,7881 je však opačná udalosť, preto ju ešte musíme odpočítať od čísla 1. Približne 21,19 % žiadostí o dotáciu bude pre úradníka náročných, nakoľko ich posudzovanie mu môže zabrať viac ako 2 hodiny.

Príklad 3.2 Výrobca automobilov chce stanoviť dĺžku záruky na počet najazdených kilometrov, ktorú zákazníkom poskytne na vyrobené automobily. Testy vozidiel a informácie z praxe ukazujú, že poruchovosť (vznik prvej poruchy) sa riadi normálnym rozdelením s priemerom 157 240 km a smerodajnou odchýlkou 15 670 km. Výrobca chce zákazníkovi poskytnúť záruku čo najväčšieho počtu najazdených kilometrov, ale tak, aby sa s prvou poruchou vrátilo na reklamáciu maximálne 5 % predaných vozidiel. Aký počet najazdených kilometrov môže ponúknuť?

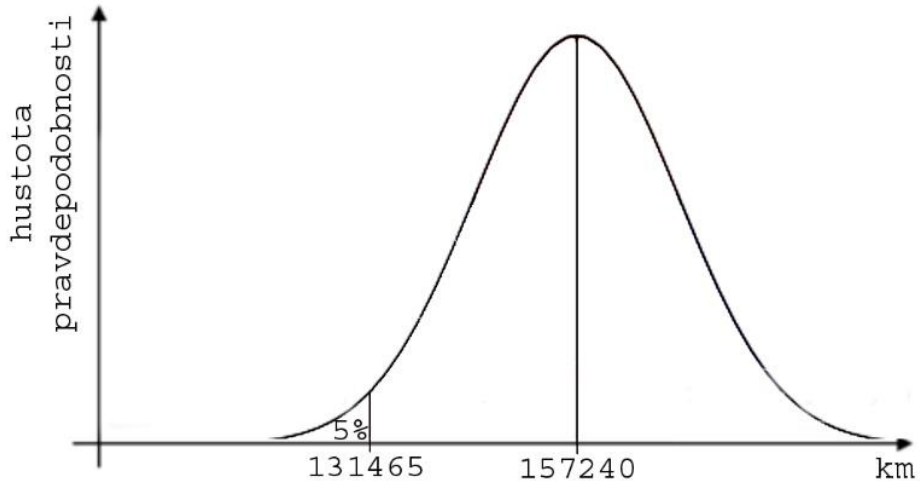
Riešenie: Na výpočet použijeme v Exceli funkciu *NORM.INV*, ako je znázornené na obrázku 3.3.



Obr. 3.3 Dialógové okno funkcie NORM.INV k príkladu 3.2 (zdroj: vlastné spracovanie)

Počet najazdených kilometrov automobilov je náhodná premenná s normálnym rozdelením. Našou úlohou je nájsť taký počet kilometrov, aby pravdepodobnosť, že náhodná premenná nadobudne túto alebo nižšiu hodnotu bola 5 %. V dialógovom okne funkcie *NORM.INV* je potrebné zadať 3 údaje. Do prvého riadku píšeme pravdepodobnosť normálneho rozdelenia, pre ktorú chceme vypočítať zodpovedajúcu hodnotu príslušného kvantilu. Pod ňou je priemer a v poslednom riadku sa nachádza smerodajná odchýlka. Po kliknutí na tlačidlo OK sa v bunke zobrazí výsledok 131 465. Aby výrobca nemusel riešiť viac ako 5 % reklamácií prvej poruchy vozidla, môže poskytnúť záruku na 131 465 najazdených kilometrov. Môže poskytnúť, samozrejme, aj záruku na menší počet najazdených kilometrov a poruchovosť neprekročí 5 %. Ale vypočítané číslo je maximálna možná hodnota, aby sa neprekročilo požadované percento očakávaných reklamácií. V praxi by sa to

číslo pravdepodobne nejako „rozumne“ zaokrúhlilo, napríklad na hodnotu 130 000 km.



Obr. 3.4 Grafické znázornenie riešenia príkladu 3.2 (zdroj: vlastné spracovanie)

Na obrázku 3.4 je znázornená grafická interpretácia riešenia.

Kontrolné otázky:

1. Čo je to pravdepodobnosť a aké definície pravdepodobnosti poznáte?
2. Prečo nás z hľadiska zamerania tejto publikácie a z praktického hľadiska najviac zaujíma štatistická definícia pravdepodobnosti?
3. Vedeli by ste uviesť nejaké príklady na diskrétne náhodné premenné a spojité náhodné premenné?
4. Kde sa v bežnom živote napríklad stretávame s pravdepodobnosťou?
5. Aký je rozdiel medzi distribučnou funkciou a funkciou hustoty?
6. Čo je to rozdelenie pravdepodobnosti náhodnej premennej?
7. Aké rozdelenia pravdepodobnosti by ste vedeli vymenovať?
8. Aké vlastnosti má normálne rozdelenie?
9. Aký tvar má Gaussova krivka?
10. Prečo je pravdepodobnosť dôležitá v štatistike?

Úlohy na riešenie:

1. V továrenskej prevádzke bola vykonaná kontrola kvality na vzorke 5000 náhodne vybraných súčiastok. Z tejto vzorky bolo identifikovaných 12 chybných súčiastok. Aká je odhadovaná pravdepodobnosť, že ak v budúcnosti náhodne vyberieme súčiastku na kontrolu, tak bude chybná? Predpokladáme konštantnú mieru chybovosti v čase.

   [0,0024]

2. Webová stránka zaznamenala v priebehu dňa 1250 návštev z rôznych zdrojov. Z nich 295 bolo z mobilných zariadení. Aká je odhadovaná pravdepodobnosť, že náhodná návšteva tejto webovej stránky v budúcnosti bude z mobilného zariadenia? Predpokladáme konštantnú mieru návštev z mobilných zariadení.

   [0,236]

3. Z archívu prednášajúceho sme náhodne vybrali 30 testov zo štatistiky, ktoré študenti v minulosti vypracovali na skúške. Každý študent dosiahol skóre v rozmedzí od 0 do 100 bodov. Určte, či sa jedná o diskrétnu alebo spojitú náhodnú premennú v súvislosti s výsledkami testov zo štatistiky.

   [diskrétna náhodná premenná]

4. Doba čakania na opravu vozidla v autoservise sa riadi normálnym rozdelením s priemerom 42 minút s smerodajnou odchýlkou 12 minút. Vypočítajte, koľko percent zákazníkov strávi v autoservise pri čakaní viac ako 60 minút?

   [6,68 %]

4 ÚVOD DO INFERENČNEJ ŠTATISTIKY

Kľúčové slová: základný súbor, výberový súbor, cenzus, zovšeobecňovanie, inferenčná (induktívna) štatistika, rozsah výberového súboru, náhodný výber, stratifikovaný výber, zámerný výber, dostupný výber

Štatistika preniká do života takmer každého človeka, bez ohľadu na oblasť, v ktorej pôsobí. Veľa našich poznatkov je založených na štatistických informáciách. Pomocou štatistických metód dokážeme dáta spracovávať, analyzovať ich a vyvodzovať z nich závery. Výskum obvykle začína plánovaním a navrhnutím vhodného postupu zberu dát. Deskriptívna štatistika, ktorou sme sa zaoberali doteraz, ponúka veľmi dobrý základ na prvotné prehľadné spracovanie údajov. Deskriptívna štatistika sa používa prakticky v každej publikácii, či už sa jedná o bakalársku, diplomovú alebo inú záverečnú prácu, vedecký článok a tak ďalej. Nie vždy si však s deskriptívnou štatistikou vystačíme. Preto je pre splnenie cieľov a výskumných úloh často potrebné použiť pokročilejšie štatistické metódy.

Množina všetkých objektov, na ktoré sa majú vzťahovať výsledky výskumu sa nazýva **základný súbor** (niekedy sa označuje slovom populácia). Napríklad ak skúmame všetky hotely v Slovenskej republike, potom tieto hotely tvoria základný súbor. Je zrejmé, že základný súbor môže byť pre výskumníka často priveľký a jeho celé preskúmanie môže byť nákladné alebo nemožné. Preto je zvyčajne potrebné vybrať zo základného súboru časť prvkov, ktoré budú základný súbor reprezentovať. Túto vybratú podmnožinu základného súboru nazývame **výberový súbor** alebo **výskumná vzorka**. Vzorka je vždy menšia ako základný súbor. Pri splnení určitých predpokladov môžeme závery zo skúmania vzorky, za použitia správnych štatistických metód, **zovšeobecniť** na celý základný súbor. Časť štatistiky, ktorá sa snaží odhadnúť vlastnosti základného súboru na základe skúmania výberového súboru sa nazýva **inferenčná** (**induktívna**) štatistika. Vo výnimočných prípadoch môže byť skúmaný aj celý základný súbor (zvyčajne ak nie je príliš veľký). Takýto výber nazývame **cenzus** (je to úplný výber).

Využitie inferenčnej štatistiky v praxi je veľmi široké. Je možné testovať hypotézy o vzťahoch medzi rôznymi ekonomickými premennými, vyvodzovať závery o celej populácii na základe vzorky, čo môže pomôcť pri rozhodovaní o rôznych investíciách, realizovať marketingové prieskumy a skúmať spotrebiteľské zvyky. Štatistické riadenie kvality sa uplatňuje aj v priemysle. Pri sériovej výrobe veľkého množstva súčiastok a výrobkov nie je možné skontrolovať každý produkt. Preto sa robí kontrola iba na náhodne vybratej vzorke a výsledky sa pomocou štatistiky zovšeobecňujú. V neposlednom rade má inferenčná štatistika svoje uplatnenie aj rôznych prieskumoch verejnej mienky, či už sa jedná o politické preferencie, hodnotové otázky alebo rôzne iné témy.

4.1 Zostavovanie výberového súboru

Cieľom pri výbere údajov je dosiahnuť čo najlepší výberový súbor, ktorý má všetky podstatné znaky základného súboru, dobre odráža jeho štruktúru, len má menej prvkov. Takýto výberový súbor môžeme nazvať **reprezentatívny**. Metodológia výskumu ponúka viac spôsobov získavania dát. Gavora a kol. (2010) rozlišuje náhodný, stratifikovaný, zámerný a dostupný výber.

Náhodný výber je taký, pri ktorom majú všetky prvky základného súboru rovnakú pravdepodobnosť, že budú vybrané. Rozlišujeme náhodné vyberanie s opakovaním (po každom výbere sa prvok vráti späť do základného súboru) a vyberanie bez opakovania (vybraté prvky ostanú pri ďalšom výbere mimo základného súboru). V prísnom zmysle slova podmienky náhodného výberu spĺňa iba vyberanie s opakovaním, lebo vtedy je výber realizovaný stále z rovnakého počtu prvkov a teda s rovnakou pravdepodobnosťou pre všetky prvky. V praxi sa však častejšie realizuje výber bez opakovania, lebo pri početnejších základných súboroch nemá zmysel rozlišovať medzi výberom s opakovaním a bez opakovania (Chráska, 2016). Pri náhodnom výbere sa neuplatňujú žiadne subjektívne názory výskumníka a z hľadiska zovšeobecniteľnosti je tento výber považovaný za najlepší.

Stratifikovaný výber patrí tiež medzi náhodné, takže je dobrým predpokladom pre zovšeobecňovanie. Základný súbor sa tu rozkladá podľa podstatných znakov (napríklad vek, pohlavie, veľkosť sídla, lokalita, vzdelanie, charakteristiky osobnosti a podobne). Proporcia vybraných subjektov musí v každom znaku zodpovedať proporcii v základnom súbore. Z každej kategórie sa subjekty vyberajú náhodným spôsobom. Typické využitie pre stratifikovaný výber sú napríklad prieskumy verejnej mienky.

Zámerný výber je charakteristický tým, že výskumník sám určí znaky (musia byť relevantné pre daný výskum), podľa ktorých bude vyberať prvky do vzorky. Je to teda dobre zvážený kvalifikovaný výber. Zovšeobecniteľnosť sa však viaže iba na tento zámerný výber, takže výsledky nie je možné zovšeobecniť na celú populáciu.

Dostupný výber je z hľadiska zovšeobecniteľnosti najslabší. Prvky do výskumu (resp. respondenti) sú vyberané na základe dostupnosti. Napríklad spolužiaci, rodina, kamaráti, ľudia z blízkeho podujatia a podobne. Študenti ho často používajú v bakalárskych a diplomových prácach.

V štatistike existujú aj iné výbery a iné členenie, ale pre potreby a rozsah tejto publikácie bude postačujúce uvedené metodologické delenie. Pre zaujímavosť spomenieme ešte tzv. metódu **snehovej gule** (**snowball sampling**), ktorá patrí medzi nepravdepodobnostné metódy a nie je založená na náhodnom výbere. Využíva existujúce (zväčša sociálne) väzby medzi štatistickými jednotkami (osobami) na získavanie potrebných údajov, a to najmä v situáciách, kde nie je ľahké dostať sa k

štatistickým jednotkám. Ide teda o populácie ťažko dosiahnuteľné, akými sú drogoví dealeri, drogovo závislí, rôzne undergroundové sociálne skupiny, skupiny organizovaného zločinu a podobne (Lyócsa a kol., 2013).

Na dosiahnutie hodnotných výsledkov a splnenie predpokladov na zovšeobecnenie nestačí len správnym spôsobom zostaviť výskumnú vzorku. Dôležitý je aj **rozsah výberového súboru**. Väčší rozsah je predpokladom lepšej zovšeobecniteľnosti. Je ťažké jednoznačne stanoviť, akú veľkosť by mala mať výskumná vzorka. Závisí to od viacerých faktorov. Napríklad od počtu premenných (viac premenných vyžaduje väčší rozsah vzorky), veľkosti základného súboru (väčší základný súbor vyžaduje väčší rozsah vzorky), ale aj od dôležitosti výsledkov. Pre výskumné práce sa zvyčajne používa vzorka o veľkosti 500 až 1500, prieskumy verejnej mienky sa zvyčajne realizujú na vzorke približne 1200 respondentov. Významné celoštátne výskumy napríklad v školstve môžu byť realizované aj na niekoľko tisíc subjektoch.

Niektorí autori (napríklad Švábová a kol., 2022) uvádzajú aj vzorce na výpočet veľkosti vzorky. U iných autorov môžeme nájsť tabuľky vyjadrujúce odhad veľkosti vzorky v závislosti od veľkosti základného súboru. Takéto tabuľky síce nie sú celkom presné, ale v tabuľke 4.1 uvádzame pre orientačnú predstavu aspoň niekoľko čísel.

Tab. 4.1 Približný vzťah medzi veľkosťou základného súboru a vzorky, (zdroj: Gavora a kol., 2010)

Veľkosť základného súboru	Odhad veľkosti výberového súboru
100	80
200	135
300	169
400	196
500	217
1 000	278
1 500	357
10 000	370

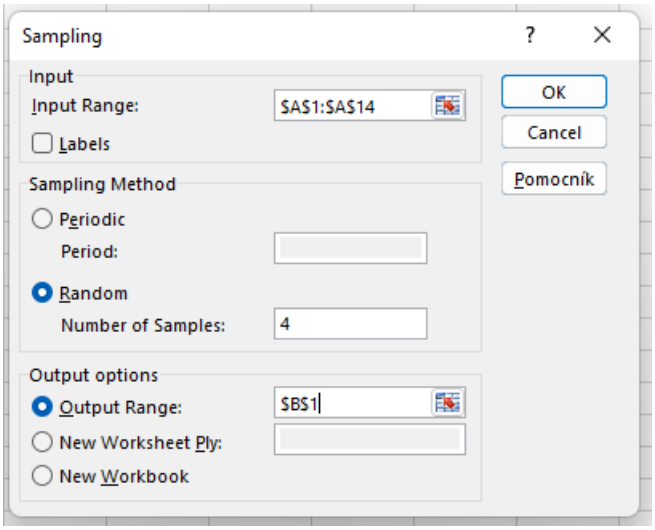
Na bakalárske a diplomové práce zvyčajne nie sú kladené také vysoké nároky a počet respondentov tam býva výrazne menší. Logicky aj zovšeobecniteľnosť výsledkov týchto prác je potom nižšia. Na druhej strane, na rozdiel od renomovaných vedeckých časopisov, to ani nie je hlavný cieľ. Zmyslom bakalárskych a diplomových prác je skôr pochopenie problematiky realizácie výskumu. Dôležité je, aby študent ovládal správne metodologické postupy získania, spracovania, analýzy a interpretácie údajov.

4.2 Realizácia náhodného výberu pomocou počítača

Pri realizácii náhodného výberu v praxi je potrebné zabezpečiť princíp náhodnosti. Najjednoduchším spôsobom je žrebovanie, vykonať ho môžeme aj manuálne napríklad s nastrihanými papierikmi. V prípade väčšej vzorky je to však zdĺhavý spôsob. Mnohé kalkulačky dnes obsahujú funkciu generovania náhodných čísel, najjednoduchšie je však využiť túto funkciu na počítači. Uvedieme príklady v programe Excel, kde je možné generovanie náhodných čísel buď pomocou nástroja „*Sampling*“ (výbery s opakovaním) alebo využitím funkcie „*RAND*“.

Príklad 4.1 Máme základný súbor, ktorý obsahuje 14 prvkov. Z neho je potrebné vybrať 4 prvky náhodným výberom s opakovaním.

Riešenie: Prvky očíslujeme od 1 do 14 a túto postupnosť prirodzených čísel napíšeme do buniek prvého stĺpca od *A1* po *A14*. V hlavnom menu vyberieme *Údaje*, potom *Data Analysis*, následne Sampling a na záver klikneme na *OK*. Otvorí sa okno nástroja *Sampling*. Do položky *Input Range* vložíme bunky *A1* až *A14*. V oblasti *Sampling Method* vyberieme *Random* a do políčka *Number of Samples* napíšeme číslo 4 (počet prvkov, ktoré je potrebné vybrať). Nakoniec vyberieme oblasť, kam má počítač umiestniť výsledky. Do políčka *Output Range* zadáme bunku *B1* (viď obrázok 4.1).



Obr. 4.1 Okno nástroja *Sampling* (zdroj: vlastné spracovanie)

Po kliknutí na *OK* sa v *B* stĺpci zobrazia štyri čísla, ktoré vznikli náhodným vyberaním a môžu sa opakovať (obr. 4.2). Čitateľ môže mať v tomto stĺpci logicky iné čísla.

	A	B	C
1	1	12	
2	2	14	
3	3	3	
4	4	2	
5	5		
6	6		
7	7		
8	8		
9	9		
10	10		
11	11		
12	12		
13	13		
14	14		
15			
16			

Obr. 4.2 Náhodne vybraté čísla s opakovaním (zdroj: vlastné spracovanie)

Poznámka: Čísla, ktoré sa nachádzajú v stĺpci *A*, nemusia tvoriť postupnosť prirodzených čísel. Ak by to bolo z nejakého dôvodu potrebné, čísla môžu byť ľubovoľné. Excel to bude brať pri generovaní do úvahy a výber realizuje iba z čísel v označenej oblasti. Zmeniť vstupné čísla na slovné označenie prvkov však nie je možné, nástroj *Sampling* je iba pre číselné hodnoty. V príkladoch 4.1 a 4.2 pracujeme s veľmi malým počtom prvkov. Príklady sú iba ilustračné, v praxi býva počet prvkov zvyčajne oveľa vyšší.

Príklad 4.2 Máme základný súbor, ktorý obsahuje 14 prvkov. Z neho je potrebné vybrať 4 prvky náhodným výberom bez opakovania.

Riešenie: Na obrázku 4.3 sú bunky v Exceli, z ktorých je vidieť riešenie. Krok za krokom teraz prejdeme postup, ako sa k nemu dostaneme.

	A	B	C
1	3	0,00212	
2	4	0,214683	
3	5	0,233896	
4	14	0,25214	
5	2	0,258284	
6	8	0,423639	
7	7	0,611422	
8	1	0,642337	
9	9	0,642884	
10	13	0,665253	
11	11	0,761549	
12	6	0,790297	
13	12	0,836639	
14	10	0,869175	
15			
16			

Obr. 4.3 Náhodne vybraté čísla bez opakovania (zdroj: vlastné spracovanie)

Prvky očíslujeme od 1 do 14 (namiesto čísel je možné použiť aj slovné označenie štatistických jednotiek) a napíšeme ich do buniek prvého stĺpca od *A1* po *A14*. Do buniek *C1* až *C14* vygenerujeme náhodné čísla pomocou príkazu *=RAND()*. Všetky tieto čísla skopírujeme a prilepíme špeciálne ako hodnoty do buniek *B1* až *B14* (aby sa hodnoty v tomto stĺpci znova nepregenerovali). Potom stĺpec *C* vymažeme (tento krok môžeme aj vynechať). Oblasť A1 až B14 zoradíme podľa hodnôt B stĺpca a nezáleží na tom, či zostupne alebo vzostupne (obr. 4.3). V bunkách A1 až A4 je náhodný výber bez opakovania.

Kontrolné otázky:

1. Aký je rozdiel medzi základným súborom a výberovým súborom?
2. Čo je to cenzus?
3. Čím sa zaoberá inferenčná štatistika?
4. Aké vlastnosti by mal spĺňať náhodný výber?
5. Ako by ste charakterizovali stratifikovaný výber?
6. Aký spôsob zostavovania výberového súboru je najlepší pre zovšeobecnenie?
7. Prečo zvyčajne nepracujeme s celým základným súborom, ale vyberáme vzorku?
8. Prečo je niekedy ťažké zrealizovať náhodný výber v praxi?
9. Aký rozsah by mal mať výberový súbor?

Úlohy na riešenie:

1. Navrhnite, aké objekty by mohli tvoriť základný súbor v záverečných prácach študentov.

2. Porozmýšľajte, aké spôsoby zostavovania výberového súboru sú v možnostiach a silách študenta v záverečnej práci vzhľadom na zvolený základný súbor.

3. Odhadnite, aký rozsah výberového súboru je študent schopný realizovať v záverečnej práci a ako to súvisí so zovšeobecniteľnosťou výsledkov.

4. Uveďte príklad, kedy je možné v záverečnej práci študenta realizovať cenzus.

5 ODHADOVANIE PARAMETROV ZÁKLADNÉHO SÚBORU

Kľúčové slová: bodový odhad, intervalový odhad, hladina významnosti, interval spoľahlivosti

V induktívnej štatistike sa riešia dve základné úlohy a to je odhadovanie parametrov základného súboru a testovanie štatistických hypotéz. V tejto kapitole sa oboznámime s prvou úlohou. Každá opisná charakteristika základného súboru (populácie) sa nazýva **parameter**. Opisná charakteristika, ktorá je vypočítaná zo vzorky, sa nazýva **štatistika**. Keďže základný súbor sa nemení, tak parametre sú pevné hodnoty, ktoré však zvyčajne nepoznáme. V praxi skúmame často veľmi veľké základné súbory, parametre sú preto pre nás neznámou konštantou. Naproti tomu štatistiky sa môžu meniť pre každý náhodný výber. Pomocou výberových údajov môžeme odhadnúť parametre základného súboru tzv. bodovým odhadom alebo intervalovým odhadom.

5.1 Bodový odhad

Podstatou bodového odhadovania parametra je odhad parametrov základného súboru na základe jedného konkrétneho bodu (konkrétnej hodnoty). Túto konkrétnu hodnotu (štatistiku) vypočítame z výberového súboru a prehlásime ju za odhad príslušného parametra. Štatistika, ktorou odhadujeme parameter, musí spĺňať viaceré kritériá. Medzi takéto základné kritériá patrí neskreslenosť (nevychýlenosť), konzistentnosť, výdatnosť, dostatočnosť a robustnosť. Podrobnejšie sa nimi nebudeme zaoberať, viac informácií na túto tému je možné nájsť v odbornej literatúre, napríklad Pacáková a kol. (2015). Pre správny výpočet bodových odhadov je potrebné vedieť nasledovné.

Strednú hodnotu základného súboru označujeme μ a odhadujeme pomocou **výberového priemeru**, ktorý označujeme $\overline{x}$.

Rozptyl základného súboru označujeme σ^2 a odhadujeme pomocou **výberového rozptylu**, ktorý označujeme s^2.

Smerodajnú odchýlku základného súboru označujeme σ a odhadujeme pomocou **výberovej smerodajnej odchýlky**, ktorú označujeme s.

Podiel kvalitatívneho znaku základného súboru označujeme π a odhadujeme pomocou **výberového podielu**, ktorý označujeme p.

Výhodou bodových odhadov je ich **jednoduchý spôsob výpočtu** a tiež vysoká **konkrétnosť**, ktorú vyjadruje **jeden bod**. Naopak **nevýhodou** bodových odhadov je ich veľmi **nízka spoľahlivosť**. Myslí sa tým podiel priaznivých javov vyjadrených jedným bodom k celkovému počtu všetkých možných výsledkov bodových odhadov.

Príklad 5.1 Pri štatistickom riadení kvality sa vo výrobnej linke kontroloval rozmer (priemer) vyrobených súčiastok. Z celej produkcie bolo náhodne vybratých 25 kusov a namerané boli tieto rozmery v milimetroch: 32,15; 32,17; 32,22; 32,11; 32,14; 32,33; 32,24; 32,12; 32,19; 32,17; 32,18; 32,09; 32,19; 32,16; 32,21; 32,24; 32,28; 32,15; 32,13; 32,22; 32,19; 32,18; 32,15; 32,23; 32,26.

Predpokladáme, že údaje pochádzajú zo súboru s normálnym rozdelením. Vypočítajte bodový odhad strednej hodnoty, rozptylu a smerodajnej odchýlky základného súboru.

Riešenie: Na výpočet bodových odhadov vypočítame výberový priemer, výberový rozptyl a výberovú smerodajnú odchýlku, ako je znázornené na obrázku 5.1.

Výberový priemer:	=AVERAGE(A1:A25)
Výberový rozptyl:	=VAR.S(A1:A25)
Výberová sm.odchýlka:	=STDEV.S(A1:A25)

Výberový priemer:	32,188
Výberový rozptyl:	0,0031
Výberová sm.odchýlka:	0,0561

Obr. 5.1 Výpočet a výsledky k príkladu 5.1 (zdroj: vlastné spracovanie)

Keďže počítame s údajmi z výberového súboru a chceme odhadnúť parametre základného súboru, tak pre výpočet výberového rozptylu a výberovej smerodajnej odchýlky použijeme v Exceli funkcie *VAR.S* resp. *STDEV.S*. Výberový priemer vypočítame zo vzorky pomocou funkcie *AVERAGE*. Výpočet na obrázku 5.1 je pre prípad, že vstupné dáta sa nachádzajú v bunkách A1 až A25. Smerodajnú odchýlku by sme mohli vypočítať aj ako druhú odmocninu z rozptylu, avšak výsledky na obrázku 5.1 sú zaokrúhlené na 4 desatinné miesta, preto by tento spôsob nemusel byť úplne presný. Na záver zhrnieme výsledky:

- Odhadom strednej hodnoty μ základného súboru (celej produkcie výrobnej linky) je výberový priemer $\overline{x} = 32{,}188\ \mathrm{mm}$.
- Odhadom rozptylu σ^2 základného súboru je výberový rozptyl $s^2 = 0{,}0031$.
- Odhadom smerodajnej odchýlky σ základného súboru je výberová smerodajná odchýlka $s = 0{,}0561$.

5.2 Intervalový odhad

Nevýhody bodového odhadu sa môžeme pokúsiť vylepšiť pomocou tzv. intervalového odhadu. V prípade **intervalového odhadu** nevypočítame iba jeden konkrétny bod, ale celý **interval** možných odhadov parametra. Tento interval s určitou pravdepodobnosťou obsahuje skutočnú hodnotu odhadovaného

parametra základného súboru a nazýva sa **interval spoľahlivosti**. Spomínaná pravdepodobnosť sa označuje $1-\alpha$ a nazýva sa spoľahlivosť odhadu, prípadne koeficient spoľahlivosti. Hladina významnosti α vyjadruje pravdepodobnosť (riziko), že odhadovaný parameter základného súboru sa nenachádza v zvolenom intervale spoľahlivosti. Interval spoľahlivosti často nazývame $100.(1-\alpha)$ percentný. Teda napríklad pre $\alpha = 0{,}05$ sa jedná o 95 percentný interval spoľahlivosti.

Pri bližšom pohľade na **výhody** a **nevýhody intervalových odhadov** zistíme, že je to naopak ako pri bodových odhadoch. Intervalové odhady sú v porovnaní s bodovými **náročnejšie na výpočet**. Majú síce **menšiu presnosť** (nie sú vyjadrené jedným konkrétnym číslom), ale ich **spoľahlivosť je vyššia**. Bodový odhad neposkytuje informáciu o tom, ako „ďaleko“ sme od skutočného parametra základného súboru. Avšak pri intervalovom odhade vieme, že skutočná hodnota parametra sa v intervale nachádza s určitou pravdepodobnosťou.

Za určitých predpokladov je možné skonštruovať intervaly spoľahlivosti pre strednú hodnotu μ, pre rozptyl σ^2, pre smerodajnú odchýlku σ, pre podiel znaku π, prípadne pre rozdiel stredných hodnôt alebo podiel dvoch rozptylov v prípade porovnávania dvoch základných súborov. My sa budeme zaoberať tvorbou **intervalového odhadu strednej hodnoty normálneho rozdelenia**, pre ktorú je v Exceli funkcia *CONFIDENCE*. Pomocou nej vypočítame najväčšiu chybu, ktorú od priemeru odpočítame, resp. pripočítame pre získanie krajných bodov intervalu spoľahlivosti. Dôsledkom tzv. centrálnej limitnej vety je možné použiť rovnaký postup výpočtu, aj keď základný súbor nemá normálne rozdelenie, ak je vzorka dostatočne veľká. Najčastejšie sa v literatúre uvádza aspoň 50 pozorovaní (Rimarčík, 2007). Interval spoľahlivosti pre **strednú hodnotu** normálneho rozdelenia so **známym rozptylom** je možné určiť pomocou funkcie ***CONFIDENCE.NORM***. Interval spoľahlivosti pre **strednú hodnotu** normálneho rozdelenia s **neznámym rozptylom** je možné získať pomocou funkcie ***CONFIDENCE.T*** (Makovička, 2016). Funkcia *CONFIDENCE.NORM* realizuje výpočet na báze normálneho rozdelenia a funkcia *CONFIDENCE.T* realizuje výpočet na báze Studentovho rozdelenia.

Šírka intervalu spoľahlivosti závistí od viacerých faktorov. Čím vyššiu spoľahlivosť budeme požadovať, tým širší (teda menej presný) bude interval spoľahlivosti. Na druhej strane, zvyšovaním presnosti (menšia šírka intervalu) sa znižuje spoľahlivosť (zvyšuje sa riziko odhadu). Je preto dôležité nájsť určitý kompromis medzi hladinou významnosti α a požadovaným rozpätím intervalu spoľahlivosti. V praxi sa hladina významnosti volí najčastejšie $\alpha = 0{,}05$, prípadne $\alpha = 0{,}01$. Na šírku intervalu spoľahlivosti má vplyv aj **veľkosť vzorky** a **variabilita dát**. Čím viac údajov máme k dispozícii, tým je interval užší. A naopak, čím menej dát obsahuje výberový súbor, tým širší je interval spoľahlivosti. Šírka intervalu rastie aj s rastúcou variabilitou dát. A naopak, čím je menší rozptyl údajov, tým užší je aj interval.

Príklad 5.2 Vypočítajte 95 percentný interval spoľahlivosti strednej hodnoty základného súboru z príkladu 5.1.

Riešenie: Keďže rozptyl dát celej produkcie výrobnej linky nepoznáme, použijeme funkciu *CONFIDENCE.T* do ktorej zadáme výberovú smerodajnú odchýlku. Výberové charakteristiky vypočítame rovnako ako v príklade 5.1. Celý výpočet je znázornený na obrázku 5.2.

Výberový priemer:	=AVERAGE(A1:A25)
Výberová sm.odchýlka:	=STDEV.S(A1:A25)
Veľkosť vzorky:	=COUNT(A1:A25)
Hodnota confidence:	=CONFIDENCE.T(0,05;0,0561;25)
Dolná hranica intervalu:	=32,188-0,0232
Horná hranica intervalu:	=32,188+0,0232

Výberový priemer:	32,188
Výberová sm.odchýlka:	0,0561
Veľkosť vzorky:	25
Hodnota confidence:	0,0232
Dolná hranica intervalu:	32,1648
Horná hranica intervalu:	32,2112

Obr. 5.2 Výpočet a výsledky k príkladu 5.2 (zdroj: vlastné spracovanie)

V ľavej časti obrázka vidíme zápis do buniek, v pravej časti obrázka sú výsledky. V dialógovom okne funkcie *CONFIDENCE.T* je potrebné zadať 3 údaje. Prvý údaj je hladina významnosti. Keďže úlohou je určiť 95 percentný interval spoľahlivosti, tak hladina významnosti je 0,05. Ďalej je ešte potrebné zadať výberovú smerodajnú odchýlku (keďže parameter základného súboru nepoznáme) a počet prvkov vo vzorke. Pomocou funkcie *CONFIDENCE.T* získame číslo 0,0232, ktoré vyjadruje akoby vzdialenosť krajných bodov intervalu spoľahlivosti na číselnej osi od výberového priemeru $\overline{x} = 32{,}188\ mm$. Na výpočet dolnej hranice intervalu je potrebné hodnotu CONFIDENCE odpočítať od výberového priemeru. Na výpočet hornej hranice intervalu je potrebné hodnotu CONFIDENCE pripočítať k výberovému priemeru. Na základe toho 95 percentný interval spoľahlivosti pre strednú hodnotu rozmerov súčiastok celej produkcie výrobnej linky (základného súboru) je:

$$32{,}1648\ mm < \mu < 32{,}2112\ mm$$

Môžeme teda povedať, že ak by sme náhodne vybrali jednu súčiastku z celej produkcie, tak s pravdepodobnosťou 95 % bude jej rozmer patriť do vypočítaného intervalu.

Príklad 5.3 Potravinárska firma dodáva do obchodného reťazca zemiaky balené v približne rovnakých baleniach. Náhodne bolo vybratých 65 balení s výberovým priemerom hmotnosti $\overline{x} = 1008$ gramov. Poznáme rozptyl základného súboru $\sigma^2 = 2500\ \mathrm{g}^2$. Predpokladáme, že dáta pochádzajú zo súboru s normálnym

rozdelením. Vypočítajte 90 percentný interval spoľahlivosti pre strednú hodnotu základného súboru balení zemiakov.

Riešenie: Keďže poznáme rozptyl základného súboru, tak na výpočet použijeme v Exceli funkciu *CONFIDENCE.NORM*, ako je znázornené na obrázku 5.3.

Hodnota confidence:	=CONFIDENCE.NORM(0,1;50;65)
Dolná hranica intervalu:	=1008-10,2
Horná hranica intervalu:	=1008+10,2

Hodnota confidence:	10,20
Dolná hranica intervalu:	997,80
Horná hranica intervalu:	1018,20

Obr. 5.3 Výpočet a výsledky k príkladu 5.3 (zdroj: vlastné spracovanie)

V ľavej časti obrázka vidíme zápis do buniek, v pravej časti obrázka sú výsledky. V dialógovom okne funkcie *CONFIDENCE.NORM* je potrebné zadať 3 údaje. Keďže úlohou je určiť 90 percentný interval spoľahlivosti, tak hladina významnosti je 0,1. Vieme, že rozptyl základného súboru je 2500, takže smerodajná odchýlka je 50 (druhá odmocnina). Počet prvkov vo vzorke zadáme 65. Pomocou funkcie *CONFIDENCE.NORM* získame číslo 10,20 (zaokrúhlené na 2 desatinné miesta). Na výpočet dolnej hranice intervalu je potrebné hodnotu CONFIDENCE odpočítať od výberového priemeru $\overline{x} = 1008$. Na výpočet hornej hranice intervalu je potrebné hodnotu CONFIDENCE pripočítať k výberovému priemeru. Na základe toho 90 percentný interval spoľahlivosti pre strednú hodnotu hmotnosti balení zemiakov (základného súboru) je:

$$997{,}8\ \text{g} < \mu < 1018{,}2\ \text{g}$$

Môžeme teda povedať, že ak by sme náhodne vybrali jedno balenie zemiakov, tak s pravdepodobnosťou 90 % bude jeho hmotnosť patriť do vypočítaného intervalu.

Kontrolné otázky:

1. Čo je to bodový odhad?
2. Ako určíme bodový odhad strednej hodnoty základného súboru?
3. Čo je to intervalový odhad?
4. Aké výhody majú bodové odhady v porovnaní s intervalovými?
5. Aké výhody majú intervalové odhady v porovnaní s bodovými?
6. Čo vyjadruje hladina významnosti?
7. Ako súvisí pri intervalových odhadoch spoľahlivosť a presnosť?
8. Aký vplyv má veľkosť vzorky na šírku intervalu spoľahlivosti?
9. Aký vplyv má variabilita dát na šírku intervalu spoľahlivosti?

Úlohy na riešenie:

1. V call centre sa analyzovala dĺžka hovorov zákazníkov. Z náhodnej vzorky 27 hovorov bol nameraný ich čas v sekundách: 120; 180; 240; 90; 150; 300; 210; 135; 270; 105; 165; 330; 240; 120; 195; 360; 225; 150; 285; 115; 180; 315; 255; 140; 210; 345; 270. Predpokladá sa, že dĺžka hovorov je normálne rozdelená. Vypočítajte bodový odhad priemernej dĺžky hovoru, rozptylu a smerodajnej odchýlky dĺžky hovorov v call centre.

   [priemer: 211,11 s; rozptyl: 6289,1 s^2; smerodajná odchýlka: 79,3 s]

2. V skleníku sa pestujú paradajky. Z náhodnej vzorky 20 rastlín bola nameraná ich výška v centimetroch: 45,2; 47,1; 46,8; 48,3; 44,9; 49,5; 47,6; 46,2; 48,7; 45,5; 47,9; 49,2; 46,4; 48,1; 45,8; 47,3; 49,8; 46,5; 48,4; 47,0. Predpokladá sa, že výška rastlín je normálne rozdelená. Vypočítajte bodový odhad priemernej výšky, rozptylu a smerodajnej odchýlky výšky rastlín v skleníku.

   [priemer: 47,31 cm; rozptyl: 2,03 cm^2; smerodajná odchýlka: 1,43 cm]

3. Vypočítajte 95 percentný interval spoľahlivosti strednej hodnoty výšky paradajok základného súboru (celej produkcie) z úlohy 2.

   [46,64 cm; 47,98 cm]

4. Pivovar analyzoval obsah alkoholu v pive svojej produkcie. Náhodne bolo vybratých 80 fliaš s výberovým priemerom obsahu alkoholu $\overline{x} = 4{,}9\ \%$. Poznáme rozptyl základného súboru $\sigma^2 = 1$. Predpokladáme, že dáta pochádzajú zo súboru s normálnym rozdelením. Vypočítajte 95 percentný interval spoľahlivosti pre strednú hodnotu základného súboru obsahu alkoholu v pive.

   [4,68 %; 5,12 %]

6 TESTOVANIE ŠTATISTICKÝCH HYPOTÉZ

Kľúčové slová: výskumná otázka, hypotéza, testovanie hypotéz, hladina významnosti, p-hodnota, nulová hypotéza, alternatívna hypotéza, chyba prvého druhu, chyba druhého druhu, sila testu

Konkrétne úlohy inferenčnej štatistiky boli vidieť už v kapitole o bodových a intervalových odhadoch. Okrem odhadovania parametrov je druhou základnou úlohou induktívnej štatistiky testovanie štatistických hypotéz. A práve touto problematikou sa budeme v nasledujúcom texte intenzívne zaoberať. Ešte poznamenajme, že výskumníci si ešte pred tvorbou hypotéz často kladú tzv. **výskumné otázky**. Z nich sa potom často hypotézy odvodzujú. Kým hypotézy bývajú zvyčajne veľmi konkrétne, tak výskumné otázky sa formulujú oveľa širšie. Nie je vhodné ich formulovať tak, aby sa na ne odpovedalo iba „áno“ alebo „nie.

6.1 Základné pojmy

Hypotéza je určitá **domnienka**, **názor výskumníka**, ktorý môže byť pravdivý, ale aj nemusí. Môžeme tiež povedať, že hypotéza je **výskumný predpoklad**, čiže určitá predikcia výskumu. Nejde však o nejaké náhodné „tipovanie“ výsledkov, ale o odborný predpoklad. Formulovaniu konkrétnych hypotéz predchádza podrobné štúdium danej problematiky, skúmanie existujúcej vedeckej teórie, konzultácie s inými odborníkmi a výskumníkmi, aplikácia skúseností z praxe a tak ďalej. Hypotézami overujeme vzťahy medzi premennými (javmi), čo je typické pre kvantitatívny výskum.

Hypotézy sú zvyčajne najskôr formulované ako tzv. **vecné hypotézy**. V nich sa na vyjadrenie jednotlivých premenných používajú vecné termíny (Chráska, 2016). Príkladom vecnej hypotézy môže byť tvrdenie:

Dievčatá na strednej škole dosahujú v ekonomických predmetoch lepšie vedomosti ako chlapci.

Premenné, ktoré sa v hypotézach vyskytujú je však potrebné presne vyjadriť (zmerať). Vo vyššie uvedenom príklade musí byť jasné, čo sa myslí pod učebným výkonom a ako sa môže merať, preto je nutné premenné **operacionalizovať**. Operacionalizácia premenných je ich vyjadrenie takým spôsobom, aby boli merateľné (alebo aspoň kategorizovateľné). V tomto prípade môžeme napríklad vedomosti z ekonomických predmetov operacionalizovať (vyjadriť) ako výsledky didaktického testu.

Napriek tomu, ak by sme chceli so stopercentnou istotou rozhodnúť o pravdivosti nejakej hypotézy, museli by sme skúmať celý základný súbor. Ako sme však už

spomínali, to je v praxi zvyčajne neekonomické alebo dokonca technicky nemožné. V silách výskumníkov je väčšinou skúmanie iba malej časti základného súboru (vzorky). Proces overovania správnosti hypotéz na základe výsledkov získaných náhodným výberom, sa nazýva **testovanie hypotéz**. Aby sme mohli hypotézu overovať (testovať) s použitím štatistických metód, vyjadruje sa v tvare tzv. **štatistickej hypotézy** (používajú sa tu štatistické pojmy). Spomínaná vecná hypotéza by ako štatistická hypotéza mohla byť formulovaná napríklad takto:

Dievčatá na strednej škole dosahujú vyšší priemerný počet bodov v didaktickom teste z ekonomických predmetov ako chlapci.

Príklad 6.1 Uveďte príklad správne aj nesprávne položenej výskumnej otázky na tému výskumu voľnočasových aktivít počas kongresovej turistiky.

Riešenie: Možností na zostavenie výskumnej otázky je veľmi veľa. Najskôr vytvoríme jednu **nie** celkom **vhodnú** formuláciu:

Je v danej lokalite dostatok možností na voľnočasové aktivity pre účastníkov kongresovej turistiky?

Na takto položenú výskumnú otázku sa dá odpovedať „áno“ resp. „nie“, čo je veľmi chudobné a stručné zistenie. Formulácia by preto mala byť komplexnejšia, **jedna z vhodných možností je** napríklad takáto:

Aké druhy voľnočasových aktivít prevládajú u návštevníkov počas kongresovej turistiky v danej lokalite?

Takto položená výskumná otázka je širšia a poskytuje výskumníkovi dostatočný základ na realizáciu výskumu.

6.2 Tvorba hypotéz a postup pri ich testovaní

Pri testovaní hypotéz kladieme oproti sebe dve hypotézy, ktoré si navzájom odporujú a zároveň pokrývajú všetky možnosti. Hypotéza, ktorej platnosť overujeme sa nazýva testovaná alebo nulová hypotéza. Oproti nej kladieme tzv. alternatívnu hypotézu.

Nulová hypotéza sa zvyčajne označuje H_0 a reprezentuje buď nejaký aktuálny stav alebo nezávislosť premenných. Často je tiež možné vychádzať z toho, že nulová hypotéza zvyčajne nepredpokladá žiadny štatisticky významný rozdiel. Najčastejšie je formulovaná v tvare rovnosti, môže vyjadrovať napríklad rovnosť neznámeho parametra s konštantou alebo rovnosť dvoch neznámych parametrov.

Alternatívna hypotéza alebo tiež **výskumná hypotéza** *(research hypothesis)* sa zvyčajne označuje H_1 a reprezentuje nejaký výskumný predpoklad alebo niečo, pravdivosť čoho chceme overiť. Výskumná hypotéza sa dokazuje vždy nepriamo takým spôsobom, že dokážeme nesprávnosť nulovej hypotézy. Po zamietnutí nulovej hypotézy ostáva už iba jediná možnosť, že správna je alternatívna hypotéza. Alternatívna hypotéza môže byť vyjadrená ako obojstranná alebo jednostranná.

Obojstranná alternatívna hypotéza je formulovaná v tvare nerovnosti (≠). Napríklad nerovnosť neznámeho parametra s konštantou alebo nerovnosť dvoch neznámych parametrov.

Jednostranná alternatívna hypotéza je formulovaná ako jednostranná nerovnosť (ľavostranná < alebo pravostranná >). Napríklad neznámy parameter je menší (väčší) ako konštanta alebo jeden parameter je menší (väčší) ako druhý parameter. V skutočnosti by sa mala aj nulová hypotéza formulovať v tvare nerovnosti tak, aby obidve hypotézy boli vzájomne úplnou negáciou. V praxi sa však najčastejšie stretávame so zápisom, že k jednostrannej alternatívnej hypotéze sa formuluje nulová hypotéza v tvare rovnosti, hoci sa jedná len o „čiastočnú negáciu“.

Nulovú a alternatívnu hypotézu je vhodné formulovať nielen formálnym zápisom, ale aj slovnou interpretáciou, ktorá sa týka priamo skúmaného problému. Záverečná interpretácia výsledku testu hypotézy je potom oveľa jasnejšia (Švábová a kol., 2022).

Príklad 6.2 Porovnávali sme mzdy zamestnancov v dvoch veľkých závodoch. Náhodne sme vybrali vzorku zamestnancov z každého závodu a zistili sme ich mzdu. Cieľom je zistiť, či je stredná hodnota mzdy zamestnancov v obidvoch závodoch rovnaká alebo sa štatisticky významne odlišuje. Ako by mohla byť formulovaná nulová a alternatívna hypotéza?

Riešenie: Formálny zápis nulovej a alternatívnej hypotézy by mohol byť nasledovný:

$H_0: \mu_1 = \mu_2$ (mzdy zamestnancov v obidvoch závodoch majú rovnaký rozptyl).

$H_1: \mu_1 \neq \mu_2$ (mzdy robotníkov v skúmaných podnikoch majú rôzny rozptyl).

Slovne môžu byť hypotézy vyjadrené nasledovne:

H_0: Stredná hodnota mzdy zamestnancov v oboch závodoch je rovnaká.

H_1: Stredná hodnota mzdy zamestnancov v oboch závodoch sa signifikantne líši.

Testovaním hypotéz chceme väčšinou (nie nevyhnutne) zistiť, či napríklad medzi skúmanými výberovými charakteristikami je **štatisticky významný** (**signifikantný**)

rozdiel. Ak je výsledok výskumu štatisticky významný (signifikantný), s veľkou pravdepodobnosťou sa jedná o systematický jav, ktorý **môžeme zovšeobecniť na celý základný súbor**. V opačnom prípade sa jedná o nevýznamný, zanedbateľný, či náhodný rozdiel. Závery testovania hypotéz majú vždy iba pravdepodobnostný charakter, pri ktorom sa môžu vyskytnúť chyby.

Chyba prvého druhu nastáva vtedy, keď je nulová hypotéza správna, ale my ju napriek tomu zamietneme. Pravdepodobnosť vzniku chyby prvého druhu sa nazýva **hladina významnosti** *(significance level)* testu a označuje sa α.

Chyba druhého druhu nastáva vtedy, keď je nulová hypotéza nesprávna, ale my ju napriek tomu nezamietneme. Pravdepodobnosť vzniku chyby druhého druhu sa označuje β. Pravdepodobnosť $1-\beta$ sa nazýva **sila testu** *(power)* a vyjadruje pravdepodobnosť, že zamietneme nulovú hypotézu, keď nie je správna. Vzťahy medzi chybami prvého a druhého druhu a pravdepodobnosťami sú zhrnuté v tabuľke 6.1.

Tab. 6.1. Vzťahy medzi chybami I. a II. druhu a ich pravdepodobnosťami pri testovaní hypotéz, (zdroj: vlastné spracovanie)

Testovanie hypotéz		**Skutočnosť**	
		H_0 je pravdivá	**H_0 je nepravdivá**
Rozhodnutie	**Nezamietame H_0** (nevýznamný výsledok)	Správne rozhodnutie s pravdepodobnosťou $1-\alpha$	Chyba II. druhu s pravdepodobnosťou β
	Zamietame H_0 (signifikantný výsledok)	Chyba I. druhu s pravdepodobnosťou α	Správne rozhodnutie s pravdepodobnosťou $1-\beta$

Veľkosť chýb spolu súvisí. Ak pri tom istom rozsahu vzorky znížime chybu α, zvýši sa tým chyba β a naopak. Preto sa α volí v nejakej „rozumnej miere“, napríklad 0,05. Zmenšenie obidvoch chýb súčasne sa dá dosiahnuť iba zväčšením rozsahu výberu.

Kedysi bolo pri testovaní hypotéz potrebné ovládať pojmy ako testovacie kritérium, či kritická hodnota. Neodmysliteľnou pomôckou boli aj štatistické tabuľky, pričom toto všetko bolo typické predovšetkým pre „ručné“ počítanie. Celý proces by bolo možné výrazne zjednodušiť, keby sme poznali tzv. **p-hodnotu** *(p-value)*. Pomocou nej je veľmi jednoduché rozhodnúť o výsledku testovania. Nepotrebujeme k tomu ďalšie pomocné výpočty ani tabuľky s kritickými hodnotami, p-hodnotu stačí

porovnať s hladinou významnosti α. Rozhodujeme sa tak, že ak je p-hodnota menšia ako hladina významnosti α, potom nulovú hypotézu H_0 zamietame. V opačnom prípade nulovú hypotézu nezamietame.

Karl Pearson ako prvý zadefinoval p-hodnotu v roku 1900 (Laplace ju počítal už dokonca v 70-tych rokoch 18. storočia), ale pre veľmi zložitý výpočet bol problém s praktickým použitím. To sa však zmenilo príchodom počítačov s príslušným softvérom a dnes za nás túto prácu vykoná stroj. V súčasnosti je to najpoužívanejší spôsob. Písmeno „p“ v slove p-hodnota pochádza z anglického *probability*, čo znamená pravdepodobnosť. P-hodnota testu je najmenšia hladina významnosti, pri ktorej by sme už nulovú hypotézu zamietli.

Celý **postup testovania hypotéz** zhrnieme prehľadne do nasledujúcich krokov:

1. Formulácia nulovej hypotézy H_0 a alternatívnej hypotézy H_1
2. Zvolenie hladiny významnosti α.
3. Výber vhodného štatistického testu.
4. Výpočet p-hodnoty testu.
5. Rozhodnutie o výsledku testu na základe porovnania p-hodnoty a hladiny významnosti α.

Podrobne bude celý postup vysvetlený neskôr na konkrétnych príkladoch. Na tomto mieste ešte spomenieme, že pri rozhodnutí o výsledku máme dve možnosti. Buď zamietneme nulovú hypotézu a prikláňame sa k platnosti alternatívnej hypotézy alebo nulovú hypotézu nezamietneme. V oboch prípadoch sa zvykne dodávať ešte slovné spojenie „na hladine významnosti (napríklad) 0,05“.

Ak **nulovú hypotézu nezamietneme**, tak to ešte **neznamená**, že sme **dokázali jej platnosť**. V praxi sa však napriek tomu často stretávame so situáciou, keď sa nulová hypotéza v prípade jej nezamietnutia považuje za správnu, čo nie je korektné. Znamená to len toľko, že na základe náhodne vybratej vzorky nemáme dostatok informácií na to, aby sme nulovú hypotézu mohli zamietnuť. Pekne to vysvetľuje Terek (2019), keď situáciu pripodobňuje k rozhodovaniu súdu v prípade osoby obvinenej z trestného činu. Vychádza sa z predpokladu, že obvinený je nevinný (H_0). Skúmajú sa dôkazy, ktoré sú v rozpore s týmto predpokladom a svedčia o vine (H_1). Ak sú dôkazy dostatočne silné a závažné, obvinený je odsúdený (zamietame H_0 a prijímame H_1). A naopak, ak sa nenájde dosť dôkazov a vina nebola preukázaná, obvinený je oslobodený (nemožno zamietnuť H_0). To ale neznamená, že bola dokázaná nevina obvineného. Bol oslobodený pre nedostatok dôkazov o vine.

6.3 Časté chyby pri aplikácii v praxi

V bakalárskych a diplomových prácach sa dá (žiaľ) často stretnúť s tým, že autor sa obmedzil iba na výpočet priemeru alebo percent (relatívnych početností) a na základe toho rozhodne o potvrdení alebo zamietnutí hypotézy a následnom zovšeobecnení na celý základný súbor bez použitia štatistických testov. Takýto postup nie je správny. Je to príliš silný záver, na ktorý nie sú štatistické dôkazy, keďže neboli použité dostatočne vhodné metódy. V tom prípade by bolo lepšie, keby vôbec neboli formulované hypotézy, ale iba výskumné otázky a celá výskumná časť by mala iba deskriptívny charakter.

Ďalšou častou praxou je použitie induktívnej štatistiky na nenáhodných výberoch alebo keď realizujeme cenzus (zvyčajne je to možné na malých základných súboroch). Induktívna štatistika je založená na teórii pravdepodobnosti a v takýchto prípadoch sa nepoužíva. Ak by sme **pri nenáhodných výberoch** aj formulovali výskumnú hypotézu, tak **výsledok sa bude vzťahovať iba na danú vzorku** a **nemožno ho zovšeobecňovať**. Ak realizujeme cenzus, tak induktívna štatistika nie je potrebná, lebo máme k dispozícii dáta za celý základný súbor (nepotrebujeme nič odhadovať, máme k dispozícii presné parametre základného súboru). V oboch prípadoch je možné formulovať jednoduchú výskumnú hypotézu o opisných charakteristikách z nenáhodného výberového súboru, resp. parametroch základného súboru (pri cenzuse). Vyhodnotiť ich môžeme aj na základe vypočítaných charakteristík (napríklad priemerov) alebo pomocou relatívnych početností (percent).

V prácach študentov sa dá tiež neraz stretnúť s nesprávnym záverom, že „hypotéza bola čiastočne potvrdená“. To však takisto nie je správne a študenti často ani sami nevedia, čo si pod tým treba predstaviť, resp. ako by sa to malo interpretovať. **Hypotéza** môže byť potvrdená alebo zamietnutá, ale **nemôže byť potvrdená iba čiastočne**. V postupe testovania hypotéz sú dané konkrétne kroky, pričom na záver máme dve možnosti a výskumník sa jednoznačne rozhodne iba pre jeden z možných záverov.

Kontrolné otázky:

1. Ako by mala byť správne formulovaná výskumná otázka?
2. Čo je to hypotéza?
3. Nulová a alternatívna hypotéza sa zhodujú alebo sú v protiklade?
4. Prečo sa pred príchodom počítačov nepoužíval postup testovania hypotéz s využitím p-hodnoty?

5. Ako sa na základe vypočítanej p-hodnoty dá rozhodnúť o výsledku?
6. Ako by ste zhrnuli postup testovania hypotéz?
7. Aké druhy chýb pri testovaní hypotéz poznáte?
8. Ako spolu tieto chyby súvisia?
9. Čo znamená, ak je výsledok signifikantný?
10. Používa sa inferenčná štatistika v prípade, ak realizujeme cenzus? A prečo?
11. Ak nezamietneme nulovú hypotézu, znamená to, že sme dokázali jej platnosť?
12. Aký je rozdiel medzi obojstrannou a jednostrannou hypotézou?
13. Čo vyjadruje sila testu?

Úlohy na riešenie:

1. Porozmýšľajte nad témou svojej záverečnej práce a navrhnite niekoľko vhodných výskumných otázok.

2. Porozmýšľajte nad témou svojej záverečnej práce a sformulujte ku konkrétnemu problému nulovú a alternatívnu hypotézu.

3. Vyberte si aktuálnu tému, o ktorej sa diskutuje v médiách. Sformulujte hypotézu o tejto téme a navrhnite, ako by sa dala otestovať.

4. Navrhnite postup, ako by ste zistili, či má nová učebná pomôcka vplyv na študijné výsledky konkrétneho predmetu na strednej škole. Sformulujte nulovú a výskumnú hypotézu.

7 VYŠETROVANIE NORMALITY ROZDELENIA

Kľúčové slová: strmosť, šikmosť, normálne rozdelenie pravdepodobnosti, Shapiro-Wilkov test, D´Agostinov test, Gaussova krivka, histogram

V predchádzajúcich kapitolách už bolo vysvetlené, čo je to normálne rozdelenie pravdepodobnosti. Pri výpočtoch bodových a intervalových odhadov sme dokonca predpokladali, že dáta pochádzajú zo súboru s normálnym rozdelením. Predpoklad o normalite rozdelenia dát je potrebný aj pri niektorých testoch, ktorými sa budeme neskôr zaoberať. Zatiaľ však nebolo ukázané, ako overiť či dáta pochádzajú zo súboru s normálnym rozdelením. A práve to je témou tejto kapitoly.

7.1 Histogram a základné charakteristiky

V rámci opisu vlastností normálneho rozdelenia už bolo spomenuté, že **modus**, **medián** a **aritmetický priemer majú tú istú hodnotu** (sú totožné) a že **Gaussova krivka má zvonovitý tvar**. Tieto vlastnosti môžu byť využité pri overovaní normality údajov z náhodného výberu.

Zo vzorky je možné vypočítať bodové odhady spomínaných parametrov základného súboru. V praxi sa síce asi nestretneme s prípadom, že by vyšli všetky tri úplne totožné, ale ak výber pochádza z normálneho rozdelenia, tak rozdiely medzi nimi by nemali byť veľké. Na základe tohto kritéria sa normalita nedá posúdiť exaktne, ale poskytne nám to aspoň približnú predstavu o dátach vo vzorke.

Vlastnosť normálneho rozdelenia, že Gaussova krivka má zvonovitý tvar, môžeme využiť ako určitú formu grafického testu. Podobne ako pri vyššie spomínaných parametroch ani toto nie je exaktný spôsob určenia normality. Poskytne nám to len približný odhad rozloženia dát. Ale čím viac informácií máme, tým lepšiu predstavu o základnom súbore si vieme vytvoriť. Grafický test môžeme zrealizovať napríklad pomocou histogramu a zistiť, či údaje z výberového súboru majú rozdelenie aspoň približne v tvare zvonu.

Excel vo verzii 2016 má v ponuke grafov aj histogram, ako už bolo ukázané v kapitole o deskriptívnej štatistike. Z hľadiska štatistiky obsahuje aj veľmi užitočné možnosti, ako napríklad nastavenie šírky priehradky (stĺpca) alebo nastavenie počtu priehradiek. Na tomto mieste je však vhodnejšie využiť doplnok Real Statistics v záložke **Desc**, položka **Histogram with Normal Curve Overlay**. Na rozdiel od bežného histogramu je tu možné nastaviť, aby graf prekrývala aj krivka normálneho rozdelenia. Vďaka tomu je celá vizualizácia názornejšia a lepšie sa posudzuje miera odlišnosti daného histogramu od Gaussovej krivky.

7.2 Strmosť a šikmosť

V kapitole o deskriptívnej štatistike boli z číselných charakteristík spomínané charakteristiky polohy a charakteristiky variability, ktoré sú v praxi najpoužívanejšie. Existuje ešte ďalší druh charakteristík, ktoré sme doteraz nespomenuli, a síce **charakteristiky tvaru**. Tie sa pri opise údajov vo výskumných správach používajú málokedy, ale dajú sa využiť pri posudzovaní, či náhodný výber pochádza zo základného súboru s normálnym rozdelením. Najznámejšie charakteristiky tvaru sú strmosť a šikmosť.

Strmosť alebo tiež špicatosť *(angl. kurtosis)* je charakteristika tvaru a poskytuje informáciu o tom, či je distribúcia údajov strmá alebo plochá, čiže ako sú hodnoty sústredené (nahustené) okolo priemeru. V literatúre ju môžeme nájsť aj pod označením špicatosť. Koeficient strmosti môžeme v Exceli vypočítať pomocou funkcie *KURT* alebo je možné využiť aj doplnok Real Statistics v záložke **Desc**, prvá položka **Descriptive Statistics and Normality**. Kladná hodnota koeficientu znamená, že väčšina údajov sa nachádza neďaleko priemeru, rozdelenie strmé (špicaté). Záporná hodnota znamená viac údajov, ktoré sú ďalej od priemeru, rozdelenie je ploché. Nulovú hodnotu má normálne rozdelenie. Práve preto je možné koeficient strmosti využiť pri overovaní, či vzorka pochádza zo súboru s normálnym rozdelením. Ak áno, hodnota koeficientu strmosti by mala byť blízko nule. Ako akceptovateľný odklon od normálneho rozdelenia (od nuly) uvádzajú niektorí autori hodnotu medzi -1 a 1 (napríklad Dancey – Reidy, 2017). U iných sa môžeme stretnúť s toleranciou od -2 po 2 (napríklad George – Mallery, 2010). Niektorí autori uvádzajú ako akceptovateľné čísla dokonca až s hranicou od -7 po 7 (Byrne, 2016). My sa prikloníme k intervalu od -2 po 2 a tieto hodnoty budeme považovať za akceptovateľné.

Šikmosť *(angl. skewness)* je charakteristika tvaru a poskytuje informáciu o tom, či je distribúcia údajov asymetricky rozložená. Koeficient šikmosti môžeme v Exceli vypočítať pomocou funkcie *SKEW* alebo je možné využiť aj doplnok Real Statistics v záložke **Desc**, prvá položka **Descriptive Statistics and Normality**. Kladná hodnota koeficientu znamená tzv. pravostrannú šikmosť, záporná hodnota ľavostrannú šikmosť a nulová hodnota symetrické rozdelenie údajov okolo priemeru. Keďže normálne rozdelenie je symetrické, hodnota koeficientu šikmosti je u normálneho rozdelenia rovná nule. Túto vlastnosť môžeme využiť pri overovaní, či vzorka pochádza zo súboru s normálnym rozdelením. Podobne ako pri strmosti, aj tu by hodnota mala byť čo najbližšia nule. Opäť sa v literatúre môžeme stretnúť s rôznymi hodnotami, ktoré autori považujú za akceptovateľný odklon od normálneho rozdelenia. Asi najčastejšie sa však stretneme s hodnotami od -2 po 2, ktoré budeme aj my považovať za akceptovateľné.

7.3 Testy normality

Existuje veľa rôznych testov, pomocou ktorých sa dá skúmať normalita základného súboru. My sa budeme zaoberať dvoma z nich. **Shapiro-Wilkov test** je vhodnejší pre menšie veľkosti vzoriek, napríklad pre náhodný výber s rozsahom $n \in \langle 7, 30 \rangle$. **D'Agostinov test** je možné použiť pre väčšie vzorky, napríklad v prípade náhodného výberu s rozsahom $n \in \langle 31, 100 \rangle$ (Jurečková – Molnárová, 2005; Markechová a kol., 2011). V rôznej literatúre dá stretnúť aj s iným odporúčaním na veľkosť vzorky pre dané testy.

Pri testoch normality vždy nulová hypotéza tvrdí, že náhodná vzorka pochádza zo základného súboru s normálnym rozdelením. Teda nulová a alternatívna hypotéza môžu byť formulované napríklad takto:

H_0: Náhodný výber pochádza zo súboru s normálnym rozdelením.

H_1: Náhodný výber nepochádza zo súboru s normálnym rozdelením.

V prípade, že vypočítaná p-hodnota testu je menšia ako hladina významnosti (napríklad 0,05), tak zamietame nulovú hypotézu a prijímame alternatívnu hypotézu. Znamená to, že údaje vo vzorke nepochádzajú zo súboru s normálnym rozdelením. Ak by vyšla p-hodnota väčšia ako hladina významnosti a nulovú hypotézu by sme nemohli zamietnuť, **nedokazuje to** automaticky **normálne rozdelenie dát**. Iba nemáme dostatok dôkazov na zamietnutie nulovej hypotézy a tým pádom na odmietnutie normality.

Na uvedené testy normality (Shapiro-Wilkov a D'Agostinov test) **má výrazný vplyv veľkosť vzorky**. So zvyšujúcim sa počtom dát rastie aj pravdepodobnosť zamietnutia nulovej hypotézy. V krajnej situácii sa teda môže stať, že pri malej vzorke sa kvôli nízkej sile testu nulová hypotéza nezamietne ani v prípade, keď sa dáta výrazne odlišujú od normálneho rozdelenia. A naopak, pri veľkej vzorke sa kvôli vysokej sile testu môže normalita zamietnuť aj v prípade, keď sú údaje veľmi blízke normálnemu rozdeleniu.

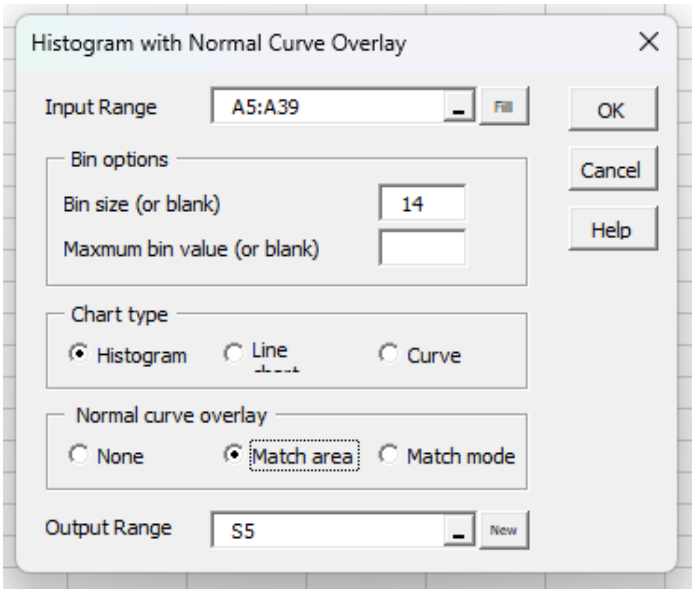
Testy normality preto nie je vhodné používať ako jediné kritérium napríklad pri rozhodovaní, či použiť typ testov, ktoré predpoklad normality vyžadujú (napríklad parametrické testy, ktorými sa neskôr budeme zaoberať). Je potrebné sa rozhodnúť na základe lepšieho poznania základného súboru z hľadiska viacerých kritérií. Testy normality sa priamo v Exceli nenachádzajú, ale na ich výpočet je možné využiť doplnok Real Statistics v záložke **Desc**, prvá položka **Descriptive Statistics and Normality**. Excel vždy vypočíta p-hodnotu pre obidva testy. Vybrať si potom treba na základe počtu údajov vo vzorke.

7.4 Overovanie normality

V tejto kapitole boli ukázané viaceré spôsoby overovania normality dát, od grafického posúdenia, cez charakteristiky polohy (modus, medián, aritmetický priemer), charakteristiky tvaru (strmosť a šikmosť) až po testy normality. Pri overovaní, či náhodný výber pochádza zo základného súboru s normálnym rozdelením by nebolo vhodné spoľahnúť sa iba na jednu z uvedených možností. Na údaje vo výberovom súbore sa vždy treba pozerať čo najkomplexnejšie, aby sme získali dobrú predstavu o základnom súbore, z ktorého pochádzajú. Preto je vhodné overovať normalitu z viacerých pohľadov. Podrobnejšie to bude teraz ukázané na konkrétnych príkladoch.

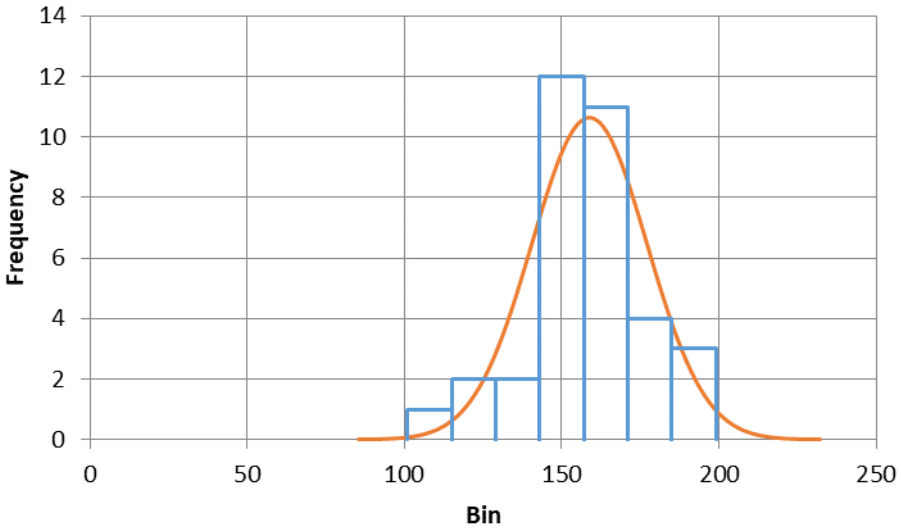
Príklad 7.1 Pri realizácii výskumu sme náhodným výberom zozbierali nasledovné dáta: 152; 155; 168; 145; 121; 115; 184; 191; 150; 161; 134; 188; 155; 165; 162; 163; 174; 154; 152; 155; 154; 199; 122; 164; 166; 153; 159; 167; 182; 174; 143; 156; 157; 158; 160. Overte, či tieto údaje pochádzajú zo súboru s normálnym rozdelením.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpca „A“. Najskôr urobíme **grafický test**, aby sme získali aspoň približnú vizuálnu predstavu o rozložení údajov. Klikneme na bunku, kam chceme umiestniť výstup. Na vytvorenie **histogramu s prekrytím krivkou normálneho rozdelenia** použijeme doplnok Real Statistics v záložke Desc, položka Histogram with Normal Curve Overlay. V dialógovom okne je potrebné vyplniť niekoľko položiek (obr. 7.1)



Obr. 7.1 Dialógové okno pre tvorbu histogramu s prekrytím krivkou normálneho rozdelenia (zdroj: vlastné spracovanie)

Do kolonky „Input Range“ vložíme oblasť buniek so vstupnými údajmi. Do položky „Bin size“ zadáme požadovanú šírku stĺpca histogramu, v tomto prípade napríklad 14. V položke „Maximum bin value“ môžeme nastaviť maximálnu hodnotu posledného stĺpca, ale túto položku nie je nutné vyplniť a môžeme ju nechať prázdnu. V „Chart type“ ponecháme predvolenú položku histogram. Nakoniec treba ešte nastaviť „Normal curve overlay“. Možnosť „Match area“ vytvorí Gaussovu krivku tak, že plocha pod krivkou je rovnaká, ako plocha histogramu. Možnosť „Match mode“ vytvorí Gaussovu krivku s rovnakou výškou ako histogram, ale nebudú zaberať rovnaké plochy. Vo všeobecnosti by sme mali vyberať možnosť „Match area“, takže ju potvrdíme aj teraz.



Obr. 7.2 Histogram s prekrytím krivkou normálneho rozdelenia (zdroj: vlastné spracovanie)

Po kliknutí na OK sa zobrazí histogram aj s Gaussovou krivkou (obr. 7.2). Na základe vizuálneho hodnotenia môžeme konštatovať, že histogram má približne tvar normálneho rozdelenia, resp. sa od neho výrazne nelíši. Tento hrubý odhad sa pokúsime podložiť ďalšími výpočtami.

Vypočítame koeficient **šikmosti**, koeficient **strmosti** a vzhľadom na veľkosť vzorky (35 údajov) zrealizujeme **D'Agostinov** test. Všetky tieto požiadavky zrealizuje doplnok Real Statistics v záložke Desc, prvá položka Descriptive Statistics and Normality. V dialógovom okne do položky „Input Range“ vložíme vstupné údaje. Z ostatných položiek necháme zaškrtnuté iba „Descriptive statistics“ a „Shapiro-Wilk“. Po kliknutí na OK sa zobrazia požadované výsledky (obr. 7.3). Čísla, ktoré nás najviac zaujímajú, sú zvýraznené farebne.

Hodnota **aritmetického priemeru** je 158,8. **Medián** má hodnotu 158 a **modus** 155. Rozdiely medzi týmito charakteristikami nie sú veľké, čo je ďalšia indícia o tom, že dáta pochádzajú zo súboru s normálnym rozdelením.

Descriptive Statistics	
	Group 1
Mean	158,8
Standard E	3,109433
Median	158
Mode	155
Standard D	18,39565
Sample Va	338,4
Kurtosis	0,78452
Skewness	-0,25384
Range	84
Maximum	199
Minimum	115

Shapiro-Wilk Test	
	Group 1
W-stat	0,953043
p-value	0,14052
alpha	0,05
normal	yes
d'Agostino-Pearson	
DA-stat	1,689023
p-value	0,429767
alpha	0,05
normal	yes

Obr. 7.3 Výpočty k príkladu 7.1 (zdroj: vlastné spracovanie)

Koeficient šikmosti bol vyčíslený na -0,25, čo je hodnota blízka nule a vzhľadom na normalitu je to akceptovateľné číslo, keďže je z intervalu od -2 po 2.

Koeficient strmosti bol vyčíslený na 0,78, čo je tiež akceptovateľné číslo, keďže je z intervalu od -2 po 2.

Pre **D'Agostinov** test normality sme formulovali nulovú a alternatívnu hypotézu:

H_0: Náhodný výber pochádza zo súboru s normálnym rozdelením.

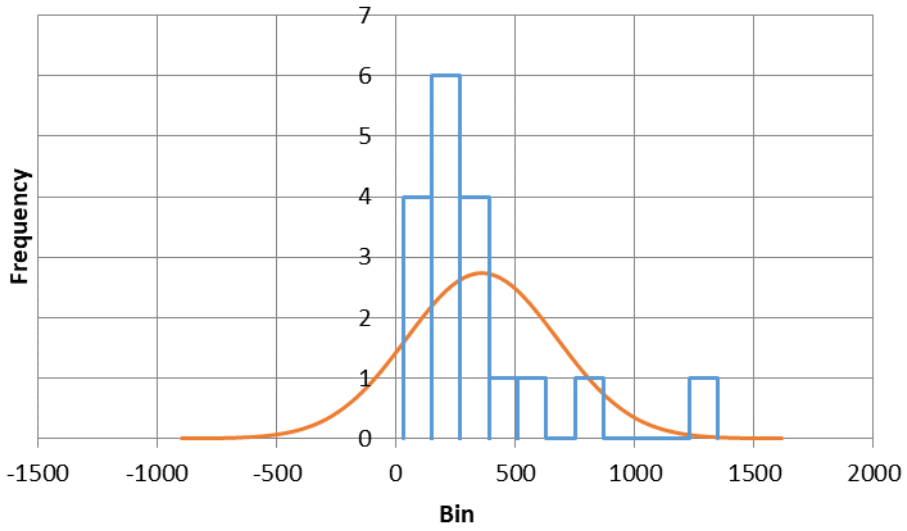
H_1: Náhodný výber nepochádza zo súboru s normálnym rozdelením.

Vypočítaná p-hodnota 0,42 je väčšia ako hladina významnosti 0,05, takže nulovú hypotézu nezamietame. Mimochodom, na obrázku 7.3 vidno, že aj Shapiro-Wilkov test vedie k rovnakému záveru, ale vzhľadom na počet údajov vo vzorke sme uprednostnili D'Agostinov test.

Záverom môžeme povedať, že všetky vykonané analýzy ukazujú, že údaje **môžeme považovať za normálne rozdelené**.

Príklad 7.2 Pri realizácii výskumu sme náhodným výberom zozbierali nasledovné dáta: 250; 350; 169; 355; 240; 501; 1350; 111; 850; 244; 300; 185; 142; 244; 355; 622; 101; 105. Overte, či tieto údaje pochádzajú zo súboru s normálnym rozdelením.

Riešenie: Uvedené dáta prepíšeme do Excelu napríklad do stĺpca „A“. Najskôr urobíme **grafický test**, aby sme získali aspoň približnú vizuálnu predstavu o rozložení údajov. Podobne ako v príklade 7.1 použijeme doplnok Real Statistics v záložke Desc, položka Histogram with Normal Curve Overlay. V dialógovom okne vložíme vstupné údaje, šírku stĺpca „Bin size“ nastavíme napríklad na hodnotu 120 a zvolíme možnosť „Match area“ pre vytvorenie krivky normálneho rozdelenia tak, aby plocha pod krivkou bola rovnaká ako plocha histogramu. Po kliknutí na OK sa zobrazí histogram aj s Gaussovou krivkou (obr. 7.4).



Obr. 7.4 Histogram s prekrytím krivkou normálneho rozdelenia (zdroj: vlastné spracovanie)

Na základe vizuálneho hodnotenia môžeme konštatovať, že histogram má výrazne odlišný tvar ako krivka normálneho rozdelenia. Tento hrubý odhad sa pokúsime podložiť ďalšími výpočtami.

Vypočítame koeficient **šikmosti**, koeficient **strmosti** a vzhľadom na veľkosť vzorky (18 údajov) zrealizujeme **Shapiro-Wilkov** test. Podobne ako v predchádzajúcom príklade použijeme doplnok Real Statistics v záložke Desc, prvá položka Descriptive Statistics and Normality. V dialógovom okne do položky „Input Range“ vložíme vstupné údaje. Z ostatných položiek necháme zaškrtnuté iba „Descriptive statistics“ a „Shapiro-Wilk“. Po kliknutí na OK sa zobrazia požadované výsledky (obr. 7.5). Čísla, ktoré nás najviac zaujímajú, sú zvýraznené farebne.

Hodnota **aritmetického priemeru** je 359,66. **Medián** má hodnotu 247 a **modus** 355. Predovšetkým medián sa dosť výrazne odlišuje od zvyšných dvoch charakteristík, čo je ďalšia indícia o tom, že dáta nepochádzajú zo súboru s normálnym rozdelením.

Descriptive Statistics			Shapiro-Wilk Test	
	Group 1			*Group 1*
Mean	359,6667		W-stat	0,746008
Standard E	73,97716		p-value	0,000288
Median	247		alpha	0,05
Mode	355		normal	no
Standard D	313,8585			
Sample Va	98507,18		ďAgostino-Pearson	
Kurtosis	5,393438			
Skewness	2,208781		DA-stat	20,41853
Range	1249		p-value	3,68E-05
Maximum	1350		alpha	0,05
Minimum	101		normal	no
Sum	6474			

Obr. 7.5 Výpočty k príkladu 7.2 (zdroj: vlastné spracovanie)

Koeficient šikmosti bol vyčíslený na 2,2 a **koeficient strmosti** 5,39. Ani jedna z týchto hodnôt nie je z intervalu od -2 po 2, čiže to nie sú akceptovateľné čísla pre tolerovaný odklon od normálneho rozdelenia. Pre **Shapiro-Wilkov** test normality sme formulovali nulovú a alternatívnu hypotézu:

H_0: Náhodný výber pochádza zo súboru s normálnym rozdelením.

H_1: Náhodný výber nepochádza zo súboru s normálnym rozdelením.

Vypočítaná p-hodnota 0,00 je menšia ako hladina významnosti 0,05, takže nulovú hypotézu zamietame a prijímame alternatívnu hypotézu. Záverom môžeme povedať, že všetky vykonané analýzy ukazujú, že údaje **nemôžeme považovať za normálne rozdelené**.

Kontrolné otázky:

1. Za akým účelom môže byť potrebné zisťovať normalitu?
2. O čom vypovedá strmosť a ako ju vypočítame v Exceli?
3. O čom vypovedá šikmosť a ako ju vypočítame v Exceli?
4. Približne aký tvar by mal mať histogram, aby sme údaje mohli považovať za normálne rozdelené?
5. Ako vieme využiť modus, medián a aritmetický priemer pri overovaní normality dát?
6. Ako je formulovaná nulová hypotéza pri testoch normality?
7. Aký vplyv má veľkosť výberového súboru na testy normality?
8. Čo treba brať do úvahy pri overovaní, či údaje pochádzajú zo súboru s normálnym rozdelením?

Úlohy na riešenie:

1. Analytik v banke skúmal mesačné zisky klientov z investovania do akciového fondu. Náhodným výberom získal údaje o mesačných ziskoch 12 klientov za posledný rok (v eurách): 120; 310; 182; 18; 305; 520; 140; 645; 1500; 50; 200; 158. Overte, či tieto údaje pochádzajú zo súboru s normálnym rozdelením.

   [údaje nemôžeme považovať za normálne rozdelené]

2. Dátový analytik skúmal výšku miezd v cestovnom ruchu. Náhodným výberom získal 32 údajov v eurách: 1285, 1050, 1922, 877, 1648, 2315, 1102, 1529, 1470, 724, 1398, 2080, 989, 1736, 1163, 2488, 1018, 912, 1345, 1572, 653, 1899, 1221, 832, 1998, 1624, 1147, 2241, 1379, 1787, 945, 1430. Overte, či tieto údaje pochádzajú zo súboru s normálnym rozdelením.

   [údaje môžeme považovať za normálne rozdelené]

3. Analytik skúmal mesačnú výšku výdavkov domácností v eurách. Náhodným výberom získal 15 údajov: 182, 345, 119, 268, 492, 221, 154, 308, 412, 235, 169, 323, 437, 251, 184. Overte, či tieto údaje pochádzajú zo súboru s normálnym rozdelením.

   [údaje môžeme považovať za normálne rozdelené]

4. Analytik skúmal mesačné výdavky vysokoškolských študentov na mobilné paušály v eurách. Náhodným výberom získal 35 údajov: 15,98; 24,32; 10,74; 19,16; 31,54; 22,42; 17,08; 28,40; 37,78; 26,16; 19,54; 30,92; 40,26; 28,68; 21,10; 33,58; 24,96; 18,34; 29,72; 41,06; 28,44; 21,82; 34,20; 42,58; 30,96; 23,34; 36,72; 45,06; 33,44; 26,82; 39,20; 47,58; 35,96; 29,34; 41,72. Overte, či tieto údaje pochádzajú zo súboru s normálnym rozdelením.

   [údaje môžeme považovať za normálne rozdelené]

8 PARAMETRICKÉ TESTY

Kľúčové slová: parametrické testy, normálne rozdelenie, F-test, jednovýberový t-test, dvojvýberový t-test, párový t-test, centrálna limitná veta, závislé vzorky, nezávislé vzorky

Testy štatistických hypotéz môžeme deliť z viacerých hľadísk. Napríklad z hľadiska formulácie hypotéz na jednostranné a obojstranné. Nás však bude viac zaujímať delenie z hľadiska toho, či sa pri testovaní uvažujú špecifické predpoklady o parametroch populácie. Vtedy delíme testy na **parametrické** a **neparametrické**.

8.1 Úvodné pojmy

Štatistické metódy, založené na predpokladoch o rozdelení pravdepodobnosti v základnom súbore, sa nazývajú **parametrické**. Slúžia na odhadovanie parametrov rozdelenia, ako napríklad stredná hodnota alebo rozptyl, a na testovanie hypotéz o nich. Pre dosiahnutie dôveryhodných výsledkov je nevyhnutné **splnenie východiskových predpokladov** o rozdelení pravdepodobnosti v základnom súbore.

S touto tematikou úzko súvisí tzv. **centrálna limitná veta**. Jej prvú verziu formuloval v roku 1810 Pierre Laplace. Centrálna limitná veta je jedným z dôvodov, pre ktoré je Gaussova normálna krivka v štatistike dôležitá (Crilly, 2011). Jej formulácií existuje viac a dajú sa nájsť v rôznej literatúre, pre naše potreby však postačí jej neformálne vysvetlenie. Ak premenná, napríklad výška, vznikla ako súčet veľkého počtu efektov nezávisle pôsobiacich príčin, má približne normálne rozdelenie. Aproximácia je tým lepšia, čím je počet prispievajúcich faktorov väčší. Z centrálnej limitnej vety tiež vyplýva, že aritmetický priemer ako náhodná premenná je za veľmi malých obmedzení asymptoticky normálne rozdelený. Jeho náhodné správanie môžeme aproximovať pomocou normálneho rozdelenia. Centrálna limitná veta tiež vysvetľuje, prečo sa s normálnym rozdelením často stretávame v prírode, ale aj v ľudskej spoločnosti. Napríklad výška človeka, výška IQ, hmotnosť a ďalšie sú ovplyvňované mnohými faktormi a tak majú rozdelenie veľmi blízke normálnemu (Hendl 2004; Rimarčík, 2007).

Dôsledkom centrálnej limitnej vety možno parametrické metódy použiť bez ohľadu na rozdelenie premennej v základnom súbore, ak sú výberové súbory dostatočne veľké. Väčšinou sa uvádza $n > 50$, môžeme sa však stretnúť aj s údajom $n > 30$. Veľký pozor treba ale dávať na extrémne hodnoty, ktoré môžu spôsobiť výrazné skreslenie výsledkov. Vyradenie jediného extrémneho pozorovania môže úplne zmeniť vypočítaný priemer a následne výsledok parametrického testu (Rimarčík, 2007).

V ďalšom texte budeme používať dôležité pojmy, ktorými sú závislé a nezávislé vzorky. Vzorky sú **nezávislé**, ak obsahujú rôzne štatistické jednotky. Napríklad

v jednej skupine sú ženy a v druhej muži. Štatistická jednotka, ktorá sa nachádza v jednej skupine nemôže byť aj v druhej.

Závislé vzorky najčastejšie pozostávajú z rovnakých štatistických jednotiek opakovane meraných v rôznych časových okamihoch (Rimarčík, 2007). Ešte podrobnejšie môžeme povedať, že o závislých výberoch hovoríme najčastejšie vtedy, ak boli hodnoty premennej X zisťované na štatistických jednotkách, ktoré tvoria „logický pár“. Ide o prípady, keď sú hodnoty tej istej premennej zisťované na:

- rovnakých štatistických jednotkách, ale za rôznych podmienok, pričom najčastejšie ide o dve rôzne časové obdobia,
- náhodne vybraných dvojiciach (pároch) štatistických jednotiek, medzi ktorými existuje určitý vzťah.

Závislé výbery sú napríklad údaje o zisku firiem, ktoré boli získané pred investovaním do reklamnej kampane a po ňom, hodnoty pevnosti náhodne vybraných testovaných odliatkov pred tepelným spracovaním a po ňom a tak ďalej (Pacáková a kol., 2015).

8.2 Jednovýberový t-test

Je to test zhody strednej hodnoty s konštantou (konkrétne číslo). Predpokladáme, že náhodná premenná X má v základnom súbore normálne rozdelenie pravdepodobnosti so strednou hodnotou μ a rozptylom σ^2. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \mu = \mu_0$ (stredná hodnota sa rovná konštante).

$H_1: \mu \neq \mu_0$ (stredná hodnota sa nerovná konštante – obojstranná alternatíva).

Ak by sme chceli použiť jednostrannú alternatívu, máme dve možnosti:

$H_1: \mu > \mu_0$ (stredná hodnota je väčšia ako konštanta).

$H_1: \mu < \mu_0$ (stredná hodnota je menšia ako konštanta).

Na tento test nie je priamo v Exceli funkcia, počítame ho pomocou doplnku Real Statistics v záložke **Misc**, prvá položka **T Test and Non-parametric Equivalents**.

Príklad 8.1 V hotelovej reštaurácii sme pre účastníkov zájazdov v priebehu mesiaca objednali 500 obedov. Mali sme dohodnutú priemernú gramáž 150 g mäsa na porciu. Za účelom kontroly sa náhodne vybralo 20 porcií, zistené údaje v gramoch boli nasledovné: 151; 152; 149; 150; 155; 146; 148; 153; 144; 150; 151; 157; 144; 145; 153; 142; 156; 152; 148; 151.

Predpokladáme, že hmotnosť je náhodná premenná, ktorá má normálne rozdelenie. Otestujme, či na základe tejto vzorky môžeme tvrdiť, že reštaurácia systematicky nedodržiava dohodnutú priemernú gramáž a dáva menej ako 150 g.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpca „A“. Vypočítame výberový priemer $\overline{x} = 149{,}85$ g, ktorý je o niečo menší ako očakávaná stredná hodnota 150 gramov. Na prvý pohľad sa zdá, že rozdiel je veľmi malý, nebudeme to však posudzovať subjektívnym dojmom, ale presným výpočtom. Zmyslom testovania hypotézy je overiť, či je tento rozdiel signifikantný (štatisticky významný) a teda či to kuchári v reštaurácii robia systematicky (zámerne). V prípade nevýznamného výsledku ide o zanedbateľný rozdiel, ktorý je skôr náhodou v tomto konkrétnom výbere. Formulácia hypotéz je takáto:

$H_0: \mu = 150$, stredná hodnota hmotnosti mäsa je 150 gramov, hotelová reštaurácia nerobí žiadnu systematickú odchýlku v hmotnosti podávaného mäsa.

$H_1: \mu < 150$, stredná hodnota hmotnosti mäsa je menej ako 150 gramov, hotelová reštaurácia systematicky (zámerne) podáva menej mäsa.

Okrem formálneho zápisu boli hypotézy formulované aj slovne, čo je veľmi vhodné a odporúčané. Záverečná interpretácia je vďaka tomu jasnejšia a jednoduchšia. Bola vybratá jednostranná alternatívna hypotéza (ľavostranná), pretože skúmame, či hmotnosť mäsa nie je systematicky menšia ako 150 gramov.

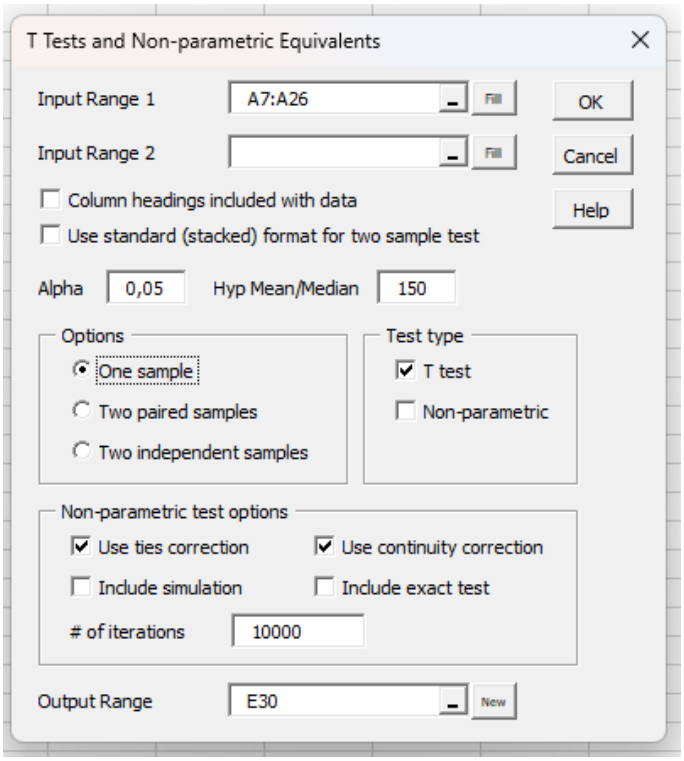
Zvolíme hladinu významnosti $\alpha = 0{,}05$.

Vstupné dáta by sme mohli preveriť z hľadiska normality, avšak v zadaní príkladu je už vyslovené **splnenie predpokladu o normálnom rozdelení**. Keďže porovnávame strednú hodnotu s konštantou, použijeme **jednovýberový t-test**.

P-hodnotu vypočítame pomocou doplnku Real Statistics v záložke Misc, prvá položka T Test and Non-parametric Equivalents. V dialógovom okne je potrebné najskôr vložiť vstupné údaje do položky „Input range 1“. Zrušíme zaškrtávacie políčko „Column headings included with data“. Použili by sme ho iba v prípade, ak by sme nad údajmi mali ešte záhlavie.

Položka „Alpha“ je hladina významnosti, napíšeme tam číslo 0,05. Dôležitá položka je „Hyp Mean/Median“, kde zadáme konštantu, s ktorou porovnávame strednú hodnotu. V tomto príklade je to číslo 150.

V časti „Options“ zvolíme „One sample“, keďže chceme jednovýberový test. V časti „Test type“ zvolíme „T test“. Ešte je vhodné nastaviť ľavú hornú bunku výstupu v políčku „Output range“ tak, aby bol dostatok miesta pre vypočítané hodnoty a aby nenastalo nevhodné prekrytie s inými bunkami. Ostatné bunky a nastavenia ponecháme bez zmeny. Správne vyplnené dialógové okno je na obrázku 8.1.



Obr. 8.1 Dialógové okno pre jednovýberový t-test (zdroj: vlastné spracovanie)

Po stlačení tlačidla OK sa zobrazí tabuľka s výpočtami (obr. 8.2). Najdôležitejší výstup je p-hodnota, ktorá je v tomto prípade 0,436. Keďže p-hodnota je väčšia ako hladina významnosti (0,436 > 0,05), tak nulovú hypotézu H_0 na hladine významnosti 0,05 **nezamietame**. Priemerná hmotnosť mäsa sa neodlišuje od predpísanej hodnoty 150 gramov. Zistené rozdiely nie sú štatisticky významné.

T Test: One Sample

SUMMARY			Alpha	0,05			
Count	*Mean*	*Std Dev*	*Std Err*	*t*	*df*	*Cohen d*	*Effect r*
20	149,85	4,132987	0,924164	-0,16231	19	0,036293	0,03721

T TEST			Hyp Mean	150	
	p-value	*t-crit*	*lower*	*upper*	*sig*
One Tail	0,436388	1,729133			no
Two Tail	0,872776	2,093024	147,9157	151,7843	no

Obr. 8.2 Výpočty k príkladu 8.1 (zdroj: vlastné spracovanie)

Poznámka: V reálnej praxi by na testovanie hypotézy a následné zovšeobecnenie na celý základný súbor bola potrebná oveľa početnejšia vzorka. Príklad je ilustračný a cieľom bolo ukázať metodický postup pri testovaní.

8.3 Dvojvýberový F-test

Je to test zhody dvoch rozptylov. Najčastejšie sa používa ako pomocný test pre dvojvýberový t-test (ukážeme neskôr), ale využitie má aj na priame riešenie úloh z praxe. Predpokladáme, že máme dva **nezávislé výbery**. Prvý výber pochádza zo súboru s normálnym rozdelením so strednou hodnotou μ_1 a rozptylom σ_1^2. Druhý výber pochádza zo súboru s normálnym rozdelením so strednou hodnotou μ_2 a rozptylom σ_2^2. Rozsah údajov vo výberových súboroch môže, ale nemusí byť rovnaký. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \sigma_1^2 = \sigma_2^2$ (rozptyly základných súborov sú rovnaké).

$H_1: \sigma_1^2 \neq \sigma_2^2$ (rozptyly základných súborov sú rôzne – obojstranná alternatíva).

Ak by sme chceli použiť jednostrannú alternatívu, máme dve možnosti:

$H_1: \sigma_1^2 > \sigma_2^2$ (rozptyl prvého súboru je väčší ako rozptyl druhého súboru).

$H_1: \sigma_1^2 < \sigma_2^2$ (rozptyl prvého súboru je menší ako rozptyl druhého súboru).

V **Exceli** na jeho realizáciu používame funkciu **F.TEST**. Prostredníctvom tejto funkcie je však možné získať p-hodnotu iba pre obojstrannú alternatívu. Ak by sme chceli výpočet pre jednostrannú alternatívu, tak je to možné prostredníctvom **Data Analysis** a **F-Test Two-sample for Variances**. Tento nástroj sa nachádza v hlavnom menu v záložke **Údaje**. Ak realizujeme výpočet prostredníctvom nástroja Data Analysis, tak ako prvé vkladáme údaje tej vzorky, ktorá má vyššiu hodnotu výberového rozptylu.

Príklad 8.2 Porovnávali sme mzdy robotníkov v dvoch veľkých podnikoch. Náhodne sme vybrali 20 robotníkov prvého podniku s týmito príjmami (€): 760; 910; 1260; 960; 980; 810; 1240; 1180; 950; 900; 1130; 1100; 770; 1300; 890; 980; 1010; 940; 880; 1050. Potom sme náhodne vybrali 17 robotníkov druhého podniku s týmito príjmami (€): 855; 930; 1050; 1080; 1020; 980; 1120; 1040; 1050; 1020; 1040; 1000; 980; 880; 940; 1030; 990. Zistite, či je medzi týmito dvoma podnikmi vo všeobecnosti rovnaká variabilita miezd alebo sa štatisticky významne odlišuje. Predpokladáme, že obidva výbery pochádzajú zo súboru s normálnym rozdelením.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpcov „A“ a „B“. Len pre zaujímavosť, ak by sme vypočítali výberové priemery, dostali by sme takmer identické hodnoty: $\overline{x_1} = 1000$ a $\overline{x_2} = 1000{,}29$. Vidíme, že v uvedených podnikoch robotníci dostávajú mzdu priemerne približne v rovnakej výške. Napríklad pomocou dvojvýberového t-testu (je vysvetlený neskôr) by bolo možné zistiť, či je rozdiel v priemernej mzde robotníkov štatisticky významný. Našou úlohou však je otestovať, či sa mzdy nelíšia variabilitou.

Formulácia hypotéz je takáto:

$H_0: \sigma_1^2 = \sigma_2^2$ (mzdy robotníkov v obidvoch podnikoch majú rovnaký rozptyl).

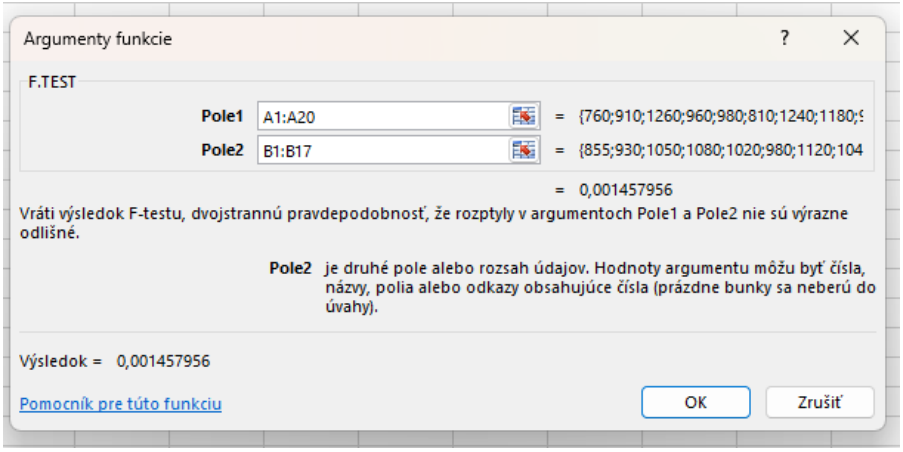
$H_1: \sigma_1^2 \neq \sigma_2^2$ (mzdy robotníkov v skúmaných podnikoch majú rôzny rozptyl).

Okrem formálneho zápisu boli hypotézy formulované aj slovne, čo je veľmi vhodné a odporúčané. Záverečná interpretácia je vďaka tomu jasnejšia a jednoduchšia. Bola vybratá obojstranná alternatívna hypotéza, pretože skúmame, či je medzi rozptylmi štatisticky významný rozdiel a nie ktorý rozptyl je väčší alebo menší.

Zvolíme hladinu významnosti $\alpha = 0{,}05$.

Vstupné dáta by sme mohli preveriť z hľadiska normality, avšak v zadaní príkladu je už vyslovené **splnenie predpokladu o normálnom rozdelení**. Keďže porovnávame zhodu dvoch rozptylov, použijeme **dvojvýberový F-test**.

Keďže sme formulovali obojstrannú alternatívu, tak **p-hodnotu** môžeme vypočítať prostredníctvom funkcie F.TEST. Dialógové okno je veľmi jednoduché. Do položiek „Pole1“ a „Pole2“ vložíme vstupné údaje z výberových súborov (obr. 8.3).



Obr. 8.3 Dialógové okno pre F-test (zdroj: vlastné spracovanie)

Po stlačení tlačidla OK sa zobrazí vypočítaná p-hodnota (obr. 8.4). Keďže je menšia ako hladina významnosti (0,0015 < 0,05), tak nulovú hypotézu H_0 na hladine významnosti 0,05 **zamietame a prijímame alternatívnu hypotézu**.

p-hodnota:	=F.TEST(A1:A20;B1:B17)	p-hodnota:	0,0015

Obr. 8.4 Výpočet a výsledky k príkladu 8.2 (zdroj: vlastné spracovanie)

Medzi rozptylmi miezd robotníkov je štatisticky významný rozdiel.

8.4 Dvojvýberový t-test

Je to test zhody dvoch stredných hodnôt pre nezávislé výbery. Podobne ako pri F-teste predpokladáme, že máme dva **nezávislé výbery**. Prvý výber pochádza zo súboru s normálnym rozdelením so strednou hodnotou μ_1 a rozptylom σ_1^2. Druhý výber pochádza zo súboru s normálnym rozdelením so strednou hodnotou μ_2 a rozptylom σ_2^2. Rozsah údajov vo výberových súboroch môže, ale nemusí byť rovnaký. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \mu_1 = \mu_2$ (stredné hodnoty základných súborov sa rovnajú).

$H_1: \mu_1 \neq \mu_2$ (stredné hodnoty sú rôzne – obojstranná alternatíva).

Ak by sme chceli použiť jednostrannú alternatívu, máme dve možnosti:

$H_1: \mu_1 > \mu_2$ (stredná hodnota prvého súboru je väčšia ako stredná hodnota druhého súboru).

$H_1: \mu_1 < \mu_2$ (stredná hodnota prvého súboru je menšia ako stredná hodnota druhého súboru).

Tento test je možné v Exceli počítať prostredníctvom funkcie T.TEST alebo pomocou doplnku Real Statistics v záložke **Misc**, prvá položka **T Test and Non-parametric Equivalents**.

Na to, aby bolo možné vypočítať p-hodnotu dvojvýberovým t-testom, je potrebné vedieť, či ide o náhodné výbery zo základných súborov s **rovnakými** alebo **rôznymi rozptylmi**. To je možné zistiť pomocou F-testu, ktorý má v tomto prípade využitie ako **pomocný test** pre dvojvýberový t-test. **Až na základe výsledku F-testu je možné správne zrealizovať dvojvýberový t-test**.

Príklad 8.3 Porovnávali sme ceny najpredávanejšieho 95-oktánového benzínu v Trnavskom a Prešovskom kraji. Náhodne sme vybrali 21 čerpacích staníc v Trnavskom kraji a zistili sme tieto ceny (€):

1,72; 1,7; 1,71; 1,69; 1,68; 1,71; 1,72; 1,75; 1,69; 1,68; 1,69; 1,69; 1,71; 1,72; 1,7; 1,71; 1,72; 1,67; 1,68; 1,72; 1,71.

Potom sme náhodne vybrali 19 čerpacích staníc v Prešovskom kraji, tu sú ceny (€):

1,68; 1,69; 1,67; 1,7; 1,66; 1,65; 1,69; 1,68; 1,72; 1,7; 1,67; 1,65; 1,68; 1,69; 1,7; 1,71; 1,69; 1,68; 1,67.

Zistite, či je stredná hodnota ceny benzínu rovnaká v Trnavskom aj v Prešovskom kraji alebo sa štatisticky významne odlišuje. Predpokladáme, že obidva výbery pochádzajú zo súboru s normálnym rozdelením. Poznámka: Príklad je iba ilustračný, údaje nie sú skutočné.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpcov „A“ a „B“. Najskôr potrebujeme zistiť, či sú rozptyly rôzne alebo ich môžeme považovať za rovnaké. Overíme to pomocou F-testu, ktorý v tomto prípade použijeme ako pomocný test. Formulácia hypotéz je takáto:

H_0: $\sigma_1^2 = \sigma_2^2$ (ceny v obidvoch krajoch majú rovnaký rozptyl).

H_1: $\sigma_1^2 \neq \sigma_2^2$ (ceny v obidvoch krajoch majú rôzny rozptyl).

Zvolíme hladinu významnosti $\alpha = 0{,}05$.

Vstupné dáta by sme mohli preveriť z hľadiska normality, avšak v zadaní príkladu je už vyslovené **splnenie predpokladu o normálnom rozdelení**. P-hodnotu vypočítame prostredníctvom funkcie F.TEST tým istým spôsobom ako v príklade 8.2. Keďže vypočítaná p-hodnota je väčšia ako hladina významnosti (0,97 > 0,05), tak nulovú hypotézu nemôžeme zamietnuť. Na testovanie stredných hodnôt teda môžeme využiť dvojvýberový t-test s rovnakými rozptylmi. Formulácia nulovej a alternatívnej hypotézy pre dvojvýberový t-test je nasledovná:

H_0: $\mu_1 = \mu_2$ (stredné hodnoty cien benzínu v obidvoch krajoch sa rovnajú).

H_1: $\mu_1 \neq \mu_2$ (stredné hodnoty cien benzínu v obidvoch krajoch sú rôzne).

Použili sme obojstrannú alternatívu, lebo v zadaní príkladu je skúmanie rozdielu v stredných hodnotách (neskúmame v ktorom kraji je stredná hodnota vyššia). Keď vypočítame výberové priemery, získame nasledovné hodnoty: $\overline{x_1} = 1{,}703$ € a $\overline{x_2} = 1{,}683$ €. Vidíme medzi krajmi rozdiel dva eurocenty. Budeme testovať, či je rozdiel štatisticky významný.

Zvolíme hladinu významnosti $\alpha = 0{,}05$.

Testujeme zhodu dvoch stredných hodnôt pre nezávislé výbery, takže použijeme **dvojvýberový t-test**.

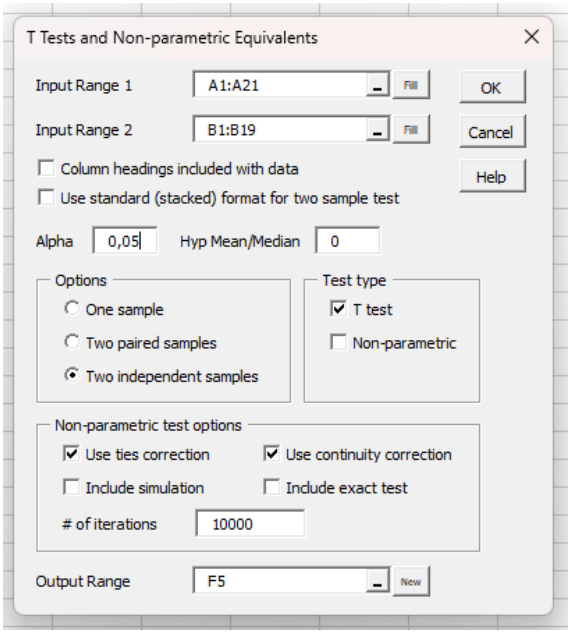
P-hodnotu vypočítame pomocou doplnku Real Statistics v záložke Misc, prvá položka T Test and Non-parametric Equivalents. V dialógovom okne je potrebné najskôr vložiť vstupné údaje do položiek „Input range 1“ a „Input range 2“. Zrušíme zaškrtávacie políčko „Column headings included with data“. Použili by sme ho iba v prípade, ak by sme nad údajmi mali ešte záhlavie.

Položka „Alpha“ je hladina významnosti, napíšeme tam číslo 0,05.

V časti „Options“ zvolíme „Two independent samples“, keďže chceme dvojvýberový test pre dva nezávislé výbery.

V časti „Test type“ zvolíme „T test“. Ešte je vhodné nastaviť ľavú hornú bunku výstupu v políčku „Output range“ tak, aby bol dostatok miesta pre vypočítané hodnoty a aby nenastalo nevhodné prekrytie s inými bunkami. Ostatné bunky

a nastavenia ponecháme bez zmeny. Správne vyplnené dialógové okno je na obrázku 8.5.



Obr. 8.5 Dialógové okno pre dvojvýberový t-test (zdroj: vlastné spracovanie)

Po stlačení tlačidla OK sa zobrazí tabuľka s výpočtami (obr. 8.6). Najdôležitejší výstup je p-hodnota, ktorá je v tomto prípade 0,0018.

T Test: Two Independent Samples

SUMMARY			Hyp Mean	0
Groups	*Count*	*Mean*	*Variance*	*Cohen d*
Group 1	21	1,703333	0,000363	
Group 2	19	1,683158	0,000356	
Pooled			0,00036	1,063448

T TEST: Equal Variances				Alpha	0,05				
	std err	*t-stat*	*df*	*p-value*	*t-crit*	*lower*	*upper*	*sig*	*effect r*
One Tail	0,006007	3,358712	38	0,000896	1,685954			yes	0,478446
Two Tail	0,006007	3,358712	38	0,001791	2,024394	0,008015	0,032336	yes	0,478446

T TEST: Unequal Variances				Alpha	0,05				
	std err	*t-stat*	*df*	*p-value*	*t-crit*	*lower*	*upper*	*sig*	*effect r*
One Tail	0,006004	3,360435	37,67566	0,000897	1,686317			yes	0,480218
Two Tail	0,006004	3,360435	37,67566	0,001794	2,024967	0,008018	0,032333	yes	0,480218

Obr. 8.6 Výpočty k príkladu 8.3 (zdroj: vlastné spracovanie)

Excel v tomto prípade ponúkal viac p-hodnôt. Bolo však potrebné vybrať t-test s rovnakými rozptylmi (Equal Variances) a obojstrannú alternatívu (Two Tail). Keďže

p-hodnota je menšia ako hladina významnosti (0,0018 < 0,05), tak nulovú hypotézu H_0 na hladine významnosti 0,05 **zamietame a prijímame alternatívnu hypotézu**. Priemerné ceny benzínu v Trnavskom a Prešovskom kraji sú rozdielne. Zistené rozdiely sú štatisticky významné (signifikantné).

8.5 Párový t-test

Je to test zhody dvoch stredných hodnôt pre **závislé výbery**. Najčastejšie ide teda o situáciu, keď sú hodnoty tej istej premennej zisťované na rovnakých štatistických jednotkách, ale za rôznych podmienok. Zvyčajne ide o dve rôzne časové obdobia, čiže zisťujeme, ako sa nejaká vlastnosť mení v čase. Prípadne môžeme skúmať náhodne vybrané dvojice (páry) štatistických jednotiek, medzi ktorými existuje určitý vzťah.

Predpokladajme, že štatistický znak meriame na každej štatistickej jednotke dvakrát, respektíve meriame ho na dvojici štatistických jednotiek, ktoré tvoria páry. Získame hodnoty náhodného výberu $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. Predpokladáme, že tento náhodný výber pochádza z dvojrozmerného základného súboru (X, Y) s dvojrozmerným normálnym rozdelením pravdepodobnosti s neznámymi parametrami $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$. Keďže cieľom testu je overiť zhodu dvoch stredných hodnôt, zisťujeme či platí $\mu_1 = \mu_2$. To je analogické tvrdeniu, že $\mu_1 - \mu_2 = 0$. Teda namiesto pôvodného tvrdenia o rovnosti stredných hodnôt dvoch základných súborov môžeme predpokladať, že výberový súbor rozdielov $D = (D_1, D_2, \dots D_n) = (X_1 - Y_1, X_2 - Y_2, \dots X_n - Y_n)$ pochádza zo základného súboru s normálnym rozdelením pravdepodobnosti. Takže čo sa týka predpokladov pre párový t-test, v praxi je potrebné overiť **splnenie predpokladu o normálnom rozdelení základného súboru diferencií** (Švábová a kol., 2022). Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \mu_1 = \mu_2$ (stredné hodnoty základných súborov sa rovnajú).

$H_1: \mu_1 \neq \mu_2$ (stredné hodnoty sú rôzne – obojstranná alternatíva).

Ak by sme chceli použiť jednostrannú alternatívu, máme dve možnosti:

$H_1: \mu_1 > \mu_2$ (stredná hodnota prvého súboru je väčšia ako stredná hodnota druhého súboru).

$H_1: \mu_1 < \mu_2$ (stredná hodnota prvého súboru je menšia ako stredná hodnota druhého súboru).

Tento test je možné v Exceli počítať prostredníctvom funkcie T.TEST alebo pomocou doplnku Real Statistics v záložke **Misc**, prvá položka **T Test and Non-parametric Equivalents**.

Príklad 8.4 Obchodný reťazec sa rozhodol zaviesť nový prvok do predaja (kupóny v mobilnej aplikácii). Manažment obchodného reťazca chce zistiť, či tento nový prvok prinesie zmenu tržieb. Náhodne bolo vybratých 18 zákazníkov, u ktorých sa počas troch mesiacov sledovalo, koľko míňajú na nákupy v uvedenom obchodnom reťazci. Boli zaznamenané tieto údaje (€): 150; 162; 199; 215; 84; 64; 242; 585; 214; 95; 143; 176; 286; 123; 101; 94; 202; 155.

Potom sa zaviedol nový prvok do predaja a počas ďalších troch mesiacov sa sledovali nákupy u tých istých zákazníkov s týmito zistenými dátami (€): 163; 175; 192; 228; 90; 72; 250; 620; 201; 96; 182; 185; 315; 135; 111; 99; 184; 190.

Predpokladáme, že je splnený predpoklad o normálnom rozdelení základného súboru diferencií, resp. že náhodný výber pochádza z dvojrozmerného základného súboru s dvojrozmerným normálnym rozdelením pravdepodobnosti.

Poznámka: V reálnej praxi by na testovanie hypotézy a následné zovšeobecnenie na celý základný súbor (všetkých zákazníkov) bola potrebná oveľa početnejšia vzorka. Príklad je ilustračný (údaje nie sú skutočné) a cieľom je iba ukázať metodický postup pri testovaní. Uvedený postup má zmysel vtedy, ak by na zmenu výdavkov nemali vplyv iné faktory (napríklad inflácia). V opačnom prípade by bolo potrebné zohľadniť aj ďalšie faktory.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpcov „A“ a „B“. Formulácia nulovej a alternatívnej hypotézy pre párový t-test je nasledovná:

$H_0: \mu_1 = \mu_2$ (stredné hodnoty výdavkov zákazníkov pred a po zavedení nového prvku do predaja sa rovnajú).

$H_1: \mu_1 \neq \mu_2$ (stredné hodnoty výdavkov zákazníkov pred a po zavedení nového prvku do predaja sú rôzne).

Použili sme obojstrannú alternatívu, lebo v zadaní príkladu je skúmanie rozdielu v stredných hodnotách (neskúmame, či nastal nárast, resp. pokles). Hoci v tomto prípade by mala zmysel aj jednostranná alternatíva, lebo manažment obchodného reťazca zaujíma nielen zmena, ale predovšetkým nárast tržieb. Keď vypočítame výberové priemery, získame nasledovné hodnoty: $\overline{x_1} = 182{,}78$ € a $\overline{x_2} = 193{,}78$ €. Z popisnej štatistiky teda vidíme rozdiel 11 eur. Budeme testovať, či je rozdiel štatisticky významný.

Zvolíme hladinu významnosti $\alpha = 0{,}05$.

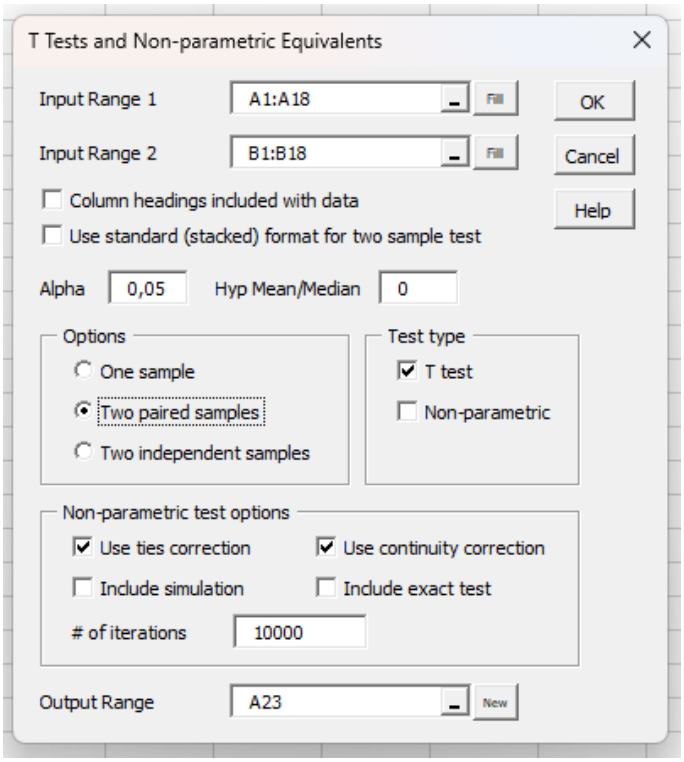
Vstupné dáta by sme mohli preveriť z hľadiska normality, avšak v zadaní príkladu je už vyslovené **splnenie predpokladu o normálnom rozdelení**. Keďže porovnávame zhodu dvoch stredných hodnôt pre závislé výbery, použijeme **párový t-test**.

P-hodnotu vypočítame pomocou doplnku Real Statistics v záložke Misc, prvá položka T Test and Non-parametric Equivalents. V dialógovom okne je potrebné najskôr vložiť vstupné údaje do položiek „Input range 1“ a „Input range 2“. Zrušíme zaškrtávacie políčko „Column headings included with data“. Použili by sme ho iba v prípade, ak by sme nad údajmi mali ešte záhlavie.

Položka „Alpha“ je hladina významnosti, napíšeme tam číslo 0,05.

V časti „Options“ zvolíme „Two paired samples“, keďže chceme párový test pre závislé výbery.

V časti „Test type“ zvolíme „T test“. Ešte je vhodné nastaviť ľavú hornú bunku výstupu v políčku „Output range“ tak, aby bol dostatok miesta pre vypočítané hodnoty a aby nenastalo nevhodné prekrytie s inými bunkami. Ostatné bunky a nastavenia ponecháme bez zmeny. Správne vyplnené dialógové okno je na obrázku 8.7.



Obr. 8.7 Dialógové okno pre párový t-test (zdroj: vlastné spracovanie)

Po stlačení tlačidla OK sa zobrazí tabuľka s výpočtami (obr. 8.8). Najdôležitejší výstup je p-hodnota, ktorá je v tomto prípade 0,0086.

T Test: Two Paired Samples

SUMMARY			Alpha	0,05		Hyp Mean	0	
Groups	*Count*	*Mean*	*Std Dev*	*Std Err*	*t*	*df*	*Cohen d*	*Effect r*
Group 1	18	182,7778	116,8733					
Group 2	18	193,7778	122,9115					
Difference	18	-11	15,71436	3,703911	-2,96983	17	0,699997	0,58446

T TEST

	p-value	*t-crit*	*lower*	*upper*	*sig*
One Tail	0,004295	1,739607			yes
Two Tail	0,008589	2,109816	-18,8146	-3,18543	yes

Obr. 8.8 Výpočty k príkladu 8.4 (zdroj: vlastné spracovanie)

Keďže p-hodnota je menšia ako hladina významnosti (0,0086 < 0,05), tak nulovú hypotézu H_0 na hladine významnosti 0,05 **zamietame a prijímame alternatívnu hypotézu**. Nový prvok v predaji priniesol zmenu tržieb. Zistené rozdiely sú štatisticky významné (signifikantné).

Kontrolné otázky:

1. Ako môžeme deliť testy štatistických hypotéz?
2. Je potrebné pri parametrických testoch dodržať nejaké východiskové predpoklady?
3. Ako by ste neformálne vysvetlili centrálnu limitnú vetu?
4. Ako súvisí centrálna limitná veta s parametrickými testami?
5. Kedy sú vzorky závislé a kedy nezávislé?
6. Vedeli by ste uviesť konkrétne príklady závislých a nezávislých vzoriek?
7. Akým parametrickým testom porovnávame strednú hodnotu s konštantou?
8. Akým parametrickým testom zisťujeme zhodu dvoch rozptylov?
9. V akej situácii používame F-test ako pomocný test?
10. Čo zisťujeme prostredníctvom dvojvýberového t-testu?
11. V akých situáciách je možné využiť párový t-test?
12. Akým spôsobom overujeme v praxi splnenie predpokladu o normálnom rozdelení základného súboru pre párový t-test?

Úlohy na riešenie:

1. Analyzovali sme výdavky turistov na týždňovej dovolenke v konkrétnej destinácii. Náhodne sme vybrali 28 turistov a zistili sme tieto výdavky v eurách: 145; 98; 112; 107; 125; 89; 133; 101; 95; 128; 110; 109; 121; 97; 116; 105; 130; 104; 123; 92; 117; 100; 131; 119; 120; 103; 99; 127. Predpokladáme, že predpoklady normality sú splnené. Otestujme na hladine významnosti 0,05, či na základe tejto vzorky môžeme tvrdiť, že priemerné výdavky turistov sú väčšie ako 110 eur.

   [p-hodnota = 0,16; nezamietame nulovú hypotézu]

2. Porovnávali sme úrokové sadzby na hypotekárnych úveroch v dvoch bankách. Náhodne sme vybrali 22 klientov z prvej banky s týmito sadzbami v percentách: 4,25; 4,50; 3,75; 4,10; 4,37; 4,00; 4,22; 4,45; 3,95; 4,18; 4,31; 4,05; 4,27; 4,40; 3,85; 4,13; 4,36; 4,02; 4,24; 4,47; 3,90; 4,15. Potom sme náhodne vybrali 19 klientov druhej banky s týmito sadzbami v percentách: 4,42; 4,57; 3,90; 4,27; 4,48; 3,82; 4,17; 4,40; 3,75; 4,12; 4,35; 3,98; 4,23; 4,46; 3,80; 4,15; 4,38; 3,70; 4,10. Zistite, či je medzi týmito dvoma bankami vo všeobecnosti rovnaká variabilita úrokových sadzieb hypotekárnych úverov alebo sa štatisticky významne odlišuje. Predpokladáme, že obidva výbery pochádzajú zo súboru s normálnym rozdelením.

   [p-hodnota = 0,26; nezamietame nulovú hypotézu]

3. Zistite, či je stredná hodnota úrokových sadzieb v bankách z úlohy číslo 2 rovnaká alebo sa štatisticky významne odlišuje.

   [p-hodnota = 0,82; nezamietame nulovú hypotézu]

4. Zisťovali sme digitálne zručnosti u 15 náhodne vybratých študentov. Dosiahli tieto bodové výsledky: 42, 78, 56, 39, 95, 28, 83, 67, 18, 52, 71, 91, 61, 47, 89. Potom sa títo študenti zúčastnili online vzdelávania a znova sme zistili ich digitálne zručnosti: 58, 87, 72, 51, 100, 43, 94, 82, 31, 64, 80, 102, 74, 59, 98. Predpokladáme, že je splnený predpoklad o normálnom rozdelení. Testujte na hladine významnosti 0,05, či sa digitálne zručnosti signifikantne zlepšili.

   [p-hodnota = 0,00; zamietame nulovú hypotézu]

9 NEPARAMETRICKÉ TESTY

Kľúčové slová: neparametrické testy, Wilcoxonov jednovýberový test, Mann-Whitneyov U test, Wilcoxonov párový test, normálne rozdelenie pravdepodobnosti

V predchádzajúcej kapitole sme sa zaoberali vybranými parametrickými testami, ktoré boli viazané na predpoklad, že základný súbor má normálne rozdelenie. Ak nemôžeme potvrdiť východiskové predpoklady parametrických metód alebo boli dokonca niektoré predpoklady zamietnuté, tak závery získané aplikáciou parametrických metód by v takomto prípade mohli viesť k nesprávnym rozhodnutiam. V prípade, že nie sú splnené predpoklady pre použitie parametrických testov, je možnosť použiť tzv. **neparametrické testy**.

Ich hlavnou **výhodou** je, že **nevyžadujú** napríklad splnenie predpokladu **normality** rozdelenia dát. Okrem toho neparametrické postupy sú vhodné pre hodnotenie **ordinálnych údajov**, takže dôvod použitia neparametrických metód môže byť ordinálny charakter údajov (Hendl, 2004; Pacáková a kol., 2015). **Nevýhodou** je, že majú menšiu silu testu, to znamená, menšiu schopnosť odhaliť neplatnosť nulovej hypotézy. **Ak sú splnené východiskové predpoklady** pre parametrické testy, tak tieto by mali byť **uprednostnené** pred neparametrickými. Dôvodom je hlavne to, že neparametrické testy využívajú menej informácií o štatistických jednotkách (napríklad len poradia usporiadaných dát vo vzorke). Dôsledkom je už spomínaná menšia sila neparametrických testov.

Hendl (2004) zhrnul dôvody pre použitie neparametrických testov, ktorými nahrádzame použitie t-testov na tieto prípady:

- dáta nie sú normálne rozdelené,
- dáta majú ordinálny charakter,
- výbery sú malé alebo existujú veľké rozdiely medzi rozsahmi výberov,
- chceme posilniť validitu (platnosť) výsledkov parametrických metód.

Niekedy sa používa aj prístup, že zároveň s t-testami sa aplikujú aj ich neparametrické alternatívy a porovnajú sa výsledky. Ak nie sú rovnaké, skúmajú sa príčiny. Takýto prístup môže odhaliť napríklad nekonzistenciu dát. Existuje veľmi veľa neparametrických testov, my sa však budeme zaoberať iba niektorými z nich.

9.1 Wilcoxonov jednovýberový test

V odbornej literatúre je niekedy nazývaný aj jednovýberový mediánový test. Používa sa namiesto parametrického jednovýberového t-testu o strednej hodnote v situácii, keď nie sú splnené predpoklady pre použitie parametrického testu. Predpokladom

je spojité symetrické rozdelenie údajov. Na rozdiel od t-testu testujeme východiskovú hypotézu, že medián základného súboru sa rovná konštante. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \tilde{X} = x_0$ (medián základného súboru sa rovná konštante).

$H_1: \tilde{X} \neq x_0$ (medián sa nerovná konštante – obojstranná alternatíva).

Ak by sme chceli použiť jednostrannú alternatívu, máme dve možnosti:

$H_1: \tilde{X} > x_0$ (medián je väčší ako konštanta).

$H_1: \tilde{X} < x_0$ (medián je menší ako konštanta).

Na tento test nie je priamo v Exceli funkcia, počítame ho pomocou doplnku Real Statistics v záložke **Misc**, prvá položka **T Test and Non-parametric Equivalents**.

Príklad 9.1 Úradníci majú za úlohu spracovať veľké množstvo žiadostí o dotácie. Zmerali sme čas posudzovania 20 náhodne vybratých žiadostí s týmito časmi spracovania v minútach: 223; 250; 28; 198; 205; 186; 28; 45; 190; 62; 35; 85; 34; 56; 75; 205; 63; 35; 82; 51.

Predpokladáme, že údaje nepochádzajú zo súboru s normálnym rozdelením. Na hladine významnosti 0,10 testujme, či na základe tejto vzorky môžeme tvrdiť, že medián doby spracovania žiadosti je rovný požadovanej hodnote 100 minút.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpca „A“. Formulácia hypotéz je takáto:

$H_0: \tilde{X} = 100$, medián základného súboru doby spracovania žiadosti je rovný 100 minút, doba spracovania žiadostí úradníkmi sa neodlišuje od požadovanej hodnoty.

$H_1: \tilde{X} \neq 100$, medián doby spracovania žiadosti nie je rovný 100 minút, doba spracovania žiadostí úradníkmi sa odlišuje od požadovanej hodnoty.

Okrem formálneho zápisu boli hypotézy formulované aj slovne, čo je veľmi vhodné a odporúčané. Záverečná interpretácia je vďaka tomu jasnejšia a jednoduchšia. Bola vybratá obojstranná alternatívna hypotéza, pretože skúmame, či sa doba spracovania žiadosti odlišuje od požadovanej hodnoty. V zadaní príkladu nebolo požadované riešiť, či je väčšia alebo menšia.

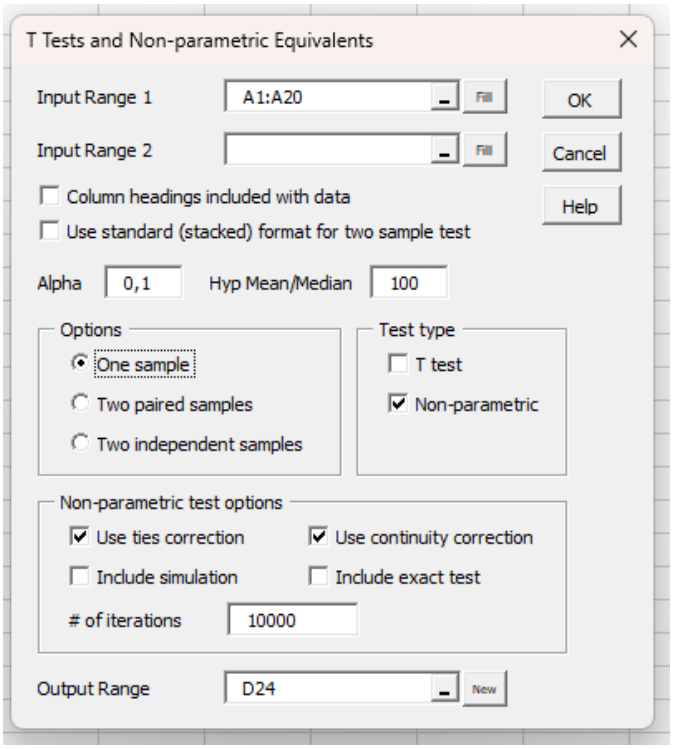
Hladina významnosti bola definovaná v zadaní príkladu $\alpha = 0{,}10$.

Vstupné dáta by sme mohli preveriť z hľadiska normality, avšak v zadaní príkladu je už vyslovené, že **nie je splnený predpoklad o normálnom rozdelení**. Keďže porovnávame medián s konštantou, použijeme **Wilcoxonov jednovýberový test**.

P-hodnotu vypočítame pomocou doplnku Real Statistics v záložke Misc, prvá položka T Test and Non-parametric Equivalents. V dialógovom okne je potrebné najskôr vložiť vstupné údaje do položky „Input range 1“. Zrušíme zaškrtávacie políčko „Column headings included with data“. Použili by sme ho iba v prípade, ak by sme nad údajmi mali ešte záhlavie.

Položka „Alpha“ je hladina významnosti, napíšeme tam číslo 0,1. Dôležitá položka je „Hyp Mean/Median“, kde zadáme konštantu, s ktorou porovnávame medián. V tomto príklade je to číslo 100.

V časti „Options“ zvolíme „One sample“, keďže chceme jednovýberový test. V časti „Test type“ zvolíme „Non-parametric“. Ešte je vhodné nastaviť ľavú hornú bunku výstupu v políčku „Output range“ tak, aby bol dostatok miesta pre vypočítané hodnoty a aby nenastalo nevhodné prekrytie s inými bunkami. Ostatné bunky a nastavenia ponecháme bez zmeny. Správne vyplnené dialógové okno je na obrázku 9.1.



Obr. 9.1 Dialógové okno pre Wilcoxonov jednovýberový test (zdroj: vlastné spracovanie)

Po stlačení tlačidla OK sa zobrazí tabuľka s výpočtami (obr. 9.2). Najdôležitejší výstup je p-hodnota, ktorá je v tomto prípade 0,62. Keďže p-hodnota je väčšia ako hladina významnosti (0,62 > 0,10), tak nulovú hypotézu H_0 na hladine významnosti 0,1 **nezamietame**. Doba spracovania žiadostí úradníkmi sa neodlišuje od požadovanej hodnoty 100 minút. Zistené rozdiely nie sú štatisticky významné.

Wilcoxon Signed-Rank Test for a Single Sample		
sample m	69	
pop medi	100	
count	20	
# unequal	20	
T+	91	
T-	119	
T	91	
	one tail	two tail
mean	105	
std dev	26,77919	ties
z-score	0,504123	yates
effect r	0,112725	
p-norm	0,307088	0,614175
p-exact	0,310757	0,621513
p-simul	N/A	N/A

Obr. 9.2 Výpočty k príkladu 9.1 (zdroj: vlastné spracovanie)

Poznámka: Príklad je ilustračný (údaje nie sú skutočné) a cieľom je iba ukázať metodický postup pri testovaní.

9.2 Mann-Whitneyov test

Tento test je neparametrickou alternatívou dvojvýberového t-testu. V prípade tohto neparametrického poradového testu pre dva nezávislé výbery sú testované východiskové hypotézy o zhode mediánov. V symetrických rozdeleniach sa mediány rovnajú stredným hodnotám. Test sa používa pre dva náhodné nezávislé výbery. Našou snahou môže byť porovnanie polohy dvoch rozdelení, ak máme údaje merané len v poradovej stupnici. Alebo sú dáta vo vyššej stupnici, ale nie sú splnené podmienky parametrických testov. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \tilde{X}_1 = \tilde{X}_2$ (mediány základných súborov sa rovnajú).

$H_1: \tilde{X}_1 \neq \tilde{X}_2$ (mediány základných súborov sa nerovnajú – obojstranná alternatíva).

Ak by sme chceli použiť jednostrannú alternatívu, máme dve možnosti:

$H_1: \tilde{X}_1 > \tilde{X}_2$ (medián prvého súboru je väčší ako medián druhého súboru).

$H_1: \tilde{X}_1 < \tilde{X}_2$ (medián prvého súboru je menší ako medián druhého súboru).

Na tento test nie je priamo v Exceli funkcia, počítame ho pomocou doplnku Real Statistics v záložke **Misc**, prvá položka **T Test and Non-parametric Equivalents**.

Príklad 9.2 Dve výrobné linky vyrábajú súčiastky. Chceme porovnať produktivitu obidvoch liniek. Náhodne bolo vybratých 15 rovnakých časových úsekov a zaznamenaná produkcia prvej linky v jednotlivých časových úsekoch (kusy): 319; 320; 320; 319; 324; 325; 319; 320; 319; 323; 324; 325; 322; 324; 320.

Produktivita druhej linky v týchto časových úsekoch bola nasledovná (kusy): 326; 319; 322; 325; 325; 320; 322; 325; 325; 326; 320; 321; 325; 326; 323.

Predpokladáme, že údaje nepochádzajú zo súboru s normálnym rozdelením. Otestujme, či na základe týchto vzoriek môžeme tvrdiť, že medián produktivity výrobných liniek je odlišný.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpcov „A“ a „B“. Na rozdiel od parametrického dvojvýberového t-testu teraz nepotrebujeme realizovať pomocný F-test. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \tilde{X}_1 = \tilde{X}_2$ (mediány produktivity obidvoch liniek sa rovnajú).

$H_1: \tilde{X}_1 \neq \tilde{X}_2$ (mediány produktivity obidvoch liniek sú rôzne).

Použili sme obojstrannú alternatívu, lebo v zadaní príkladu je skúmanie rozdielu (neskúmame ktorá linka má vyššiu produktivitu). Budeme testovať, či je rozdiel v produktivite medzi linkami štatisticky významný.

Vstupné dáta by sme mohli preveriť z hľadiska normality, avšak v zadaní príkladu je už vyslovené, že **nie je splnený predpoklad o normálnom rozdelení**. Keďže porovnávame zhodu mediánov dvoch nezávislých náhodných výberov, použijeme **Mann-Whitneyov test**. V tomto prípade je počet údajov v obidvoch vzorkách rovnaký, čo nie je podmienka pre použitie tohto testu.

Zvolíme hladinu významnosti $\alpha = 0{,}05$.

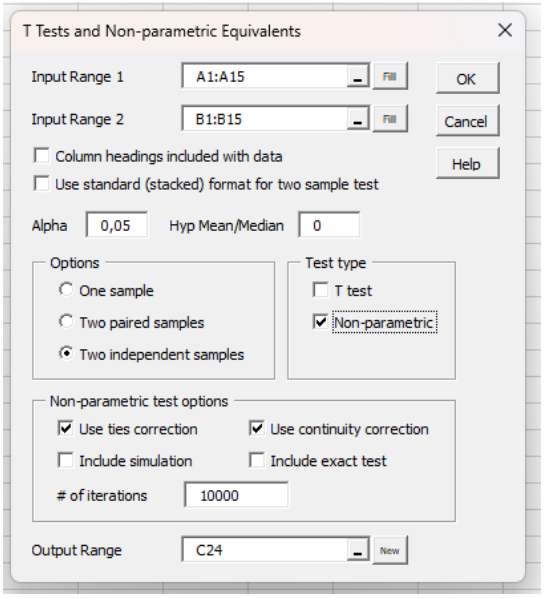
P-hodnotu vypočítame pomocou doplnku Real Statistics v záložke Misc, prvá položka T Test and Non-parametric Equivalents. V dialógovom okne je potrebné najskôr vložiť vstupné údaje do položiek „Input range 1“ a „Input range 2“. Zrušíme zaškrtávacie políčko „Column headings included with data“. Použili by sme ho iba v prípade, ak by sme nad údajmi mali ešte záhlavie.

Položka „Alpha“ je hladina významnosti, napíšeme tam číslo 0,05.

V časti „Options“ zvolíme „Two independent samples“, keďže chceme dvojvýberový test pre dva nezávislé výbery.

V časti „Test type“ zvolíme „Non-parametric“. Ešte je vhodné nastaviť ľavú hornú bunku výstupu v políčku „Output range“ tak, aby bol dostatok miesta pre vypočítané hodnoty a aby nenastalo nevhodné prekrytie s inými bunkami. Ostatné bunky

a nastavenia ponecháme bez zmeny. Správne vyplnené dialógové okno je na obrázku 9.3.



Obr. 9.3 Dialógové okno pre Mann-Whitneyov test (zdroj: vlastné spracovanie)

Po stlačení tlačidla OK sa zobrazí tabuľka s výpočtami (obr. 9.4). Najdôležitejší výstup je p-hodnota, ktorá je v tomto prípade 0,03.

Mann-Whitney Test for Two Independent Samples		
	Sample 1	Sample 2
count	15	15
median	320	325
rank sum	181,5	283,5
U	163,5	61,5
	one tail	two tail
U	61,5	
mean	112,5	
std dev	23,77426	ties
z-score	2,124146	yates
effect r	0,387814	
p-norm	0,016829	0,033658
p-exact	0,016472	0,032944
p-simul	N/A	N/A

Obr. 9.4 Výpočty k príkladu 9.2 (zdroj: vlastné spracovanie)

Keďže p-hodnota je menšia ako hladina významnosti (0,03 < 0,05), tak nulovú hypotézu H_0 na hladine významnosti 0,05 **zamietame a prijímame alternatívnu hypotézu**. Produktivita obidvoch výrobných liniek je odlišná. Zistené rozdiely sú štatisticky významné (signifikantné).

9.3 Wilcoxonov párový test

Wilcoxonov párový test *(Wilcoxon signed-rank test for paired samples)* sa používa na overenie zhody dvoch mediánov pri párových pozorovaniach. Je neparametrickou obdobou parametrického párového t-testu o stredných hodnotách pre dva závislé výbery. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \tilde{X}_1 = \tilde{X}_2$ (mediány sa rovnajú).

$H_1: \tilde{X}_1 \neq \tilde{X}_2$ (mediány sa nerovnajú – obojstranná alternatíva).

Ak by sme chceli použiť jednostrannú alternatívu, máme dve možnosti:

$H_1: \tilde{X}_1 > \tilde{X}_2$ (medián prvého súboru je väčší ako medián druhého súboru).

$H_1: \tilde{X}_1 < \tilde{X}_2$ (medián prvého súboru je menší ako medián druhého súboru).

Na tento test nie je priamo v Exceli funkcia, počítame ho pomocou doplnku Real Statistics v záložke **Misc**, prvá položka **T Test and Non-parametric Equivalents**.

Príklad 9.3 Energetická spoločnosť menila cenu elektriny pre domácnosti. Manažment spoločnosti chce zistiť, či zmena ceny ovplyvnila spotrebu elektriny v domácnostiach. Preto bolo náhodne vybratých 23 zákazníkov. U týchto zákazníkov sa sledovala spotreba elektrickej energie pred zmenou a po zmene ceny za dve rovnako dlhé obdobia. Pred zmenou ceny boli zaznamenané u jednotlivých domácností tieto údaje o spotrebe (kWh): 1093; 852; 965; 1052; 1121; 985; 1230; 1121; 988; 896; 922; 1030; 1050; 1042; 956; 1122; 1098; 1053; 998; 1011; 1035; 1154; 942.

Po zmene ceny boli zaznamenané tieto údaje o spotrebe (kWh): 1080; 844; 964; 1010; 1050; 942; 1180; 1205; 985; 901; 902; 1025; 1042; 1035; 910; 1035; 1101; 1009; 954; 921; 987; 1102; 940.

Predpokladáme, že nie je splnený predpoklad o normálnom rozdelení základného súboru diferencií, resp. že náhodný výber nepochádza z dvojrozmerného základného súboru s dvojrozmerným normálnym rozdelením pravdepodobnosti.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpcov „A“ a „B“. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \tilde{X}_1 = \tilde{X}_2$ (mediány spotreby elektriny domácností pred a po zmene ceny sa rovnajú).

$H_1: \tilde{X}_1 \neq \tilde{X}_2$ (mediány spotreby elektriny domácností pred a po zmene ceny sú rôzne).

Použili sme obojstrannú alternatívu, lebo v zadaní príkladu je skúmanie zmeny spotreby (neskúmame, či nastal nárast, resp. pokles).

Zvolíme hladinu významnosti $\alpha = 0{,}05$.

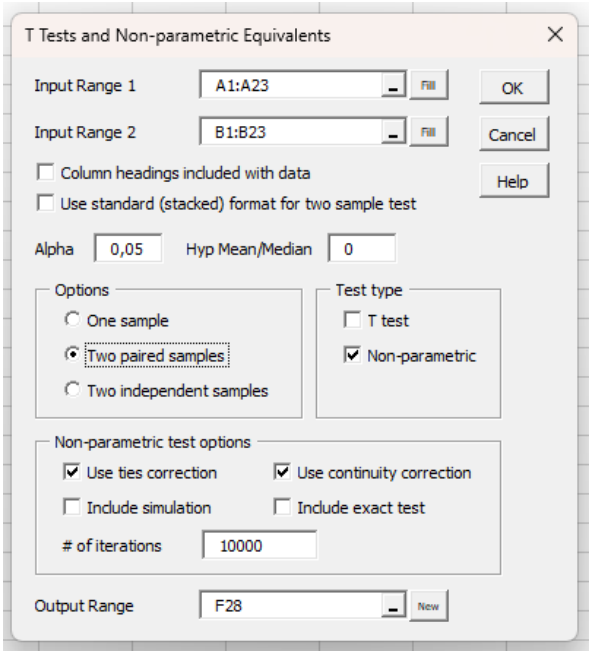
Vstupné dáta by sme mohli preveriť z hľadiska normality, avšak v zadaní príkladu je už vyslovené, že **nie je splnený predpoklad o normálnom rozdelení**. Keďže porovnávame zhodu dvoch mediánov pre závislé výbery, použijeme **Wilcoxonov párový test**.

P-hodnotu vypočítame pomocou doplnku Real Statistics v záložke Misc, prvá položka T Test and Non-parametric Equivalents. V dialógovom okne je potrebné najskôr vložiť vstupné údaje do položiek „Input range 1“ a „Input range 2“. Zrušíme zaškrtávacie políčko „Column headings included with data“. Použili by sme ho iba v prípade, ak by sme nad údajmi mali ešte záhlavie.

Položka „Alpha“ je hladina významnosti, napíšeme tam číslo 0,05.

V časti „Options“ zvolíme „Two paired samples“, keďže chceme párový test pre závislé výbery.

V časti „Test type“ zvolíme „Non-parametric“. Ešte je vhodné nastaviť ľavú hornú bunku výstupu v políčku „Output range“ tak, aby bol dostatok miesta pre vypočítané hodnoty a aby nenastalo nevhodné prekrytie s inými bunkami. Ostatné bunky a nastavenia ponecháme bez zmeny. Správne vyplnené dialógové okno je na obrázku 9.5.



Obr. 9.5 Dialógové okno pre Wilcoxonov párový test (zdroj: vlastné spracovanie)

Po stlačení tlačidla OK sa zobrazí tabuľka s výpočtami (obr. 9.6). Najdôležitejší výstup je p-hodnota, ktorá je v tomto prípade 0,0005.

Wilcoxon Signed-Rank Test for Paired Samples		
	Sample 1	Sample 2
median	1035	1009
count	23	
# unequal	23	
T+	30	
T-	246	
T	30	
	one tail	two tail
mean	138	
std dev	32,87096	ties
z-score	3,270364	yates
effect r	0,482189	
p-norm	0,000537	0,001074
p-exact	0,000238	0,000475
p-simul	N/A	N/A

Obr. 9.6 Výpočty k príkladu 9.3 (zdroj: vlastné spracovanie)

Keďže p-hodnota je menšia ako hladina významnosti (0,0005 < 0,05), tak nulovú hypotézu H_0 na hladine významnosti 0,05 **zamietame a prijímame alternatívnu hypotézu**. Zmena ceny ovplyvnila spotrebu domácností. Zistené rozdiely sú štatisticky významné (signifikantné).

Poznámka: Príklad je ilustračný (údaje nie sú skutočné) a cieľom je iba ukázať metodický postup pri testovaní.

Kontrolné otázky:

1. Kedy sa na testovanie hypotéz používajú neparametrické testy?
2. Aké výhody majú neparametrické testy v porovnaní s parametrickými?
3. Neparametrické testy nevyžadujú predpoklad normality, prečo ich teda nepoužívame v každom výskume?
4. V akých situáciách sa používa Wilcoxonov jednovýberový test?
5. Aký parametrický test je možné nahradiť Mann-Whitneyovým testom?
6. Aký parametrický test je možné nahradiť Wilcoxonovým párovým testom?
7. Môže mať charakter údajov vplyv na výber medzi parametrickým a neparametrickým testom?
8. Môže mať počet údajov vo vzorke vplyv na výber medzi parametrickým a neparametrickým testom?

Úlohy na riešenie:

1. Náhodne sme na univerzite vybrali 15 študentov a zistili sme ich znalosti z finančnej gramotnosti. Študenti vo vedomostnom teste získali nasledovné počty bodov: 92; 17; 5; 84; 2; 68; 100; 31; 4; 75; 1; 98; 23; 6; 89. Predpoklady normality nie sú splnené. Testujte na hladine významnosti 0,05, či sa medián získaných bodov signifikantne odlišuje od hodnoty 50, ktorá bola zistená na inej univerzite.

   [p-hodnota = 0,68; nezamietame nulovú hypotézu]

2. Analyzovali sme príjmy zamestnancov s vysokoškolským vzdelaním (VŠ) a stredoškolským vzdelaním (SŠ). Náhodným výberom sme zistili nasledovné príjmy v eurách. VŠ: 1150; 1230; 1300; 2050; 2110; 1170; 1200; 1730; 1850; 1100. SŠ: 850; 900; 1500; 1460; 880; 1410; 1650; 960; 980. Predpoklady normality nie sú splnené. Otestujme na hladine významnosti 0,05, či je medián príjmov u zamestnancov s VŠ vzdelaním signifikantne vyšší ako u zamestnancov so SŠ vzdelaním.

   [p-hodnota = 0,047; zamietame nulovú hypotézu]

3. Skúmali sme zmenu produktivity zamestnancov po zavedení nového softvéru. Náhodným výberom sme získali údaje o 12 zamestnancoch. Počet dokončených úloh so starým softvérom: 15; 12; 21; 16; 35; 40; 39; 15; 24; 21; 19; 16. Počet dokončených úloh tých istých zamestnancov za rovnaký čas s novým softvérom: 17; 15; 22; 15; 32; 42; 30; 15; 25; 23; 20; 20. Predpoklady normality nie sú splnené. Otestujme na hladine významnosti 0,05, či je medián dokončených úloh s novým softvérom signifikantne vyšší ako predtým so starým softvérom.

   [p-hodnota = 0,18; nezamietame nulovú hypotézu]

4. Porozmýšľajte nad témou svojej záverečnej práce a ak je to možné, navrhnite, ako by sa tam dali aplikovať neparametrické testy.

10 KORELAČNÁ ANALÝZA

Kľúčové slová: veda, výskum, prieskum, kvalitatívny a kvantitatívny výskum, znak

V predchádzajúcich kapitolách boli riešené problémy, pri ktorých sme napríklad zisťovali štatisticky významné rozdiely medzi rôznymi skupinami z hľadiska variability, stredných hodnôt a podobne. V štatistike však často skúmame aj vzťahy alebo súvislosti medzi premennými. **Závislosť** dvoch náhodných premenných sa nazýva **korelácia** *(correlation)*. Pomocou korelačnej analýzy môžeme skúmať závislosť dvoch ľubovoľných **kvantitatívnych** premenných. Často je zámerom určiť **silu** (**intenzitu**) tejto závislosti a jej **smer** (či je pozitívna alebo negatívna, resp. priama alebo nepriama).

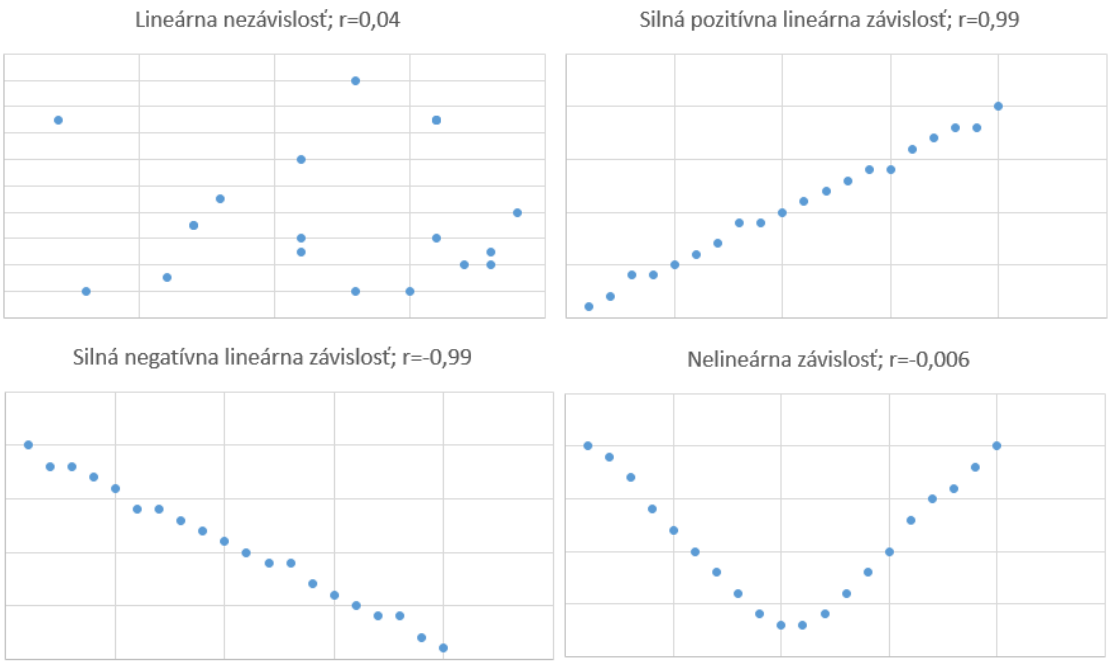
O **funkčnej závislosti** hovoríme vtedy, keď každej hodnote jednej premennej zodpovedá práve jedna hodnota druhej premennej. Funkčnú závislosť môžeme všeobecne zapísať ako $y = f(x)$, kde x je hodnota nezávisle premennej a y je hodnota závisle premennej. Funkčné závislosti sú typické predovšetkým pre prírodné zákony. V sociálnej, pedagogickej alebo aj ekonomickej oblasti sa často stretávame s tzv. **štatistickou (stochastickou) závislosťou**. O nej hovoríme vtedy, keď jednej hodnote danej premennej nezodpovedá vždy len jedna hodnota druhej premennej, ale celý obor hodnôt druhej premennej (Chráska, 2016).

Pozitívna (kladná) **korelácia** nastáva vtedy, keď nárastom hodnoty jednej premennej možno očakávať aj nárast hodnoty druhej premennej alebo keď s poklesom hodnoty jednej premennej možno očakávať aj pokles hodnoty druhej premennej.

Negatívna (záporná) **korelácia** nastáva vtedy, keď nárastom hodnoty jednej premennej možno očakávať pokles hodnoty druhej premennej, resp. keď poklesom hodnoty jednej premennej možno očakávať nárast hodnoty druhej premennej.

Pomocou **korelačnej analýzy** je možné skúmať vzťahy medzi premennými jednak **graficky**, ale aj pomocou rôznych **mier závislostí**. Tie nazývame **korelačné koeficienty**. Korelačným koeficientom kvantifikujeme intenzitu závislosti medzi dvoma kvantitatívnymi premennými. Môže nadobúdať hodnoty medzi –1 a 1 (vrátane krajných hodnôt). Na grafickú analýzu sa najviac využívajú **bodové diagramy**. V ňom každý bod predstavuje obraz dvojice hodnôt na vodorovnej a zvislej osi. Jedná sa o hodnoty dvoch premenných na tých istých objektoch. Z tvaru grafu a rozptýlenosti bodov je možné odhadnúť mieru závislosti premenných. Grafické skúmanie závislostí síce nie je celkom presné a neposkytuje konkrétne výsledky v číselnej podobe, je to však veľmi dobrý úvodný krok na získanie základnej

vizuálnej predstavy o premenných. Na obrázku 10.1 sú štyri príklady grafov rôznych bodových zoskupení s prislúchajúcimi korelačnými koeficientami.



Obr. 10.1 Grafy s rôznymi hodnotami koeficientu korelácie (zdroj: vlastné spracovanie)

Ak existuje medzi premennými korelácia (hoci aj silná), **neznamená to**, že medzi týmito premennými existuje aj **príčinná** súvislosť. Silná korelácia síce často môže odrážať príčinný vzťah, ale nie je to pravidlo. Je teda dôležité si pamätať, že **korelácia nie je to isté ako kauzalita** (príčinnosť). Vhodnosť použitia korelačnej analýzy a skutočné vzťahy a príčiny, je nutné vždy zvážiť na základe logickej analýzy, niekedy je možné overiť kauzalitu pomocou experimentu (manipulácia s nezávisle premennou a sledovanie prípadných zmien druhej (závisle) premennej. Takýto kvalitatívny rozbor je veľmi dôležitý. Často sa totiž stáva, že **meranie závislosti nie je v niektorých situáciách vhodné a zmysluplné**. Môže to mať viacero príčin, na niektoré sa teraz pozrieme podrobnejšie.

Korelačné vzťahy môžu byť ovplyvnené napríklad **nehomogenitou** údajov. Ďalším problémom môže byť tzv. **formálna korelácia**, ktorá vzniká napríklad vtedy, keď skúmame závislosť percentuálnych charakteristík, ktoré sa navzájom dopĺňajú do 100 %. Napríklad korelácia percentuálneho vyjadrenia trhového podielu dvoch firiem, ktoré majú s určitým produktom alebo službou spolu 100 percentný podiel na trhu.

Ďalšia situácia, v ktorej korelačná analýza neprináša zmysluplné výsledky, je **pôsobenie spoločnej príčiny**. Môže tak vzniknúť **zdanlivá korelácia**, či dokonca

nezmyselná korelácia *(nonsense correlation)*. Ak by sme napríklad skúmali konzumáciu kaviáru a počet ľudí, ktorí nosia okuliare, mohli by sme zistiť silnú kladnú koreláciu. V krajinách, kde je vysoká konzumácia kaviáru, nosí veľa ľudí okuliare. Rozhodne to však nedokazuje, že konzumácia kaviáru zhoršuje zrak. Spoločnou príčinou môže byť vyspelá ekonomika (vyšší životný štandard), ktorá zvyšuje u obyvateľov dopyt po luxusných potravinách a zároveň umožňuje lepšiu zdravotnú starostlivosť (vrátane starostlivosti o zrak).

Interpretácia korelačného koeficientu na základe vypočítanej hodnoty môže byť taká, ako je uvedené v tabuľke 10.1. Interpretácia však závisí od viacerých faktorov, napríklad aj od charakteru a rozsahu údajov. Z týchto dôvodov nie je stupnica interpretácie úplne jednotná a v literatúre sa môžeme stretnúť aj s mierne odlišnou verziou.

Tab. 10.1 Interpretácia korelačného koeficientu, (zdroj: vlastné spracovanie)

Hodnota koeficientu	Interpretácia
$\langle \mathbf{0},\mathbf{8};\mathbf{1}\rangle$	Silná pozitívna lineárna závislosť
$\langle \mathbf{0},\mathbf{4};\mathbf{0},\mathbf{8})$	Stredná pozitívna lineárna závislosť
$\langle \mathbf{0},\mathbf{1};\mathbf{0},\mathbf{4})$	Slabá pozitívna lineárna závislosť
$(-\mathbf{0},\mathbf{1};\mathbf{0},\mathbf{1})$	Lineárna nezávislosť (nekorelovanosť)
$(-\mathbf{0},\mathbf{4};\ -\mathbf{0},\mathbf{1}\rangle$	Slabá negatívna lineárna závislosť
$(-\mathbf{0},\mathbf{8};\ -\mathbf{0},\mathbf{4}\rangle$	Stredná negatívna lineárna závislosť
$\langle -\mathbf{1};\ -\mathbf{0},\mathbf{8}\rangle$	Silná negatívna lineárna závislosť

Čím je hodnota korelačného koeficientu bližšie k číslam ± 1, tým je lineárna závislosť silnejšia. A naopak, čím je hodnota bližšie k nule, tým je závislosť slabšia. Ak je hodnota koeficientu korelácie rovná nule, znamená to úplnú lineárnu nezávislosť (nekorelovanosť), ale neznamená to, že by tam nemohol existovať iný druh závislosti (napríklad kvadratická, exponenciálna a podobne). Krajné hodnoty 1 a -1 znamenajú, že všetky body ležia na priamke, závislosť je funkčná.

Okrem korelačného koeficientu je užitočné doplniť výpočet aj o tzv. **koeficient determinácie**. Počíta sa veľmi jednoducho, umocnením korelačného koeficientu na druhú. Koeficient determinácie poskytuje informáciu o tom, akú veľkú časť zmien jednej premennej je možné vysvetliť vplyvom druhej premennej. Zvyšok potom pripadá na iné, neznáme faktory. V rámci korelačnej analýzy je možné za závisle premennú považovať ľubovoľnú z dvoch skúmaných premenných, takže vplyv je prípustné interpretovať aj opačným spôsobom, ak je to zmysluplné.

10.1 Pearsonov korelačný koeficient

Pearsonov korelačný koeficient vyjadruje mieru lineárneho vzťahu medzi premennými. Hodnota korelačného koeficientu vypočítaná z vzorky sa nazýva výberový korelačný koeficient (odhad). Pearsonov (výberový) korelačný koeficient by sme mali používať iba vtedy, ak údaje vo vzorke, ktorú používame na výpočet, pochádzajú zo súboru s **dvojrozmerným normálnym rozdelením a** ak sú tieto údaje **metrické** (intervalové alebo pomerové), prípadne sa môžeme stretnúť s označením kardinálne údaje.

Pri **intervalovom** meraní čísla vyjadrujú rozdiely medzi meranými objektmi. Merané hodnoty nemajú prirodzený nulový bod, nula je definovaná iba arbitrárne. Tieto čísla môžeme sčitovať a odčitovať, ale nemôžeme ich deliť a násobiť. Zisťovanie úrovne vedomostí študentov pomocou didaktických testov môže byť príkladom intervalového merania. Pri **pomerovom** meraní existuje prirodzená nula, ktorá reprezentuje absenciu meranej vlastnosti, a namerané hodnoty je možné ľubovoľne sčítať, odčítať, násobiť a deliť. Výsledky merania môžeme porovnávať na základe otázok „o koľko“ alebo „koľkokrát“. Meranie hmotnosti môže slúžiť ako príklad pomerového merania. Pearsonov korelačný koeficient môžeme v **Exceli** vypočítať pomocou funkcie CORREL alebo je možné využiť doplnok Real Statistics v záložke **Corr**, prvá položka **Correlation Tests**. Tento doplnok ponúka aj výpočet dvoch neparametrických koeficientov a síce Spearmanov korelačný koeficient a Kendallov korelačný koeficient. V tejto publikácii sa nimi však nebudeme bližšie zaoberať. Aj napriek určitým nedostatkom je Pearsonov korelačný koeficient jeden z najpoužívanejších.

10.2 Testy významnosti koeficientov korelácie

Ak počítame koeficient korelácie znakov X a Y zo všetkých prvkov základného súboru, označujeme ho zvyčajne ρ. V praxi sa však často stretávame s výpočtom korelačného koeficientu len z výberového súboru. Písmenom r zvyčajne označujeme výberový koeficient korelácie a písmenom R výberový koeficient poradovej korelácie. Otázkou je, či je možné výsledky výpočtu korelačného koeficientu z výberového súboru **zovšeobecniť** na celý základný súbor. Na zistenie odpovede slúžia **testy významnosti** koeficientov korelácie. Zisťujeme, či vypočítaná hodnota korelačného koeficientu je dostatočná na to, aby sme vzťah mohli považovať za štatisticky významný (systematický). Postupujeme podobne ako pri iných testoch. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \rho = 0$ (korelačný koeficient sa rovná nule, znaky sú lineárne nezávislé).

$H_1: \rho \neq 0$ (korelačný koeficient je rôzny od nuly, existuje lineárna závislosť).

Nedá sa presne povedať, od akej hodnoty korelačného koeficientu je korelácia významná (či už pozitívna alebo negatívna). Závisí to aj od počtu údajov, ktoré máme k dispozícii. So zvyšujúcim sa počtom pozorovaní sa hraničná hodnota znižuje. Hraničnou hodnotou rozumieme takú hodnotu, pri ktorej môžeme povedať, že sa jedná o významný vzťah. Ak je údajov málo, tak výpočty sú nepresné, preto na potvrdenie významnosti potrebujeme silný dôkaz (vysokú absolútnu hodnotu korelačného koeficientu). A naopak, ak máme veľké množstvo dát, tak výpočty sú presnejšie, teda vypočítaný korelačný koeficient sa viac približuje realite a môžeme mu dávať väčšiu dôležitosť aj pri jeho nižších absolútnych hodnotách (Janáček, 2022). Preto pre posúdenie, či je hodnota korelácie významná používame špeciálne štatistické testy významnosti koeficientov korelácie. Doplnok Real Statistics tieto testy ponúka na rovnakom mieste ako výpočet koeficientov korelácie, teda v záložke **Corr**, prvá položka **Correlation Tests**.

Príklad 10.1 Náhodne sme vybrali 12 pozemkov z ponuky realitných kancelárií a sledovali sme na každom pozemku dva údaje. Zistili sme tieto ceny (€): 152 000; 135 000; 85 000; 74 000; 168 000; 215 000; 175 000; 120 000; 95 000; 115 000; 190 000; 150 000. Výmery (plocha pozemkov v m^2) boli nasledovné: 1 500; 1 100; 450; 350; 850; 1 650; 1 355; 980; 400; 580; 930; 740. Predpokladáme, že sú splnené podmienky pre použitie Pearsonovho korelačného koeficientu. Chceme zistiť, či medzi týmito údajmi existuje závislosť a aká je jej sila.

Riešenie: Uvedené dáta prepíšeme do Excelu (alebo ich už tam máme) napríklad do stĺpcov „A“ a „B“. Formulácia nulovej a alternatívnej hypotézy je nasledovná:

$H_0: \rho = 0$ (korelačný koeficient sa rovná nule, znaky sú lineárne nezávislé).

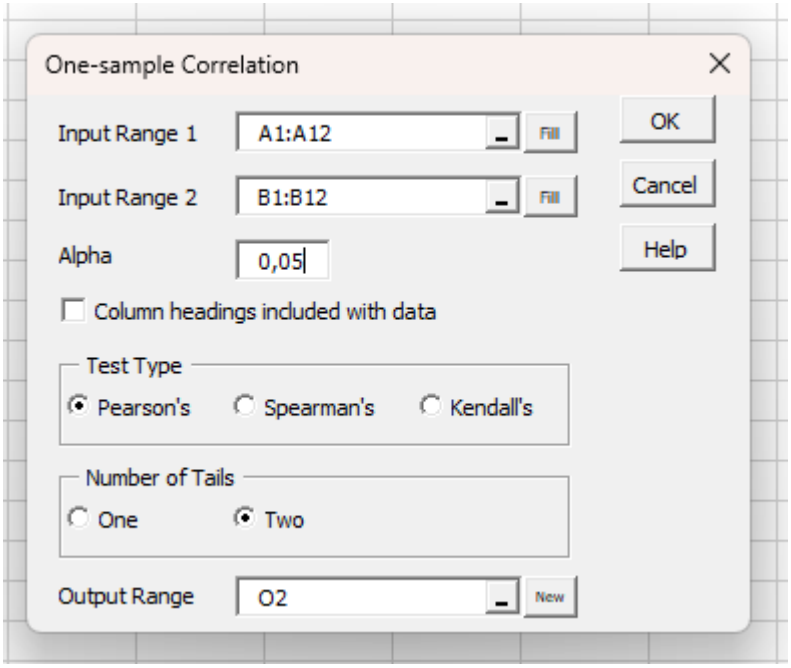
$H_1: \rho \neq 0$ (korelačný koeficient je rôzny od nuly, existuje lineárna závislosť).

P-hodnotu vypočítame pomocou doplnku Real Statistics. V záložke **Corr**, vyberieme prvú položku **Correlation Tests**. V dialógovom okne je potrebné najskôr vložiť vstupné údaje do položiek „Input range 1“ a „Input range 2“. Zrušíme zaškrtávacie políčko „Column headings included with data“. Použili by sme ho iba v prípade, ak by sme nad údajmi mali ešte záhlavie.

Položka „Alpha“ je hladina významnosti, napíšeme tam číslo 0,05.

V časti „Test Type“ zvolíme „Pearsons“, keďže chceme počítať test významnosti pre Pearsonov korelačný koeficient.

Ostatné bunky a nastavenia ponecháme bez zmeny. Správne vyplnené dialógové okno je na obrázku 10.1.



Obr. 10.1 Dialógové okno pre výpočet korelácie (zdroj: vlastné spracovanie)

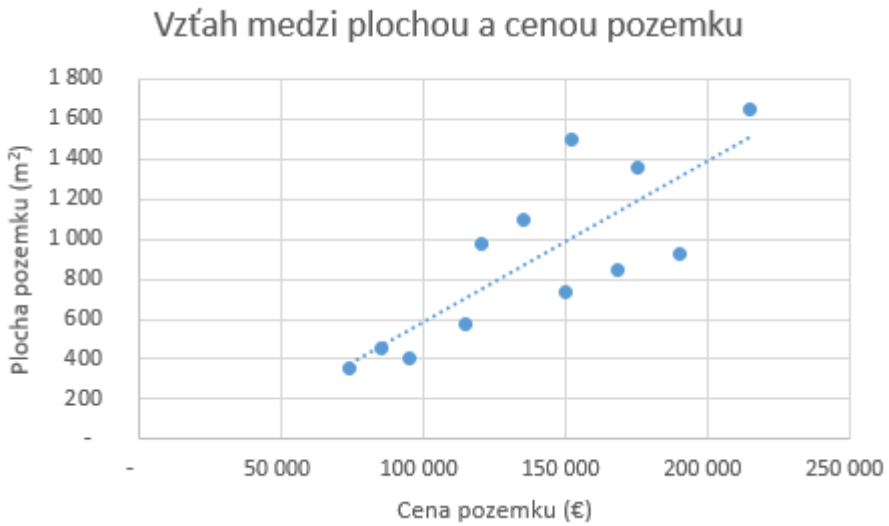
Po stlačení tlačidla OK sa zobrazí tabuľka s výpočtami (obr. 10.2). Najdôležitejší výstup je p-hodnota, ktorá je v tomto prípade 0,00158.

Correlation Coefficients	
Pearson	0,805196
Spearman	0,797203
Kendall	0,666667
Pearson's coeff (t test)	
Alpha	0,05
Tails	2
corr	0,805196
std err	0,187526
t	4,293785
p-value	0,001577
lower	0,387362
upper	1,22303

Obr. 10.2 Výpočty k príkladu 10.1 (zdroj: vlastné spracovanie)

Keďže p-hodnota je menšia ako hladina významnosti (0,00158 < 0,05), tak nulovú hypotézu H_0 na hladine významnosti 0,05 **zamietame a prijímame alternatívnu hypotézu**. Teda cena pozemku a jeho výmera sú lineárne závislé na hladine významnosti 0,05. Potvrdili sme existenciu významného systematického vzťahu medzi cenou a plochou pozemkov. Na obrázku 10.2 vidíme aj veľkosť Pearsonovho korelačného koeficientu. Hodnota 0,805 znamená silnú pozitívnu závislosť. To bol aj očakávaný výsledok. Hoci korelačná analýza nedokazuje príčinné súvislosti, v tomto prípade na základe logickej analýzy môžeme predpokladať, že výmera pozemkov je nezávislá premenná a cena je závislá premenná. Pretože logicky cena závisí od veľkosti pozemku a nie naopak. Ukázali sme teda, že čím väčší je pozemok, tým je vyššia jeho cena.

Veľmi vhodný a názorný je v tomto prípade bodový graf, ktorý môžeme vytvoriť aj s trendovou spojnicou (obr. 10.3).



Obr. 10.3 Bodový graf s trendovou spojnicou k príkladu 10.1 (zdroj: vlastné spracovanie)

Kontrolné otázky:

1. Čo je to korelácia?
2. Ako sa nazýva korelácia, keď nárastom hodnoty jednej premennej možno očakávať aj nárast hodnoty druhej premennej?
3. Ako sa nazýva korelácia, keď nárastom hodnoty jednej premennej možno očakávať pokles hodnoty druhej premennej?
4. Dokazuje silná korelácia existenciu príčinnej súvislosti?

5. Vedeli by ste uviesť príklad na zdanlivú koreláciu?
6. Ako interpretujeme korelačné koeficienty?
7. Na čo slúžia testy významnosti koeficientov korelácie?
8. Je presne stanovené, od akej hodnoty korelačného koeficientu je korelácia štatisticky významná?

Úlohy na riešenie:

1. Náhodne sme vybrali 12 respondentov a u každého sme sledovali dva údaje. Výška (cm): 180; 185; 160; 175; 177; 192; 176; 162; 181; 173; 170; 179. Hmotnosť (kg): 80; 81; 68; 72; 90; 88; 75; 68; 71; 80; 75; 82. Predpokladáme, že sú splnené podmienky pre použitie Pearsonovho korelačného koeficientu. Chceme zistiť, či medzi týmito údajmi existuje závislosť a aká je jej sila.
   [p-hodnota = 0,01; zamietame nulovú hypotézu; korelačný koeficient: 0,7]

2. Náhodne sme vybrali 13 študentov a o každom sme zistili dva údaje. Počet hodín prípravy na skúšku: 5; 0; 12; 65; 30; 25; 15; 70; 24; 15; 38; 40; 50. Počet bodov získaných na skúške: 25; 2; 25; 100; 55; 68; 30; 95; 88; 49; 75; 80; 82. Predpokladáme, že sú splnené podmienky pre použitie Pearsonovho korelačného koeficientu. Chceme zistiť, či medzi týmito údajmi existuje závislosť a aká je jej sila.
   [p-hodnota = 0,00; zamietame nulovú hypotézu; korelačný koeficient: 0,88]

3. Navrhnite príklady, kde by sa dala v ekonomickej oblasti použiť korelačná analýza a porozmýšľajte nad tým, aké riziká hrozia pri interpretácii výsledkov.

4. Porozmýšľajte nad témou svojej záverečnej práce a ak je to možné, navrhnite, ako by sa tam dala aplikovať korelačná analýza.

ZÁVER

Publikácia je určená predovšetkým študentom vysokých škôl, ktorí majú záujem o štúdium ekonomickej štatistiky. Nie je to však jediná cieľová skupina, pre ktorú by mohla byť užitočná. Zaujímavým zdrojom informácií môže byť aj pre odbornú verejnosť, poslucháčov vzdelávacích kurzov a všetkých, ktorí sa pri svojej odbornej činnosti stretávajú s problematikou štatistiky, spracovania a vyhodnotenia dát. Sú tu spracované témy od deskriptívnej štatistiky cez základy pravdepodobnosti až po vybrané časti induktívnej štatistiky. Pre lepšie zapamätanie dôležitých pojmov sa v úvode každej kapitoly nachádzajú kľúčové slová. Na konci sú vždy kontrolné otázky, ktoré slúžia na overenie vedomostí a pôsobia tiež ako námet na zamyslenie nad konkrétnymi aspektmi problematiky.

Osobitný dôraz v tejto publikácii je kladený na praktické využitie a aplikáciu v ekonomickej oblasti. Ku každej téme sú uvedené vzorovo vyriešené príklady, ktoré umožňujú lepšie porozumieť uplatneniu teoretických konceptov v praxi.

Publikácia si nekladie za cieľ poskytnúť vyčerpávajúce informácie z oblasti ekonomickej štatistiky. Táto problematika je veľmi rozsiahla a bolo by ju ťažké obsiahnuť rozumným rozsahom publikácie. Navyše, náročnosť kladená na čitateľa by sa výrazne zvýšila. Preto sú v knihe rozoberané iba vybrané dôležité témy takým spôsobom, aby čitateľ nebol zneistený a odradený priveľkým rozsahom, ale zároveň aby získal základné zručnosti a spôsobilosti pre využitie štatistiky v ekonomickej praxi. Okrem toho bol veľký dôraz kladený na zrozumiteľnosť textov a jednoduchosť pri riešení konkrétnych príkladov. Či sa spomínané ciele podarilo naplniť, necháme na láskavom posúdení čitateľa.

Keďže texty sú primárne orientované na prax, často chýba podrobné odvodzovanie postupov, dokazovanie vzorcov, či vyčerpávajúce definovanie všetkých pojmov, aby boli dosiahnuté spomínané požiadavky na rozsah. Avšak prípadní záujemcovia si tieto informácie môžu ľahko nájsť v inej literatúre, pričom ako inšpirácia môže poslúžiť zoznam bibliografických odkazov na konci knihy. Po dôkladnom preštudovaní tejto publikácie by mal čitateľ porozumieť základným pojmom a princípom v deskriptívnej a induktívnej štatistike a mal by byť schopný samostatne riešiť vybrané praktické problémy z ekonomickej štatistiky s použitím programu Microsoft Excel.

LITERATÚRA

1. Ballová Mikušková, E. (2021) Štatistika pre začiatočníkov. Základy štatistických analýz pre študentov učiteľstva, Nitra: PF UKF v Nitre, 100 s., ISBN 978-80-558-1824-5
2. Byrne, B. M. (2016) Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming, Third Edition (3rd ed.), Routledge
3. Crilly, T. (2011) Matematika 50 myšlienok, ktoré by ste mali poznať, Bratislava: Slovart, 207 s., ISBN 978-80-556-0294-3
4. Čukan, J. a kol. (2017) Metódy a techniky výskumu v kultúre a cestovnom ruchu, Nitra: FF UKF v Nitre, 60 s., ISBN 978-80-558-1151-2
5. Dancey, C., Reidy, J. (2017) Statistics without maths for psychology, Pearson.
6. Ferjenčík, J. (2009) Základy štatistických metód v sociálnych vedách, Košice: Univerzita Pavla Jozefa Šafárika v Košiciach, 186 s., ISBN 978-80-7097-739-2
7. Gavora, P. (2012) Tvorba výskumného nástroja pre pedagogické bádanie, Bratislava: Slovenské pedagogické nakladateľstvo, 105 s.
8. Gavora, P. a kol. (2010) Elektronická učebnica pedagogického výskumu. [online]. Bratislava: Univerzita Komenského. Dostupné na: http://www.e-metodologia.fedu.uniba.sk/ ISBN 978–80–223–2951–4
9. Gáll, J. a kol. (2021) Vybrané štatistické metódy v cestovnom ruchu, Praha: Wolters Kluwer ČR, 104 s., ISBN 978-80-7676-129-2
10. George, D., Mallery, M. (2010) SPSS for Windows Step by Step: A Simple Guide and Reference, 11th edition, Boston: Pearson
11. Giovannini, E. (2010) Ekonomická statistika srozumitelně, Wolters Kluwer ČR, 208 s., ISBN 9788073575366
12. Hendl, J. (2004) Přehled statistických metod zpracování dat, Praha: Portál, 583 s., ISBN 80-7178-820-1
13. Hendl, J. a kol. (2014) Statistika v aplikacích, Praha: Portál, 456 s., ISBN 978-80-262-0700-9
14. Hendl, J., Remr, J. (2017) Metody výzkumu a evaluace, Praha: Potrál, 376 s., ISBN 978-80-262-1192-1
15. Hindls, R. a kol. (2007) Statistika pro ekonomy, 8. vydání, Praha: Professional Publishing, 418 s., ISBN 978-80-86946-43-6
16. Chajdiak, J. (2013) Analýza dotazníkových údajov, Bratislava: STATIS, 108 s., ISBN 978-80-85659-76-4
17. Chajdiak, J. (2010) Štatistika jednoducho, 3. vydanie, Bratislava: STATIS, 194 s., ISBN 978-80-85659-60-3
18. Chajdiak, J. (2013) Štatistika jednoducho v Exceli, Bratislava: STATIS, 341 s., ISBN 978-80-85659-74-0
19. Chajdiak, J. (2009) Štatistika v Exceli 2007, Bratislava: STATIS, 302 s., ISBN 978-80-85659-49-8

20. Chráska, M. (2016) Metody pedagogického výzkumu, Grada Publishing, 256 s., ISBN 978-80-247-5326-3
21. Janáček, J. (2022) Statistika jednoduše, Grada Publishing, 120 s., ISBN 978-80-271-1738-3
22. Janiga, I., Gabková, J. (2016) Základy štatistickej analýzy, Bratislava: STU, 157 s., ISBN 978-80-227-4533-8
23. Jurečková, M., Molnárová, I. (2005) Štatistika s Excelom, Liptovský Mikuláš: Akadémia ozbrojených síl gen. M.R. Štefánika, 234 s., ISBN 80-8040-257-4
24. Klein, D. (2020) Pokročilé štatistické metódy, Košice: Univerzita Pavla Jozefa Šafárika v Košiciach, Vydavateľstvo ŠafárikPress, 136 s., ISBN 978-80-8152-915-3
25. Lyócsa, Š. a kol. (2013) Kvantitatívne metódy v ekonómii II., Košice: elfa, s.r.o., 460 s., ISBN 978-80-8086-210-7
26. Magnello, E., Van Loon, B. (2010) Statistika, Praha: Portál, 192 s., ISBN 978-80-7367-753-4
27. Makovička, J. (2016) Excel pro přírodovědce, Univerzita Karlova v Praze, Nakladatelství Karolinum, 173 s., ISBN 978-80-246-3139-4
28. Marek, L. a kol. (2015) Statistika v příkladech, 2. vydání, Praha: Professional Publishing, 426 s., ISBN 978-80-7431-153-6
29. Markechová, D., Tirpáková, A., Stehlíková, B. (2011) Základy štatistiky pre pedagógov, UKF v Nitre, edícia Prírodovedec, ISBN 978-80-8094-899-3
30. Pacáková, V. a kol. (2015) Štatistická indukcia pre ekonómov a manažérov, Bratislava: Wolters Kluwer, 368 s., ISBN 978-80-8168-081-6
31. Pacáková, V. a kol. (2005) Štatistika pre ekonómov, Bratislava: IURA EDITION, 268 s., ISBN 80-8078-033-1
32. Pavlík, T., Dušek, L. (2012) Biostatistika, Brno: AKADEMICKÉ NAKLADATELSTVÍ CERM, s.r.o., 132 s., ISBN 978-80-7204-782-6
33. Rimarčík, M. (2007) Štatistika pre prax, 200 s., ISBN 978-80-969813-1-1
34. Silverman, D. (2005) Ako robiť kvalitatívny výskum, Bratislava: Ikar, 327 s., ISBN 80-551-0904-4
35. Somorčík, J., Teplička, I. (2015) Štatistika zrozumiteľne, Enigma Publishing, 244 s., ISBN 978-80-8133-042-1
36. Šoltés, E. a kol. (2018) Štatistické metódy pre ekonómov, 2. vydanie, Bratislava: Wolters Kluwer, 368 s., ISBN 978-80-8168-767-9
37. Švábová, L., Ďurana, P., Ďurica, M. (2022) Deskriptívna a induktívna štatistika, Žilinská univerzita v Žiline, 588 s., ISBN 978-80-554-1839-1
38. Terek, M. (2017) Interpretácia štatistiky a dát, 5. vydanie, EQUILIBRIA, 244 s., ISBN 978-80-8143-212-5
39. Terek, M. (2019) Dotazníkové prieskumy a analýzy získaných dát, EQUILIBRIA, 202 s., ISBN 978-80-8143-247-7

40. Tirpáková, A., Malá, D. (2007) Základy štatistiky pre pedagógov, psychológov a sociológov s popisom postupu práce v programe Excel, Nitra: PF UKF v Nitre, 166 s., ISBN 978-80-8094-220-5
41. Tirpáková, A., Markechová, D. (2008) Štatistika v praxi, Nitra: FPV UKF v Nitre, 390 s., ISBN 978-80-8094-283-0
42. Tomšik, R. (2017) Kvantitatívny výskum v pedagogických vedách, Nitra: PF UKF v Nitre, 505 s., ISBN 978-80-558-1206-9
43. Zaiontz, C. (2023) Real statistics using excel, cit. 15.6.2023, dostupné na: www.real-statistics.com

Názov:	**Ekonomická štatistika**
Autor:	Milan Maroš
Vydavateľ:	Univerzita Konštantína Filozofa v Nitre
Edícia:	Prírodovedec č. 862
Schválené:	Edičnou komisiou FPVaI UKF v Nitre dňa 25.10.2024
Obálka:	Ing. Michal Levický, PhD.; obrázok: Clint Post (Pixabay)
Ikony:	flaticon.com
Formát:	B5
Rok vydania:	2024
Miesto vydania:	Nitra
Počet strán:	100
Počet kusov:	elektronická publikácia
Počet AH:	4,85

9 788055 821986

ISBN 978-80-558-2198-6