

Cvičenie 7 – Základné spracovanie dát, opisná štatistika

Obsah tohto dokumentu obsahom presahuje prebrané učivo, v priebehu semestra sa však k niektorým bodom ešte vrátíme.

Dokument je zameraný na „ručné“ prepočítanie a ozrejmienie si pojmov. Na „fyzických“ cvičeniach sa zameriame viac na využitie softvéru (iný dokument).

Pojmy a ukážky:

→ Dátový vektor

Dátový vektor získané údaje (súbor dát) uložené v tvare vektora

$$x = [1 \ 5 \ 7 \ 4 \ 5 \ 6 \ 8 \ 4 \ 1 \ 1 \ 2 \ 7]$$

Usporiadaný dátový vektor dáta zoradené vo vektore podľa veľkosti

$$x_s = [1 \ 1 \ 1 \ 2 \ 4 \ 4 \ 5 \ 5 \ 6 \ 7 \ 7 \ 8]$$

Tabuľka početností hodnoty v dátovom vektore a ich početnosti

x_k	1	2	4	5	6	7	8
n_k	3	1	2	2	1	2	1

Tabuľka relatívnych početností početnosti hodnôt delené celkovým počtom dát

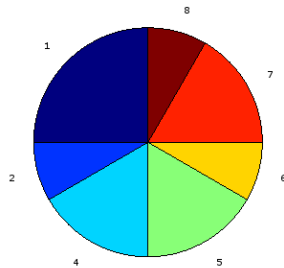
x_k	1	2	4	5	6	7	8
p_k	3/12	1/12	2/12	2/12	1/12	2/12	1/12

Koláčový diagram

slúži najmä na grafické vyjadrenie relatívnej početnosti
(zobrazuje hmatateľne tzv. podiel na celku)

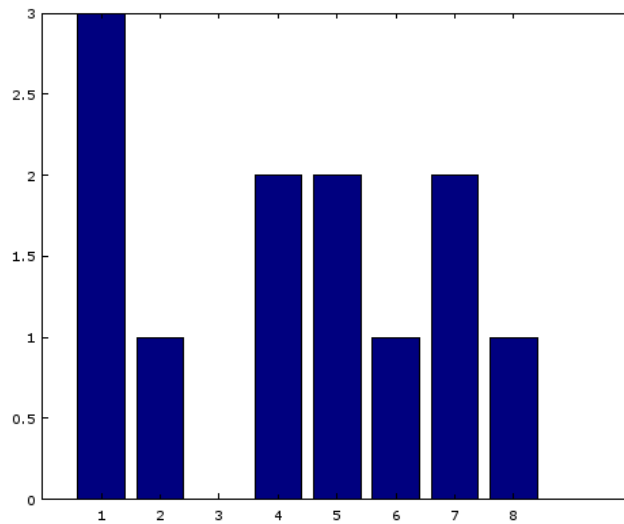
$p = [3 \ 1 \ 2 \ 2 \ 1 \ 2 \ 1] / 12$

`pie(p, {'1' '2' '4' '5' '6' '7' '8'})`



Stĺpcový diagram

slúži najmä na grafické vyjadrenie početnosti



Hodnota 3 sa medzi získanými dátami nenachádza. Je na našom posúdení situácie, či ju chceme / máme spomínať a zobrazovať v diagrame.

Priemer

Výpočet z vektora: $x = [1 \ 5 \ 7 \ 4 \ 5 \ 6 \ 8 \ 4 \ 1 \ 1 \ 2 \ 7]$

$$\bar{x} = \text{mean}(x) = (1+5+7+4+5+6+8+4+1+1+2+7) / 12 = 4.25$$

Výpočet z tabuľky relatívnych početností:

x_k	1	2	4	5	6	7	8
p_k	3/12	1/12	2/12	2/12	1/12	2/12	1/12

$$\text{mean}(v) = 1*3/12 + 2*1/12 + 4*2/12 + 5*2/12 + 6*1/12 + 7*2/12 + 8*1/12$$

Rozptyl

Výpočet z vektora x

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = ((1-4.25)^2 + (5-4.25)^2 + \dots + (7-4.25)^2) / 12 = 5.8542$$

Na tomto mieste treba rozlíšiť niekoľko základných situácií:

- Náhodná veličina/premenná ako „živel“, šedá eminencia v pozadí, ktorá má svoje charakteristiky a v súlade s nimi je schopná v experimente/pozorovaní/ankete a pod. sa prejavovať a generovať náhodné výsledky. Toto sme preberali do minulého týždňa.
- Vektor dát, ktoré predstavujú „celú realitu“ = *populáciu* (napr. všetky výsledky z pretekov, známky všetkých žiakov). Tento vektor vieme opísať podobnými charakteristikami ako náhodnú veličinu. O tom je dnešné cviko.
- Výberový náhodný vektor predstavuje výsledok experimentu, v ktorom sa náhodná veličina prejavila v určitom počte dát, avšak tieto dáta nevyčerpávajú celú realitu, iba ju lepšie či slabšie reprezentujú (napr. preferencie niekoľkých miliónov spotrebiteľov zistené na náhodnej tisícke respondentov). O tomto bude reč v nasledujúcich týždňoch.

Šikmost'

$$\begin{aligned}\gamma_1 &= E((X - E(X))^3) / (D(X))^{3/2} = \\ &= ((1-4.25)^3 + (5-4.25)^3 + \dots + (7-4.25)^3) / 12 / (5.8542)^{3/2} = -0.08163\end{aligned}$$

Napriek prvotnému zdaniu sú údaje naklonené viac vpravo.

Špicatost'

$$\begin{aligned}\gamma_2 &= E((X - E(X))^4) / (D(X))^2 - 3 = \\ &= -1.3405\end{aligned}$$

Údaje sú rozložené viac sploštene v porovnaní s ideálnou krivkou normálneho rozdelenia.

Kvantily

Kvantil je pozícia prvku/hodnoty v kontexte všetkých hodnôt usporiadaných podľa veľkosti, meraná na škále od 0 po 1. Definícií kvantilu je viacero a vedú k nie vždy celkom zhodným výsledkom (pri malom počte dát), ale tie rozdiely sú zanedbateľné.

Dátový vektor

Pri dátovom vektore dávame prednosť takej definícii kvantilu, ktorá v rámci možnosti nevyrába nové hodnoty, než sú tie, ktoré už máme k dispozícii. Ak teda nejaký kvantil vychádza medzi dvoma hodnotami dátového vektora, prikloníme sa k tej, ktorá je bližšie. Iba v prípade, že by to bolo naozaj nerozhodne (presný stred), siahneme za aritmetickým priemerom tých hodnôt.

Definícia

Je daný dátový vektor x s n údajmi. Pre číslo p z intervalu $[0, 1]$ je p -kvantil (q_p) taká hodnota, o ktorej platí:

- aspoň $p \cdot n$ hodnôt z vektora x je menších alebo rovných ako q_p
- aspoň $(1-p) \cdot n$ hodnôt z vektora x je väčších alebo rovných ako q_p
- ak uvedeným podmienkam vyhovuje celý interval hodnôt, za kvantil q_p budeme považovať jeho stred

Slovo „aspoň“ prakticky znamená zaokrúhlenie nahor na najbližšie celé číslo.

Ilustrácia:

$$xs = [1 \ 1 \ 1 \ 2 \ 4 \ 4 \ 5 \ 5 \ 6 \ 7 \ 7 \ 8]$$

Usporiadaný dátový vektor má 12 prvkov. Hľadáme 0.7-kvantil. Vypočítame a zaokrúhlime nahor:

$$0.7 \cdot 12 = 8.4 \approx 9$$

$$0.3 \cdot 12 = 3.6 \approx 4$$

Vidíme, že $9+4 = 13$, o 1 viac ako je dĺžka vektora. Tak to má byť. Štvrtá hodnota sprava a deviata zľava, to sa stretne akurát na čísle 6. Hľadaný kvantil je teda $q_{0.7} = 6$. Ak teda pôjdete po rade od najmenších hodnôt, pri čísle 6 si môžete povedať, že aspoň 70 percent údajov ste prešli.

Hľadáme teraz 0.25-kvantil:

$$0.25 \cdot 12 = 3 \approx 3$$

$$0.75 \cdot 12 = 9 \approx 9$$

Vyhovujú všetky čísla medzi 1 (tretie zľava) až 2 (deviate sprava). Aplikujeme teda tretie pravidlo a za hľadaný kvantil vyhlásime stred nájdeného intervalu, teda $q_{0.25} = 1.5$.

Špeciálne kvantily:

Medzi všetkými kvantilmi sú najzaujímavejšie a najčastejšie sa hľadajú:

$Q_L = q_{0.25}$ – dolný (prvý) kvartil

$Med = q_{0.5}$ – medián (druhý kvartil)

$Q_U = q_{0.75}$ – horný (tretí) kvartil

Boxplot

Boxplot je spôsob znázornenia súboru dát, ktorý sugestívne a prvoplánovo naznačuje rozloženie jednotlivých údajov a ich vzťah k celku.

Pravidlá kreslenia:

- Obrázok sa orientuje zľava doprava (alebo zdola hore).
- Obdĺžnikom sa vymedzí „mainstream“, teda údaje medzi Q_L a Q_U . Vyznačí sa tiež medián, prípadne aj priemer.
- Vypočítame $R = Q_U - Q_L$. Obdĺžniku nakreslíme „fúzy“ (*whiskers*) s dĺžkou $1.5 \cdot R$. V tomto rozmedzí, od $Q_L - 1.5R$ po $Q_U + 1.5R$, sa nachádzajú všetky „normálne“ údaje, teda aj tie, ktoré síce nie sú mainstream, ale prirodzene a očakávateľne sa vyskytujú v realite a sú súčasťou celku.
- Údaje mimo uvedeného rozsahu sa považujú za extrémny (outlier), z hľadiska štatistiky ide o prípady, ktoré sa berú ako výnimky alebo je pri nich podozrenie na chybu. Pri výpočte charakteristík sa extrémne údaje spravidla neberú do úvahy.
- Ak najmenší alebo najväčší z údajov (tiež treba vyznačiť) leží v zóne určenej fúzmi, môže sa fúz skrátiť.

Ukážka:

$xs = [1, 4, 4, 4, 5 \mid 5, 6, 7, 8, 13]$

Priemer $E = 5.7$

Med = 5 (stred medzi 5 a 5),

$Q_L = 4$,

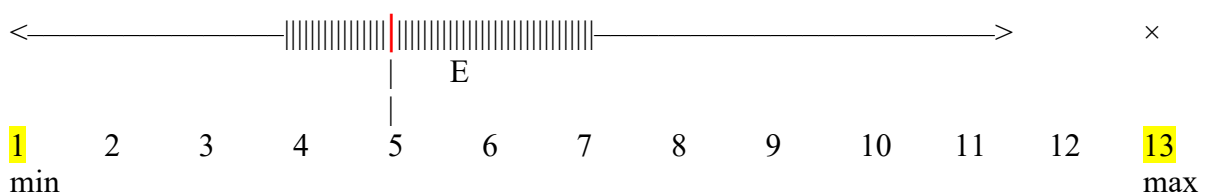
$Q_U = 7$, $R = 3$, $1.5R = 4.5$.

Obdĺžnik (box) bude od 4 po 7.

Fúzy by mali byť od $4 - 4.5 = -0.5$ po $7 + 4.5 = 11.5$.

Keďže minimum medzi údajmi je 1, ľavý fúz stačí ťahať len po 1.

Hodnota 13 je outlier a zároveň maximum.



→ Triedené dáta

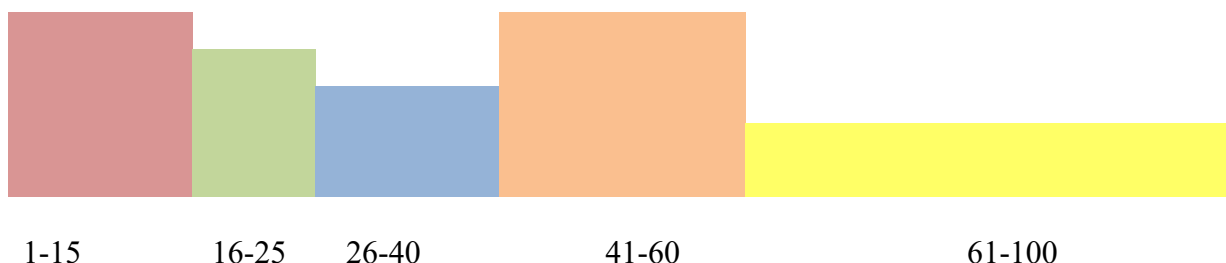
Triedy pri veľkom počte rozličných dát môže byť príliš podrobný pohľad prekážkou k rozpoznaniu štruktúry ich rozloženia. V záujme zvýraznenia štruktúry je vhodné dáta roztriediť a ďalej pracovať len s triedami a ich početnosťami. Niekedy už pri samotnom zbere dát sa výsledky priamo zatriedujú bez uchovávania presnejšej informácie (napr. keď v ankete neudávate svoj presný vek, ale len odkliknete vekovú kategóriu, kam patríte).

Vek	1-15	16-25	26-40	41-60	61-100
Počet	15	8	9	20	16

Histogram špeciálny typ stĺpcového diagramu na zobrazenie triedených dát. Zohľadňuje sa tu šírka tried – trieda 61-100 s počtom 22 ľudí je síce bohatšia než trieda 16-25 s 8 ľuďmi, ale v skutočnosti 8 ľudí na rozsah 10 je lepšia „úroda“ ako 16 ľudí na rozsah 40, pretože $8/10 > 16/40$.

V histograme je preto početnosť vyjadrená nie výškou stĺpca, ale jeho plošným obsahom. Výška stĺpca je podiel početnosti a šírky triedy.

Vek	1-15	16-25	26-40	41-60	61-100
Šírka triedy	15	10	15	20	40
Počet	15	8	9	20	16
Výška stĺpca	$15/15 = 1$	$8/10 = 0.8$	$9/15 = 0.6$	$20/20 = 1$	$16/40 = 0.4$

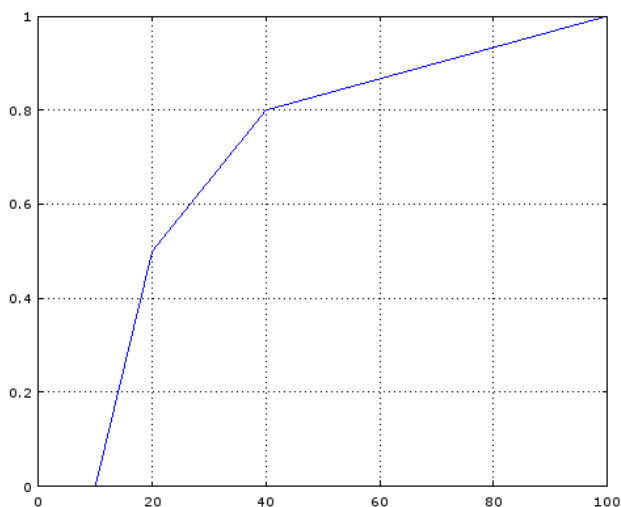


Kvantily pri triedených dátach

Počítanie kvantilov pri triedených dátach je sťažené tým, že nemáme k dispozícii samotné dáta, iba triedy a početnosti. Najjednoduchším východiskom je pracovať s hypotézou rovnomerného rozloženia dát v rámci každej triedy. Vysvetlíme to na jednoduchom príklade:

Prekročenie MPR o x km/h	(10,20]	(20,40]	(40,100]
Počet zaznamenaných priestupkov	50	30	20
Relatívna početnosť	0.5	0.3	0.2
Kumulatívna rel. poč.	0.5	0.8	1

Rovnomerné rozloženie dát v triede znamená, že kumulatívna RP v prvej triede porastie rovnomerne (lineárne) od 0 po 0.5, v druhej triede od 0.5 po 0.8 a v tretej od 0.8 po 1.



Poznámka: obrázok sme vykreslili v Octave pomocou nasledujúcej „vetvičkovej“ funkcie, ktorá vyjadruje presne to, čo sme slovne opísali:

$$F(x) = (10 < x) * (x \leq 20) * (x - 10) * 0.05 \\ + (20 < x) * (x \leq 40) * (0.5 + (x - 20) * 0.3 / 20) \\ + (40 < x) * (x \leq 100) * (0.8 + (x - 40) * 0.2 / 60)$$

Označenie funkcie ako F naznačuje, že ide o analógiu distribučnej funkcie. Hľadanie kvantilov znamená hľadať také x, aby platilo $F(x) = p$. Ideálne je nájsť inverznú funkciu k F – ak to raz urobíme, už nám bude sypať kvantily na želanie. Ak však máme zistiť len zopár kvantilov, môže byť rýchlejšie ich nájsť skusmo.

Z obrázku (a zo zadania) vidíme, že medián je 20. Dolný kvartil je 15 (to vidno voľným okom) a horný kvartil 36.666666... (trochu poskúšame).

Avšak nájsť inverznú funkciu ku F nie je až tak ťažké, aby sme na to rezignovali:

$$F^{-1}(x) = (0 < p) * (p \leq 0.5) * (20 * p + 10) \\ + (0.5 < p) * (p \leq 0.8) * (20 * (p - 0.5) / 0.3 + 20) \\ + (0.8 < p) * (p \leq 1) * (60 * (p - 0.8) / 0.2 + 40)$$

A už stačí dosadzovať želané hodnoty p.

Riešené príklady

Príklad 1:

Je daný súbor dát, zapísaný vo vektore:

[1 4 5 7 2 3 4 8 6 2 4 1 8 5 3]

- Nájdite priemer, rozptyl, medián, kvartily a 0.15, 0.6 a 0.9-kvantily.
- Zostavte tabuľku početností.
- Nakreslite stĺpcový diagram a boxplot.

Riešenie:

- a) Vektor má dĺžku 15. Priemer a rozptyl vieme vypočítať s pomocou softvérov. Napr. takto:

```
mean(x)
ans = 4.2000
```

```
var(x,1)
ans = 4.9600
```

Na ďalšie kroky si musíme vektor usporiadať:

x = [1 1 2 2 3 3 4 4 4 5 5 6 7 8 8]

Med = 4

Kvartily: $0.25 \cdot 15 = 3.75 \approx 4$
 $0.75 \cdot 15 = 11.25 \approx 12$

$Q_L = 2$ (4-tá hodnota zľava, 12-ta sprava)

$Q_U = 6$ (4-tá hodnota sprava, 12-ta zľava)

Kvantily:

x = [1 1 2 2 3 3 4 4 4 5 5 6 7 8 8]

$0.15 \cdot 15 = 2.25 \approx 3$

$q_{0.15} = 2$

$0.6 \cdot 15 = 9 \approx 9$, $0.4 \cdot 15 = 6 \approx 6$

$q_{0.6} = (4+5)/2 = 4.5$

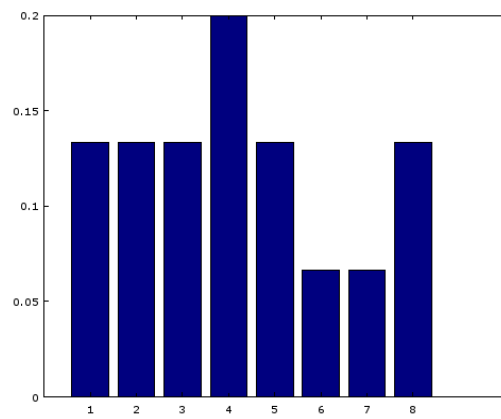
$0.9 \cdot 15 = 13.5 \approx 14$

$q_{0.9} = 8$

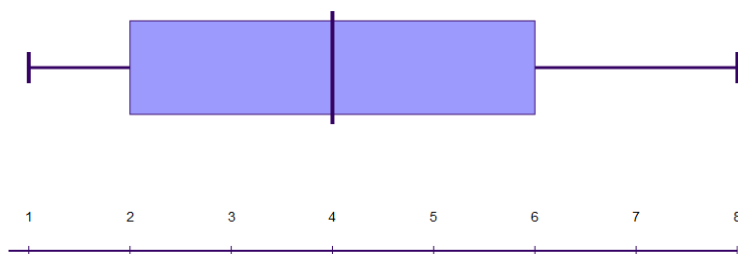
b)

x	1	2	3	4	5	6	7	8
poč.	2	2	2	3	2	1	1	2
rel.poč.	2/15	2/15	2/15	3/15	2/15	1/15	1/15	2/15

c)



$$R = 6 - 2 = 4, 1.5 \cdot R = 6$$



Poznámka – skúste si nájsť online nástroje na kreslenie boxplotu.
Jeden z tých primitívnejších:

<https://www.meta-chart.com/box-and-whisker#/data>

Príklad 2:

100 študentov hrdinsky bojovalo na skúške z Pivovarníckej štatistiky. Po prvom termíne boli výsledky v AIS nasledovné:

body	[0,50)	[50,64)	[64,76)	[76,86)	[86,94)	[94,100)
známka	fx	e	d	c	b	a
počet	40	21	15	12	8	4

Poznámka: Bodové hodnotenie nemuselo byť celočíselné, intervaly chápeme ako podmnožiny $\mathbb{R}$.

- a) Nakreslite histogram.
b) Vypočítajte medián, kvartily a zopár kvantilov.

Riešenie:

a) Doplníme ďalšie užitočné čísla do tabuľky:

známka	fx	e	d	c	b	a
šírka triedy	50	14	12	10	8	6
výška stĺpca	0.8	1.5	1.2	1.2	1	0.66666

Histogram:



b)

body	[0,50)	[50,64)	[64,76)	[76,86)	[86,94)	[94,100)
známka	fx	e	d	c	b	a
počet	40	21	15	12	8	4
rel. poč.	0.4	0.21	0.15	0.12	0.08	0.04
kumul. rp	0.4	0.61	0.76	0.88	0.96	1

Aproximativná distribučná funkcia (po intervaloch):

$$\begin{aligned}
 F(x) = & \quad x \cdot 0.4 / 50 & \text{na } [0, 50) \\
 & 0.4 + (x - 50) / 14 \cdot 0.21 & \text{na } [50, 64) \\
 & 0.61 + (x - 64) / 12 \cdot 0.15 & \text{na } [64, 76) \\
 & 0.76 + (x - 76) / 10 \cdot 0.12 & \text{na } [76, 86) \\
 & 0.88 + (x - 86) / 8 \cdot 0.08 & \text{na } [86, 94) \\
 & 0.96 + (x - 94) / 6 \cdot 0.04 & \text{na } [94, 100)
 \end{aligned}$$

Kvantilová funkcia:

$$\begin{aligned}
 F^{-1}(p) = & \quad p \cdot 50 / 0.4 & \text{na } [0, 0.4) \\
 & 50 + (p - 0.4) \cdot 14 / 0.21 & \text{na } [0.4, 0.61) \\
 & 64 + (p - 0.61) \cdot 12 / 0.15 & \text{na } [0.61, 0.76) \\
 & 76 + (p - 0.76) \cdot 10 / 0.12 & \text{na } [0.76, 0.88) \\
 & 86 + (p - 0.88) \cdot 8 / 0.08 & \text{na } [0.88, 0.96) \\
 & 94 + (p - 0.96) \cdot 6 / 0.04 & \text{na } [0.96, 1]
 \end{aligned}$$

A môžeme dosadzovať:

$$\begin{aligned}
 \text{Medián} & \quad F^{-1}(0.5) = 50 + (0.5 - 0.4) \cdot 14 / 0.21 = 56.6666 \\
 \text{Dolný kvartil} & \quad F^{-1}(0.25) = 0.25 \cdot 50 / 0.4 = 31.25 \\
 \text{Horný kvartil} & \quad F^{-1}(0.75) = 64 + (0.75 - 0.61) \cdot 12 / 0.15 = 75.2 \\
 0.98\text{-kvantil} & \quad F^{-1}(0.98) = 94 + (0.98 - 0.96) \cdot 6 / 0.04 = 97
 \end{aligned}$$

Neriešené príklady

1. 7 študentov sa pýtali, koľkokrát za týždeň využili MHD. Výsledok je zapísaný v podobe dátového vektora:

$$x = [3 \ 9 \ 7 \ 3 \ 1 \ 5 \ 7]$$

- a) Znázorníte výsledok stĺpcovým diagramom.
- b) Vypočítajte priemer a rozptyl. (5, 6.8571)
- c) Nájdite medián a oba kvartily. (5, 3, 7)
- d) Nájdite 0.2, 0.6 a 0.9- kvantily. (3, 7, 9)
- e) Nakreslite boxplot.

2. To isté pre vektor $x = [3 \ 9 \ 7 \ 3 \ 1 \ 5 \ 7 \ 9 \ 3 \ 1]$

3. Na pretekoch *cross-country* / *bežec + pes* (to sa reálne koná napr. na trase Rača-Kamzík) vyzerali časy na výsledkovej listine (v celých minútach) nasledovne:

$$[76, 78, 87, 87, 89, 91, 97, 99, 105, 113, 114, 114, 144, 236]$$

- a) Vypočítajte priemer a rozptyl. (109.3, 1530.1)
- c) Nájdite medián a oba kvartily. (98, 87, 114)
- d) Nájdite 0.2, 0.6 a 0.9- kvantily. (87, 105, 144)
- e) Nakreslite boxplot.

4. Ankety sa zúčastnilo 200 respondentov. Do vekových tried sa začlenili takto:

0-18	54
18-25	70
25-40	45
40-65	30
65-105	1

- a) Znázorníte uvedené hodnoty v histograme.
- b) Nájdite medián, kvartily a 0.11, 0.24 a 0.72-kvantily. (22.6, 16.6666, 33.6666, 7.3333, 16, 31.6666)

5. Medzi prvými majiteľmi áut značky @#\$ zisťovali, po koľkých rokoch sa rozhodli svoje auto predať (alebo zošrotovať). Výsledky sú uvedené v tabuľke.

Vek predaného vozidla	[0,2)	[2,6)	[6,12)	[12,15)	[15,20)	[20,30)
Počet	27	84	277	61	19	3

- a) Znázorníte uvedené hodnoty v histograme.
- b) Nájdite medián, kvartily a 0.15, 0.33 a 0.9-kvantily.