PODATKI, FREKVENČNE PORAZDELITVE IN NJIHOV OPIS: MERE SREDNJE VREDNOSTI IN RAZPRŠENOSTI

1. KAKO NAREDIMO FREKVENČNO PORAZDELITEV

Recimo, da so nam na razpolago podatki (npr. število prijateljev, s katerimi se študentke in študentje videvejo vsak dan):

1, 2, 11, 3, 4, 8, 4, 7, 4, 5, 6, 0, 6, 1, 2, 7, 5, 9, 4, 3,

ali če jih uredimo:

0, 1, 1, 2, 2, 3, 3, 4, 4, 4, 4, 5, 5, 6, 6, 7, 7, 8, 9, 11.

Razdelimo jih v razrede (predstavljajte si posode ali škatle):

od 0 do 2, od 3 do 5, od 6 do 8 in od 9 do 11.

Skica:

Med razredi so "luknje" (od 2 do 3, od 5 do 6 itd.), zato premaknemo meje razredov tako, da te "luknje" pokrijemo. Natančna meja med dvema razredoma je sredina med njunima mejama:

zgornjo mejo nižjega razreda prištejemo spodnji meji višjega in seštevek delimo z 2.

$$\text{meja} = (x_{0\text{gornja}} + x_{1\text{spodnja}}) / 2 = x_{0\text{max}} = x_{1\text{min}}$$

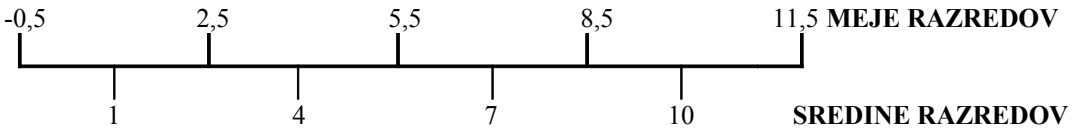
Spodnjo mejo najnižjega razreda in zgornjo mejo najvišjega razreda izberemo tako, da so vsi razredi enako široki. Širina razrede je razlika med natančno zgornjo in natančno spodnjo mejo razreda:

$$\text{širina} = x_{0\text{max}} - x_{0\text{min}},$$

njegova sredina pa polovica seštevka natančne zgornje in natančne spodnje meje razreda:

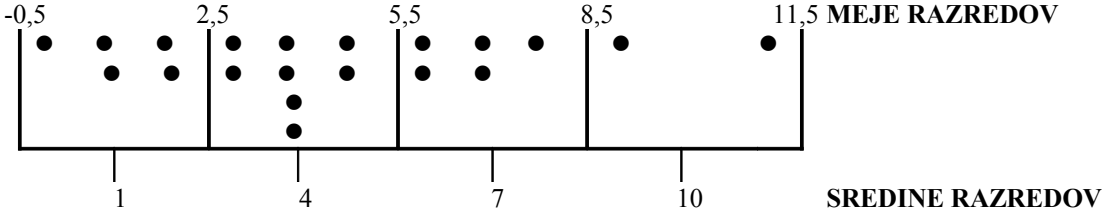
$$\text{sredina} = (x_{0\text{max}} + x_{0\text{min}}) / 2.$$

Skica:



V te razrede razvrstimo naše enote:

Skica:



Za vsak razred ugotovimo frekvenco (pogostost), enot v njem tako, da jih preprosto preštejemo in to zapišemo v tabelo. Izračunamo še odstotke, kumulativne frekvence (potrebne pri ugotavljanju median

in kvartilov), ki jih dobimo tako, da frekvenco danega razreda prištejemo kumulativni frekvenci prejšnjega razreda, in kumulativne odstotke. Dobili smo

Frekvenčno porazdelitev:

	f	f%	F	F%
· 0 - 2	5	25	5	25
· 3 - 5	8	40	13	65
· 6 - 8	5	25	18	90
· 9 - 11	2	10	20	100
Skupaj	20	100		

2. MERE SREDNJE VREDNOSTI

Za naše podatke: 1, 2, 11, 3, 4, 8, 4, 7, 4, 5, 6, 0, 6, 1, 2, 7, 5, 9, 4, 3 bomo določili mere srednje vrednosti.

2.1. ARITMETIČNA SREDINA

Neurejeni podatki:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{20} x_i}{20} = \frac{1+2+11+3+4+8+4+7+4+5+6+0+6+1+2+7+5+9+4+3}{20} = \frac{92}{20} = 4{,}6$$

x_i je vsak podatek (i je zaporedna številka podatka in je od 1 do n, v našem primeru $n = 20$, kot piše na znaku za sumacijo Σ), n je število podatkov.

Iz frekvenčne porazdelitve:

	f	f%	F	F%
0 - 2	5	25	5	25
3 - 5	8	40	13	65
6 - 8	5	25	18	90
9 - 11	2	10	20	100
Skupaj	20	100		

$$\bar{x} = \frac{\sum_{k=1}^{r} (x_k \cdot f_k)}{n} = \frac{\sum_{k=1}^{4} (x_k \cdot f_k)}{20} = \frac{(1 \cdot 5) + (4 \cdot 8) + (7 \cdot 5) + (10 \cdot 2)}{20} = \frac{5 + 32 + 35 + 20}{20} = \frac{92}{20} = 4{,}6$$

x_k so sredine razredov (k je zaporedna številka razreda od 1 do r, v našem primeru $r = 4$ kot piše na znaku za sumacijo Σ), n je število podatkov.

2.2. MEDIANA

Neurejeni podatki:

Iz neurejenih podatkov mediane ne računamo, temveč jo določamo. Podatke uredimo v ranžirno vrsto (uredimo po velikosti ali učeno po rangu):

0, 1, 1, 2, 2, 3, 3, 4, 4, 4, 4, 5, 5, 6, 6, 7, 7, 8, 9, 11

Nato določimo točko, ki razdeli podatke točno na polovico:

tule
⬇
0, 1, 1, 2, 2, 3, 3, 4, 4, 4 4, 5, 5, 6, 6, 7, 7, 8, 9, 11

Če imamo liho število podatkov, je to realni podatek, če je število podatkov sodo, pa izračunamo sredino med podatkoma, v našem primeru (4 + 4) / 2 = 4.

Iz frekvenčne porazdelitve:

	f	f%	F	F%
0 - 2	5	25	5	25
3 - 5	8	40	13	65
6 - 8	5	25	18	90
9 - 11	2	10	20	100
Skupaj	20	100		

Najprej najdemo razred, v katerem leži mediana. To je razred, kjer kumulativni procent **F%** prvič preseže 50%. Temu razredu pravimo medianski razred. V našem primeru je to razred 3 – 5.

$$Me = x_{0,\min} + \frac{\frac{n}{2} - F_{-1}}{f_0} \cdot i = 2{,}5 + \frac{10 - 5}{8} \cdot 3 = 2{,}5 + \frac{5}{8} \cdot 3 = 2{,}5 + 0{,}625 \cdot 3 = 2{,}5 + 1{,}875 = 4{,}375 \cong 4{,}4$$

$x_{0,min}$ je natančna spodnja meja (spomnimo se, da meje vpisane v tabeli niso zmeraj natančne) medianskega razreda, torej razreda, v katerem je mediana (prvi razred, katerega kumulativni odstotek je večji od 50%), F_{-1} je kumulativna frekvenca prejšnjega razreda (kar označi indeks $_{-1}$), f_0 je frekvenca medianskega razreda in i je širina razreda (razdalja med natančno spodnjo in natančno zgornjo mejo razreda).

Kot vidimo, se mediana izračunana na podlagi frekvenčne porazdelitve razlikuje od mediane iz neurejenih podatkov. Razlike lahko nastanejo tudi pri aritmetični sredini in modusu in so posledica izgube informacije, ko ne poznamo več pravih vrednosti podatkov znotraj razredov.

2.3. MODUS

Modus je vrednost, ki se najpogosteje pojavlja ali vrednost, okoli katere so vrednosti najgostejše.

Neurejeni podatki:

Najpametneje je znova podatke urediti v ranžirno vrsto (po velikosti) in najti najpogostejši podatek.

0, 1, 1, 2, 2, 3, 3, **4, 4, 4, 4,** 5, 5, 6, 6, 7, 7, 8, 9, 11

V našem primeru je to 4.

Iz frekvenčne porazdelitve:

	f	f%	F	F%
0 - 2	5	25	5	25
3 - 5	8	40	13	65
6 - 8	5	25	18	90
9 - 11	2	10	20	100
Skupaj	20	100		

Najprej najdemo razred, v katerem je največ enot.
Nato določimo natančni položaj modusa znotraj razreda po naslednji formuli:

$$Mo = x_{0,\min} + \frac{f_0 - f_{-1}}{(f_0 - f_{-1}) + (f_0 - f_{+1})} \cdot 3 = 2{,}5 + \frac{8 - 5}{(8 - 5) + (8 - 5)} \cdot 3 = 2{,}5 + \frac{3}{6} \cdot 3 = 2{,}5 + 1{,}5 = 4$$

$x_{0,min}$ je natančna spodnja meja (spomnimo se znova, da meje vpisane v tabeli niso zmeraj natančne) modalnega razreda, torej razreda, v katerem je modus (v katerega je največ enot), f_0 je frekvenca tega razreda, f_{-1} je frekvenca prejšnjega razreda, f_{+1} je frekvenca naslednjega razreda in i je širina razreda (že vemo: razdalja med natančno spodnjo in natančno zgornjo mejo razreda).

Naj ne zavede: frekvenca prejšnjega razreda in frekvenca naslednjega razreda sta zgolj slučajno enaki. Drugače to ni pravilo, potrebno je natančno vpisati frekvence razredov v formulo.

3. MERE RAZPRŠENOSTI

Za naše podatke: 1, 2, 11, 3, 4, 8, 4, 7, 4, 5, 6, 0, 6, 1, 2, 7, 5, 9, 4, 3, bomo določili še mere razpršenosti.

3.1. KVARTILNI RAZMIK

Da bi določili kvartilni razmik, moramo najprej določiti **kvartile** podatkov. Kvartile določamo tako kot mediano, le da so kvartili **vrednosti**, ki podatke **razdelijo na četrtine**:

Neurejeni podatki:

Za izračun kvartilov neurejene podatke najprej uredimo v ranžirno vrsto, nato pa določimo kvartile.

0, 1, 1, 2, 2 $\boldsymbol{Q_1}$ 3, 3, 4, 4, 4 $\boldsymbol{Q_2}$ 4, 5, 5, 6, 6 $\boldsymbol{Q_3}$ 7, 7, 8, 9, 11

$\boldsymbol{Q_1 = 2{,}5}$ $\boldsymbol{Q_2 = 4}$ $\boldsymbol{Q_3 = 6{,}5}$

Nato preprosto odštejemo prvi kvartil od tretjega in dobimo kvartilni razmik:

$Q = Q_3 - Q_1 = 6{,}5 - 2{,}5 = 4.$

V skladu z utečeno prakso je lepo izračunati še polovični razmik,

$$\frac{Q}{2} = \frac{Q_3 - Q_1}{2} = \frac{6{,}5 - 2{,}5}{2} = \frac{4}{2} = 2$$

ki nam pove, kako daleč od **mediane navzdol in navzgor** (±) se razteza polovica podatkov.

Iz frekvenčne porazdelitve:

	f	f%	F	F%
0 - 2	5	25	5	25
3 - 5	8	40	13	65
6 - 8	5	25	18	90
9 - 11	2	10	20	100
Skupaj	20	100		

Iz podatkov urejenih v frekvenčno porazdelitev izračunamo kvartile po podobni formuli kot mediano, le da določimo kvartilne razrede (za Q_1 prvi razred s kumulativno frekvenco nad eno četrtino, za Q_3 pa prvi razred s kumulativno frekvenco nad tri četrtine) in v formuli ustrezno spremenimo ulomek.

Izračune za naš primer kažejo naslednje formule

$$Q_1 = x_{0,\min} + \frac{\frac{n}{4} - F_{-1}}{f_0} \cdot i = -0{,}5 + \frac{\frac{20}{4} - 0}{5} \cdot 3 = -0{,}5 + \frac{5}{5} \cdot 3 = -0{,}5 + 3 = 2{,}5$$

$$Q_2 = Me = x_{0\min} + \frac{\frac{n}{2} - F_{-1}}{f_0} \cdot i = 2{,}5 + \frac{\frac{20}{2} - 5}{8} = 2{,}5 + \frac{10 - 5}{8} \cdot 3 =$$

$$\frac{5}{8} \cdot 3 = 2{,}5 + 0{,}675 \cdot 3 = 2{,}5 + 1{,}875 = 4{,}375$$

$$Q_3 = x_{0\min} + \frac{\frac{3 \cdot n}{4} - F_{-1}}{f_0} \cdot i = 5{,}5 + \frac{\frac{20 \cdot 3}{4} - 13}{5} \cdot 3 = 5{,}5 + \frac{15 - 13}{5} \cdot 3 =$$

$$5{,}5 + \frac{2}{5} \cdot 3 = 5{,}5 + 0{,}4 \cdot 3 = 5{,}5 + 1{,}2 = 6{,}7$$

Nato znova odštejemo prvi kvartil od tretjega in dobimo kvartilni razmik:

$$Q = Q_3 - Q_1 = 6{,}7 - 2{,}5 = 4{,}2.$$

In znova je v skladu z utečeno prakso lepo izračunati še popovični razmik,

$$\frac{Q}{2} = \frac{Q_3 - Q_1}{2} = \frac{6{,}7 - 2{,}5}{2} = \frac{4{,}2}{2} = 2{,}1$$

.

3.1. POVPREČNI ABSOLUTNI ODKLON

Neurejeni podatki:

Prvi način je podlaga za izračun povprečnega absolutnega odklona. Seštejemo samo absolutne vrednosti odklonov, nato pa jih delimo s številom podatkov (numerusom), ali s formulo:

$$AD = \frac{\sum |x_i - \bar{x}|}{n} = \frac{|1-4{,}6|+|2-4{,}6|+|11-4{,}6|+|3-4{,}6|+|4-4{,}6|+|8-4{,}6|}{20} +$$
$$+ \frac{|4-4{,}6|+|7-4{,}6|+|4-4{,}6|+|5-4{,}6|+|6-4{,}6|+|0-4{,}6|+|6-4{,}6|+|1-4{,}6|}{20} +$$
$$+ \frac{+|2-4{,}6|+|7-4{,}6|+|5-4{,}6|+|9-4{,}6|+|4-4{,}6|+|3-4{,}6|}{20} =$$
$$\frac{3{,}6+2{,}6+6{,}4+1{,}6+0{,}6+3{,}4+0{,}6+2{,}4+0{,}6+0{,}4}{20} +$$
$$+ \frac{1{,}4+4{,}6+1{,}4+3{,}6+2{,}6+2{,}4+0{,}4+4{,}4+0{,}6+1{,}6}{20} = \frac{42{,}2}{20} = 2{,}11$$

kjer pokončni črti pomenita, da upoštevamo le absolutno vrednost izraza med njima.

Kot lahko opazimo, je x_i vsak posamezen podatek, i njegova zaporedna številka, n število podatkov, $\bar{x}$ pa aritmetična sredina naših podatkov.

Za izračun se splača podatke urediti v tabelo:

	x	x-M	|x-M|
1	1	-3,6	3,6
2	2	-2,6	2,6
3	11	6,4	6,4
4	3	-1,6	1,6
5	4	-0,6	0,6
6	8	3,4	3,4
7	4	-0,6	0,6
8	7	2,4	2,4
9	4	-0,6	0,6
10	5	0,4	0,4
11	6	1,4	1,4
12	0	-4,6	4,6
13	6	1,4	1,4
14	1	-3,6	3,6
15	2	-2,6	2,6
16	7	2,4	2,4
17	5	0,4	0,4
18	9	4,4	4,4
19	4	-0,6	0,6
20	3	-1,6	1,6
Skupaj	92	0	45,2

M=92/20= 4,6 AD=45,2/20= 2,26

Iz frekvenčne porazdelitve:

	F	f%	F	F%
0 - 2	5	25	5	25
3 - 5	8	40	13	65
6 - 8	5	25	18	90
9 - 11	2	10	20	100
Skupaj	20	100		

Pri računanju absolutnega odklona iz frekvenčne porazdelitve uporabljamo odklone sredin razredov od aritmetične sredine. Za izračun bomo uporabili aritmetično sredino izračunano iz frekvenčne porazdelitve.

$$AD = \frac{\sum_{k=1}^{r} (|x_k - \bar{x}| \cdot f_k)}{n} = \frac{\sum_{k=1}^{4} (|x_k - \bar{x}| \cdot f_k)}{20} = \frac{(|1 - 4{,}6| \cdot 5) + (|4 - 4{,}6| \cdot 8) + (|7 - 4{,}6| \cdot 5) + (|10 - 4{,}6| \cdot 2)}{20} =$$

$$\frac{(|-3{,}6| \cdot 5) + (|-0{,}6| \cdot 8) + (|2{,}4| \cdot 5) + (5{,}4 \cdot 2)}{20} = \frac{18 + 4{,}8 + 12 + 10{,}8}{20} = \frac{45{,}6}{20} = 2{,}28$$

x_k so sredine razredov (k je zaporedna številka razreda od 1 do r, v našem primeru r= 4 kot piše na znaku za sumacijo Σ), n je število podatkov, pokončni črti pa znova pomenita, da zanemarimo predznak in vzamemo absolutno vrednost odklona.

3.3. VARIANCA IN STANDARDNI ODKLON

Neurejeni podatki:

Drugi način (kvadriranje) je sicer nekaj zamudnejši, vendar se je izkazal kot bolj uporaben za primerjanje večih spremenljivk. Na podlagi kvadratov odklonov izračunamo varianco:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{(1 - 4{,}6)^2 + (2 - 4{,}6)^2 + (11 - 4{,}6)^2 + (3 - 4{,}6)^2 + (4 - 4{,}6)^2 + (8 - 4{,}6)^2}{20} +$$

$$\frac{(4 - 4{,}6)^2 + (7 - 4{,}6)^2 + (4 - 4{,}6)^2 + (5 - 4{,}6)^2 + (6 - 4{,}6)^2 + (0 - 4{,}6)^2 + (6 - 4{,}6)^2 + (1 - 4{,}6)^2}{20} +$$

$$+ \frac{+(2 - 4{,}6)^2 + (7 - 4{,}6)^2 + (5 - 4{,}6)^2 + (9 - 4{,}6)^2 + (4 - 4{,}6)^2 + (3 - 4{,}6)^2}{20} =$$

$$\frac{3{,}6^2 + 2{,}6^2 + 6{,}4^2 + 1{,}6^2 + 0{,}6^2 + 3{,}4^2 + 0{,}6^2 + 2{,}4^2 + 0{,}6^2 + 0{,}4^2 + 1{,}4^2 + 4{,}6^2}{20} +$$

$$+ \frac{1{,}4^2 + 3{,}6^2 + 2{,}6^2 + 2{,}4^2 + 0{,}4^2 + 4{,}4^2 + 0{,}6^2 + 1{,}6^2}{20} = \frac{154{,}8}{20} = 7{,}74$$

Pozor! **Najprej** kvadriramo odklone, **nato** jih seštejemo in **nato** vsoto delimo s številom enot.

Znova je x_i vsak posamezen podatek, i njegova zaporedna številka, n število podatkov, $\bar{x}$ pa aritmetična sredina naših podatkov. Za razliko od absolutnega odklona se predznakov znebimo s kvadriranjem.

Varianca je sestavljena iz kvadratov odklonov, zato je njena enota kvadrat enote podatkov. Da dobimo znova iste enote, kot so v podatkih, jo korenimo. Dobimo standardni odklon:

$$SD = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{7{,}74} = 2{,}78$$

,

ki je splošno sprejeta mera razpršenosti za zvezne (intervalne in razmernostne) podatke.

Naj tu opozorimo, da so te mere razpršenosti le aritmetične sredine asolutnih odklonov ali njihovih kvadratov.

Tudi tu je ureditev podatkov v tabelo pri izračunu veliko ugodnejši način od neposrednega računa, kot je prikazan zgoraj.

	x	x-M	$(x-M)^2$
1	1	-3,6	12,96
2	2	-2,6	6,76
3	11	6,4	40,96
4	3	-1,6	2,56
5	4	-0,6	0,36
6	8	3,4	11,56
7	4	-0,6	0,36
8	7	2,4	5,76
9	4	-0,6	0,36
10	5	0,4	0,16
11	6	1,4	1,96
12	0	-4,6	21,16
13	6	1,4	1,96
14	1	-3,6	12,96
15	2	-2,6	6,76
16	7	2,4	5,76
17	5	0,4	0,16
18	9	4,4	19,36
19	4	-0,6	0,36
20	3	-1,6	2,56
Skupaj	92	0	154,80

M= 4,6 Var= 7,74

SD= 2,78

Iz frekvenčne porazdelitve:

	f	**f%**	**F**	**F%**
0 - 2	5	25	5	25
3 - 5	8	40	13	65
6 - 8	5	25	18	90
9 - 11	2	10	20	100
Skupaj	20	100		

Tudi pri računanju variance in standardnega odklona iz frekvenčne porazdelitve uporabljamo odklone sredin razredov od aritmetične sredine. Za izračun bomo uporabili aritmetično sredino izračunano iz frekvenčne porazdelitve.

$$s^2 = \frac{\sum_{k=1}^{r} ((x_k - \bar{x})^2 \cdot f_k)}{n} = \frac{\sum_{k=1}^{4} ((x_k - \bar{x})^2 \cdot f_k)}{20} =$$

$$\frac{((1-4{,}6)^2 \cdot 5) + ((4-4{,}6)^2 \cdot 8) + ((7-4{,}6)^2 \cdot 5) + ((10-4{,}6)^2 \cdot 2)}{20} =$$

$$\frac{(-3{,}6^2 \cdot 5) + (-0{,}6^2 \cdot 8) + (2{,}4^2 \cdot 5) + (5{,}4^2 \cdot 2)}{20} = \frac{(12{,}96 \cdot 5) + (0{,}36 \cdot 8) + (5{,}76 \cdot 5) + (29{,}16 \cdot 2)}{20} =$$

$$\frac{64{,}8 + 2{,}88 + 28{,}8 + 58{,}32}{20} = \frac{154{,}8}{20} = 7{,}74$$

x_k so sredine razredov (k je zaporedna številka razreda od 1 do r, v našem primeru r= 4 kot piše na znaku za sumacijo Σ), n je število podatkov. Tudi tu se predznakov znebimo s kvadriranjem.

In standardni odklon je seveda koren iz variance:

$$SD = \sqrt{s^2} = \sqrt{\frac{\sum ((x_k - \bar{x})^2 \cdot f_k}{n}} = \sqrt{7{,}74} = 2{,}78$$