Použitie niektorých štatistických metód
MUDr. T. Baška, PhD

Úlohou štatistiky pri hodnotení informácií je spracovanie hromadných pozorovaní, ich interpertácia a analýza. Možno ju rozdeliť na deskriptívnu štatistiku, ktorá sa zaoberá spôsobmi opisu, sumarizácie a prezentácie dát a na analytickú (inferenčnú), ktorá umožňuje vytvárať štatistické odhady, t.j. generalizovať údaje získané zo sledovaných súborov na celú cieľovú populáciu. Jej ďalšou úlohou je odlíšiť podiel náhody a zákonitosti na vzniku zistených výsledkov

Deskriptívna štatistika
Jej úlohou je vhodným spôsobom opísať a zosumarizovať zistené údaje. Hoci často nejde o ťažkú úlohu, tvorí základ pre ďalšiu analýzu dát. Jej správne vykonanie je preto podmienkou neskreslenenej a výstižnej interpretácie výsledkov a vyslovenia správnych záverov. Spôsob spracovania dát závisí od toho, akého druhu sú zistené dáta. Podľa charakteru informácie, ktorú v sebe obsahujú, existujú tri základné druhy dát (štatistických veličín):
1. Kvalitatívne (nominálne) veličiny popisujú kvalitatívne vlastnosti zistených javov. Na základe prítomnosti, či neprítomnosti jednotlivých vlastností možno v rámci jednej štatistickej veličiny vždy určiť aspoň dve rôzne kategórie zistení. Napr. štatistická veličina pohlavie môže nadobúdať dve kategórie vlastností - mužské a ženské. Prítomnosť sledovaných príznakov choroby, ako iný príklad štatistickej veličiny obsahuje kategórie: pozitivita a negativita. Sledované kvalitatívne veličiny možno opísať početnosťou ich jednotlivých kategórií. Tá sa najčastejšie vyjadruje vo forme percent, ale aj iných relatívnych ukazovateľov. Napr. v demografii sa používajú často počty ľudí s určitou vlastnosťou (chorí, zomrelí a pod.) prepočítané na 100 000 obyvateľov. Pri uvádzaní relatívnych ukazovateľov, najmä percentuálnych údajov je však dôležité zároveň uviesť aj veľkosť sledovaného súboru, z ktorého boli vypočítané, aby si čitateľ mohol urobiť dojem o spoľahlivosti zistených výsledkov (čím je súbor väčší, tým možno považovať zistený výskyt danej vlastnosti za spoľahlivejší). Preto z dôvodu nízkej spoľahlivosti, ak je súbor menší ako cca 50 jedincov, je zavádzajúce vyjadrovať výsledky v percentách. Pri tak malých počtoch je vhodnejšie udávať absolútne počty.
2. Poradové veličiny možno tiež zatrieďovať do kategórií, ale na rozdiel od kvalitatívnych dát možno tieto kategórie jednoznačne zoradiť do poradia. Hoci možno takýmto spôsobom kategórie číslovať, čísla nevyjadrujú vzdialenosti medzi jednotlivými kategóriami a preto nemá zmysel počítať z nich napr. priemery a pod. Príkladom takejto štatistickej veličiny je klasifikácia nádorov. Jednotlivé štádia nádorového ochorenia, definované svojimi špecifickými kvalitatívnymi vlastnosťami, možno zoradiť do poradia. Je však nezmyselné počítať napr. priemerné štádium, alebo iným spôsobom sa pokúšať „merať“ použité rádové číslovky. Iným príkladom často používaných poradových dát je Apgarovej skóre novorodenca. Takéto dáta možno opisovať tak isto ako kvalitatívne, t.j. jednak absolútnymi počtami, ako aj rôznymi relatívnymi ukazovateľmi, najmä percentami.
3. Kvantitatívne (numerické, číselné) veličiny obsahujú merateľné číselné hodnoty. Ich rozloženie v súbore, resp. populácii možno graficky znázorniť tak, že zvislá os označuje častosť výskytu (t.j. počet jedincov) a vodorovná os predstavuje stupnicu meraných hodnôt danej veličiny. Rozloženie sa takto zobrazuje vo forme tzv. distribučnej krivky. Ak má táto krivka symetrický zvonovitý tvar, matematicky podobný Gaussovej krivke, hovoríme o normálnom rozdelení. Jednou z jeho charakteristík je to, že so vzdialenosťou od stredu krivky klesá početnosť príslušných hodnôt, t.j. čím sa hodnota viac odlišuje od priemeru, tým sa

vzácnejšie vyskytuje. Keďže tvar krivky normálneho rozdelenia je známy, na jeho popis postačujú 2 údaje: ukazovateľ stredu, udávajúci miesto, kde leží stred, príp. vrchol krivky a ukazovateľ rozptylu, udávajúci rozloženie krivky do šírky. Najčastejšie používané ukazovatele stredu:

- Aritmetický priemer. Počíta sa ako súčet všetkých nameraných hodnôt vydelený počtom meraných hodnôt. Ide o najčastejšie používaný ukazovateľ stredu. Jeho výhodou je, že dobre charakterizuje vlastnosti sledovanej cieľovej populácie (je reprezentatívny) a zohľadňuje veľkosť všetkých nameraných hodnôt v súbore.
- Medián. Ak sa všetky namerané hodnoty zoradia podľa veľkosti, medián predstavuje hodnotu ležiacu v strede (ak ide o párny počet hodnôt, medián je priemerom dvoch stredových hodnôt). Na rozdiel od priemeru, medián nezohľadňuje veľkosť hodnôt ležiacich mimo stredu, v dôsledku čoho stráca ako ukazovateľ na reprezentatívnosti.
- Modus je najčastejšie sa vyskytujúca hodnota v sledovanej populácii. Používa sa menej často, pretože nezohľadňuje veľkosti nameraných hodnôt a zo všetkých ukazovateľov stredu ide o najmenej reprezentatívny ukazovateľ. Jeho výhodou je, že ho možno použiť okrem číselných aj u poradových veličín.

Jedným zo znakov normálneho rozdelenia hodnôt v súbore je to, že priemer, medián a modus ležia v jednom spoločnom bode (v skutočnosti blízko seba). Ak sú tieto ukazovatele od seba výrazne vzdialené, znamená to, že rozdelenie veličiny nie je normálne, t.j. jej distribučná krivka nemá symetrický zvonovitý tvar. Túto skutočnosť je potrebné zohľadňovať pri výbere vhodného testu na hodnotenie štatistickej významnosti zistených rozdielov.

Medzi ukazovatele rozptylu patrí:

- Rozsah predstavuje rozdiel medzi najväčšou a najmenšou nameranou hodnotou súboru. Jeho hlavnou nevýhodou je, že výlučne závisí len od dvoch krajných hodnôt a nezohľadňuje ostatné hodnoty. Rozsah možno použiť najmä na popis menších súborov, napr. pri pilotných štúdiách, pokiaľ nie sú prítomné extrémne hodnoty výrazne sa odlišujúce od ostatných
- Štandardná (smerodajná) odchýlka (SD – standard deviation) graficky predstavuje vodorovnú vzdialenosť medzi stredom krivky a bodom, kde sa mení zakrivenie krivky z dutého na vypuklé (bod inflexie). Matematicky je to odmocnený súčet štvorcov odchýlok jednotlivých hodnôt (x) od priemeru ($\bar{x}$)vydelený veľkosťou súboru (n) zmenšenou o jednu:

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}.$$

Štandardná odchýlka je dôležitým ukazovateľom charakteru danej číselnej veličiny súboru. Pomocou nej možno hodnotiť relatívnu veľkosť odchýlok jednotlivých nameraných hodnôt od priemeru, t.j. či ide o bežné hodnoty, ležiace blízko stredu, alebo o odchýlené hodnoty, ktoré sa vyskytujú zriedkavo. Vo vzdialenosti do jednej SD od priemeru leží asi 68% hodnôt. Tieto sa zároveň označujú ako normálne (bežné) hodnoty. Hodnoty ležiace jednu až dve SD od priemeru sú mierne zvýšené, resp. znížené hodnoty. Hodnoty ležiace dve až tri SD od stredu sú abnormálne a hodnoty ležiace viac ako tri SD od priemeru označujeme ako extrémne. Napr. priemerné pH ľudskej krvi je 7,4 so smerodajnou odchýlkou 0,04. Hodnotu pH 7,38 môžeme preto považovať ešte za normálnu. Naproti tomu, pH 7,49 je už hodnotou abnormálnou. SD slúži zároveň ako prostriedok pri počítaní väčšiny štatistických analýz.

- Percentil označuje relatívnu veľkosť meranej hodnoty na základe toho, aké veľké percento tvoria jednotlivci v populácii, u ktorých je hodnota danej veličiny menšia. Ak je distribúcia symetrická, 50. percentil leží v jej strede, pretože od stredu smerom dole leží polovica nameraných hodnôt. Od 25. percentilu má menšie hodnoty 25% cieľovej populácie a 75% má zase väčšie. U 75. percentilu je situácia opačná, t.j. 75% hodnôt je menších a 25% hodnôt je väčších atď. Pomocou percentilu možno označovať relatívnu veľkosť meranej veličiny vzhľadom na populáciu. V praxi sa používa napr. v pediatrii na hodnotenie telesného vývoja

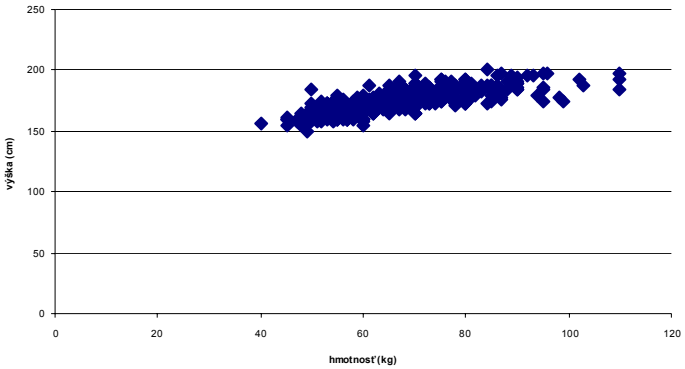
detí – v špeciálnych percentilových diagramoch je možné odčítať percentil zodpovedajúci danej zistenej hodnote (napr. výška dieťaťa) a tým ju posúdiť vzhľadom na celú populáciu detí daného veku.

Grafická a tabuľková prezentácia zistených údajov

Neoddeliteľnou súčasťou štatistického spracovania zistených údajov je prehľadná a názorná prezentácia prostredníctvom tabuliek a grafov. Najmä grafy predstavujú veľmi účinný spôsob prezentácie. Ak sú správne zostavené, môžu výstižne poukázať na dôležité zistenia rôzneho druhu, pomôcť odhaliť niektoré skutočnosti a názorné zobrazenie výsledkov často dodá na ich presvedčivosti a podčiarkne ich význam.

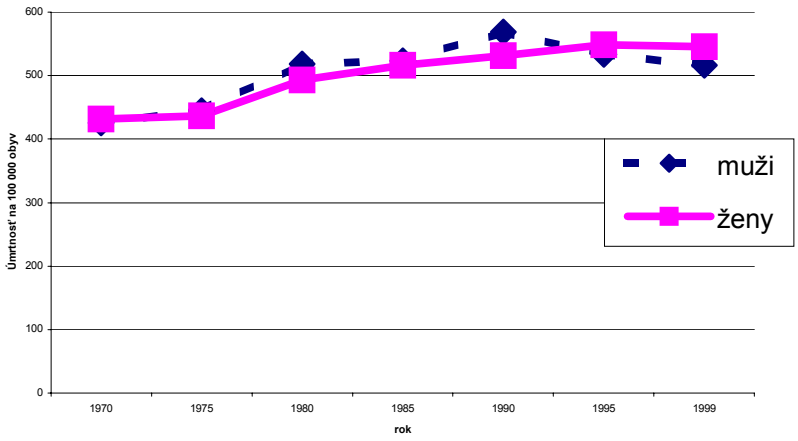
Existuje niekoľko druhov grafov vhodných pre rôzne typy zobrazovaných výsledkov:

1. Bodový graf (Obr. 1) sa používa najčastejšie na prezentáciu vzťahu dvoch číselných premenných (korelácie), kde jedna os predstavuje stupnicu prvej meranej veličiny a druhá – kolmá na ňu – stupnicu druhej meranej veličiny. Každý jednotlivec v súbore predstavuje jeden bod, ktorého polohu na ploche grafu určujú hodnoty oboch meraných veličín. Zoskupenia bodov ukazuje vzájomné asociácie medzi veličinami. Elipsovitý tvar svedčí pre existenciu vzťahu (čím je elipsa užšia, tým je asociácia výraznejšia). Graf možno doplniť aj regresnou priamkou.



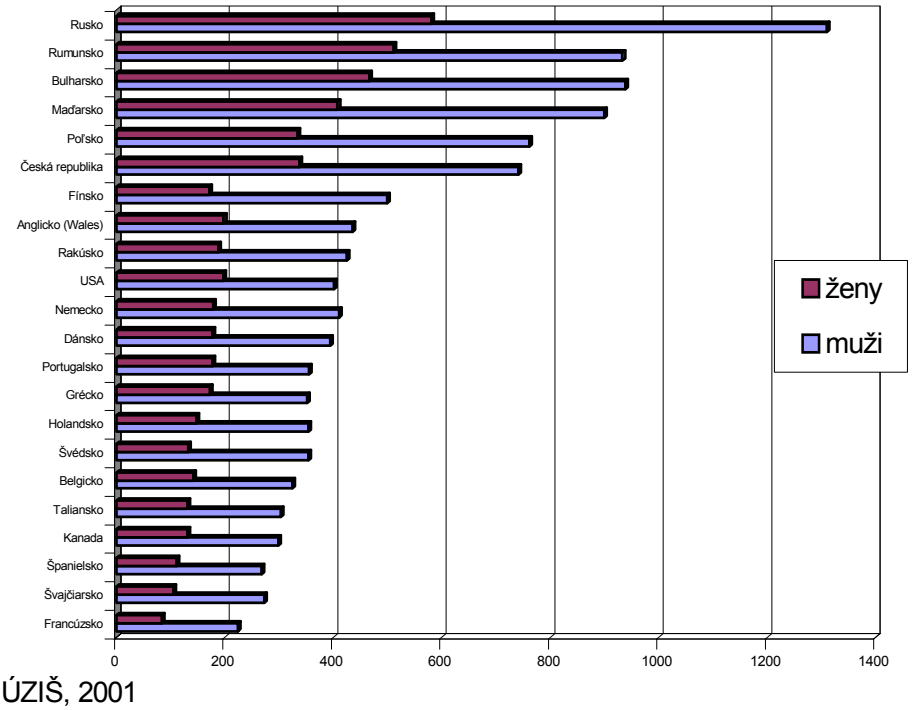
Obr. 1: Vzťah medzi telesnou výškou a hmotnosťou (bodový graf)

2. Spojnicový graf (Obr. 2) sa používa často na vyjadrenie dynamiky zmien sledovanej veličiny v čase. Na vodorovnej osi býva zvyčajne vyznačená časová stupnica a sledovaná veličina na zvislej osi. U číselných (kvantitatívnych) premenných možno pomocou tohto grafu zobraziť výskyt jednotlivých hodnôt premennej (vodorovná os predstavuje hodnoty meranej veličiny a zvislá počet jednotlivcov, u ktorých bola daná veličina meraná)



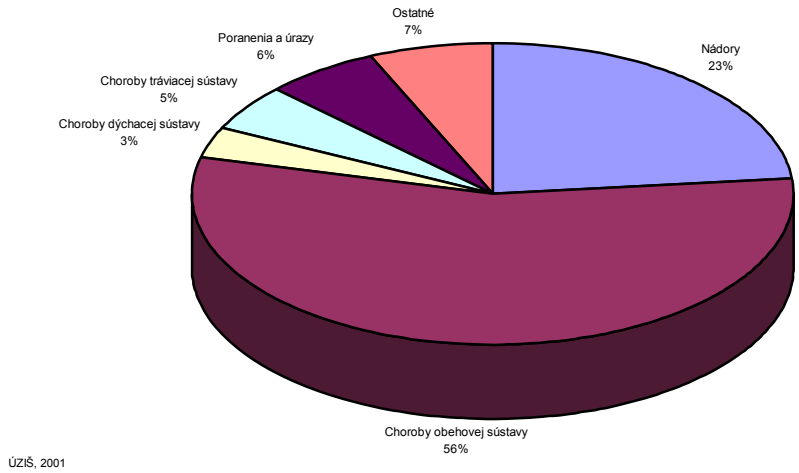
Obr. 2: Kardiovaskulárna úmrtnosť na Slovensku v rokoch 1970-1999 (spojnicový graf)

3. Stĺpcový graf (Obr. 3) sa používa v rôznych obmenách veľmi často. Je výhodný na zobrazovanie početnosti kategórií kvalitatívnych veličín, časových trendov ako aj početnosti jednotlivých intervalov hodnôt číselných (kvantitatívnych) veličín (podobne ako spojnicový graf). Stĺpce je možné doplniť chybovými úsečkami znázorňujúcimi intervaly spoľahlivosti, prípadne smerodajnú odchýlku.



Obr. 3: Úmrtnosť na kardiovaskulárne choroby (na 100 000 obyv.) vo vybraných krajinách (stĺpcový graf)

4. Kruhový (koláčový) graf (Obr. 4) výstižne zobrazuje štruktúru súboru, t.j. relatívnu početnosť jednotlivých podskupín súboru (kategórií, intervalov hodnôt a pod.). Vzhľadom na to, že kruh (t.j. 360°) predstavuje celý súbor, 1% časť súboru zodpovedá 3,6° výseku kruhu.



Obr. 4. Úmrtnosť na Slovensku podľa príčin smrti v roku 1999 (koláčový graf)

Pri používaní grafov platia niektoré všeobecné zásady:

- Každý jeden graf, obrázok, prípadne tabuľka by mali byť označené jasným, a výstižným nadpisom. Mali by obsahovať vhodnú legendu a popisy informujúce o význame jednotlivých položiek tak, aby boli v hlavných rysoch zrozumiteľné aj bez ďalšieho sprievodného textu (samovysvetľujúce).
- V grafe a tabuľke by mala byť uvedená aj veľkosť súboru, najmä ak sa údaje zobrazujú v percentách (ako súčasť nadpisu, legendy a pod.)
- Graf by mal byť čo najjednoduchší, aby bol jasne zrozumiteľný. Pokiaľ by išlo množstvo prezentovaných údajov na úkor zrozumiteľnosti ich znázornenia, je vždy lepšie použiť niekoľko jednoduchých, ako jeden komplikovaný graf. Tabuľka v rámci písaného textu môže byť aj rozsiahlejšia, avšak stále by mala zostávať prehľadná a zrozumiteľná. Na druhej strane, tabuľky ako súčasť prezentácie prednášky by nemali obsahovať príliš veľa údajov, pretože poslucháč (divák) by ich nestíhal sledovať.
- V tabuľkách sa často používajú štandardne používané značky: „x“ – z logických dôvodov údaj nie je možné získať údaj, „.“ – nie sú známe údaje, prípadne sú nespoľahlivé, „-„ – nenašli sa žiadne údaje, „0“ – zistený údaj je menší ako polovica použitej mernej jednotky.

Inferenčná štatistika (analytická štatistika)

Pri zisťovaní jednotlivých veličín máme len málokedy možnosť preskúmať celú populáciu a väčšinou sa spoliehame na výsledky získané zo vzoriek. Tieto však nie sú totožné s výsledkami, aké by sa získali u celej populácie, t.j. sú vždy zaťažené určitou odchýlkou (chybou), ktorá je prejavom náhodnej variability (kolísania) výsledkov. Napr. hoci prieskumy verejnej mienky dokážu dobre odhadovať preferencie jednotlivých politických strán, konečné výsledky volieb sú vždy trochu prekvapením a najmä politici ich vyhlásenie s napätím očakávajú. Z toho dôvodu hovoríme o údajoch získaných zo vzoriek ako o štatistických odhadoch. Ak sú vyjadrené jedným číslom napr. veľkosť priemeru, percentuálny výskyt a pod. ide o tzv. bodové odhady. Ich nevýhodou je, že neumožňujú vyjadriť stupeň presnosti odhadu, resp jeho mieru neurčitosti. Väčšiu informačnú hodnotu majú preto intervalové odhady, ktoré sa vyjadrujú dvomi číslami určujúcimi spodnú a hornú hranicu intervalu. Tento interval vyjadruje mieru neurčitosti, resp. nepresnosti výsledku. Z intervalových odhadov sa najčastejšie používa interval spoľahlivosti 95% (CI 95% - confidence interval 95%) kde hranice intervalu určujú rozpätie, v ktorom môžeme s 95% istotou predpokladať hodnotu platnú pre celú populáciu. Z toho vyplýva, že čím je interval spoľahlivosti užší, tým pokladáme získaný výsledok za presnejší. Šírka intervalu spoľahlivosti u číselných veličín závisí od smerodajnej odchýlky a veľkosti súboru (vzdialenosť hranice CI95% od priemeru $= 1{,}96\frac{SD}{\sqrt{n}}$, SD – smerodajná odchýlka, n – veľkosť súboru) Čím je súbor väčší, tým je interval spoľahlivosti užší.

*Napr. pri meraní body mass indexu v 150 člennom súbore sme namerali jeho priemernú hodnotu 23,5 so smerodajnou odchýlkou 2,8. Priemer je od hraníc intervalu spoľahlivosti 95% vzdialený 1,96*2,8/12,25=0,45; t.j. BMI(CI95%)=23,05-23,95.*

Hodnotenie zistených rozdielov

Pri analýze zistených výsledkov je často potrebné ich vzájomné porovnávanie a hodnotenie rozdielov medzi nimi. V takomto prípade je úlohou štatistickej analýzy rozhodnúť, či sú zistené rozdiely reálne, t.j. dôsledkom odlišných podmienok v porovnávaných súboroch (napr. či pri porovnávaní dvoch súborov pacientov, ktorí dostávali odlišné lieky, boli zistené rozdiely spôsobené odlišnou účinnosťou jednotlivých podávaných liekov), alebo možno tieto

rozdiely pripísať na vrub náhode. V rámci hodnotenia rozdielov možno na začiatku stanoviť dve rôzne hypotézy:

1. Nulová hypotéza, podľa ktorej zistené rozdiely spôsobila len náhoda a možno ich hodnotiť ako prejavy náhodnej variability medzi vzorkami z tej istej populácie.
2. Alternatívna hypotéza odmieta pripisovať zistené rozdiely hre náhody a sú prejavom reálne existujúcej odlišnosti.

O tom, ktorá z hypotéz sa nakoniec prijme, pomôžu rozhodnúť tzv. štatistické signifikančné testy. Sú to matematické algoritmy, pomocou ktorých možno vypočítať pravdepodobnosť náhodného výskytu sledovaného rozdielu. Ak je rozdiel len malý a nevýrazný, možnosť jeho náhodného vzniku je vysoko pravdepodobná. Ak je rozdiel výrazný a nápadný, je málo pravdepodobné, aby vznikol len náhodou. Štatistické testy udávajú výsledok pomocou tzv. p hodnoty, ktorá môže byť v rozsahu 0 až 1 a vyjadruje pravdepodobnosť náhodného vzniku daného rozdielu. Napr. p=0,5 znamená, že na 50% je pravdepodobné, že sledovaný rozdiel vznikol len náhodou. Ak sa p blíži k hodnote 1, pravdepodobnosť vzniku sledovaného rozdielu je veľmi vysoká, prípadne sa blíži k nevyhnutnosti. Na druhej strane, veľmi malé hodnoty p, blížiace sa hodnote 0 znamenajú veľmi malú pravdepodobnosť. Ak je p hodnota vysoká (vysoká pravdepodobnosť náhody), akceptuje sa nulová hypotéza. Ak je p hodnota malá, znamená to aj malú pravdepodobnosť náhody a oprávnenie akceptovať alternatívnu hypotézu. Hraničnou hodnotou určujúcou, či je pravdepodobnosť náhody „malá“ alebo „veľká“, je tzv. α hodnota. Vzhľadom na ňu sa hodnotí veľkosť p hodnoty. U všetkých rozdielov, u ktorých vyšla pomocou signifikančného testu p hodnota menšia ako α, akceptujeme alternatívnu hypotézu a hovoríme, že rozdiel je štatisticky významný (signifikantný). Ak je naopak, p hodnota väčšia ako α, je potrebné prijať nulovú hypotézu a rozdiel označiť za štatisticky nevýznamný (nesignifikantný). Ak je p hodnota blízko hladiny α, možno hovoriť o rozdiele na hranici štatistickej významnosti (napr. p=0,056). Hladina α zároveň určuje veľkosť rizika nesprávne zamietnutej nulovej a mylné prijatie alternatívnej hypotézy, t.j. vyslovenie záveru, že súbory sú odlišné, avšak v skutočnosti pozorovaný rozdiel spôsobila len náhoda (takéhoto omylu sa dopúšťa v praktickom živote ten, kto u výhercu žrebovacej súťaže nechce pripustiť že mal jednoducho šťastie a je presvedčený o nejakom podvode). Takýto druh nesprávneho záveru s nazýva chyba I. typu. Inou možnosťou omylu je zamietnutie alternatívnej hypotézy, keď rozdiel v skutočnosti existuje a nesprávne prijatie nulovej hypotézy. V takomto prípade hovoríme o chybe II. typu. Táto vzniká najčastejšie pri použití malých súborov, keď ich nedostatočná veľkosť neumožňuje jasne odlíšiť rozdiely medzi nimi. Pravdepodobnosť chyby II. typu sa označuje ako β hodnota. S ňou úzko súvisí aj tzv. sila štúdie. Je to pravdepodobnosť správneho prijatia alternatívnej hypotézy, t.j. zistenie prítomnosti rozdielu, ak tento skutočne existuje. Počíta sa ako 1- β. T.j. ak je napr. β=0,1, sila štúdie je rovná 0,9, resp. 90%. Znamená to, že zistený rozdiel s 90% pravdepodobnosťou skutočne existuje. K výpočtu β hladiny je potrebné poznať veľkosť súborov a rozdiel, ktorý sa medzi nimi zistil. Tento výpočet však možno vykonať aj opačne. Ešte pred začatím samotného skúmania si výskumník určí potrebnú silu štúdie (najčastejšie sa používa 90%), ako aj to, aký veľký by mal byť pozorovaný rozdiel (napr. pri skúmaní efektu nového lieku je potrebné vopred určiť, o koľko má byť účinnejší v porovnaní s doteraz používaným liekom, aby bol rozdiel klinicky významný a opodstatnil jeho zavedenie do praxe). Pomocou týchto vstupných požiadaviek možno vypočítať potrebnú veľkosť súborov, aby sa, ak je reálne prítomný, daný rozdiel ukázal ako štatisticky signifikantný. Takto možno na jednej strane ušetriť výskumné prostriedky, aby sa nesledovali zbytočne veľké súbory a na druhej strane eliminovať možnosť, že rozdiel sa nezistí v dôsledku príliš malých súborov.

Výber vhodného štatistického testu závisí od druhu štatistických veličín, veľkosti súborov, distribúcie sledovanej premennej (normálna distribúcia, prípadne iný druh distribúcie), počtu porovnávaných súborov a pod.

V prípade numerických veličín a dvoch porovnávaných súboroch za predpokladu ich normálnej distribúcie sa používa veľmi často Studentov t-test.
Napr. pri hodnotení rozdielu systolického krvného tlaku dvoch súboroch pacientov liečených rôznymi antihypertenzívami možno využiť bežne dostupný program Excel. Jednotlivé namerané hodnoty krv. tlaku vpíšeme pod seba do stĺpcov (každý súbor do iného stĺpca. Cez položku VLOŽIŤ hlavného menu zvolíme možnosť FUNKCIA a vyberieme funkciu TTEST. Ako prvé pole (array) označíme hodnoty v prvom stĺpci a ako druhé hodnoty v druhom stĺpci. Vzhľadom na to, že zisťujeme rozdiel bez ohľadu na to, v ktorom súbore je priemerná hodnota väčšia, zvolíme obojstrannú distribúciu. Pre tento druh výpočtu je vhodný dvojvýberový typ testu (s rovnakými, alebo rozdielnymi rozptylmi). Po správnom zadaní všetkých hodnôt funkcia ukáže priamo hodnotu p. Ak je väčšia ako 0,05, výsledok hodnotíme tak, že nebol zistený štatisticky signifikantný rozdiel priemerného systolického tlaku v súboroch pacientov liečených dvoma rozdielnymi druhmi antihypertenzív. V opačnom prípade hodnotíme rozdiel ako štatisticky významný resp. signifikantný.

V prípade kvalitatívnych a poradových veličín je na porovnávanie vhodný Chí-kvadrátový test. Jeho najjednoduchším variantom je porovnávanie výskytu jedného znaku medzi dvoma súbormi.
Napr. sa porovnával výskyt fajčiarov v súbore mužov a žien. Údaje zistené dotazníkovým prieskumom zapíšeme do tabuľky:

	fajčiari	*nefajčiari*
muži	*43*	*68*
ženy	*31*	*92*

Dôležité je, že do tabuľky možno vpisovať len absolútne počty, nemožno používať percentuálne hodnoty. Druhým krokom je zostavenie tabuľky tzv. očakávaných hodnôt. Ide o hodnoty, aké by sme mohli očakávať v prípade, že by pri zachovaní daných veľkostí súborov a celkového výskytu fajčiarov neexistovali žiadne rozdiely medzi mužmi a ženami. Hodnota pre každé políčko tabuľky sa vypočíta ako súčin všetkých hodnôt v danom riadku (v tomto prípade je to 111 mužov, resp. 123 žien) a v danom stĺpci (v tomto prípade 74 fajčiarov, resp. 160 nefajčiarov) vydelený počtom všetkých sledovaných jednotlivcov (v tomto prípade 234). Dostávame tieto očakávané hodnoty:

	fajčiari	*nefajčiari*
muži	*33,1*	*71,6*
ženy	*36,7*	*79,4*

(Vzhľadom na to, že ide o hypotetické hodnoty, nevadí, že nejde o celé čísla). Na výpočet možno opäť použiť program Excel. Obe tabuľky vpíšeme do príslušných hárkov (na zostavenie tabuľky očakávaných čísel možno využiť počítanie pomocou vložených matematických algoritmov, ktoré tento program umožňuje). Podobne ako v predchádzajúcom prípade zvolíme funkciu CHITEST. Do poľa Actual_Range vložíme tabuľku skutočných počtov a do poľa Expected_Range tabuľku očakávaných počtov. V prípade správneho zadania príslušných údajov funkcia ukáže priamo hodnotu p, v tomto prípade 0,014. Vzhľadom na to, že táto hodnota je menšia ako α=0,05, výsledok hodnotíme ako štatisticky významný, t.j. v daných súboroch bol medzi mužmi signifikantne vyšší výskyt fajčiarov ako u žien.
Jednou z podmienok správneho použitia Chí-kvadrátového testu je, aby ani jedna z očakávaných hodnôt nebola menšia ako 5.
Podrobný popis ostatných testov a podmienky ich použitia je mimo rozsahu tejto kapitoly. Všeobecnou zásadou je použiť najjednoduchšiu metodiku, aká je možné v danom prípade použiť. Výskumníkov často zvádza hľadať pre spracovanie svojich výsledkov tzv. „čo najlepšie“ testy. Zvyčajne sa pod týmto pojmom skrýva snaha použiť taký test, ktorý dá čo najmenšiu p hodnotu, ktorá by umožnila interpretovať výsledky ako pozitívne, t.j. dokázať

existenciu rozdielu. Avšak realitu nemožno znásilniť a prispôsobiť podľa výsledku testu. Aj keby vyšiel pri testovaní nového lieku štatisticky významný rozdiel, jeho farmakologickú účinnosť to nijako neovplyvní. Dobrý test je preto taký, ktorý čo najviac odzrkadľuje skutočnú situáciu, aj keby nebola momentálne pre výskumníka najpriaznivejšia. V každom inom prípade klame výskumník sám seba a odhalenie skutočnej reality, ku ktorému skôr či neskôr vždy dôjde, je o to nepríjemnejšie.

Stanovenie asociácie medzi číselnými a poradovými premennými

V mnohých prípadoch sa sleduje viac ako jedna štatistická veličina a hľadá sa ich vzájomná asociácia (súvislosť), t.j. či sú veličiny na sebe vzájomne závislé.

V prípade dvoch číselných, alebo poradových veličín sa takáto vzájomná závislosť nazýva koreláciou. Jej veľkosť, t.j. mieru, do akej sledované premenné navzájom súvisia, možno počítať pomocou tzv. korelačného koeficientu. Ak ide o dve číselné (kvantitatívne) premenné, používa sa Pearsonov korelačný koeficient. Jeho veľkosť môže byť v rozmedzí od –1 do +1. Ak má korelačný koeficient kladnú hodnotu, ide o pozitívnu koreláciu, čo znamená, že so zvyšovaním hodnoty jednej premennej možno očakávať zvyšovanie aj druhej (Napr. zvyšovanie miery obezity vyjadrenej body mass indexom prináša so sebou aj zvýšenie systolického tlaku krvi). Čím sa viac blíži k jednej, tým je vzťah tesnejší a korelácia silnejšia. Záporné hodnoty korelačného koeficienta znamenajú negatívnu koreláciu, t.j. so zvyšovaním jednej premennej možno očakávať u druhej premennej jej znižovanie. Ak je korelačný koeficient blízko nuly, znamená to neprítomnosť korelácie a premenné sú vzájomne nezávislé.

V prípade poradových dát sa používa Spearmanov korelačný koeficient. Podobne ako v predchádzajúcom prípade, sa jeho veľkosť pohybuje od –1 do +1.

Ak sa berie do úvahy vzájomný vplyv viac ako dvoch číselných premenných, využíva sa mnohonásobná korelácia. Keď chceme zistiť vzťah medzi dvoma premennými pri vylúčení spolupôsobiaceho vplyvu ďalších faktorov, používame parciálnu (čiastkovú) koreláciu.

Pri regresnej analýze sa jedna premenná posudzuje ako závislá premenná a jej hodnota závisí od jednej, alebo viacerých nezávislých premenných. Regresná priamka (v prípade lineárnej regresie), zostrojená na základe tejto závislosti umožňuje interpoláciu, resp. extrapoláciu a tým aj odhadovať jednu premennú na základe inej premennej. Existujú aj regresné modely, ktoré počítajú s nelineárnym vzťahom medzi premennými.

Literatúra:

1. Barker, D.J.P.-Rose, G.: Epidemiology in Medical Practice. 4th Edition, Edinburgh, London, Melbourne and New York, Churchill Livingstone, 1990, 163 s.
2. Dawson-Saunders, B.-Trapp, R.G.: Basic and Clinical Biostatistcs. 2nd Edition, London, Prentice-Hall International, 1994, 344 s.
3. Hill,B.H.-Hill,I.D.: Bradford Hill's Principles of Medical Statistics. 12th Edition, London, Melbourne, Auckland, Edward Arnold, 1991, 339 s.
4. Stallings, M.C., Hewitt, J.K., Beresford, T. et al.: A twin study of drinking and smoking onset and latencies from first use to regular use. Behav Genet, 1999, 29(6), s. 409-421.
5. Ticháček, B.: Základy epidemiologie. 1. vydanie, Galén, Praha, 1997, 230 s.
6. Ústav zdravotníckych informácií a štatistiky: Analýza štandardizovanej úmrtnosti na choroby obehovej sústavy v SR. ÚZIŠ, Bratislava, 2001, 175 s.

MUDr. Tibor Baška, PhD., Ústav verejného zdravotníctva, Oddelenie epidemiológie, JLF UK, Sklabinská 26, 037 53 Martin, e-mail: baska@jfmed.uniba.sk