StatSoft

Typy dat pro statistické šetření

V tomto bloku se budeme věnovat statistickému zpracování dat pěkně od začátku. Začněme tedy daty.

Statistické zpracování dat mělo vždy svůj význam ve vědě, lékařské praxi a některých technických oblastech. Pro ostatní obory nebyl přínos příliš hmatatelný, protože vše se točilo na úrovni teorie. Chyběla totiž reálná data, data z praxe. S rozvojem informačních technologií se situace postupně měnila, velký výpočetní výkon umožnil relačním databázím v reálném čase zpracovávat velké množství dotazů a náklady na uložení velkého množství dat se také postupem času snížily. Firma tak může zálohovat téměř vše, co se jí zamane. Flash disky jsou dnes rozdávány zadarmo jako reklamní předměty a jejich velikost je v řádech stovek až tisíců větší než kapacita pevného disku kdysi běžného počítače 486. S jakými typy uložených dat se tedy můžeme setkat?

Typy dat pro statistické zpracování

Na data se lze dívat z různých úhlů pohledu. Marketingové šetření rozděluje data podle jejich vzniku na softdata a harddata. V jednom pojetí bude pojem primární a sekundární data mít stejnou interpretaci a v jiném nikoliv. Ponechme stranou marketingové a další business pohledy a zaměřme se na dělení, které je podstatné pro volbu přístupu k jejich zpracování, resp. pro volbu statistických testů a technik. V těchto základech se totiž často chybuje a výsledné statistiky potom nemají žádnou vypovídací hodnotu. Co že tím myslíme?

Máme-li v datech reprezentován stupeň vzdělání číslem, má smysl počítat průměr z této proměnné? Má smysl počítat průměrnou dojivost a nerozlišovat přitom jednotlivá plemena?

$$\text{ø Kráva} = \frac{\textit{Hereford} + \textit{Highland} + \textit{Galloway} + \textit{Česká straka}}{50} = ?$$

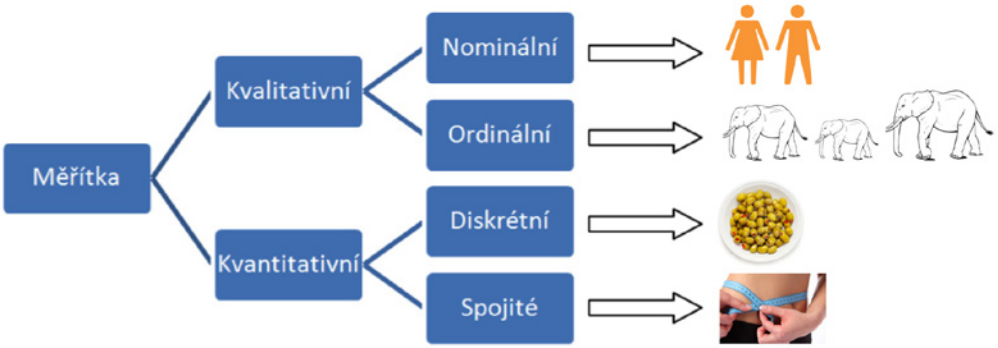
Průměrný počet dětí v daném souboru už je lepší, tabulka četností by však byla zajímavější. Z hlediska statistického zpracování dat hovoříme o statistické jednotce a statistickém znaku. Na statistické jednotce měříme různé statistické znaky – vzdělání, pohlaví, plat, tlak, rozměry, počet chyb apod.

Statistickou jednotkou může být tedy pacient, sklo do brýlí ve výrobě, anebo náhodně vybraný vzorek limonády, na kterém měříme kyselost a další parametry, které určují její vlastnosti.

	List1						
	1 Pohlaví	2 Vzdělání	3 Výška	4 Počet dětí	5 Sport (h/m)	6 Popis	7 Body 1. kolo
1	Muž	1	177	0	12,3	dobrá kondice	25
2	Žena	3	161	3	9,5	sportovec	33
3	Žena	3	162	1	14	sportovec	53
4	Žena	3	180	1	4,2	nesportovní	12
5	Muž	3	168	2	14,9	pod kondicí	4
6	Žena	1	170	2	12,6	dobrá kondice	28
7	Žena	1	162	0	12,8	dobrá kondice	24
8	Žena	3	158	2	15,6	sportovec	48
9	Muž	2	181	5	11,2	nesportovní	5
10	Muž	1	182	2	11,6	sportovec	56

Statistická jednotka

Pro statistické zpracování se měřítka dat v základu dělí na kvalitativní a kvantitativní, další dělení zachycuje následující obrázek:



Nominálním znakem je například plemeno skotu, fakulta (technická, ekonomická…), odrůda brambor, jde tedy o názvy kategorií. Ordinální veličiny (pořadové) jsou na první pohled podobné nominálním, měli bychom však být schopni je nějakým způsobem seřadit, např. stupeň kouření lze seřadit podle intenzity (*silný kuřák* > *střední* > *slabý* > *občasný* > *nekuřák*). Mezi další ordinální proměnné patří například vzdělání, intenzita bolesti atd. Jak si poradíme s proměnnou pohlaví – patří také do této kategorie? Někteří se mohou domnívat, že ano. Z hlediska statistického zpracování dat však tuto proměnnou zařadíme do kategorie Nominální. Pokud můžeme u ordinální proměnné navíc počítat, o kolik je jedna hodnota větší/menší než předchozí, lze hovořit o proměnné rozdílové, případně podílové, přidáme-li možnost spočítat, kolikrát je hodnota větší/menší (počet zaměstnanců v pobočkách). Tyto dvě proměnné jsou v praxi souhrnně definovány jako numerické, resp. kvantitativní. Pro naše potřeby budeme tyto proměnné dělit na diskrétní, které nabývají pouze celočíselných hodnot (počet dětí, počet chyb stroje, počet válců automobilu), a spojité (metrické), které nabývají libovolných hodnot (věk, příjem, teplota, cena).

V tomto díle se bavíme o typech dat v našem vzorku. Výsledky analýzy jsou jen tak dobré, jak dobrý je samotný vzorek, proto je tato fáze velmi důležitá. Reprezentativnost a náhodnost vzorku spolu s jeho velikostí patří mezi klíčové faktory ovlivňující věrohodnost našeho závěru. Statistika není schopna činit závěry o jevech, které vzorek neobsahuje. Před vlastním sběrem dat je nutné zformulovat otázky, na které hledáme odpověď, a určit cílovou skupinu výzkumu. Nedostatečnost vzorku je nejčastější chybou při statistickém šetření, někdy však nemáme na výběr a musíme se spokojit i s velmi omezeným vzorkem (například v lékařské praxi). V příští části se podíváme na to, jak udělat přehled o našem souboru dat, tedy jak vystihnout a co nejlépe popsat vlastnosti konkrétního vzorku dat, budeme se tak věnovat popisné statistice. Nebudeme zobecňovat na populaci, nebudeme testovat, ale vhodně aplikujeme charakteristiky polohy a variability na různé typy dat.

StatSoft CR s.r.o.
Ringhofferova 115/1, 155 21 Praha 5 – Zličín
t +420 233 325 006, **f** +420 233 324 005
e info@statsoft.cz