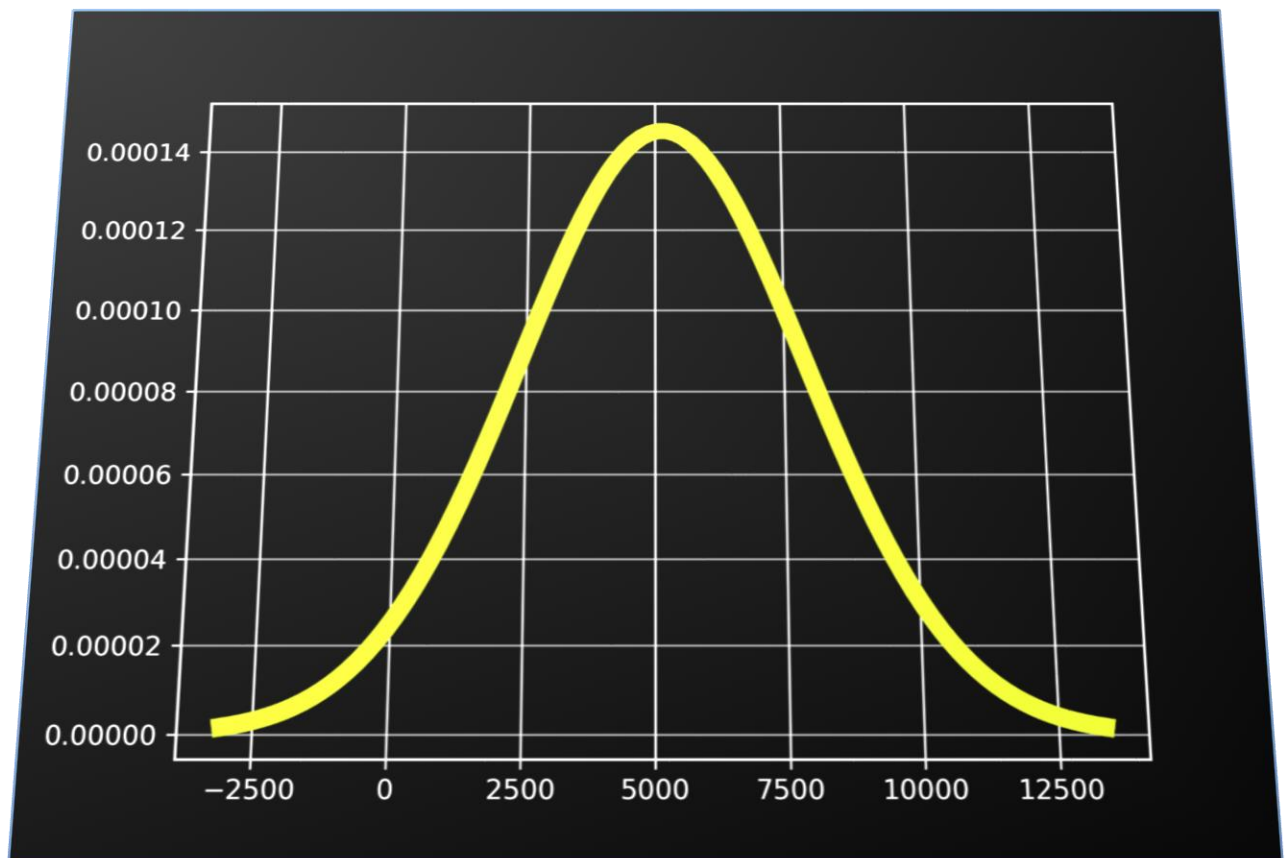


SEMINARSKI RAD IZ KOLEGIJA

# OSNOVE VJEROJATNOSTI I STATISTIKE

Izradili:

Marino Linić,  
Dorian Manjarić,  
Fran Poje,  
Luka Krulčić,  
Ivan Matejčić



*Normalna distribucija nasumično generiranih podataka iz uzorka u radu*

*Rijeka, 19. lipnja 2021.*

# Tablica sadržaja

Korišteni Python moduli i biblioteke.....	3
Uvod.....	4
Deskriptivna statistika.....	5
Tablica frekvencija.....	5
Histogram frekvencija i relativnih frekvencija .....	6
Kumulativne frekvencije.....	8
Mjere centralne tendencije.....	9
Položajne mjere centralne tendencije .....	11
Raspon varijacije .....	12
Kvantili .....	13
Percentili.....	14
Suma apsolutnih vrijednosti odstupanja od srednje vrijednosti.....	15
Prosječno apsolutno odstupanje od aritmetičke sredine.....	15
Varijanca.....	15
Standarda devijacija .....	16
Korigirana varijanca .....	16
Korigirana standardna devijacija.....	16
Kutijasti (Box-plot) dijagram .....	17
Koeficijent asimetrije.....	18
Pearsonova mjera asimetrije .....	18
Bowleyjeva mjera asimetrije .....	19
Mjera zaobljenosti.....	20
Linijski graf za distribuciju podataka .....	21
Histogrami frekvencija sa svim mjerama uzorka.....	22
Intervali pouzdanosti .....	23
Zaključak .....	24
Cijeli ispis u konzoli.....	25
Kôd .....	25
Tablica opisa slika .....	26

# Korišteni Python moduli i biblioteke

Za izradu seminarskog rada, kao i za rad sa uzorkom, korištena je nekolicina python-ovih biblioteka i modula. Ispod su navedene knjižice, popraćene objašnjenjem.

**Python Statistics** – biblioteka koja omogućuje rad sa standardnim statističkim funkcionalnostima poput: aritmetičke sredine, geometrijske sredine, harmonijske sredine, itd.

**Python Scipy** – biblioteka koja za znanstvene i tehničke izračune poput: linearne algebre, integrale, signalne izračune, statističke izračune i sl.

**Python Numpy** – biblioteka koja omogućuje rad sa višedimenzionalnim listama, vektorima, matricama i dr.

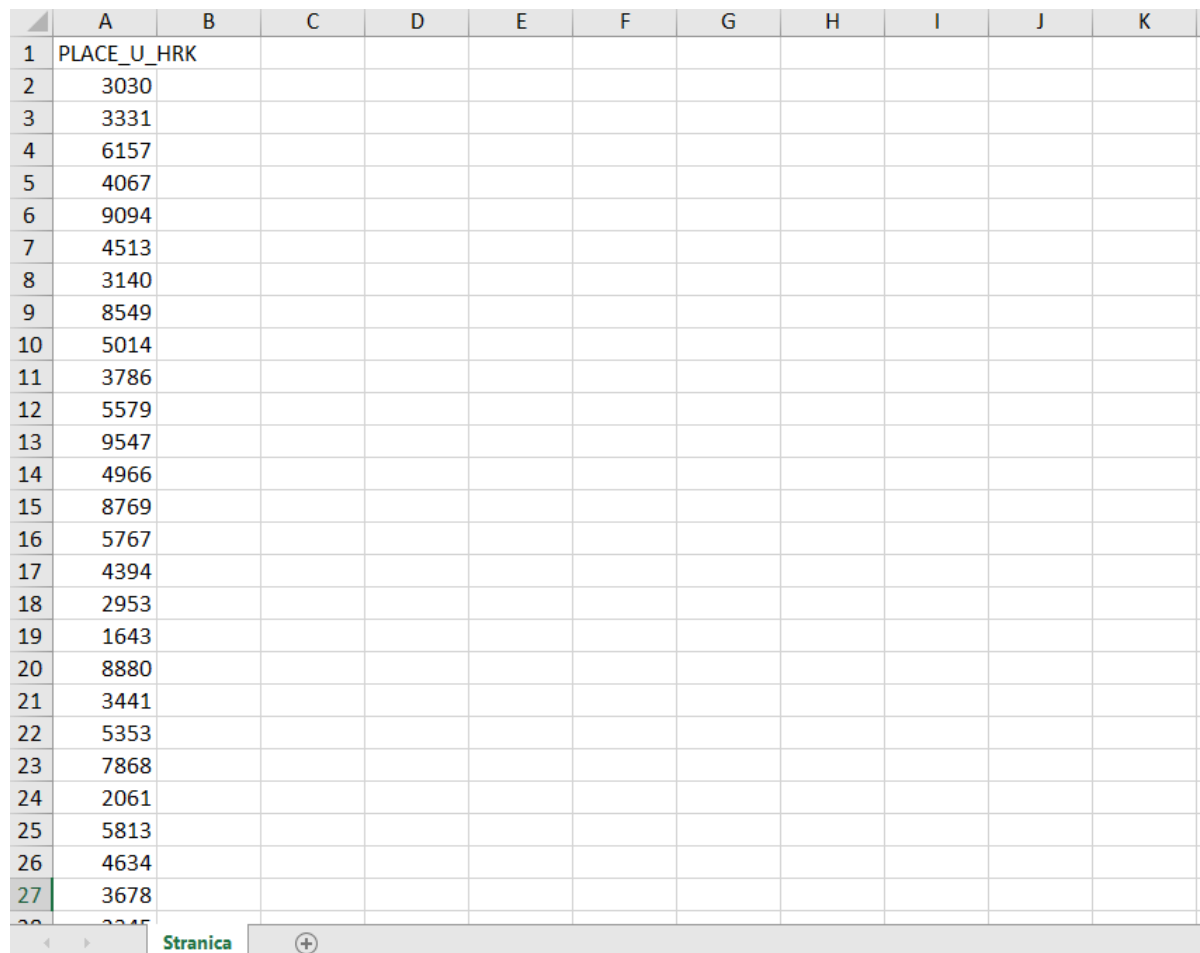
**Python Pandas** – biblioteka koja omogućuje funkcionalnosti za rad s podacima.

**Python Matplotlib** – biblioteka pomoću koje možemo crtati grafove i generirati slike grafova.

**Python Tabulate** – biblioteka koja nam omogućuje formatiranje generiranih podataka u uređenu tablicu.

# Uvod

Za ovaj rad uzeli smo izmišljeni i nasumično generirani uzorak od 200 podataka u iznosima plaća sa web stranice <https://randat.com> i spremili smo ga u datoteku **place.xml**. U ovom istraživanju promatramo numeričke varijable, odnosno kvantitativne.



	A	B	C	D	E	F	G	H	I	J	K
1	PLACE_U_HRK										
2	3030										
3	3331										
4	6157										
5	4067										
6	9094										
7	4513										
8	3140										
9	8549										
10	5014										
11	3786										
12	5579										
13	9547										
14	4966										
15	8769										
16	5767										
17	4394										
18	2953										
19	1643										
20	8880										
21	3441										
22	5353										
23	7868										
24	2061										
25	5813										
26	4634										
27	3678										
28	3345										

Slika 1. XML dokument "place.xml" s našim uzorkom od 200 podataka.

Kôd u programskom jeziku Python pisali smo zajedno u online okruženju <https://replit.com>. Ujedno i preporučamo pokretanje kôda u istom okruženju. Sve slike i tablice, ukoliko nisu dovoljne veličine, mogu se povećati korištenjem raznih preglednih alata, ali nalaze se i u zapakiranoj RAR datoteci.

# Deskriptivna statistika

## Tablica frekvencija

Tablicu frekvencija smo ispisali u konzolu i u datoteku „**tablica.txt**“.

1. „*Grupa (razred)*“ stupac – označava 10 razreda s jednakom udaljenosti ((maksimalna vrijednost – minimalna vrijednost) / br. razreda) po iznosu plaća, s obzirom na to da želimo sažeti informacije o plaćama i intuitivno ih analizirati (bez grupacije bi tablica bila samo granično korisnija od sirovih podataka).
2. „*Iznos plaće u razredu*“ stupac – označava prosječnu plaću u razredu (pojam prosjeka ćemo kasnije navesti pod podnaslovom „Mjere centralnih tendencija“ i nadalje *aritmetičkom sredinom*)
3. „*Frekvencija*“ stupac – označava broj podataka u razredu, odnosno frekvenciju pojavljivanja razreda unutar uzorka. Konkretnije, ako je vrijednost podatka iz uzorka unutar ranga razreda, onda se pridodaje frekventnosti toga razreda.
4. „*Relativna frekvencija*“ stupac – označava frekvencije razreda unutar skupa razreda, odnosno da je totalna frekvencija = 1.

Grupa (razred)	Iznos plaće u razredu	Frekvencija	Relativna frekvencija
1	1483.9	18	0.09
2	2427.8	24	0.12
3	3371.7	22	0.11
4	4315.6	23	0.115
5	5259.5	20	0.1
6	6203.4	18	0.09
7	7147.3	18	0.09
8	8091.2	17	0.085
9	9035.1	19	0.095
10	9979	20	0.1

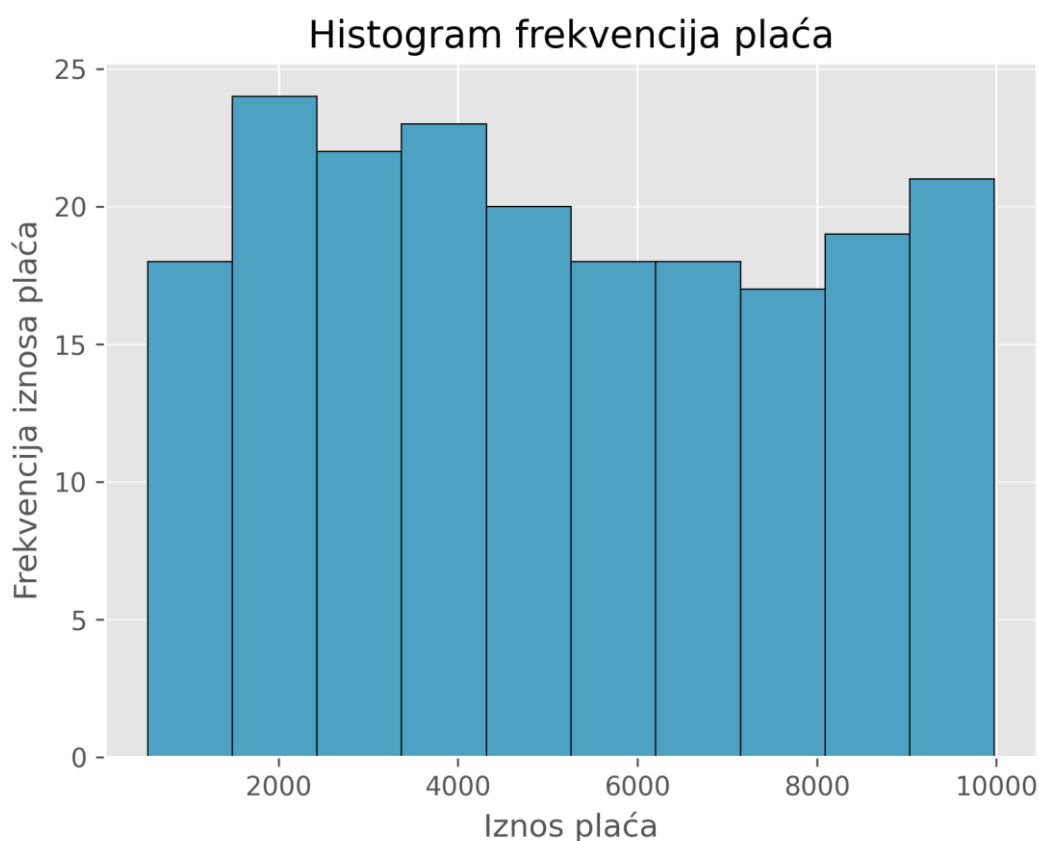
Slika 2. Tablica frekvencija



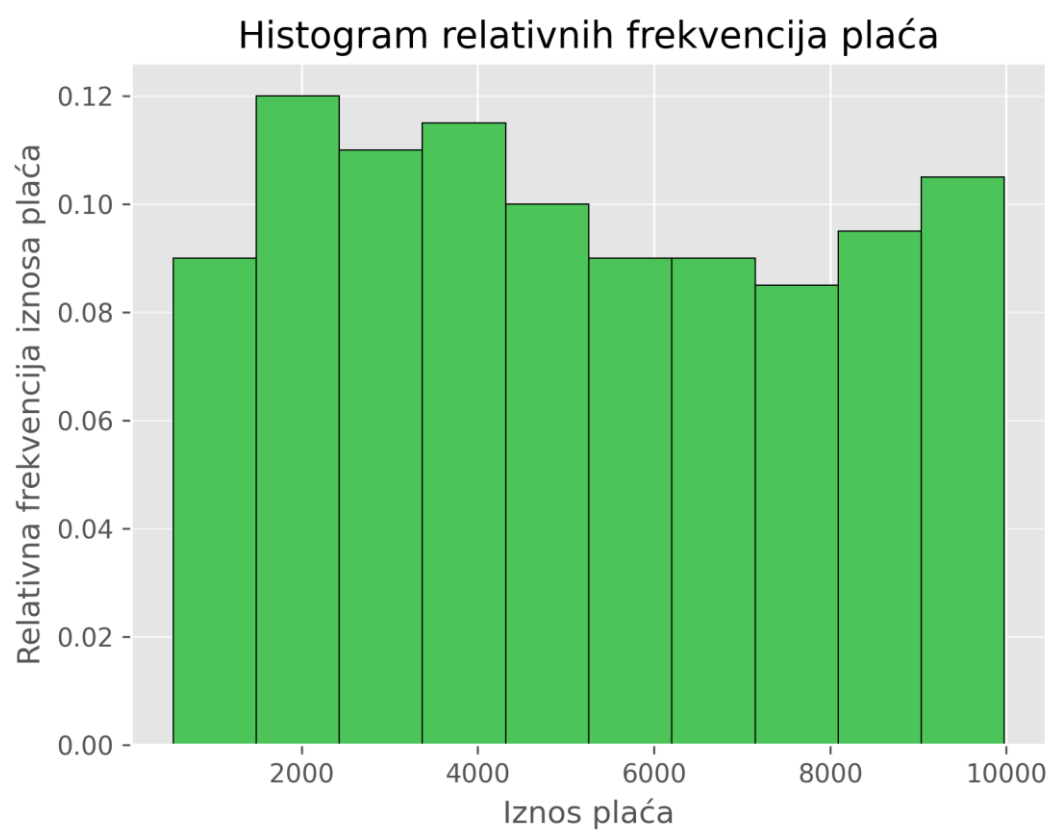
tablica.txt

## Histogram frekvencija i relativnih frekvencija

Histogram je grafički prikaz odnosa podataka i njihovih frekvencija. Izgleda poput stupčastog grafikona. U našem slučaju, podaci su grupirani u razrede od 10 na X osi, te frekvencije do 25 na Y osi. Sve grafičke prikaze smo definirali koristeći Python biblioteku „Matplotlib“, stoga je nakon ovoga nećemo više navoditi. U kôdu smo naveli veličinu razreda na x osi pomoću jedne od funkcija biblioteke.



*Slika 3. Histogram frekvencija plaća*



Slika 4. Histogram relativnih frekvencija plaća

Kraj 1. zadatka.

## Kumulativne frekvencije

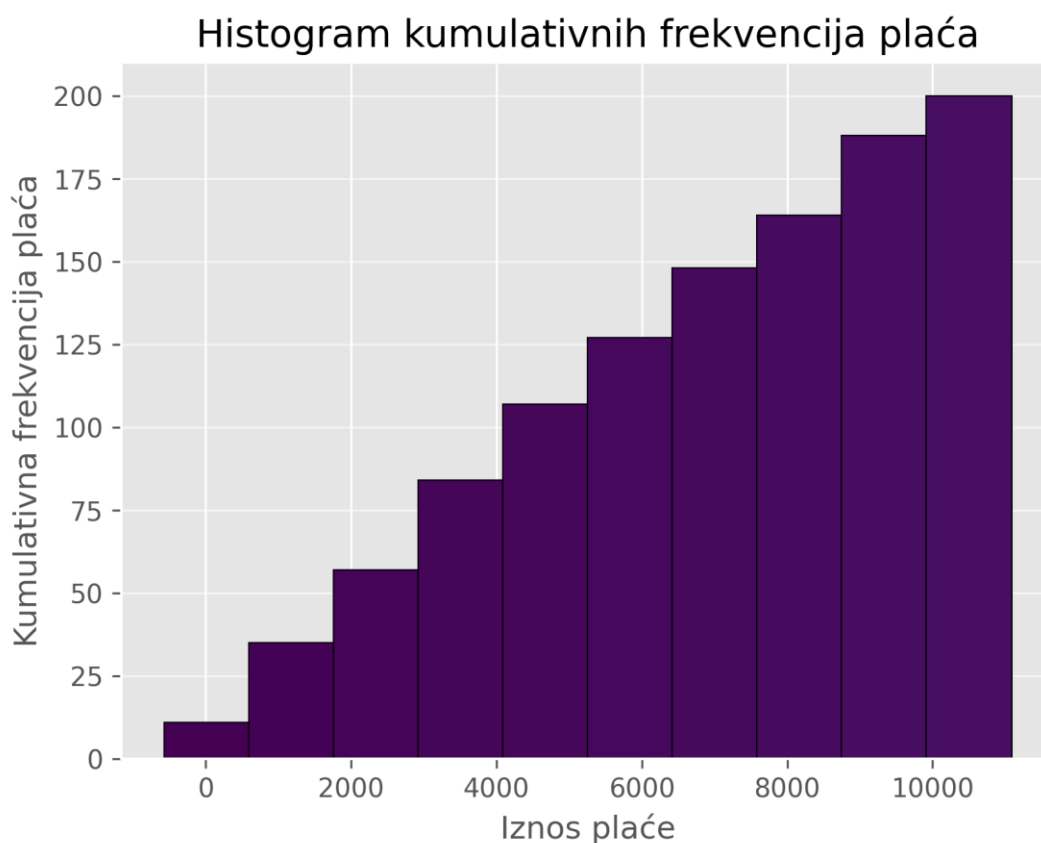
Kumulativna frekvencija distribucija je po kojoj se analizira i zbraja frekventnost ispod i unutar razreda (odnosno raspona razreda). Frekventnost se ne može smanjivati kako raspon vrijednosti raste.

Razred	Kumulativna frekvencija
15.6111	11
1180.92	35
2346.23	57
3511.54	84
4676.85	107
5842.15	127
7007.46	148
8172.77	164
9338.08	188
10503.4	200

Slika 5. Tablica kumulativnih frekvencija



kumulativne\_vrijednosti.txt



Slika 6. Graf kumulativnih frekvencija

Kraj 2. zadatka.



## Mjere centralne tendencije

Aritmetička sredina je jedna od središnjih vrijednosti koje se koriste u statistici i računa se za neki kvocijent sume članova skupa i broja članova tog skupa. Aritmetička sredina je najčešće korištena mjera centralne tendencije i računa se po formuli:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Aritmetička sredina za uzorak iz zadatka iznosi: **5145.635**

Geometrijska sredina je također jedna od mjere središnje tendencije, a pretežno se koristi kao mjera prosječne brzine nekih promjena. Računa se po formuli:

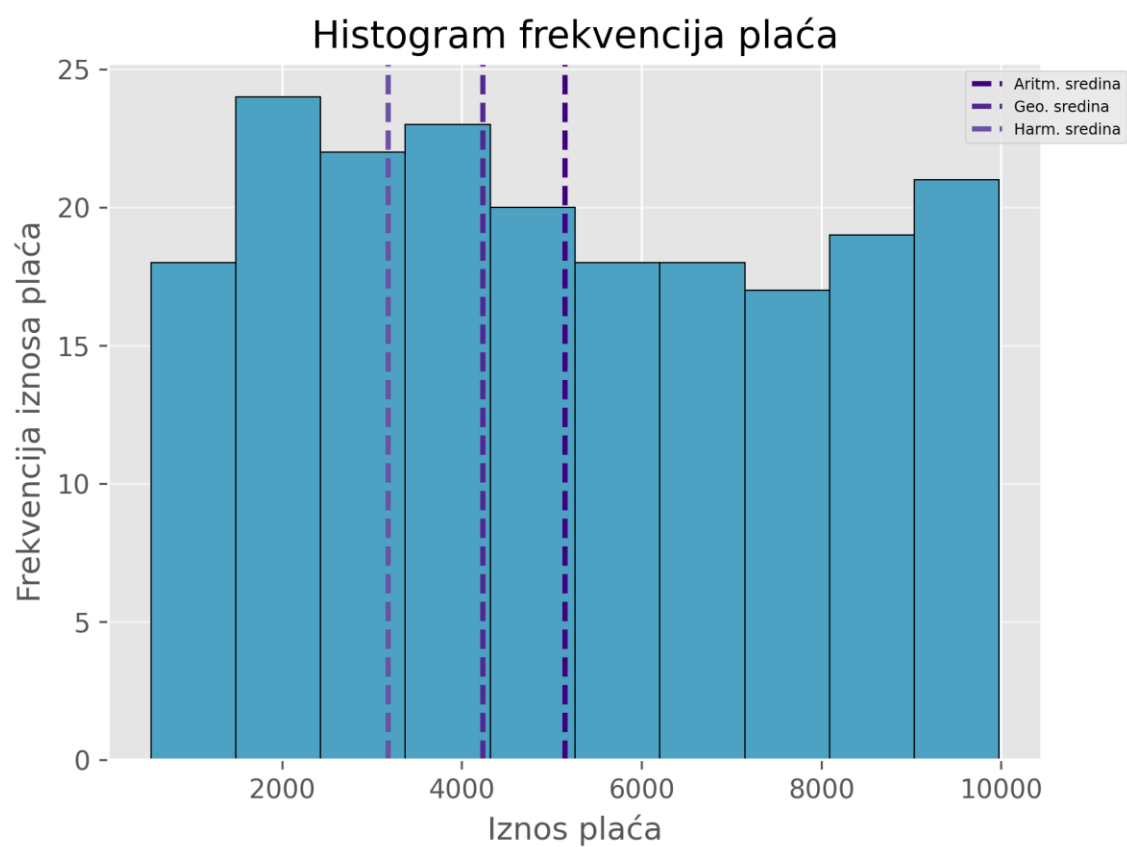
$$G = (a_1 \cdot a_2 \cdots a_n)^{\frac{1}{n}} \text{ ili } G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

Geometrijska sredina za uzorak iz zadatka iznosi: **4231.869**

Harmonijska sredina je srednja vrijednost koja je recipročna vrijednost aritmetičke sredine recipročnih vrijednosti zadanih vrijednosti. Računa se po formuli:

$$H = \frac{n}{\frac{1}{a_1} + \frac{1}{a_2} + \dots + \frac{1}{a_n}}$$

Harmonijska sredina za uzorak iz zadatka iznosi: **3181.014**



Slika 7. Histogram frekvencija plaća s mjerama centralne tendencije

Tu možemo i vizualizirati mjere centralne tendencije u našem uzorku.

Kraj 3. zadatka.

## Položajne mjere centralne tendencije

Modus ili skraćeno mod je još jedna mjera centralne tendencije i predstavlja najčešću vrijednost u nekom nizu tako zvanu sredinu distribucije.

Nadalje, ako je distribucija numeričke varijable grupirana u prave razrede, onda se razred s najvećom korigiranom frekvencijom  $b$  naziva se modalni razred. Ako je  $L_1$  donja granica tog razreda, a  $l$  njegova veličina te  $a$  frekvencija razreda koji prethodi modalnom, a  $c$  frekvencija razreda koji slijedi iza modalnog, onda se Mod aproksimira formulom

$$Mod = L_1 + \frac{b - a}{(b - a) + (b - c)} l.$$

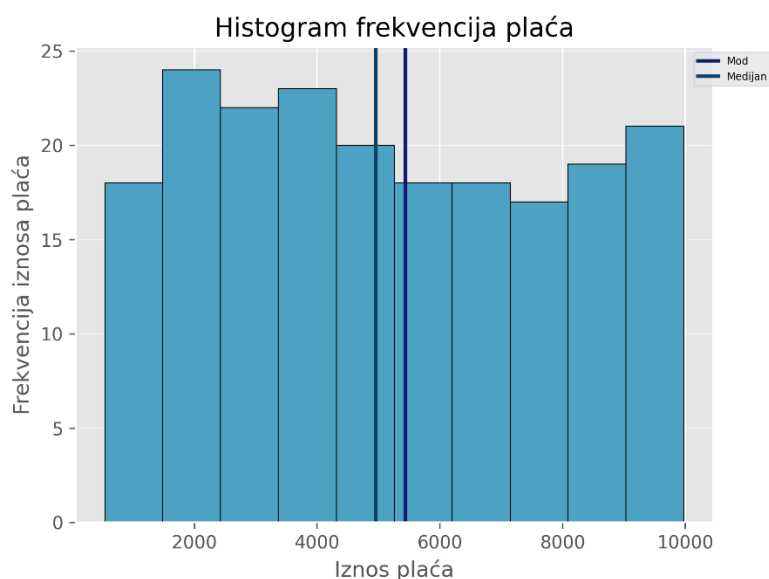
Mod za uzorak iz zadatka iznosi: **5440**

Medijan je vrijednost nekog skupa za koju vrijedi da je 50% podataka manje ili jednako toj vrijednosti i 50% podataka je veće ili jednako njoj. Ako je broj promatranih podataka paran onda medijan dobijemo s polu zbrojem središnjih podataka.

U slučaju da je  $N$  (članovi populacije) paran, Medijan se računa kao poluzbroj vrijednosti središnjih članova po formuli

$$R = x_{max} - x_{min}.$$

Medijan za uzorak iz zadatka iznosi: **4961.5**



Slika 8. Histogram frekvencije plaća s položajnim mjerama centralne tendencije

Kraj 4. zadatka.

## Raspon varijacije

Raspon podataka je mjera koja prikazuje koliko su podaci raspršeni. Računa se kao razlika najveće i najmanje vrijednosti u zadanom uzorku.

$$R = X_{max} - X_{min}.$$

Raspon za uzorak u zadatku iznosi: **9439**

Najveći element skupa je **9979**.

Najmanji element skupa je **540**.

*Kraj 5. zadatka.*

## Kvantili

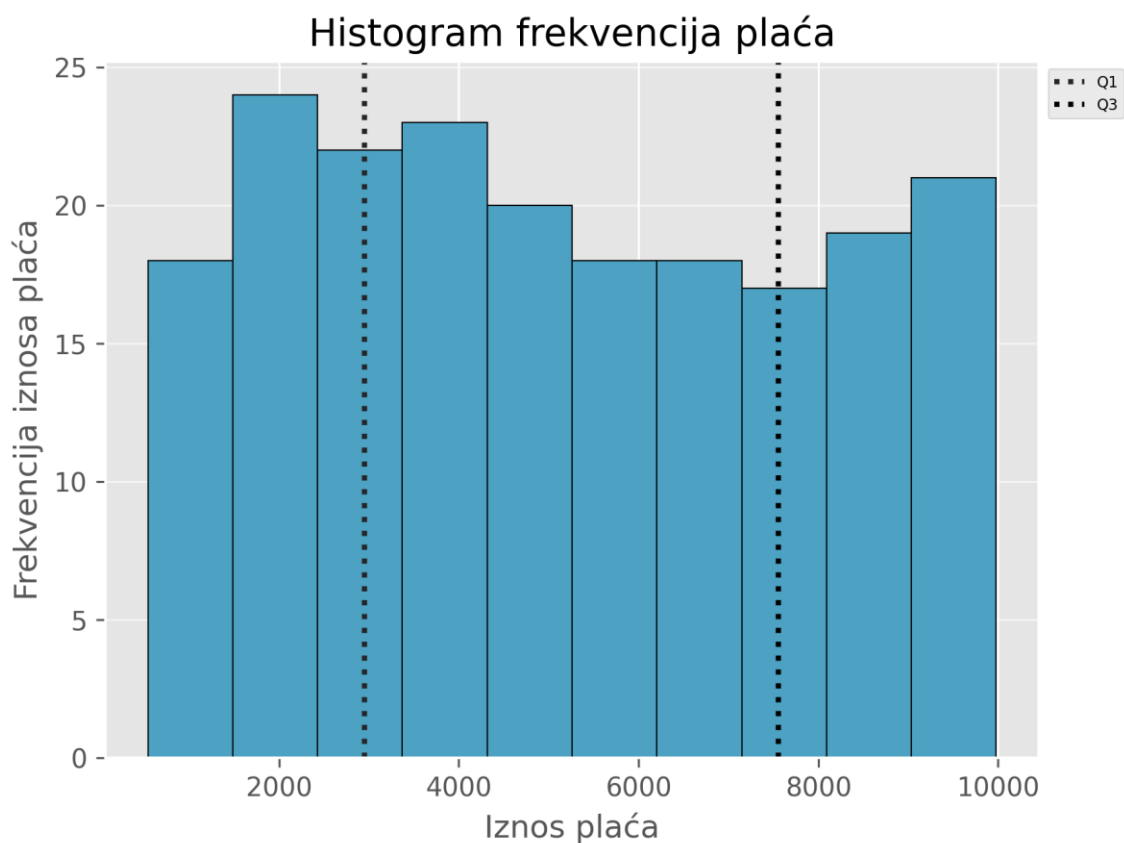
Kvantili su položajne vrijednosti obilježja koje uređene statističke nizove dijele na jednake dijelove. U praksi se najčešće koriste neki specifični kvantili: četvrtog reda (kvantili), desetog reda (decili) i stotog reda (centili/percentili). Kvantili su položajne vrijednosti koje uređene statističke nizove dijele na četiri jednaka dijela.

Prvi ili donji kvartil je broj od kojeg je 25% podataka manje ili jednako tom broju. Računa se tako da nađemo koji broj se nalazi na 25% cijelog promatranog uzorka.

Prvi kvartil za uzorak iz zadatka iznosi: **2947.25**

Treći ili gornji kvartil je broj od kojega je 75% manje ili jednako tom broju. Računa se tako da nađemo koji broj se nalazi na 75% cijelog promatranog uzorka.

Treći kvartil za uzorak iz zadatka iznosi: **7554.25**



Slika 9. Histogram frekvencija plaća s prvim i trećim kvartilom

Kraj 6. zadatka.

## Percentili

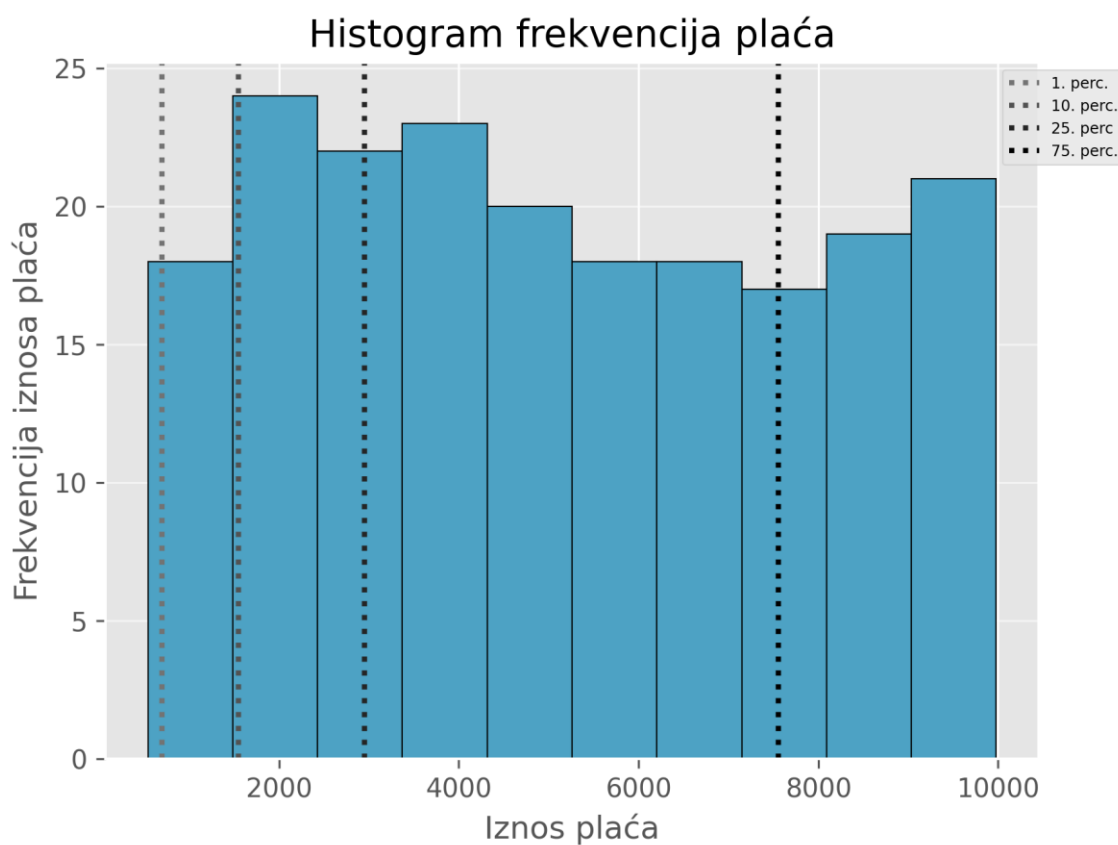
Percentil je statistička mjera koja spada u mjere lokacije. Percentil je broj koji predstavlja koliki je postotak brojeva manji ili jednak njemu.

1. percentil: **699.24**

10. percentil: **1548.4**

25. percentil je ujedno i 1. kvartil te iznosi: **2947.25**

75. percentil je ujedno i 3. kvartil te iznosi: **7554.25**



Slika 10. Histogram frekvencija plaća s percentilima

Kraj 7. zadatka.

## Suma apsolutnih vrijednosti odstupanja od srednje vrijednosti

Suma apsolutnih vrijednosti odstupanja služi da vidimo rasipanje podataka oko srednje vrijednosti. Računa se po formuli :

$$S\ AO = |x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|$$

Suma apsolutnih vrijednosti odstupanja za uzorak iz zadatka iznosi: **475335.350**

## Prosječno apsolutno odstupanje od aritmetičke sredine

Prosječno apsolutno odstupanje od aritmetičke sredine računamo tako da sumu apsolutnih vrijednosti odstupanja podijelimo s brojem uzoraka:

$$P\ AO = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{n}$$

Prosječno apsolutno odstupanje od aritmetičke sredine za uzorak iz zadatka iznosi: **2376.677**

## Varijanca

Varijanca uzorka definira se kao prosječno kvadratno odstupanje od prosjeka i računa se po formuli:

$$(s')^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n}$$

Varijanca za uzorak iz zadanog zadatka iznosi: **7519289.372**

*Kraj 8. zadatka.*

## Standardna devijacija

Standardna devijacija je mjera koja se koristi da kvantificira iznos varijacije ili disperzije vrijednosti podataka. Računa se po formuli:

$$\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2}$$

Standardna devijacija za uzorak iz zadanog zadatka iznosi: **2742.132**

## Korigirana varijanca

Korigirana varijanca jednaka je zbroju svih kvadrata razlike elemenata uzorka i aritmetičke sredine podijeljenih s duljinom uzorka umanjenom za jedan. Računa se po formuli:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Korigirana varijanca za uzorak u zadanom zadatku iznosi: **7557074.746**

## Korigirana standardna devijacija

Korigirana standardna devijacija uzorka koristi se za procjenu standardne devijacije populacije. Računa se po formuli:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

Korigirana standardna devijacija uzorka iz zadatka iznosi: **2749.013**

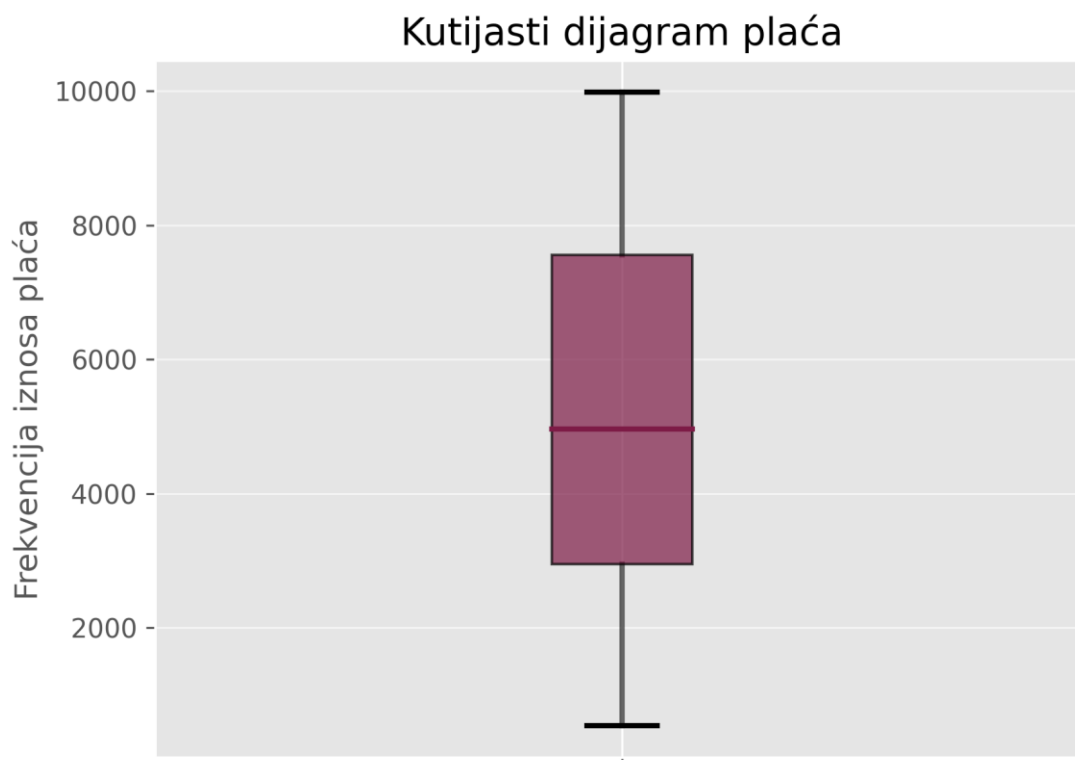
*Kraj 9. zadatka.*



## Kutijasti (Box-plot) dijagram

Box-plot ili kutijasti dijagram sastoji se od box-a odnosno pravokutnika koji predstavlja podatke od 1. odnosno donjeg kvartila do gornjeg odnosno 4. kvartila.

Sve točke koje se nalaze izvan područja koje predstavlja podatke od 1. do 4. kvartila nazivaju se „stršeće vrijednosti” koje predstavljaju vrijednosti sa malom frekvencijom pojavljivanja i velikim odklonom od srednje vrijednosti, medijana i moda.



*Slika 11. Kutijasti dijagram plaća*

*Kraj 10. zadatka.*

## Koeficijent asimetrije

Koeficijent asimetrije koristi sva odstupanja vrijednosti numeričke varijable od aritmetičke sredine i po tome je potpuna mjera asimetrije. Za mjerenje asimetrije polazna je veličina aritmetička sredina na treću potenciju odnosno treći moment oko sredine koja se računa po formuli:

$$\mu_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n} \text{ - za negrupirane vrijednosti}$$

$$\mu_3 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^3}{n} \text{ - za grupirane vrijednosti}$$

A koeficijent asimetrije  $\alpha_3$  računa se po formuli:

$$\alpha_3 = \frac{\mu_3}{\sigma^3}$$

Koeficijent asimetrije za uzorak iz zadatka iznosi: **0.113** i zbog toga je distribucija podataka pozitivno asimetrična.

## Pearsonova mjera asimetrije

Pearsonova mjera asimetrije je standardizirano odstupanje medijana ili moda od aritmetičke sredine. Pearsonova mjera je nepotpuna mjera asimetrije i manje je informativna od koeficijenata asimetrije, ali je jednostavnija i brže se izračuna. U pravilu vrijednosti su joj u intervalu  $[-3, 3]$ . Računa se po formuli:

$$S_{k_1} = \frac{\bar{x} - M_0}{\sigma} \text{ ili } S_{k_2} = \frac{3(\bar{x} - M_e)}{\sigma}$$

Pearsonova mjera asimetrije uzorka iz zadatka iznosi: **-0.108** i **0.201**

## Bowleyjeva mjera asimetrije

Bowleyjeva mjera asimetrije je mjera asimetrije koja se temelji na odnosima kvartila i medijana. U pravilu zauzima vrijednosti u intervalu  $[-1, 1]$ . Kao i Pearsonova mjera, Bowleyjeva mjera je također nepotpuna mjera asimetrije. Računa se po formuli:

$$s_{kQ} = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1}$$

Bowleyjeva mjera asimetrije za uzorak iz zadatka iznosi: **10499.346**

Distribucija uzorka iz zadatka je pozitivno asimetrična distribucija zato što je razlika 4. kvartila i medijana veća od razlike medijana i 1. kvartila.

*Kraj 11. zadatka.*

## Mjera zaobljenosti

Mjera zaobljenosti mjeri zaobljenost modalnog vrha distribucije. Izračunava se kao omjer četvrtog momenta oko sredine i standardne devijacije na četvrtu potenciju. Za različite rezultate koji se uspoređuju s brojem 3 postoje različite distribucije podataka. Računa se po formuli:

$$\alpha_4 = \frac{\mu_4}{\sigma_4^4}$$

$$\mu_4 = \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{n} \quad \text{- za negrupirane vrijednosti}$$

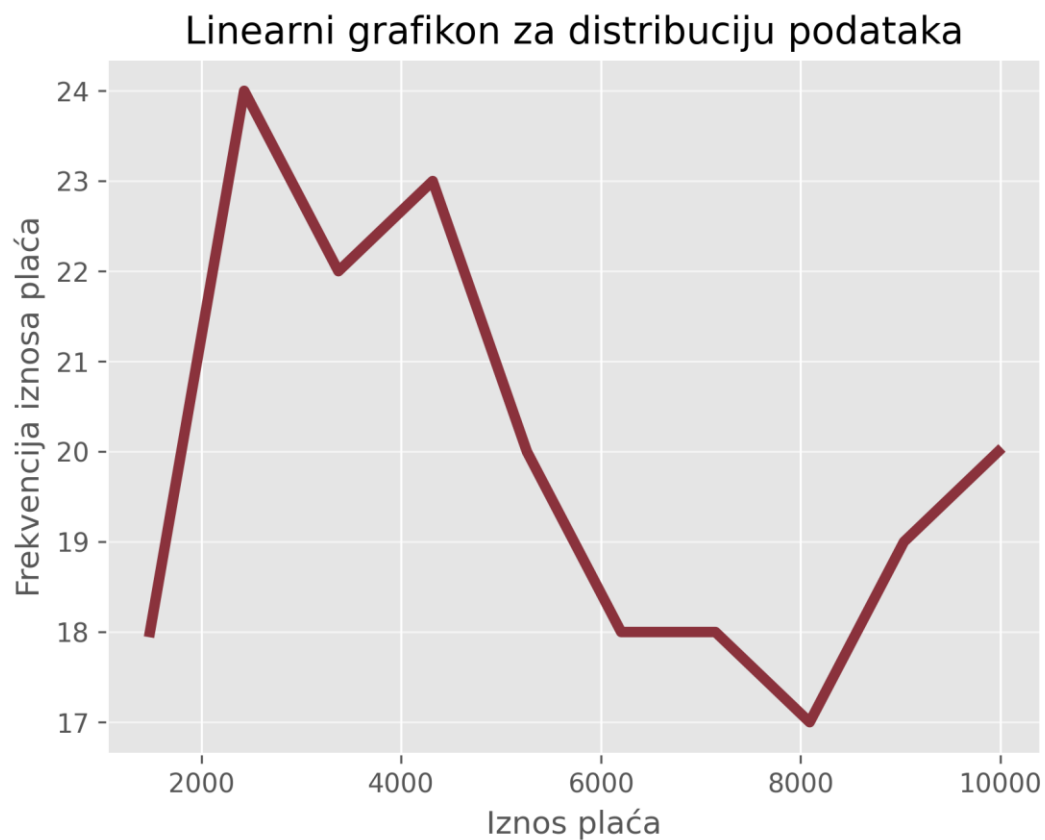
$$\mu_4 = \frac{\sum_{i=1}^n f_i (x_i - \bar{x})^4}{n} \quad \text{- za grupirane vrijednosti}$$

Mjera zaobljenosti za uzorak iz zadatka iznosi: **-1.198**

*Kraj 12. zadatka.*

## Linijski graf za distribuciju podataka

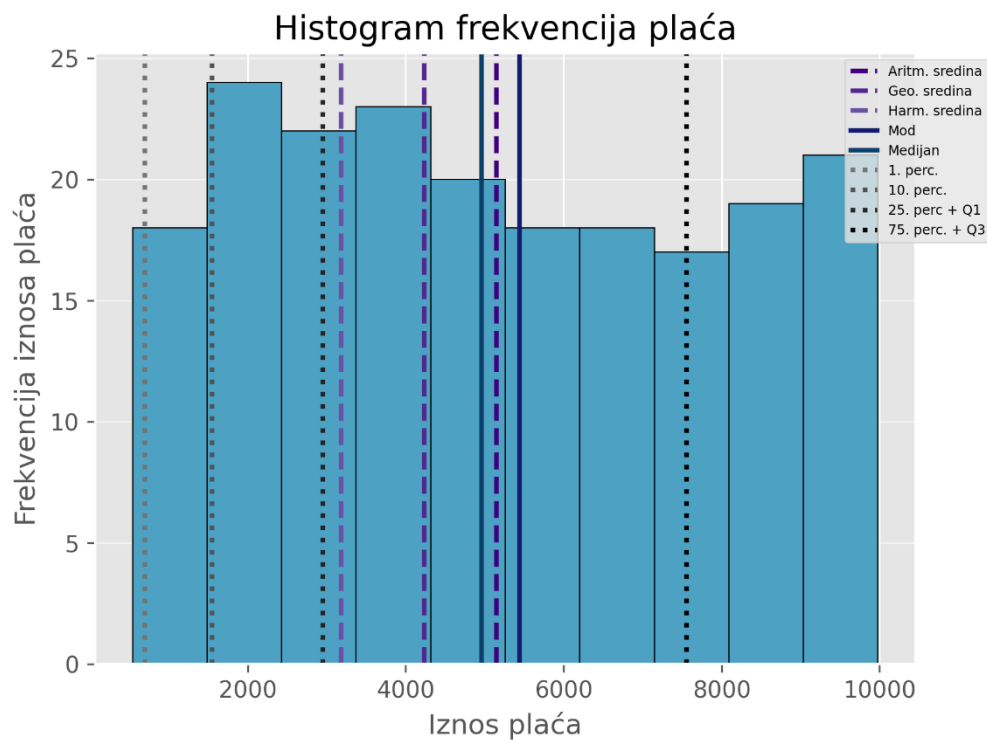
Na ovom grafu prikazana je distribucija plaća po njihovim frekvencijama i iznosima. Frekvencije iznosa plaća su predstavljeni pomoću Y osi, a iznosi placa X osi.



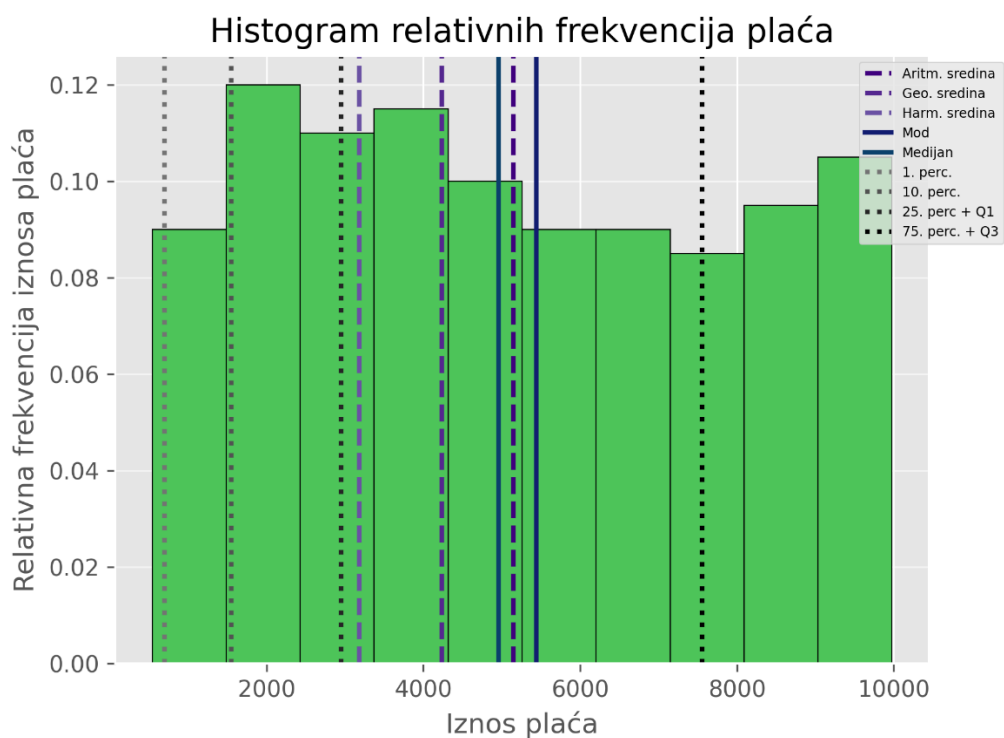
Slika 12. Linearni graf distribucije podataka

Kraj 13. zadatka.

## Histogrami frekvencija sa svim mjerama uzorka



Slika 13. Histogram frekvencija plaća sa svim mjerama uzorka



Slika 14. Histogram relativnih frekvencija plaća sa svim mjerama uzorka

## Intervali pouzdanosti

Interval pouzdanosti je interval u kojem se s predodređenoj pouzdanosti nalaze vrijednosti određenog parametra s odgovarajućom vjerojatnosti. To ne znači da se u, primjerice, intervalu pouzdanosti od 95% nalazi 95% podataka, nego da se u tom intervalu nalazi traženi parametar s 95% pouzdanosti. U našem slučaju, taj parametar je *očekivanje danog uzorka*.

Interval pouzdanosti koristimo da bi uzeli u obzir činjenicu da ne analiziramo cijelu populaciju uzorka te da imamo pogrešku uzorkovanja. Pouzdanost će rasti s povećanjem intervala, zato što obuhvaćanje više podataka znači i manji prostor za pogrešku, ali inverzno onda imamo i manju korist od našeg intervala.

Formula za interval pouzdanosti slijedi:

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

Gdje je:

CI = interval pouzdanosti

$\bar{x}$  = aritmetička sredina uzorka (očekivanje)

z = vrijednost pouzdanosti

s = standardna devijacija

n = broj podataka u uzorku

Interval pouzdanosti za očekivanje danog uzorka uz pouzdanost od 95%:

**4762.31 ≤ x ≤ 5528.95**, gdje x označava interval pouzdanosti.

Interval pouzdanosti za očekivanje danog uzorka uz pouzdanost od 85%:

**4864.73 ≤ x ≤ 5426.54**, gdje x označava interval pouzdanosti.

*Kraj 1. i 2. zadatka intervala pouzdanosti.*

# Zaključak

Nimalo začuđujuće s obzirom na to da su podaci generirani specifično u okviru plaća na web stranici s koje su preuzeti, naši podaci općenito indiciraju standardno variranje i ostale parametre analize našeg uzorka. Unatoč tome, plaće se svejedno mogu razlikovati skoro deseterostruko unutar samo 200 podataka, i relativno su jednako raspoređene po razredima.



## Cijeli ispis u konzoli

Grupa (razred)	Iznos plaće u razredu	Frekvencija	Relativna frekvencija
1	1483.9	18	0.09
2	2427.8	24	0.12
3	3371.7	22	0.11
4	4315.6	23	0.115
5	5259.5	20	0.1
6	6203.4	18	0.09
7	7147.3	18	0.09
8	8091.2	17	0.085
9	9035.1	19	0.095
10	9979	20	0.1

Aritmetička sredina uzorka: 5145.635  
 Geometrijska sredina uzorka: 4231.868919870252  
 Harmonijska sredina uzorka: 3181.0140941660175

Mod uzorka: 5440  
 Medijan uzorka: 4961.5

Najveci element u uzorku: 9979  
 Najmanji element u uzorku: 540  
 Raspon uzorka: 9439

Prvi kvartil: 2947.25  
 Treci kvartil: 7554.25

1. percentil: 699.24  
 10. percentil: 1548.4  
 25. percentil: 2947.25  
 75. percentil: 7554.25

Suma apsolutnih odstupanja: 475335.35000000003  
 Prosječno apsolutno odstupanje: 2376.67675

Varijanca uzorka: 7519289.371775  
 Standardna devijacija uzorka: 2742.132267374242  
 Korigirana varijanca uzorka: 7557074.745502513  
 Korigirana standardna devijacija uzorka: 2749.0134131179702

Koeficijent asimetrije uzorka: 0.11251893445469131  
 Pearsonova mjera asimetrije S.k1 uzorka: -0.10734894282903142  
 Pearsonova mjera asimetrije S.k2 uzorka: 0.20145089519294485  
 Bowleyeva mjera asimetrije uzorka: 10499.346103755155

Mjera zaobljenosti: -1.1983644649780543

Procijenjena očekivana vrijednost numeričkog obilježja populacije: 5145.635  
 Interval pouzdanosti 95%: (4762.317010803881, 5528.952989196119)  
 Interval pouzdanosti 85%: (4864.727986411072, 5426.542013588928)

## Kôd



main--linic\_krulcic\_manjaric\_matejcic\_poje.py

# Tablica opisa slika

Slika 1. XML dokument "place.xml" s našim uzorkom od 200 podataka. ....	4
Slika 2. Tablica frekvencija.....	5
Slika 3. Histogram frekvencija plaća.....	6
Slika 4. Histogram relativnih frekvencija plaća .....	7
Slika 5. Tablica kumulativnih frekvencija .....	8
Slika 6. Graf kumulativnih frekvencija.....	8
Slika 7. Histogram frekvencija plaća s mjerama centralne tendencije .....	10
Slika 9. Histogram frekvencije plaća s položajnim mjerama centralne tendencije ....	11
Slika 10. Histogram frekvencija plaća s prvim i trećim kvartilom .....	13
Slika 11. Histogram frekvencija plaća s percentilima .....	14
Slika 12. Kutijasti dijagram plaća.....	17
Slika 13. Linearni graf distribucije podataka .....	21
Slika 14. Histogram frekvencija plaća sa svim mjerama uzorka .....	22
Slika 15. Histogram relativnih frekvencija plaća sa svim mjerama uzorka.....	22