Základy štatistiky vo vedeckej práci

(toto nie je kompletný študijný materiál, podávajúci vyčerpávajúci prehľad štatistiky, ale obsahuje len definíciu a vysvetlenie vybraných základných pojmov)

MUDr. T. Baška, PhD

Štatistika – veda, ktorá sa zaoberá hromadnými náhodnými javmi, čiže javmi, ktoré nemôžeme presne predvídať
Pravdepodobnosť – miera častosti výskytu daného javu
Deskriptívna štatistika – sumarizácia dát, ich prezentácia a popis
Inferenčná (analytická) štatistika – analýza dát za účelom vyslovenia záverov, prognóz, hodnotení a pod. V zásade ide o: a) odhad charakteristík väčšieho súboru na základe údajov získaných s malej časti tohto súboru (vzorky) – generalizácia; b) oddelenie náhody od zákonitosti, čiže hodnotenie významnosti rozdielov

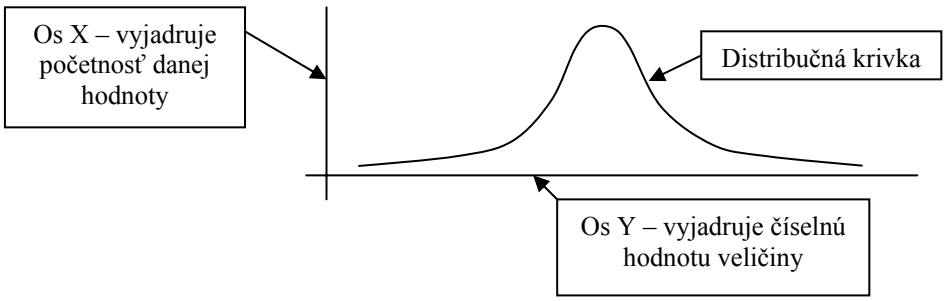
Typy dát (štatistické veličiny):

a) ***kvalitatívne (nominálne)*** – vyjadrujú kvalitu javu a je možné ich zaraďovať do kategórií. Nemajú číselnú hodnotu.
b) ***poradové*** – podobné, ako kvalitatívne, ale kategórie možno zoradiť.
c) ***kvantitatívne (číselné)*** – majú číselnú hodnotu, je možné ich merať. Delia sa na spojité (môžu nadobúdať akúkoľvek hodnotu v rámci určitého intervalu) a nespojité (môžu mať len určité hodnoty, obyčajne sú to celé čísla).

Prezentovanie kvalitatívnych a poradových dát:

- ***proporcie***: p=k/n (p=proporcia, k=počet pozorovaní, kde nastal sledovaný jav, n=počet všetkých pozorovaní); vyjadruje vlastne pravdepodobnosť sledovaného javu
- ***percento***: proporcia*100; percento by sa malo používať len vtedy, ak je počet pozorovaní (n) rádovo 100, minimálne 40-50
- ***relatívne čísla***: incidencia, prevalencia, mortalita, letalita, morbidita a pod.

Prezentovanie kvantitatívnych dát:
Štatistická veličina môže nadobúdať rôzne kvantitatívne hodnoty. Ich početnosť graficky najlepšie vyjadruje ***distribučná krivka***:



Pokiaľ je krivka bilaterálne súmerná, má zvonovitý tvar a podobá sa Gaussovej krivke, hovoríme o normálnej distribúcii, t.j. že veličina je normálne distribuovaná. Takúto distribúciu možno jednoducho definovať pomocou: a)ukazovateľov stredu – udávajú, kde leží stred krivky b) ukazovateľov rozptylu – udávajú, ako je krivka „roztiahnutá“ do šírky.

Ukazovatele stredu:

1. ***aritmetický priemer*** – nemožno ho používať pre poradové dáta.
2. ***medián*** – ak sa všetky pozorované hodnoty zoradia podľa veľkosti, medián je hodnota, ktorá leží v strede (v prípade párneho počtu hodnôt je medián priemer dvoch stredných hodnôt.
3. ***modus*** – najčastejšie sa vyskytujúca hodnota
4. ***geometrický priemer*** – výhodný u logaritmických škál ($GM = \sqrt[n]{(X_1)(X_2)...(X_n)}$; GM – geometrický priemer, n – počet pozorovaných hodnôt, X1...Xn – jednotlivé pozorované hodnoty).

Ukazovatele rozptylu:

1. ***rozsah*** – rozdiel medzi najväčšou a najmenšou hodnotou. Výlučne závisí od dvoch krajných hodnôt, preto jej informačný význam je obmedzaný
2. ***štandardná (smerodajná) odchýlka*** – graficky je to vodorovná vzdialenosť od vrcholu distribučnej krivky po bod inflexie (miesto, v ktorom distr. krivka mení svoje zakrivenie. Je vhodná na definovanie normálnych a abnormálnych hodnôt, t.j. hodnoty, ktoré ležia od priemeru menej ako jednu štand. odchýlku, považujeme za normálne ($s = \sqrt{\frac{\sum (X - \overline{X})^2}{n - 1}}$; s – štandardná odchýlka, $\sum (X - \overline{X})^2$ - súčet štvorcov odchýlok jednotlivých hodnôt od priemeru, n – celkový počet pozorovaných hodnôt).
3. ***x. percentil*** – miesto na distribučnej krivke, od ktorého smerom doľava leží x percent pozorovaných hodnôt (napr. 50. percentil leží v mieste vrcholu distr. krivky, pretože ak je krivka symetrická, od stredu smerom doľava leží polovica, t.j. 50% pozorovaných hodnôt)

Štatistické odhady

Populácia (základný súbor) – skupina ľudí, ale aj iných objektov, ktoré možno definovať niektorými spoločnými vlastnosťami

Vzorka – časť populácie, ktorú skúmame na základe zistených charakteristík odhadujeme charakteristiky celej populácie

Štandardná chyba priemeru – miera odchýlky priemeru vzorky od priemeru celej populácie (miera presnosti odhadu). Pri opakovanom výbere rôznych vzoriek tej istej populácie predstavuje štandardnú odchýlku distribúcie priemerov týchto vzoriek.

Interval spoľahlivosti (CI-confidence interval) – interval hodnôt v ktorom s určitou pravdepodobnosťou leží skutočná hodnota platná pre celú populáciu. Napr. CI 95% znamená, že sa môžeme spoľahnúť na 95%, že hodnota platná pre celú populáciu leží v tomto intervale.

Pozor! - Štandardná odchýlka zisťovaného javu je vlastnosťou skúmanej populácie a nehovorí nič o presnosti merania (nemala by sa výrazne meniť so zväčšovaním vzorky). Na druhej strane, štandardná chyba a interval spoľahlivosti vyjadrujú mieru presnosti merania, nie sú vlastnosťou populácie a sú tým menšie, čím je vzorka väčšia.

Hodnotenie významnosti rozdielov

nulová hypotéza – zistené rozdiely medzi súbormi sú spôsobené len náhodou. V skutočnosti medzi súbormi rozdiel neexistuje.

alternatívna hypotéza – zistené rozdiely sú natoľko veľké, že nepredpokladáme, že vznikli len náhodou. Sú spôsobené tým, že medzi súbormi skutočne existuje rozdiel.

To, či sa prikloníme pri hodnotení k nulovej, alebo alternatívnej hypotéze závisí od toho, aká veľká je pravdepodobnosť, že daný rozdiel vznikol len náhodou. Túto pravdepodobnosť náhody je možné vypočítať pomocou tzv. ***štatistických signifikančných testov***. Tieto dnes existujú najčastejšie vo forme počítačových programov. Zadajú sa vstupné dáta a počítač vypočíta tzv. p hodnotu. Ak je p hodnota veľká, akceptuje sa nulová hypotéza a odmieta alternatívna (pravdepodobnosť náhody je príliš veľká). Ak je p hodnota malá, akceptuje sa alternatívna a zamieta nulová hypotéza (je veľmi málo pravdepodobné, že daný rozdiel vznikol len náhodou)

Hranica, ktorá určuje, či je p hodnota malá, alebo veľká, sa nazýva ***hladina štatistickej signifikantnosti (alfa hodnota).*** Najčastejšie sa používa alfa hladina p=0,05. Čiže, ak je pravdepodobnosť náhody vypočítaná signifikančným testom menšia ako 0,05, napr. p=0,0002, akceptujeme alternatívnu hypotézu a hovoríme, že rozdiel je štatisticky signifikantný (významný). Naopak, ak je napr. vypočítané p=0,242, akceptujeme nulovú hypotézu a hovoríme, že rozdiel je nevýznamný.

Vypracoval: MUDr. Tibor Baška, PhD., Ústav verejného zdravotníctva JLF UK, Odd. epidemiológie, Sklabinská 26, 037 53 Martin, tel.: 043/4134715, e-mail: baska@jfmed.uniba.sk