I Typy dat

Data, se kterými pracujeme, tedy realizace nějaké náhodné veličiny, mohou mít různý charakter, který nám omezuje možnosti, jak s nimi dále pracovat. Klíčová je informace, zda-li hodnoty umíme uspořádat a měřit mezi nimi vzdálenost.

- Kvalitativní (měkká) data – tato data můžeme rozdělit do konečně mnoha kategorií, ale nemá smysl kategoriím přiřazovat číselné ohodnocení. Dále je dělíme na :
  - Nominální data – Kategorie mezi sebou nelze porovnávat, např. pohlaví jedince, fakulta, kterou člověk studuje nebo barva vlasů.
  - Ordinální data – Kategorie umíme uspořádat (nejvyšší dosažené vzdělání, stupnice „souhlasím - spíše souhlasím - spíše nesouhlasím - nesouhlasím“ nebo známkování A-F.
- Kvantitativní (tvrdá) data – realizace jsou čísla. Opět je můžeme rozdělit na:
  - Diskrétní data – nabývají pouze celočíselných hodnot, např. počet dětí v rodině, počet zaměstnanců ve firmě.
  - Spojitá data – mohou nabývat všech hodnot z nějakého intervalu, např. výška člověka, cena výrobku, čas ...

II Tabulka rozdělení četností

Obsahuje všechny informace z dat zaznamenané tak, aby se v nich dalo jednoduše orientovat. Data nejprve vytřídíme, a pak je zaneseme do tabulky. Ta podle typu dat může mít různé počty sloupců. Nejjednodušší je tzv. prosté třídění, které používáme, pokud jsou data kvalitativní nebo diskrétní s malým počtem tříd. Výsledná tabulka může vypadat např. takto:

Počty studentů jednotlivých fakult MU v roce 2015/2016		
Fakulta	počet studenů	relativní četnost
Právnická fakulta	3267	0,102
Lékařská fakulta	4529	0,142
Přírodovědecká fakulta	3526	0,110
Filozofická fakulta	7422	0,232
Pedagogická fakulta	4607	0,144
Ekonomicko-správní fakulta	2589	0,081
Fakulta informatiky	1797	0,056
Fakulta sociálních studií	3063	0,096
Fakulta sportovních studií	1214	0,038
součet	32 014	1

Třídy (kategorie) v této tabulce jsou jednotlivé fakulty – právnická fakulta, lékařská fakulta.... Absolutní četnosti se značí n_i a v tomto příkladě jsou to počty studenů jednotlivých fakult, $n_1 = 3267$ znamená, že právnicá fakulta má 3267 studentů. Celkový součet $n = 32014$ je počet všech studentů MU. Poslední sloupec, relativní četnosti, se značí r_i. Udávají, jakou

část z celkového počtu studentů MU má i-tá fakulta. Spočítáme je ze vztahu $r_i = \frac{n_i}{n}$, může se uvádět i v procentech.

Jednotlivé třídy řadíme buď abecedně, podle absolutních četností nebo existuje-li zažité pořadí (dny v týdnu, měsíce…), jako v tomto případě.
Pokud data dokážeme seřadit (jsou ordinálního nebo kvantitativního typu), má smysl počítat kumulativní četnosti.

Tabulka rozdělení četnosti známek BPM_STA2, jaro 2015

Známka	Absolutní četnost	Kumulativní četnost	Relativní četnost	Kumulativní relativní četnost
A	7	7	0,025	$\frac{7}{282} \doteq 0,025$
B	43	7+43=50	0,152	$\frac{50}{282} \doteq 0,177$
C	52	50+52=102	0,184	$\frac{102}{282} \doteq 0,362$
D	67	102+67=169	0,238	$\frac{169}{282} \doteq 0,599$
E	61	169+61=230	0,216	$\frac{230}{282} \doteq 0,816$
F	52	230+52=282	0,184	$\frac{282}{282} = 1$
součet	282	-	1	-

Zde kumulativní četnosti udávají, kolik studentů dostalo danou známku nebo lepší, relativní udává podíl studentů, kteří dostali danou známku nebo lepší v celém ročníku. Kdybychom kategorie dat nedokázali seřadit, nemělo by smysl uvažovat „lepší“ známku.

Pokud jsou data diskrétní, můžeme tabulku četností seřadit stejně, jako v předchozím případě. Pokud je ale tříd velké množství, byla by tabulka nepřehledná – budeme-li se respondentů ptát na počet sourozenců, dostaneme odpovědi „0“,„1“,…,„5 a více“. Pokud nás ale bude zajímat věk, můžeme získat i několik desítek tříd. Proto je přehlednější některé z těchto tříd sloučit do intervalů. Mějme například tato data (celkem 250 záznamů) udávající věk respondenta:

22 66 18 36 24 22 25 56 57 65 21 31 48 51 62 68 34 30 44 43 20 49 45 23 41 18 52 25 30
28 61 28 61 61 44 21 49 30 47 19 57 40 25 28 38 26 62 33 32 46 29 64 38 47 25 63 30 20 34 59 44
29 47 69 29 39 22 53 20 25 49 43 63 18 45 21 22 66 55 20 46 20 24 46 22 50 46 53 28 61 26 26 25
32 39 58 48 29 47 40 32 20 61 26 52 28 40 70 35 30 36 23 21 50 40 25 30 68 19 42 49 35 23 25
31 70 65 19 26 61 34 18 28 36 67 20 32 24 27 28 53 51 25 18 28 52 28 21 29 35 63 28 69 60 28
59 38 24 56 40 20 54 19 20 53 30 28 70 25 30 20 27 21 53 23 38 64 27 30 22 49 33 23 66 24 26
46 26 33 27 42 20 24 46 44 23 30 45 30 20 53 21 68 58 25 61 24 26 22 42 62 70 20 19 29 49 26
27 46 27 30 24 29 26 63 53 55 36 68 26 30 34 30 51 20 30 56 23 27 49 54 47 30 28 39 67 54 30 56 66

Mohli bychom každé z tříd 18 až 70 let vyhradit vlastní řádek, ale to by bylo velmi nepřehledné. Proto je výhodnější vždy spojit několik tříd, které se stanou intervaly. Musíme ale určit počet intervalů a jejich šířku. Pro počet intervalů k můžeme použít některou z formulí:

- Sturgesovo pravidlo: $k \approx 1 + 3,3 \log n$.
- Jednoduché (odmocninové pravidlo): $k \approx \sqrt{n}$
- Vlastní, podle potřeby a účelu zpracování dat: věk po desítkách, ceny zboží po stokorunách apod., cílem je přehledost. Také je třeba zvážit, jak budeme dále s daty pracovat.

Máme-li určen počet intervalů, určit jejich šířku je snadné - šířka intervalu se spočítá následovně

$$\text{šířka intervalu} = \frac{\text{maximální hodnota} - \text{minimální hodnota}}{\text{počet intervalů } k}$$

Důležité je, aby intervaly na sebe navazovaly, nepřekrývaly se, pokrývaly všechny hodnoty a byly, pokud možno, stejně široké. Pokud data obsahují extrémní hodnoty(objevil by se nám v datech např. věk 95 let), pak je můžeme schovat do krajních intervalů formulací „65 a více“. Jak tedy uspořádat naše data do přehledné tabulky? Určíme počet intervalů – Sturgesovo kritérium nám dává $k \approx 1 + 3,3 \log 250 \doteq 9$ intervalů, jejich šířka by pak byla $\frac{70-18}{9} \approx= 5,8$. Pro srovnání, prostým odmocninovým kritériem bychom pracovali s $k \approx \sqrt{250} \doteq 16$ intervaly o šířce $\frac{70-18}{16} \doteq 3,25$ roku. Jako kompromis můžeme zvolit šířku intervalu 5 let, se kterou se bude dobře pracovat.

Protože je délka intervalu 5 let, nejmenší věk v datech 18 let a nejvyšší 70 let, nemohou být intervaly stejně široké. První interval může být buď $\langle 18, 23)$, čímž by poslední interval byl $\langle 68, 73\rangle$, nebo můžeme první interval posunout tak, aby končil ve 20, čímž bychom získali hezké uspořádání. První interval tedy může být i $\langle 18, 20)$ nebo $\langle 15, 20)$, podle původu dat – mohly se v datech objevit i osoby mladší 18 let? Pak má smysl uvažovat interval $\langle 15, 20)$. Pokud ne (data byla sbírána pouze mezi dospělou populací), je vhodnjší použít menší interval $\langle 18, 20)$ – použitím většího by mohlo dojít ke zkreslení. Při práci s daty je třeba vždy mít na paměti jejich původ a účel, za kterým je zpracováváme. Výsledná tabulka vypadá následovně:

Věk respondenta					
Věk	Střed intervalu	Absolutní četnost	Kumulativní četnost	Relativní četnost	Kumulativní relativní četnost
$\langle 18, 20)$	19	10	10	0,040	0,040
$\langle 20, 25)$	22,5	43	53	0,172	0,212
$\langle 25, 30)$	27,5	49	102	0,196	0,408
$\langle 30, 35)$	32,5	30	132	0,120	0,528
$\langle 35, 40)$	37,5	14	146	0,056	0,584
$\langle 40, 45)$	42,5	15	161	0,060	0,644
$\langle 45, 50)$	45,5	24	185	0,096	0,740
$\langle 50, 55)$	52,5	18	203	0,072	0,812
$\langle 55, 60)$	57,5	12	215	0,048	0,860
$\langle 60, 65)$	62,5	17	232	0,068	0,928
$\langle 65, 70\rangle$	67,5	18	250	0,072	1
součet	282	-	1689	-	1

Tato tabulka je přehledná a použitelná do prezentace (při obhajobě, v práci...) nebo do textu (článku, bakalářky...).