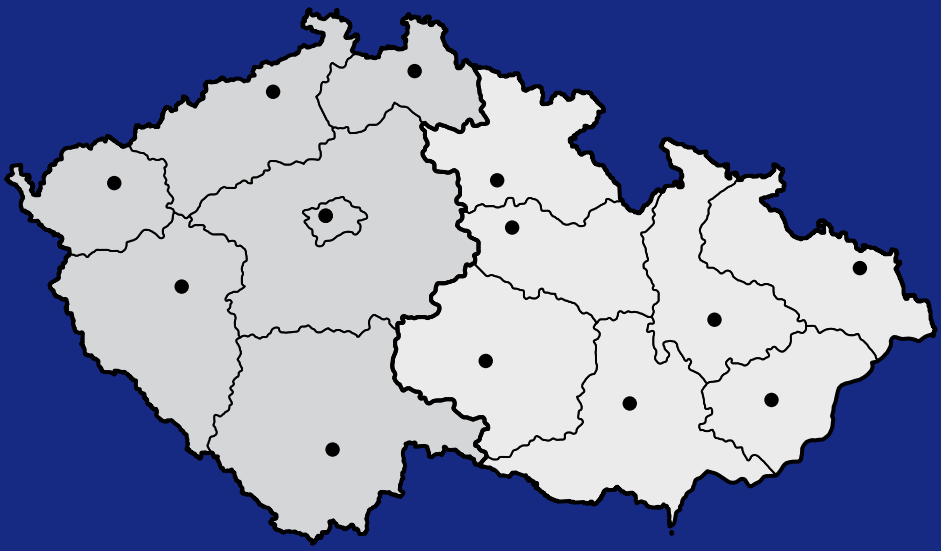
Biomedicínská statistika

IV.

ZÁKLADY STATISTIKY V PROSTŘEDÍ **R**

Karel Zvára



EuroMISE

Biomedicínská statistika IV.
Jana Zvárová (editor)

Základy statistiky v prostředí R
Karel Zvára

———

Recenzovali:
prof. RNDr. Jiří Anděl, DrSc.
RNDr. Patrície Martinková, Ph.D.

Obálka Anna Schlenker
Sazba pomocí LaTeX Karel Zvára
1. vydání

ISBN 978-80-246-2245-3
ISBN 978-80-246-2447-1 (online : pdf)

Univerzita Karlova v Praze
Nakladatelství Karolinum 2013

http://www.cupress.cuni.cz

Obsah

Předmluva

Text je určen studentům přírodovědecké fakulty UK a nejen jim. Vychází z dlouholetých přednášek pro biology, studenty učitelství, v posledních letech pro geografy, demografy... Doufám, že bude užitečný také studentům lékařských fakult a doktorandům biomedicínských oborů. Různé části knihy jsou nestejně obtížné, ale věřím, že si studenti dokáží najít to, co je pro jejich studium důležité. První kapitola je věnována popisným statistikám, druhá pak základním pravděpodobnostním pojmům. Za nejobtížnější, ale svým způsobem nejdůležitější, považuji třetí kapitolu, jejímž úkolem je přiblížit princip statistického uvažování. Zbývající kapitoly uvádějí nejběžnější statistické úlohy a metody k jejich řešení. Příloha A je určena jako minimální úvod těm, kteří s R právě začínají. Tento úvod je možno doplnit některým z manuálů, které lze nalézt na internetu, především na některém ze zrcadel CRAN (viz přílohu A). Příloha B stručně popisuje data, s nimiž se v knížce pracuje a která jsou umístěna na přiloženém cédéčku. Věřím, že toto cédéčko usnadní práci s knížkou. Jsou na něm po jednotlivých kapitolách uloženy všechny erkové výpočty uvedené v textu, v adresáři `data` jsou uložena všechna data.

Vzorečky, na které si někteří studenti stěžují, jsem se snažil omezit. Pravda, k vlastnímu počítání dnes tak často vzorečky nepotřebujeme, počítače je zpravidla znají spolehlivěji, ale vzorečky také umožňují stručný a přesný zápis mnoha myšlenek, principů, algoritmů... Kde vidím možnost, tam takový vzoreček doprovázím podrobným, a jak doufám, snad také srozumitelným vysvětlením.

Text je prostoupen ukázkami výpočtů v prostředí R s erkovými příkazy. Je to pomůcka, kterou vřele doporučuji. Pravda, seznamování se s tímto programem je poněkud náročnější, než třeba začátek práce v Excelu. Student, který se nebojí samostatného myšlení, po čase začne komunikovat s prostředím R zcela samozřejmě. Toto prostředí umožní svému uživateli logicky

hledat odpovědi na položené otázky. Znalost angličtiny práci v tomto prostředí usnadní. Další pomůckou při práci v prostředí R bude, jak doufám, speciální rejstřík erkových funkcí použitých v této knize. Nicméně, nepsal jsem příručku programu R, psal jsem knížku o statistice.

Výklad je doprovázen řadou příkladů, které jsou v každé kapitole průběžně číslovány. Mnohé spolu souvisejí, navazují na sebe, což se pozná podle stručného označení řešené úlohy, které je uvedeno vždy za číslem příkladu. Tato označení lze nalézt také v rejstříku. Doporučuji příklady při studiu nepřeskakovat. Naopak, jsou tam uváděny interpretační komentáře důležité pro osvojení základů statistického myšlení. Studium příkladů je užitečné i tehdy, když si čtenář výstupy z R jen prohlédne. Konec příkladu čtenář snadno rozpozná podle symbolu ◯ umístěného na jeho konci. Podobně poznámky, které obsahují další vysvětlení, jsou odlišeny menším písmem.

Ještě upozornění určené zejména těm, kteří znají moji podobně zaměřenou knížku nazvanou Biostatistika (Zvára, 1998), která pak vyšla v mírně upravené verzi během deseti let ještě několikrát. Místo tam používaných kritických hodnot používám zásadně kvantily, takže například místo 2,5% kritické hodnoty standardního normálního rozdělení rovné 1,96 je tu hodnotou identický 97,5% kvantil téhož rozdělení. Vysvětlení obou těchto pojmů najde čtenář na str. 73.

Děkuji oběma recenzentům textu, profesoru Jiřimu Andělovi, DrSc, a RNDr. Patrícii Martinkové, Ph.D., za pečlivé přečtení rukopisu, za upozornění na nedopatření v něm obsažená a zejména za řadu podnětů k jeho zlepšení. Podobně děkuji docentu RNDr. Josefu Ježkovi, CSc., který mi kolegiálně, bez oficiálního pověření, pomohl při konečných úpravách textu. Dík patří i docentu RNDr. Martinu Ouředníčkovi, PhD., za náměty týkající se odstavce 1.4.

Na závěr se omlouvám za všechny překlepy a jiná nedopatření, které přes veškerou snahu v textu zůstaly. Prosím laskavé čtenáře, aby mne na takové případy upozornili, abych se jich mohl příště vyvarovat.

V Praze a v Českém Dubě v lednu roku 2013

Karel Zvára

Kapitola 1

Popisné statistiky

1.1 Měřítka

Nejprve si ujasníme, čím se budeme zabývat. Máme nějaký soubor **statistických jednotek** a u každé z nich zjišťujeme jeden či několik údajů. Jednotlivým údajům říkáme **znaky**. Mohou to být tělesná výška či hmotnost, měsíční příjem nebo také pohlaví, národnost či strana, kterou daný jedinec hodlá volit. U každého znaku musíme určit měřítko, v němž budeme zjišťované hodnoty vyjadřovat. Je zřejmé, že nevystačíme s jediným druhem měřítka, že měřítko pro národnost bude zcela jiného druhu než měřítko pro výšku postavy.

Nejjednodušším druhem měřítka je **nominální měřítko**. K jeho zavedení stačí vyjmenovat hodnoty, jichž může daný znak nabývat. Je nutné, aby se jednotlivé hodnoty navzájem vylučovaly a abychom vždycky vhodnou hodnotu mohli najít. Hodnota je pak dána jednoznačně. Příkladem může být pohlaví, národnost nebo barva očí. Pokud měřítko obsahuje pouhé dvě možné hodnoty, hovoří se někdy o **měřítku nula-jedničkovém**.

Podrobněji vypovídá **měřítko ordinální**. Má všechny vlastnosti nominálního měřítka, ale navíc jsou jeho hodnoty uspořádány. Patří sem například nejvyšší dosažené vzdělání, stupeň bolesti nebo třeba barva v barevném spektru duhy. Jak uvidíme později, někdy na znaky s ordinálním měřítkem můžeme pohlížet jako na znaky s měřítkem nominálním a příslušné uspořádání pro jednoduchost pomíjíme.

Dalším stupněm složitosti měřítek je **měřítko intervalové**, které stejně jako ordinální měřítko předpokládá uspořádané hodnoty měřeného znaku.

Navíc předpokládá, že hodnoty jsou pravidelně rozmístěny. Stačí pak jednu z hodnot označit za nultou a některou větší jako jedničku, následující (stejně vzdálenou od jedničky, jako je jednička od nuly) jako dvojku atd. Interval mezi nulou a jedničkou můžeme libovolně dělit (polovina, desetiny, tisíciny) či násobit (stem nebo také záporným číslem). Hodnoty pak ukazují „vzdálenost“ od zvoleného počátku, případně i směr, jakým se k dané hodnotě od počátku dostaneme. Typickou otázkou je, „o kolik se dvě hodnoty liší“. Příkladem může být Celsiova teplotní stupnice nebo rok narození.

Nejsložitější je **měřítko poměrové**. Má opět všechny vlastnosti předcházejícího měřítka (intervalového), ale navíc počátek (nulová hodnota) není libovolný. V poměrovém měřítku porovnáváme naměřenou hodnotu s předem definovanou jednotkovou hodnotou. Výška postavy je údaj, který porovnává skutečný fyzikální rozměr lidského těla se zvolenou fyzikální jednotkou. Říká, kolikrát je člověk delší, než je 1 m, což byla, jak jsem se v polovině minulého století učil ve škole, vzdálenost mezi dvěma ryskami na tyči ze slitiny platiny a iridia umístěné v Sèvres u Paříže. Má zde smysl stejná otázka jako u měřítka intervalového (o kolik km je delší cesta z A do B, když jedeme přes C a ne přes D), navíc má smysl také otázka: „Kolikrát je cesta z A do B přes C delší než cesta z A do B přes D?“ Příkladem může sloužit každé měření délky či hmotnosti nebo také údaj o věku dané osoby.

V některých souvislostech vystačíme s hrubším rozlišením na **měřítka kvalitativní** (**kategoriální**) a **měřítka kvantitativní** (**číselná**). K prvním se zpravidla zařazují měřítka nominální a ordinální, ke druhým pak měřítko intervalové a poměrové. Z tohoto hrubšího rozdělení vychází i dva způsoby, jak výsledek měření modelujeme matematicky, jak zavádíme pojem **veličiny**. Protože u kvantitativních měřítek používáme k vyjádření hodnot čísla, bude veličina s tímto číslem přímo ztotožněna. Naproti tomu u kvalitativních znaků se bez čísel můžeme obejít, jednotlivé hodnoty mají často jen slovní popis. K číslům se dostaneme, když zjistíme **četnosti**, tedy počty případů, kolikrát se ta která hodnota vyskytla.

1.2 Kvantitativní znak

1.2.1 Míry polohy

V rámci již zmíněného rozsáhlého sledování mužů středního věku s ohledem na výskyt aterosklerózy bylo vyšetřeno více než tisíc mužů, kteří byli podle stupně rizika aterosklerózy rozděleni do několika skupin. Podejme si příslušný datový soubor a připravme ukázky dat.

```
> data(Stulong)
> names(Stulong)
 [1] "ID"       "výška"    "váha"     "syst1"    "syst2"
 [6] "chlst"    "Víno"     "cukr"     "bmi"      "věk"
[11] "KOURrisk" "Skupina"

> table(Stulong$Skupina)

NS&NSS    RSI    RSK     PS
   204    329    348     73

> bmiZdraví <- with(Stulong,bmi[Skupina=="NS&NSS"])
> length(bmiZdraví)

[1] 204

> bmiNemocní <- with(Stulong,bmi[Skupina=="PS"])
> length(bmiNemocní)

[1] 73
```

Ve skupině označené jako normální (pro jednoduchost jim budeme říkat zdraví) máme úplná pozorování 204 mužů. Zjištěné hodnoty indexu BMI jsou uvedeny v tabulce 1.1. Rádi bychom úroveň BMI těchto mužů charakterizovali jediným číslem. Půjde o **míru polohy**. Bude to hodnota v nějakém smyslu prostřední? Nebo hodnota, která se vyskytla nejčastěji? Nebo půjde o jakési těžiště hodnot?

Prohlédněme si naměřené hodnoty. V tabulce 1.1 jsou uvedeny všechny hodnoty, v tabulce 1.2 jsou tytéž hodnoty, avšak vzestupně uspořádané pomocí funkce `sort()`. Takové uspořádání číselných hodnot se nazývá **variační řada**. Ve variační řadě snadno identifikujeme nejmenší a největší naměřenou hodnotu (**minimum**, **maximum**). Lze však očekávat, že to nikterak nemusí být hodnoty pro daná měření typické. Proto první dvě z nabízených možností jak charakterizovat úroveň všech měření rychle zavrhneme. Máme-li jedinou hodnotou charakterizovat našich 204 čísel, asi nás hned napadne **průměr**, přesněji **aritmetický průměr**. V našem případě je průměr po zaokrouhlení roven 24,79.

Průměr je číselná charakteristika známá z běžného života. Z novin známe například průměrnou měsíční mzdu zaměstnanců. Ta se spočítá zhruba řečeno tak, že se sečtou všechny jejich (hrubé) mzdy a součet se vydělí počtem zaměstnanců. Je to tedy mzda, kterou by měl každý zaměstnanec, kdyby při stejné celkově vyplacené částce všichni zaměstnanci brali stejně. Formálně

Tabulka 1.1: Hodnoty BMI zdravých mužů v původním pořadí

25,43	27,14	26,01	23,26	21,46	24,62	22,28	26,32	24,57	25,01
24,30	26,22	22,22	26,51	24,02	27,75	24,57	21,46	25,61	21,27
22,60	25,83	26,88	24,45	27,76	24,76	25,88	28,39	24,22	25,46
24,42	25,31	25,01	26,75	22,60	25,83	26,42	26,49	25,01	26,51
24,11	23,32	26,22	25,62	26,03	24,34	21,46	25,98	22,55	23,99
28,09	26,12	24,68	26,83	19,73	23,37	23,55	22,60	22,20	23,36
22,53	24,98	25,83	26,00	21,15	25,43	23,81	22,79	26,77	23,18
22,59	27,17	24,73	25,65	25,51	27,40	22,89	23,57	22,72	23,36
23,18	23,66	24,84	25,11	24,45	22,53	24,09	24,86	26,09	25,16
27,34	22,86	22,23	21,55	25,06	23,78	24,97	26,15	27,72	24,73
25,25	25,18	22,53	24,15	24,44	24,57	26,12	24,16	27,08	25,00
23,36	25,90	27,68	27,12	26,09	23,99	26,53	24,39	24,49	23,32
24,72	23,67	27,77	23,71	23,46	26,99	27,68	25,22	23,51	27,43
24,88	25,73	21,98	26,51	24,54	25,06	24,11	24,90	24,98	26,03
27,61	23,24	24,39	22,91	25,08	25,99	25,28	23,04	25,26	24,30
24,34	22,99	22,10	25,71	23,25	24,38	25,54	24,03	23,46	22,53
25,56	25,14	25,91	22,72	25,76	25,88	23,88	25,56	21,56	25,46
23,67	25,66	25,66	27,36	26,59	27,31	27,08	26,85	24,51	27,68
22,99	24,76	27,46	27,98	25,40	27,06	23,15	22,72	24,67	26,57
25,01	26,73	24,16	23,66	22,94	23,45	26,79	24,84	25,95	22,88
22,74	26,88	20,66	24,96						

vyjádřeno je průměr $\bar{x}$ dán vztahem

$$\boxed{\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.} \tag{1.1}$$

Je taková charakteristika spravedlivá i v případě, že někteří mají třeba jen poloviční úvazek? Neměli bychom mzdu odpovídající částečnému úvazku vzít v úvahu jen „částečně“? Přinejmenším při hodnocení jak dobře je zaplacena odvedená práce je to jistě na místě. Řešením je vážený průměr.

Příklad 1.1 (V4) Na internetové stránce Českého statistického úřadu lze nalézt informace o zemích Evropské unie vztažené ke konci prvního desetiletí tohoto století (Český statistický úřad, 2012). Data jsou také uvedena na přiloženém disku. Všimněme si údajů o HDP vztaženém na jednoho obyvatele a vyjádřeném ve standardu kupní síly čtyř států Visegrádské skupiny

Tabulka 1.2: Uspořádané hodnoty BMI zdravých mužů (variační řada)

19,73	20,66	21,15	21,27	21,46	21,46	21,46	21,55	21,56	21,98
22,10	22,20	22,22	22,23	22,28	22,53	22,53	22,53	22,53	22,55
22,59	22,60	22,60	22,60	22,72	22,72	22,72	22,74	22,79	22,86
22,88	22,89	22,91	22,94	22,99	22,99	23,04	23,15	23,18	23,18
23,24	23,25	23,26	23,32	23,32	23,36	23,36	23,36	23,37	23,45
23,46	23,46	23,51	23,55	23,57	23,66	23,66	23,67	23,67	23,71
23,78	23,81	23,88	23,99	23,99	24,02	24,03	24,09	24,11	24,11
24,15	24,16	24,16	24,22	24,30	24,30	24,34	24,34	24,38	24,39
24,39	24,42	24,44	24,45	24,45	24,49	24,51	24,54	24,57	24,57
24,57	24,62	24,67	24,68	24,72	24,73	24,73	24,76	24,76	24,84
24,84	24,86	24,88	24,90	24,96	24,97	24,98	24,98	25,00	25,01
25,01	25,01	25,01	25,06	25,06	25,08	25,11	25,14	25,16	25,18
25,22	25,25	25,26	25,28	25,31	25,40	25,43	25,43	25,46	25,46
25,51	25,54	25,56	25,56	25,61	25,62	25,65	25,66	25,66	25,71
25,73	25,76	25,83	25,83	25,83	25,88	25,88	25,90	25,91	25,95
25,98	25,99	26,00	26,01	26,03	26,03	26,09	26,09	26,12	26,12
26,15	26,22	26,22	26,32	26,42	26,49	26,51	26,51	26,51	26,53
26,57	26,59	26,73	26,75	26,77	26,79	26,83	26,85	26,88	26,88
26,99	27,06	27,08	27,08	27,12	27,14	27,17	27,31	27,34	27,36
27,40	27,43	27,46	27,61	27,68	27,68	27,68	27,72	27,75	27,76
27,77	27,98	28,09	28,39						

(ČR, Maďarsko, Polsko, Slovensko) za rok 2010. Jsou to po řadě hodnoty 19 400, 15 800, 15 300 a 18 000. Máme-li tuto skupinu států, označovanou jako V4, charakterizovat jedinou hodnotou, jistě to nebude obyčejný průměr $(19\,400 + 15\,800 + 15\,300 + 18\,000)/4 = 17\,125$, ale průměr vážený, přičemž váhy jsou dány počtem obyvatel:

$$\frac{\begin{array}{r}19\,400 \cdot 10\,517\,247 + 15\,800 \cdot 9\,976\,062 \\ + 15\,300 \cdot 38\,441\,588 + 18\,000 \cdot 5\,477\,038\end{array}}{10\,517\,247 + 9\,976\,062 + 38\,441\,588 + 5\,477\,038},$$

což dá hodnotu 16 276,48, která je téměř o 850 jednotek menší, než prostý průměr. Při výpočtu váženého průměru jsme v čitateli spočítali za každou zemi celkovou hodnotu HDP a tyto hodnoty sečetli, takže jsme dostali celkový hrubý domácí produkt za celou skupinu V4. Ten jsme vydělili celkovým počtem obyvatel V4 a dostali tak ukazatel vztažený na jednoho obyvatele.

Výsledná hodnota je menší, než prostý průměr, protože počty obyvatel jednotlivých zemí jsou poměrně nevyrovnané, nejlidnatější Polsko má nejmenší hodnotu HDP na obyvatele. ◯

Příklad 1.2 (příjmy) Řekněme, že každý ze tří zaměstnanců s měsíčními platy 17, 23 resp. 69 tisíc pracuje na celý úvazek. Další dva zaměstnanci mají jen poloviční úvazky na místech, kde je při plném úvazku plat 19 resp. 32 tisíc. To znamená, že jejich měsíční příjmy jsou 9,5 a 16 tisíc. Mechanicky spočítaný průměr z celých úvazků by byl $(17+23+69+19+32)/5 = 32$ tisíc. Ten nás ale nezajímá, protože vlastně nic reálného o skutečnosti nevypovídá. Důležitější je průměr měsíčních příjmů, tedy skutečně vyplácených částek, totiž $(17+23+69+9{,}5+16)/5 = 26{,}9$ tisíce. Ovšem nejmenší příjmy mají dva zaměstnanci s pouhým poloviční úvazkem. Pro zaměstnavatele je to vlastně totéž, jako by na jejich místě měl jediného zaměstnance s měsíčním platem $0{,}5 \cdot 32 + 0{,}5 \cdot 19 = 25{,}5$ tisíc. Průměrný měsíční plat *vztažený na celý úvazek* je roven

$$(17 + 23 + 69 + 0{,}5 \cdot 32 + 0{,}5 \cdot 19)/(1 + 1 + 1 + 0{,}5 + 0{,}5) = 33{,}625 \text{ tisíc.}$$

Příjem vyplácený lidem s poloviční úvazkem jsme vzali v úvahu jen s poloviční vahou. ◯

Obecně zapíšeme **vážený průměr** hodnot $x_1, x_2, \ldots, x_n$ s nezápornými vahami $w_1, w_2, \ldots, w_n$ jako

$$\boxed{\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}.} \tag{1.2}$$

K zavedení další charakteristiky polohy se znovu pokusíme využít variační řadu. K výkladu opět použijeme příklad s hodnotami BMI. V tabulce 1.2 obsahující variační řadu hodnot BMI zdravých mužů se pokusíme pracovat s číslem (čísly), které je (jsou) uprostřed. Celkem je v tabulce 204 hodnot. Když variační řadu o 204 prvcích rozdělíme na dvě stejně velké části, bude její 102. prvek, totiž hodnota 24,86, posledním prvkem poloviny s menšími hodnotami a následující 103. prvek variační řady rovný 24,88 bude první hodnotou v polovině větších hodnot. Průměr z těchto dvou hodnot $\tilde{x} = 24{,}87$ má tu vlastnost, že dělí variační řadu na *dvě stejně velké části*. Na hodnoty, které jsou *menší (nebo stejné)* jako $\tilde{x}$ a hodnoty které jsou *větší (nebo stejné)* jako $\tilde{x}$. Číslo s uvedenou vlastností se nazývá **medián**. (Čtenář si možná uvědomil, že zmíněnou vlastnost má každé číslo větší než 24,86 a současně menší než 24,88. Abychom měli zaveden medián jednoznačně, volíme průměr těchto hodnot.)

Tabulka 1.3: Uspořádané hodnoty BMI nemocných mužů (variační řada)

20,34	20,99	21,45	21,47	21,53	21,87	22,46	22,86	23,45	23,78
23,84	23,88	24,11	24,22	24,62	24,69	24,80	24,81	24,86	25,01
25,10	25,26	25,47	25,54	25,62	25,69	25,69	25,76	25,83	25,83
25,88	26,02	26,09	26,12	26,18	26,22	26,30	26,37	26,45	26,47
26,49	26,54	26,57	26,70	26,75	26,78	26,99	27,12	27,12	27,18
27,28	27,36	27,44	27,47	27,64	27,76	28,06	28,34	28,40	29,41
29,41	29,63	30,00	30,13	30,86	30,99	31,38	31,55	32,49	32,49
33,61	34,33	44,96							

V tabulce 1.3 je uvedena variační řada hodnot BMI mužů, které lékaři označili jako nemocné (přesněji, z jejich pohledu šlo o patologickou skupinu). Délka této řady je jen 73, takže přesně uprostřed variační řady je 37. hodnota, která je rovna 26,30. Stejný počet hodnot stojí ve variační řadě od této hodnoty nalevo jako napravo. Pro medián tedy platí $\tilde{x} = 26{,}30$.

Zapišme zjištěné obecně. Naměřené číselné hodnoty jsme již dříve označili jako $x_1, x_2, \ldots, x_n$. **Variační řadu** odlišíme od pouhého seznamu naměřených hodnot tím, že indexy rozlišující její jednotlivé prvky napíšeme do závorky. Nejmenší z hodnot $x_1, x_2, \ldots, x_n$ má tedy označení $x_{(1)}$, druhá nejmenší hodnota označení $x_{(2)}$, atd. až největší je $x_{(n)}$:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}. \tag{1.3}$$

Medián pak můžeme definovat s rozlišením sudého a lichého počtu členů variační řady jako

$$\tilde{x} = \begin{cases} x_{(\frac{n+1}{2})} & \text{pro } n \text{ liché,} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{pro } n \text{ sudé.} \end{cases} \tag{1.4}$$

S uspořádanou posloupností (variační řadou) souvisí také pojem **pořadí**. Nejmenší zjištěná hodnota, tedy $x_{(1)}$, má pořadí 1, druhá nejmenší hodnota $x_{(2)}$ má pořadí 2 atd. až největší zjištěná hodnota $x_{(n)}$ má pořadí n. Pokud někde došlo ke shodě (sousední prvky variační řady jsou stejné), přidělíme takovým hodnotám průměrné pořadí z těch, která by dostaly, kdyby stejné nebyly, ale nějak málo se lišily. Například hodnoty 6, 3, 5, 3, 3, 8, 6, 9 mají po řadě pořadí 5,5, 2, 4, 2, 2, 7, 5,5, 8. Nejmenší je tu trojka, a to ve třech

exemplářích, které se dělí o první, druhé a třetí místo. Proto všechny trojky dostaly pořadí 2.

Pojem mediánu můžeme zobecnit. Připomeňme, že medián odděluje polovinu menších hodnot od druhé poloviny, od hodnot větších. Chceme-li místo poloviny oddělit jen čtvrtinu nejmenších hodnot, nazveme onu oddělující konstantu **dolní kvartil**. Podobně tři čtvrtiny nejmenších hodnot od těch ostatních odděluje **horní kvartil**. Pořadí kvartilů odpovídá i jejich běžné označení symboly Q_1 a Q_3. Způsob výpočtu kvartilů upřesníme za chvíli, nejprve pojem kvartilu ještě zobecníme.

Výběrový kvantil x_p (p-tý **percentil**) odděluje dané procento nejmenších hodnot od hodnot větších. Existuje řada postupů, jak percentil vyčíslit, zde uvedeme ten, který standardně používá erková funkce `quantile()`. Percentil x_p z variační řady délky n pro dané p, $0 \le p \le 1$, určíme jako vážený průměr sousedních hodnot $x_{(k)}, x_{(k+1)}$ variační řady. Index k určující jejich umístění ve variační řadě spočítáme jako celou část výrazu $1+(n-1)p$, tedy $k = \lfloor 1+(n-1)p \rfloor$. Zlomkovou část tohoto výrazu označíme symbolem q. Platí tedy $q = 1 + (n-1)p - \lfloor 1 + (n-1)p \rfloor$ resp. $1 + (n-1)p = k + q$. Percentil je dán vztahem

$$x_p = (1-q)x_{(k)} + qx_{(k+1)}, \tag{1.5}$$

tedy je to vážený průměr prvků $x_{(k)}, x_{(k+1)}$ variační řady, přičemž vahami jsou čísla $1-q$ a q. Na čtenáři ponechám ověření, že medián je vlastně percentil $x_{0,5}$, kdežto kvartily jsou percentily $x_{0,25}$ a $x_{0,75}$.

Poznámka Při výpočtu kvartilů Q_1, Q_3 použitých při kreslení **krabicového diagramu** používá R poněkud jiný postup (jde vlastně o odhad populačního kvantilu).

Příklad 1.3 (BMI) Spočítejme oba kvartily z hodnot BMI zdravých mužů. Jak víme, je $n = 204$. Pro výpočet dolního kvartilu Q_1 zvolíme $p = 1/4$. Je tedy $1 + (n-1) \cdot p = 1 + (204-1) \cdot 0{,}25 = 51{,}75 = 51 + 0{,}75$. Celá část čísla 51,75 je $k = 51$, jeho zlomková část pak $q = 0{,}75$. Je tedy

$$Q_1 = (1-q) \cdot x_{(k)} + q \cdot x_{(k+1)} = 0{,}25 \cdot 23{,}46 + 0{,}75 \cdot 23{,}46 = 23{,}46.$$

Podobně pro horní kvartil: $k = \lfloor 1 + (204-1) \cdot 0{,}75 \rfloor = \lfloor 153{,}25 \rfloor = 153$, $q = 0{,}25$, tedy

$$Q_3 = (1-q) \cdot x_{(k)} + q \cdot x_{(k+1)} = 0{,}75 \cdot 26{,}00 + 0{,}25 \cdot 26{,}01 \doteq 26{,}00.$$

○

Speciálním případem váženého průměru je **useknutý průměr** (angl. trimmed, odtud označení v (1.6)). Zvolí se díl nejmenších pozorování, která vynecháme. Často se volí $\alpha = 0{,}1$. Vynechá se také stejný díl největších pozorování a ze zbývajících hodnot se spočítá aritmetický průměr. Jde vlastně o vážený průměr. Vezmeme celou část k z hodnoty $\alpha \cdot n$ a zvolíme $w_1 = \ldots = w_k = w_{n-k+1} = \ldots = w_n = 0$, ostatní váhy jsou $w_i = 1$ a vzorec (1.2) pro vážený průměr použijeme na variační řadu. Dostaneme

$$\bar{x}_{\text{trim}} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}. \tag{1.6}$$

Příklad 1.4 (BMI) Spočítejme useknutý průměr hodnot BMI zdravých mužů. Pro standardní $\alpha = 0{,}1$ dostaneme $\alpha \cdot n = 0{,}1 \cdot 204 = 20{,}4$, takže celá část je $k = 20$. Z variační řady (tabulka 1.2) tedy vynecháme prvních 20 a posledních 20 hodnot a useknutý průměr bude roven

$$\bar{x}_{\text{trim}} = (22{,}59 + 22{,}60 + \ldots + 27{,}08 + 27{,}08)/(204 - 2 \cdot 20) \doteq 24{,}81.$$

Připomeňme, že aritmetický průměr je roven $\bar{x} = 24{,}79$. Oba průměry se tentokrát liší jen nepatrně. V případě 73 nemocných mužů je však rozdíl mezi oběma průměry větší. Aritmetický průměr je $\bar{x} \doteq 26{,}74$, kdežto useknutý průměr pro $\alpha = 0{,}1$ je 26,46, tedy znatelně menší. Rozdíl je způsoben hlavně největší hodnotou 44,96, která ovlivní pouze aritmetický průměr, nikoliv průměr useknutý. Ještě méně je velkými hodnotami ovlivněn medián, který u nemocných mužů vyšel $\tilde{x} = 26{,}30$.

Později (příklad 6.8) uvidíme, že někdy je vhodné naměřené hodnoty vyjádřit v jiném měřítku, například logaritmickém. Pro logaritmy původních hodnot dostaneme průměr 3,278 a useknutý průměr 3,273. Poměr mezi useknutým průměrem a průměrem je v tomto případě 0,9985, kdežto před logaritmováním byl tento poměr roven 0,9895, tedy dál od jedničky. Je patrné, že logaritmování potlačilo vliv mimořádně velké maximální hodnoty. Pro zajímavost, medián je tentokrát roven hodnotě 3,269, kteroužto hodnotu můžeme také dostat logaritmováním mediánu netransformovaných hodnot. Zamyslete se nad tím, proč tomu tak je, že to není jen shodou okolností. ○

Jako poslední uvedeme **modus** $\hat{x}$. Je to nejčastěji se vyskytující hodnota. Modus nemusí být dán jednoznačně. Není příliš vhodný pro spojité veličiny, u nichž pak velmi záleží na tom, jak přesně jednotlivé hodnoty při zápisu zaokrouhlujeme.

Příklad 1.5 (BMI) Vezměme hodnoty BMI skupiny nemocných mužů. Při zaokrouhlení hodnot na celé číslo dostaneme $\hat{x} = 26$, při vyjádření hodnot s přesností na jedno desetinné místo bude maximální četnost u hodnot 21,5, 25,8 a 26,5. Při zaokrouhlení na dvě desetinná místa bude maximální četnost u hodnot 25,69, 25,83, 27,12, 29,41 a 32,49. ◯

1.2.2 Výpočet pomocí R

Ukažme nyní výpočet popisných charakteristik (měr) polohy v prostředí R. Předpokládáme, že data jsou již přítomna v pracovním prostoru R. Hodnoty BMI zdravých mužů jsou označeny jako bmiZdraví. Variační řadu spočítáme pomocí funkce `sort()` a uložíme jako bmiZdravíSRT. Vnější závorky u příkazu generujícího bmiZdravíSRT zajistí, že se výsledná variační řada objeví v okně konzole. V hranatých závorkách je na začátku každého řádku uveden pořadový index prvního prvku variační řady v daném řádku:

```
> (bmiZdravíSRT <- sort(bmiZdraví))
  [1] 19,73 20,66 21,15 21,27 21,46 21,46 21,46 21,55 21,56 21,98
 [11] 22,10 22,20 22,22 22,23 22,28 22,53 22,53 22,53 22,53 22,55
 [21] 22,59 22,60 22,60 22,60 22,72 22,72 22,72 22,74 22,79 22,86
 [31] 22,88 22,89 22,91 22,94 22,99 22,99 23,04 23,15 23,18 23,18
 [41] 23,24 23,25 23,26 23,32 23,32 23,36 23,36 23,36 23,37 23,45
 [51] 23,46 23,46 23,51 23,55 23,57 23,66 23,66 23,67 23,67 23,71
 [61] 23,78 23,81 23,88 23,99 23,99 24,02 24,03 24,09 24,11 24,11
 [71] 24,15 24,16 24,16 24,22 24,30 24,30 24,34 24,34 24,38 24,39
 [81] 24,39 24,42 24,44 24,45 24,45 24,49 24,51 24,54 24,57 24,57
 [91] 24,57 24,62 24,67 24,68 24,72 24,73 24,73 24,76 24,76 24,84
[101] 24,84 24,86 24,88 24,90 24,96 24,97 24,98 24,98 25,00 25,01
[111] 25,01 25,01 25,01 25,06 25,06 25,08 25,11 25,14 25,16 25,18
[121] 25,22 25,25 25,26 25,28 25,31 25,40 25,43 25,43 25,46 25,46
[131] 25,51 25,54 25,56 25,56 25,61 25,62 25,65 25,66 25,66 25,71
[141] 25,73 25,76 25,83 25,83 25,83 25,88 25,88 25,90 25,91 25,95
[151] 25,98 25,99 26,00 26,01 26,03 26,03 26,09 26,09 26,12 26,12
[161] 26,15 26,22 26,22 26,32 26,42 26,49 26,51 26,51 26,51 26,53
[171] 26,57 26,59 26,73 26,75 26,77 26,79 26,83 26,85 26,88 26,88
[181] 26,99 27,06 27,08 27,08 27,12 27,14 27,17 27,31 27,34 27,36
[191] 27,40 27,43 27,46 27,61 27,68 27,68 27,68 27,72 27,75 27,76
[201] 27,77 27,98 28,09 28,39
```

Minimum, maximum, průměr i medián spočítáme snadno pomocí příslušných funkcí:

```
> min(bmiZdraví)
```

```
[1] 19,73
>  max(bmiZdraví)
[1] 28,39
>  mean(bmiZdraví)
[1] 24,78946
>  median(bmiZdraví)
[1] 24,87
```

Obecně najdeme výběrové kvantily (percentily) pomocí funkce `quantile()`, v níž v parametru probs uvedeme, které kvantily chceme počítat:

```
> quantile(bmiZdraví,probs=c(0,25,50,75,100)/100)
     0%     25%     50%     75%    100%
19,7300 23,4600 24,8700 26,0025 28,3900
```

Při výpočtu useknutého průměru stačí ve funkci `mean()` změnit implicitní nastavení parametru trim z nuly na příslušný podíl. Chceme-li vynechat deset procent nejmenších a deset procent největších hodnot, použijeme

```
> mean(bmiZdraví,trim=0.1)
[1] 24,81238
```

Doplňme ještě výpočet průměru a useknutého průměru z logaritmů naměřených hodnot:

```
> mean(log(bmiZdraví))
[1] 3,208017
> mean(log(bmiZdraví),trim=0.1)
[1] 3,210122
```

Zmiňovali jsme se také o závislosti modu spojité veličiny na zaokrouhlování hodnot. Tabulku četností při zaokrouhlování na jedno desetinné místo spočítáme, pro další použití uložíme do proměnné tab a hned vytiskneme pomocí

```
> print(tab <- table(round(bmiNemocní,1)))
20,3   21 21,4 21,5 21,9 22,5 22,9 23,4 23,8 23,9 24,1 24,2 24,6
   1    1    1    2    1    1    1    1    2    1    1    1    1
24,7 24,8 24,9   25 25,1 25,3 25,5 25,6 25,7 25,8 25,9   26 26,1
   1    2    1    1    1    1    2    1    2    3    1    1    2
26,2 26,3 26,4 26,5 26,6 26,7 26,8   27 27,1 27,2 27,3 27,4 27,5
```

```
   2    1    2    3    1    1    2    1    2    1    1    2    1
27,6 27,8 28,1 28,3 28,4 29,4 29,6   30 30,1 30,9   31 31,4 31,6
   1    1    1    1    1    2    1    1    1    1    1    1    1
32,5 33,6 34,3   45
   2    1    1    1
```

Funkce `table()` v našem případě vytvořila vektor (posloupnost čísel) četností jednotlivých hodnot, přičemž jako název dané četnosti (horní řádek) použila vždy příslušnou hodnotu. Pouze položky tabulky s největšími četnostmi vybereme z vektoru `tab` pomocí

```
> tab[tab==max(tab)]
25,8 26,5
   3    3
```

1.2.3 Co mají míry polohy společné?

Pokusme se popsat obecné vlastnosti, které jsou společné právě uvedeným charakteristikám. Kdybychom v příkladu s měsíčními platy zaměstnanců každý měsíční plat zvětšili o deset tisíc, byl by jejich nový aritmetický průměr o stejných deset tisíc větší. To platí i pro vážený průměr. Kdybychom stejné platy vyjádřili nikoliv v tisících korun, ale přímo v korunách, byl by nový průměr numericky také tisíckrát větší. Obecně můžeme uvedená pravidla zapsat jako požadavky na obecnou **míru polohy** μ (též **míra centrální tendence**):

$$\mu(a + X) = a + \mu(X), \tag{1.7}$$

$$\mu(b \cdot X) = b \cdot \mu(X), \qquad (b > 0) \tag{1.8}$$

přičemž a a b jsou nějaké konstanty a X je symbolické označení celého seznamu číselných hodnot. S jistou námahou lze ověřit, že oba uvedené požadavky splňují všechny dosud uvedené charakteristiky. Omezení na kladné b je nutné pouze pro percentily s výjimkou percentilu $x_{0,5}$, je tedy nutné pro oba kvartily, není nutné pro medián.

1.2.4 Míry variability

Míry polohy nám říkají, zda jsou naměřené hodnoty malé nebo velké, ukazují na jejich úroveň. Nyní se budeme věnovat jiné vlastnosti – variabilitě.