2 Bodové a intervalové rozdělení četností

2.1 Jednorozměrné bodové rozdělení četností

Dataset: 17-anova-newborns-2.txt
Máme k dispozici údaje o porodní hmotnosti novorozenců z okresní nemocnice získané v období jednoho roku a současně máme k dispozici údaje o počtu starších biologických sourozenců novorozence, pohlaví novorozence a vzdělání matky (Alánová, 2008; soubor 17-anova-newborns-2.txt).

Popis proměnných v datasetu:

- edu.M – vzdělání matky (1 – základní, 2 – střední bez maturity, 3 – střední s maturitou, 4 – vysokoškolské);
- prch.N – počet biologických starších sourozenců (0–9);
- sex.C – pohlaví dítěte (m – muž, f – žena);
- weight.C – porodní hmotnost dítěte (g);
- weight.K – porodní hmotnost dítěte (1 = nízká (nižší než 2 500 g), 2 = norma (2 500 − 4 200 g), 3 = vysoká (větší než 4 200 g))

Příklad 2.1. Načtení datového souboru
Načtěte dataset 17-anova-newborns-2.txt do proměnné data a vypište prvních 5 řádků z načteného souboru. Zjistěte, zda soubor obsahuje neznámé (NA) hodnoty a pokud ano, tak je odstraňte. Potom zjistěte dimenzi datové tabulky data.

Řešení příkladu 2.1

```
  edu.M prch.N sex.C weight.C weight.K
1     2      0     m     3470        2
2     2      0     m     3240        2
3     2      0     f     2980        2
4     1      0     m     3280        2
5     3      0     m     3030        2
```

Načtená datová tabulka obsahuje údaje o znacích: vzdělání matky (edu.M), počet starších sourozenců novorozence (prch.N), pohlaví novorozence (sex.C), porodní hmotnost novorozence (weight.C) a kategoriální porodní hmotnost novorozence (weight.K). Datový soubor obsahuje celkem .... NA hodnot. Tabulka data má po odstranění NA hodnot celkem .... řádků a .... sloupců. V tabulce jsou tedy po odstranění NA hodnot uloženy údaje o .... **objektech**, přičemž u každého objektu máme záznamy o .... **znacích**.

★

Příklad 2.2. Úprava datového souboru
Upravte označení jednotlivých variant kategorického znaku *porodní hmotnost* tak, aby bylo na první pohled zřejmé, jakou hmotnost novorozenec má (1 = nizka, 2 = norma, 3 = vysoka). Analogicky upravte označení jednotlivých variant znaku *vzdělání matky* (1 – ZS, 2 – SS, 3 – SSm, 4 – VS).

Řešení příkladu 2.2

```
  edu.M prch.N sex.C weight.C weight.K
1    SS      0     m     3470    norma
2    SS      0     m     3240    norma
3    SS      0     f     2980    norma
4    ZS      0     m     3280    norma
5   SSm      0     m     3030    norma
6    SS      1     m     3650    norma
```

★

Příklad 2.3. Variační řada
Vytvořte variační řadu znaku X = *vzdělání matky* a variační řadu kategorického znaku Y = *porodní hmotnost novorozence.*

Řešení příkladu 2.3
Zaměřme se nejprve na znak X = *vzdělání matky.* Znak má celkem čtyři varianty: ...., ..., ... a .... Variační řada je tabulka obsahující pro každou (j-tou) variantu znaku X (a) absolutní četnost .... ; (b) relativní četnost ....; (c) absolutní kumulativní četnost ....; (d) relativní kumulatiní četnost ....

```
     nj     pj   Nj     Fj
ZS  417 0.3020  417 0.3020
SS  448 0.3244  865 0.6264
SSm 435 0.3150 1300 0.9413
VS   81 0.0587 1381 1.0000
```

Interpretace výsledků: Datový soubor obsahuje údaje o celkovém počtu .... novorozenců, přičemž v 417 případech (30.20 %) bylo nejvyšší dosažené vzdělání matky ...., v .... případech (............. %) bylo nejvyšší dosažené vzdělání matky středoškolské bez maturity, apod. Celkem .... (............. %) matek novorozenců v datovém souboru získalo středoškolské vzdělání bez maturity nebo nižší, celkem 1300 (94.13 %) matek novorozenců získalo .... nebo .... vzdělání.

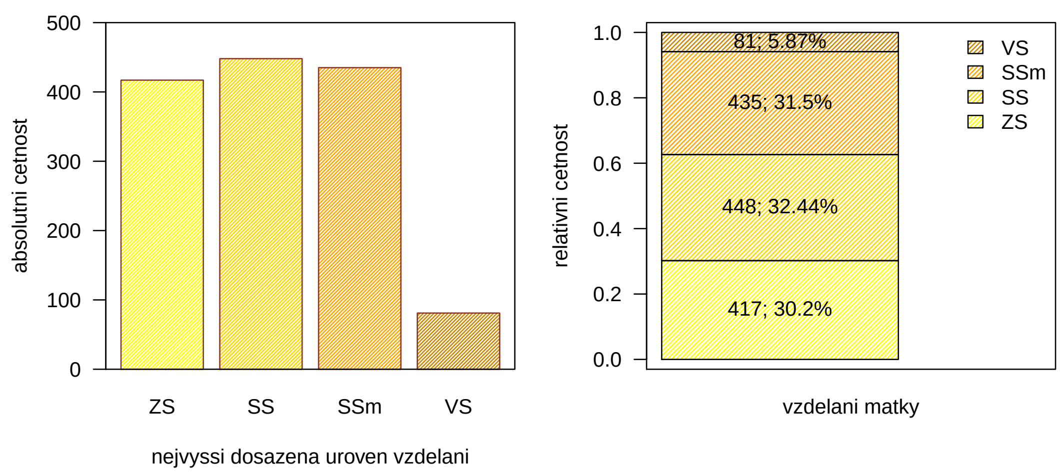
Zaměřme se nyní na znak Y = *porodní hmotnost novorozence.* Protože variační řadu má smysl sestrojovat pouze pro kategoriální / spojitý znak, použijeme k vytvoření variační řady proměnnou weight.C / weight.K. Znak Y má .... varianty: nízká porodní hmotnost, norma a vysoká porodní hmotnost.

```
         nj     pj   Nj     Fj
nizka   266 0.1926  266 0.1926
norma  1071 0.7755 1337 0.9681
vysoka   44 0.0319 1381 1.0000
```

Interpretace výsledků: Porodní hmotnost novorozenců v datovém souboru se v .... případech (............. %) pohybovala v normě. Celkem .... novorozenců (............. %) mělo porodní hmotnost nižší nebo rovnu normě a .... novorozenců (............. %) mělo porodní hmotnost vysokou, v normě, nebo nižší. ★

Příklad 2.4. Sloupcový graf absolutních a relativních četností
Nakreslete sloupcový graf absolutních četností a sloupcový graf relativních četností pro znak X = *vzdělání matky.*

Řešení příkladu 2.4



★

Dvourozměrné bodové rozdělení četností

Příklad 2.5. Kontingenční tabulka absolutních a relativních simultánních četností
Zaměřme se nyní na oba znaky X = *vzdělání matky* a Y = *kategorizovaná porodní hmotnost novorozence* najednou. Z předchozího textu víme, že znak X má čtyři varianty, znak Y má tři varianty. Celkem tedy můžeme získat $4*3 = 12$ různých kombinací variant znaků X a Y. Sestrojte kontingenční tabulku simultánních absolutních četností a kontingenční tabulku simultánních relativních četností znaků X a Y.

Řešení příkladu 2.5
Kontingenční tabulka simultánních absolutních četností bude tabulka o velikosti $(4+1) \times (3+1) = 5 \times 4$ ve tvaru

	nizka	norma	vysoka	suma
ZS	n_{11}	n_{12}	n_{13}	$n_{1.}$
SS	n_{21}	n_{22}	n_{23}	$n_{2.}$
SSm	n_{31}	n_{32}	n_{33}	$n_{3.}$
VS	n_{41}	n_{42}	n_{43}	$n_{4.}$
suma	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

kde n_{jk}, $j = 1, \ldots, 4$ a $k = 1, \ldots, 3$ je *simultánní absolutní četnost* j-té varianty znaku X a k-té varianty znaku Y, $n_{j.}$ (resp. $n_{.k}$) je *marginální absolutní četnost* j-té varianty znaku X (resp. k-té varianty znaku Y) a n je celkový počet objektů v datovém souboru.

Kontingenční tabulka simultánních absolutních četností

```
     nizka norma vysoka suma
ZS      97   312      8  417
SS      82   346     20  448
SSm     74   349     12  435
VS      13    64      4   81
suma   266  1071     44 1381
```

Interpretace výsledků: V datovém souboru se vyskytuje celkem 97 novorozenců, kteří mají .... porodní hmotnost a jejichž matka má .... vzdělání, a .... novorozenců, jejichž porodní hmotnost je v normě a jejichž matka má středoškolské vzdělání s maturitou. Celkem 81 novorozenců se narodilo matkám s .... vzděláním.

Kontingenční tabulka simultánních relativních četností

```
      nizka  norma vysoka   suma
ZS   0.0702 0.2259 0.0058 0.3020
SS   0.0594 0.2505 0.0145 0.3244
SSm  0.0536 0.2527 0.0087 0.3150
VS   0.0094 0.0463 0.0029 0.0587
suma 0.1926 0.7755 0.0319 1.0000
```

Interpretace výsledků: V datovém souboru se vyskytuje celkem 7.02 % novorozenců, kteří mají .... porodní hmotnost a jejichž matka má .... vzdělání. V datovém souboru se vyskytuje celkem ....% novorozenců, jejichž porodní hmotnost je v normě a jejichž matka má středoškolské vzdělání s maturitou. Celkem 3.19,% novorozenců v datovém souboru má .... porodní hmotnost. ★

Příklad 2.6. Kontingenční tabulka řádkově a sloupcově podmíněných relativních četností
Zaměřte se nyní opět na oba znaky X = *vzdělání matky* a Y = *kategorizovaná porodní hmotnost novorozence* najednou. Vytvořte kontingenční tabulku řádkově podmíněných relativních četností a kontingenční tabulku sloupcově podmíněných relativních četností.

Řešení příkladu 2.6
Kontingenční tabulka řádkově podmíněných relativních četností

```
     wei
edu    nizka  norma vysoka
  ZS  0.2326 0.7482 0.0192
  SS  0.1830 0.7723 0.0446
  SSm 0.1701 0.8023 0.0276
  VS  0.1605 0.7901 0.0494
```

Interpretace výsledků: Ze všech novorozenců v datovém souboru, jejichž matka má dokončené středoškolské vzdělání zakončené maturitou, má 17.01 % .... porodní hmotnost a 2.76 % .... porodní hmotnost. Ze všech novorozenců v datovém souboru, jejichž matka má dokončené vysokoškolské vzdělání, má .... % **nízkou** porodní hmotnost a .... % **vysokou** porodní hmotnost.

Kontingenční tabulka sloupcově podmíněných relativních četností

```
     wei
edu    nizka  norma vysoka
  ZS  0.3647 0.2913 0.1818
  SS  0.3083 0.3231 0.4545
  SSm 0.2782 0.3259 0.2727
  VS  0.0489 0.0598 0.0909
```

Interpretace výsledků: Ze všech novorozenců v datovém souboru, jejichž porodní hmotnost byla nízká, se 36.47 % narodilo matkám s ukončeným .... vzděláním. Ze všech novorozenců v datovém souboru, jejichž porodní hmotnost byla v normě, se .... % se narodilo matkám s dokončeným středoškolským vzděláním bez maturity. ★

2.2 Jednorozměrné intervalové rozdělení četností

Dataset: 01-one-sample-mean-skull-mf.txt
Z archivních materiálů (Schmidt, 1888; soubor 01-one-sample-mean-skull-mf.txt) máme k dispozici původní kraniometrické údaje o délce a šířce mozkovny a ze starověké egyptské populace.

Popis proměnných v datasetu:

- id – pořadové číslo;
- pop – populace (egant – egyptská starověká);
- sex – pohlaví (m – muž, f – žena);
- skull.L – největší délka mozkovny (mm), t.j. přímá vzdálenost kraniometrických bodů *glabella* a *opisthocranion*;
- skull.B – největší šířka mozkovny (mm), t.j. vzdálenost obou kraniometrických bodů *euryon*.

Příklad 2.7. Načtení datového souboru
Načtěte dataset 01-one-sample-mean-skull-mf.txt a vypište první čtyři řádky z načteného souboru. Prozkoumejte, zda soubor obsahuje neznámé hodnoty a případně je ze souboru odstraňte. Potom zjistěte dimenzi datové tabulky.

Řešení příkladu 2.7

```
   id   pop sex skull.L skull.B
1 416 egant   m     188     145
2 417 egant   m     172     139
3 420 egant   m     176     138
4 421 egant   m     184     128
```

V datovém souboru se vyskytuje celkem .... neznámých (NA) hodnot.
Po odstranění NA pozorování nám zůstala datová tabulka o velikosti .... řádků a .... sloupců. Celkem tedy máme údaje o 325 .... přičemž pro každý objekt máme .... identifikační proměnnou id a údaje o .... znacích: populaci (pop), pohlaví skeletu (sex), největší délce mozkovny (skull.L) a největší šířce mozkovny (skull.B). ★

Příklad 2.8. Histogram a krabicový diagram
V následující analýze se zaměříme primárně na znak X = *největší šířka mozkovny* u skeletů mužského pohlaví. Proveďte prvotní náhled na znak X = *největší šířka mozkovky* u mužů pomocí (a) histogramu; (b) krabicového diagramu.

Řešení příkladu 2.8
Celkem máme údaje o největší šířce mozkovny u .... mužských skeletů. Hodnoty největší šířky mozkovny v datovém souboru se pohybují v rozmezí ....–............ mm.

Jelikož je sledovaný znak X spojitého typu, je potřeba naměřené hodnoty roztřídit do stejně dlouhých tzv. *třídicích intervalů*. V praxi to znamená, že vytvoříme intervaly pokrývající svým rozsahem celou reálnou osu, tj.

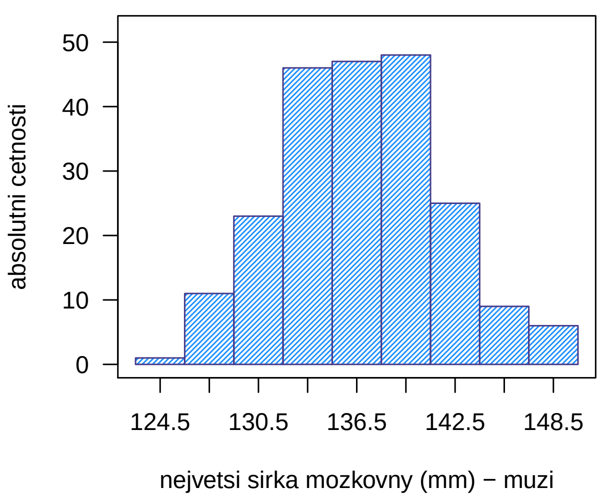
$$(\infty; u_1\rangle\ , (u_1; u_2\rangle, \ldots, (u_r; u_{r+1}\rangle, (u_{r+1}; \infty\rangle,$$

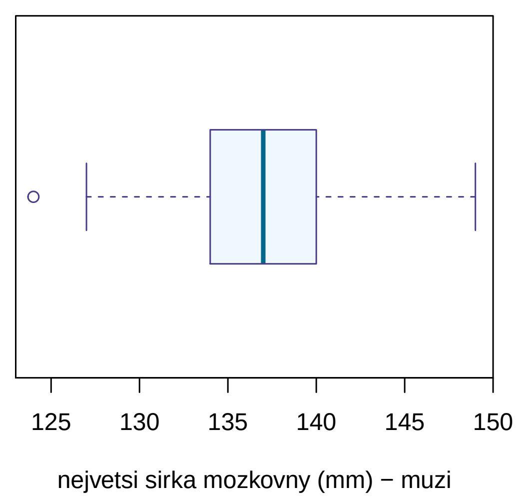
kde $(u_j; u_{j+1}\rangle$, $j = 1, \ldots, J$ je j-tý třídicí interval. Krajní intervaly $(\infty; u_1\rangle$ a $(u_{r+1}; \infty\rangle$ jako třídicí intervaly neuvažujeme, nikdy neobsahují žádné pozorování a slouží jako doplnění celé reálné osy. Počet třídicích intervalů se mění v závislosti na počtu pozorování, které máme k dispozici. Přesný počet třídicích intervalů r v konkrétním případě stanovíme pomocí tzv. Sturgesova pravidla

$$r \approx 1 + 3.3 \log_{10} n. \tag{2.1}$$

Podle Sturgersova pravidla je optimální počet třídicích intervalů pro znak X = největší šířka mozkovny roven .... Minimální naměřená hodnota znaku X je ...., maximální hodnota je .... Rozsah hodnot mezi minimální a maximální hodnotou je ....

Optimální šířka třídicího intervalu pro znak X je .... mm. Vynásobíme-li počet třídicích intervalů optimálním rozsahem jednoho intervalu, zjistíme, že rozsah třídicích intervalů je $9 \times 3 = 27$. Rozsah hodnot 124–149 je však pouze 25. Proto dolní hranici prvního třídicího intervalu u_1 stanovíme jako 123, $u_2 = 126, \ldots, u_9 = 150$.





★