2

Дескриптивна статистика

Задача 2.1. Нека $x_1, ..., x_n$ и $y_1, ..., y_n$ се две низи од податоци кои одговараат на обележјата X и Y соодветно, и за кои важи $y_i = ax_i + b$, $i = 1, 2, ..., n$, каде $a, b = const$. Покажи дека $\overline{y} = a\overline{x} + b$ и $\overline{s}_y^2 = a^2\overline{s}_x^2$.

Решение. Користејќи ги дефинициите на аритметичка средина и дисперзија на податоците, имаме

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b) = \frac{1}{n}\sum_{i=1}^{n} ax_i + \frac{1}{n}\sum_{i=1}^{n} b = a\overline{x} + \frac{1}{n}\cdot nb = a\overline{x} + b,$$

$$\overline{s}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(ax_i + b - (a\overline{x} + b))^2 = \frac{1}{n}\sum_{i=1}^{n} a^2(x_i - \overline{x})^2 = a^2\overline{s}_x^2,$$

што требаше да се покаже. ∎

Задача 2.2. По првите n_1 мерења $x'_1, ..., x'_{n_1}$ кои одговараат на статистичкото обележје X добиена е средина $\overline{x}_1$ и варијанса $\overline{s}_1^2$, а од новите n_2 мерења $x''_1, ..., x''_{n_2}$ добиена е средина $\overline{x}_2$ и варијанса $\overline{s}_2^2$. Ако сите $n_1 + n_2 = n$ мерења се сфатат како една низа од статистички податоци, тогаш се добива средина $\overline{x}$ и варијанса $\overline{s}_0^2$. Покажи дека

а) $\overline{x} = \frac{1}{n}(n_1\overline{x}_1 + n_2\overline{x}_2)$,

б) $\overline{s}_0^2 = \frac{1}{n}(n_1(\overline{s}_1^2 + (\overline{x}_1 - \overline{x})^2) + n_2(\overline{s}_2^2 + (\overline{x}_2 - \overline{x})^2))$.

Решение. За првото тврдење имаме

$$\overline{x} = \frac{1}{n}(\sum_{i=1}^{n_1} x'_i + \sum_{i=1}^{n_2} x''_i) = \frac{1}{n}(n_1\overline{x}_1 + n_2\overline{x}_2),$$

додека за второто тврдење имаме

$$\begin{aligned}
\overline{s}_0^2 &= \frac{1}{n}\Big(\sum_{i=1}^{n_1}(x_i'-\overline{x})^2+\sum_{i=1}^{n_2}(x_i''-\overline{x})^2\Big)= \\
&= \frac{1}{n}\Big(\sum_{i=1}^{n_1}(x_i'-\overline{x}_1+\overline{x}_1-\overline{x})^2+\sum_{i=1}^{n_2}(x_i''-\overline{x}_2+\overline{x}_2-\overline{x})^2\Big)= \\
&= \frac{1}{n}\Big(\sum_{i=1}^{n_1}((x_i'-\overline{x}_1)^2-2(x_i'-\overline{x}_1)(\overline{x}_1-\overline{x})+(\overline{x}_1-\overline{x})^2)+ \\
&\quad +\sum_{i=1}^{n_2}((x_i''-\overline{x}_2)^2-2(x_i''-\overline{x}_2)(\overline{x}_2-\overline{x})+(\overline{x}_2-\overline{x})^2)\Big)= \\
&= \frac{1}{n}(n_1\overline{s}_1^2+n_1(\overline{x}_1-\overline{x})^2+n_2\overline{s}_2^2+n_2(\overline{x}_2-\overline{x})^2)= \\
&= \frac{1}{n}(n_1(\overline{s}_1^2+(\overline{x}_1-\overline{x})^2)+n_2(\overline{s}_2^2+(\overline{x}_2-\overline{x})^2)),
\end{aligned}$$

затоа што $\sum_{i=1}^{n_1}(x_i'-\overline{x}_1)=0$ и $\sum_{i=1}^{n_2}(x_i''-\overline{x}_2)=0$. ∎

Задача 2.3. По анкетирањето на 100 возачи во врска со просечната дневна потрошувачка на гориво (во литри) добиените резултати прикажани се во Табела 2.1.

потрошувачка (во литри)	2	4	5	6	8	10	11	12	13	14
број на возачи	5	10	10	12	18	12	8	10	9	6

Табела 2.1: Просечна дневна потрошувачка на гориво.

Обработи ги дадените податоци (одреди ги бројните карактеристики и направи ги соодветните графички прикази).

Решение. Бројните карактеристики на дадените податоци се

$$\begin{aligned}
\overline{x} &= \frac{1}{100}(5\cdot 2+10\cdot 4+10\cdot 5+12\cdot 6+18\cdot 8+12\cdot 10+ \\
&\quad +8\cdot 11+10\cdot 12+9\cdot 13+6\cdot 14)=8,45,
\end{aligned}$$

$$\begin{aligned}
\overline{s}^2 &= \frac{1}{100}(5\cdot 2^2+10\cdot 4^2+10\cdot 5^2+12\cdot 6^2+18\cdot 8^2+12\cdot 10^2+ \\
&\quad +8\cdot 11^2+10\cdot 12^2+9\cdot 13^2+6\cdot 14^2)-8,45^2=11,7875,
\end{aligned}$$

Ирена Стојковска

$$s^2 = \frac{n}{n-1}\overline{s}^2 = \frac{100}{100-1} \cdot 11,7875 = 11,9066.$$

Модата како вредност која се прима со најголема честота, е вредноста 8 (честотата е 18). Бидејќи $n = 100 = 2 \cdot 50$ е парен број, медијаната m се пресметува според формулата

$$m = \frac{1}{2}(x'_{\frac{n}{2}} + x'_{\frac{n}{2}+1}) = \frac{1}{2}(x'_{50} + x'_{51}) = \frac{1}{2}(8+8) = 8.$$

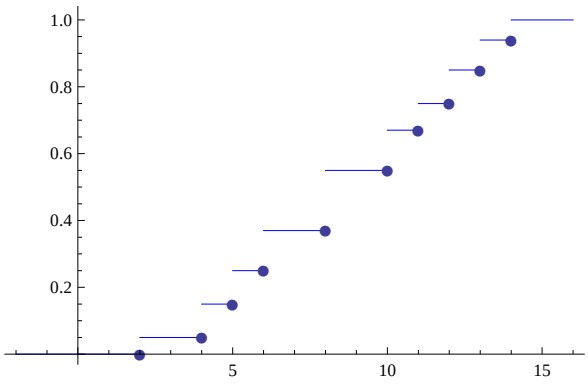
Бидејќи $n = 100 = 4 \cdot 25$ е број од облик $n = 4k$, кварталите Q_1 и Q_3 се пресметуваат според формулите

$$Q_1 = \frac{1}{4}(x'_{\frac{n}{4}} + 3x'_{\frac{n}{4}+1}) = \frac{1}{4}(x'_{25} + 3x'_{26}) = \frac{1}{4}(5 + 3 \cdot 6) = 5,75,$$

$$Q_3 = \frac{1}{4}(3x'_{\frac{3n}{4}} + x'_{\frac{3n}{4}+1}) = \frac{1}{4}(3x'_{75} + x'_{76}) = \frac{1}{4}(3 \cdot 11 + 12) = 11,25.$$

Емпириската функција на распределба е $F_n(x) = \frac{n_x}{n}$, $x \in \mathbb{R}$, каде n_x е бројот на податоци со вредност помала од x. Така, за $x \leq 2$ имаме дека $n_x = 0$, па $F_n(x) = 0$. За $2 < x \leq 4$ имаме дека $n_x = 5$, па $F_n(x) = 5/100 = 0,05$. За $4 < x \leq 5$ имаме дека $n_x = 5 + 10 = 15$, па $F_n(x) = 15/100 = 0,15$. За $5 < x \leq 6$ имаме дека $n_x = 5 + 10 + 10 = 25$, па $F_n(x) = 25/100 = 0,25$ и.т.н. Функцијата $F_n(x)$ заедно со нејзиниот график се прикажани подолу.

$$F_n(x) = \begin{cases} 0 & , x \leq 2 \\ 0,05 & , 2 < x \leq 4 \\ 0,15 & , 4 < x \leq 5 \\ 0,25 & , 5 < x \leq 6 \\ 0,37 & , 6 < x \leq 8 \\ 0,55 & , 8 < x \leq 10 \\ 0,67 & , 10 < x \leq 11 \\ 0,75 & , 11 < x \leq 12 \\ 0,85 & , 12 < x \leq 13 \\ 0,94 & , 13 < x \leq 14 \\ 1 & , x > 14 \end{cases}$$

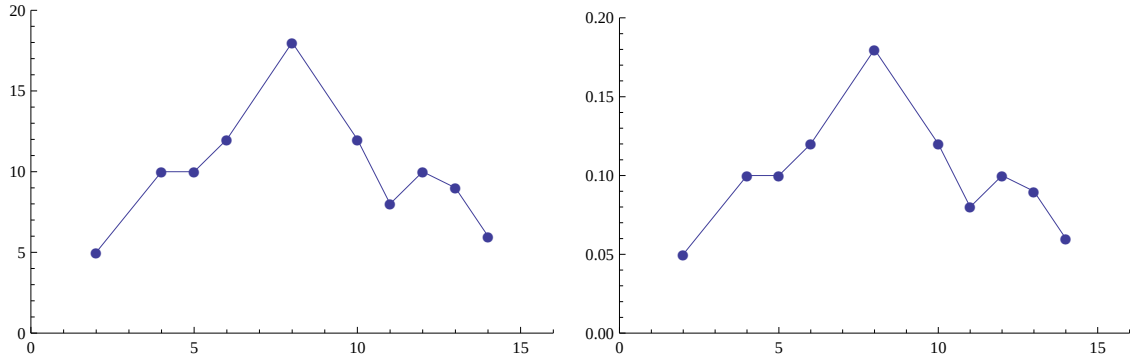


Соодветните релативни честоти на податоците дадени со Табела 2.1 се во Табела 2.2. Полигонот на честоти и полигонот на релативни честоти прикажани се на Слика 2.1. Од овие графички прикази се согледува дека распределбата на податоците е скоро симетрична.

Доколку сакаме да ги прикажеме дадените податоци графички со хистограми, потребно е претходно да ги поделиме во интервали.

потрошувачка (во литри)	2	4	5	6	8	10	11	12	13	14
релативни честоти	0,05	0,10	0,10	0,12	0,18	0,12	0,08	0,10	0,09	0,06

Табела 2.2: Релативни честоти на податоците од Табела 2.1.



(а) (б)

Слика 2.1: (а) Полигон на честоти, Табела 2.1, (б) Полигон на релативни честоти, Табела 2.2

Бројот на интервали r може да се одреди на повеќе начини:

$$r \approx \sqrt{n} = \sqrt{100} = 10,$$

$$r \approx 1 + 3,21 \log n1 + 3,21 \log 100 = 7,42 \approx 7,$$

$$r \approx 5 \log n = 5 \log 100 = 10.$$

Треба да се земе во предвид и препорачливата референца $5-10\%$ од вкупниот број на податоци n и не повеќе од 30% од n, што во нашиот случај ($n = 100$) значи дека препорачливо е r да е број меѓу 5 и 10, но не поголем од 30. Од до сега кажаното, може да се одлучиме за $r = 7$.

Потоа, го прошируваме интервалот во кој се наоѓаат податоците, тоа е интервалот $[2, 14]$. Едно проширување е $[1, 15]$. За ова проширување, должината на секој од подинтервалите ќе изнесува

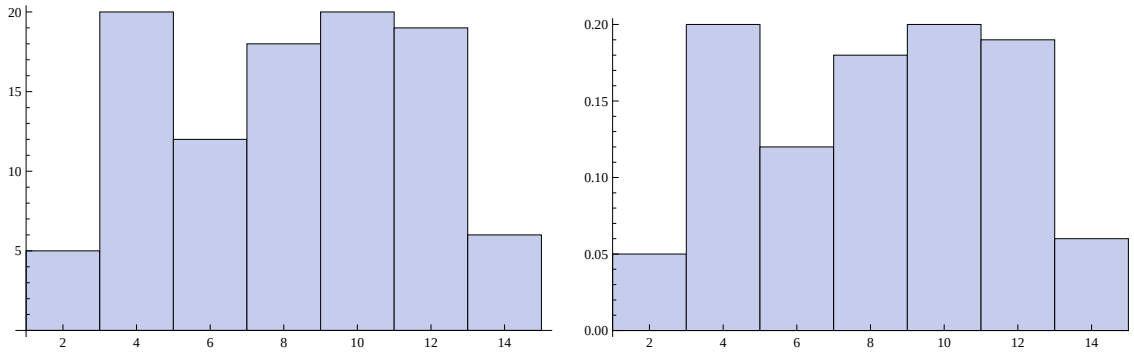
$$h = \frac{15-1}{7} = \frac{14}{7} = 2.$$

Со Табела 2.3 дадени се податоците од Табела 2.1 групирани во интервали, заедно со релативните честоти. Соодветните графички прикази со хистограм на честоти и хистограм на релативни честоти прикажани се на Слика 2.2. Добиените хистограми (Слика 2.2) не укажуваат на некоја поголема симетричност на распределбата на податоците.

Ирена Стојковска

интервал	[1,3]	(3,5]	(5,7]	(7,9]	(9,11]	(11,13]	(13,15]
честоти	5	20	12	18	20	19	6
релативни честоти	0,05	0,20	0,12	0,18	0,20	0,19	0,06

Табела 2.3: Податоците од Табела 2.1 групирани во интервали.



(а) (б)

Слика 2.2: (а) Хистограм на честоти, Табела 2.3, (б) Хистограм на релативни честоти, Табела 2.3

Друг начин на поделба на интервали ќе даде друга претстава за распределбата на податоците. На пример, ако се одлуциме за $r = 10$ интервали и ако за проширен интервал го земеме интервалот $[1,5-14,5]$, тогаш должината на секој од подинтервалите ќе биде

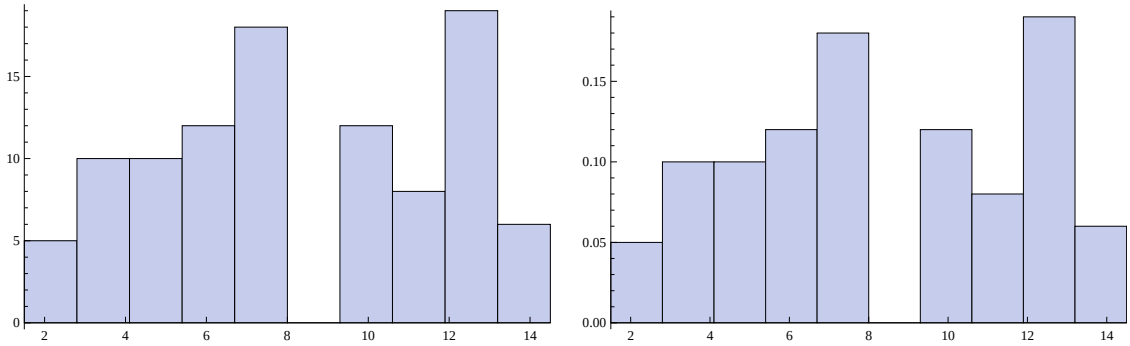
$$h = \frac{14,5 - 1,5}{10} = \frac{13}{10} = 1,3.$$

Групираните податоци во 10 интервали заедно со релативните честоти се прикажани во Табела 2.4. Соодветните графички прикази со хистограм на честоти и хистограм на релативни честоти прикажани се на Слика 2.3. Хистограмите (Слика 2.3) даваат различна претстава за распределбата на истите податоци, но многу поблиска до онаа слика која ја даваат полигоните. ■

интервал	[1,5-2,8]	(2,8-4,1]	(4,1-5,4]	(5,4-6,7]	(6,7-8]	(8-9,3]
честоти	5	10	10	12	18	0
релативни честоти	0,05	0,10	0,10	0,12	0,18	0

интервал	(9,3-10,6]	(10,6-11,9]	(11,9-13,2]	(13,2-14,5]
честоти	12	8	19	6
релативни честоти	0,12	0,08	0,19	0,06

Табела 2.4: Податоците од Табела 2.1 групирани во интервали.



(а) (б)

Слика 2.3: (а) Хистограм на честоти, Табела 2.3, (б) Хистограм на релативни честоти, Табела 2.3

Задача 2.4. На случаен начин се избрани 200 стебла на кои им е измерен дијаметарот на напречниот пресек (во cm) и добиените резултати прикажани се во Табела 2.5.

дијаметар (во cm)	40-43	43-46	46-49	49-52	52-55	55-58	58-61
број на стебла	2	7	40	87	58	5	1

Табела 2.5: Дијаметар на напречниот пресек на стеблото.

Обработи ги дадените податоци (одреди ги бројните карактеристики и направи ги соодветните графички прикази).

Решение. Податоците кои се дадени во оваа задача се веќе групирани во интервали, така да пресметувањето на $\overline{x}$, $\overline{s}^2$ е преку средините на интервалите (Табела 2.6).

$$\begin{aligned}\overline{x} &= \frac{1}{200}(2\cdot 41,5 + 7\cdot 44,5 + 40\cdot 47,5 + 87\cdot 50,5 + \\ &\quad +58\cdot 53,5 + 5\cdot 56,5 + 1\cdot 59,5) = 50,665,\end{aligned}$$

$$\begin{aligned}\overline{s}^2 &= \frac{1}{200}(2\cdot 41,5^2 + 7\cdot 44,5^2 + 40\cdot 47,5^2 + 87\cdot 50,5^2 + \\ &\quad +58\cdot 53,5^2 + 5\cdot 56,5^2 + 1\cdot 59,5^2) - 50,665^2 = 7,76,\end{aligned}$$

$$s^2 = \frac{n}{n-1}\overline{s}^2 = \frac{200}{200-1}\cdot 7,76 = 7,799.$$

Модата е средината на оној интервал кој содржи најголем број на податоци, тоа е интервалот $(49,52]$ кој содржи 87 податоци, па модата е 50,5.

Ирена Стојковска

Дадената табела со честоти, Табела 2.5, ја дополнуваме со средините на интервалите, кумулативни честоти, релативни честоти, и кумулативни релативни честоти. Се добива Табела 2.6.

интервал	40-43	43-46	46-49	49-52	52-55	55-58	58-61
средина на инт.	41,5	44,5	47,5	50,5	53,5	56,5	59,5
честоти	2	7	40	87	58	5	1
кумулат. честоти	2	9	49	136	194	199	200
релативни честоти	0,01	0,035	0,2	0,435	0,29	0,025	0,005
кумул. рел. честоти	0,01	0,045	0,245	0,68	0,97	0,995	1

Табела 2.6: Дополнета Табела 2.5.

Медијаната и кварталите Q_1 и Q_3 се пресметуваат на следниот начин. Бидејќи медијаната е $p = 50$-ти процентил, таа се наоѓа на

$$1 + (n-1) \cdot p\% = 1 + (200-1) \cdot 0,50 = 100,5\text{-тото}$$

место, што значи дека се наоѓа во интервалот $(49, 52]$ и изнесува

$$m = 49 + (100,5 - 49) \cdot \frac{52-49}{87} = 50,77586.$$

Слично, кварталот Q_1 се наоѓа на $1 + (200-1) \cdot 0,25 = 50,75$-тото место, значи повторно во интервалот $(49, 52]$, и изнесува

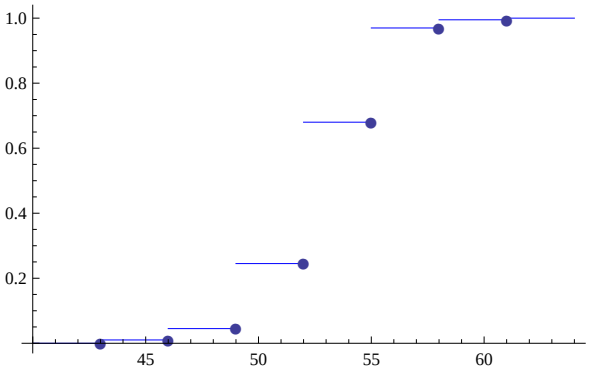
$$Q_1 = 49 + (50,75 - 49) \cdot \frac{52-49}{87} = 49,06034.$$

Кварталот Q_3 се наоѓа на $1 + (200-1) \cdot 0,75 = 150,25$-тото место, значи во интервалот $(52, 55]$, и изнесува

$$Q_3 = 52 + (150,25 - 136) \cdot \frac{55-52}{58} = 52,73707.$$

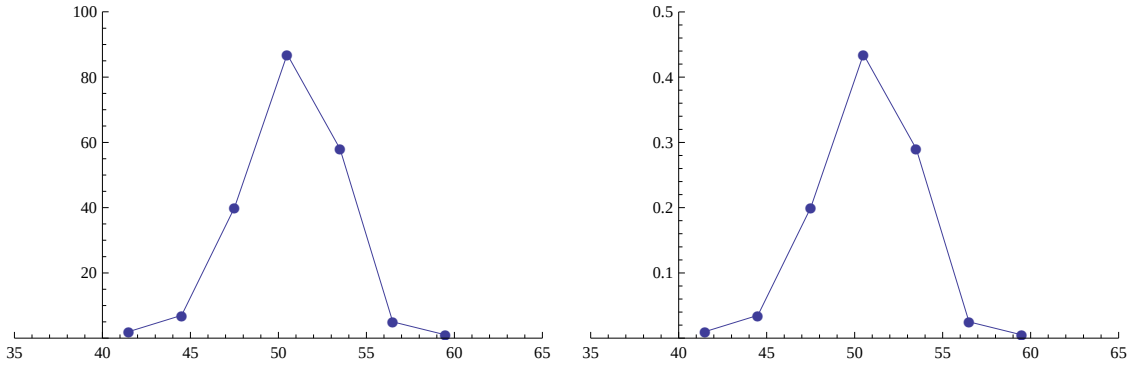
Емпириската функција на распределба $F_n(x) = \frac{n_x}{n}$, $x \in \mathbb{R}$, каде n_x е бројот на податоци со вредност помала од x, заедно со нејзиниот график се прикажани подолу.

$$F_n(x) = \begin{cases} 0 & , x \leq 43 \\ 0,01 & , 43 < x \leq 46 \\ 0,045 & , 46 < x \leq 49 \\ 0,245 & , 49 < x \leq 52 \\ 0,68 & , 52 < x \leq 55 \\ 0,97 & , 55 < x \leq 58 \\ 0,995 & , 58 < x \leq 61 \\ 1 & , x > 61 \end{cases}$$



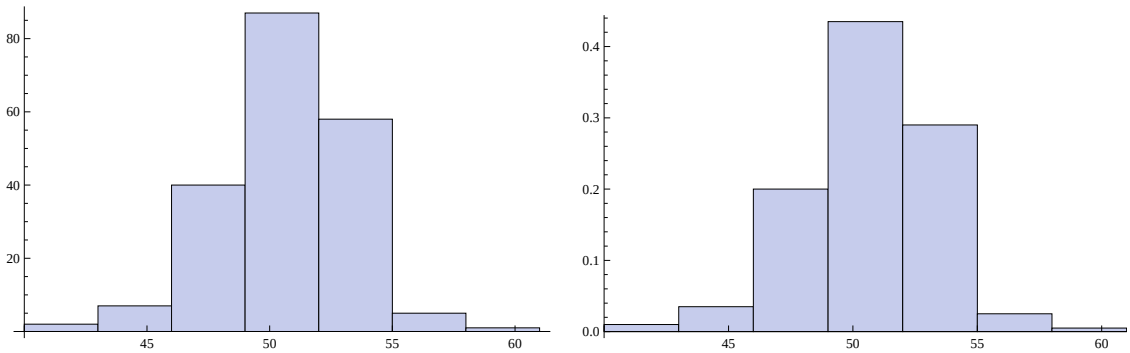
Ирена Стојковска

Полигонот на честоти и полигонот на релативни честоти дадени се на Слика 2.4. Хистограмот на честоти и хистограмот на релативни честоти дадени се на Слика 2.5. Од овие графички прикази се согледува симетричноста на распределбата на податоците. ■



(а) (б)

Слика 2.4: (а) Полигон на честоти, Табела 2.6, (б) Полигон на релативни честоти, Табела 2.6



(а) (б)

Слика 2.5: (а) Хистограм на честоти, Табела 2.6, (б) Хистограм на релативни честоти, Табела 2.6

Задача 2.5. Направено е испитување кај 150 студенти за да се увиди врската меѓу оценката по математика во последната година од средното образование (X) и оценката на испитот по математика на факултет (Y). Добиените резултати прикажани се во Табела 2.7.

X \ Y	5	6	7	8	9	10
2	2	1	0	1	0	0
3	5	35	2	5	0	1
4	3	1	10	15	6	6
5	1	0	0	6	18	32

Табела 2.7: Оценки по предметот математика во последната година од средното образование (X) и оценката на испитот по математика на факултет (Y).

Обработи ги дадените податоци (одреди ги бројните карактеристики и испитај ја зависноста меѓу обележјата).

Решение. За наоѓање на бројните карактеристики, најнапред ги одредуваме маргиналните распределби на фрекфенциите.

X	2	3	4	5
g_i	4	48	41	57

Y	5	6	7	8	9	10
h_j	11	37	12	27	24	39

Табела 2.8: Маргинални фрекфенции.

Тогаш, имаме

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{r} g_i a_i = \frac{1}{150}(4\cdot 2 + 48\cdot 3 + 41\cdot 4 + 57\cdot 5) = 4,006667,$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{s} h_i b_i = \frac{1}{150}(11\cdot 5 + 37\cdot 6 + 12\cdot 7 + 27\cdot 8 + 24\cdot 9 + 39\cdot 10) = 7,886667,$$

$$\overline{s}_x^2 = \frac{1}{n}\sum_{i=1}^{r} g_i a_i^2 - \overline{x}^2 = 0,8066222,$$

$$\overline{s}_y^2 = \frac{1}{n}\sum_{i=1}^{s} h_i b_i^2 - \overline{y}^2 = 2,913822,$$

$$\overline{s}_{xy} = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} a_i b_j f_{ij} - \overline{x}\,\overline{y} = 1,03404,$$

Ирена Стојковска

па коефициентот на корелација е

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = 0,6744826,$$

што значи дека постои статистички сигнификантна корелација ($|r_{xy}| \geq 0,5$) меѓу оценките по математика во последната година од средното образование и истите на испитот по математика на факултет. Понатаму, за да ја испитаме статистичката зависност меѓу обележјата, ги пресметуваме оцтапувањето од статистичката независност

$$f^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{f_{ij}^2}{g_i h_j} - 1 = 1,077948,$$

и степенот на статистичката зависност

$$o = \frac{f^2}{\min\{r,s\}-1} = \frac{1,077948}{4-1} = 0,359827 \approx 36\%,$$

што значи дека меѓу оценката по математика во последната година од средното образование и оценката на испитот по математика на факултет постои статистичка зависност од 36%.

За одредување на кривите на регресија, ги бараме најнапред условните распределби на фрекфенциите, за да ги најдеме соодветните аритметички средини на податоците од условните распределби.

$X\|Y=5$	2	3	4	5
честота	2	5	3	1

$X\|Y=6$	2	3	4	5
честота	1	35	1	0

$X\|Y=7$	2	3	4	5
честота	0	2	10	0

$X\|Y=8$	2	3	4	5
честота	1	5	15	6

$X\|Y=9$	2	3	4	5
честота	0	0	6	18

$X\|Y=10$	2	3	4	5
честота	0	1	6	32

Табела 2.9: Условни распределби на фрекфенции за обележјето X.

Соодветните аритметички средини на податоците од условните распределби за X се

$$\overline{x}(5) = 3,272727,\ \overline{x}(6) = 3,102564,\ \overline{x}(7) = 3,75,$$

$$\overline{x}(8) = 3,962963,\ \overline{x}(9) = 4,428571,\ \overline{x}(10) = 4,658537.$$

Ирена Стојковска

$Y\|X=2$	5	6	7	8	9	10
честота	2	1	0	1	0	0

$Y\|X=3$	5	6	7	8	9	10
честота	5	35	2	5	0	1

$Y\|X=4$	5	6	7	8	9	10
честота	3	1	10	15	6	6

$Y\|X=5$	5	6	7	8	9	10
честота	1	0	0	6	18	32

Табела 2.10: Условни распределби на фрекфенции за обележјето Y.

Соодветните аритметички средини на податоците од условните распределби за Y се

$$\overline{y}(2) = 6,\ \overline{y}(3) = 6,34,\ \overline{y}(4) = 7,92683,\ \overline{y}(5) = 9,196721.$$

Графичкиот приказ на кривата на регресија на Y во зависност од X е искршена линија со темиња во $(a_i, \overline{y}(a_i))$, $i = 1, ..., r$, а кривата на регресија на X во зависност од Y е искршена линија со темиња во $(\overline{x}(b_i), b_j)$, $j = 1, ..., s$. Правите на регресија кои ги апроксимираат овие две криви се

$$y = \overline{y} + \frac{s_{xy}}{s_x^2}(x - \overline{x}) = 1,2819x + 2,7504,$$

$$x = \overline{x} + \frac{s_{xy}}{s_y^2}(y - \overline{y}) = 0,3549y + 1,2079,$$

соодветно.

Задача 2.6. При едно испитување направено на пазарот за посетеноста на супермаркетите, во еден супермаркет, во 40 случајно одбрани денови, измерени се просечниот процент на намалувањето на цените (x) и бројот на посетители во минута (y) и добиени се следните дводимензионални податоци:

(0, 12), (0, 13), (0, 13), (1, 14), (1, 14), (2, 14), (2, 14), (2, 16),
(3, 17), (3, 17), (4, 18), (4, 18), (4, 18), (4, 18), (5, 19), (5, 19),
(6, 19), (6, 19), (6, 19), (7, 19), (7, 20), (7, 20), (7, 20), (7, 20),
(7, 20), (7, 20), (8, 21), (9, 21), (10, 21), (11, 21), (11, 22), (13, 22),
(13, 22), (14, 23), (15, 25), (16, 26), (17, 26), (18, 27), (22, 29), (26, 30)

Обработи ги дадените податоци (одреди ги бројните карактеристики и испитај ја зависноста меѓу обележјата).