SVEUČILIŠTE U SPLITU
SVEUČILIŠNI ODJEL ZA STRUČNE STUDIJE

Diplomski stručni studij Primijenjeno računarstvo

Predmet: Statistika

S E M I N A R S K I R A D

Student: Anamarija Papić

Naslov rada: Vrijeme provedeno na Zavodu za zapošljavanje čekajući posao

Mentor: Nada Roguljić, viši predavač

Split, siječanj 2024.

Sadržaj

1. Uvod

Svrha izrade ovog seminarskog rada/projektnog zadatka jest da student koristeći programsko okruženje po izboru ovlada osnovama deskriptivne statistike i bar jednom metodom inferencijalne statistike.

Za ovaj seminarski rad odabran je skup podataka sadržan u datoteci `p10.xlsx` koji predstavlja vrijeme provedeno na Zavodu za zapošljavanje čekajući posao (izraženo u danima, uzorak 300 nezaposlenih). Dobiveni skup podataka može se vidjeti među dodatcima u tablici 4.

Analiza je provedena u programskom jeziku **Python** koristeći se bibliotekama / paketima `pandas`, `numpy`, `matplotlib`, `scipy`.

U poglavljima koja slijede opisat će se statistička analiza podataka provedena na odabranom uzorku podataka, uz kratak pregled teorijskih osnova provedenih postupaka i interpretaciju dobivenih vrijednosti te će se priložiti kôd te tablice i grafovi. Analiza obuhvaća deskriptivnu statistiku, razdiobu frekvencija, grafičke prikaze, izračun intervala povjerenja te testiranje hipoteza.

2. Deskriptivna statistika

Deskriptivna (opisna) statistika je grana statistike koja se bavi obrađivanjem dobivenih podataka, te u kojoj se isti opisuju te predočuju tablicama i grafikonima.

Iz **populacije** je izdvojen **uzorak (podskup)** veličine 300 nezaposlenih na kojem je promatrano samo jedno svojstvo - statističko obilježje $X = "vrijeme\ (u\ danima)"$ te je statistička varijabla X numerička (kvantitativna).

U ispisu 1 prikazano je čitanje podataka iz Excel datoteke i konverzija istih u numeričke vrijednosti.

```
# Citanje podataka iz Excel datoteke
data = pd.read_excel('p10.xlsx', usecols=[0], names=['Vrijeme'])

# Konverzija stupca 'Vrijeme' u numericki format
data['Vrijeme'] = pd.to_numeric(data['Vrijeme'], errors='coerce')
```

Ispis 1: Pristup podatcima iz `p10.xlsx` Excel datoteke

U ovom koraku provedene analize, izračunate su osnovne mjere središnje tendencije: aritmetička sredina (srednja vrijednost), mod, medijan te mjere raspršenosti (disperzije): varijanca, standardna devijacija, interkvartilni raspon, raspon uzorka.

Aritmetička sredina (srednja vrijednost) jest suma svih vrijednosti u skupu podataka podijeljena s brojem tih vrijednosti. Predstavlja "prosječnu" vrijednost u skupu.

Mod je vrijednost koja se najčešće pojavljuje u skupu podataka.

Medijan je središnja vrijednost u uređenom skupu podataka. Polovina podataka je ispod, a polovina iznad medijana. Otporna je na ekstremne vrijednosti.

Varijanca je mjera koja prikazuje koliko su pojedinačne vrijednosti u skupu podataka raspršene u odnosu na aritmetičku sredinu. Izračunava se kao prosječna kvadratna udaljenost svake vrijednosti od aritmetičke sredine.

Standardna devijacija je kvadratni korijen varijance. Praktičnija je za interpretaciju jer je izražena u istim jedinicama kao i podatci.

Interkvartilni raspon je raspon vrijednosti između prvog kvartila (25% podataka) i trećeg kvartila (75% podataka) u uređenom skupu podataka.

Raspon uzorka označava razliku između najveće i najmanje vrijednosti u skupu podataka.

Ove mjere pružaju uvid u različite aspekte distribucije podataka i omogućuju bolje razumijevanje njihove središnje tendencije i raspršenosti. Aritmetička sredina, mod i medijan daju informacije o središnjoj tendenciji, dok varijanca, standardna devijacija, interkvartilni raspon i raspon uzorka pružaju informacije o raspršenosti podataka.

Za izračun i ispis prethodno spomenutih vrijednosti napisan je kôd prikazan u ispisu 2.

```
# 1. Deskriptivna statistika
mean = np.mean(data['Vrijeme'])
mode = stats.mode(data['Vrijeme'])
median = np.median(data['Vrijeme'])
five_number_summary = np.percentile(data['Vrijeme'], [0, 25, 50, 75,
    100])
variance = np.var(data['Vrijeme'])
std_deviation = np.std(data['Vrijeme'])
interquartile_range = np.percentile(data['Vrijeme'], 75) - np.percentile(
    data['Vrijeme'], 25)
data_range = np.max(data['Vrijeme']) - np.min(data['Vrijeme'])

print(f"Srednja vrijednost: {mean}")
print(f"Mod: {mode} sto znaci da je vrijednost {mode[0]} najfrekventnija
    vrijednost i pojavljuje se {mode[1]} put.",)
print(f"Medijan: {median}")
print(f"Karakteristicna petorka uzorka: {five_number_summary}")
print(f"Varijanca: {variance}")
print(f"Standardna devijacija: {std_deviation}")
print(f"Interkvartil: {interquartile_range}")
print(f"Raspon uzorka: {data_range}")
```

Ispis 2: Izračun i ispis traženih osnovnih pojmova statističke analize podataka

Dobivene su sljedeće vrijednosti:

$$\begin{aligned}
\text{Srednja vrijednost} &: 14.4 \\
\text{Mod} &: 14 \text{ (pojavljuje se 21 put)} \\
\text{Medijan} &: 14.0 \\
\text{Karakteristična petorka uzorka} &: [0.0, 8.0, 14.0, 20.0, 30.0] \\
\text{Varijanca} &: 60.27333333333325 \\
\text{Standardna devijacija} &: 7.7635902347646635 \\
\text{Interkvartil} &: 12.0 \\
\text{Raspon uzorka} &: 30.0
\end{aligned}$$

Dakle, dobiveni podatci imaju srednju vrijednost od 14.4, što znači da je prosječna vrijednost skupa podataka oko 14.4. Najčešća vrijednost (mod) u skupu podataka je 14, pojavljujući se čak 21 put, što ukazuje na izraženu koncentraciju podataka oko te vrijednosti.

Srednja vrijednost i medijan su bliski, pri čemu je medijan jednak 14.0. To sugerira da je polovina podataka manja od 14.0, dok je druga polovina veća od te vrijednosti.

Karakteristična petorka uzorka ([0.0, 8.0, 14.0, 20.0, 30.0]) pruža ključne vrijednosti koje karakteriziraju raspodjelu podataka, uključujući minimum (0.0), prvi kvartil (Q1 - 8.0), medijan (14.0), treći kvartil (Q3 - 20.0) i maksimum (30.0).

Raspršenost podataka je izražena varijancom od 60.27 i standardnom devijacijom od 7.76. Ove vrijednosti ukazuju na to da su podaci razmjerno raspršeni u odnosu na srednju vrijednost.

Interkvartilni raspon iznosi 12.0, što znači da se srednja polovina podataka nalazi unutar tog raspona. Raspon uzorka, koji označava razliku između maksimalne i minimalne vrijednosti, iznosi 30.0.

3. Razdioba frekvencija

Nakon što je prvi korak statističke analize proveden, napravljena je razdioba frekvencija podataka. Podatci su grupirani u razrede koristeći intervale širine 5 dana, te su izračunate frekvencije, relativne frekvencije i kumulativne relativne frekvencije (prikazane tablično u tablicama 1 - 3).

Frekvencija predstavlja broj puta koliko se određena vrijednost pojavljuje ili se nalazi unutar određenog raspona u skupu podataka.

Relativna frekvencija je omjer frekvencije određene vrijednosti i ukupnog broja podataka u skupu. Izražava se kao postotak ili decimalni broj.

Relativna kumulativna frekvencija predstavlja kumulativni postotak ili udio podataka koji su manji ili jednaki određenoj vrijednosti ili intervalu. To se postiže zbrajanjem relativnih frekvencija određenih vrijednosti i svih prethodnih vrijednosti. Relativna kumulativna frekvencija za posljednju vrijednost uvijek će biti 1 (ili 100% ako se izražava postotkom). Ova mjera pruža uvid u distribuciju podataka na kumulativnoj skali.

Frekvencijska tablica je organizirani prikaz podataka koji pokazuje koliko često se pojedine vrijednosti ili rasponi vrijednosti pojavljuju u skupu podataka. Svaki red tablice predstavlja određeni interval ili kategoriju, a odgovarajuća vrijednost u tom retku predstavlja broj pojavljivanja ili frekvenciju tog intervala. U kontekstu statističke analize, frekvencijske tablice često se koriste za prikazivanje distribucije podataka.

Relativna frekvencijska tablica je varijacija frekvencijske tablice u kojoj se frekvencije izražavaju kao udio ili postotak u odnosu na ukupan broj podataka. Ovo omogućuje usporedbu distribucije podataka između različitih skupova koji mogu imati različite veličine. Relativne frekvencije se dobivaju dijeljenjem frekvencije pojedinog intervala s ukupnim brojem podataka.

Relativna kumulativna frekvencijska tablica proširuje koncept relativne frekvencijske tablice dodajući kumulativnu vrijednost. Kumulativna relativna frekvencija za svaki interval predstavlja zbroj relativnih frekvencija tog intervala i svih prethodnih intervala. Ova tablica pomaže u vizualizaciji kumulativne distribucije podataka, što znači koliko postotaka podataka leži ispod ili unutar određenog intervala.

Tablica 1: Frekvencijska tablica

Razred	Frekvencija
$(-0.001, 5.0]$	44
$(5.0, 10.0]$	54
$(10.0, 15.0]$	70
$(15.0, 20.0]$	64
$(20.0, 25.0]$	39
$(25.0, 30.0]$	29

Tablica 2: Tablica relativnih frekvencija

Razred	Relativna frekvencija
$(-0.001, 5.0]$	0.146667
$(5.0, 10.0]$	0.180000
$(10.0, 15.0]$	0.233333
$(15.0, 20.0]$	0.213333
$(20.0, 25.0]$	0.130000
$(25.0, 30.0]$	0.096667

Tablica 3: Tablica kumulativnih relativnih frekvencija

Razred	Kumulativna relativna frekvencija
$(-0.001, 5.0]$	0.146667
$(5.0, 10.0]$	0.326667
$(10.0, 15.0]$	0.560000
$(15.0, 20.0]$	0.773333
$(20.0, 25.0]$	0.903333
$(25.0, 30.0]$	1.000000

Ove tablice prikazuju frekvencije i relativne frekvencije za svaki interval, kao i kumulativne relativne frekvencije koje se akumuliraju kako se krećemo kroz intervale.

Ovi rezultati pružaju uvid u distribuciju podataka prema vremenskim intervalima u trajanju 5 dana.

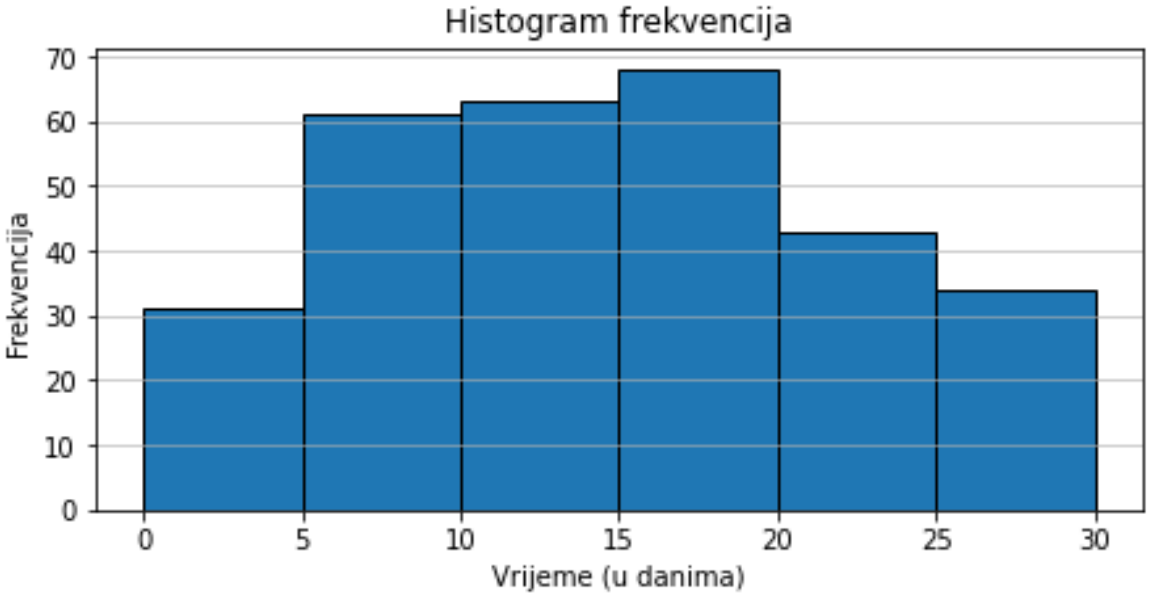
Prethodno prezentirane tablice rezultat su kôda prikazanog u ispisu 3.

```
# 2. Razdioba frekvencija
bins = range(0, 35, 5)
frequency_table = pd.Series(pd.cut(data['Vrijeme'], bins=bins,
    include_lowest=True)).value_counts().sort_index()
relative_frequency = frequency_table / len(data)
cumulative_relative_frequency = relative_frequency.cumsum()

print("\nFrekvencijska tablica:")
print(frequency_table)

print("\nRelativna frekvencija:")
print(relative_frequency)

print("\nKumulativna relativna frekvencija:")
print(cumulative_relative_frequency)
```

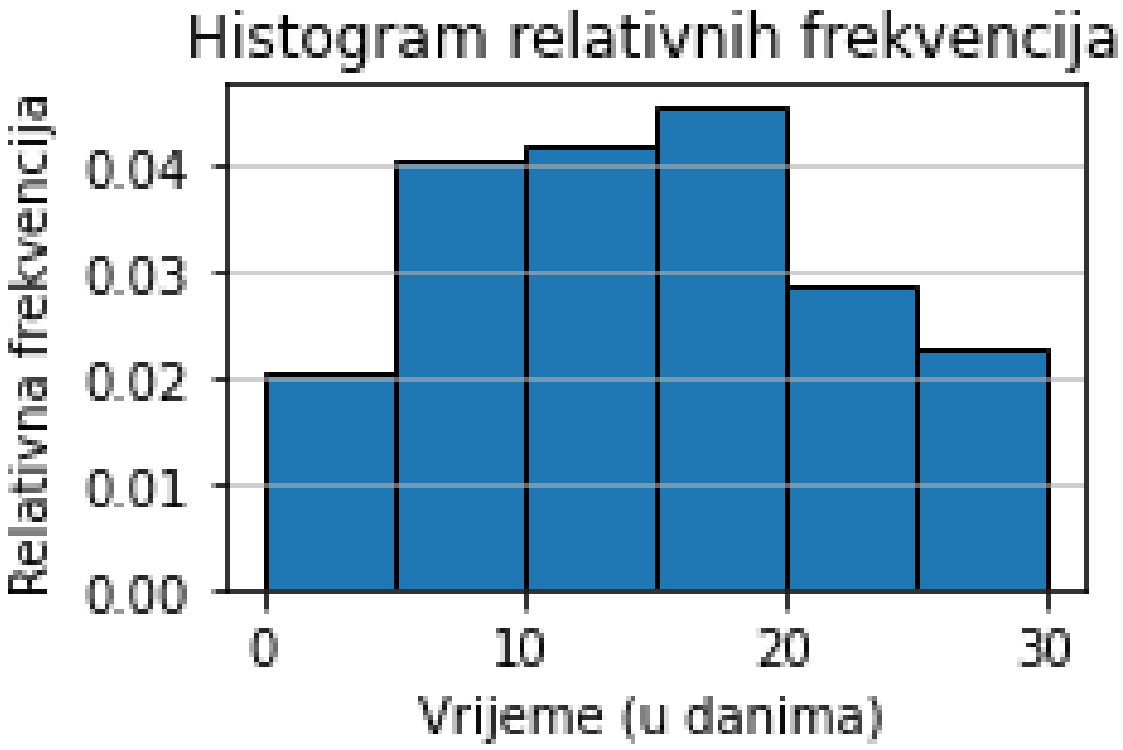
Ispis 3: Razdioba frekvencija

4. Grafički prikazi

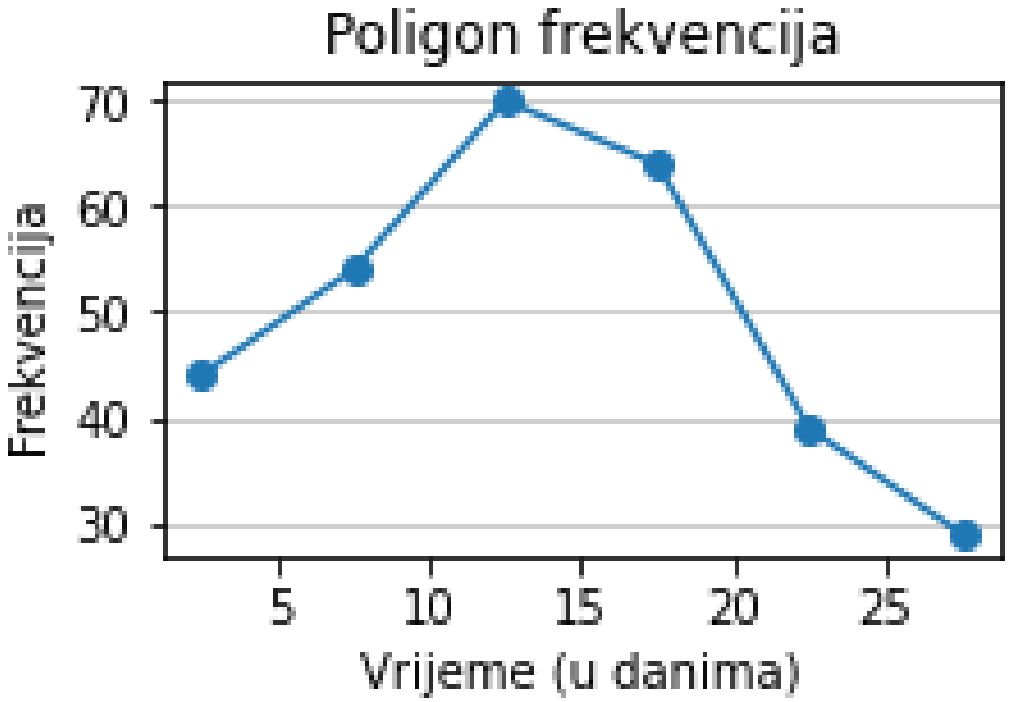
U nastavku su prikazane različite grafičke reprezentacije podataka kako bi se dobila bolja vizualna interpretacija distribucije.



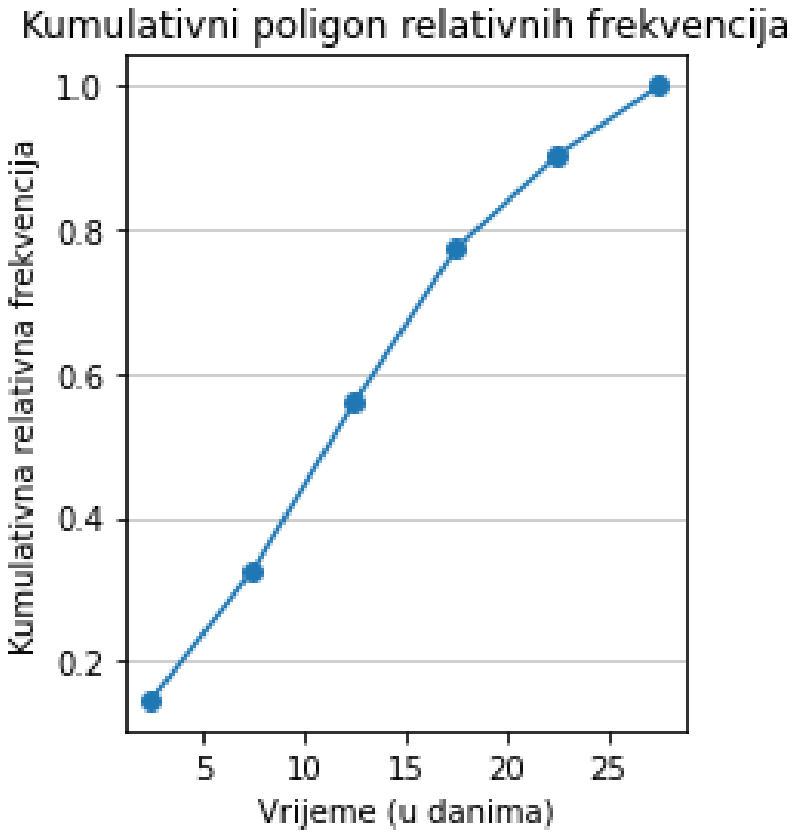
Slika 1: Histogram frekvencija



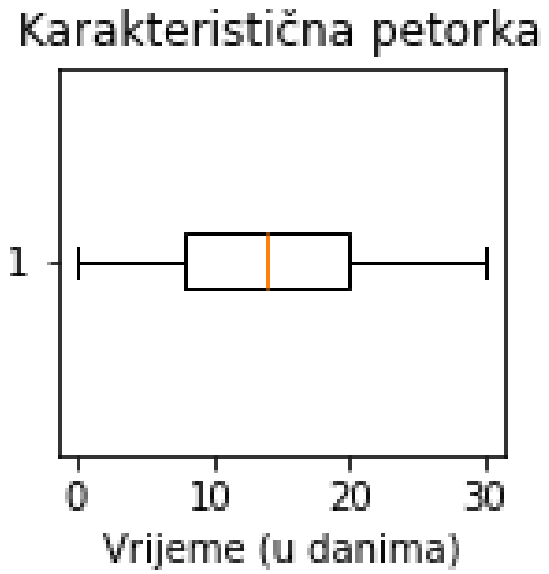
Slika 2: Histogram relativnih frekvencija



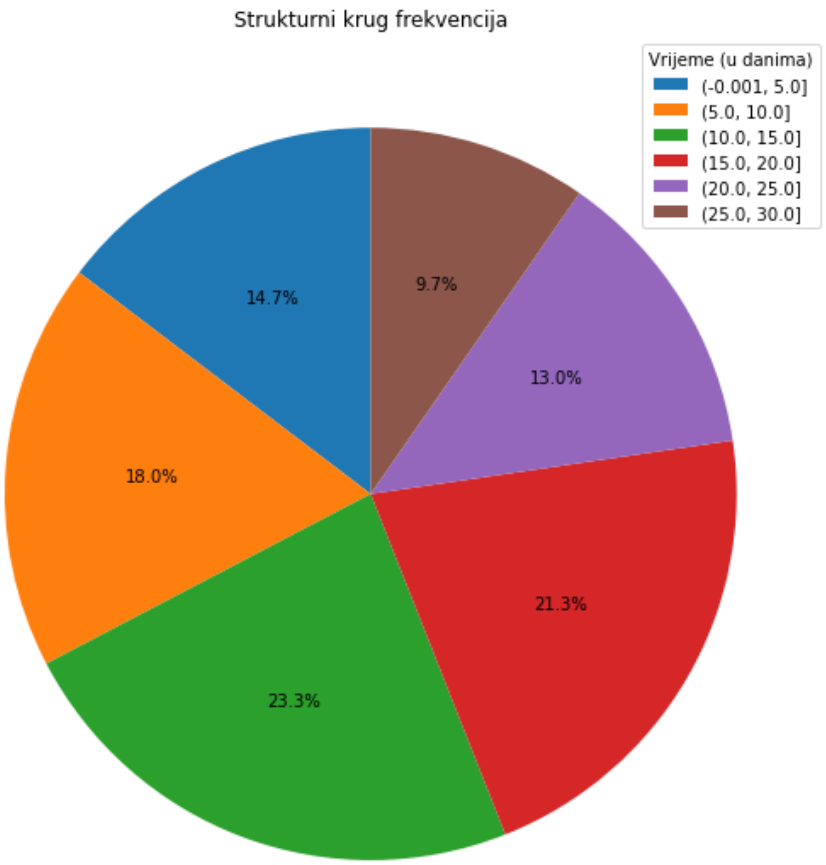
Slika 3: Poligon frekvencija



Slika 4: Kumulativni poligon relativnih frekvencija



Slika 5: Karakteristična petorka, prikazana kutijastim dijagramom tzv. *box plot*



Slika 6: Strukturni krug frekvencija, prikazan kružnim dijagramom tzv. *pie chart*

U ispisu 4 prikazan je način na koji je bilo potrebno konfigurirati i postaviti *plot* iz biblioteke `matplotlib` kako bi se što jasnije prethodno izračunate vrijednosti grafički prikazale te kôd koji je zaslužan za crtanje svakog od prikazanih grafikona.

```
# Racunanje sredisnjih tocki za iscrtavanje na grafu
midpoints = [interval.mid for interval in frequency_table.index]

plt.subplot(2, 2, 1)
plt.hist(data['Vrijeme'], bins=bins, edgecolor='black')
plt.title('Histogram frekvencija')
plt.xlabel('Vrijeme (u danima)')
plt.ylabel('Frekvencija')
plt.grid(axis='y', alpha=0.75)
plt.tight_layout()
plt.show()

plt.subplot(2, 2, 2)
plt.hist(data['Vrijeme'], bins=bins, density=True, edgecolor='black')
plt.title('Histogram relativnih frekvencija')
plt.xlabel('Vrijeme (u danima)')
plt.ylabel('Relativna frekvencija')
plt.grid(axis='y', alpha=0.75)
plt.tight_layout()
plt.show()

plt.subplot(2, 2, 3)
plt.plot(midpoints, frequency_table.values, marker='o')
plt.title('Poligon frekvencija')
plt.xlabel('Vrijeme (u danima)')
plt.ylabel('Frekvencija')
plt.grid(axis='y', alpha=0.75)
plt.tight_layout()
plt.show()

plt.subplot(1, 2, 1)
plt.plot(midpoints, cumulative_relative_frequency.values, marker='o')
plt.title('Kumulativni poligon relativnih frekvencija')
plt.xlabel('Vrijeme (u danima)')
```

```
plt.ylabel('Kumulativna relativna frekvencija')
plt.grid(axis='y', alpha=0.75)
plt.tight_layout()
plt.show()

plt.subplot(2, 3, 6)
plt.boxplot(data['Vrijeme'], vert=False)
plt.title('Karakteristicna petorka')
plt.xlabel('Vrijeme (u danima)')
plt.tight_layout()
plt.show()

plt.figure(figsize=(8, 8))
wedges, texts, autotexts = plt.pie(frequency_table, autopct='%1.1f%%',
    startangle=90)
plt.title('Strukturni krug frekvencija')
plt.legend(wedges, frequency_table.index, title='Vrijeme (u danima)')
plt.tight_layout()
plt.show()
```

Ispis 4: Način konstruiranja grafičkih prikaza koristeći biblioteku `matplotlib`

5. Interval povjerenja

Interval povjerenja je statistički pojam koji se koristi za procjenu nesigurnosti u vezi s procjenama parametara populacije na temelju uzorka podataka. To je raspon vrijednosti koji sadrži procijenjenu vrijednost parametra s određenom vjerojatnošću. Interval povjerenja ovisi o razinama pouzdanja, a uobičajene razine pouzdanja su 90%, 95% ili 99%, ovisno o željenoj razini sigurnosti. Na način prikazan u ispisu 5 izračunati su sljedeći intervali povjerenja za očekivanje populacije s pouzdanošću od 90%, 95% i 99%:

```
# 4. Interval povjerenja
confidence_interval = stats.norm.interval(0.90, loc=mean, scale=
    std_deviation/np.sqrt(len(data)))
print(f"\nInterval povjerenja (90%): {confidence_interval}")

confidence_interval = stats.norm.interval(0.95, loc=mean, scale=
    std_deviation/np.sqrt(len(data)))
print(f"Interval povjerenja (95%): {confidence_interval}")

confidence_interval = stats.norm.interval(0.99, loc=mean, scale=
    std_deviation/np.sqrt(len(data)))
print(f"Interval povjerenja (99%): {confidence_interval}")
```

Ispis 5: Izračun intervala povjerenja (90%, 95% i 99%)

$$\text{Interval povjerenja (90\%)} : (13.662725463940534, 15.137274536059467)$$

$$\text{Interval povjerenja (95\%)} : (13.521483204512723, 15.278516795487278)$$

$$\text{Interval povjerenja (99\%)} : (13.24543322054612, 15.55456677945388)$$

Ovi intervali pružaju informaciju o tome gdje se s velikom vjerojatnošću nalazi stvarna vrijednost očekivanja populacije. Općenito, razina pouzdanja odražava koliko smo sigurni u točnost intervala povjerenja. Što je razina pouzdanja viša, to je interval širi, ali pruža veću sigurnost da sadrži stvarnu vrijednost parametra. Obratno, niža razina pouzdanja rezultira užim intervalom, ali manjom sigurnošću.

6. Testiranje hipoteza

Postavljena je hipoteza o očekivanju populacije, te je proveden t-test za testiranje hipoteze.

Kôd je prikazan u ispisu 6.

```
# 5. Testiranje hipoteze
# Npr. testiranje hipoteze o prosjecnom vremenu provedenom na Zavodu za
    zaposljavanje cekajuci posao
# Koristi se t-test
alpha = 0.05
t_statistic, p_value = stats.ttest_1samp(data['Vrijeme'], popmean=15)

print("\nTestiranje hipoteza:")
print(f"T-statistika: {t_statistic}")
print(f"P-vrijednost: {p_value}")

if p_value < alpha:
    print("Odbacujemo nultu hipotezu.")
else:
    print("Ne mozemo odbaciti nultu hipotezu.")
```

Ispis 6: Testiranje hipoteza

Dakle, proveli smo testiranje hipoteza vezano uz srednju vrijednost populacije (μ). Hipoteze su sljedeće:

$$H_0 : \mu = 15$$

$$H_1 : \mu \neq 15$$

Rezultati testa su sljedeći:

$$\text{T-statistika} : -1.3363623743324509$$

$$\text{P-vrijednost} : 0.18244692949663507$$

1. **T-statistika:**

   T-statistika predstavlja mjeru koliko se srednja vrijednost uzorka razlikuje od srednje vrijednosti populacije iz nulte hipoteze, izraženo u standardnim devijacijama. Negativna vrijednost ukazuje na to da je srednja vrijednost uzorka manja od pretpostavljene srednje vrijednosti populacije.

2. **P-vrijednost:**

   P-vrijednost je vjerojatnost da bismo dobili rezultate testa (ili ekstremnije) ako je nulta hipoteza istinita. Ako je p-vrijednost mala (obično manja od odabrane razine značajnosti, npr. 0.05), obično odbacujemo nultu hipotezu. Ako je p-vrijednost visoka, zadržavamo nultu hipotezu.

Budući da je p-vrijednost (0.182) veća od odabrane razine značajnosti $\alpha = 0.05$, ne možemo odbaciti nultu hipotezu. To sugerira da nema dovoljno statističkih dokaza da se tvrdi da je očekivanje populacije različito od 15 dana.

7. Zaključak

U ovom seminarskom radu temeljito je proučen i analiziran skup podataka koji predstavlja vrijeme provedeno na Zavodu za zapošljavanje čekajući posao za uzorak od 300 nezaposlenih osoba. Cilj je bio ovladati osnovama deskriptivne statistike i primijeniti barem jednu metodu inferencijalne statistike korištenjem programskog jezika Python.

Prvo je izvršeno učitavanje podataka iz datoteke `p10.xlsx` i predstavljen uzorak u tabličnom obliku. Nakon toga, provedena je opsežna statistička analiza koja uključuje deskriptivnu statistiku, razdiobu frekvencija, grafičke prikaze, izračun intervala povjerenja te testiranje hipoteza.

Kroz provedene postupke, stečen je uvid u različite aspekte distribucije podataka, mjere središnje tendencije i raspršenosti, oblik raspodjele, te su doneseni zaključci o vremenu koje nezaposleni provode čekajući posao.

Daljnje istraživanje i proširenje analize moglo bi uključivati dublje istraživanje uzorka ili primjenu drugih statističkih tehnika kako bi se dobila potpunija slika o temi.

Dodatci

Tablica 4: Sadržaj datoteke `p10.xlsx` u kojoj se nalaze dobiveni podatci - "Vrijeme (u danima) provedeno na Zavodu za zapošljavanje čekajući posao, uzorak 300 nezaposlenih"

Vrijeme (u danima)
15
17
10
9
18
20
25
18
16
9
23
13
18
19
15
14
12
17
7
9
14
6
5
12
19
11
15

17
14
6
5
12
19
11
15
14
12
17
7
9
14
6
5
12
19
11
15
17
14
6
19
11
15
14
12
17
7
9
14

6
5
12
19
11
15
17
15
17
14
6
19
11
15
14
12
17
7
9
14
6
5
12
19
11
15
17
5
17
13
4
16

17
14
23
19
11
15
14
13
18
17
22
17
1
8
7
28
1
27
15
22
9
2
25
12
8
18
10
8
17
0
10
4

14
8
28
22
20
24
8
4
11
28
9
1
19
26
0
16
16
12
4
26
14
17
4
25
13
28
9
20
12
18
18
14

25
14
9
17
0
13
24
5
8
5
4
20
20
14
27
7
2
3
12
1
14
9
19
11
28
6
8
16
21
2
26
0

3
28
3
5
24
16
18
15
12
29
7
6
13
24
22
23
28
23
3
26
12
30
27
11
13
25
24
29
30
9
10
9

25
18
10
8
15
8
1
25
29
3
29
5
12
15
3
13
23
17
13
26
7
4
5
16
10
20
19
10
15
7
4
6

11
27
30
17
15
19
15
4
9
25
2
21
24
24
15
2
28
11
4
8
8
24
5
18
9
10
17
10
19
27
0
19

8
28
6
2
22
0
25
25
26
18
13
13
23
26
16
13
5

Popis slika

Popis tablica

Popis ispisa kôda