2.2 Statistika

Statistika je znanstvena (matematička ?) disciplina koja se bavi prikupljanjem podataka, njihovim sređivanjem i analiziranjem, zatim tumačenjem fenomena na koje se podaci odnose, te konačno praktičnom primjenom rezultata analize.

Opasnost: Često kriva analiza vodi do krivih zaključaka (uzrečica: laž, gnjusna laž, statistika).

Osnovna podjela:

- **Deskriptivna statistika** - manipuliranje statističkim podacima i njihovo opisivanje (grupiranje, grafički prikaz, numerička obrada do određenog nivoa);
- **Matematička statistika** - matematička apstrakcija "konkretnih" empiričkih pojmova;
- **Statističko zaključivanje** - vrlo teško i suptilno.

Osnovni pojmovi:

Statistički skup - skup čiji elementi imaju neko zajedničko obilježje X (ili više obilježja). Dakle, to je skup na kojem vršimo neko istraživanje ili mjerenje. Veličinu X nazivamo **statističko obilježje**:

- **numerička statistička obilježja** (opisujemo brojevima);
- **atributivna statistička obilježja** (opisujemo nekim drugim informacijama).

Numerička statistička obilježja

Ako uzimamo jedan po jedan element iz statističkog skupa i mjerimo veličinu X dobivamo niz podataka: x_1, x_2, x_3, ... Taj niz brojeva nazivamo **niz statističkih podataka.**

Razlikujemo dvije vrste numeričkih statističkih obilježja:

- **diskretno** (statistički podaci su iz diskretnog skupa, npr. $\mathbb{N} \cup \{0\}$, $A = \{-11, -5, 1, 13, ..\}$);
- **kontinuirano** (statistički podaci su iz kontinuiranog skupa, npr. $\mathbb{R}$, intervala $A = (1.2,\ 5.8)$).

Diskretna statistička obilježja

Budući da statističkih podaci x_1, x_2, x_3, $...x_N$ poprimaju vrijednosti a_i iz diskretnog skupa A, onda se za svaki $a_j \in A$ može uočiti broj f_j njegova ponavljanja u nizu x_1, x_2, x_3, $...x_N$. Broj f_j, $f_j \in \{1, 2, ..., r\}$, $r \leq N$, se naziva **frekvencija**, a broj $r_j = \frac{f_j}{N}$ **relativna frekvencija** vrijednosti $a_j \in A$ u nizu od N statističkih podataka x_1, x_2, x_3, $...x_N$ (sugerira način sortiranja podataka u tablici). Vrijedi

$$f_1 + f_2 + ... + f_r = \sum_{j=1}^{r} f_j = N,$$

$$r_1 + r_2 + ... + r_r = \sum_{j=1}^{r} r_j = \sum_{j=1}^{r} \frac{f_j}{N} = \frac{1}{N} \sum_{j=1}^{r} f_j = \frac{N}{N} = 1.$$

Na temelju tabličnog prikaza izrađuju se različiti grafički prikazi (u koordinatnom sustavu):

- **grafikon frekvencija**;
- **grafikon relativnih frekvencija.**

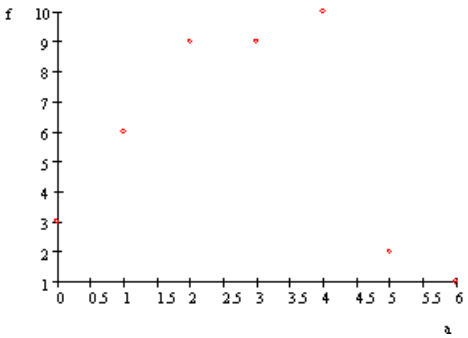
Svakoj različitoj vrijednosti $a_j \in \{a_1, a_2, ..., a_r\}$ obilježja X pridružujemo točku u koordinatnom sustavu tako da joj je apscisa vrijednost a_j, a pripadna ordinata- odgovarajuće frekvencija f_j (relativna frekvencija r_j). Spajanjem tako dobivenih točaka dobiva se **poligon frekvencija (relativnih frekvencija)**.

Primjer 2.8 Broj odsutnih učenika jednog razreda na satu matematike tijekom jednog polugodišta (ukupno 40 sati) dan je nizom statističkih podataka: 2, 5, 1, 1, 3, 4, 4, 4, 2, 3, 3, 4, 0, 0, 4, 4, 3, 6, 1, 4, 2, 2, 4, 3, 2, 1, 3, 2, 2, 5, 4, 0, 3, 2, 1, 2, 4, 1, 3, 3. Odredite tablicu frekvencija i relativnih frekvencija.

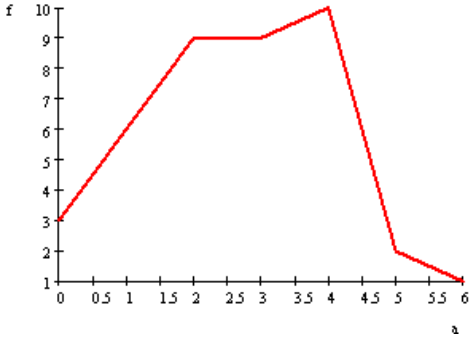
Teoretski, ako je u razredu $n = 30$ učenika, onda je $A = \{0, 1, 2, 3, 4, 5, .., 30\}$. Nadalje, imamo ukupno $N = 40$ podataka, od čega $r = 7$ različitih. Frekvencije i relativne frekvencije su dane tablicom

a_j	f_j	$r_j = \frac{f_j}{N}$
0	3	$\frac{3}{40} = 0.075$
1	6	$\frac{6}{40} = 0.15$
2	9	$\frac{9}{40} = 0.225$
3	9	$\frac{9}{40} = 0.225$
4	10	$\frac{10}{40} = 0.25$
5	2	$\frac{2}{40} = 0.05$
6	1	$\frac{1}{40} = 0.025$
	$N = \sum = 40$	$\sum = 1$

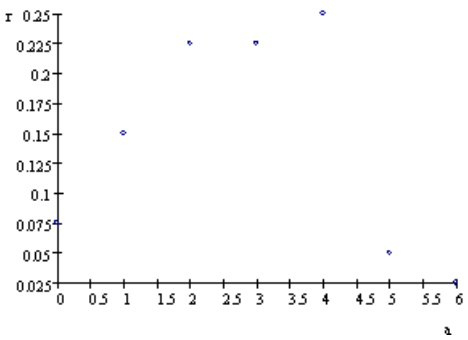
.



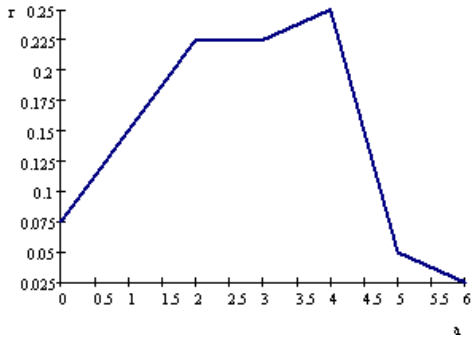
Grafiki prikaz frekvencija



Poligon frekvencija



Grafiki prikaz relativnih frekvencija



Poligon relativnih frekvencija

Parametri niza statističkih podataka (numeričke karakteristike).

Parametri niza statističkih podataka bitna svojstva promatranog statističkog obilježja X izražavaju sažetije.

Aritmetička sredina ili **srednja vrijednost (prosjek)** se definira formulom

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + ... + x_N) = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

ili

$$\bar{x} = \frac{1}{N}(a_1 f_1 + a_2 f_2 + ... + a_r f_r) = \frac{1}{N}\sum_{j=1}^{r} a_j f_j$$

Svojstva aritmetičke sredine:

- Vrijedi

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + ... + (x_N - \bar{x}) = \sum_{i=1}^{N}(x_i - \bar{x}) = 0,$$

  što znači da je zbroj svih odstupanja podataka x_i od njihovog prosjeka $\bar{x}$ nula.

- $$\sum_{i=1}^{N}(x_i - c)^2 > \sum_{i=1}^{N}(x_i - \bar{x})^2, \qquad c \neq \bar{x}$$

  što znači da je zbroj kvadrata svih odstupanja podataka x_i od njihovog prosjeka $\bar{x}$ manji od zbroja kvadrata svih odstupanja podataka x_i od bili kojeg drugog broja $c \neq \bar{x}$.

Varijanca ili disperzija se definira formulom

$$s_0^2 = \frac{1}{N}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_N - \bar{x})^2\right] = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2,$$

ili

$$s_0^2 = \frac{1}{N}\sum_{j=1}^{r}(a_j - \bar{x})^2 f_j = \frac{1}{N}\sum_{j=1}^{r} a_j^2 f_j - \bar{x}^2.$$

Standardna devijacija ili **standardno odstupanje** se definira kao

$$\sigma = \sqrt{s_0^2}.$$

Dakle, varijanca i standardna devijacija mjere rasipanje podataka. Varijanca je prosječno kvadratno odstupanje od prosjeka, dok standardna devijacija daje uvid u položaj i rasipanje danih podataka.

Primjer 2.9 Za podatke iz Primjera 2.8 treba odrediti $\bar{x}$, s_0^2, σ. Dopunimo tablicu iz Primjera 2.8

a_j	f_j	$r_j = \frac{f_j}{N}$	$a_j f_j$	$a_j^2 f_j$
0	3	0.075	0	0
1	6	0.15	6	6
2	9	0.225	18	36
3	9	0.225	27	81
4	10	0.25	40	160
5	2	0.05	10	50
6	1	0.025	6	36
	$\sum = 40$	$\sum = 1$	$\sum = 107$	$\sum = 369$

.

Sada je

$$\begin{aligned}
\bar{x} &= \frac{1}{N}\sum_{j=1}^{r} a_j f_j = \frac{107}{40} = 2.675; \\
s_0^2 &= \frac{1}{N}\sum_{j=1}^{r} a_j^2 f_j - \bar{x}^2 = \frac{369}{40} - (2.675)^2 = 2.069\,4; \\
\sigma &= \sqrt{s_0^2} = \sqrt{2.069\,4} = 1.438\,5.
\end{aligned}$$

Primjer 2.10 Mjerenjem statističkog obilježja X dobiveni su podaci $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, $x_5 = 5$, a mjerenjem statističkog obilježja Y dobiveni su podaci $y_1 = y_2 = y_3 = y_4 = y_5 = 3$. Imamo

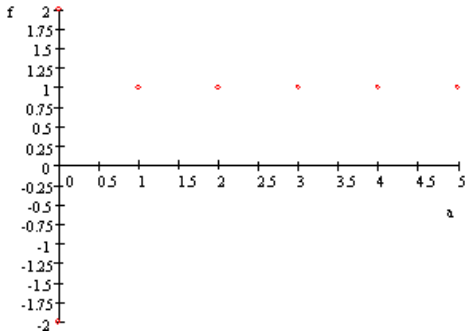
$$\bar{x} = \frac{1 \cdot 1 + 2 \cdot 1 + 3 \cdot 1 + 4 \cdot 1 + 5 \cdot 1}{5} = 3$$

$$\bar{y} = \frac{3 \cdot 5}{5} = 3$$

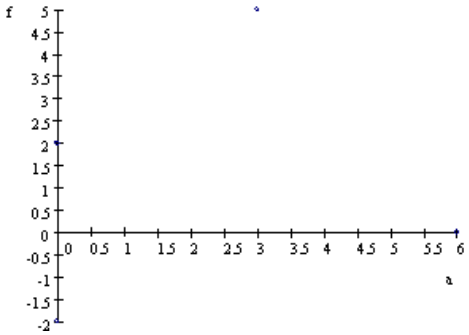
Vidimo da su aritmetičke sredine jednake. Dakle, aritmetičke sredine nam ništa ne govore o raspodjeli podataka. Odredimo zato varijance i disperzije

$$\begin{aligned} s_{0X}^2 &= \frac{1}{N}\sum_{j=1}^{r} a_j^2 f_j - \bar{x}^2 = \frac{1}{5}\sum_{j=1}^{5} a_j^2 f_j - \bar{x}^2 = \\ &= \frac{1}{5}\left(1^2 \cdot 1 + 2^2 \cdot 1 + 3^3 \cdot 1 + 4^4 \cdot 1 + 5^5 \cdot 1\right) - 3^2 = 2, \\ \sigma_X &= \sqrt{s_{0X}^2} = \sqrt{2} \approx 1.41 \\ s_{0Y}^2 &= \frac{1}{N}\sum_{j=1}^{r} a_j^2 f_j - \bar{y}^2 = \frac{1}{5}\sum_{j=1}^{1} a_j^2 f_j - \bar{y}^2 = \frac{1}{5}\left(3^2 \cdot 5\right) - 3^2 = 0, \\ \sigma_Y &= \sqrt{s_{0Y}^2} = \sqrt{0} = 0. \end{aligned}$$

Varijanca $s_{0X}^2 = 2$ i disperzija $\sigma_X \approx 1.41$ nam govore da statističko obilježje X ima veliko rasipanje oko aritmetičke sredine $\bar{x} = 3$, dok varijanca $s_{0Y}^2 = 0$ i disperzija $\sigma_Y = 0$ nam govore da statističko obilježje Y uopće nema rasipanja oko aritmetičke sredine $\bar{y} = 3$.



Prikaz frekvencija stat. obilj. X



Prikaz frekvencija stat. obilj. Y

Primjer 2.11 Rad jednog stroja kontroliran je tako što su u određenim vremenskim razmacima uzimani su uzorci od 50 proizvoda. Za svaki uzorak stanovljeno je koliko sadrži neispravnih proizvoda. Na 20 uzoraka dobiveni su sljedeći podaci: 0 neispravnih proizvoda bilo je u 2 uzorka, 1 neispravan proizvod u 7 uzoraka, 2 u 8, 3 u 2 i 4 neispravna proizvoda u jednom uzorku.

Dakle, statističko obilježje X je broj neispravnih proizvoda u uzorku. Imamo

a_j	f_j	$r_j = \frac{f_j}{N}$	$a_j f_j$	$a_j^2 f_j$
0	2	$\frac{2}{20}$	0	0
1	7	$\frac{7}{20}$	7	7
2	8	$\frac{8}{20}$	16	32
3	2	$\frac{2}{20}$	6	18
4	1	$\frac{1}{20}$	4	16
	$\sum = 20$	$\sum = 1$	$\sum = 33$	$\sum = 73$

,

pa je

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{r} a_j f_j = \frac{33}{20} = 1.\,65,$$

$$s_0^2 = \frac{1}{N}\sum_{j=1}^{r} a_j^2 f_j - \bar{x}^2 = \frac{73}{20} - (1.65)^2 = 0.927\,5,$$

$$\sigma = \sqrt{s_0^2} = \sqrt{0.927\,5} = 0.963\,07,$$

Kontinuirana statistička obilježja

Ako statističko obilježje X poprima vrijednosti iz nekog intervala skupa realnih brojeva $\mathbb{R}$, onda govorimo o **kontinuiranom statističkom obilježju**. (Npr. kontinuirana statistička obilježja su visina, težina, vrijeme, površina,...)

Problem: U nizu statističkih podataka nema među usobno jednakih vrijednosti (nema smisla govoriti o frekvenciji).

Primjer 2.12 Neka je statističko obilježje X visina petnaestogodišnjaka (u cm):

173.2, 165.1, 181.3, 178.3, 190, 142.1, 157.7, 197.6, 174.8, 159, 170.5, 172.1, 146.3, 184.4, 171.5, 165.3, 167.8, 172.2, 182, 159.9, 164.2, 176.6, 191.1, 170.1, 165.3.

Uočimo: nema istih podataka, pa ih treba na neki način grupirati.

Grupiranje:

1. Treba uočiti najmanju vrijednost $x_{\min}$ i najveću vrijednost $x_{\max}$ u nizu statističkih podataka. Dakle, svi podaci nalaze se u intervalu $[x_{\min},\ x_{\max}]$. (U Primjeru 2.12: $x_{\min} = x_6 = 142.1$, a $x_{\max} = x_8 = 197.6$, pa je $[x_{\min},\ x_{\max}] = [142.1,\ 197.6]$).

2. Zbog praktičnosti uzimamo nešto veći interval $[a_0,\ a_r]$ (U Primjeru 2.12: $[142.1,\ 197.6] \subset [140,\ 200] = [a_0,\ a_r]$) i podijelimo taj interval na r manjih inervala $[a_0,\ a_1)$, $[a_1,\ a_2)$, ..., $[a_{r-1}, a_r]$ koje nazivamo **razredi**. Duljina svakog od ovih intervala se naziva **širina razreda** d_j. Obično se uzima, zbog jednostavnosti, da je širina svih razreda jednaka, tj. $d_1 = d_2 = ... = d_r = d$. Broj razreda r ovisi o tome koliko detaljnu analizu želimo (obično uzimamo da je r od $5\% - 10\%$ do 30% broja statističkih podataka N).
   (U Primjeru 2.12: $r = 10$ (40% od $N = 25$), tj. $d = 6$. Naime, $d = \frac{a_r - a_0}{r} = \frac{200-140}{10} = 6$)

3. Određujemo frekvenciju f_j svakog razreda, tj. broj onih podataka u nizu statističkih podataka koji upadaju u interval $[a_{j-1},\ a_j)$, a zatim formiramo tablicu frekvencija za grupirane podatke promatranog statističkog obilježja X. U tablici se navodi i sredina svakog razreda $\overline{a_j} = \frac{a_{j-1}+a_j}{2}$ ($a_{j+1} = a_j + d$, pa je $\overline{a_{j+1}} = \overline{a_j} + d$), te $\overline{a_j} f_j$ i $\overline{a_j}^2 f_j$ (potrebno za računanje parametara stat. podataka).

Primjer 2.12

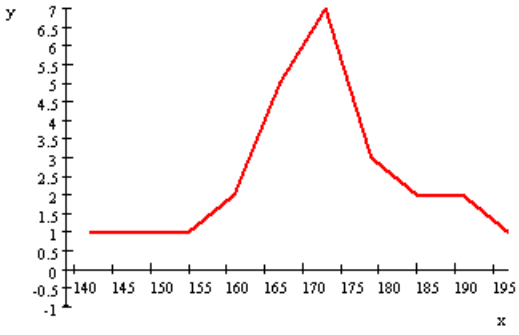
Br. raz.	$[a_{j-1},\ a_j)$ $a_{j-1} - a_j$	$\overline{a_j}$	f_j	$r_j = \frac{f_j}{N}$	$\overline{a_j} f_j$	$\overline{a_j}^2 f_j$
1.	$140 - 146$	143	1	$\frac{1}{25}$	143	20449
2.	$146 - 152$	149	1	$\frac{1}{25}$	149	22201
3.	$152 - 158$	155	1	$\frac{1}{25}$	155	24025
4.	$158 - 164$	161	2	$\frac{2}{25}$	322	51842
5.	$164 - 170$	167	5	$\frac{5}{25}$	835	139445
6.	$170 - 176$	173	7	$\frac{7}{25}$	1211	209503
7.	$176 - 182$	179	3	$\frac{3}{25}$	537	96123
8.	$182 - 188$	185	2	$\frac{2}{25}$	370	68450
9.	$188 - 194$	191	2	$\frac{2}{25}$	382	72962
10.	$194 - 200$	197	1	$\frac{1}{25}$	197	38809
			$\sum = 25$	$\sum = 1$	$\sum = 4301$	$\sum = 743809$

Na temelju tabličnog prikaza izrađuju se različiti grafički prikazi (u koordinatnom sustavu):

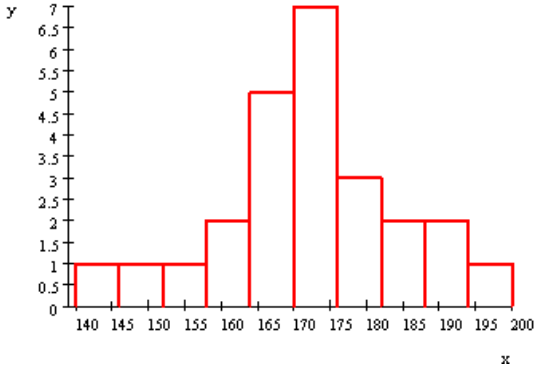
- Ako se na apscisu nanesu sredine razreda $\overline{a_j}$, a za pripadne ordinate uzmu odgovarajuće frekvencije (relativne frekvencije), spajanjem tako dobivenih točaka dobiva se odgovarajući **poligon frekvencija** (**relativnih frekvencija**) grupiranih podataka.

- Ako se na apscisu nanesu rubovi razreda i zatim iznad svakog razreda ucrta pravokutnik visine jednake odgovarajućoj frekvenciji (relativnoj frekvenciji) tog razreda, dobivamo tzv. **histogram frekvencija** (**relativnih frekvencija**) grupiranih podataka.

Primjer 2.12



Poligon frekvencija



Histogram frekvencija

Parametri niza statističkih podataka (numeričke karakteristike)

Aritmetička sredina ili srednja vrijednost (prosjek) grupiranih podataka se definira formulom

$$\tilde{x} = \frac{1}{N}\left(\overline{a_1}f_1 + \overline{a_2}f_2 + ... + \overline{a_r}f_r\right) = \frac{1}{N}\sum_{j=1}^{r}\overline{a_j}f_j$$

Budući se $\tilde{x}$ i obična aritmetička sredina

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i$$

redovito neznatno razlikuju, obično se, zbog praktičnosti uzima $\tilde{x}$ umjesto $\bar{x}$.

Primjer 2.12

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N}x_i = \frac{173.2 + 165.1 + ... + 165.3}{25} = 171.13$$

$$\tilde{x} = \frac{1}{N}\sum_{j=1}^{r}\overline{a_j}f_j = (\text{tablica}) = \frac{4301}{25} = 172.\,04$$

Varijanca ili disperzija grupiranih podataka se definira formulom

$$\begin{aligned}\tilde{s}_0^2 &= \frac{1}{N}\left[(\overline{a_1}-\tilde{x})^2 f_1 + (\overline{a_2}-\tilde{x})^2 f_2 + ... + (\overline{a_r}-\bar{x})^2 f_r\right] = \\ &= \frac{1}{N}\sum_{j=1}^{r}(\overline{a_j}-\tilde{x})^2 f_j\end{aligned}$$

tj.

$$\tilde{s}_0^2 = \frac{1}{N}\sum_{j=1}^{r}(\overline{a_j}-\tilde{x})^2 f_j = \frac{1}{N}\sum_{j=1}^{r}\overline{a}_j^2 f_j - \tilde{x}^2.$$

Budući se $\tilde{s}_0^2$ i obična varijanca $s_0^2 = \frac{1}{N}\sum\limits_{i=1}^{N}(x_i-\bar{x})^2$ redovito neznatno razlikuju, obično se, zbog praktičnosti uzima $\tilde{s}_0^2$ umjesto s_0^2.

Standardna devijacija ili standardno odstupanje grupiranih podataka se definira kao

$$\tilde{\sigma} = \sqrt{\tilde{s}_0^2}.$$

Primjer 2.12

$$\tilde{s}_0^2 = \frac{1}{N}\sum_{j=1}^{r}\overline{a}_j^2 f_j - \tilde{x}^2 = (\text{tablica}) = \frac{743809}{25} - (172.\,04)^2 = 154.60$$

$$\tilde{\sigma} = \sqrt{\tilde{s}_0^2} = \sqrt{154.60} = 12.\,434$$

Primjer 2.13 Težine (u kg) novorođenčadi rođene tijekom jednog tjedna u nekom rodilištu grupirane su na sljedeći način

$a_{j-1}-a_j$	$2-2.5$	$2.5-3$	$3-3.5$	$3.5-4$	$4-4.5$
f_j	2	3	13	17	5

Odredimo $\tilde{x}$, $\tilde{s}_0^2$, $\tilde{\sigma}$.

$a_{j-1} - a_j$	f_j	$\bar{a}_j$	$\bar{a}_j \cdot f_j$	$\bar{a}_j^2 \cdot f_j$
$2 - 2.5$	2	2.25	4.5	10.13
$2.5 - 3$	3	2.75	8.25	22.69
$3 - 3.5$	13	3.25	42.25	137.31
$3.5 - 4$	17	3.75	63.75	239.06
$4 - 4.5$	5	4.25	21.25	90.31
	$\sum = 40$		$\sum = 140$	$\sum = 499.5$

$d = 0.5, \;\; N = \sum f_j = 40,$

$\tilde{x} = \frac{\sum \bar{a}_j \cdot f_j}{N} = \frac{140}{40} = 3.5$

$\tilde{s}_0^2 = \frac{\sum \bar{a}_j^2 \cdot f_j}{N} - (\tilde{x})^2 = \frac{499.5}{40} - (3.\,5)^2 = 0.237\,5$

$\sigma = \sqrt{\tilde{s}_0^2} = \sqrt{0.237\,5} = 0.487\,34 \approx 0.49$