Statistika

2. predavanje

Barbara Boldin

Fakulteta za matematiko, naravoslovje in informacijske tehnologije
Univerza na Primorskem

Frekvenčna porazdelitev

Naj ima opisna spremenljivka X končen nabor vrednosti, npr.: $a_1, \ldots, a_n$.

Npr.: za spremenljivko X = krvna skupina so možne vrednosti $0^+, A^+, B^+, 0^-, A^-, B^-, AB^+, AB^-$

Število pojavljanj vrednosti a_j ($j = 1, \ldots, n$) je **frekvenca** enote a_j,

$$\boxed{f_j = \#(X = a_j)}$$

Relativna frekvenca j-te enote je

$$\boxed{f_j^\circ = \frac{f_i}{f_1 + \ldots + f_n}}$$

izražena v odstotkih

$$\boxed{f_j\% = 100 \cdot \frac{f_i}{f_1 + \ldots + f_n}}$$

Porazdelitev frekvenc predstavimo v tabeli

Vrednosti	Frekvenca (f_j)	Relativna frekvenca ($f_j\%$)
a_1	f_1	$f_1\%$
a_2	f_2	$f_2\%$
$\vdots$	$\vdots$	$\vdots$
a_n	f_n	$f_n\%$

in jih grafično predstavimo s **stolpičnim** ali **tortnim** diagramom.

Primer. 60 naključno izbranih povprašamo po njihovem najljubšem okusu sladoleda. Dobimo naslednje odgovore:

V, Č, J, V, P, J, L, Č, P, V, J, Č, V, P, Č, L, Č, J, Č, V, P, L, V, Č, P, Č, J, Č, Č, P, J, Č, V, L, Č, P, V, Č, Č, L, J, Č, P, V, V, J, L, V, P, Č, Č, P, Č, V, L, Č, V, J, V, P

Č = čokolada, V = vanilija, J = jagoda, L= lešnik, P = pistacija

Okus	f_j	$f_j\%$
Čokolada		
Vanilija		
Jagoda		
Lešnik		
Pistacija		

Primer. 60 naključno izbranih povprašamo po njihovem najljubšem okusu sladoleda. Dobimo naslednje odgovore:

V, Č, J, V, P, J, L, Č, P, V, J, Č, V, P, Č, L, Č, J, Č, V, P, L, V, Č, P, Č, J, Č, Č, P, J, Č, V, L, Č, P, V, Č, Č, L, J, Č, P, V, V, J, L, V, P, Č, Č, P, Č, V, L, Č, V, J, V, P

Č = čokolada, V = vanilija, J = jagoda, L= lešnik, P = pistacija

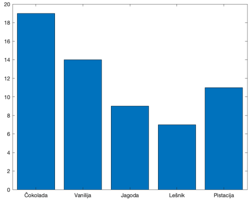
Okus	f_j	$f_j\%$
Čokolada	19	
Vanilija	14	
Jagoda	9	
Lešnik	7	
Pistacija	11	

Primer. 60 naključno izbranih povprašamo po njihovem najljubšem okusu sladoleda. Dobimo naslednje odgovore:

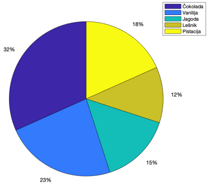
V, Č, J, V, P, J, L, Č, P, V, J, Č, V, P, Č, L, Č, J, Č, V, P, L, V, Č, P, Č, J, Č, Č, P, J, Č, V, L, Č, P, V, Č, Č, L, J, Č, P, V, V, J, L, V, P, Č, Č, P, Č, V, L, Č, V, J, V, P

Č = čokolada, V = vanilija, J = jagoda, L= lešnik, P = pistacija

Okus	f_j	$f_j\%$
Čokolada	19	32
Vanilija	14	23
Jagoda	9	15
Lešnik	7	12
Pistacija	11	18



Stolpični diagram



Tortni diagram

Frekvenčna porazdelitev

Urejenostne in številske spremenljivke

Naj ima urejenostna ali številska spremenljivka X končen nabor vrednosti. Le te lahko uredimo po velikosti, npr.: $a_1 < a_2 < \cdots < a_n$ ter poleg frekvenc f_j izračunamo še **kumulativne frekvence**

$$\boxed{F_j = \#(X \leq a_j) = f_1 + \ldots + f_j}$$

F_j je torej število enot, ki imajo vrednost X največ a_j.

Relativne kumulativne frekvence (v odstotkih) so

$$\boxed{F_j\% = 100 \cdot \frac{F_j}{f_1 + \ldots + f_n}.}$$

$F_j\%$ je odstotek enot, ki imajo vrednost X največ a_j.

Frekvenčna tabela:

Vrednosti	Frekvenca	Relativna frekvenca	F_j	$F_j\%$
a_1	f_1	$f_1\%$	$F_1 = f_1$	$F_1\% = f_1\%$
a_2	f_2	$f_2\%$	$F_2 = F_1 + f_2$	$F_2\% = F_1\% + f_2\%$
$\vdots$	$\vdots$	$\vdots$		
a_n	f_n	$f_n\%$	$F_n = F_{n-1} + f_n$	$F_n\% = F_{n-1}\% + f_n\% = 100$

Primer. Dopolnimo frekvenčno tabelo za rezultate 1. izpita Statistike:

Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22			
6	17			
7	14			
8	9			
9	6			
10	7			
Skupaj				

Primer. Dopolnimo frekvenčno tabelo za rezultate 1. izpita Statistike:

Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22	29.3		
6	17	22.7		
7	14	18.7		
8	9	12		
9	6	8		
10	7	9.3		
Skupaj	75			

Primer. Dopolnimo frekvenčno tabelo za rezultate 1. izpita Statistike:

Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22	29.3	22	29.3
6	17	22.7	39	52
7	14	18.7	53	70.7
8	9	12	62	82.7
9	6	8	68	90.7
10	7	9.3	75	100
Skupaj	75			

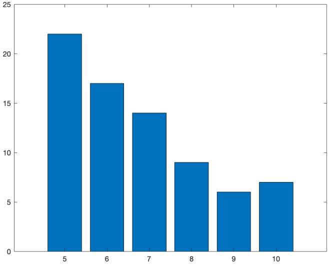
- Kakšen odstotek študentov je dosegel oceno največ 8?
- Kakšen odstotek študentov je dosegel oceno več kot 7?
- Kakšen odstotek študentov je dobil oceno 7 ali 8?

Primer. Dopolnimo frekvenčno tabelo za rezultate 1. izpita Statistike:

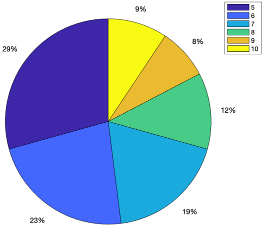
Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22	29.3	22	29.3
6	17	22.7	39	52
7	14	18.7	53	70.7
8	9	12	62	82.7
9	6	8	68	90.7
10	7	9.3	75	100
Skupaj	75			

- Kakšen odstotek študentov je dosegel oceno največ 8? 82.7%
- Kakšen odstotek študentov je dosegel oceno več kot 7? 29.3%
- Kakšen odstotek študentov je dobil oceno 7 ali 8? 30.7%

Grafična predstavitev



Stolpični diagram



Tortni diagram

Tortni diagram je v primeru urejenostnih spremenljivk manj primeren, saj ne prikazuje urejenosti.

Kadar je možnih vrednosti veliko (ali pa X zavzame kontinuum vrednosti), vrednosti razdelimo v nekaj intervalov.
Recimo, da so vrednosti številske spremenljivke razvrščene v K intervalov. Potem frekvenčno porazdelitev opremimo še s

- **spodnjo in zgorno mejo razredov** $x_{i,min}$ in $x_{i,max}$, da je za $i = 1, \ldots, K$

$$x_{i,max} = x_{i+1,min}.$$

  Z besedo: zgornja meja razreda je spodnja meja naslednjega razreda.

- **sredino razredov**

$$x_i = \frac{x_{i,max} + x_{i,min}}{2}.$$

- **širino razredov**

$$d_i = x_{i,max} - x_{i,min}.$$

  Kadar so razredi različno široki, frekvence po razredih niso primerljive. Tedaj za vsak razred izračunamo **gostoto frekvenc**

$$g_i = \frac{f_i}{d_i}.$$

Primer. Za 30 učencev imamo podatke o številu ur odsotnosti v zadnjem šolskem letu:

$$70, 54, 29, 73, 72, 47, 41, 43, 59, 97, 43, 52, 67, 42, 73$$
$$84, 74, 60, 80, 71, 42, 69, 37, 64, 78, 63, 59, 72, 72, 69$$

Odločimo se za razrede: $20 - 39, 40 - 59, 60 - 79, 80 - 99$ in zapišemo frekvenčno tabelo:

Odsotnost (h)	$x_{i,min}$	$x_{i,max}$	d_i	x_i	f_i	$f_i\%$	F_i	$F_i\%$
20-39								
40-59								
60-79								
80-99								

Primer. Za 30 učencev imamo podatke o številu ur odsotnosti v zadnjem šolskem letu:

$$70, 54, 29, 73, 72, 47, 41, 43, 59, 97, 43, 52, 67, 42, 73$$
$$84, 74, 60, 80, 71, 42, 69, 37, 64, 78, 63, 59, 72, 72, 69$$

Odločimo se za razrede: $20 - 39, 40 - 59, 60 - 79, 80 - 99$ in zapišemo frekvenčno tabelo:

Odsotnost (h)	$x_{i,min}$	$x_{i,max}$	d_i	x_i	f_i	$f_i\%$	F_i	$F_i\%$
20-39	19.5	39.5	20	29.5				
40-59	39.5	59.5	20	49.5				
60-79	59.5	79.5	20	69.5				
80-99	79.5	99.5	20	89.5				

Primer. Za 30 učencev imamo podatke o številu ur odsotnosti v zadnjem šolskem letu:

$$70, 54, 29, 73, 72, 47, 41, 43, 59, 97, 43, 52, 67, 42, 73$$
$$84, 74, 60, 80, 71, 42, 69, 37, 64, 78, 63, 59, 72, 72, 69$$

Odločimo se za razrede: $20-39, 40-59, 60-79, 80-99$ in zapišemo frekvenčno tabelo:

Odsotnost (h)	$x_{i,min}$	$x_{i,max}$	d_i	x_i	f_i	$f_i\%$	F_i	$F_i\%$
20-39	19.5	39.5	20	29.5	2	6.7		
40-59	39.5	59.5	20	49.5	10	33.3		
60-79	59.5	79.5	20	69.5	15	50		
80-99	79.5	99.5	20	89.5	3	10		

Primer. Za 30 učencev imamo podatke o številu ur odsotnosti v zadnjem šolskem letu:

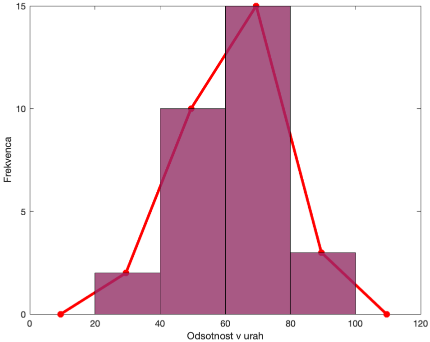
$$70, 54, 29, 73, 72, 47, 41, 43, 59, 97, 43, 52, 67, 42, 73$$

$$84, 74, 60, 80, 71, 42, 69, 37, 64, 78, 63, 59, 72, 72, 69$$

Odločimo se za razrede: $20 - 39, 40 - 59, 60 - 79, 80 - 99$ in zapišemo frekvenčno tabelo:

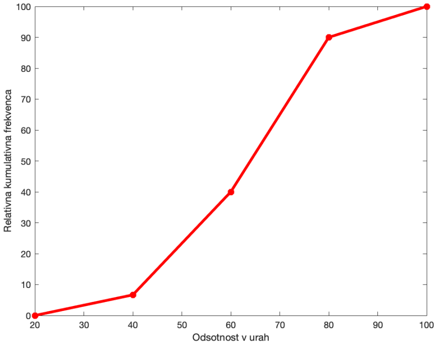
Odsotnost (h)	$x_{i,min}$	$x_{i,max}$	d_i	x_i	f_i	$f_i\%$	F_i	$F_i\%$
20-39	19.5	39.5	20	29.5	2	6.7	2	6.7
40-59	39.5	59.5	20	49.5	10	33.3	12	40
60-79	59.5	79.5	20	69.5	15	50	27	90
80-99	79.5	99.5	20	89.5	3	10	30	100

Grafično lahko podatke predstavimo s **histogramom** ali s **poligonom**.



Poligon dobimo tako, da za vsak razred narišemo točko (x_i, f_i), dodamo še točki $(x_0, 0)$, $(x_{K+1}, 0)$ ter točke povežemo v linijski grafikon.

Še en grafični prikaz je **ogiva**, pri kateri na abscisno os narišemo zgornje meje razredov, na ordinatno os pa pripadajoče relativne kumulativne frekvence. Za vsak razred narišemo točko $(x_{i,max}, F_i\%)$, dodamo še točko $(x_{1,min}, 0)$ in povežemo točke.



Iz ogive lahko razberemo (približne) odgovore na vprašanja kot so: kakšen odstotek učencev ima največ 50 ur odsotnosti? Katero število ur odsotnosti preseže vsaj 30% učencev? Itd.

Ranžirna vrsta in rangi

Če vrednosti $x_1, \dots, x_n$ urejenostne oz. številske spremenljivke x uredimo po velikosti rečemo, da smo jih razvrstili v **ranžirno vrsto**

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}.$$

Elementi $x_{(i)}$ so **vrstilne statistike**.

Rang dane vrednosti je njen položaj v ranžirni vrsti: rang vrednosti x (oznaka $R(x)$) je enak i, če je $x = x_{(i)}$. Ali je rang enolično določen?

Primer. $x_1 = 5, x_2 = 3, x_3 = 9, x_4 = 6, x_5 = 12, x_6 = 1$. Potem je

$$x_{(1)} = 1,\ x_{(2)} = 3,\ x_{(3)} = 5, x_{(4)} = 6,\ x_{(5)} = 9,\ x_{(6)} = 12$$

ter $R(1) = 1, R(3) = 2, R(5) = 3, R(6) = 4, R(9) = 5, R(12) = 6$.

Primer. $x_1 = 1, x_2 = 2, x_3 = 1, x_4 = 0, x_5 = 1, x_6 = 0$. Potem je

$$x_{(1)} = 0,\ x_{(2)} = 0,\ x_{(3)} = 1, x_{(4)} = 1,\ x_{(5)} = 1,\ x_{(6)} = 2.$$

Velja $R(2) = 6$, rang elementa 0 je lahko 1 ali 2, rang elementa 1 pa je lahko 3, 4 ali 5.

Ranžirna vrsta in rangi

Če vrednosti $x_1, \ldots, x_n$ urejenostne oz. številske spremenljivke x uredimo po velikosti rečemo, da smo jih razvrstili v **ranžirno vrsto**

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Elementi $x_{(i)}$ so **vrstilne statistike**.

Rang dane vrednosti je njen položaj v ranžirni vrsti: rang vrednosti x (oznaka $R(x)$) je enak i, če je $x = x_{(i)}$. Ali je rang enolično določen?

Primer. $x_1 = 5, x_2 = 3, x_3 = 9, x_4 = 6, x_5 = 12, x_6 = 1$. Potem je

$$x_{(1)} = 1,\ x_{(2)} = 3,\ x_{(3)} = 5, x_{(4)} = 6,\ x_{(5)} = 9,\ x_{(6)} = 12$$

ter $R(1) = 1, R(3) = 2, R(5) = 3, R(6) = 4, R(9) = 5, R(12) = 6$.

Primer. $x_1 = 1, x_2 = 2, x_3 = 1, x_4 = 0, x_5 = 1, x_6 = 0$. Potem je

$$x_{(1)} = 0,\ x_{(2)} = 0,\ x_{(3)} = 1, x_{(4)} = 1,\ x_{(5)} = 1,\ x_{(6)} = 2.$$

Velja $R(2) = 6$, rang elementa 0 je lahko 1 ali 2, rang elementa 1 pa je lahko 3, 4 ali 5.

Ranžirna vrsta in rangi

Če vrednosti $x_1, \ldots, x_n$ urejenostne oz. številske spremenljivke x uredimo po velikosti rečemo, da smo jih razvrstili v **ranžirno vrsto**

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Elementi $x_{(i)}$ so **vrstilne statistike**.

Rang dane vrednosti je njen položaj v ranžirni vrsti: rang vrednosti x (oznaka $R(x)$) je enak i, če je $x = x_{(i)}$. Ali je rang enolično določen?

Primer. $x_1 = 5, x_2 = 3, x_3 = 9, x_4 = 6, x_5 = 12, x_6 = 1$. Potem je

$$x_{(1)} = 1,\ x_{(2)} = 3,\ x_{(3)} = 5, x_{(4)} = 6,\ x_{(5)} = 9,\ x_{(6)} = 12$$

ter $R(1) = 1, R(3) = 2, R(5) = 3, R(6) = 4, R(9) = 5, R(12) = 6$.

Primer. $x_1 = 1, x_2 = 2, x_3 = 1, x_4 = 0, x_5 = 1, x_6 = 0$. Potem je

$$x_{(1)} = 0,\ x_{(2)} = 0,\ x_{(3)} = 1, x_{(4)} = 1,\ x_{(5)} = 1,\ x_{(6)} = 2.$$

Velja $R(2) = 6$, rang elementa 0 je lahko 1 ali 2, rang elementa 1 pa je lahko 3, 4 ali 5.

Rang torej ni nujno enolično določen. Vsem možnim rangom dane vrednosti rečemo *surovi* rangi. *Spodnji rang* je najmanjši surovi rang, *zgornji rang* pa največji surovi rang.

Vezani rang $R(x)$ je artimetična sredina zgornjega in spodnjega ranga.

$$\boxed{R(x) = \frac{\text{spodnji surovi rang} + \text{zgornji surovi rang}}{2}}$$

Primer. $x_1 = 1, x_2 = 2, x_3 = 1, x_4 = 0, x_5 = 1, x_6 = 0$. Potem je

$$x_{(1)} = 0,\ x_{(2)} = 0,\ x_{(3)} = 1, x_{(4)} = 1,\ x_{(5)} = 1,\ x_{(6)} = 2.$$

Velja $R(0) = 1.5$, $R(1) = 4$, $R(2) = 6$

Relativni ali kvartilni rang elementa x je

$$\boxed{r(x) = \frac{R(x) - \frac{1}{2}}{n}}$$

Za prejšnji primer je $r(0) = \frac{1}{6}$, $r(1) = \frac{7}{12}$, $r(2) = \frac{11}{12}$.

Rang torej ni nujno enolično določen. Vsem možnim rangom dane vrednosti rečemo *surovi* rangi. *Spodnji rang* je najmanjši surovi rang, *zgornji rang* pa največji surovi rang.

Vezani rang $R(x)$ je artimetična sredina zgornjega in spodnjega ranga.

$$\boxed{R(x) = \frac{\text{spodnji surovi rang} + \text{zgornji surovi rang}}{2}}$$

Primer. $x_1 = 1, x_2 = 2, x_3 = 1, x_4 = 0, x_5 = 1, x_6 = 0$. Potem je

$$x_{(1)} = 0,\ x_{(2)} = 0,\ x_{(3)} = 1, x_{(4)} = 1,\ x_{(5)} = 1,\ x_{(6)} = 2.$$

Velja $R(0) = 1.5$, $R(1) = 4$, $R(2) = 6$

Relativni ali kvartilni rang elementa x je

$$\boxed{r(x) = \frac{R(x) - \frac{1}{2}}{n}}$$

Za prejšnji primer je $r(0) = \frac{1}{6}$, $r(1) = \frac{7}{12}$, $r(2) = \frac{11}{12}$.

Rang torej ni nujno enolično določen. Vsem možnim rangom dane vrednosti rečemo *surovi* rangi. *Spodnji rang* je najmanjši surovi rang, *zgornji rang* pa največji surovi rang.

Vezani rang $R(x)$ je artimetična sredina zgornjega in spodnjega ranga.

$$\boxed{R(x) = \frac{\text{spodnji surovi rang} + \text{zgornji surovi rang}}{2}}$$

Primer. $x_1 = 1, x_2 = 2, x_3 = 1, x_4 = 0, x_5 = 1, x_6 = 0$. Potem je

$$x_{(1)} = 0,\ x_{(2)} = 0,\ x_{(3)} = 1, x_{(4)} = 1,\ x_{(5)} = 1,\ x_{(6)} = 2.$$

Velja $R(0) = 1.5$, $R(1) = 4$, $R(2) = 6$

Relativni ali kvartilni rang elementa x je

$$\boxed{r(x) = \frac{R(x) - \frac{1}{2}}{n}}$$

Za prejšnji primer je $r(0) = \frac{1}{6}$, $r(1) = \frac{7}{12}$, $r(2) = \frac{11}{12}$.

Kako določimo vrstilne statistike in range za frekvenčne porazdelitve, npr.:

Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22	29	22	29
6	17	23	39	52
7	14	19	53	71
8	9	12	62	83
9	6	8	68	91
10	7	9	75	100

Vrstilne statistike razberemo iz kumulativnih frekvenc in sicer

$$\boxed{x_{(i)} = a_j, \text{ če je } 1 + F_{j-1} \leq i \leq F_j}$$

Za i-to vrstilno statistiko moramo torej poiskati prvo kumulativno frekvenco, ki je enaka vsaj i.

V zgornjem primeru dobimo npr.:: $x_{(50)} = 7$, $x_{(60)} = 8$, $x_{(70)} = 10$.

Kako določimo vrstilne statistike in range za frekvenčne porazdelitve, npr.:

Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22	29	22	29
6	17	23	39	52
7	14	19	53	71
8	9	12	62	83
9	6	8	68	91
10	7	9	75	100

Vrstilne statistike razberemo iz kumulativnih frekvenc in sicer

$$\boxed{x_{(i)} = a_j, \text{ če je } 1 + F_{j-1} \leq i \leq F_j}$$

Za i-to vrstilno statistiko moramo torej poiskati prvo kumulativno frekvenco, ki je enaka vsaj i.

V zgornjem primeru dobimo npr.:: $x_{(50)} = 7, x_{(60)} = 8$, $x_{(70)} = 10$.

Tudi range določimo iz kumulativnih frekvenc: vrednost a_j ima surove range od $F_{j-1}+1$ do F_j in **vezani rang**

$$\boxed{R(a_j) = \frac{F_{j-1}+F_j+1}{2}}$$

ter **relativni rang**

$$\boxed{r(a_j) = \frac{F_{j-1}+F_j}{2n}}$$

Primer.

Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22	29	22	29
6	17	23	39	52
7	14	19	53	71
8	9	12	62	83
9	6	8	68	91
10	7	9	75	100

Dobimo vezane range $R(5) = 11.5$, $R(6) = 31$, $R(7) = 46.5$, $R(8) = 58$, $R(9) = 65.5$, $R(10) = 72$ in relativne range $r(5) = 0.147$, $r(6) = 0.407$, $r(7) = 0.613$, $r(8) = 0.767$, $r(9) = 0.867$, $r(10) = 0.953$.

Tudi range določimo iz kumulativnih frekvenc: vrednost a_j ima surove range od $F_{j-1}+1$ do F_j in **vezani rang**

$$\boxed{R(a_j) = \frac{F_{j-1}+F_j+1}{2}}$$

ter **relativni rang**

$$\boxed{r(a_j) = \frac{F_{j-1}+F_j}{2n}}$$

Primer.

Ocene	f_j	$f_j\%$	F_j	$F_j\%$
5	22	29	22	29
6	17	23	39	52
7	14	19	53	71
8	9	12	62	83
9	6	8	68	91
10	7	9	75	100

Dobimo vezane range $R(5) = 11.5, R(6) = 31, R(7) = 46.5, R(8) = 58, R(9) = 65.5, R(10) = 72$ in relativne range $r(5) = 0.147, r(6) = 0.407, r(7) = 0.613, r(8) = 0.767, r(9) = 0.867, r(10) = 0.953$.

Kvantili

Kvantil statistične spremenljivke za določen delež je vrednost, pod katero leži približno dani delež podatkov. Najbolj pomembni kvantili so:

- Kvantil za delež $\frac{1}{2}$, ki mu pravimo mediana in ga označimo z Me.
  Mediana razdeli ranžirno vrsto na dve polovici: približno polovica podatkov leži pod mediano, približno polovico nad njo.
- Kvantila za deleža $\frac{1}{3}$ in $\frac{2}{3}$ imenujemo tercila.
  Tercila razdelita ranžirno vrsto na tri približno enake dele: približno tretjina vrednosti leži pod 1. tercilom, približno tretjina med 1. in 2. tercilom in približno tretjina nad 2. tercilom.
- Kvantili za deleže $\frac{1}{4}$, $\frac{2}{4}$ in $\frac{3}{4}$ so kvartili.
  Drugi kvartil je enak mediani.
- Kvantilom za deleže $\frac{j}{10}$ $(j = 1, \ldots, 9)$ pravimo decili.
- Kvantilom za deleže $\frac{j}{100}$ $(j = 1, \ldots, 99)$ pravimo centili.

Mediana je eno od meril *srednje vrednosti* oz. *centralne tendence*. Tudi aritmetična sredina in modus (vrednost oz. razred z najvišjo frekvenco) sta meri centralne tendence.

Definicija kvantila žal ni enotna. Matematična definicija: rečemo, da je vrednost Q_γ **kvantil spremenljivke X za delež γ**,če je

$$\boxed{\frac{\#(X < Q_\gamma)}{n} \leq \gamma \ \text{ in } \ \frac{\#(X \leq Q_\gamma)}{n} \geq \gamma}.$$

Primer. Za ranžirno vrsto

$$1, 2, 3, 3, 7, 15, 32, 47, 69$$

na oko opazimo, da je $Me = 7$. Izračun po definiciji pove, da je mediana število, ki zadošča pogojema $\frac{\#(X < Q_{0.5})}{9} \leq 0.5$ in $\frac{\#(X \leq Q_{0.5})}{9} \geq 0.5$, oz. $\#(X < Q_{0.5}) \leq 4.5$ in $\#(X \leq Q_{0.5}) \geq 4.5$. Prvi pogoj izpolnjujejo vrednosti do vključno 7, drugega pa od vključno 7 dalje. Edina možnost za mediano je torej 7.

Primer. Za ranžirno vrsto

$$1, 2, 3, 3, 7, 15, 32, 47, 69, 91$$

je mediana število, ki zadošča pogojema $\#(X < Q_{0.5}) \leq 5$ in $\#(X \leq Q_{0.5}) \geq 5$. Prvi pogoj izpolnjujejo vrednosti do vključno 15, drugega pa od vključno 7 dalje. Vsako število med 7 in 15 je torej lahko mediana. Kvantili torej niso enolično določeni.

Kaj opazite za ranžirne vrste s sodim/lihim številom podatkov?

Definicija kvantila žal ni enotna. Matematična definicija: rečemo, da je vrednost Q_γ **kvantil spremenljivke X za delež γ**, če je

$$\boxed{\frac{\#(X < Q_\gamma)}{n} \leq \gamma \ \text{ in } \ \frac{\#(X \leq Q_\gamma)}{n} \geq \gamma}.$$

Primer. Za ranžirno vrsto

$$1, 2, 3, 3, 7, 15, 32, 47, 69$$

na oko opazimo, da je $Me = 7$. Izračun po definiciji pove, da je mediana število, ki zadošča pogojema $\frac{\#(X < Q_{0.5})}{9} \leq 0.5$ in $\frac{\#(X \leq Q_{0.5})}{9} \geq 0.5$, oz. $\#(X < Q_{0.5}) \leq 4.5$ in $\#(X \leq Q_{0.5}) \geq 4.5$. Prvi pogoj izpolnjujejo vrednosti do vključno 7, drugega pa od vključno 7 dalje. Edina možnost za mediano je torej 7.

Primer. Za ranžirno vrsto

$$1, 2, 3, 3, 7, 15, 32, 47, 69, 91$$

je mediana število, ki zadošča pogojema $\#(X < Q_{0.5}) \leq 5$ in $\#(X \leq Q_{0.5}) \geq 5$. Prvi pogoj izpolnjujejo vrednosti do vključno 15, drugega pa od vključno 7 dalje. Vsako število med 7 in 15 je torej lahko mediana. Kvantili torej niso enolično določeni.

Kaj opazite za ranžirne vrste s sodim/lihim številom podatkov?

Primer. Izračunajmo kvartile za ranžirno vrsto:

$$1, 2, 3, 3, 7, 15, 32, 47, 69, 91, 112, 250, 327, 512, 694$$

Uporabimo dejstvo $R = \gamma n + 0.5$ in z upoštevanjem $n = 15$

- $R(Q_{0.25}) = 0.25 \cdot 15 + 0.5 = 4.25$. Prvi kvartil je torej katerokoli število na intervalu $[3, 7]$.
- $R(Q_{0.5}) = 0.5 \cdot 15 + 0.5 = 8$. Drugi kvartil (oz. mediana) je torej število 47.
- $R(Q_{0.75}) = 0.75 \cdot 15 + 0.5 = 11.75$. Tretji kvartil je torej katerokoli število na intervalu $[112, 250]$.

Medkvartilni razmik je $Q = Q_{0.75} - Q_{0.25}$. Interval $[Q_{0.25}, Q_{0.75}]$ vsebuje torej približno polovico sredinskih vrednosti.

Kadar kvantil ni enolično določen lahko predstavnika izberemo z *interpolacijo*: če kvantil leži med vrednostima a in b in za rang kvantila velja $R(a) < R(Q) < R(b) = R(a) + 1$, potem Q določimo z

$$\boxed{Q = a + (R(Q) - R(a))(b - a)}.$$

Za prejšnji primer je predstavnik 1. kvartila $Q_{0.25} = 3 + (4.25 - 4)(7 - 3) = 4$, predstavnik 3. kvartila pa $Q_{0.75} = 112 + (11.75 - 11)(250 - 112) = 215.5$.

Primer. Izračunajmo kvartile za ranžirno vrsto:

$$1, 2, 3, 3, 7, 15, 32, 47, 69, 91, 112, 250, 327, 512, 694$$

Uporabimo dejstvo $R = \gamma n + 0.5$ in z upoštevanjem $n = 15$

- $R(Q_{0.25}) = 0.25 \cdot 15 + 0.5 = 4.25$. Prvi kvartil je torej katerokoli število na intervalu $[3, 7]$.
- $R(Q_{0.5}) = 0.5 \cdot 15 + 0.5 = 8$. Drugi kvartil (oz. mediana) je torej število 47.
- $R(Q_{0.75}) = 0.75 \cdot 15 + 0.5 = 11.75$. Tretji kvartil je torej katerokoli število na intervalu $[112, 250]$.

Medkvartilni razmik je $Q = Q_{0.75} - Q_{0.25}$. Interval $[Q_{0.25}, Q_{0.75}]$ vsebuje torej približno polovico sredinskih vrednosti.

Kadar kvantil ni enolično določen lahko predstavnika izberemo z *interpolacijo*: če kvantil leži med vrednostima a in b in za rang kvantila velja $R(a) < R(Q) < R(b) = R(a) + 1$, potem Q določimo z

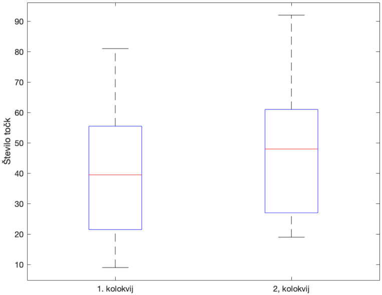
$$\boxed{Q = a + (R(Q) - R(a))(b - a)}.$$

Za prejšnji primer je predstavnik 1. kvartila $Q_{0.25} = 3 + (4.25 - 4)(7 - 3) = 4$, predstavnik 3. kvartila pa $Q_{0.75} = 112 + (11.75 - 11)(250 - 112) = 215.5$.

Urejenostne spremenljivke lahko grafično predstavimo s **škatlo z brki** (angl. box plot), kjer so prikazani minimalna in maksimalna vrednost ter kvartili.

Primer. Rezultati obeh kolokvijev Statistike so:

1. kolokvij: 9, 11, 12, 15, 16, 17, 26, 30, 34, 35, 36, 37, 42, 45, 49, 50, 51, 54, 57, 62, 63, 65, 69, 81
2. kolokvij: 19, 19, 20, 24, 27, 27, 36, 45, 47, 47, 48, 48, 49, 57, 57, 60, 61, 63, 64, 65, 69, 92



- Določi število točk na 1. kolokviju, ki ga je presegla ena četrtina študentov.
- Med katerimi vrednostimi najdemo sredinskih 50% rezultatov 1. in 2. kolokvija?

Kadar so podatki dani le s frekvenčno tabelo, potem osnovnih podatkov nimamo in kvantile izračunamo iz kumulativnih frekvenc s pomočjo interpolacije.

Primer.

Odsotnost (h)	$x_{i,min}$	$x_{i,max}$	d_i	x_i	f_i	$f_i\%$	F_i	$F_i\%$
20-39	19.5	39.5	20	29.5	2	6.7	2	6.7
40-59	39.5	59.5	20	49.5	10	33.3	12	40
60-79	59.5	79.5	20	69.5	15	50	27	90
80-99	79.5	99.5	20	89.5	3	10	30	100

Jasno je, da je 1. kvartil med 40 in 59, mediana in 3. kvartil pa med 60 in 79.

Če kvantil Q_γ leži v razredu s spodnjo mejo $x_{0,min}$, zgornjo mejo $x_{0,max}$, frekvenco f_0 ter širino d_0, potem Q_γ izračunamo z

$$\boxed{Q_\gamma = x_{0,min} + \frac{F(Q_\gamma)-F(x_{0,min})}{f_0} \cdot d_0}$$

kjer je

$$\boxed{F(Q_\gamma) = n \cdot \gamma + 0.5}.$$

Za zgornji primer: za prvi kvartil $Q_{0.25}$ je $x_{0,min} = 39.5$, $x_{0,max} = 59.5$, $d_0 = 20$, $f_0 = 10$ in $F(Q_{0.25}) = 8$, torej $Q_{0.25} = 51.5$. Dobimo še $Q_{0.5} = Me = 64.2$ in $Q_{0.75} = 74.2$. Interpretacija?

Kadar so podatki dani le s frekvenčno tabelo, potem osnovnih podatkov nimamo in kvantile izračunamo iz kumulativnih frekvenc s pomočjo interpolacije.

Primer.

Odsotnost (h)	$x_{i,min}$	$x_{i,max}$	d_i	x_i	f_i	$f_i\%$	F_i	$F_i\%$
20-39	19.5	39.5	20	29.5	2	6.7	2	6.7
40-59	39.5	59.5	20	49.5	10	33.3	12	40
60-79	59.5	79.5	20	69.5	15	50	27	90
80-99	79.5	99.5	20	89.5	3	10	30	100

Jasno je, da je 1. kvartil med 40 in 59, mediana in 3. kvartil pa med 60 in 79.

Če kvantil Q_γ leži v razredu s spodnjo mejo $x_{0,min}$, zgornjo mejo $x_{0,max}$, frekvenco f_0 ter širino d_0, potem Q_γ izračunamo z

$$\boxed{Q_\gamma = x_{0,min} + \frac{F(Q_\gamma) - F(x_{0,min})}{f_0} \cdot d_0}$$

kjer je

$$\boxed{F(Q_\gamma) = n \cdot \gamma + 0.5}.$$

Za zgornji primer: za prvi kvartil $Q_{0.25}$ je $x_{0,min} = 39.5$, $x_{0,max} = 59.5$, $d_0 = 20$, $f_0 = 10$ in $F(Q_{0.25}) = 8$, torej $Q_{0.25} = 51.5$. Dobimo še $Q_{0.5} = Me = 64.2$ in $Q_{0.75} = 74.2$. Interpretacija?